P

Painlevé Equations

N Joshi, University of Sydney, Sydney, NSW, Australia

Introduction

The Painlevé equations P_I–P_{VI} are six classical second-order ordinary differential equations that appear widely in modern physical applications. Their conventional forms (governing $y(x)$ with derivatives $y' = dy/dx$, $y'' = d^2y/dx^2$) are:

$$P_I: \quad y'' = 6y^2 + x$$

$$P_{II}: \quad y'' = 2y^3 + xy + \alpha$$

$$P_{III}: \quad y'' = \frac{y'^2}{y} - \frac{y'}{x} + \frac{1}{x}(\alpha y^2 + \beta) + \gamma y^3 + \frac{\delta}{y}$$

$$P_{IV}: \quad y'' = \frac{y'^2}{2y} + \frac{3}{2}y^3 + 4xy^2 + 2(x^2 - \alpha)y + \frac{\beta}{y}$$

$$P_V: \quad y'' = \left(\frac{1}{2y} + \frac{1}{y-1}\right)y'^2 - \frac{y'}{x} + \frac{(y-1)^2}{x^2}\left(\alpha y + \frac{\beta}{y}\right) + \frac{\gamma y}{x} + \frac{\delta y(y+1)}{y-1}$$

$$P_{VI}: \quad y'' = \frac{1}{2}\left(\frac{1}{y} + \frac{1}{y-1} + \frac{1}{y-x}\right)y'^2 - \left(\frac{1}{x} + \frac{1}{x-1} + \frac{1}{y-x}\right)y' + \frac{y(y-1)(y-x)}{x^2(x-1)^2}\left\{\alpha + \frac{\beta x}{y^2} + \frac{\gamma(x-1)}{(y-1)^2} + \frac{\delta x(x-1)}{(y-x)^2}\right\}$$

where α, β, γ, δ are constants. They were identified and studied by Painlevé and his school in their search for ordinary differential equations (in the class $y'' = R(x, y, y')$, where R is rational in y', y and analytic in x) that define new transcendental functions. Painlevé focussed his search on equations that possess what is now known as the Painlevé property: that all solutions are single-valued around all movable singularities (a singularity is "movable" if its location changes with initial conditions).

For the Painlevé equations, all movable singularities are poles. For P_I and P_{II}, all solutions are meromorphic functions. However, the solutions of each of the remaining equations have other singularities called "fixed" singularities, with locations that are determined by the singularities of the coefficient functions of the equation. P_{III}–P_{VI} have a fixed singularity at $x = \infty$. P_{III} and P_V have additional fixed singularities at $x = 0$, and P_{VI} has them at $x = 0$ and 1. Although each solution of P_{III}–P_{VI} is single-valued around a movable singularity, it may be multivalued around a fixed singularity.

Painlevé's school considered canonical classes of ordinary differential equations equivalent under linear fractional transformations of y and x. Of the fifty canonical classes of equations they found, all except six were found to be solvable in terms of already known functions. These six lead to the Painlevé equations P_I–P_{VI} as their canonical representatives.

A resurgence of interest in the Painlevé equations came about from the observation (due to Ablowitz and Segur) that they arise as similarity reductions of well-known integrable partial differential equations (PDEs), or soliton equations, such as the Korteweg–de Vries equation, the sine-Gordon equation, and the self-dual Yang–Mills equations.

As this connection suggests, the Painlevé equations possess many of the special properties that are commonly associated with soliton equations. They have associated linear problems (i.e., Lax pairs) for which they act as compatibility conditions. There exist special transformations (called Bäcklund transformations) mapping a solution of one equation to a solution of another Painlevé equation (or the same equation with changed parameters). There exist Hamiltonian forms that are related to existence of tau-functions, that are analytic everywhere except at the fixed singularities. They also possess multilinear forms (or Hirota forms) that are satisfied by tau-functions. In the following subsections, for conciseness, we give examples of these properties for the first or second Painlevé equations and briefly indicate differences, in any, with other Painlevé equations.

Complex Analytic Structure of Solutions

Consider the two-(complex-)parameter manifold of solutions of a Painlevé equation. Each solution is globally determined by two initial values given at a regular point of the solution. However, the solution can also be determined by two pieces of data given at a movable pole. The location x_0 of such a pole provides one of the two free parameters. The other free parameter occurs as a coefficient in the Laurent expansion of the solution in a domain punctured at x_0. For P_I, the Laurent expansion of a solution at a movable singularity x_0 is

$$y(x) = \frac{1}{(x-x_0)^2} + \frac{x_0}{10}(x-x_0)^2 + \frac{1}{6}(x-x_0)^3 + c_I(x-x_0)^4 + \cdots \qquad [1]$$

where c_I is arbitrary. This second free parameter is normally called a "resonance parameter." For P_{II}, the Laurent expansion of a solution at a movable singularity x_0 is

$$y(x) = \frac{\pm 1}{(x-x_0)} + \frac{\mp x_0}{6}(x-x_0) + \frac{\mp 1-\alpha}{4}(x-x_0)^2 + c_{II}(x-x_0)^3 + \cdots \qquad [2]$$

where c_{II} is arbitrary. The symmetric solution of P_I that has a pole at the origin and corresponding resonance parameter $c_I = 0$ has a distribution of poles in the complex x-plane shown in Figure 1. (This figure was obtained by searching for zeros of truncated Taylor expansions of the tau-function τ_I described in the section "Bäcklund and Miura transformations." One hundred and sixty numerical zeros are shown. The two pairs of closely spaced zeros near the imaginary axis (between $8 < \pm\Im x < 12$) may be numerical artifacts. We used the command NSolve to 32 digits in MATHEMATICA4.)

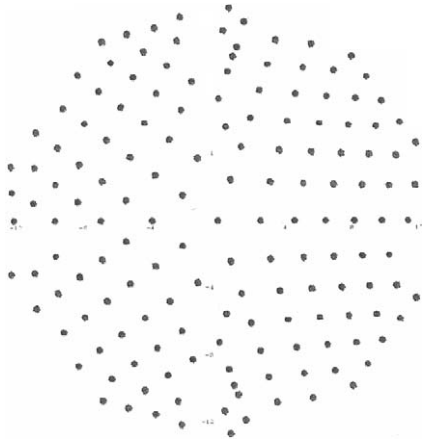


Figure 1 Poles of a symmetric solution of P_I in the complex x-plane, with a pole at the origin and zero corresponding resonance parameter, i.e., $x_0 = 0$, $c_I = 0$.

The rays of symmetry evident in Figure 1 reflect discrete symmetries of P_I. The solutions of P_I and P_{II} are invariant under the respective discrete symmetries,

$$P_I: \quad y_n(x) = e^{2\pi i n/5} y(e^{4\pi i n/5} x), \quad n = \pm 1, \pm 2$$

$$P_{II}: \quad y_n(x) = e^{\pi i n/3} y(e^{2\pi i n/3} x), \quad \alpha \mapsto e^{-\pi i n}\alpha, \quad n = \pm 1, \pm 2, 3$$

The rays of angle $2\pi n/5$ for P_I and $\pi n/3$ for P_{II} related to these symmetries play special roles in the asymptotic behaviors of the corresponding solutions for $|x| \to \infty$.

Linear Problems

The Painlevé equations are regarded as completely integrable because they can be solved through an associated system of linear equations (Jimbo and Miwa 1981).

$$\frac{d\varphi}{d\zeta} = L(x,\zeta)\varphi \qquad [3a]$$

$$\frac{d\varphi}{dx} = M(x,\zeta)\varphi \qquad [3b]$$

The compatibility condition, that is,

$$L_x - M_\zeta + [L, M] = 0 \qquad [4]$$

is equivalent to the corresponding Painlevé equation. The matrices L, M for P_I and P_{II} are listed below:

$$P_I: \quad L_I(x,\zeta) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\zeta^2 + \begin{pmatrix} 0 & y \\ 4 & 0 \end{pmatrix}\zeta + \begin{pmatrix} -z & y^2 + x/2 \\ -4y & z \end{pmatrix}$$

$$M_I(x,\zeta) = \begin{pmatrix} 0 & 1/2 \\ 0 & 0 \end{pmatrix}\zeta + \begin{pmatrix} 0 & y \\ 2 & 0 \end{pmatrix}$$

$$\text{where} \quad z = y', \quad z' = 6y^2 + x$$

$$P_{II}: \quad L_{II}(x,\zeta) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\zeta^2 + \begin{pmatrix} 0 & u \\ -2z/u & 0 \end{pmatrix}\zeta + \begin{pmatrix} z + x/2 & -uy \\ -2(\vartheta + zy)/u & -(z + x/2) \end{pmatrix}$$

$$M_{II}(x,\zeta) = \begin{pmatrix} 1/2 & 0 \\ 0 & -1/2 \end{pmatrix}\zeta + \begin{pmatrix} 0 & u/2 \\ -z/u & 0 \end{pmatrix}$$

$$\text{where} \quad u' = -uy, \quad z = y' - y^2 - x/2$$

$$\vartheta := \frac{1}{2} - \alpha$$

Alternative linear problems also exist for each equation. For example, for P_{II}, an alternative choice of L and M is (Flaschka and Newell 1980):

$$P_{II}: \quad L_{II'}(x,\zeta) = \begin{pmatrix} -4\,i & 0 \\ 0 & 4\,i \end{pmatrix}\zeta^2 + \begin{pmatrix} 0 & 4y \\ 4y & 0 \end{pmatrix}\zeta + \begin{pmatrix} -i(x+2y^2) & 2\,iy' \\ -2\,iy' & i(x+2y^2) \end{pmatrix} + \frac{\alpha}{\zeta}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$M_{II'}(x,\zeta) = \begin{pmatrix} -i\zeta & y \\ y & i\zeta \end{pmatrix}$$

The matrix L for each Painlevé equation is singular at a finite number of points $a_i(x)$ in the ζ-plane. For the above choices of L for P_I and P_{II}, the point $\zeta = \infty$ is clearly a singularity. For $L_{II'}$, the origin $\zeta = 0$ is also a singularity. The analytic continuation of a fundamental matrix of solutions Φ around a_i gives a new solution $\widetilde{\Phi}$ which must be related to the original solution: $\widetilde{\Phi} = \Phi A$. A is called the monodromy matrix and its trace and determinant are called the monodromy data. In general, the data will change with x. However, eqn [4] ensures that the monodromy data remain constant in x. For this reason, the system [3] is called an isomonodromy problem.

Bäcklund and Miura Transformations

Bäcklund transformations are those that map a solution of a Painlevé equation with one choice of parameter to a solution of the same equation with different parameters. For P_I no such transformation is known. For P_{II}, there is one Bäcklund transformation. Let $y = y(x;\alpha)$ denote a solution of P_{II} with parameter α. Then $\tilde{y} = y(x;\alpha-1)$, which solves P_{II} with parameter $\alpha - 1$, is given by

$$\tilde{y} := -y + \frac{\alpha - \frac{1}{2}}{y' - y^2 - x/2} \quad \text{if } \alpha \neq 1/2 \qquad [5]$$

If $\alpha = 1/2$, then $y' = y^2 + x/2$ and $\tilde{y} = -y$ (see the next section for this case). Combined with the symmetry $y \mapsto -y, \alpha = -\alpha$, we can write down another version of this Bäcklund transformation which maps y to $\hat{y} = y(x;\alpha+1)$:

$$\hat{y} := -y - \frac{\alpha + \frac{1}{2}}{y' + y^2 + x/2}, \quad \text{if } \alpha \neq -\frac{1}{2} \qquad [6]$$

If we parametrize α by $c + n$ for arbitrary c, and denote the solution for corresponding parameter as y_n, we can write a difference equation relating y_{n-1} and y_{n+1} (by eliminating y' from the two transformations $\tilde{y}$, $\hat{y}$) as

$$\frac{c + \frac{1}{2} + n}{y_{n+1} + y_n} + \frac{c - \frac{1}{2} + n}{y_{n-1} + y_n} + 2y_n^2 + x = 0$$

This is an example of a discrete Painlevé equation (called "alternate" dP_I in the literature). In such a discrete Painlevé equation, x is fixed while n varies. Another lesser known Bäcklund transformation for P_{II} is

$$y' - y^2 - \frac{x}{2} - \delta v^2 = 0 \qquad [7]$$

$$v' + y v = 0 \qquad [8]$$

between P_{II} with $\alpha = 1/2$ and

$$v'' + \delta v^3 + \frac{x}{2} v = 0$$

which can be scaled (take $v(x) = y(\sqrt{2}x)/\sqrt{\sqrt{2}\delta}$) to the usual form of P_{II} with $\alpha = 0$.

Miura transformations are those that map a solution of a Painlevé equation to another equation in the 50 canonical types classified by Painlevé's school. If y is a solution of P_{II} with parameter $\alpha \neq 1/2$, then

$$(2\alpha - 1)w = 2(y' - y^2 - x/2), \quad y = \frac{1 - w'}{2w}$$

maps between P_{II} and

$$w'' = \frac{(w')^2}{2w} - (2\alpha - 1)\,w^2 - xw - \frac{1}{2w}$$

which represents the 34th canonical class in the Painlevé classification listed in Ince (1927).

The Painlevé equations do not possess continuous symmetries other than Bäcklund and Miura transformations described here. However, they do possess discrete symmetries described in the section "Complex analytic structure of solutions."

Classical Special Solutions

Painlevé showed that there can be no explicit first integral that is rational in y and y' for his eponymous equations. It is known that this statement can be extended to say that no such algebraic first integral exists. But the question whether the Painlevé equations define new transcendental functions remained open until recently.

Form a class of functions consisting of those satisfying linear second-order differential equations, such as the Airy, Bessel, and hypergeometric functions, as well as rational, algebraic, and exponential functions. Extend this class to include arithmetic operations, compositions under such functions, and

solutions of linear equations with these earlier functions as coefficients. Members of this class are called classical functions. For general values of the constants $\alpha, \beta, \gamma, \delta$, it is now known (Umemura 1990, Umemura and Watanabe 1997) that the six Painlevé equations cannot be solved in terms of classical functions. However, there are special values of the constant parameters α, β, γ, δ for which classical functions do solve the Painlevé equations. Each Painlevé equation, except P_I, has special solutions given by classical functions when the parameters in the Painlevé equation take on special values. For P_{II}, with $\alpha = 1/2$ we have the special integral

$$I_{1/2} \equiv y' - y^2 - \frac{x}{2} = 0 \qquad [9]$$

which, modulo P_{II} with $\alpha = 1/2$, satisfies the relation

$$\left(\frac{\mathrm{d}}{\mathrm{d}x} + 2y\right) I_{1/2} = 0$$

The Riccati eqn [9] can be linearized via $y = -\psi'/\psi$ to yield

$$\psi'' + \frac{x}{2}\psi = 0$$

which gives

$$\psi(x) = a\,\mathrm{Ai}(-2^{-1/3}x) + b\,\mathrm{Bi}(-2^{-1/3}x)$$

for arbitrary constants a and b, that is, the well-known Airy function solutions of P_{II}. Iterations of the Bäcklund transformations $\tilde{y}$ and $\hat{y}$, [5]–[6] give further classical solutions in terms of Airy functions for the case when $\alpha = (2N+1)/2$ for integer N.

Similarly, there is a sequence of rational solutions of the family of equations P_{II} with $\alpha = N$, for integer N, if we iterate the Bäcklund transformations $\tilde{y}, \hat{y}$ by starting with the trivial solution $y \equiv 0$ for the case $\alpha = 0$. For example, for $\alpha = 1$, we have $\hat{y} = -1/x$. The transformations [7]–[8] give a mapping that shows that this family of rational solutions and the above family of Airy-type solutions of P_{II} both exist for the cases when α is half-integer and when it is integer.

Hamiltonians and Tau-Functions

Each Painlevé equation has a Hamiltonian form. For P_I and P_{II}, these can be found by integrating each equation after multiplying by y'. These give

$$P_I: \qquad \frac{y'^2}{2} = 2y^3 + xy - \int^x y(\xi)\mathrm{d}\xi + E_I$$

$$P_{II}: \qquad \frac{y'^2}{2} = \frac{y^4}{2} + \frac{x}{2}y^2 - \frac{1}{2}\int^x y(\xi)^2\mathrm{d}\xi + \alpha y + E_{II}$$

where E_{II} and E_{II} are constants. We choose canonical variables $q_1(t) = y(x), p_1(t) = y'(x)$, where $t = x$. Furthermore, for P_I, we take

$$q_2(t) = x, \qquad p_2(t) = \int^x y(\xi)\mathrm{d}\xi$$

and the Hamiltonian

$$H_I := \frac{p_1^2}{2} - 2{q_1}^3 - q_2 q_1 + p_2$$

so that the Hamiltonian equations of motion $\dot{q}_i = \partial H/\partial p_i$ and $\dot{p}_i = -\partial H/\partial q_i$ are satisfied. For P_{II}, we take

$$q_2(t) = x/2, \qquad p_2(t) = \int^x y(\xi)^2\mathrm{d}\xi$$

and the Hamiltonian

$$H_{II} := \frac{{p_1}^2}{2} - \frac{{q_1}^4}{2} - q_2{q_1}^2 + \frac{1}{2}p_2 - \alpha q_1$$

We note that these Hamiltonians govern systems with two degrees of freedom and each is conserved. However, no explicit second conserved quantity is known (see comments on first integrals in the last section).

Painlevé's viewpoint of the transcendental solutions of the Painlevé equations as natural generalizations of elliptic functions also led him to search for entire functions that play the role of theta functions in this new setting. He found that analogous functions could be defined which have only zeros at the locations of the movable singularities of the Painlevé transcendents. These functions are now commonly known as tau-functions (also denoted τ-functions). For P_I and P_{II}, the corresponding tau-functions are entire functions (i.e., they are analytic everywhere in the complex x-plane). However, for the remaining Painlevé equations, they are singular at the fixed singularities of the respective equation.

For P_I, all movable singularities of P_I are double poles of strength unity (see eqn [1]). Therefore, the function given by

$$P_I: \quad \tau_I(x) = \exp\left(-\int^x\int^s y(t)\mathrm{d}t\mathrm{d}s\right)$$

has Taylor expansion with leading term $(x - x_0)$. In other words, $\tau_I(x)$ is analytic at all the poles of the corresponding solution of P_I. Since $y(x)$ has no other singularity (other than at infinity), $\tau_I(x)$ must be analytic everywhere in the complex x-plane. Differentiation and substitution of P_I shows that $\tau_I(x)$ satisfies the fourth-order equation

$$P_I: \quad \tau_I^{(4)}(x)\tau_I(x) = 4\tau_I'(x)\tau_I^{(3)}(x) - 3\tau_I''(x)^2 - x\tau_I(x)^2$$

Note that this equation is bilinear in τ and its derivatives. Such bilinear, or in general, multilinear, equations are called Hirota-type forms of the Painlevé equations. The special nature of such equations is most simply expressed in terms of the Hirota $D(\equiv D_x)$ operator, an antisymmetric differential operator defined here on products of functions of x:

$$D^n f \cdot g = (\partial_\xi - \partial_\eta)^n f(\xi) g(\eta)|_{\xi=\eta=x}$$

Notice that

$$D^2\tau\cdot\tau = \tau\tau'' - \tau'^2,$$
$$D^4\tau\cdot\tau = \tau\tau^{(4)} - 4\tau'\tau^{(3)} + 3\tau''^2$$

Hence the equation satisfied by $\tau_{\mathrm{I}}(x)$ can be rewritten more succinctly as

$$(D^4 + x)\tau_{\mathrm{I}}\cdot\tau_{\mathrm{I}} = 0$$

For $\mathrm{P_{II}}$, a generic solution $y(x)$ has movable simple poles of residue ± 1 (see eqn [2]). Painlevé pointed out that if we square the function $y(x)$, multiply by -1 and integrate twice, we obtain a function with Taylor expansion with leading term $(x - x_0)$. However, the square is not invertible and to construct an invertible mapping to entire functions, we need two τ-functions. We denote these by $\tau(x)$ and $\sigma(x)$:

$$\mathrm{P_{II}}: \qquad \tau_{\mathrm{II}}(x) = \exp\left(-\int^x\int^s y(t)^2\,\mathrm{d}t\,\mathrm{d}s\right)$$
$$\sigma_{\mathrm{II}}(x) = y(x)\tau_{\mathrm{II}}(x)$$

The equations satisfied by these tau-functions are

$$\mathrm{P_{II}}: \qquad \tau''(x)\tau(x) = \tau'(x)^2 - \sigma(x)^2$$
$$\sigma''(x)\tau(x)^2 = 2\tau(x)\tau'(x)\sigma'(x) - \tau'(x)^2\tau(x) + \sigma(x)^3 + x\tau(x)^2\sigma(x) + \alpha\tau(x)^3$$

Hierarchies

Each Painlevé equation is associated with at least one infinite sequence of ordinary differential equations (ODEs) indexed by order. These sequences are called hierarchies and arise from symmetry reductions of PDE hierarchies that are associated with soliton equations.

Define the operator $\mathcal{L}_n\{v(z)\}$ (the Lenard recursion operator) recursively by

$$\frac{\mathrm{d}}{\mathrm{d}z}\mathcal{L}_{n+1}\{v\} = \left(\frac{\mathrm{d}^3}{\mathrm{d}z^3} + 4v\frac{\mathrm{d}}{\mathrm{d}z} + 2v'\right)\mathcal{L}_n\{v\}$$
$$\mathcal{L}_1\{v\} = v$$

where primes denote z-derivatives. Note that

$$\mathcal{L}_2\{v\} = v'' + 3v^2$$
$$\mathcal{L}_3\{v\} = v^{(4)} + 10vv'' + 5v'^2 + 10v^3$$

This operator is intimately related to the Korteweg–de Vries equation. (It was first discovered as a method of generating the infinite number conservation laws associated with this soliton equation.)

The scaling $v(z) = \lambda y(\mu x)$, with $\lambda = (-2)^{1/3}$, $\mu = (-2)^{-1/3}$, shows that the case $n = 2$ of the sequence of ODEs defined recursively by

$$\mathcal{L}_n\{v\} = z$$

is $\mathrm{P_I}$. Hence this is called the first Painlevé hierarchy.

A second Painlevé hierarchy is given recursively by

$$\left(\frac{\mathrm{d}}{\mathrm{d}x} + 2y\right)\mathcal{L}_n\{y' - y^2\} = xy + \alpha_n, \quad n \geq 1$$

where α_n are constants.

Each Painlevé equation may arise as a reduction of more than one PDE. Since different soliton equations have different hierarchies, this means that more than one hierarchy may be associated with each Painlevé equation.

See also: Bäcklund Transformations; Integrable Discrete Systems; Integrable Systems: Overview; Isomonodromic Deformations; Ordinary Special Functions; Riemann–Hilbert Methods in Integrable Systems; Riemann–Hilbert Problem; Solitons and Kac–Moody Lie Algebras; Two-Dimensional Ising Model; WDVV Equations and Frobenius Manifolds.

Further Reading

Flaschka H and Newell AC (1980) Monodromy- and spectrum-preserving deformations. *Communications in Mathematical Physics* 76: 65–116.

Ince EL (1927) *Ordinary Differential Equations*, London: Longmans, Green and Co. (Reprinted in 1956 – New York: Dover).

Jimbo M and Miwa T (1981) Monodromy preserving deformation of linear ordinary differential equations with rational coefficients. II. *Physica D* 2: 407–448.

Umemura H (1990) Second proof of the irreducibility of the first differential equation of Painlevé. *Nagoya Mathematical Journal* 117: 125–171.

Umemura H and Watanabe H (1997) Solutions of the second and fourth Painlevé equations. I. *Nagoya Mathematical Journal* 148: 151–198.

Partial Differential Equations: Some Examples

R Temam, Indiana University, Bloomington, IN, USA

Introduction

Many physical laws are mathematically expressed in terms of partial differential equations (PDEs); this is, for instance, the case in the realm of classical mechanics and physics of the laws of conservation of angular momentum, mass, and energy.

The object of this short article is to provide an overview and make a few comments on the set of PDEs appearing in classical mechanics, which is tremendously rich and diverse. From the mathematical point of view the PDEs appearing in mechanics range from well-understood PDEs to equations which are still at the frontier of sciences as far as their mathematical theory is concerned. The mathematical theory of PDEs deals primarily with their "well-posedness" in the sense of Hadamard. A well-posed PDE problem is a problem for which existence and uniqueness of solutions in suitable function spaces and continuous dependence on the data have been proved.

For simplicity, let us restrict ourselves to space dimension 2. Several interesting and important PDEs are of the form

$$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} = 0 \qquad [1]$$

Here a, b, c may depend on x and y or they may be constants, and then eqn [1] is linear: they may also depend on u, $\partial u/\partial x$, and $\partial u/\partial y$, in which case the equation is nonlinear.

Such an equation is

- elliptic when (where) $b^2 - 4ac < 0$,
- hyperbolic when (where) $b^2 - 4ac > 0$,
- parabolic when (where) $b^2 - 4ac = 0$.

Among the simplest linear equations, we have the elliptic equation

$$\Delta u = 0 \qquad [2]$$

which governs the following phenomena: equation for the potential or stream function of plane, incompressible irrotational fluids; equation for some potential in linear elasticity, or the equation for the temperature in suitable conditions (stationary case; see below for the time-dependent case).

Another eqn of the form [1] is the hyperbolic equation

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0 \qquad [3]$$

which governs, for example, linear acoustics in one dimension (sound pipes) or the propagation of an elastic wave along an elastic string.

A third equation of type [1] is the linear parabolic equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \qquad [4]$$

also called the heat equation, which governs, under appropriate circumstances, the temperature ($u(x,t) =$ temperature at x at time t).

All these equations are well understood from the mathematical viewpoint and many well-posedness results are available. A fundamental difference between eqns [2], [3], and [4] is that for [2] and [4] the solution is as smooth as allowed by the data (forcing terms, boundary data not mentioned here), whereas the solutions of [3] usually present some discontinuities corresponding to the propagation of a wave or wave front.

A considerable jump of complexity occurs if we consider the equation of transonic flows in which

$$\begin{aligned} a &= \left(1 - \frac{1}{v^2}\left(\frac{\partial u}{\partial x}\right)^2\right) \\ b &= -\frac{2}{v^2}\frac{\partial u}{\partial x}\frac{\partial u}{\partial y} \\ c &= 1 - \frac{1}{v^2}\left(\frac{\partial u}{\partial y}\right)^2 \end{aligned} \qquad [5]$$

where $v = v(x,y)$ is the local speed of sound. This is a mixed second-order equation: it is elliptic in the subsonic region where $M < 1, M$ the Mach number being the ratio of the velocity

$$|\text{grad } u| = \left(\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right)^{1/2}$$

to the local velocity of sound $v = v(x,y)$; eqn [1] (with [5]) is hyperbolic in the supersonic region, where $M > 1$ and parabolic on the sonic line $M = 1$. Essentially no result of well-posedness is available for this problem, and it is not even totally clear what are the boundary conditions that one should associate to [1]–[5] to obtain a well-posed problem.

Intermediate mathematical situations are encountered with the Navier–Stokes and Euler equations, which govern the motion of fluids in the viscous and inviscid cases, respectively. A number of mathematical results are available for these equations (*see* Compressible Flows: Mathematical Theory, Incompressible Euler Equations: Mathematical Theory, Viscous Incompressible Fluids: Mathematical Theory, Inviscid Flows); but other questions are still open, including the famous Clay prize problem, which is: to show that the solutions of the (viscous, incompressible) Navier–Stokes equations, in space dimension three, remain smooth for all time, or to exhibit an example of appearance of singularity. A prize of US$ 1 million will be awarded by the Clay Foundation for the solution of this problem.

For compressible fluids, the Navier–Stokes equations expressing conservation of angular momentum and mass read

$$\rho\left(\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u}\right) - \mu\Delta\boldsymbol{u} + \nabla p - (\lambda+\mu)\nabla(\nabla\cdot\boldsymbol{u}) = 0 \qquad [6]$$

$$\frac{\partial \rho}{\partial t} + \nabla(\rho\boldsymbol{u}) = 0 \qquad [7]$$

Here $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x}, t)$ is the velocity at $\boldsymbol{x}$ at time t, $p = p(\boldsymbol{x}, t)$ the pressure, ρ the density; λ, μ are viscosity coefficients, $\mu > 0, 3\lambda + 2\mu \geq 0$. When $\mu = \lambda = 0$, we obtain the Euler equation (*see* Compressible Flows: Mathematical Theory). If the fluid is incompressible and homogeneous, then the density is constant, $\rho = \rho_0$ and

$$\nabla\cdot\boldsymbol{u} = 0 \qquad [8]$$

so that eqn [8] replaces eqn [7] and eqn [6] simplifies accordingly.

Finally, let us mention still different nonlinear PDEs corresponding to nonlinear wave phenomena, namely the Korteweg–de Vries (*see* Korteweg–de Vries Equation and Other Modulation Equations)

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0 \qquad [9]$$

and the nonlinear Schrödinger equation (*see* Nonlinear Schrödinger Equations)

$$\frac{\partial u}{\partial z} + i\beta\frac{\partial^2 A}{\partial t^2} - \mathrm{i}\,\gamma|A|^2 A + \alpha A = 0 \qquad [10]$$

$\alpha, \gamma > 0$. These equations are very different from eqns [1]–[8] and are reasonably well understood from the mathematical point of view; they produce and describe the amazing physical wave phenomenon known as the soliton (*see* Solitons and Kac–Moody Lie Algebras).

This article is based on the Appendix of the book by Miranville and Temam quoted below, with the authorization of Cambridge University Press.

See also: Compressible Flows: Mathematical Theory; Elliptic Differential Equations: Linear Theory; Evolution Equations: Linear and Nonlinear; Fluid Mechanics: Numerical Methods; Fractal Dimensions in Dynamics; Image Processing: Mathematics; Incompressible Euler Equations: Mathematical Theory; Integrable Systems and the Inverse Scattering Method; Interfaces and Multicomponent Fluids; Inviscid Flows; Korteweg–de Vries Equation and Other Modulation Equations; Leray–Schauder Theory and Mapping Degree; Magnetohydrodynamics; Newtonian Fluids and Thermohydraulics; Nonlinear Schrödinger Equations; Solitons and Kac–Moody Lie Algebras; Stochastic Hydrodynamics; Symmetric Hyperbolic Systems and Shock Waves; Viscous Incompressible Fluids: Mathematical Theory; Non-Newtonian Fluids.

Further Reading

Brezis H and Browder F (1998) Partial differential equations in the 20th century. *Advances in Mathematics*, 135: 76–144.

Evans LC (1998) *Partial Differential Equations*. Providence, RI: American Mathematical Society.

Miranville A and Temam R (2001) *Mathematical Modelling in Continuum Mechanics*. Cambridge: Cambridge University Press.

Path Integral Methods *see* Functional Integration in Quantum Physics; Feynman Path Integrals

Path Integrals in Noncommutative Geometry

R Léandre, Université de Bourgogne, Dijon, France

Introduction

Let us recall that there are basically two algebraic infinite-dimensional distribution theories:

- The first one is white-noise analysis (Hida *et al.* 1993, Berezansky and Kondratiev 1995), and uses Fock spaces and the algebra of creation and annihilation operators.
- The second one is the noncommutative differential geometry of Connes (1988) and uses the entire cyclic complex.

If we disregard the differential operations, these two distribution theories are very similar. Let us recall quickly their background on geometrical examples. Let V be a compact Riemannian manifold and E a Hermitian bundle on it. We consider an elliptic Laplacian Δ_E acting on sections ω of this bundle. We consider the Sobolev space $H_k, k > 0$, of sections ω of E such that:

$$\int_V \left\langle \left(\Delta_E^k + 1\right)\omega, \omega \right\rangle \mathrm{d}m_V < \infty \qquad [1]$$

where $\mathrm{d}m_V$ is the Riemannian measure on V and $\langle , \rangle$ the Hermitian structure on V. H_{k+1} is included in H_k and the intersection of all H_k is nothing other than the space of smooth sections of the bundle E, by the Sobolev embedding theorem.

Let us quickly recall Connes' distribution theory: let $\alpha(n)$ be a sequence of real strictly positive numbers. Let

$$\sigma = \sum \sigma_n \qquad [2]$$

where σ_n belongs to $H_k^{\otimes n}$ with the Hilbert structure naturally inherited from the Hilbert structure of H_k. We put, for $C > 0$,

$$\|\sigma\|_{1,C,k} = \sum C^n \alpha(n) \|\sigma_n\|_{H_k^{\otimes n}} \qquad [3]$$

The set of σ such that $\|\sigma\|_{1,C,k} < \infty$ is a Banach space called $\mathrm{Co}_{C,k}$. The space of Connes functionals $\mathrm{Co}_{\infty -}$ is the intersection of these Banach spaces for $C > 0$ and $k > 0$ endowed with its natural topology. Its topological dual $\mathrm{Co}_{-\infty}$ is the space of distributions in Connes' sense.

Remark We do not give the original version of the space of Connes where tensor products of Banach algebras appear but we use here the presentation of Jones and Léandre (1991).

Let us now quickly recall the theory of distributions in the white-noise sense. The main tools are Fock spaces. We consider interacting Fock spaces (Accardi and Bożejko (1998)) constituted of σ written as in [2] such that

$$\|\sigma\|_{2,C,k}^2 = \sum C^n \alpha(n)^2 \|\sigma_n\|_{H_k^{\otimes n}}^2 < \infty \qquad [4]$$

The space of white-noise functionals $\mathrm{WN}_{\infty -}$ is the intersection of these interacting Fock spaces $\Lambda_{k,C}$ for $C > 0, k > 0$. Its topological dual $\mathrm{WN}_{-\infty}$ is called the space of white-noise distributions.

Traditionally, in white-noise analysis, one considers in [2] the case where σ_n belongs to the symmetric tensor product of H_k endowed with its natural Hilbert structure. We get a symmetric Fock space $\Lambda_{C,k}^s$ and another space of white-noise distributions $\mathrm{WN}_{s,-\infty}$. The interest in considering symmetric Fock spaces, instead of interacting Fock spaces, arises from the characterization theorem of Potthoff–Streit. For the sake of simplicity, let us consider the case where $\alpha(n) = 1$. If ω if a smooth section of E, we can consider its exponential $\exp[\omega] = \sum n!^{-1} \omega^{\otimes n}$. If we consider an element Φ of $\mathrm{WN}_{s,-\infty}$, $\langle \Phi, \exp[\omega] \rangle$ satisfies two natural conditions:

1. $|\langle \Phi, \exp[\omega] \rangle| \leq C \exp[C\|\omega\|_{H_k}^2]$ for some $k > 0$.
2. $z \to \langle \Phi, \exp[\omega_1 + z\omega_2] \rangle$ is entire.

The Potthoff–Streit theorem states the opposite: a functional which sends a smooth section of V into a Hilbert space and which satisfies the two previous requirements defines an element of $\mathrm{WN}_{s,-\infty}$ with values in this Hilbert space. Moreover, if the functional depends holomorphically on a complex parameter, then the distribution depends holomorphically on this complex parameter as well.

The Potthoff–Streit theorem allows us to define flat Feynman path integrals as distributions. It is the opposite point of view, from the traditional point of view of physicists, where generally path integrals are defined by convergence of the finite-dimensional lattice approximations. Hida–Streit have proposed replacing the approach of physicists by defining path integrals as infinite-dimensional distributions, and by using Wiener chaos. Getzler was the first who thought of replacing Wiener chaos by other functionals on path spaces, that is, Chen iterated integrals. In this article, we review the recent

developments of path integrals in this framework. We will mention the following topics:

- infinite-dimensional volume element
- Feynman path integral on a manifold
- Bismut–Chern character and path integrals
- fermionic Brownian motion

The reader who is interested in various rigorous approaches to path integrals should consult the review of Albeverio (1996).

Infinite-Dimensional Volume Element

Let us recall that the Lebesgue measure does not exist generally as a measure in infinite dimensions. For instance, the Haar measure on a topological group exists if and only if the topological group is locally compact. Our purpose in this section is to define the Lebesgue measure as a distribution.

We consider the set $C^\infty(M;N)$ of smooth maps $x(.)$ from a compact Riemannian manifold M into a compact Riemannian manifold N endowed with its natural Fréchet topology. S is the generic point of M and x the generic point of N. We would like to say that the law of $x(S_i)$ for a finite set of n different points S_i under the formal Lebesgue measure $\mathrm{d}D(x(.))$ on $C^\infty(M;N)$ is the product law of $n\,\mathrm{d}m_N$ (This means that the Lebesgue measure on $C^\infty(M;N)$ is a cylindrical measure). Let us consider a smooth function σ_n from $(M\times N)^n$ into C. We introduce the associated functional $F(\sigma_n)(x(.))$ on $C^\infty(M;N)$:

$$F(\sigma_n)(x(.)) = \int_{M^n} \sigma_n(S_1,\ldots,S_n,x(S_1),\ldots,x(S_n))\mathrm{d}m_{M^n} \quad [5]$$

If we use formally the Fubini formula, we get

$$\int_{C^\infty(M;N)} F(\sigma_n)(x(.))\mathrm{d}D(x(.)) = \int_{M^n\times N^n} F(S_1,\ldots,S_n,x_1,\ldots,x_n)\mathrm{d}m_{M^n\times N^n} \quad [6]$$

We will interpret this formal remark in the framework of the distribution theories of the introduction. We consider $V = M\times N$ and E the trivial complex line bundle endowed with the trivial metric and $\alpha(n)=1$. We can define the associated algebraic spaces $\mathrm{Co}_{-\infty}$ and $\mathrm{WN}_{-\infty}$ and we can extend to $\mathrm{Co}_{\infty-}$ and $\mathrm{WN}_{\infty-}$ the map F of [5]. F sends elements of $\mathrm{Co}_{\infty-}$ and $\mathrm{WN}_{\infty-}$ into the set of continuous bounded maps of $C^\infty(M;N)$ where we can extend [6]. We obtain:

Theorem 1 *$\sigma \to \int_{C^\infty(M;N)} F(\sigma)(x(.))\mathrm{d}D(x(.))$ defines an element of $\mathrm{Co}_{-\infty}$ or $\mathrm{WN}_{-\infty}$.*

Feynman Path Integral on a Manifold

Let us introduce the flat Brownian motion $s\to B(s)$ in $\boldsymbol{R}^d$ starting from 0. It has formally the Gaussian law

$$Z^{-1}\exp\left[-\frac{1}{2}\int_0^1\left|\frac{\mathrm{d}}{\mathrm{d}s}B(s)\right|^2\mathrm{d}s\right]\mathrm{d}D(B(.))$$

where $\mathrm{d}D(B(.))$ is the formal Lebesgue measure on finite-energy paths starting from 0 in $\boldsymbol{R}^d$ (the partition function Z is infinite!). Let N be a compact Riemannian manifold of dimension d endowed with the Levi–Civita connection. The stochastic parallel transport on semimartingales for the Levi-Civita connection exists almost surely (Ikeda and Watanabe 1981). Let us introduce the Laplace–Beltrami operator Δ_N on N and the Eells–Elworthy–Malliavin equation starting from x (Ikeda and Watanabe 1981):

$$\mathrm{d}x_s(x) = \tau_s(x)\mathrm{d}B(s) \quad [7]$$

where $B(.)$ is a Brownian motion in $T_x(M)$ starting from 0 and $s\to\tau_s(x)$ is the stochastic parallel transport associated to the solution. $s\to x_s(x)$ is called the Brownian motion on N. The heat semigroup associated to Δ_N satisfies $\exp[-t\Delta_N]f(x) = E[f(x_t(x))]$ for f continuous on N. Formally, there is a Jacobian which appears in the transformation of the formal path integral which governs $B(.)$ into the formal path integral which governs $x_.(x)$

$$\mathrm{d}\mu_x(1) = Z_x^{-1}\exp[-I(x_.(x))/2]\mathrm{d}D(x_.(x)) \quad [8]$$

It was shown by B DeWitt, in a formal way, that the action in [8] is not the energy of the path and that there are some counter-terms in the action where the scalar curvature K of N appears (see Andersson and Driver (1999) and Sidorova *et al.* (2004) for rigorous results). In order to describe Feynman path integrals, we perform, as it is classical in physics, analytic continuation on the semigroup and on the "measure" $\mathrm{d}\mu_x(1)$ such that we get a distribution $\mathrm{d}\mu_x(\alpha)$ which depends holomorphically on $\alpha, \mathrm{Re}\,\alpha\geq 0$.

In order to return to the formalism of the introduction, we consider $V=N, E$ the trivial complex line bundle and the symmetric Fock space and $\alpha(n)=1$. To $\sigma_n/n!$ belonging to $H_k^{\otimes_{\mathrm{sym}}n}$ we associate the functional on $P(N)$, the smooth path space on N:

$$F(\sigma_n/n!)(x(.)) = \int_{\Delta_n}\sigma_n(x(s_1),\ldots,x(s_n))\mathrm{d}s_1\cdots\mathrm{d}s_n \quad [9]$$

where Δ_n is the n-dimensional simplex of $[0,1]^n$ constituted of times $0<s_1<\cdots<s_n<1$ (Léandre (2003)). We remark that F maps $\mathrm{WN}_{s,\infty-}$ into the set of bounded continuous functionals on $P(N)$. We

introduce an element h of $L^2(N)$. The map which to ω, a smooth function on N, associates $\exp[\alpha(\Delta_N+\omega)]h(\mathrm{Re}\,\alpha\leq 0)$ satisfies the requirements (1) and (2) of the introduction and depends holomorphically on α. This defines by the Potthoff–Streit theorem a distribution Φ_α which depends holomorphically on α, $\mathrm{Re}\,\alpha\leq 0$ with values in $L^2(N)$. By uniqueness of analytic continuation, we obtain:

Theorem 2 *If $P_x(N)$ is the space of smooth paths starting from x in N, we have*

$$\langle\Phi_\alpha,\sigma\rangle=\left\{x\to\int_{P_x(N)}F(\sigma)h(x(1))\mathrm{d}\mu_x(-\alpha)\right\} \quad [10]$$

Instead of taking functions, we can consider as bundle E the space of complex 1-forms on N. We then consider Chen (1973) iterated integrals:

$$F(\sigma_n)(x(.))=\int_{\Delta_n}\langle\sigma_n(x(s_1),\ldots,x(s_n)),\mathrm{d}x(s_1),\ldots,\mathrm{d}x(s_n)\rangle \quad [11]$$

such that F maps $\mathrm{WN}_{s,\infty-}$ into the set of measurable maps on $P(N)$. These maps are generally not bounded. Namely,

$$F(\exp[\omega])=\exp\left[\int_0^1\langle\omega(x(s)),\mathrm{d}x(s)\rangle\right] \quad [12]$$

instead of $\exp[\int_0^1\omega(x(s))\mathrm{d}s]$ in the previous case. By using the Cameron–Martin–Girsanov–Maruyama formula and Kato perturbation theory, we get an analog of Theorem 2 for Chen iterated integrals, but for $\mathrm{Re}\,\alpha<0$, because we have to deal with a perturbation of Δ_N by a drift when we want to check (1) and (2). The interest of this formalism is that the parallel transport belongs in some sense to the domain of the distribution and that we get the flat Feynman path integral from the curved one by using an analog of [7].

Bismut–Chern Character and Path Integrals

Since we are concerned in this part with index theory, we replace the free path space of N by the free smooth loop space $L(N)$. We consider the case where $V=N$ is a compact oriented Riemannian spin manifold and $E=E_-\oplus E_+$. E_- is the bundle of complexified odd forms and E_+ is the bundle of complexified even forms. To $\sigma_n=n!^{-1}(\omega_1+\omega_1^1)\otimes\cdots\otimes(\omega_n+\omega_n^1)$, we associate the even Chen (1973) iterated integral

$$F(\sigma_n)=\int_{\Delta_n}\left(\omega_1(\mathrm{d}x(s_1),.)+\omega_1^1\mathrm{d}s_1\right)\wedge\cdots\wedge\left(\omega_n(\mathrm{d}x(s_n),.)+\omega_n^1\mathrm{d}s_n\right) \quad [13]$$

where $s\to x(s)$ is a smooth loop in N, ω_i is of odd degree and ω_i^1 is of even degree. Let us recall that even forms on the free loop space commute. $F(\sigma_n)$ is built from even forms on the free loop space, which commute. This explains why we have to consider the symmetric Fock space. Therefore, if σ belongs to $\mathrm{WN}_{s,\infty-}$, then $F(\sigma)=\sum F^{2r}(\sigma)$, where $F^{2r}(\sigma)$ is a measurable form on $L(N)$ of degree $2r$ (see Jones and Léandre (1991) for an analogous statement in the stochastic context).

Let us explain why the free loop space is important in this context. Let $\mathrm{d}\nu_x(1)$ be the law of the Brownian bridge on N starting from x and coming back at x at time 1: this is the law of the Brownian motion $x_.(x)$ subject to return in time 1 at its departure. Let $p_t(x,y)$ be the heat kernel associated with $x_t(x)$: the law of $x_t(x)$ is namely $p_t(x,y)\,\mathrm{d}m_N(y)$ (Ikeda and Watanabe 1981). We consider the Bismut–Høegh–Krohn measure on the continuous free loop space $L_0(N)$:

$$\mathrm{d}P=p_1(x,x)\mathrm{d}x\otimes\mathrm{d}\nu_x(1) \quad [14]$$

This satisfies

$$\mathrm{tr}[\exp[-s_1\Delta_N]f_1\cdots f_n\exp[-(1-s_n)\Delta_N]]=\int_{L_0(N)}f_1(x(s_1))\cdots f_n(x(s_n))\mathrm{d}P \quad [15]$$

(We are interested in the trace of the heat semigroup instead of the heat semigroup itself unlike in the previous section.)

Since N is spin, we can consider the spin bundle $\mathrm{Sp}=\mathrm{Sp}_+\oplus\mathrm{Sp}_-$ on it, the Clifford bundle Cl on it with its natural $\mathbb{Z}/2\mathbb{Z}$ gradation (Gilkey 1995). Let us recall that the Clifford algebra acts on the spinors. A form ω can be associated with an element $\tilde{\omega}$ of the Clifford bundle (Gilkey 1995). We consider the Brownian loop $x(.)$ associated to the Bismut–Høegh–Krohn measure. If $s<t$, we can define the stochastic parallel transport $\tilde{\tau}_{s,t}$ from $x(t)$ to $x(s)$ (we identify a loop to a path from $[0,1]$ into N with the same end values). We remark that with the notations of [13]

$$\int_{\Delta_n}\tilde{\tau}_{0,s_1}(\tilde{\omega}_1(\mathrm{d}x(s_1))+\tilde{\omega}_1\mathrm{d}s_1)\tilde{\tau}_{s_1,s_2}\ldots\times\tilde{\tau}_{s_{n-1},s_n}(\tilde{\omega}_n(\mathrm{d}x(s_n))+\tilde{\omega}_n^1\mathrm{d}s_n)\tilde{\tau}_{s_n,1}=A \quad [16]$$

is a random almost surely defined even element of the Clifford bundle over $x(0)$. Acting on $\mathrm{Sp}(x(0))$, it thus preserves the gradation. We consider its supertrace $\mathrm{tr}_sA=\mathrm{tr}_{\mathrm{Sp}_+}A-\mathrm{tr}_{\mathrm{Sp}_-}A$. This becomes a random variable on $L_0(N)$. We introduce the scalar curvature K of the Levi–Civita connection on N, whose introduction arises from the Lichnerowicz formula given the square of the Dirac operator in terms of the horizontal

Laplacian on the spin bundle (Gilkey 1995). We consider the expression $\int_{L_0(N)} \exp[-\int_0^1 K(x(s)\,ds/8]\,\mathrm{tr}_s A\,dP$. This expression can be extended to $\mathrm{WN}_{s,\infty-}$ and therefore defines an element Wi of $\mathrm{WN}_{s,-\infty}$ called by Getzler (Léandre 2002) the Witten current.

Bismut has introduced a Hermitian bundle ξ on M. He deduces a bundle ξ_∞ on $L(N)$: the fiber on a loop $x(.)$ is the space of smooth sections along the loop of ξ. We can suppose that ξ is a sub-bundle given by a projector p of a trivial bundle. We can suppose that the Hermitian connection on ξ is the projection connection $A = p\mathrm{d}p$ such that its curvature is $R = p\mathrm{d}p \wedge p\mathrm{d}p$. Bismut (1985, 1987) has introduced the Bismut–Chern character:

$$\mathrm{Ch}(\xi_\infty) = \mathrm{tr}\left(\int_{\Delta_n} (A\mathrm{d}x(s_1) - R\mathrm{d}s_1) \wedge \cdots \wedge (A\mathrm{d}x(s_n) - R\mathrm{d}s_n) \right) \quad [17]$$

$\mathrm{Ch}(\xi_\infty)$ is a collection of even forms equal to $F(\sigma(\xi))$, where $\sigma(\xi)$ belongs to $\mathrm{WN}_{s,\infty-}$. We obtain:

Theorem 3 *Let us consider the index* $\mathrm{Ind}(D_\xi)$ *of the Dirac operator on N with auxiliary bundle* ξ (Hida *et al.* 1993). *We have*

$$\langle \mathrm{Wi}, \sigma(\xi) \rangle = \mathrm{Ind}\ D_\xi \quad [18]$$

The proof arises from the Lichnerowicz formula, the matricial Feynman–Kac formula, and the decomposition of the solution of a stochastic linear equation into the sum of iterated integrals.

By using the Potthoff–Streit theorem, we can do the analytic continuation of [18], as is suggested by the path-integral interpretation of Atiyah (1985) or Bismut (1985, 1987) of [18], motivated by the Duistermaat–Heckman or Berline–Vergne localization formulas on the free loop space. For this, these authors consider the Atiyah–Witten even form on the free loop space given by $I(x(.)) = \int_{S^1} |(\mathrm{d}/\mathrm{d}s)x(s)|^2 \mathrm{d}s + \mathrm{d}X_\infty$, where $\mathrm{d}X_\infty$ is the exterior derivative of the Killing form X_∞ which to a vector $X(.)$ on the loop associates $\langle X_\infty, X(.) \rangle = \int_{S^1} \langle X(s), \mathrm{d}x(s) \rangle$. We should obtain the heuristic formula

$$\langle \mathrm{Wi}, \sigma \rangle = Z^{-1} \int_{L(M)} F(\sigma) \wedge \exp\left[-\frac{1}{2} I(x(.)) \right] \quad [19]$$

We refer to Léandre (2002) for details.

Let us remark that Bismut (1987) and Léandre (2003) has continued his formal considerations to the case of the index theorem for a family of Dirac operators. We consider a fibration $\pi : N \to B$ of compact manifolds. Bismut replaces [19] by an integral of forms on the set of loops of N which project to a given loop of B. Bismut remarks that this integration in the fiber is related to filtering theory in stochastic analysis.

Fermionic Brownian Motion

Alvarez-Gaumé has given a supersymmetric proof of the index theorem: the path representation of the index of the Dirac operator involves infinite-dimensional Berezin integrals, while in the previous section only integrals of forms on the free loop space were concerned. Rogers (1987) has given an interpretation of the work of Alvarez-Gaumé, which begins with the study of fermionic Brownian motion. Let us interpret the considerations of Rogers (1987) in this framework.

We consider $\mathbf{C}^d$. H is the space of L^2-maps from $[0,1]$ into $\mathbf{C}^d$. We denote such a path by $\phi(s) = (\phi_1(s), \ldots, \phi_d(s))$, where $\phi_i(s) = q_i(s) + \sqrt{-1} p_i(s)$. $p_i(s)$ is the ith momentum and $q_i(s)$ the ith position. We denote by $\hat{\Lambda}(H)$ the fermionic Fock space associated with H.

We introduce the bilinear antisymmetric form on H:

$$\Omega(\phi^1, \phi^2) = \sqrt{-1} \sum_{i=1}^{d} \int_0^1 -p_i^1(s)\,\mathrm{d}q_i^2(s) + p_i^2(s)\,\mathrm{d}q_i^1(s) \quad [20]$$

and we consider the formal expression $\exp[\Omega] = \sum_{n=0}^{\infty} n!^{-1} \Omega^{\wedge n}$. We define a state on $\hat{\Lambda}^2(H)$ by $\omega(\phi^1 \wedge \phi^2) = \Omega(\phi^1, \phi^2)$. We put $\hat{\phi}_i(s) = 1_{[0,s]} + \sqrt{-1} 1_{[0,s]}$ where we take the ith coordinate in $\mathbf{C}^d$. We obtain, if $s_1 < s_2$,

$$\omega(\hat{\phi}_i(s_1) \wedge \hat{\phi}_j(s_2)) = -\sqrt{-1}\delta_{i,j} \quad [21]$$

where $\delta_{i,j}$ is the Kronecker symbol. We change the sign if $s_2 > s_1$ and we write 0 if $s_1 = s_2$.

We consider the finite-dimensional space Pol of fermionic polynomials on $\mathbf{C}^d$. Pol is endowed with a suitable norm, and we consider $\mathrm{Pol}^{\otimes n}$ endowed with the induced norm. We consider a formal series $\sigma = \sum \sigma_n$, where σ_n belongs to $\mathrm{Pol}^{\otimes n}$. In order to simplify the treatment, we suppose that our fermionic polynomials do not contain constant terms. We introduce the following Banach norm:

$$\|\sigma\|_C = \sum \frac{C^n}{n!} \|\sigma_n\| \quad [22]$$

We obtain the notion of Connes space $\mathrm{Co}_{\infty-}$ in this simpler context: σ belongs to $\mathrm{Co}_{\infty-}$ if $\|\sigma\|_C < \infty$ for all C. If $\sigma_n = P_1 \otimes \cdots \otimes P_n$, we associate

$$F(\sigma_n) = \int_{\Delta_n} P_1(\hat{\phi}(s_1)) \wedge \cdots \wedge P_n(\hat{\phi}(s_n))\,\mathrm{d}s_1 \cdots \mathrm{d}s_n \quad [23]$$

F can be extended in an injective continuous map from $\mathrm{Co}_{\infty-}$ into $\hat{\Lambda}(H)$. By using [21], we get:

Theorem 4 $\exp[\Omega]$ *is a distribution in the sense of Connes.*

We have only to use the formula [21] and

$$\langle \exp[\Omega], \phi^1 \wedge \phi^2 \cdots \wedge \phi^{2n} \rangle = \mathrm{Pf}\,\{\omega(\phi^i \wedge \phi^j)\} \quad [24]$$

and to estimate the obtained Pfaffians when $n \to \infty$.

Theorem 4 allows us to give a rigorous interpretation of the fermionic Feynman–Kac formula of Rogers (1987). We refer to Roepstorff (1994) for details.

$\exp[\Omega]$ should give a rigorous interpretation to the Gaussian Berezin integral with formal density $\exp\left[\sqrt{-1}\int_0^1 \sum p_i(s)\,\mathrm{d}q_i(s)\right]$.

See also: Equivariant Cohomology and the Cartan Model; Feynman Path Integrals; Functional Integration in Quantum Physics; Hopf Algebras and *q*-Deformation Quantum Groups; Index Theorems; Measure on Loop Spaces; Positive Maps on C^*-Algebras; Stationary Phase Approximation; Stochastic Differential Equations; Supermanifolds; Supersymmetric Quantum Mechanics.

Further Reading

Accardi L and Bożejko M (1998) Interacting Fock spaces and Gaussianization of probability measures. *Infinite Dimensional Analysis, Quantum Probability and Related Topics* 1: 663–670.

Albeverio S (1996) Wiener and Feynman path integrals and their applications. In: Masani PR (eds.) *Norbert Wiener Centenary Congress*, Proc. Symp. Appl. Math. vol. 52, pp. 163–194. Providence, RI: American Mathematical Society.

Andersson L and Driver B (1999) Finite dimensional approximation to Wiener measure and path integral formulas on manifolds. *Journal of Functional Analysis* 165: 430–498.

Atiyah M (1985) Circular symmetry and stationary phase approximation. In: *Colloque en l'honneur de L. Schwartz*, vol. 131, pp. 43–59. Paris: Asterisque.

Bismut JM (1985) Index theorem and equivariant cohomology on the loop space. *Communications in Mathematical Physics* 98: 213–237.

Berezansky YM and Kondratiev YO (1995) *Spectral Methods in Infinite-Dimensional Analysis*, vols. I, II. Dordrecht: Kluwer.

Bismut JM (1986) Localization formula, superconnections and the index theorem for families. *Communications in Mathematical Physics* 103: 127–166.

Bismut JM (1987) Filtering equation, equivariant cohomology and the Chern character. In: Seneor R (ed.) *Proc. VIIIth Int. Cong. Math. Phys.*, pp. 17–56. Singapore: World Scientific.

Chen KT (1973) Iterated path integrals of differential forms and loop space homology. *Annals of Mathematics* 97: 213–237.

Connes A (1988) Entire cyclic cohomology of Banach algebras and character of θ-summable Fredholm modules. *K-Theory* 1: 519–548.

Cuntz J (2001) Cyclic theory, bivariant K-theory and the bivariant Chern–Connes character. In: *Cyclic Homology in Noncommutative Geometry*, pp. 2–69. Encyclopedia of Mathematical Sciences, 121. Heidelberg: Springer.

Gilkey P (1995) *Invariance Theory, the Heat Equation and the Atiyah–Singer Theorem*. Boca Raton: CRC Press.

Hida T, Kuo HH, Potthoff J, and Streit L (1993) *White Noise: An Infinite Dimensional Calculus*. Dordrecht: Kluwer.

Ikeda N and Watanabe S (1981) *Stochastic Differential Equations and Diffusion Processes*. Amsterdam: North-Holland.

Jones JDS and Léandre R (1991) L^p Chen forms on loop spaces. In: Barlow M and Bingham N (eds.) *Stochastic Analysis*, pp. 104–162. Cambridge: Cambridge University Press.

Léandre R (2002) White noise analysis, filtering equation and the index theorem for families. In: Heyer H and Saitô (eds.) *Infinite Dimensional Harmonic Analysis* (to appear).

Léandre R (2003) Theory of distributions in the sense of Connes–Hida and Feynman path integral on a manifold. *Inf. Dim. Anal. Quant. Probab. Rel. Top.* 6: 505–517.

Roepstorff G (1994) *Path Integral Approach to Quantum Physics. An Introduction*. Heidelberg: Springer.

Rogers A (1987) Fermionic path integration and Grassmann Brownian motion. *Communications in Mathematical Physics* 113: 353–368.

Sidorova N, Smolyanov O, von Weizsaecker H, and Wittich O (2004) The surface limit of Brownian motion in tubular neighborhood of an embedded Riemannian manifold. *Journal of Functional Analysis* 206: 391–413.

Szabo R (2000) Equivariant cohomology and localization of path integrals in physics, *Lecture Notes in Physics* M63. Berlin: Springer.

Peakons

D D Holm, Imperial College, London, UK

Introduction

Peakons are singular solutions of the dispersionless Camassa–Holm (CH) shallow-water wave equation in one spatial dimension. These are reviewed in the context of asymptotic expansions and Euler–Poincaré (EP) variational principles. The dispersionless CH equation generalizes to the EPDiff equation (defined subsequently in this article), whose singular solutions are peakon wave fronts in higher dimensions. The reduction of these singular solutions of CH and EPDiff to canonical Hamiltonian dynamics on lower-dimensional sets may be understood, by realizing that their solution ansatz is a momentum map, and momentum maps are Poisson.

Camassa and Holm (1993) discovered the "peakon" solitary traveling-wave solution for a shallow-water wave:

$$u(x,t) = c\mathrm{e}^{-|x-ct|/\alpha} \quad [1]$$

whose fluid velocity u is a function of position x on the real line and time t. The peakon traveling wave

moves at a speed equal to its maximum height, at which it has a sharp peak (jump in derivative). Peakons are an emergent phenomenon, solving the initial-value problem for a partial differential equation (PDE) derived by an asymptotic expansion of Euler's equations using the small parameters of shallow-water dynamics. Peakons are nonanalytic solitons, which superpose as

$$u(x,t) = \sum_{a=1}^{N} p_a(t) e^{-|x-q_a(t)|/\alpha} \qquad [2]$$

for sets $\{p\}$ and $\{q\}$ satisfying canonical Hamiltonian dynamics. Peakons arise for shallow-water waves in the limit of zero linear dispersion in one dimension. Peakons satisfy a PDE arising from Hamilton's principle for geodesic motion on the smooth invertible maps (diffeomorphisms) with respect to the H^1 Sobolev norm of the fluid velocity. Peakons generalize to higher dimensions, as well. We explain how peakons were derived in the context of shallow-water asymptotics and describe some of their remarkable mathematical properties.

Shallow-Water Background for Peakons

Euler's equations for irrotational incompressible ideal fluid motion under gravity with a free surface have an asymptotic expansion for shallow-water waves that contains two small parameters, ϵ and δ^2, with ordering $\epsilon \geq \delta^2$. These small parameters are $\epsilon = a/h_0$ (the ratio of wave amplitude to mean depth) and $\delta^2 = (h_0/l_x)^2$ (the squared ratio of mean depth to horizontal length, or wavelength). Euler's equations are made nondimensional by introducing $x = l_x x'$ for horizontal position, $z = h_0 z'$ for vertical position, $t = (l_x/c_0)t'$ for time, $\eta = a\eta'$ for surface elevation, and $\varphi = (g l_x a/c_0)\varphi'$ for velocity potential, where $c_0 = \sqrt{gh_0}$ is the mean wave speed and g is the constant gravity. The quantity $\sigma = \sigma'/(h_0 \rho c_0^2)$ is the dimensionless Bond number, in which ρ is the mass density of the fluid and σ' is its surface tension, both of which are taken to be constants. After dropping primes, this asymptotic expansion yields the nondimensional Korteweg–de Vries (KdV) equation for the horizontal velocity variable $u = \varphi_x(x,t)$ at "linear" order in the small dimensionless ratios ϵ and δ^2, as the left-hand side of

$$u_t + u_x + \frac{3\epsilon}{2} u u_x + \frac{\delta^2}{6}(1 - 3\sigma) u_{xxx} = O(\epsilon\delta^2) \qquad [3]$$

Here, partial derivatives are denoted using subscripts, and boundary conditions are $u = 0$ and $u_x = 0$ at spatial infinity on the real line. The famous $\mathrm{sech}^2(x-t)$ traveling-wave solutions (the solitons) for KdV [3] arise in a balance between its (weakly) nonlinear steepening and its third-order linear dispersion, when the quadratic terms in ϵ and δ^2 on its right-hand side are neglected.

In eqn [3], a normal-form transformation due to Kodama (1985) has been used to remove the other possible quadratic terms of order $O(\epsilon^2)$ and $O(\delta^4)$. The remaining quadratic correction terms in the KdV equation [3] may be collected at order $O(\epsilon\delta^2)$. These terms may be expressed, after introducing a "momentum variable,"

$$m = u - \nu\delta^2 u_{xx} \qquad [4]$$

and neglecting terms of cubic order in ϵ and δ^2, as

$$m_t + m_x + \frac{\epsilon}{2}(u m_x + b m u_x) + \frac{\delta^2}{6}(1 - 3\sigma) u_{xxx} = 0 \qquad [5]$$

In the momentum variable $m = u - \nu\delta^2 u_{xx}$, the parameter ν is given by Dullin *et al.* (2001):

$$\nu = \frac{19 - 30\sigma - 45\sigma^2}{60(1 - 3\sigma)} \qquad [6]$$

Thus, the effects of δ^2-dispersion also enter the nonlinear terms. After restoring dimensions in eqn [5] and rescaling velocity u by $(b+1)$, the following "b-equation" emerges,

$$m_t + c_0 m_x + u m_x + b m u_x + \Gamma u_{xxx} = 0 \qquad [7]$$

where $m = u - \alpha^2 u_{xx}$ is the dimensional momentum variable, and the constants α^2 and Γ/c_0 are squares of length scales. When $\alpha^2 \to 0$, one recovers KdV from the b-equation [7], up to a rescaling of velocity. Any value of the parameter $b \neq -1$ may be achieved in eqn [7] by an appropriate Kodama transformation (Dullin *et al.* 2001).

As already emphasized, the values of the coefficients in the asymptotic analysis of shallow-water waves at quadratic order in their two small parameters only hold, modulo the Kodama normal-form transformations. Hence, these transformations may be used to advance the analysis and thereby gain insight, by optimizing the choices of these coefficients. The freedom introduced by the Kodama transformations among asymptotically equivalent equations at quadratic order in ϵ and δ^2 also helps to answer the perennial question, "Why are integrable equations so ubiquitous when one uses asymptotics in modeling?"

Integrable Cases of the b-equation [7]

The cases $b = 2$ and $b = 3$ are special values for which the b-equation becomes a completely integrable Hamiltonian system. For $b = 2$, eqn [7]

specializes to the integrable CH equation of Camassa and Holm (1993). The case $b=3$ in [7] recovers the integrable equation of Degasperis and Procesi (1999) (henceforth DP equation). These two cases exhaust the integrable candidates for [7], as was shown using Painlevé analysis. The b-family of eqns [7] was also shown in Mikhailov and Novikov (2002) to admit the symmetry conditions necessary for integrability, only in the cases $b=2$ for CH and $b=3$ for DP.

The b-equation [7] with $b=2$ was first derived in Camassa and Holm (1993) by using asymptotic expansions directly in the Hamiltonian for Euler's equations governing inviscid incompressible flow in the shallow-water regime. In this analysis, the CH equation was shown to be bi-Hamiltonian and thereby was found to be completely integrable by the inverse-scattering transform (IST) on the real line. Reviews of IST may be found, for example, in Ablowitz and Clarkson (1991), Dubrovin (1981), and Novikov *et al.* (1984). For discussions of other related bi-Hamiltonian equations, see Degasperis and Procesi (1999).

Camassa and Holm (1993) also discovered the remarkable peaked soliton (peakon) solutions of [1], [2] for the CH equation on the real line, given by [7] in the case $b=2$. The peakons arise as solutions of [7], when $c_0=0$ and $\Gamma=0$ in the absence of linear dispersion. Peakons move at a speed equal to their maximum height, at which they have a sharp peak (jump in derivative). Unlike the KdV soliton, the peakon speed is independent of its width (α). Periodic peakon solutions of CH were treated in Alber *et al.* (1999). There, the sharp peaks of periodic peakons were associated with billiards reflecting at the boundary of an elliptical domain. These billiard solutions for the periodic peakons arise from geodesic motion on a triaxial ellipsoid, in the limit that one of its axes shrinks to zero length.

Before Camassa and Holm (1993) derived their shallow-water equation, a class of integrable equations existed, which was later found to contain eqn [7] with $b=2$. This class of integrable equations was derived using hereditary symmetries in Fokas and Fuchssteiner (1981). However, eqn [7] was not written explicitly, nor was it derived physically as a shallow-water equation and its solution properties for $b=2$ were not studied before Camassa and Holm (1993). (See Fuchssteiner (1996) for an insightful history of how the shallow-water equation [7] in the integrable case with $b=2$ relates to the mathematical theory of hereditary symmetries.)

Equation [7] with $b=2$ was recently re-derived as a shallow-water equation by using asymptotic methods in three different approaches in Dullin *et al.* (2001), in Fokas and Liu (1996), and also in Johnson [2002]. All the three derivations used different variants of the method of asymptotic expansions for shallow-water waves in the absence of surface tension. Only the derivation in Dullin *et al.* (2001) used the Kodama normal-form transformations to take advantage of the nonuniqueness of the asymptotic expansion results at quadratic order.

The effects of the parameter b on the solutions of eqn [7] were investigated in Holm and Staley (2003), where b was treated as a bifurcation parameter, in the limiting case when the linear dispersion coefficients are set to $c_0=0$ and $\Gamma=0$. This limiting case allows several special solutions, including the peakons, in which the two nonlinear terms in eqn [7] balance each other in the "absence" of linear dispersion.

Peakons: Singular Solutions without Linear Dispersion in One Spatial Dimension

Peakons were first found as singular soliton solutions of the completely integrable CH equation. This is eqn [7] with $b=2$, now rewritten in terms of the velocity as

$$u_t + c_0 u_x + 3uu_x + \Gamma u_{xxx} = \alpha^2(u_{xxt} + 2u_x u_{xx} + uu_{xxx}) \qquad [8]$$

Peakons were found in Camassa and Holm (1993) to arise in the absence of linear dispersion. That is, they arise when $c_0=0$ and $\Gamma=0$ in CH [8]. Specifically, peakons are the individual terms in the peaked N-soliton solution of CH [8] for its velocity

$$u(x,t) = \sum_{b=1}^{N} p_b(t) e^{-|x-q_b(t)|/\alpha} \qquad [9]$$

in the absence of linear dispersion. Each term in the sum is a soliton with a sharp peak at its maximum, hence the name "peakon." Expressed using its momentum, $m=(1-\alpha^2\partial_x^2)u$, the peakon velocity solution [9] of dispersionless CH becomes a sum over a delta functions, supported on a set of points moving on the real line. Namely, the peakon velocity solution [9] implies

$$m(x,t) = 2\alpha \sum_{b=1}^{N} p_b(t)\delta(x - q_b(t)) \qquad [10]$$

because of the relation $(1-\alpha^2\partial_x^2)e^{-|x|/\alpha} = 2\alpha\delta(x)$. These solutions satisfy the b-equation [7] for any value of b, provided $c_0=0$ and $\Gamma=0$.

Thus, peakons are "singular momentum solutions" of the dispersionless b-equation, although

they are not stable for every value of b. From numerical simulations (Holm and Staley 2003), peakons are conjectured to be stable for $b > 1$. In the integrable cases $b = 2$ for CH and $b = 3$ for DP, peakons are stable singular soliton solutions. The spatial velocity profile $e^{-|x|/\alpha}/2\alpha$ of each separate peakon in [9] is the Green's function for the Helmholtz operator on the real line, with vanishing boundary conditions at spatial infinity. Unlike the KdV soliton, whose speed and width are related, the width of the peakon profile is set by its Green's function, independently of its speed.

Integrable Peakon Dynamics of CH

Substituting the peakon solution ansatz [9] and [10] into the dispersionless CH equation

$$m_t + um_x + 2mu_x = 0, \quad m = u - \alpha^2 u_{xx} \qquad [11]$$

yields Hamilton's canonical equations for the dynamics of the discrete set of peakon parameters $q_a(t)$ and $p_a(t)$:

$$\dot{q}_a(t) = \frac{\partial h_N}{\partial p_a} \quad \text{and} \quad \dot{p}_a(t) = -\frac{\partial h_N}{\partial q_a} \qquad [12]$$

for $a = 1, 2, \ldots, N$, with Hamiltonian given by (Camassa and Holm 1993):

$$h_N = \frac{1}{2}\sum_{a,\,b=1}^{N} p_a p_b e^{-|q_a - q_b|/\alpha} \qquad [13]$$

Thus, one finds that the points $x = q^a(t)$ in the peakon solution [9] move with the flow of the fluid velocity u at those points, since $u(q^a(t), t) = \dot{q}^a(t)$. This means the $q^a(t)$ are Lagrangian coordinates. Moreover, the singular momentum solution ansatz [10] is the Lagrange-to-Euler map for an invariant manifold of the dispersionless CH equation [11]. On this finite-dimensional invariant manifold for the PDE [11], the dynamics is canonically Hamiltonian.

With Hamiltonian [13], the canonical equations [12] for the $2N$ canonically conjugate peakon parameters $p_a(t)$ and $q_a(t)$ were interpreted in Camassa and Holm (1993) as describing "geodesic motion" on the N-dimensional Riemannian manifold whose co-metric is $g^{ij}(\{q\}) = e^{-|q_i - q_j|/\alpha}$. Moreover, the canonical geodesic equations arising from Hamiltonian [13] comprise an integrable system for any number of peakons N. This integrable system was studied in Camassa and Holm (1993) for solutions on the real line, and in Alber *et al.* (1999) and Mckean and Constantin (1999) and references therein, for spatially periodic solutions.

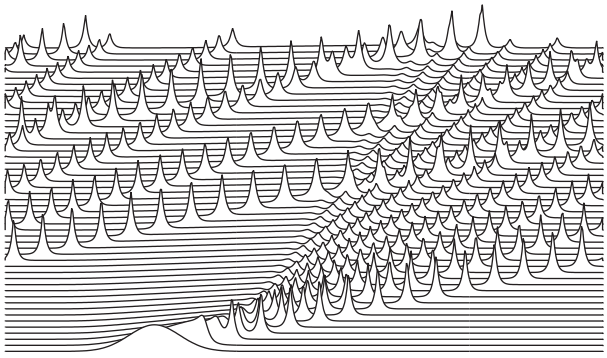


Figure 1 A smooth localized (Gaussian) initial condition for the CH equation breaks up into an ordered train of peakons as time evolves (the time direction being vertical). The peakon train eventually wraps around the periodic domain, thereby allowing the leading peakons to overtake the slower emergent peakons from behind in collisions that cause phase shifts as discussed in Camassa and Holm (1993). Courtesy of Staley M.

Being a completely integrable Hamiltonian soliton equation, the continuum CH equation [8] has an associated isospectral eigenvalue problem, discovered in Camassa and Holm (1993) for any values of its dispersion parameters c_0 and Γ. Remarkably, when $c_0 = 0$ and $\Gamma = 0$, this isospectral eigenvalue problem has a purely "discrete" spectrum. Moreover, in this case, each discrete eigenvalue corresponds precisely to the time-asymptotic velocity of a peakon. This discreteness of the CH isospectrum in the absence of linear dispersion implies that only the singular peakon solutions [10] emerge asymptotically in time, in the solution of the initial-value problem for the dispersionless CH equation [11]. This is borne out in numerical simulations of the dispersionless CH equation [11], starting from a smooth initial distribution of velocity (Fringer and Holm 2001, Holm and Staley 2003).

Figure 1 shows the emergence of peakons from an initially Gaussian velocity distribution and their subsequent elastic collisions in a periodic one-dimensional domain. This figure demonstrates that singular solutions dominate the initial-value problem and, thus, that it is imperative to go beyond smooth solutions for the CH equation; the situation is similar for the EPDiff equation.

Peakons as Mechanical Systems

Being governed by canonical Hamiltonian equations, each N-peakon solution can be associated with a mechanical system of moving particles. Calogero (1995) further extended the class of mechanical systems of this type. The r-matrix approach was applied to the Lax pair formulation of the N-peakon system for CH by Ragnisco and Bruschi (1996), who also pointed out the connection

of this system with the classical Toda lattice. A discrete version of the Adler–Kostant–Symes factorization method was used by Suris (1996) to study a discretization of the peakon lattice, realized as a discrete integrable system on a certain Poisson submanifold of gl(N) equipped with an r-matrix Poisson bracket. Beals *et al.* (1999) used the Stieltjes theorem on continued fractions and the classical moment problem for studying multipeakon solutions of the CH equation. Generalized peakon systems are described for any simple Lie algebra by Alber *et al.* (1999).

Pulsons: Generalizing the Peakon Solutions of the Dispersionless b-Equation for Other Green's Functions

The Hamiltonian h_N in eqn [13] depends on the Green's function for the relation between velocity u and momentum m. However, the singular momentum solution ansatz [10] is "independent" of this Green's function. Thus, as discovered in Fringer and Holm (2001), the singular momentum solution ansatz [10] for the dispersionless equation

$$m_t + um_x + 2mu_x = 0, \quad \text{with } u = g * m \qquad [14]$$

provides an invariant manifold on which canonical Hamiltonian dynamics occurs, for any choice of the Green's function g relating velocity u and momentum m by the convolution $u = g * m$.

The fluid velocity solutions corresponding to the singular momentum ansatz [10] for eqn [14] are the "pulsons". Pulsons are given by the sum over N velocity profiles determined by the Green's function g, as

$$u(x,t) = \sum_{a=1}^{N} p_a(t) g(x, q_a(t)) \qquad [15]$$

Again for [14], the singular momentum ansatz [10] results in a finite-dimensional invariant manifold of solutions, whose dynamics is canonically Hamiltonian. The Hamiltonian for the canonical dynamics of the $2N$ parameters $p_a(t)$ and $q_a(t)$ in the "pulson" solutions [15] of eqn [14] is

$$h_N = \frac{1}{2} \sum_{a,\, b=1}^{N} p_a\, p_b\, g(q_a, q_b) \qquad [16]$$

Again, for the pulsons, the canonical equations for the invariant manifold of singular momentum solutions provide a phase-space description of geodesic motion, this time with respect to the co-metric given by the Green's function g. Mathematical analysis and numerical results for the dynamics of these pulson solutions are given in Fringer and Holm (2001). These results describe how the collisions of pulsons [15] depend upon their shape.

Compactons in the $1/\alpha^2 \to 0$ Limit of CH

As mentioned earlier, in the limit that $\alpha^2 \to 0$, the CH equation [8] becomes the KdV equation. In contrast, when $1/\alpha^2 \to 0$, CH becomes the Hunter–Zheng equation (Hunter and Zheng 1994):

$$(u_t + uu_x)_{xx} = \tfrac{1}{2}\left(u_x^2\right)_x$$

This equation has "compacton" solutions, whose collision dynamics was studied numerically and put into the present context in Fringer and Holm (2001). The corresponding Green's function satisfies $-\partial_x{}^2 g(x) = 2\delta(x)$, so it has the triangular shape, $g(x) = 1 - |x|$ for $|x| < 1$, and vanishes otherwise, for $|x| \geq 1$. That is, the Green's function in this case has compact support, hence the name "compactons" for these pulson solutions, which as a limit of the integrable CH equations are true solitons, solvable by IST.

Pulson Solutions of the Dispersionless b-Equation

Holm and Staley (2003) give the pulson solutions of the traveling-wave problem and their elastic collision properties for the dispersionless b-equation:

$$m_t + um_x + bmu_x = 0, \quad \text{with } u = g * m \qquad [17]$$

with any (symmetric) Green's function g and for any value of the parameter b. Numerically, pulsons and peakons are both found to be stable for $b > 1$ (Holm and Staley 2003). The reduction to "noncanonical" Hamiltonian dynamics for the invariant manifold of singular momentum solutions [10] of the other integrable case $b = 3$ with peakon Green's function $g(x,y) = \mathrm{e}^{-|x-y|/\alpha}$ is found in Degasperis and Procesi (1999) and Degasperis *et al.* (2002).

Euler–Poincaré Theory in More Dimensions

Generalizing the Peakon Solutions of the CH Equation to Higher Dimensions

In Holm and Staley (2003), weakly nonlinear analysis and the assumption of columnar motion in the variational principle for Euler's equations are found to produce the two-dimensional generalization of the dispersionless CH equation [11]. This generalization is the EP equation (Holm *et al.* 1998a, b) for the Lagrangian consisting of the kinetic energy:

$$\ell = \frac{1}{2} \int \left[|\boldsymbol{u}|^2 + \alpha^2 (\operatorname{div} \boldsymbol{u})^2 \right] \mathrm{d}x\, \mathrm{d}y \qquad [18]$$

in which the fluid velocity $\boldsymbol{u}$ is a two-dimensional vector. Evolution generated by kinetic energy in

Hamilton's principle results in geodesic motion, with respect to the velocity norm $\|\boldsymbol{u}\|$, which is provided by the kinetic-energy Lagrangian. For ideal incompressible fluids governed by Euler's equations, the importance of geodesic flow was recognized by Arnol'd (1966) for the L^2 norm of the fluid velocity. The EP equation generated by any choice of kinetic-energy norm without imposing incompressibility is called "EPDiff," for "Euler–Poincaré equation for geodesic motion on the diffeomorphisms." EPDiff is given by (Holm *et al.* 1998a):

$$\left(\frac{\partial}{\partial t} + \boldsymbol{u}\cdot\nabla\right)\boldsymbol{m} + \nabla\boldsymbol{u}^{\mathrm{T}}\cdot\boldsymbol{m} + \boldsymbol{m}(\operatorname{div}\boldsymbol{u}) = 0 \qquad [19]$$

with momentum density $\boldsymbol{m} = \delta\ell/\delta\boldsymbol{u}$, where $\ell = (1/2)\|\boldsymbol{u}\|^2$ is given by the kinetic energy, which defines a norm in the fluid velocity $\|\boldsymbol{u}\|$, yet to be determined. By design, this equation has no contribution from either potential energy or pressure. It conserves the velocity norm $\|\boldsymbol{u}\|$ given by the kinetic energy. Its evolution describes geodesic motion on the diffeomorphisms with respect to this norm (Holm *et al.* 1998a).

An alternative way of writing the EPDiff equation [19] in either two or three dimensions is

$$\frac{\partial}{\partial t}\boldsymbol{m} - \boldsymbol{u}\times\operatorname{curl}\boldsymbol{m} + \nabla(\boldsymbol{u}\cdot\boldsymbol{m}) + \boldsymbol{m}(\operatorname{div}\boldsymbol{u}) = 0 \qquad [20]$$

This form of EPDiff involves all three differential operators: curl, gradient, and divergence. For the kinetic-energy Lagrangian ℓ given in [18], which is a norm for "irrotational" flow (with $\operatorname{curl}\boldsymbol{u} = 0$), we have the EPDiff equation [19] with momentum $\boldsymbol{m} = \delta\ell/\delta\boldsymbol{u} = \boldsymbol{u} - \alpha^2\nabla(\operatorname{div}\boldsymbol{u})$.

EPDiff [19] may also be written intrinsically as

$$\frac{\partial}{\partial t}\frac{\delta\ell}{\delta\boldsymbol{u}} = -\mathrm{ad}^*_{\boldsymbol{u}}\frac{\delta\ell}{\delta\boldsymbol{u}} \qquad [21]$$

where ad^* is the L^2 dual of the ad-operation (commutator) for vector fields (see Arnol'd and Khesin (1998) and Marsden and Ratiu (1999) for additional discussions of the beautiful geometry underlying this equation).

Reduction to the Dispersionless CH Equation in One Dimension

In one dimension, the EPDiff equations [19]–[21] with Lagrangian ℓ given in [18] simplify to the dispersionless CH equation [11]. The dispersionless limit of the CH equation appears, because potential energy and pressure have been ignored.

Strengthening the Kinetic-Energy Norm to Allow for Circulation

The kinetic-energy Lagrangian [18] is a norm for irrotational flow, with $\operatorname{curl}\boldsymbol{u} = 0$. However, inclusion of rotational flow requires the kinetic-energy norm to be strengthened to the H^1_α norm of the velocity, defined as

$$\begin{aligned}\ell &= \frac{1}{2}\int\left[|\boldsymbol{u}|^2 + \alpha^2(\operatorname{div}\boldsymbol{u})^2 + \alpha^2(\operatorname{curl}\boldsymbol{u})^2\right]\mathrm{d}x\,\mathrm{d}y \\ &= \frac{1}{2}\int\left[|\boldsymbol{u}|^2 + \alpha^2|\nabla\boldsymbol{u}|^2\right]\mathrm{d}x\,\mathrm{d}y = \frac{1}{2}\|\boldsymbol{u}\|^2_{H^1_\alpha}\end{aligned} \qquad [22]$$

Here, we assume boundary conditions that give no contributions upon integrating by parts. The corresponding EPDiff equation is [19] with $\boldsymbol{m} \equiv \delta\ell/\delta\boldsymbol{u} = \boldsymbol{u} - \alpha^2\Delta\boldsymbol{u}$. This expression involves inversion of the familiar Helmholtz operator in the (nonlocal) relation between fluid velocity and momentum density. The H^1_α norm $\|\boldsymbol{u}\|^2_{H^1_\alpha}$ for the kinetic energy [22] also arises in three dimensions for turbulence modeling based on Lagrangian averaging and using Taylor's hypothesis that the turbulent fluctuations are "frozen" into the Lagrangian mean flow (Foias *et al.* 2001).

Generalizing the CH Peakon Solutions to *n* Dimensions

Building on the peakon solutions [9] for the CH equation and the pulsons [15] for its generalization to other traveling-wave shapes in Fringer and Holm (2001), Holm and Staley (2003) introduced the following measure-valued singular momentum solution ansatz for the n-dimensional solutions of the EPDiff equation [19]:

$$\boldsymbol{m}(\boldsymbol{x},t) = \sum_{a=1}^{N}\int \boldsymbol{P}^a(s,t)\delta(\boldsymbol{x} - \boldsymbol{Q}^a(s,t))\,\mathrm{d}s \qquad [23]$$

These singular momentum solutions, called "diffeons," are vector density functions supported in $\mathbb{R}^n$ on a set of N surfaces (or curves) of codimension $(n-k)$ for $s \in \mathbb{R}^k$ with $k < n$. They may, for example, be supported on sets of points (vector peakons, $k = 0$), one-dimensional filaments (strings, $k = 1$), or two-dimensional surfaces (sheets, $k = 2$) in three dimensions.

Figure 2 shows the results for the EPDiff equation when a straight peakon segment of finite length is created initially moving rightward (East). Because of propagation along the segment in adjusting to the condition of zero speed at its ends and finite speed in its interior, the initially straight segment expands outward as it propagates and curves into a peakon "bubble."

Figure 3 shows an initially straight segment whose velocity distribution is exponential in the transverse

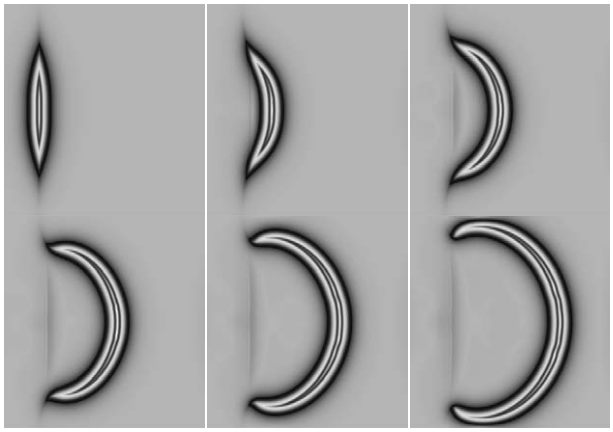


Figure 2 A peakon segment of finite length is initially moving rightward (east). Because its speed vanishes at its ends and it has fully two-dimensional spatial dependence, it expands into a peakon "bubble" as it propagates. (The various shades indicate different speeds. Any transverse slice will show a wave profile with a maximum at the center of the wave, which falls exponentially with distance away from the center.)

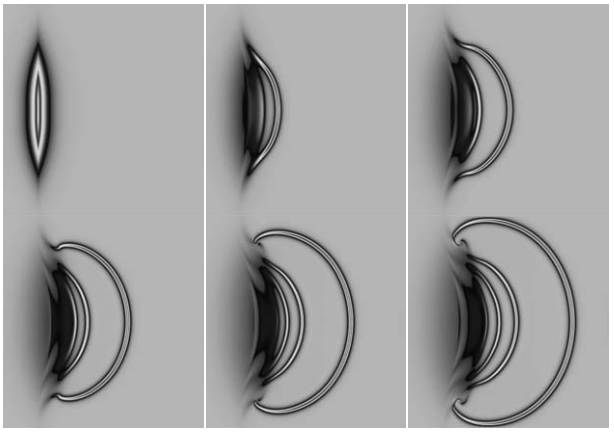


Figure 3 An initially straight segment of velocity distribution whose exponential profile is wider than the width α for the peakon solution breaks up into a train of curved peakon "bubbles," each of width α. This example illustrates the emergent property of the peakon solutions in two dimensions.

direction, but is wider than α for the peakon solution. This initial-velocity distribution evolves under EPDiff to separate into a train of curved peakon "bubbles," each of width α. This example illustrates the emergent property of the peakon solutions in two dimensions. This phenomenon is observed in nature, for example, as trains of internal wave fronts in the South China Sea (Liu *et al.* 1998).

Substitution of the singular momentum solution ansatz [23] into the EPDiff equation [19] implies the following integro-partial-differential equations (IPDEs) for the evolution of the parameters $\{\boldsymbol{P}\}$ and $\{\boldsymbol{Q}\}$:

$$\begin{aligned}\frac{\partial}{\partial t}\boldsymbol{Q}^a(s,t) &= \sum_{b=1}^{N}\int \boldsymbol{P}^b(s',t)\,G\Big(\boldsymbol{Q}^a(s,t) \\ &\quad - \boldsymbol{Q}^b(s',t)\Big)\,\mathrm{d}s' \\ \frac{\partial}{\partial t}\boldsymbol{P}^a(s,t) &= -\sum_{b=1}^{N}\int \Big(\boldsymbol{P}^a(s,t)\cdot\boldsymbol{P}^b(s',t)\Big) \\ &\quad \times \frac{\partial}{\partial \boldsymbol{Q}^a(s,t)}\,G\Big(\boldsymbol{Q}^a(s,t) \\ &\quad - \boldsymbol{Q}^b(s',t)\Big)\,\mathrm{d}s'\end{aligned} \qquad [24]$$

Importantly for the interpretation of these solutions, the coordinates $s \in \mathbb{R}^k$ turn out to be Lagrangian coordinates. The velocity field corresponding to the momentum solution ansatz [23] is given by

$$\begin{aligned}\boldsymbol{u}(\boldsymbol{x},t) &= G * \boldsymbol{m} \\ &= \sum_{b=1}^{N}\int \boldsymbol{P}^b(s',t)\,G\Big(\boldsymbol{x} - \boldsymbol{Q}^b(s',t)\Big)\,\mathrm{d}s'\end{aligned} \qquad [25]$$

for $\boldsymbol{u} \in \mathbb{R}^n$. When evaluated along the curve $\boldsymbol{x} = \boldsymbol{Q}^a(s,t)$, this velocity satisfies

$$\begin{aligned}\boldsymbol{u}(\boldsymbol{Q}^a(s,t),t) &= \sum_{b=1}^{N}\int \boldsymbol{P}^b(s',t) \\ &\quad \times G\Big(\boldsymbol{Q}^a(s,t) - \boldsymbol{Q}^b(s',t)\Big)\,\mathrm{d}s' \\ &= \frac{\partial \boldsymbol{Q}^a(s,t)}{\partial t}\end{aligned} \qquad [26]$$

Consequently, the lower-dimensional support sets defined on $\boldsymbol{x} = \boldsymbol{Q}^a(s,t)$ and parametrized by coordinates $s \in \mathbb{R}^k$ move with the fluid velocity. This means that the $s \in \mathbb{R}^k$ are Lagrangian coordinates. Moreover, eqns [24] for the evolution of these support sets are canonical Hamiltonian equations:

$$\frac{\partial}{\partial t}\boldsymbol{Q}^a(s,t) = \frac{\delta H_N}{\delta \boldsymbol{P}^a}, \qquad \frac{\partial}{\partial t}\boldsymbol{P}^a(s,t) = -\frac{\delta H_N}{\delta \boldsymbol{Q}^a} \qquad [27]$$

The corresponding Hamiltonian function $H_N : (\mathbb{R}^n \times \mathbb{R}^n)^N \to \mathbb{R}$ is

$$\begin{aligned}H_N &= \frac{1}{2}\int\int \sum_{a,b=1}^{N}\Big(\boldsymbol{P}^a(s,t)\cdot\boldsymbol{P}^b(s',t)\Big) \\ &\quad \times G(\boldsymbol{Q}^a(s,t),\boldsymbol{Q}_j(s',t))\,\mathrm{d}s\,\mathrm{d}s'\end{aligned} \qquad [28]$$

This is the Hamiltonian for geodesic motion on the cotangent bundle of a set of curves $\boldsymbol{Q}^a(s,t)$ with respect to the metric given by G. This dynamics was investigated numerically in Holm and Staley (2003) which can be referred to for more details of the solution properties. One important result found "numerically" in Holm and Staley (2003) is that only codimension-1 singular momentum solutions

appear to be stable under the evolution of the EPDiff equation. Thus,

Stability for codimension-1 solutions: *the singular momentum solutions of EPDiff are stable*, *as points on the line* (*peakons*), *as curves in the plane* (*filaments*, *or wave fronts*), *or as surfaces in space* (*sheets*).

Proving this stability result analytically remains an outstanding problem. The stability of peakons on the real line is proven in Constantin and Strauss (2000).

Reconnections in Oblique Overtaking Collisions of Peakon Wave Fronts

Figures 4 and 5 show results of oblique wave front collisions producing reconnections for the EPDiff equation in two dimensions. Figure 4 shows a single oblique overtaking collision, as a faster expanding peakon wave front overtakes a slower one and reconnects with it at the collision point. Figure 5 shows a series of reconnections involving the oblique overtaking collisions of two trains of curved peakon filaments, or wave fronts.

The Peakon Reduction is a Momentum Map

As shown in Holm and Marsden (2004), the singular solution ansatz [23] is a momentum map from the cotangent bundle of the smooth embeddings of lower-dimensional sets $\mathbb{R}^s \in \mathbb{R}^n$, to the dual of the Lie algebra of vector fields defined on these sets. (Momentum maps for Hamiltonian dynamics are reviewed in Marsden and Ratiu (1999), for example.) This geometric feature underlies the remarkable reduction properties of the EPDiff equation, and it also explains why the reduced equations must be Hamiltonian on the invariant manifolds of the singular solutions; namely, because momentum maps are Poisson maps. This geometric feature also underlies the singular momentum solution [23] and its associated velocity [25] which generalize the peakon solutions, both to higher dimensions and to arbitrary kinetic-energy metrics. The result that the singular solution ansatz [23] is a momentum map helps to organize the theory, to explain previous results, and to suggest new avenues of exploration.

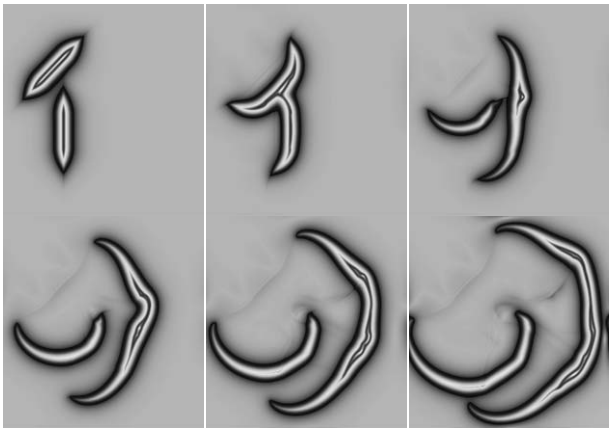


Figure 4 A single collision is shown involving reconnection as the faster peakon segment initially moving southeast along the diagonal expands, curves, and obliquely overtakes the slower peakon segment initially moving rightward (east). This reconnection illustrates one of the collision rules for the strongly two-dimensional EPDiff flow.

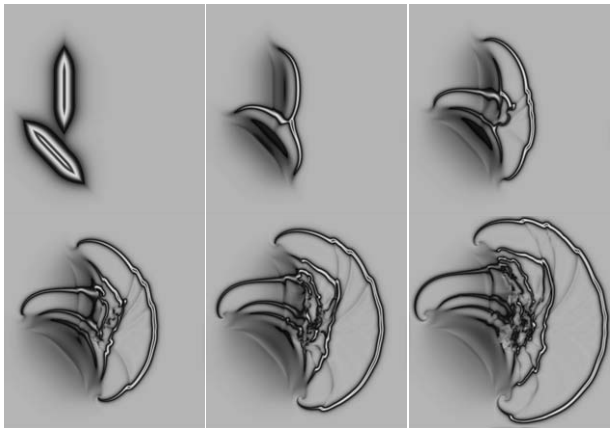


Figure 5 A series of multiple collisions is shown involving reconnections as the faster wider peakon segment initially moving northeast along the diagonal expands, breaks up into a wave train of peakons, each of which propagates, curves, and obliquely overtakes the slower wide peakon segment initially moving rightward (east), which is also breaking up into a train of wave fronts. In this series of oblique collision, the now-curved peakon filaments exchange momentum and reconnect several times.

Acknowledgment

The author is grateful to R Camassa, J E Marsden, T S Ratiu, and A Weinstein for their collaboration, help, and inspiring discussions over the years. He also thanks M F Staley for providing the figures obtained from his numerical simulations in the collaborations. US DOE provided partial support, under contract W-7405-ENG-36 for Los Alamos National Laboratory, and Office of Science ASCAR/AMS/MICS.

See also: Hamiltonian Systems: Obstructions to Integrability; Integrable Systems: Overview; Wave Equations and Diffraction.

Further Reading

Ablowitz MJ and Clarkson PA (1991) *Solitons, Nonlinear Evolution Equations and Inverse Scattering*. Cambridge: Cambridge University Press.

Ablowitz MJ and Segur H (1981) *Solitons and the Inverse Scattering Transform*. Philadelphia: SIAM.

Alber M, Camassa R, Fedorov Y, Holm D, and Marsden JE (1999) On billiard solutions of nonlinear PDEs. *PhysicsLetters A* 264: 171–178.

Alber M, Camassa R, Fedorov Y, Holm D, and Marsden JE (2001) The complex geometry of weak piecewise smooth solutions of integrable nonlinear PDE's of shallow water and Dym type. *Communications in Mathematical Physics* 221: 197–227.

Alber M, Camassa R, Holm D, and Marsden JE (1994) The geometry of peaked solitons and billiard solutions of a class of integrable PDEs. *Letters in Mathematical Physics* 32: 137–151.

Alber MS, Camassa R, and Gekhtman M (2000) On billiard weak solutions of nonlinear PDE's and Toda flows. *CRM Proceedings and Lecture Notes* 25: 1–11.

Arnol'd VI (1966) Sur la géométrie differentielle des groupes de Lie de dimenson infinie et ses applications à l'hydrodynamique des fluids parfaits. *Annales de l'Institut Fourier, Grenoble* 16: 319–361.

Arnol'd VI and Khesin BA (1998) *Topological Methods in Hydrodynamics*. Springer: New York.

Beals R, Sattinger DH, and Szmigielski J (1999) Multi-peakons and a theorem of Stietjes. *Inverse Problems* 15: L1–4.

Beals R, Sattinger DH, and Szmigielski J (2000) Multipeakons and the classical moment problem. *Advances in Mathematics* 154: 229–257.

Beals R, Sattinger DH, and Szmigielski J (2001) Peakons, strings, and the finite Toda lattice. *Communications in Pure and Applied Mathematics* 54: 91–106.

Calogero F (1995) An integrable Hamiltonian system. *Physics Letters A* 201: 306–310.

Calogero F and Francoise J-P (1996) A completely integrable Hamiltonian system. *Journal of Mathematics* 37: 2863–2871.

Camassa R and Holm DD (1993) An integrable shallow water equation with peaked solitons. *Physical Review Letters* 71: 1661–1664.

Camassa R, Holm DD, and Hyman JM (1994) A new integrable shallow water equation. *Advances in Applied Mechanics* 31: 1–33.

Constantin A and Strauss W (2000) Stability of peakons. *Communications on Pure and Applied Mathematics* 53: 603–610.

Degasperis A and Procesi M (1999) Asymptotic integrability. In: Degasperis A and Gaeta G (eds.) *Symmetry and Perturbation Theory*, pp. 23–37. Singapore: World Scientific.

Degasperis A, Holm DD, and Hone ANW (2002) A new integrable equation with peakon solutions. *Theoretical and Mathematical Physics* 133: 1463–1474.

Dubrovin B (1981) Theta functions and nonlinear equations. *Russian Mathematical Surveys* 36: 11–92.

Dubrovin BA, Novikov SP, and Krichever IM (1985) Integrable systems. *I. Itogi Nauki i Tekhniki. Sovr. Probl. Mat. Fund. Naprav*. 4. VINITI (Moscow) (Engl. transl. (1989) *Encyclopaedia of Mathematical Sciences*, vol. 4. Berlin: Springer).

Dullin HR, Gottwald GA, and Holm DD (2001) An integrable shallow water equation with linear and nonlinear dispersion. *Physical Review Letters* 87: 194501–04.

Dullin HR, Gottwald GA, and Holm DD (2003) Camassa–Holm, Korteweg–de Vries-5 and other asymptotically equivalent equations for shallow water waves. *Fluid Dynamics Research* 33: 73–95.

Dullin HR, Gottwald GA, and Holm DD (2004) On asymptotically equivalent shallow water wave equations. *Physica D* 190: 1–14.

Foias C, Holm DD, and Titi ES (2001) The Navier–Stokes-alpha model of fluid turbulence. *Physica D* 152: 505–519.

Fokas AS and Fuchssteiner B (1981) Bäcklund transformations for hereditary symmetries. *Nonlinear Analysis Transactions of the American Mathematical Society* 5: 423–432.

Fokas AS and Liu QM (1996) Asymptotic integrability of water waves. *Physical Review Letters* 77: 2347–2351.

Fuchssteiner B (1996) Some tricks from the symmetry-toolbox for nonlinear equations: generalization of the Camassa–Holm equation. *Physica D* 95: 229–243.

Fringer O and Holm DD (2001) Integrable vs. nonintegrable geodesic soliton behavior. *Physica D* 150: 237–263.

Holm DD (2005) The Euler–Poincaré variational framework for modeling fluid dynamics. In: Montaldi J and Ratiu T (eds.) *Geometric Mechanics and Symmetry: The Peyresq Lectures*, pp. 157–209. London Mathematical Society Lecture Notes Series 306. Cambridge: Cambridge University Press.

Holm DD and Marsden JE (2004) Momentum maps and measure-valued solutions (peakons, filaments and sheets) for the EPDiff equation. In: Marsden JE and Ratiu TS (eds.) *The Breadth of Symplectic and Poisson Geometry, A Festshcrift for Alan Weinstein*, pp. 203–235, Progress in Mathematics, vol. 232. Boston: Birkhäuser.

Holm DD, Marsden JE, and Ratiu TS (1998a) The Euler–Poincaré equations and semidirect products with applications to continuum theories. *Advances in Mathematics* 137: 1–81.

Holm DD, Marsden JE, and Ratiu TS (1998b) Euler–Poincaré models of ideal fluids with nonlinear dispersion. *Physical Review Letters* 349: 4173–4177.

Holm DD and Staley MF (2003) Nonlinear balance and exchange of stability in dynamics of solitons, peakons, ramps/cliffs and leftons in a 1+1 nonlinear evolutionary PDE. *Physical Letters A* 308: 437–444.

Holm DD and Staley MF (2003) Wave structures and nonlinear balances in a family of evolutionary PDEs. *SIAM Journal of Applied Dynamical Systems* 2(3): 323–380.

Holm DD and Staley MF (2004) Interaction dynamics of singular wave fronts in nonlinear evolutionary fluid equations (in preparation).

Hunter JK and Zheng Y (1994) On a completely integrable hyperbolic variational equation. *Physica D* 79: 361–386.

Johnson RS (2002) Camassa–Holm, Korteweg–de Vries models for water waves. *Journal of Fluid Mechanics* 455: 63–82.

Kodama Y (1985) On integrable systems with higher order corrections. *Physics Letters A* 107: 245–249.

Kodama Y (1985a) Normal forms for weakly dispersive wave equations. *Physics Letters A* 112: 193–196.

Kodama Y (1987) On solitary-wave interaction. *Physics letters A* 123: 276–282.

Liu AK, Chang YS, Hsu M-K, and Liang NK (1998) Evolution of nonlinear internal waves in the east and south China Seas. *Journal of Geophysical Research* 103: 7995–8008.

Marsden JE and Ratiu TS (1999) *Introduction to Mechanics and Symmetry*. Texts in Applied Mathematics, 2nd edn. vol. 17, Berlin: Springer.

McKean HP and Constantin A (1999) A shallow water equation on the circle. *Communications on Pure and Applied Mathematics* 52: 949–982.

Mikhailov AV and Novikov VS (2002) Perturbative symmetry approach. *Journal of Physics A* 35: 4775–4790.

Novikov SP, Manakov SV, Pitaevski LP, and Zakharov VE (1984) *Theory of Solitons. The Inverse Scattering Method*, Comtemporary Soviet Mathematics. Consultants Bureau (translated from Russian). New York: Plenum.

Ragnisco O and Bruschi M (1996) Peakons, *r*-matrix and Toda lattice. *Physica A* 228: 150–159.

Suris YB (1996) A discrete time peakons lattice. *Physics Letters A* 217: 321–329.

Penrose Inequality *see* Geometric Flows and the Penrose Inequality

Percolation Theory

V Beffara, Ecole Normale Supérieure de Lyon, Lyon, France
V Sidoravicius, IMPA, Rio de Janeiro, Brazil

Introduction

Percolation as a mathematical theory was introduced by Broadbent and Hammersley (1957), as a stochastic way of modeling the flow of a fluid or gas through a porous medium of small channels which may or may not let gas or fluid pass. It is one of the simplest models exhibiting a phase transition, and the occurrence of a critical phenomenon is central to the appeal of percolation. Having truly applied origins, percolation has been used to model the fingering and spreading of oil in water, to estimate whether one can build nondefective integrated circuits, and to model the spread of infections and forest fires. From a mathematical point of view, percolation is attractive because it exhibits relations between probabilistic and algebraic/topological properties of graphs.

To make the mathematical construction of such a system of channels, take a graph $\mathcal{G}$ (which originally was taken as $\mathbb{Z}^d$), with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$, and make all the edges independently open (or passable) with probability p or closed (or blocked) with probability $1-p$. Write P_p for the corresponding probability measure on the set of configurations of open and closed edges – that model is called bond percolation. The collection of open edges thus forms a random subgraph of $\mathcal{G}$, and the original question stated by Broadbent was whether the connected component of the origin in that subgraph is finite or infinite.

A path on $\mathcal{G}$ is a sequence $v_1, v_2, \ldots$ of vertices of $\mathcal{G}$, such that for all $i \geq 1$, v_i and v_{i+1} are adjacent on $\mathcal{G}$. A path is called open if all the edges $\{v_i, v_{i+1}\}$ between successive vertices are open. The infiniteness of the cluster of the origin is equivalent to the existence of an unbounded open path starting from the origin.

There is an analogous model, called "site percolation," in which all edges are assumed to be passable, but the vertices are independently open or closed with probability p or $1-p$, respectively. An open path is then a path along which all vertices are open. Site percolation is more general than bond percolation in the sense that the existence of a path for bond percolation on a graph $\mathcal{G}$ is equivalent to the existence of a path for site percolation on the covering graph of $\mathcal{G}$. However, site percolation on a given graph may not be equivalent to bond percolation on any other graph.

All graphs under consideration will be assumed to be connected, locally finite and quasitransitive. If $A, B \subset \mathcal{V}$, then $A \leftrightarrow B$ means that there exists an open path from some vertex of A to some vertex of B; by a slight abuse of notation, $u \leftrightarrow v$ will stand for the existence of a path between sites u and v, that is, the event $\{u\} \leftrightarrow \{v\}$. The open cluster $C(v)$ of the vertex v is the set of all open vertices which are connected to v by an open path:

$$C(v) = \{u \in \mathcal{V} : u \leftrightarrow v\}$$

The central quantity of the percolation theory is the percolation probability:

$$\theta(p) := P_p\{\mathbf{0} \leftrightarrow \infty\} = P_p\{|C(\mathbf{0})| = \infty\}$$

The most important property of the percolation model is that it exhibits a phase transition, that is, there exists a threshold value $p_c \in [0,1]$, such that the global behavior of the system is substantially different in the two regions $p < p_c$ and $p > p_c$. To make this precise, observe that θ is a nondecreasing function. This can be seen using Hammersley's joint construction of percolation systems for all $p \in [0,1]$ on $\mathcal{G}$: let $\{U(v), v \in \mathcal{V}\}$ be independent random variables, uniform in [0,1]. Declare v to be p-open if $U(v) \leq p$, otherwise it is declared p-closed. The configuration of p-open vertices has the distribution P_p for each $p \in [0,1]$. The collection of p-open vertices is nondecreasing in p, and therefore $\theta(p)$ is nondecreasing as well. Clearly, $\theta(0) = 0$ and $\theta(1) = 1$ (Figure 1).

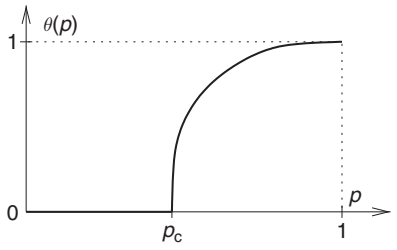


Figure 1 The behavior of $\theta(p)$ around the critical point (for bond percolation).

The critical probability is defined as

$$p_c := p_c(\mathcal{G}) = \sup\{p\colon \theta(p) = 0\}$$

By definition, when $p < p_c$, the open cluster of the origin is P_p-a.s. finite; hence, all the clusters are also finite. On the other hand, for $p > p_c$ there is a strictly positive P_p-probability that the cluster of the origin is infinite. Thus, from Kolmogorov's zero–one law it follows that

$$P_p\{|C(v)| = \infty \text{ for some } v \in \mathcal{V}\} = 1 \quad \text{for } p > p_c$$

Therefore, if the intervals $[0, p_c)$ and $(p_c, 1]$ are both nonempty, there is a phase transition at p_c.

Using a so-called Peierls argument it is easy to see that $p_c(\mathcal{G}) > 0$ for any graph $\mathcal{G}$ of bounded degree. On the other hand, Hammersley proved that $p_c(\mathbb{Z}^d) < 1$ for bond percolation as soon as $d \geq 2$, and a similar argument works for site percolation and various periodic graphs as well. But for some graphs $\mathcal{G}$, it is not so easy to show that $p_c(\mathcal{G}) < 1$. One says that the system is in the subcritical (resp. supercritical) phase if $p < p_c$ (resp. $p > p_c$).

It was one of the most remarkable moments in the history of percolation when Kesten (1980) proved, based on results by Harris, Russo, Seymour and Welsh, that the critical parameter for bond percolation on $\mathbb{Z}^2$ is equal to 1/2. Nevertheless, the exact value of $p_c(\mathcal{G})$ is known only for a handful of graphs, all of them periodic and two dimensional – see below.

Percolation in $\mathbb{Z}^d$

The graph on which most of the theory was originally built is the cubic lattice $\mathbb{Z}^d$, and it was not before the late twentieth century that percolation was seriously considered on other kinds of graphs (such as Cayley graphs), on which specific phenomena can appear, such as the coexistence of multiple infinite clusters for some values of the parameter p. In this section, the underlying graph is thus assumed to be $\mathbb{Z}^d$ for $d \geq 2$, although most of the results still hold in the case of a periodic d-dimensional lattice.

The Subcritical Regime

When $p < p_c$, all open clusters are finite almost surely. One of the greatest challenges in percolation theory has been to prove that $\chi(p) := E_p\{|C(v)|\}$ is finite if $p < p_c$ (E_p stands for the expectation with respect to P_p). For that one can define another critical probability as the threshold value for the finiteness of the expected cluster size of a fixed vertex:

$$p_T(\mathcal{G}) := \sup\{p\colon \chi(p) < \infty\}$$

It was an important step in the development of the theory to show that $p_T(\mathcal{G}) = p_c(\mathcal{G})$. The fundamental estimate in the subcritical regime, which is a much stronger statement than $p_T(\mathcal{G}) = p_c(\mathcal{G})$, is the following:

Theorem 1 (Aizenman and Barsky, Menshikov). *Assume that $\mathcal{G}$ is periodic. Then for $p < p_c$ there exist constants $0 < C_1, C_2 < \infty$, such that*

$$P_p\{|C(v)| \geq n\} \leq C_1\, e^{-C_2 n}$$

The last statement can be sharpened to a "local limit theorem" with the help of a subadditivity argument: for each $p < p_c$, there exists a constant $0 < C_3(p) < \infty$, such that

$$\lim_{n\to\infty} -\frac{1}{n} \log P_p\{|C(v)| = n\} = C_3(p)$$

The Supercritical Regime

Once an infinite open cluster exists, it is natural to ask how it looks like, and how many infinite open clusters exist. It was shown by Newman and Schulman that for periodic graphs, for each p, exactly one of the following three situations prevails: if $N \in \mathbb{Z}_+ \cup \{\infty\}$ is the number of infinite open clusters, then $P_p(N = 0) = 1$, or $P_p(N = 1) = 1$, or $P_p(N = \infty) = 1$.

Aizenman, Kesten, and Newman showed that the third case is impossible on $\mathbb{Z}^d$. By now several proofs exist, perhaps the most elegant of which is due to Burton and Keane, who prove that indeed there cannot be infinitely many infinite open clusters on any amenable graph. However, there are some graphs, such as regular trees, on which coexistence of several infinite clusters is possible.

The geometry of the infinite open cluster can be explored in some depth by studying the behavior of a random walk on it. When $d = 2$, the random walk is recurrent, and when $d \geq 3$ is a.s. transient. In all dimensions $d \geq 2$, the walk behaves diffusively, and the "central limit theorem" and the "invariance principle" were established in both the annealed and quenched cases.

Wulff droplets In the supercritical regime, aside from the infinite open cluster, the configuration contains finite clusters of arbitrary large sizes. These large finite open clusters can be thought of as droplets swimming in the areas surrounded by an infinite open cluster. The presence at a particular location of a large finite cluster is an event of low probability, namely, on $\mathbb{Z}^d, d \geq 2$, for $p > p_c$, there exist positive constants $0 < C_4(p), C_5(p) < \infty$, such that

$$C_4(p) \leq -\frac{1}{n^{(d-1)/d}} \log P_p\{|C(v)| = n\} \leq C_5(p)$$

for all large n. This estimate is based on the fact that the occurrence of a large finite cluster is due to a surface effect. The typical structure of the large finite cluster is described by the following theorem:

Theorem 2 *Let $d \geq 2$, and $p > p_c$. There exists a bounded, closed, convex subset W of $\mathbb{R}^d$ containing the origin, called the* normalized *Wulff crystal of the Bernoulli percolation model, such that, under the conditional probability $P_p\{\cdot \mid n^d \leq |C(0)| < \infty\}$, the random measure*

$$\frac{1}{n^d} \sum_{x \in C(0)} \delta_{x/n}$$

(where δ_x denotes a Dirac mass at x) converges weakly in probability toward the random measure $\theta(p)\mathbb{1}_W(x - M)\,dx$ (where M is the rescaled center of mass of the cluster $C(0)$). The deviation probabilities behave as $\exp\{-cn^{d-1}\}$ (i.e., they exhibit large deviations of surface order; in dimensions 4 and more it holds up to re-centering).

This result was proved in dimension 2 by Alexander *et al.* (1990), and in dimensions 3 and more by Cerf (2000).

Percolation Near the Critical Point

Percolation in Slabs The main macroscopic observable in percolation is $\theta(p)$, which is positive above $p_c, 0$ below p_c, and continuous on $[0,1]\backslash\{p_c\}$. Continuity at p_c is an open question in the general case; it is known to hold in two dimensions (cf. below) and in high enough dimension (at the moment $d \geq 19$ though the value of the critical dimension is believed to be 6) using lace expansion methods. The conjecture that $\theta(p_c) = 0$ for $3 \leq d \leq 18$ remains one of the major open problems.

Efforts to prove that led to some interesting and important results. Barsky, Grimmett, and Newman solved the question in the half-space case, and simultaneously showed that the slab percolation and half-space percolation thresholds coincide. This was complemented by Grimmett and Marstrand showing that

$$p_c(\text{slab}) = p_c(\mathbb{Z}^d)$$

Critical exponents In the subcritical regime, exponential decay of the correlation indicates that there is a finite correlation length $\xi(p)$ associated to the system, and defined (up to constants) by the relation

$$P_p(0 \leftrightarrow nx) \approx \exp\left(-\frac{n\varphi(x)}{\xi(p)}\right)$$

where φ is bounded on the unit sphere (this is known as Ornstein–Zernike decay). The phase transition can then also be defined in terms of the divergence of the correlation length, leading again to the same value for p_c; the behavior at or near the critical point then has no finite characteristic length, and gives rise to scaling exponents (conjecturally in most cases).

The most usual critical exponents are defined as follows, if $\theta(p)$ is the percolation probability, C the cluster of the origin, and $\xi(p)$ the correlation length:

$$\begin{aligned}
\frac{\partial^3}{\partial p^3} E_p[|C|^{-1}] &\approx |p - p_c|^{-1-\alpha} \\
\theta(p) &\approx (p - p_c)_+^{\beta} \\
\chi^f(p) := E_p[|C|\mathbb{1}_{|C|<\infty}] &\approx |p - p_c|^{-\gamma} \\
P_{p_c}[|C| = n] &\approx n^{-1-1/\delta} \\
P_{p_c}[x \in C] &\approx |x|^{2-d-\eta} \\
\xi(p) &\approx |p - p_c|^{\nu} \\
P_{p_c}[\text{diam}(C) = n] &\approx n^{-1-1/\rho} \\
\frac{E_p[|C|^{k+1}\mathbb{1}_{|C|<\infty}]}{E_p[|C|^{k}\mathbb{1}_{|C|<\infty}]} &\approx |p - p_c|^{-\Delta}
\end{aligned}$$

These exponents are all expected to be universal, that is, to depend only on the dimension of the lattice, although this is not well understood at the mathematical level; the following scaling relations between the exponents are believed to hold:

$$2 - \alpha = \gamma + 2\beta = \beta(\delta + 1), \quad \Delta = \delta\beta,\ \gamma = \nu(2 - \eta)$$

In addition, in dimensions up to $d_c = 6$, two additional hyperscaling relations involving d are strongly conjectured to hold:

$$d\rho = \delta + 1, \qquad d\nu = 2 - \alpha$$

while above d_c the exponents are believed to take their mean-field value, that is, the ones they have for percolation on a regular tree:

$$\alpha = -1,\ \beta = 1,\ \gamma = 1,\ \delta = 2$$
$$\eta = 0,\ \nu = \tfrac{1}{2},\ \rho = \tfrac{1}{2},\ \Delta = 2$$

Not much is known rigorously on critical exponents in the general case. Hara and Slade (1990) proved that mean field behavior does happen above dimension 19, and the proof can likely be extended to treat the case $d \geq 7$. In the two-dimensional case on the other hand, Kesten (1987) showed that, assuming that the exponents δ and ρ exist, then so do β, γ, η, and ν, and they satisfy the scaling and hyperscaling relations where they appear.

The incipient infinite cluster When studying long-range properties of a critical model, it is useful to have an object which is infinite at criticality, and such is not the case for percolation clusters. There are two ways to condition the cluster of the origin to

be infinite when $p = p_c$: The first one is to condition it to have diameter at least n (which happens with positive probability) and take a limit in distribution as n goes to infinity; the second one is to consider the model for parameter $p > p_c$, condition the cluster of 0 to be infinite (which happens with positive probability) and take a limit in distribution as p goes to p_c. The limit is the same in both cases; it is known as the incipient infinite cluster.

As in the supercritical regime, the structure of the cluster can be investigated by studying the behavior of a random walk on it, as was suggested by de Gennes; Kesten proved that in two dimensions, the random walk on the incipient infinite cluster is subdiffusive, that is, the mean square displacement after n steps behaves as $n^{1-\varepsilon}$ for some $\varepsilon > 0$.

The construction of the incipient infinite cluster was done by Kesten (1986) in two dimensions, and a similar construction was performed recently in high dimension by van der Hofstad and Jarai (2004).

Percolation in Two Dimensions

As is the case for several other models of statistical physics, percolation exhibits many specific properties when considered on a two-dimensional lattice: duality arguments allow for the computation of p_c in some cases, and for the derivation of *a priori* bounds for the probability of crossing events at or near the critical point, leading to the fact that $\theta(p_c) = 0$. On another front, the scaling limit of critical site percolation on the two-dimensional triangular lattice can be described in terms of Stochastic Loewner evolutions (SLE) processes.

Duality, Exact Computations, and RSW Theory

Given a planar lattice $\mathcal{L}$, define two associated graphs as follows. The dual lattice $\mathcal{L}'$ has one vertex for each face of the original lattice, and an edge between two vertices if and only if the corresponding faces of $\mathcal{L}$ share an edge. The star graph $\mathcal{L}^*$ is obtained by adding to $\mathcal{L}$ an edge between any two vertices belonging to the same face ($\mathcal{L}^*$ is not planar in general; $(\mathcal{L}, \mathcal{L}^*)$ is commonly known as a matching pair). Then, a result of Kesten is that, under suitable technical conditions,

$$p_c^{\text{bond}}(\mathcal{L}) + p_c^{\text{bond}}(\mathcal{L}') = p_c^{\text{site}}(\mathcal{L}) + p_c^{\text{site}}(\mathcal{L}^*) = 1$$

Two cases are of particular importance: the lattice $\mathbb{Z}^2$ is isomorphic to its dual; the triangular lattice $\mathcal{T}$ is its own star graph. It follows that

$$p_c^{\text{bond}}(\mathbb{Z}^2) = p_c^{\text{site}}(\mathcal{T}) = \tfrac{1}{2}$$

The only other critical parameters that are known exactly are $p_c^{\text{bond}}(\mathcal{T}) = 2\sin(\pi/18)$ (and hence also p_c^{bond} for $\mathcal{T}'$, i.e., the hexagonal lattice) and p_c^{bond} for the bow-tie lattice which is a root of the equation $p^5 - 6p^3 + 6p^2 + p - 1 = 0$. The value of the critical parameter for site percolation on $\mathbb{Z}^2$ might, on the other hand, never be known; it is even possible that it is "just a number" without any other signification.

Still using duality, one can prove that the probability, for bond percolation on the square lattice with parameter $p = 1/2$, that there is a connected component crossing an $(n+1) \times n$ rectangle in the longer direction is exactly equal to 1/2. This and clever arguments involving the symmetry of the lattice lead to the following result, proved independently by Russo and by Seymour and Welsh and known as the RSW theorem:

Theorem 3 (Russo 1978, Seymour and Welsh 1978). *For every $a, b > 0$ there exist $\eta > 0$ and $n_0 > 0$ such that for every $n > n_0$, the probability that there is a cluster crossing an $\lfloor na \rfloor \times \lfloor nb \rfloor$ rectangle in the first direction is greater than η.*

The most direct consequence of this estimate is that the probability that there is a cluster going around an annulus of a given modulus is bounded below independently of the size of the annulus; in particular, almost surely there is some annulus around 0 in which this happens, and that is what allows to prove that $\theta(p_c) = 0$ for bond percolation on $\mathbb{Z}^2$ (**Figure 2**).

The Scaling Limit

RSW-type estimates give positive evidence that a scaling limit of the model should exist; it is indeed essentially sufficient to show convergence of the crossing probabilities to a nontrivial limit as n goes to infinity. The limit, which should depend only on the ratio a/b, was predicted by Cardy using conformal field theory methods. A celebrated result of Smirnov is the proof of Cardy's formula in the case of site percolation on the triangular lattice $\mathcal{T}$:

Theorem 4 (Smirnov (2001)). *Let Ω be a simply connected domain of the plane with four points a, b, c, d (in that order) marked on its boundary. For every $\delta > 0$, consider a critical site-percolation*

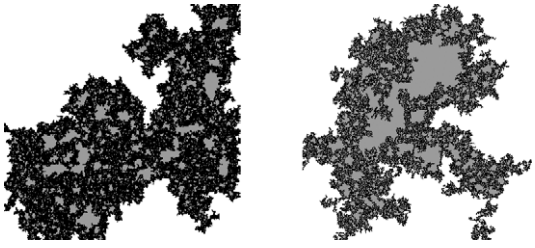


Figure 2 Two large critical percolation clusters in a box of the square lattice (first: bond percolation, second: site percolation).

model on the intersection of Ω *with* $\delta\mathcal{T}$ *and let* $f_\delta(ab, cd; \Omega)$ *be the probability that it contains a cluster connecting the arcs* ab *and* cd. *Then*:

(i) $f_\delta(ab, cd; \Omega)$ *has a limit* $f_0(ab, cd; \Omega)$ *as* $\delta \to 0$;
(ii) *the limit is conformally invariant, in the following sense: if* Φ *is a conformal map from* Ω *to some other domain* $\Omega' = \Phi(\Omega)$, *and maps* a *to* a', b *to* b', c *to* c' *and* d *to* d', *then* $f_0(ab, cd; \Omega) = f_0(a'b', c'd'; \Omega')$; *and*
(iii) *in the particular case when* Ω *is an equilateral triangle of side length* 1 *with vertices* a, b *and* c, *and if* d *is on* (ca) *at distance* $x \in (0, 1)$ *from* c, *then* $f_0(ab, cd; \Omega) = x$.

Point (iii) in particular is essential since it allows us to compute the limiting crossing probabilities in any conformal rectangle. In the original work of Cardy, he made his prediction in the case of a rectangle, for which the limit involves hypergeometric functions; the remark that the equilateral triangle gives rise to nicer formulae is originally due to Carleson.

To precisely state the convergence of percolation to its scaling limit, define the random curve known as the percolation exploration path (see **Figure 3**) as follows: In the upper half-plane, consider a site-percolation model on a portion of the triangular lattice and impose the boundary conditions that on the negative real half-line all the sites are open, while on the other half-line the sites are closed. The exploration curve is then the common boundary of the open cluster spanning from the negative half-line, and the closed cluster spanning from the positive half-line; it is an infinite, self-avoiding random curve in the upper half-plane.

As the mesh of the lattice goes to 0, the exploration curve then converges in distribution to the trace of an SLE process, as introduced by Schramm, with parameter $\kappa = 6$ – see **Figure 4**. The limiting curve is not simple anymore (which corresponds to the existence of pivotal sites on large critical percolation clusters), and it has Hausdorff dimension 7/4. For more details on SLE processes, see, for example, the related entry in the present volume.

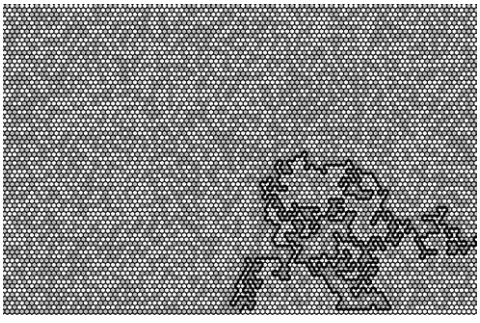


Figure 3 A percolation exploration path. Figure courtesy Schramm O (2000) Scaling limits of loop-erased random walks and uniform spanning trees. *Israel Journal of Mathematics* 118: 221–228.

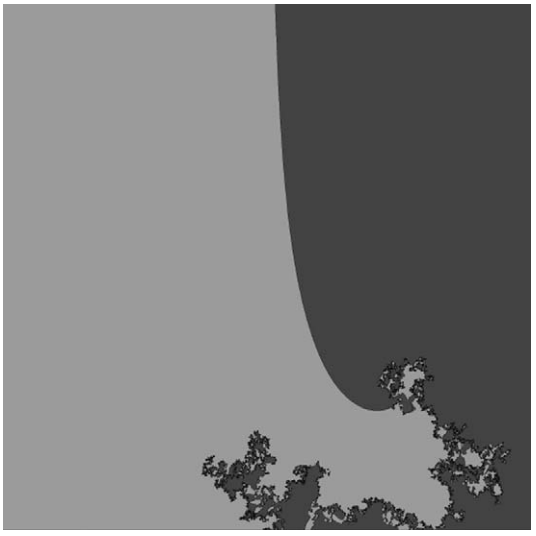


Figure 4 An SLE process with parameter $\kappa = 6$ (infinite time, with the driving process stopped at time 1).

As an application of this convergence result, one can prove that the critical exponents described in the previous section do exist (still in the case of the triangular lattice), and compute their exact values, except for α, which is still listed here for completeness:

$$\left[\alpha = -\frac{2}{3}\right], \beta = \frac{5}{36}, \gamma = \frac{43}{18}, \delta = \frac{91}{5}$$
$$\eta = \frac{5}{24}, \nu = \frac{4}{3}, \rho = \frac{48}{5}, \Delta = \frac{91}{36}$$

These exponents are expected to be universal, in the sense that they should be the same for percolation on any two-dimensional lattice; but at the time of this writing, this phenomenon is far from being understood on a mathematical level.

The rigorous derivation of the critical exponents for percolation is due to Smirnov and Werner (2001); the dimension of the limiting curve was obtained by Beffara (2004).

Other Lattices and Percolative Systems

Some modifications or generalizations of standard Bernoulli percolation on $\mathbb{Z}^d$ exhibit an interesting behavior and as such provide some insight into the original process as well; there are too many mathematical objects which can be argued to be percolative in some sense to give a full account of all

of them, so the following list is somewhat arbitrary and by no means complete.

Percolation on Nonamenable Graphs

The first modification of the model one can think of is to modify the underlying graph and move away from the cubic lattice; phase transition still occurs, and the main difference is the possibility for infinitely many infinite clusters to coexist. On a regular tree, such is the case whenever $p \in (p_c, 1)$, the first nontrivial example was produced by Grimmett and Newman as the product of $\mathbb{Z}$ by a tree: there, for some values of p the infinite cluster is unique, while for others there is coexistence of infinitely many of them. The corresponding definition, due to Benjamini and Schramm, is then the following: if N is as above the number of infinite open clusters,

$$p_u := \inf\{p : P_p(N = 1) = 1\} \geq p_c$$

The main question is then to characterize graphs on which $0 < p_c < p_u < 1$.

A wide class of interesting graphs is that of Cayley graphs of infinite, finitely generated groups. There, by a simultaneous result by Häggström and Peres and by Schonmann, for every $p \in (p_c, p_u)$ there are P_p-a.s. infinitely many infinite cluster, while for every $p \in (p_u, 1]$ there is only one – note that this does not follow from the definition since new infinite components could appear when p is increased. It is conjectured that $p_c < p_u$ for any Cayley graph of a nonamenable group (and more generally for any quasitransitive graph with positive Cheeger constant), and a result by Pak and Smirnova is that every infinite, finitely generated, nonamenable group has a Cayley graph on which $p_c < p_u$; this is then expected not to depend on the choice of generators. In the general case, it was recently proved by Gaboriau that if the graph $\mathcal{G}$ is unimodular, transitive, locally finite, and supports nonconstant harmonic Dirichlet functions (i.e., harmonic functions whose gradient is in ℓ^2), then indeed $p_c(\mathcal{G}) < p_u(\mathcal{G})$.

For reference and further reading on the topic, the reader is advised to refer to the review paper by Benjamini and Schramm (1996), the lecture notes of Peres (1999), and the more recent article of Gaboriau (2005).

Gradient Percolation

Another possible modification of the original model is to allow the parameter p to depend on the location; the porous medium may for instance have been created by some kind of erosion, so that there will be more open edges on one side of a given domain than on the other. If p still varies smoothly, then one expects some regions to look subcritical and others to look supercritical, with interesting behavior in the vicinity of the critical level set $\{p = p_c\}$. This particular model was introduced by Sapoval *et al.* (1978) under the name of gradient percolation (see Figure 5).

Figure 5 Gradient percolation in a square. In black is the cluster spanning from the bottom side of the square.

The control of the model away from the critical zone is essentially the same as for usual Bernoulli percolation, the main question being how to estimate the width of the phase transition. The main idea is then the same as in scaling theory: if the distance between a point v and the critical level set is less than the correlation length for parameter $p(v)$, then v is in the phase transition domain. This of course makes sense only asymptotically, say in a large $n \times n$ square with $p(x, y) = 1 - y/n$ as is the case in the figure: the transition then is expected to have width of order n^a for some exponent $a > 0$.

First-Passage Percolation

First-passage percolation (also known as Eden or Richardson model) was introduced by Hammersley and Welsh (1965) as a time-dependent model for the passage of fluid through a porous medium. To define the model, with each edge $e \in \mathcal{E}(\mathbb{Z}^d)$ is associated a random variable $T(e)$, which can be interpreted as being the time required for fluid to flow along e. The $T(e)$ are assumed to be independent non-negative random variables having common distribution F. For any path π we define the passage time $T(\pi)$ of π as

$$T(\pi) := \sum_{e \in \pi} T(e)$$

The first passage time $a(x, y)$ between vertices x and y is given by

$$a(x, y) = \inf\{T(\pi) : \pi \text{ a path from } x \text{ to } y\}$$

and we can define

$$W(t) := \{x \in \mathbb{Z}^d : a(0, x) \leq t\}$$

the set of vertices reached by the liquid by time t. It turns out that $W(t)$ grows approximately linearly as time passes, and that there exists a nonrandom limit set B such that either B is compact and

$$(1 - \varepsilon)B \subseteq \frac{1}{t}\widetilde{W}(t) \subseteq (1 + \varepsilon)B, \text{ eventually a.s.}$$

for all $\epsilon > 0$, or $B = \mathbb{R}^d$, and

$$\{x \in \mathbb{R}^d : |x| \leq K\} \subseteq \frac{1}{t}\widetilde{W}(t), \text{ eventually a.s.}$$

for all $K > 0$. Here $\widetilde{W}(t) = \{z + [-1/2, 1/2]^d : z \in W(t)\}$.

Studies of first-passage percolation brought many fascinating discoveries, including Kingman's celebrated subadditive ergodic theorem. In recent years interest has been focused on study of fluctuations of the set $\widetilde{W}(t)$ for large t. In spite of huge effort and some partial results achieved, it still remains a major task to establish rigorously conjectures predicted by Kardar–Parisi–Zhang theory about shape fluctuations in first passage percolation.

Contact Processes

Introduced by Harris and conceived with biological interpretation, the contact process on $\mathbb{Z}^d$ is a continuous-time process taking values in the space of subsets of $\mathbb{Z}^d$. It is informally described as follows: particles are distributed in $\mathbb{Z}^d$ in such a way that each site is either empty or occupied by one particle. The evolution is Markovian: each particle disappears after an exponential time of parameter 1, independently from the others; at any time, each particle has a possibility to create a new particle at any of its empty neighboring sites, and does so with rate $\lambda > 0$, independently of everything else.

The question is then whether, starting from a finite population, the process will die out in finite time or whether it will survive forever with positive probability. The outcome will depend on the value of λ, and there is a critical value λ_c, such that for $\lambda \leq \lambda_c$ process dies out, while for $\lambda > \lambda_c$ indeed there is survival, and in this case the shape of the population obeys a shape theorem similar to that of first-passage percolation.

The analogy with percolation is strong, the corresponding percolative picture being the following: in $\mathbb{Z}_+^{d+1}$, each edge is open with probability $p \in (0, 1)$, and the question is whether there exists an infinite oriented path π (i.e., a path along which the sum of the coordinates is increasing), composed of open edges. Once again, there is a critical parameter customarily denoted by p_c, at which no such path exists (compare this to the open question of the continuity of the function θ at p_c in dimensions $3 \leq d \leq 18$). This variation of percolation lies in a different universality class than the usual Bernoulli model.

Invasion Percolation

Let $X(e) : e \in \mathcal{E}$ be independent random variables indexed by the edge set $\mathcal{E}$ of $\mathbb{Z}^d$, $d \geq 2$, each having uniform distribution in $[0, 1]$. One constructs a sequence $C = \{C_i, i \geq 1\}$ of random connected subgraphs of the lattice in the following iterative way: the graph C_0 contains only the origin. Having defined C_i, one obtains C_{i+1} by adding to C_i an edge e_{i+1} (with its outer lying end-vertex), chosen from the outer edge boundary of C_i so as to minimize $X(e_{i+1})$. Still very little is known about the behavior of this process.

An interesting observation, relating $\theta(p_c)$ of usual percolation with the invasion dynamics, comes from CM Newman:

$$\theta(p_c) = 0 \Leftrightarrow P\{x \in C\} \to 0 \text{ as } |x| \to \infty$$

Further Remarks

For a much more in-depth review of percolation on lattices and the mathematical methods involved in its study, and for the proofs of most of the results we could only point at, we refer the reader to the standard book of Grimmett (1999); another excellent general reference, and the only place to find some of the technical graph-theoretical details involved, is the book of Kesten (1982). More information in the case of graphs that are not lattices can be found in the lecture notes of Peres (1999).

For curiosity, the reader can refer to the first mention of a problem close to percolation, in the problem section of the first volume of the *American Mathematical Monthly* (problem 5, June 1894, submitted by D V Wood).

See also: Determinantal Random Fields; Stochastic Loewner Evolutions; Two-Dimensional Ising Model; Wulff Droplets.

Further Reading

Alexander K, Chayes JT, and Chayes L (1990) The Wulff construction and asymptotics of the finite cluster distribution for two-dimensional Bernoulli percolation. *Communications in Mathematical Physics* 131: 1–50.

Beffara V (2004) Hausdorff dimensions for SLE_6. *Annals of Probability* 32: 2606–2629.

Benjamini I and Schramm O (1996) Percolation beyond $\mathbb{Z}^d$, many questions and a few answers. *Electronic Communications in Probability* 1(8): 71–82 (electronic).

Broadbent SR and Hammersley JM (1957) Percolation processes, I and II. *Mathematical Proceedings of the Cambridge Mathematical Society* 53, pp. 629–645.

Cerf R (2000) Large deviations for three dimensional supercritical percolation. *Astérisque, SMF* vol. 267.

Gaboriau D (2005) Invariant percolation and harmonic Dirichlet functions. *Geometric and Functional Analysis* (in print).

Grimmett G (1999) Percolation. *Grundlehren der Mathematischen Wissenschaften*, 2nd edn, vol. 321. Berlin: Springer.

Hammersley JM and Welsh DJA (1965) First-passage percolation, subadditive processes, stochastic networks, and generalized renewal theory. In: *Proc. Internat. Res. Semin*, Statist. Lab., Univ. California, Berkeley, CA, pp. 61–110. New York: Springer.

Hara T and Slade G (1990) Mean-field critical behaviour for percolation in high dimensions. *Communications in Mathematical Physics* 128: 333–391.

Kesten H (1980) The critical probability of bond percolation on the square lattice equals 1/2. *Communications in Mathematical Physics* 74: 41–59.

Kesten H (1982) *Percolation Theory for Mathematicians, Progress in Probability and Statistics*, vol. 2. Boston, MA: Birkhäuser.

Kesten H (1986) The incipient infinite cluster in two-dimensional percolation. *Probability Theory and Related Fields* 73: 369–394.

Kesten H (1987) Scaling relations for 2D-percolation. *Communications in Mathematical Physics* 109: 109–156.

Peres Y (1999) *Probability on Trees: An Introductory Climb*, Lectures on Probability Theory and Statistics (Saint-Flour, 1997), Lecture Notes in Math, vol. 1717, pp. 193–280. Berlin: Springer.

Russo L (1978) A note on percolation. *Zeitschift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 43: 39–48.

Sapoval B, Rosso M, and Gouyet J (1985) The fractal nature of a diffusion front and the relation to percolation. *Journal de Physique Lettres* 46: L146–L156.

Seymour PD and Welsh DJA (1978) Percolation probabilities on the square lattice. *Annals of Discrete Mathematics* 3: 227–245.

Smirnov S (2001) Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits. *Comptes-rendus de l'Académie des Sciences de Paris, Série I Mathématiques* 333: 239–244.

Smirnov S and Werner W (2001) Critical exponents for two-dimensional percolation. *Mathematical Research Letters* 8: 729–744.

van der Hofstad R and Járai AA (2004) The incipient infinite cluster for high-dimensional unoriented percolation. *Journal of Statistical Physics* 114: 625–663.

Perturbation Theory and Its Techniques

R J Szabo, Heriot-Watt University, Edinburgh, UK

Introduction

There are several equivalent formulations of the problem of quantizing an interacting field theory. The list includes canonical quantization, path-integral (or functional) techniques, stochastic quantization, "unified" methods such as the Batalin–Vilkovisky formalism, and techniques based on the realizations of field theories as low-energy limits of string theory. The problem of obtaining an exact nonperturbative description of a given quantum field theory is most often a very difficult one. Perturbative techniques, on the other hand, are abundant, and common to all of the quantization methods mentioned above is that they admit particle interpretations in this formalism.

The basic physical quantities that one wishes to calculate in a relativistic $(d+1)$-dimensional quantum field theory are the S-matrix elements

$$S_{ba} = {}_{\text{out}}\langle\psi_b(t)|\psi_a(t)\rangle_{\text{in}} \qquad [1]$$

between in and out states at large positive time t. The scattering operator S is then defined by writing [1] in terms of initial free-particle (descriptor) states as

$$S_{ba} =: \langle\psi_b(0)|\mathsf{S}|\psi_a(0)\rangle \qquad [2]$$

Suppose that the Hamiltonian of the given field theory can be written as $H = H_0 + H'$, where H_0 is the free part and H' the interaction Hamiltonian. The time evolutions of the in and out states are governed by the total Hamiltonian H. They can be expressed in terms of descriptor states which evolve in time with H_0 in the interaction picture and correspond to free-particle states. This leads to the Dyson formula

$$\mathsf{S} = \mathrm{T}\exp\left(-\mathrm{i}\int_{-\infty}^{\infty}\mathrm{d}t\,H_I(t)\right) \qquad [3]$$

where T denotes time ordering and $H_I(t) = \int \mathrm{d}^d\boldsymbol{x}\,\mathcal{H}_{\text{int}}(\boldsymbol{x},t)$ is the interaction Hamiltonian in the interaction picture, with $\mathcal{H}_{\text{int}}(\boldsymbol{x},t)$ the interaction Hamiltonian density, which deals with essentially free fields. This formula expresses S in terms of interaction-picture operators acting on free-particle states in [2] and is the first step towards Feynman perturbation theory.

For many analytic investigations, such as those which arise in renormalization theory, one is interested instead in the Green's functions of the quantum field theory, which measure the response of the system to an external perturbation. For definiteness, let us consider a free real scalar field theory in $d+1$ dimensions with Lagrangian density

$$\mathcal{L} = \tfrac{1}{2}\partial_\mu\phi\partial^\mu\phi - \tfrac{1}{2}m^2\phi^2 + \mathcal{L}_{\text{int}} \qquad [4]$$

where $\mathcal{L}_{\text{int}}$ is the interaction Lagrangian density which we assume has no derivative terms. The interaction Hamiltonian density is then given by $\mathcal{H}_{\text{int}} = -\mathcal{L}_{\text{int}}$. Introducing a real scalar source $J(x)$, we define the normalized "partition function" through the vacuum expectation values,

$$Z[J] = \frac{\langle 0|\mathsf{S}[J]|0\rangle}{\langle 0|\mathsf{S}[0]|0\rangle} \qquad [5]$$

where $|0\rangle$ is the normalized perturbative vacuum state of the quantum field theory given by (4) (defined to be destroyed by all field annihilation operators), and

$$\mathsf{S}[J] = \mathrm{T}\exp\left(\mathrm{i}\int \mathrm{d}^{d+1}x(\mathcal{L}_{\text{int}} + J(x)\phi(x))\right) \qquad [6]$$

from the Dyson formula. This partition function is the generating functional for all Green's functions of the quantum field theory, which are obtained from [5] by taking functional derivatives with respect to the source and then setting $J(x) = 0$. Explicitly, in a formal Taylor series expansion in J one has

$$Z[J] = \sum_{n=0}^{\infty}\frac{\mathrm{i}^n}{n!}\prod_{i=1}^{n}\int \mathrm{d}^{d+1}x_i J(x_i)\, G^{(n)}(x_1,\ldots,x_n) \qquad [7]$$

whose coefficients are the Green's functions

$$G^{(n)}(x_1,\ldots,x_n) := \frac{\langle 0|\mathrm{T}[\exp\left(\mathrm{i}\int \mathrm{d}^{d+1}x\mathcal{L}_{\text{int}}\right)\phi(x_1)\cdots\phi(x_n)]|0\rangle}{\langle 0|\mathrm{T}\exp\left(\mathrm{i}\int \mathrm{d}^{d+1}x\mathcal{L}_{\text{int}}\right)|0\rangle} \qquad [8]$$

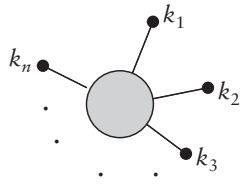
It is customary to work in momentum space by introducing the Fourier transforms

$$\begin{aligned}\tilde{J}(k) &= \int \mathrm{d}^{d+1}x\, \mathrm{e}^{\mathrm{i}k\cdot x}J(x)\\ \tilde{G}^{(n)}(k_1,\ldots,k_n) &= \prod_{i=1}^{n}\int \mathrm{d}^{d+1}x_i\, \mathrm{e}^{\mathrm{i}k_i\cdot x_i}G^{(n)}(x_1,\ldots,x_n)\end{aligned} \qquad [9]$$

in terms of which the expansion [7] reads

$$Z[J] = \sum_{n=0}^{\infty}\frac{\mathrm{i}^n}{n!}\prod_{i=1}^{n}\int\frac{\mathrm{d}^{d+1}k_i}{(2\pi)^{d+1}}\,\tilde{J}(-k_i) \times \tilde{G}^{(n)}(k_1,\ldots,k_n) \qquad [10]$$

The generating functional [10] can be written as a sum of Feynman diagrams with source insertions. Diagrammatically, the Green's function is an infinite series of graphs which can be represented symbolically as

$\tilde{G}^{(n)}(k_1,\ldots,k_n) =$ 

[11]

where the n external lines denote the source insertions of momenta k_i and the bubble denotes the sum over all Feynman diagrams constructed from the interaction vertices of $\mathcal{L}_{\text{int}}$.

This procedure is, however, rather formal in the way that we have presented it, for a variety of reasons. First of all, by Haag's theorem, it follows that the interaction representation of a quantum field theory does not exist unless a cutoff regularization is introduced into the interaction term in the Lagrangian density (this regularization is described explicitly below). The addition of this term breaks translation covariance. This problem can be remedied via a different definition of the regularized Green's functions, as we discuss below. Furthermore, the perturbation series of a quantum field theory is typically divergent. The expansion into graphs is, at best, an asymptotic series which is Borel summable. These shortcomings will not be emphasized any further in this article. Some mathematically rigorous approaches to perturbative quantum field theory can be found in the bibliography.

The Green's functions can also be used to describe scattering amplitudes, but there are two important differences between the graphs [11] and those which appear in scattering theory. In the present case, external lines carry propagators, that is, the free-field Green's functions

$$\begin{aligned}\Delta(x-y) &= \langle 0|\,\mathrm{T}[\phi(x)\,\phi(y)]\,|0\rangle\\ &= \left\langle x\middle|\left(\Box + m^2\right)^{-1}\middle|y\right\rangle\\ &= \int\frac{\mathrm{d}^{d+1}p}{(2\pi)^{d+1}}\frac{\mathrm{i}}{p^2 - m^2 + \mathrm{i}\epsilon}\,\mathrm{e}^{-\mathrm{i}p\cdot(x-y)}\end{aligned} \qquad [12]$$

where $\epsilon \to 0^+$ regulates the mass shell contributions, and their momenta k_i are off-shell in general ($k_i^2 \neq m^2$). By the LSZ theorem, the S-matrix element

is then given by the multiple on-shell residue of the Green's function in momentum space as

$$\begin{aligned}&\langle k'_1,\ldots,k'_n|\,\mathsf{S}-\mathbb{1}|k_1,\ldots,k_l\rangle\\&\quad=\lim_{\substack{k'_1,\ldots,k'_n\to m^2\\k_1,\ldots,k_l\to m^2}}\prod_{i=1}^{n}\frac{1}{\mathrm{i}\sqrt{c'_i}}\left(k'^2_i-m^2\right)\prod_{j=1}^{l}\frac{1}{\mathrm{i}\sqrt{c_j}}\left(k_j^2-m^2\right)\\&\qquad\times\tilde{G}^{(n+m)}\left(-k'_1,\ldots,-k'_n,k_1,\ldots,k_l\right)\end{aligned}\qquad[13]$$

where $\mathrm{i}c'_i,\mathrm{i}c_j$ are the residues of the corresponding particle poles in the exact two-point Green's function.

This article deals with the formal development and computation of perturbative scattering amplitudes in relativistic quantum field theory, along the lines outlined above. Initially we deal only with real scalar field theories of the sort [4] in order to illustrate the concepts and technical tools in as simple and concise a fashion as possible. These techniques are common to most quantum field theories. Fermions and gauge theories are then separately treated afterwards, focusing on the methods which are particular to them.

Diagrammatics

The pinnacle of perturbation theory is the technique of Feynman diagrams. Here we develop the basic machinery in a quite general setting and use it to analyze some generic features of the terms comprising the perturbation series.

Wick's Theorem

The Green's functions [8] are defined in terms of vacuum expectation values of time-ordered products of the scalar field $\phi(x)$ at different spacetime points. Wick's theorem expresses such products in terms of normal-ordered products, defined by placing each field creation operator to the right of each field annihilation operator, and in terms of two-point Green's functions [12] of the free-field theory (propagators). The consequence of this theorem is the Haffnian formula

$$\begin{aligned}&\langle 0|\,\mathrm{T}[\phi(x_1)\cdots\phi(x_n)]|0\rangle\\&\quad=\begin{cases}0 & \\ \qquad n=2k-1 & \\ \displaystyle\sum_{\pi\in S_{2k}}\prod_{i=1}^{k}\langle 0|\,\mathrm{T}\left[\phi(x_{\pi(2i-1)})\,\phi(x_{\pi(2i)})\right]|0\rangle & \\ \qquad n=2k & \end{cases}\end{aligned}\qquad[14]$$

The formal Taylor series expansion of the scattering operator S may now be succinctly summarized into a diagrammatic notation by using Wick's theorem. For each spacetime integration $\int \mathrm{d}^{d+1}x_i$ we introduce a vertex with label i, and from each vertex there emanate some lines corresponding to field insertions at the point x_i. If the operators represented by two lines appear in a two-point function according to [14], that is, they are contracted, then these two lines are connected together. The S operator is then represented as a sum over all such Wick diagrams, bearing in mind that topologically equivalent diagrams correspond to the same term in S. Two diagrams are said to have the same pattern if they differ only by a permutation of their vertices. For any diagram $\mathfrak{D}$ with $n(\mathfrak{D})$ vertices, the number of ways of interchanging vertices is $n(\mathfrak{D})!$. The number of diagrams per pattern is always less than this number. The symmetry number $S(\mathfrak{D})$ of $\mathfrak{D}$ is the number of permutations of vertices that give the same diagram. The number of diagrams with the pattern of $\mathfrak{D}$ is then $n(\mathfrak{D})!/S(\mathfrak{D})$.

In a given pattern, we write the contribution to S of a single diagram $\mathfrak{D}$ as

$$\frac{1}{n(\mathfrak{D})!}:\theta(\mathfrak{D}):$$

where the combinatorial factor comes from the Taylor expansion of S, the large colons denote normal ordering of quantum operators, and $:\theta(\mathfrak{D}):$ contains spacetime integrals over normal-ordered products of the fields. Then all diagrams with the pattern of $\mathfrak{D}$ contribute $:\theta(\mathfrak{D}):/S(\mathfrak{D})$ to S. Only the connected diagrams $\mathfrak{D}_r, r\in\mathbb{N}$ (those in which every vertex is connected to every other vertex) contribute and we can write the scattering operator in a simple form which eliminates contributions from all disconnected diagrams as

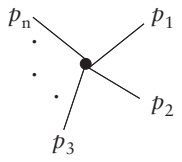
$$\mathsf{S}=:\exp\left(\sum_{r=1}^{\infty}\frac{\theta(\mathfrak{D}_r)}{S(\mathfrak{D}_r)}\right):\qquad[15]$$

Feynman Rules

Feynman diagrams in momentum space are defined from the Wick diagrams above by dropping the labels on vertices (and also the symmetry factors $S(\mathfrak{D})^{-1}$), and by labeling the external lines by the momenta of the initial and final particles that the corresponding field operators annihilate. In a spacetime interpretation, external lines

represent on-shell physical particles while internal lines of the graph represent off-shell virtual particles ($k^2 \neq m^2$). Physical particles interact via the exchange of virtual particles. An arbitrary diagram is then calculated via the Feynman rules:

$$\overset{p}{\longrightarrow} = \int \frac{\mathrm{d}^{d+1}p}{(2\pi)^{d+1}} \; \frac{\mathrm{i}}{p^2 - m^2 + \mathrm{i}\epsilon}$$

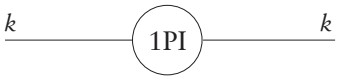


$$= \mathrm{i}g\,(2\pi)^{d+1}\,\delta^{(d+1)}(p_1 + \cdots + p_n) \qquad [16]$$

for a monomial interaction $\mathcal{L}_{\mathrm{int}} = (g/n!)\phi^n$.

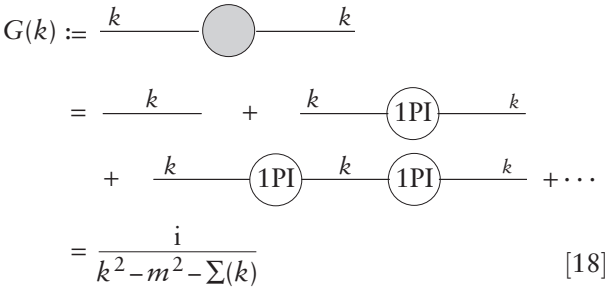
Irreducible Green's Functions

A one-particle irreducible (1PI) or proper Green's function is given by a sum of diagrams in which each diagram cannot be separated by cutting one internal line. In momentum space, it is defined without the overall momentum conservation delta-function factors and without propagators on external lines. For example, the two particle 1PI Green's function



$$=: \Sigma(k) \qquad [17]$$

is called the self-energy. If $G(k)$ is the complete two-point function in momentum space, then one has



$$= \frac{\mathrm{i}}{k^2 - m^2 - \Sigma(k)} \qquad [18]$$

and thus it suffices to calculate only 1PI diagrams.

The 1PI effective action, defined by the Legendre transformation $\Gamma[\phi] := -\mathrm{i}\ln Z[J] - \int \mathrm{d}^{d+1}x J(x)\phi(x)$ of [5], is the generating functional for proper vertex functions and it can be represented as a functional of only the vacuum expectation value of the field ϕ, that is, its classical value. In the semiclassical (WKB) approximation, the one-loop effective action is given by

$$\begin{aligned}\Gamma[\phi] &= S[\phi] + \frac{\mathrm{i}\hbar}{2}\mathrm{Tr}\ln(1 + \Delta V''[\phi]) + O(\hbar^2) \\ &= S[\phi] + \mathrm{i}\hbar \sum_{n=1}^{\infty} \frac{(-1)^n}{2n} \\ &\quad \times \prod_{i=1}^{n} \int \mathrm{d}^{d+1}x_i \Delta(x_i - x_{i+1}) V''[\phi(x_{i+1})] \\ &\quad + O(\hbar^2) \end{aligned} \qquad [19]$$

where we have denoted $S[\phi] = \int \mathrm{d}^{d+1}x\mathcal{L}$ and $V[\phi] = -\mathcal{L}_{\mathrm{int}}$, and for each term in the infinite series we define $x_{n+1} := x_1$. The first term in [19] is the classical contribution and it can be represented in terms of connected tree diagrams. The second term is the sum of contributions of one-loop diagrams constructed from n propagators $-\mathrm{i}\Delta(x - y)$ and n vertices $-\mathrm{i}V''[\phi]$. The expansion may be carried out to all orders in terms of connected Feynman diagrams, and the result of the above Legendre transformation is to select only the one-particle irreducible diagrams and to replace the classical value of ϕ by an arbitrary argument. All information about the quantum field theory is encoded in this effective action.

Parametric Representation

Consider an arbitrary proper Feynman diagram $\mathfrak{D}$ with n internal lines and v vertices. The number, ℓ, of independent loops in the diagram is the number of independent internal momenta in $\mathfrak{D}$ when conservation laws at each vertex have been taken into account, and it is given by $\ell = n + 1 - v$. There is an independent momentum integration variable k_i for each loop, and a propagator for each internal line as in [16]. The contribution of $\mathfrak{D}$ to a proper Green's function with r incoming external momenta p_i, with $\sum_{i=1}^{r} p_i = 0$, is given by

$$\begin{aligned}\tilde{I}_{\mathfrak{D}}(p) = \frac{V(\mathfrak{D})}{S(\mathfrak{D})} \prod_{i=1}^{n} \int \frac{\mathrm{d}^{d+1}k_i}{(2\pi)^{d+1}} \frac{\mathrm{i}}{k_i^2 - m^2 + \mathrm{i}\epsilon} \\ \times \prod_{j=1}^{v} (2\pi)^{d+1} \delta^{(d+1)}\left(P_j - K_j\right) \end{aligned} \qquad [20]$$

where $V(\mathfrak{D})$ contains all contributions from the interaction vertices of $\mathcal{L}_{\mathrm{int}}$, and P_j (resp. K_j) is the sum of incoming external momenta p_{l_j} (resp. internal momenta k_{l_j}) at vertex j with respect to a fixed chosen orientation of the lines of the graph. After resolving the delta-functions in terms of independent internal loop momenta $k_1, \ldots, k_\ell$ and dropping the overall momentum conservation

delta-function along with the symmetry and vertex factors in [20], one is left with a set of momentum space integrals

$$I_{\mathfrak{D}}(p) = \prod_{i=1}^{\ell} \int \frac{\mathrm{d}^{d+1}k_i}{(2\pi)^{d+1}} \prod_{j=1}^{n} \frac{\mathrm{i}}{a_j(k,p) + \mathrm{i}\epsilon} \qquad [21]$$

where $a_j(k,p)$ are functions of both the internal and external momenta.

It is convenient to exponentiate propagators using the Schwinger parametrization

$$\frac{\mathrm{i}}{a_j + \mathrm{i}\epsilon} = \int_0^{\infty} \mathrm{d}\alpha_j \, \mathrm{e}^{\mathrm{i}\alpha_j(a_j + \mathrm{i}\epsilon)} \qquad [22]$$

and after some straightforward manipulations one may write the Feynman parametric formula

$$\prod_{j=1}^{n} \frac{\mathrm{i}}{a_j(k,p) + \mathrm{i}\epsilon} = (n-1)! \prod_{j=1}^{n} \int_0^1 \mathrm{d}\alpha_j \frac{\delta\left(1 - \sum_j \alpha_j\right)}{D_{\mathfrak{D}}(k;\alpha,p)^n} \qquad [23]$$

where $D_{\mathfrak{D}}(k;\alpha,p) := \sum_j \alpha_j [a_j(k,p) + \mathrm{i}\epsilon]$ is generically a quadratic form

$$D_{\mathfrak{D}}(k;\alpha,p) = \frac{1}{2} \sum_{i,j=1}^{\ell} k_i \cdot Q_{ij}(\alpha) k_j + \sum_{i=1}^{\ell} L_i(p) \cdot k_i + \lambda(p^2) \qquad [24]$$

The positive symmetric matrix Q_{ij} is independent of the external momenta p_l, invertible, and has nonzero eigenvalues $Q_1, \ldots, Q_\ell$. The vectors L_i^μ are linear combinations of the p_j^μ, while $\lambda(p^2)$ is a function of only the Lorentz invariants p_i^2. After some further elementary manipulations, the loop diagram contribution [21] may be written as

$$I_{\mathfrak{D}}(p) = (n-1)! \prod_{j=1}^{n} \int_0^1 \mathrm{d}\alpha_j \prod_{i=1}^{\ell} \frac{1}{Q_i(\alpha)^2} \int \frac{\mathrm{d}^{d+1}k_i}{(2\pi)^{d+1}} \delta\left(1 - \sum_j \alpha_j\right) \times \left(\frac{1}{2} \sum_i k_i^2 + \lambda(p^2) - \frac{1}{2} \sum_{i,j} L_i(p) \cdot Q^{-1}(\alpha)_{ij} L_j(p) \right)^{-n} \qquad [25]$$

Finally, the integrals over the loop momenta k_i may be performed by Wick-rotating them to Euclidean space and using the fact that the combination of ℓ integrations in $\mathbb{R}^{d+1}$ has $O((d+1)\ell)$ rotational invariance. The contribution from the entire Feynman diagram $\mathfrak{D}$ thereby reduces to the calculation of the parametric integrals:

$$I_{\mathfrak{D}}(p) = \frac{\Gamma\left(n - \frac{(d+1)\ell}{2}\right)}{(2\pi)^{\frac{(d+1)\ell}{2}} \mathrm{i}^{d\ell}} \prod_{j=1}^{n} \int_0^1 \mathrm{d}\alpha_j \prod_{i=1}^{\ell} \frac{1}{Q_i(\alpha)^2} \times \frac{\delta\left(1 - \sum_j \alpha_j\right)}{\left(\lambda(p^2) - \frac{1}{2} \sum_{i,j} L_i(p) \cdot Q^{-1}(\alpha)_{ij} L_j(p)\right)^{\frac{n-(d+1)\ell}{2}}} \qquad [26]$$

where $\Gamma(s)$ is the Euler gamma-function.

Regularization

The parametric representation [26] is generically convergent when $2n - (d+1)\ell > 0$. When divergent, the infinities arise from the lower limits of integration $\alpha_j \to 0$. This is just the parametric representation of the large-k divergence of the original Feynman amplitude [20]. Such ultraviolet divergences plague the very meaning of a quantum field theory and must be dealt with in some way. We will now quickly tour the standard methods of ultraviolet regularization for such loop integrals, which is a prelude to the renormalization program that removes the divergences (in a renormalizable field theory). Here we consider regularization simply as a means of justification for the various formal manipulations that are used in arriving at expressions such as [26].

Momentum Cutoff

Cutoff regularization introduces a mass scale Λ into the quantum field theory and throws away the Fourier modes of the fields for spatial momenta $\boldsymbol{k}$ with $|\boldsymbol{k}| > \Lambda$. This regularization spoils Lorentz invariance. It is also nonlocal. For example, if we restrict to a hypercube in momentum space, so that $|k_i| < \Lambda$ for $i = 1, \ldots, d$, then

$$\int_{|\boldsymbol{k}|>\Lambda} \frac{\mathrm{d}^d \boldsymbol{k}}{(2\pi)^d} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} = \prod_{i=1}^{d} \frac{\sin(\Lambda x^i)}{\pi x^i}$$

which is a delta-function in the limit $\Lambda \to \infty$ but is nonlocal for $\Lambda < \infty$. The regularized field theory is finite order by order in perturbation theory and depends on the cutoff Λ.

Lattice Regularization

We can replace the spatial continuum by a lattice $\mathfrak{L}$ of rank d and define a Lagrangian on $\mathfrak{L}$ by

$$L_{\mathfrak{L}} = \frac{1}{2}\sum_{i\in \mathrm{S}(\mathfrak{L})} \dot{\phi}_i^2 + J \sum_{\langle i,j\rangle \in \mathrm{L}(\mathfrak{L})} \phi_i\phi_j + \sum_{i\in \mathrm{S}(\mathfrak{L})} V(\phi_i) \quad [27]$$

where $\mathrm{S}(\mathfrak{L})$ is the set of sites i of the lattice on each of which is situated a time-dependent function ϕ_i, and $L_{\mathfrak{L}}$ is the collection of links connecting pairs $\langle i,j\rangle$ of nearest-neighbor sites i,j on $\mathfrak{L}$. The regularized field theory is now local, but still has broken Lorentz invariance. In particular, it suffers from broken rotational symmetry. If $\mathfrak{L}$ is hypercubic with lattice spacing a, that is, $\mathfrak{L} = (\mathbb{Z}a)^d$, then the momentum cutoff is at $\Lambda = a^{-1}$.

Pauli–Villars Regularization

We can replace the propagator $\mathrm{i}(k^2 - m^2 + \mathrm{i}\epsilon)^{-1}$ by $\mathrm{i}(k^2 - m^2 + \mathrm{i}\epsilon)^{-1} + \mathrm{i}\sum_{j=1}^{N} c_j(k^2 - M_j^2 + \mathrm{i}\epsilon)^{-1}$, where the masses $M_j \gg m$ are identified with the momentum cutoff as $\min\{M_j\} = \Lambda \to \infty$. The mass-dependent coefficients c_j are chosen to make the modified propagator decay rapidly as $(k^2)^{-N-1}$ at $k \to \infty$, which gives the N equations $(m^2)^i + \sum_j c_j (M_j^2)^i = 0, i = 0, 1, \ldots, N-1$. This regularization preserves Lorentz invariance (and other symmetries that the field theory may possess) and is local in the following sense. The modified propagator can be thought of as arising through the alteration of the Lagrangian density [4] by N additional scalar fields φ_j of masses M_j with

$$\mathcal{L}_{\mathrm{PV}} = \tfrac{1}{2}\partial_\mu\phi\partial^\mu\phi - \tfrac{1}{2}m^2\phi^2 + \sum_{j=1}^{N}\left(\tfrac{1}{2}\partial_\mu\varphi_j\partial^\mu\varphi_j - \tfrac{1}{2}M_j^2\varphi_j^2\right) + \mathcal{L}_{\mathrm{int}}[\Phi] \quad [28]$$

where $\Phi := \phi + \sum_j \sqrt{c_j}\varphi_j$. The contraction of the Φ field thus produces the required propagator. However, the c_j's as computed above are generically negative numbers and so the Lagrangian density [28] is not Hermitian (as $\Phi \neq \Phi^\dagger$). It is possible to make [28] formally Hermitian by redefining the inner product on the Hilbert space of physical states, but this produces negative-norm states. This is no problem at energy scales $E \ll M_j$ on which the extra particles decouple and the negative probability states are invisible.

Dimensional Regularization

Consider a Euclidean space integral $\int \mathrm{d}^4k(k^2 + a^2)^{-r}$ arising after Wick rotation from some loop diagram in $(3+1)$-dimensional scalar field theory. We replace this integral by its D-dimensional version

$$\int \frac{\mathrm{d}^D k}{(k^2 + a^2)^r} = \frac{\pi^{D/2}(a^2)^{D/2-r}}{(r-1)!}\Gamma\left(r - \frac{D}{2}\right) \quad [29]$$

This integral is absolutely convergent for $D < 2r$. We can analytically continue the result of this integration to the complex plane $D \in \mathbb{C}$. As an analytic function, the only singularities of the Euler function $\Gamma(z)$ are poles at $z = 0, -1, -2, \ldots$. In particular, $\Gamma(z)$ has a simple pole at $z = 0$ of residue 1. If we write $D = 4 + \epsilon$ with $|\epsilon| \to 0$, then the integral [29] is proportional to $\Gamma(r - 2 - \epsilon/2)$ and ϵ plays the role of the regulator here. This regularization is Lorentz invariant (in D dimensions) and is distinguished as having a dimensionless regularization parameter ϵ. This parameter is related to the momentum cutoff Λ by $\epsilon^{-1} = \ln(\Lambda/m)$, so that the limit $\epsilon \to 0$ corresponds to $\Lambda \to \infty$.

Infrared Divergences

Thus far we have only considered the ultraviolet behavior of loop amplitudes in quantum field theory. When dealing with massless particles ($m = 0$ in [4]) one has to further worry about divergences arising from the $k \to 0$ regions of Feynman integrals. After Wick rotation to Euclidean momenta, one can show that no singularities arise in a given Feynman diagram as some of its internal masses vanish provided that all vertices have superficial degree of divergence $d + 1$, the external momenta are not exceptional (i.e., no partial sum of the incoming momenta p_i vanishes), and there is at most one soft external momentum. This result assumes that renormalization has been carried out at some fixed Euclidean point. The extension of this property when the external momenta are continued to physical on-shell values is difficult. The Kinoshita–Lee–Nauenberg theorem states that, as a consequence of unitarity, transition probabilities in a theory involving massless particles are finite when the sum over all degenerate states (initial and final) is taken. This is true order by order in perturbation theory in bare quantities or if minimal subtraction renormalization is used (to avoid infrared or mass singularities in the renormalization constants).

Fermion Fields

We will now leave the generalities of our pure scalar field theory and start considering the extensions of our previous considerations to other types of particles. Henceforth we will primarily deal with the case of $(3+1)$-dimensional spacetime. We begin by indicating how the rudiments of perturbation

theory above apply to the case of Dirac fermion fields. The Lagrangian density is

$$\mathcal{L}_F = \overline{\psi}(\mathrm{i}\partial\!\!\!/ - m)\psi + \mathcal{L}' \qquad [30]$$

where ψ are four-component Dirac fermion fields in $3+1$ dimensions, $\overline{\psi} := \psi^\dagger\gamma^0$ and $\partial\!\!\!/ = \gamma^\mu\partial_\mu$ with γ^μ the generators of the Clifford algebra $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$. The Lagrangian density $\mathcal{L}'$ contains couplings of the Dirac fields to other field theories, such as the scalar field theories considered previously.

Wick's theorem for anticommuting Fermi fields leads to the Pfaffian formula

$$\langle 0|\mathrm{T}[\psi(1)\cdots\psi(n)]|0\rangle = \begin{cases} 0, & n = 2k-1 \\ \dfrac{1}{2^k k!}\displaystyle\sum_{\pi\in S_{2k}} \mathrm{sgn}(\pi) & \\ \quad\times \displaystyle\prod_{i=1}^{k}\langle 0|\mathrm{T}[\psi(\pi(2i-1))\psi(\pi(2i))]|0\rangle & \\ n = 2k & \end{cases} \qquad [31]$$

where for compactness we have written in the argument of $\psi(i)$ the spacetime coordinate, the Dirac index, and a discrete index which distinguishes ψ from $\overline{\psi}$. The nonvanishing contractions in [31] are determined by the free-fermion propagator

$$\begin{aligned}\Delta_{\mathrm{F}}(x-y) &= \langle 0|\mathrm{T}[\psi(x)\overline{\psi}(y)]|0\rangle \\ &= \left\langle x\middle|(\mathrm{i}\partial\!\!\!/ - m)^{-1}\middle|y\right\rangle \\ &= \mathrm{i}\int\frac{\mathrm{d}^4p}{(2\pi)^4}\frac{p\!\!\!/ + m}{p^2 - m^2 + \mathrm{i}\epsilon}\mathrm{e}^{-\mathrm{i}p\cdot(x-y)}\end{aligned} \qquad [32]$$

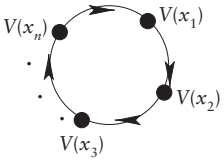
Perturbation theory now proceeds exactly as before. Suppose that the coupling Lagrangian density in [30] is of the form $\mathcal{L}' = \overline{\psi}(x)V(x)\psi(x)$. Both the Dyson formula [3] and the diagrammatic formula [15] are formally the same in this instance. For example, in the formal expansion in powers of $\int \mathrm{d}^4x\mathcal{L}'$, the vacuum-to-vacuum amplitude (the denominator in [5]) will contain field products of the form

$$\prod_{i=1}^{n}\int \mathrm{d}^4x_i\langle 0|\mathrm{T}[\overline{\psi}(x_i)V(x_i)\psi(x_i)]|0\rangle$$

which correspond to fermion loops. Before applying Wick's theorem, the fields must be rearranged as

$$\mathrm{tr}\prod_{i=1}^{n}V(x_i)\psi(x_i)\overline{\psi}(x_{i+1})$$

(with $x_{n+1} := x_1$), where tr is the 4×4 trace over spinor indices. This reordering introduces the familiar minus sign for a closed fermion loop, and one has



$$= (-)\prod_{i=1}^{n}\int \mathrm{d}^4x_i \times \mathrm{tr}\prod_{j=1}^{n-1}\Delta_F(x_j - x_{j+1}) \times V(x_{j+1})\,\Delta_F(x_{j+1} - x_{j+2}) \qquad [33]$$

Feynman rules are now described as follows. Fermion lines are oriented to distinguish a particle from its corresponding antiparticle, and carry both a four-momentum label p as well as a spin polarization index $r = 1, 2$. Incoming fermions (resp. antifermions) are described by the wave functions $u_p^{(r)}$ (resp. $\overline{v}_p^{(r)}$), while outgoing fermions (resp. antifermions) are described by the wave functions $\overline{u}_p^{(r)}$ (resp. $v_p^{(r)}$). Here $u_p^{(r)}$ and $v_p^{(r)}$ are the classical spinors, that is, the positive and negative-energy solutions of the Dirac equation $(p\!\!\!/ - m)u_p^{(r)} = (p\!\!\!/ + m)v_p^{(r)} = 0$. Matrices are multiplied along a Fermi line, with the head of the arrow on the left. Closed fermion loops produce an overall minus sign as in [33], and the multiplication rule gives the trace of Dirac matrices along the lines of the loop. Unpolarized scattering amplitudes are summed over the spins of final particles and averaged over the spins of initial particles using the completeness relations for spinors

$$\sum_{r=1,2}u_p^{(r)}\overline{u}_p^{(r)} = p\!\!\!/ + m, \quad \sum_{r=1,2}v_p^{(r)}\overline{v}_p^{(r)} = p\!\!\!/ - m \qquad [34]$$

leading to basis-independent results. Polarized amplitudes are computed using the spinor bilinears $\overline{u}_p^{(r)}\gamma^\mu u_p^{(s)} = \overline{v}_p^{(r)}\gamma^\mu v_p^{(s)} = 2p^\mu\delta^{rs}$, $\overline{u}_p^{(r)}u_p^{(s)} = -\overline{v}_p^{(r)}v_p^{(s)} = 2m\delta^{rs}$, and $\overline{u}_p^{(r)}v_p^{(s)} = 0$.

When calculating fermion loop integrals using dimensional regularization, one utilizes the Dirac algebra in D dimensions

$$\begin{aligned}\gamma^\mu\gamma_\mu &= \eta^\mu_\mu = D \\ \gamma_\mu p\!\!\!/\,\gamma^\mu &= (2-D)p\!\!\!/ \\ \gamma^\mu p\!\!\!/\,k\!\!\!/\,\gamma_\mu &= 4p\cdot k + (D-4)p\!\!\!/\,k\!\!\!/ \\ \gamma^\mu p\!\!\!/\,k\!\!\!/\,q\!\!\!/\,\gamma_\mu &= -2q\!\!\!/\,k\!\!\!/\,p\!\!\!/ - (D-4)p\!\!\!/\,k\!\!\!/\,q\!\!\!/ \\ \mathrm{tr}\mathbb{1} = 4, \mathrm{tr}\gamma^{\mu_1}\cdots\gamma^{\mu_{2k-1}} &= 0, \mathrm{tr}\gamma^\mu\gamma^\nu = 4\eta^{\mu\nu} \\ \mathrm{tr}\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma &= 4(\eta^{\mu\nu}\eta^{\rho\sigma} - \eta^{\mu\rho}\eta^{\nu\sigma} \\ &\quad + \eta^{\mu\sigma}\eta^{\nu\rho})\end{aligned} \qquad [35]$$

Specific to $D=4$ dimensions are the trace identities

$$\begin{aligned} \mathrm{tr}\gamma^5 &= \mathrm{tr}\gamma^\mu\gamma^\nu\gamma^5 = 0, \\ \mathrm{tr}\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma\gamma^5 &= -4\mathrm{i}\epsilon^{\mu\nu\rho\sigma} \end{aligned} \quad [36]$$

where $\gamma^5 := \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3$. Finally, loop diagrams evaluated with the fermion propagator [32] require a generalization of the momentum space integral [29] given by

$$\begin{aligned} &\int \frac{\mathrm{d}^D k}{(2\pi)^D} \frac{1}{(k^2+2k\cdot p+a^2+\mathrm{i}\epsilon)^r} \\ &= \frac{\mathrm{i}(-\pi)^{D/2}\Gamma\left(r-\frac{D}{2}\right)}{(2\pi)^D(r-1)!} \frac{1}{(a^2-p^2+\mathrm{i}\epsilon)^{r-D/2}} \end{aligned} \quad [37]$$

From this formula we can extract expressions for more complicated Feynman integrals which are tensorial, that is, which contain products of momentum components k^μ in the numerators of their integrands, by differentiating [37] with respect to the external momentum p^μ.

Gauge Fields

The issues we have dealt with thus far have interesting difficulties when dealing with gauge fields. We will now discuss some general aspects of the perturbation expansion of gauge theories using as prototypical examples quantum electrodynamics (QED) and quantum chromodynamics (QCD) in four spacetime dimensions.

Quantum Electrodynamics

Consider the QED Lagrangian density

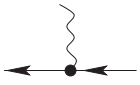
$$\begin{aligned} \mathcal{L}_{\mathrm{QED}} &= -\tfrac{1}{4}F_{\mu\nu}F^{\mu\nu} \\ &\quad + \tfrac{1}{2}\mu^2 A_\mu A^\mu + \overline{\psi}(\mathrm{i}\not{\partial} - e\not{A} - m)\psi \end{aligned} \quad [38]$$

where A_μ is a U(1) gauge field in $3+1$ dimensions and $F_{\mu\nu}=\partial_\mu A_\nu - \partial_\nu A_\mu$ is its field strength tensor. We have added a small mass term $\mu \to 0$ for the gauge field, which at the end of calculations should be taken to vanish in order to describe real photons (as opposed to the soft photons described by [38]). This is done in order to cure the infrared divergences generated in scattering amplitudes due to the masslessness of the photon, that is, the long-range nature of the electromagnetic interaction. The Bloch–Nordsieck theorem in QED states that infrared divergences cancel for physical processes, that is, for processes with an arbitrary number of undetectable soft photons.

Perturbation theory proceeds in the usual way via the Dyson formula, Wick's theorem, and Feynman diagrams. The gauge field propagator is given by

$$\begin{aligned} &\langle 0|\mathrm{T}\left[A_\mu(x)A_\nu(y)\right]|0\rangle \\ &= \langle x|\left[\eta_{\mu\nu}(\Box+\mu^2) - \partial_\mu\partial_\nu\right]^{-1}|y\rangle \\ &= \mathrm{i}\int \frac{\mathrm{d}^4 p}{(2\pi)^4} \frac{-\eta_{\mu\nu}+\frac{p_\mu p_\nu}{\mu^2}}{p^2-\mu^2+\mathrm{i}\epsilon} \mathrm{e}^{-\mathrm{i}p\cdot(x-y)} \end{aligned} \quad [39]$$

and is represented by a wavy line. The fermion–fermion–photon vertex is



$$= -\mathrm{i}e\gamma_\mu \quad [40]$$

An incoming (resp. outgoing) soft photon of momentum k and polarization r is described by the wave function $e_\mu^{(r)}(k)$ (resp. $e_\mu^{(r)}(k)^*$), where the polarization vectors $e_\mu^{(r)}(k), r=1,2,3$ solve the vector field wave equation $(\Box+\mu^2)A_\mu = \partial_\mu A^\mu = 0$ and obey the orthonormality and completeness conditions

$$\begin{aligned} e^{(r)}(k)^*\cdot e^{(s)}(k) &= -\delta^{rs} \\ \sum_{r=1}^{3} e_\mu^{(r)}(k)e_\nu^{(r)}(k)^* &= -\eta_{\mu\nu} + \frac{k_\mu k_\nu}{\mu^2} \end{aligned} \quad [41]$$

along with $k\cdot e^{(r)}(k)=0$. All vector indices are contracted along the lines of the Feynman graph. All other Feynman rules are as previously.

Quantum Chromodynamics

Consider nonabelian gauge theory in $3+1$ dimensions minimally coupled to a set of fermion fields $\psi^A, A=1,\ldots,N_f$, each transforming in the fundamental representation of the gauge group G whose generators T^a satisfy the commutation relations $[T^a,T^b]=f^{abc}T^c$. The Lagrangian density is given by

$$\begin{aligned} \mathcal{L}_{\mathrm{QCD}} &= -\frac{1}{4}F^a_{\mu\nu}F^{a\,\mu\nu} + \frac{1}{2\alpha}\left(\partial_\mu A^{a\,\mu}\right)^2 + \partial_\mu\overline{\eta}\,D^\mu\eta \\ &\quad + \sum_{A=1}^{N_f}\overline{\psi}^A(\mathrm{i}\not{D} - m_A)\psi^A \end{aligned} \quad [42]$$

where $F^a_{\mu\nu}=\partial_\mu A^a_\nu - \partial_\nu A^a_\mu + f^{abc}A^b_\mu A^c_\nu$ and $D_\mu = \partial_\mu + \mathrm{i}eR(T^a)A^a_\mu$, with R the pertinent representation of G ($R(T^a)_{bc}=f^a_{bc}$ for the adjoint representation and $R(T^a)=T^a$ for the fundamental representation). The first term is the Yang–Mills Lagrangian density, the second term is the covariant gauge-fixing term, and the third term contains the Faddeev–Popov ghost fields η which transform in the adjoint representation of the gauge group.

Feynman rules are straightforward to write down and are given in Figure 1 where wavy lines

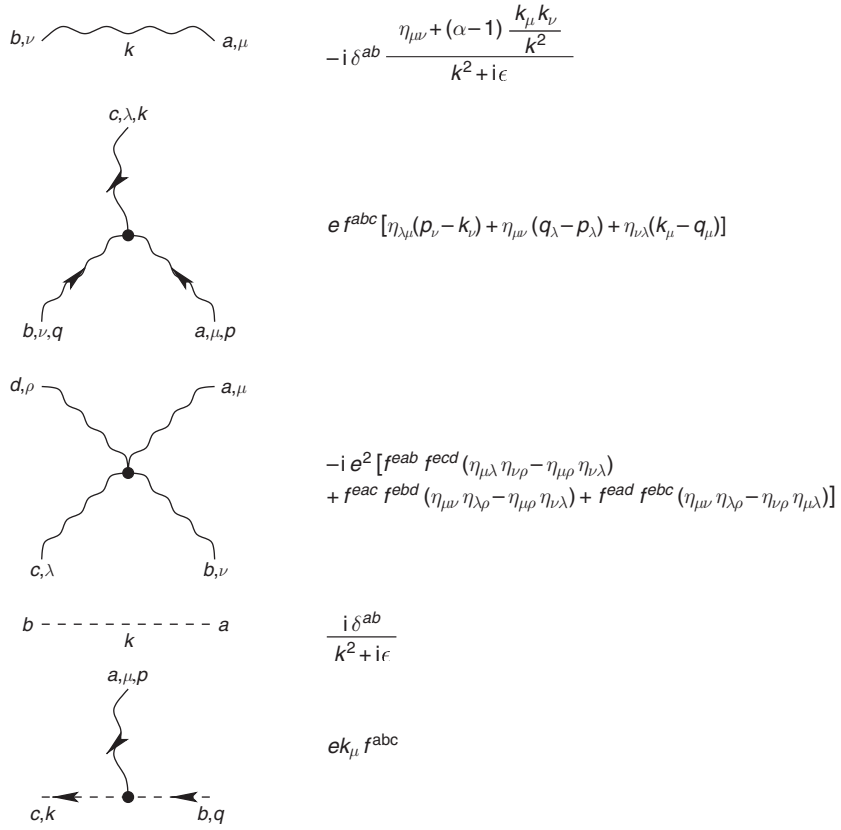


Figure 1 Feynman rules.

represent gluons and dashed lines represent ghosts. Feynman rules for the fermions are exactly as before, except that now the vertex [40] is multiplied by the color matrix T^a. All color indices are contracted along the lines of the Feynman graph. Color factors may be simplified by using the identities

$$\begin{aligned}&\mathrm{Tr}\,R^a\,R^b=\frac{\dim R}{\dim G}\,C_2(R)\delta^{ab},\quad R^a\,R^a=C_2(R)\\&R^a\,R^b\,R^a=\left(C_2(R)-\frac{1}{2}\,C_2(G)\right)R^b\end{aligned}\qquad[43]$$

where $R^a:=R(T^a)$ and $C_2(R)$ is the quadratic Casimir invariant of the representation R (with value $C_2(G)$ in the adjoint representation). For $G=\mathrm{SU}(N)$, one has $C_2(G)=N$ and $C_2(N)=(N^2-1)/2N$ for the fundamental representation.

The cancellation of infrared divergences in loop amplitudes of QCD is far more delicate than in QED, as there is no analog of the Bloch–Nordsieck theorem in this case. The Kinoshita–Lee–Nauenberg theorem guarantees that, at the end of any perturbative calculation, these divergences must cancel for any appropriately defined physical quantity. However, at a given order of perturbation theory, a physical quantity typically involves both virtual and real emission contributions that are separately infrared divergent. Already at two-loop level these divergences have a highly intricate structure. Their precise form is specified by the Catani color-space factorization formula, which also provides an efficient way of organizing amplitudes into divergent parts, which ultimately drop out of physical quantities, and finite contributions.

The computation of multigluon amplitudes in nonabelian gauge theory is rather complicated when one uses polarization states of vector bosons. A much more efficient representation of amplitudes is provided by adopting a helicity (or circular polarization) basis for external gluons. In the spinor–helicity formalism, one expresses positive and negative-helicity polarization vectors in terms of massless Weyl spinors $|k^\pm\rangle:=\frac{1}{2}(\mathbb{1}\pm\gamma_5)u_k=\frac{1}{2}(\mathbb{1}\pm\gamma_5)v_k$ through

$$e_\mu^\pm(k;q)=\pm\frac{\langle q^\mp|\gamma_\mu|k^\mp\rangle}{\sqrt{2}\langle q^\mp|k^\pm\rangle}\qquad[44]$$

where q is an arbitrary null reference momentum which drops out of the final gauge-invariant amplitudes. The spinor products are crossing symmetric, antisymmetric in their arguments, and satisfy the identities

$$\left\langle k_i^- | k_j^+ \right\rangle \left\langle k_j^+ | k_i^- \right\rangle = 2k_i \cdot k_j$$
$$\left\langle k_i^- \middle| k_j^+ \right\rangle \langle k_l^- | k_r^+ \rangle = \langle k_i^- | k_r^+ \rangle \left\langle k_l^- | k_j^+ \right\rangle + \langle k_i^- | k_l^+ \rangle \left\langle k_j^- | k_r^+ \right\rangle \quad [45]$$

Any amplitude with massless external fermions and vector bosons can be expressed in terms of spinor products. Conversely, the spinor products offer the most compact representation of helicity amplitudes which can be related to more conventional amplitudes described in terms of Lorentz invariants. For loop amplitudes, one uses a dimensional regularization scheme in which all helicity states are kept four dimensional and only internal loop momenta are continued to $D = 4 + \epsilon$ dimensions.

Computing Loop Integrals

At the very heart of perturbative quantum field theory is the problem of computing Feynman integrals for multiloop scattering amplitudes. The integrations typically involve serious technical challenges and for the most part are intractable by straightforward analytical means. We will now survey some of the computational techniques that have been developed for calculating quantum loop amplitudes which arise in the field theories considered previously.

Asymptotic Expansion

In many physical instances one is interested in scattering amplitudes in certain kinematical limits. In this case one may perform an asymptotic expansion of multiloop diagrams whose coefficients are typically nonanalytic functions of the perturbative expansion parameter $\hbar$. The main simplification which arises comes from the fact that the expansions are done before any momentum integrals are evaluated. In the limits of interest, Taylor series expansions in different selected regions of each loop momentum can be interpreted in terms of subgraphs and co-subgraphs of the original Feynman diagram.

Consider a Feynman diagram $\mathfrak{D}$ which depends on a collection $\{Q_i\}$ of large momenta (or masses), and a collection $\{m_i, q_i\}$ of small masses and momenta. The prescription for the large-momentum asymptotic expansion of $\mathfrak{D}$ may be summarized in the diagrammatic formula

$$\lim_{Q\to\infty} \mathfrak{D}(Q; m, q) = \sum_{\mathfrak{d}\subset\mathfrak{D}} (\mathfrak{D}/\mathfrak{d})(m, q) \star \left(\mathcal{T}_{\{m_\mathfrak{d}, q_\mathfrak{d}\}}\mathfrak{d}\right)(Q; m_\mathfrak{d}, q_\mathfrak{d}) \quad [46]$$

where the sum runs through all subgraphs $\mathfrak{d}$ of $\mathfrak{D}$ which contain all vertices where a large momentum enters or leaves the graph and is one-particle irreducible after identifying these vertices. The operator $\mathcal{T}_{\{m_\mathfrak{d}, q_\mathfrak{d}\}}$ performs a Taylor series expansion before any integration is carried out, and the notation $(\mathfrak{D}/\mathfrak{d}) \star (\mathcal{T}_{\{m_\mathfrak{d}, q_\mathfrak{d}\}}\mathfrak{d})$ indicates that the subgraph $\mathfrak{d} \subset \mathfrak{D}$ is replaced by its Taylor expansion in all masses and external momenta of $\mathfrak{d}$ that do not belong to the set $\{Q_i\}$. The external momenta of $\mathfrak{d}$ which become loop momenta in $\mathfrak{D}$ are also considered to be small. The loop integrations are then performed only after all these expansions have been carried out. The diagrams $\mathfrak{D}/\mathfrak{d}$ are called co-subgraphs.

The subgraphs become massless integrals in which the scales are set by the large momenta. For instance, in the simplest case of a single large momentum Q one is left with integrals over propagators. The co-subgraphs may contain small external momenta and masses, but the resulting integrals are typically much simpler than the original one. A similar formula is true for large-mass expansions, with the vertex conditions on subdiagrams replace by propagator conditions. For example, consider the asymptotic expansion of the two-loop double bubble diagram (Figure 2) in the region $q^2 \ll m^2$, where m is the mass of the inner loop. The subgraphs (to the right of the stars) are expanded in all external momenta including q and reinserted into the fat vertices of the co-subgraphs (to the left of the stars). Once such asymptotic expansions are carried out, one may attempt to reconstruct as much information as possible about the given scattering amplitude

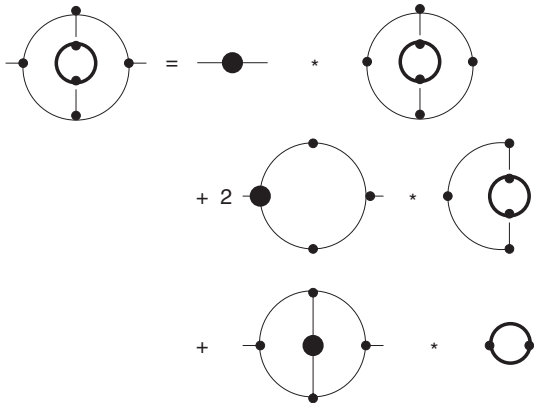


Figure 2 Asymptotic expansion of the two-loop double bubble diagram.

by using the method of Padé approximation which requires knowledge of only part of the expansion of the diagram. By construction, the Padé approximation has the same analytic properties as the exact amplitude.

Brown–Feynman Reduction

When considering loop diagrams which involve fermions or gauge bosons, one encounters tensorial Feynman integrals. When these involve more than three distinct denominator factors (propagators), they require more than two Feynman parameters for their evaluation and become increasingly complicated. The Brown–Feynman method simplifies such higher-rank integrals and effectively reduces them to scalar integrals which typically require fewer Feynman parameters for their evaluation.

To illustrate the idea behind this method, consider the one-loop rank-3 tensor Feynman integral

$$J^{\mu\nu\lambda} = \int \frac{\mathrm{d}^D k}{(2\pi)^D} \times \frac{k^\mu k^\nu k^\lambda}{k^2(k^2-\mu^2)(q-k)^2((k-q)^2+\mu^2)(k^2+2k\cdot p)} \quad [47]$$

where p and q are external momenta with the mass-shell conditions $p^2=(p-q)^2=m^2$. By Lorentz invariance, the general structure of the integral [47] will be of the form

$$J^{\mu\nu\lambda} = a^{\mu\nu}p^\lambda + b^{\mu\nu}q^\lambda + c^\mu s^{\nu\lambda} + c^\nu s^{\mu\lambda} \quad [48]$$

where $a^{\mu\nu}, b^{\mu\nu}$ are tensor-valued functions and c^μ a vector-valued function of p and q. The symmetric tensor $s^{\mu\nu}$ is chosen to project out components of vectors transverse to both p and q, i.e., $p_\mu s^{\mu\nu} = q_\mu s^{\mu\nu} = 0$, with the normalization $s_\mu{}^\mu = D-2$. Solving these constraints leads to the explicit form

$$s^{\mu\nu} = \eta^{\mu\nu} - \frac{m^2 q^\mu q^\nu + q^2 p^\mu p^\nu - (p\cdot q)(q^\mu p^\nu + p^\mu q^\nu)}{m^2 q^2 - (p\cdot q)^2} \quad [49]$$

To determine the as yet unknown functions $a^{\mu\nu}, b^{\mu\nu}$ and c^μ above, we first contract both sides of the decomposition [48] with p^μ and q^μ to get

$$\begin{aligned} 2p_\lambda J^{\mu\nu\lambda} &= 2m^2 a^{\mu\nu} + 2(p\cdot q)b^{\mu\nu} \\ 2q_\lambda J^{\mu\nu\lambda} &= 2(p\cdot q)a^{\mu\nu} + 2q^2 b^{\mu\nu} \end{aligned} \quad [50]$$

Inside the integrand of [47], we then use the trivial identities

$$\begin{aligned} 2k\cdot p &= (k^2+2k\cdot p) - k^2 \\ 2q\cdot k &= k^2 + q^2 - (k-q)^2 \end{aligned} \quad [51]$$

to write the left-hand sides of [50] as the sum of rank-2 Feynman integrals which, with the exception of the one multiplied by q^2 from [51], have one less denominator factor. This formally determines the coefficients $a^{\mu\nu}$ and $b^{\mu\nu}$ in terms of a set of rank-2 integrations. The vector function c^μ is then found from the contraction

$$J^{\mu\nu}{}_\nu = p_\nu a^{\mu\nu} + q_\nu b^{\mu\nu} + (D-2)c^\mu \quad [52]$$

This contraction eliminates the k^2 denominator factor in the integrand of [47] and produces a vector-valued integral. Solving the system of algebraic equations [50] and [52] then formally determines the rank-3 Feynman integral [47] in terms of rank-1 and rank-2 Feynman integrals. The rank-2 Feynman integrals thus generated can then be evaluated in the same way by writing a decomposition for them analogous to [48] and solving for them in terms of vector-valued and scalar-valued Feynman integrals. Finally, the rank-1 integrations can be solved for in terms of a set of scalar-valued integrals, most of which have fewer denominator factors in their integrands.

Generally, any one-loop amplitude can be reduced to a set of basic integrals by using the Passarino–Veltman reduction technique. For example, in supersymmetric amplitudes of gluons any tensor Feynman integral can be reduced to a set of scalar integrals, that is, Feynman integrals in a scalar field theory with a massless particle circulating in the loop, with rational coefficients. In the case of $\mathcal{N}=4$ supersymmetric Yang–Mills theory, only scalar box integrals appear.

Reduction to Master Integrals

While the Brown–Feynman and Passarino–Veltman reductions are well suited for dealing with one-loop diagrams, they become rather cumbersome for higher-loop computations. There are other more powerful methods for reducing general tensor integrals into a basis of known integrals called master integrals. Let us illustrate this technique on a scalar example. Any scalar massless two-loop Feynman integral can be brought into the form

$$I(p) = \int \frac{\mathrm{d}^D k}{(2\pi)^D} \int \frac{\mathrm{d}^D k'}{(2\pi)^D} \prod_{j=1}^{t} \Delta_j^{-l_j} \prod_{i=1}^{q} \Sigma_i^{n_i} \quad [53]$$

where Δ_j are massless scalar propagators depending on the loop momenta k, k' and the external momenta $p_1, \ldots, p_n$, and Σ_i are scalar products of a loop momentum with an external momentum or of the two loop momenta. The topology of the corresponding Feynman diagram is uniquely determined by specifying the set $\Delta_1, \ldots, \Delta_t$ of t distinct

propagators in the graph, while the integral itself is specified by the powers $l_j \geq 1$ of all propagators, by the selection $\Sigma_1, \ldots, \Sigma_q$ of q scalar products and by their powers $n_i \geq 0$.

The integrals in a class of diagrams of the same topology with the same denominator dimension $r = \sum_j l_j$ and same total scalar product number $s = \sum_i n_i$ are related by various identities. One class follows from the fact that the integral over a total derivative with respect to any loop momentum vanishes in dimensional regularization as

$$\int \frac{\mathrm{d}^D k}{(2\pi)^D} \frac{\partial J(k)}{\partial k^\mu} = 0$$

where $J(k)$ is any tensorial combination of propagators, scalar products and loop momenta. The resulting relations are called integration-by-parts identities and for two-loop integrals can be cast into the form

$$\begin{aligned} \int \frac{\mathrm{d}^D k}{(2\pi)^D} \int \frac{\mathrm{d}^D k'}{(2\pi)^D} v^\mu \frac{\partial f(k,k',p)}{\partial k^\mu} &= 0 \\ = \int \frac{\mathrm{d}^D k}{(2\pi)^D} \int \frac{\mathrm{d}^D k'}{(2\pi)^D} v^\mu \frac{\partial f(k,k',p)}{\partial k'^\mu} & \qquad [54] \end{aligned}$$

where $f(k,k',p)$ is a scalar function containing propagators and scalar products, and v^μ is any internal or external momentum. For a graph with ℓ loops and n independent external momenta, this results in a total of $\ell(n+\ell)$ relations.

In addition to these identities, one can also exploit the fact that all Feynman integrals [53] are Lorentz scalars. Under an infinitesimal Lorentz transformation $p^\mu \to p^\mu + \delta p^\mu$, with $\delta p^\mu = p^\nu \delta\epsilon^\mu_\nu$, $\delta\epsilon^\mu_\nu = -\delta\epsilon^\nu_\mu$, one has the invariance condition $I(p+\delta p) = I(p)$, which leads to the linear homogeneous differential equations

$$\sum_{i=1}^{n} \left(p_i^\nu \frac{\partial}{\partial p_{i\mu}} - p_i^\mu \frac{\partial}{\partial p_{i\nu}} \right) I(p) = 0 \qquad [55]$$

This equation can be contracted with all possible antisymmetric combinations of $p_{i\mu} p_{j\nu}$ to yield linearly independent Lorentz invariance identities for (53).

Using these two sets of identities, one can either obtain a reduction of integrals of the type (53) to those corresponding to a small number of simpler diagrams of the same topology and diagrams of simpler topology (fewer denominator factors), or a complete reduction to diagrams with simpler topology. The remaining integrals of the topology under consideration are called irreducible master integrals. These momentum integrals cannot be further reduced and have to be computed by different techniques. For instance, one can apply a Mellin–Barnes transformation of all propagators given by

$$\frac{1}{(k^2+a)^l} = \frac{1}{(l-1)!} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \frac{\mathrm{d}z}{2\pi\mathrm{i}} \frac{a^z}{(k^2)^{l+z}} \Gamma(l+z)\Gamma(-z) \qquad [56]$$

where the contour of integration is chosen to lie to the right of the poles of the Euler function $\Gamma(l+z)$ and to the left of the poles of $\Gamma(-z)$ in the complex z-plane. Alternatively, one may apply the negative-dimension method in which D is regarded as a negative integer in intermediate calculations and the problem of loop integration is replaced with that of handling infinite series. When combined with the above methods, it may be used to derive powerful recursion relations among scattering amplitudes. Both of these techniques rely on an explicit integration over the loop momenta of the graph, their differences occurring mainly in the representations used for the propagators.

The procedure outlined above can also be used to reduce a tensor Feynman integral to scalar integrals, as in the Brown–Feynman and Passarino–Veltman reductions. The tensor integrals are expressed as linear combinations of scalar integrals of either higher dimension or with propagators raised to higher powers. The projection onto a tensor basis takes the form [53] and can thus be reduced to master integrals.

String Theory Methods

The realizations of field theories as the low-energy limits of string theory provides a number of powerful tools for the calculation of multiloop amplitudes. They may be used to provide sets of diagrammatic computational rules, and they also work well for calculations in quantum gravity. In this final part we shall briefly sketch the insights into perturbative quantum field theory that are provided by techniques borrowed from string theory.

String Theory Representation

String theory provides an efficient compact representation of scattering amplitudes. At each loop order there is only a single closed string diagram, which includes within it all Feynman graphs along with the contributions of the infinite tower of massive string excitations. Schematically, at one-loop order, the situation is as shown in **Figure 3**. The terms arising from the heavy string modes are removed by taking the low-energy limit in which all external momenta lie well below the energy scale set by the string tension. This limit picks out the regions of integration in the string diagram corresponding to particle-like graphs, but with different diagrammatic rules.

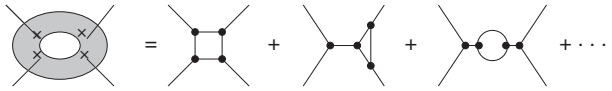


Figure 3 String theory representation at one-loop order.

Given these rules, one may formulate a purely field-theoretic framework which reproduces them. In the case of QCD, a key ingredient is the use of a special gauge originally derived from the low-energy limit of tree-level string amplitudes. This is known as the Gervais–Neveu gauge and it is defined by the gauge-fixing Lagrangian density

$$\mathcal{L}_{\mathrm{GN}} = -\frac{1}{2}\mathrm{Tr}\left(\partial_\mu A^\mu - \frac{ie}{\sqrt{2}}A_\mu A^\mu\right)^2 \qquad [57]$$

This gauge choice simplifies the color factors that arise in scattering amplitudes. The string theory origin of gauge theory amplitudes is then most closely mimicked by combining this gauge with the background field gauge, in which one decomposes the gauge field into a classical background field and a fluctuating quantum field as $A_\mu = A_\mu^{\mathrm{cl}} + A_\mu^{\mathrm{qu}}$, and imposes the gauge-fixing condition $D_\mu^{\mathrm{cl}} A^{\mathrm{qu}\mu} = 0$, where D_μ^{cl} is the background field covariant derivative evaluated in the adjoint representation of the gauge group. This hybrid gauge is well suited for computing the effective action, with the quantum part describing gluons propagating around loops and the classical part describing gluons emerging from the loops. The leading loop momentum behavior of one-particle irreducible graphs with gluons in the loops is very similar to that of graphs with scalar fields in the loops.

Supersymmetric Decomposition

String theory also suggests an intimate relationship with supersymmetry. For example, at tree level, QCD is effectively supersymmetric because a multigluon tree amplitude contains no fermion loops, and so the fermions may be taken to lie in the adjoint representation of the gauge group. Thus, pure gluon tree amplitudes in QCD are identical to those in supersymmetric Yang–Mills theory. They are connected by supersymmetric Ward identities to amplitudes with fermions (gluinos) which drastically simplify computations. In supersymmetric gauge theory, these identities hold to all orders of perturbation theory.

At one-loop order and beyond, QCD is not supersymmetric. However, one can still perform a supersymmetric decomposition of a QCD amplitude for which the supersymmetric components of the amplitude obey the supersymmetric Ward identities. Consider, for example, a one-loop multigluon scattering amplitude. The contribution from a fermion propagating in the loop can be decomposed into the contribution of a complex scalar field in the loop plus a contribution from an $\mathcal{N}=1$ chiral supermultiplet consisting of a complex scalar field and a Weyl fermion. The contribution from a gluon circulating in the loop can be decomposed into contributions of a complex scalar field, an $\mathcal{N}=1$ chiral supermultiplet, and an $\mathcal{N}=4$ vector supermultiplet comprising three complex scalar fields, four Weyl fermions and one gluon all in the adjoint representation of the gauge group. This decomposition assumes the use of a supersymmetry-preserving regularization.

The supersymmetric components have important cancellations in their leading loop momentum behavior. For instance, the leading large loop momentum power in an n-point 1PI graph is reduced from $|k|^n$ down to $|k|^{n-2}$ in the $\mathcal{N}=1$ amplitude. Such a reduction can be extended to any amplitude in supersymmetric gauge theory and is related to the improved ultraviolet behavior of supersymmetric amplitudes. For the $\mathcal{N}=4$ amplitude, further cancellations reduce the leading power behavior all the way down to $|k|^{n-4}$. In dimensional regularization, $\mathcal{N}=4$ supersymmetric loop amplitudes have a very simple analytic structure owing to their origins as the low-energy limits of superstring scattering amplitudes. The supersymmetric Ward identities in this way can be used to provide identities among the nonsupersymmetric contributions. For example, in $\mathcal{N}=1$ supersymmetric Yang–Mills theory one can deduce that fermion and gluon loop contributions are equal and opposite for multigluon amplitudes with maximal helicity violation.

Scattering Amplitudes in Twistor Space

The scattering amplitude in QCD with n incoming gluons of the same helicity vanishes, as does the amplitude with $n-1$ incoming gluons of one helicity and one gluon of the opposite helicity for $n \geq 3$. The first nonvanishing amplitudes are the maximal helicity violating (MHV) amplitudes involving $n-2$ gluons of one helicity and two gluons of the opposite helicity. Stripped of the momentum conservation delta-function and the group theory factor, the tree-level amplitude for a pair of gluons of negative helicity is given by

$$\mathcal{A}(k) = e^{n-2}\langle k_r^- | k_s^+ \rangle \prod_{i=1}^{n} \langle k_i^- | k_{i+1}^+ \rangle^{-1} \qquad [58]$$

This amplitude depends only on the holomorphic (negative chirality) Weyl spinors. The full MHV amplitude (with the momentum conservation delta-function) is invariant under the conformal group $\mathrm{SO}(4,2) \cong \mathrm{SU}(2,2)$ of four-dimensional

Minkowski space. After a Fourier transformation of the positive-chirality components, the complexification $SL(4,\mathbb{C})$ has an obvious four-dimensional representation acting on the positive- and negative-chirality spinor products. This representation space is isomorphic to $\mathbb{C}^4$ and is called twistor space. Its elements are called twistors.

Wave functions and amplitudes have a known behavior under the $\mathbb{C}^\times$-action which rescales twistors, giving the projective twistor space $\mathbb{CP}^3$ or $\mathbb{RP}^3$ according to whether the twistors are complex valued or real valued. The Fourier transformation to twistor space yields (due to momentum conservation) the localization of an MHV amplitude to a genus-0 holomorphic curve $\mathbb{CP}^1$ of degree 1 in $\mathbb{CP}^3$ (or to a real line $\mathbb{RP}^1 \subset \mathbb{RP}^3$). It is conjectured that, generally, an ℓ-loop amplitude with p gluons of positive helicity and q gluons of negative helicity is supported on a holomorphic curve in twistor space of degree $q + \ell - 1$ and genus $\leq \ell$. The natural interpretation of this curve is as the world sheet of a string. The perturbative gauge theory may then be described in terms of amplitudes arising from the couplings of gluons to a string. This twistor string theory is a topological string theory which gives the appropriate framework for understanding the twistor properties of scattering amplitudes. This framework has been used to analyze MHV tree diagrams and one-loop $\mathcal{N} = 4$ supersymmetric amplitudes of gluons.

See also: Constructive Quantum Field Theory; Dispersion Relations; Effective Field Theories; Gauge Theories from Strings; Hopf Algebra Structure of Renormalizable Quantum Field Theory; Perturbative Renormalization Theory and BRST; Quantum Chromodynamics; Renormalization: General Theory; Scattering, Asymptotic Completeness and Bound States; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools; Stationary Phase Approximation; Supersymmetric Particle Models.

Further Reading

Bern Z, Dixon L, and Kosower DA (1996) Progress in one-loop QCD computations. *Annual Review of Nuclear and Particle Science* 46: 109–148.

Brown LM and Feynman RP (1952) Radiative corrections to Compton scattering. *Physical Review* 85: 231–244.

Cachazo F and Svrček P (2005) Lectures on twistor strings and perturbative Yang–Mills theory. Preprint arXiv:hep-th/0504194.

Catani S (1998) The singular behaviour of QCD amplitudes at two-loop order. *Physics Letters B* 427: 161–171.

Chetyrkin KG and Tkachov FV (1981) Integration by parts: the algorithm to calculate beta functions in 4 loops. *Nuclear Physics B* 192: 159–204.

Dixon L (1996) Calculating scattering amplitudes efficiently. In: Soper DE *QCD and Beyond*, pp. 539–584. Singapore: World Scientific.

Fleischer J, Jegerlehner F, and Tarasov OV (2000) Algebraic reduction of one-loop Feynman graph amplitudes. *Nuclear Physics B* 566: 423–440.

Gehrmann T and Remiddi E (2000) Differential equations for two-loop four-point functions. *Nuclear Physics B* 580: 485–518.

Halliday IG and Ricotta RM (1987) Negative dimensional integrals 1: Feynman graphs. *Physics Letters B* 193: 241–246.

Itzykson C and Zuber J-B (1980) *Quantum Field Theory*. New York: McGraw-Hill.

Mangano ML and Parke SJ (1991) Multiparton amplitudes in gauge theories. *Physics Reports* 200: 301–367.

Magnen J and Seneor R (1979) Expansion and summability methods in constructive field theory. *Mathematical Problems in Theoretical Physics, Proc. Lausanne* 1979: 217–223.

Passarino G and Veltman MJG (1979) One-loop corrections for e^+e^- annihilation into $\mu^+\mu^-$ in the Weinberg model. *Nuclear Physics B* 160: 151–207.

Peskin ME and Schroeder DV (1995) *An Introduction to Quantum Field Theory*. Reading, MA: Addison-Wesley.

Pokorski S (1987) *Gauge Field Theories*. Cambridge: Cambridge University Press.

Steinhauser M (2002) Results and techniques of multiloop calculations. *Physics Reports* 364: 247–357.

Steinmann O (1995) Axiomatic approach to perturbative quantum field theory. *Annales de Poincare en Physique Theorique*. 63: 399–409.

Weinberg S (1995) *The Quantum Theory of Fields, Volume 1: Foundations*. Cambridge: Cambridge University Press.

Perturbative Renormalization Theory and BRST

K Fredenhagen, Universität Hamburg, Hamburg, Germany
M Dütsch, Universität Zürich, Zürich, Switzerland

Main Problems in the Perturbative Quantization of Gauge Theories

Gauge theories are field theories in which the basic fields are not directly observable. Field configurations yielding the same observables are connected by a gauge transformation. In the classical theory, the Cauchy problem is well posed for the observables, but in general not for the nonobservable gauge-variant basic fields, due to the existence of time-dependent gauge transformations.

Attempts to quantize the gauge-invariant objects directly have not yet been completely satisfactory. Instead, one modifies the classical action by adding a gauge-fixing term such that standard techniques of perturbative quantization can be applied and such that the dynamics of the gauge-invariant classical fields is not changed. In perturbation theory, this

problem shows up already in the quantization of the free gauge fields (see the section "Quantization of free gauge fields"). In the final (interacting) theory the physical quantities should be independent on how the gauge fixing is done ("gauge independence").

Traditionally, the quantization of gauge theories is mostly analyzed in terms of path integrals (e.g., by Faddeev and Popov), where some parts of the arguments are only heuristic. In the original treatment of Becchi, Rouet, and Stora (cf. also Tyutin) (which is called "BRST-quantization"), a restriction to purely massive theories was necessary; the generalization to the massless case by Lowenstein's method is cumbersome.

The BRST quantization is based on earlier work of Feynman, Faddeev, and Popov (introduction of "ghost fields"), and of Slavnov. The basic idea is that after adding a term to the Lagrangian which makes the Cauchy problem well posed but which is not gauge-invariant one enlarges the number of fields by infinitesimal gauge transformations ("ghosts") and their duals ("anti-ghosts"). One then adds a further term to the Lagrangian which contains a coupling of the anti-ghosts and ghosts. The BRST transformation acts as an infinitesimal gauge transformation on the original fields and on the gauge transformations themselves and maps the anti-ghosts to the gauge-fixing terms. This is done in such a way that the total Lagrangian is invariant and that the BRST transformation is nilpotent. The hard problem in the perturbative construction of gauge theories is to show that BRST symmetry can be maintained during renormalization (see the section on perturbative renormalization). By means of the "quantum action principle" of Lowenstein (1971) and Lam (1972, 1973) a cohomological classification of anomalies was worked out (an overview is given, e.g., in the book of Piguet and Sorella (1995)). For more details, *see* BRST Quantization.

The BRST quantization can be carried out in a transparent way in the framework of algebraic quantum field theory (AQFT, *see* Algebraic Approach to Quantum Field Theory). The advantage of this formulation is that it allows one to separate the three main problems of perturbative gauge theories:

1. the elimination of unphysical degrees of freedom,
2. positivity (or "unitarity"), and
3. the problem of infrared divergences.

In AQFT, the procedure is the following: starting from an algebra of all local fields, including the unphysical ones, one shows that after perturbative quantization the algebra admits the BRST transformation as a graded nilpotent derivation. The algebra of observables is then defined as the cohomology of the BRST transformation. To solve the problem of positivity, one has to show that the algebra of observables, in contrast to the algebra of all fields, has a nontrivial representation on a Hilbert space. Finally, one can attack the infrared problem by investigating the asymptotic behavior of states. The latter problem is nontrivial even in quantum electrodynamics (since an electron is accompanied by a "cloud of soft photons") and may be related to confinement in quantum chromodynamics.

The method of BRST quantization is by no means restricted to gauge theories, but applies to general constrained systems. In particular, massive vector fields, where the masses are usually generated by the Higgs mechanism, can alternatively be treated directly by the BRST formalism, in close analogy to the massless case (cf. the section on quantization of free gauge fields).

Local Operator BRST Formalism

In AQFT, the principal object is the family of operator algebras $\mathcal{O} \to \mathcal{A}(\mathcal{O})$ (where $\mathcal{O}$ runs, e.g., through all double cones in Minkowski space), which fulfills the Haag–Kastler axioms (cf. Algebraic Approach to Quantum Field Theory). To construct these algebras, one considers the algebras $\mathcal{F}(\mathcal{O})$ generated by all local fields including ghosts u and anti-ghosts $\tilde{u}$. Ghosts and anti-ghosts are scalar fermionic fields. The algebra gets a $\mathbb{Z}_2$ grading with respect to even and odd ghost numbers, where ghosts get ghost numbers +1 and anti-ghosts ghost number –1. The BRST transformation s acts on these algebras as a $\mathbb{Z}_2$-graded derivation with $s^2 = 0$, $s(\mathcal{F}(\mathcal{O})) \subset \mathcal{F}(\mathcal{O})$, and $s(F^*) = -(-1)^{\delta_F} s(F)^*$, δ_F denoting the ghost number of F.

The observables should be s-invariant and may be identified if they differ by a field in the range of s. Since the range $\mathcal{A}_{00}$ of s is an ideal in the kernel $\mathcal{A}_0$ of s, the algebra of observables is defined as the quotient

$$\mathcal{A} := \mathcal{A}_0/\mathcal{A}_{00} \qquad [1]$$

and the local algebras $\mathcal{A}(\mathcal{O}) \subset \mathcal{A}$ are the images of $\mathcal{A}_0 \cap \mathcal{F}(\mathcal{O})$ under the quotient map $\mathcal{A}_0 \to \mathcal{A}$.

To prove that $\mathcal{A}$ admits a nontrivial representation by operators on a Hilbert space, one may use the BRST operator formalism (Kugo and Ojima 1979, Dütsch and Fredenhagen 1999): one starts from a representation of $\mathcal{F}$ on an inner-product space $(\mathcal{K}, \langle \cdot, \cdot \rangle)$ such that $\langle F^*\phi, \psi \rangle = \langle \phi, F\psi \rangle$

and that s is implemented by an operator Q on $\mathcal{K}$, that is,

$$s(F) = [Q, F] \qquad [2]$$

with $[\cdot\,,\cdot]$ denoting the graded commutator, such that Q is symmetric and nilpotent. One may then construct the space of physical states as the cohomology of Q, $\mathcal{H} := \mathcal{K}_0/\mathcal{K}_{00}$, where $\mathcal{K}_0$ is the kernel and $\mathcal{K}_{00}$ the range of Q. The algebra of observables now has a natural representation π on $\mathcal{H}$:

$$\pi([A])[\phi] := [A\phi] \qquad [3]$$

(where $A \in \mathcal{A}_0, \phi \in \mathcal{K}_0, [A] := A + \mathcal{A}_{00}, [\phi] := \phi + \mathcal{K}_{00}$). The crucial question is whether the scalar product on $\mathcal{H}$ inherited from $\mathcal{K}$ is positive definite.

In free quantum field theories $(\mathcal{K}, \langle\cdot\,,\cdot\rangle)$ can be chosen in such a way that the positivity can directly be checked by identifying the physical degrees of freedom (see next section). In interacting theories (see the section on perturbative construction of gauge theories), one may argue in terms of scattering states that the free BRST operator on the asymptotic fields coincides with the BRST operator of the interacting theory. This argument, however, is invalidated by infrared problems in massless gauge theories. Instead, one may use a stability property of the construction.

Namely, let $\tilde{\mathcal{F}}$ be the algebra of formal power series with values in $\mathcal{F}$, and let $\tilde{\mathcal{K}}$ be the vector space of formal power series with values in $\mathcal{K}$. $\tilde{\mathcal{K}}$ possesses a natural inner product with values in the ring of formal power series $\mathbb{C}[[\lambda]]$, as well as a representation of $\tilde{\mathcal{F}}$ by operators. One also assumes that the BRST transformation $\tilde{s}$ is a formal power series $\tilde{s} = \sum_n \lambda^n s_n$ of operators s_n on $\mathcal{F}$ and that the BRST operator $\tilde{Q}$ is a formal power series $\tilde{Q} = \sum_n \lambda^n Q_n$ of operators on $\mathcal{K}$. The algebraic construction can then be done in the same way as before, yielding a representation $\tilde{\pi}$ of the algebra of observables $\tilde{\mathcal{A}}$ by endomorphisms of a $\mathbb{C}[[\lambda]]$ module $\tilde{\mathcal{H}}$, which has an inner product with values in $\mathbb{C}[[\lambda]]$.

One now assumes that at $\lambda = 0$ the inner product is positive, in the sense that

(Positivity)

(i) $\langle\phi,\phi\rangle \geq 0 \quad \forall\phi \in \mathcal{K}$ with $Q_0\phi = 0$, and
(ii) $Q_0\phi = 0 \wedge \langle\phi,\phi\rangle = 0 \Longrightarrow \phi \in Q_0\mathcal{K}$ [4]

Then the inner product on $\tilde{\mathcal{H}}$ is positive in the sense that for all $\tilde{\phi} \in \tilde{\mathcal{H}}$ the inner product with itself, $\langle\tilde{\phi},\tilde{\phi}\rangle$, is of the form $\tilde{c}^*\tilde{c}$ with some power series $\tilde{c} \in \mathbb{C}[[\lambda]]$, and $\tilde{c} = 0$ iff $\tilde{\phi} = 0$.

This result guarantees that, within perturbation theory, the interacting theory satisfies positivity, provided the unperturbed theory was positive and BRST symmetry is preserved.

Quantization of Free Gauge Fields

The action of a classical free gauge field A,

$$S_0(A) = -\frac{1}{4}\int \mathrm{d}x\, F^{\mu\nu}(x)F_{\mu\nu}(x)$$

$$= \frac{1}{2}\int \mathrm{d}k \hat{A}_\mu(k)^* M^{\mu\nu}(k)\hat{A}_\nu(k) \qquad [5]$$

(where $F^{\mu\nu} := \partial^\mu A^\nu - \partial^\nu A^\mu$ and $M^{\mu\nu}(k) := k^2 g^{\mu\nu} - k^\mu k^\nu$) is unsuited for quantization because $M^{\mu\nu}$ is not invertible: due to $M^{\mu\nu}k_\mu = 0$, it has an eigenvalue 0. Therefore, the action is usually modified by adding a Lorentz-invariant gauge-fixing term: $M^{\mu\nu}$ is replaced by $M^{\mu\nu}(k) + \lambda k^\mu k^\nu$, where $\lambda \in \mathbb{R} \setminus \{0\}$ is an arbitrary constant. The corresponding Euler–Lagrange equation reads

$$\Box A^\mu - (1-\lambda)\partial^\mu\partial_\nu A^\nu = 0 \qquad [6]$$

For simplicity, let us choose $\lambda = 1$, which is referred to as Feynman gauge. Then the algebra of the free gauge field is the unital $\star$-algebra generated by elements $A^\mu(f)$, $f \in \mathcal{D}(\mathbb{R}^4)$, which fulfill the relations:

$$f \mapsto A^\mu(f) \quad \text{is linear} \qquad [7]$$

$$A^\mu(\Box f) = 0 \qquad [8]$$

$$A^\mu(f)^* = A^\mu(\bar{f}) \qquad [9]$$

$$[A^\mu(f), A^\nu(g)] = \mathrm{i}g^{\mu\nu}\int \mathrm{d}x\,\mathrm{d}y\, f(x)D(x-y)g(y) \qquad [10]$$

where D is the massless Pauli–Jordan distribution.

This algebra does not possess Hilbert space representations which satisfy the microlocal spectrum condition, a condition which in particular requires the singularity of the two-point function to be of the so-called Hadamard form. It possesses, instead, representations on vector spaces with a nondegenerate sequilinear form, for example, the Fock space over a one-particle space with scalar product

$$\langle\phi,\psi\rangle = (2\pi)^{-3}\int \frac{\mathrm{d}^3\boldsymbol{p}}{2|\boldsymbol{p}|}\overline{\phi^\mu(p)}\psi_\mu(p)|_{p^0=|\boldsymbol{p}|} \qquad [11]$$

Gupta and Bleuler characterized a subspace of the Fock space on which the scalar product is semidefinite; the space of physical states is then obtained

by dividing out the space of vectors with vanishing norm.

After adding a mass term

$$\frac{m^2}{2}\int \mathrm{d}x A_\mu(x)A^\mu(x)$$

to the action [5], it seems to be no longer necessary to add also a gauge-fixing term. The fields then satisfy the Proca equation

$$\partial_\mu F^{\mu\nu} + m^2 A^\nu = 0 \qquad [12]$$

which is equivalent to the equation $(\Box + m^2)A^\mu = 0$ together with the constraint $\partial_\mu A^\mu = 0$. The Cauchy problem is well posed, and the fields can be represented in a positive-norm Fock space with only physical states (corresponding to the three physical polarizations of A). The problem, however, is that the corresponding propagator admits no power-counting renormalizable perturbation series.

The latter problem can be circumvented in the following way: for the algebra of the free quantum field, one takes only the equation $(\Box + m^2)A^\mu = 0$ into account (or, equivalently, one adds the gauge-fixing term $(1/2)(\partial_\mu A^\mu)^2$ to the Lagrangian) and goes over from the physical field A^μ to

$$B^\mu := A^\mu + \frac{\partial^\mu \phi}{m} \qquad [13]$$

where ϕ is a real scalar field, to the same mass m where the sign of the commutator is reversed ("bosonic ghost field" or "Stückelberg field"). The propagator of B^μ yields a power-counting renormalizable perturbation series; however, B^μ is an unphysical field. One obtains four independent components of B which satisfy the Klein–Gordon equation. The constraint $0 = \partial_\mu A^\mu = \partial_\mu B^\mu + m\phi$ is required for the expectation values in physical states only. So, quantization in the case $m > 0$ can be treated in analogy with [8]–[10] by replacing A^μ by B^μ, the wave operator by the Klein–Gordon operator $(\Box + m^2)$ in [8], and D by the corresponding massive commutator distribution Δ_m in [10]. Again, the algebra can be nontrivially represented on a space with indefinite metric, but not on a Hilbert space.

One can now use the method of BRST quantization in the massless as well as in the massive case. One introduces a pair of fermionic scalar fields (ghost fields) $(u, \tilde{u})$. $u, \tilde{u}$, and (for $m > 0$) ϕ fulfill the Klein–Gordon equation to the same mass $m \geq 0$ as the vector field B. The free BRST transformation reads

$$\begin{aligned} s_0(B^\mu) &= \mathrm{i}\partial^\mu u, \qquad & s_0(\phi) &= \mathrm{i}mu \\ s_0(u) &= 0, & s_0(\tilde{u}) &= -\mathrm{i}(\partial_\nu B^\nu + m\phi) \end{aligned} \qquad [14]$$

(see, e.g., Scharf (2001)). It is implemented by the free BRST charge

$$Q_0 = \int_{x^0=\text{const.}} \mathrm{d}^3 x j_0^{(0)}(x^0, \boldsymbol{x}) \qquad [15]$$

where

$$j_\mu^{(0)} := (\partial_\nu B^\nu + m\phi)\partial_\mu u - \partial_\mu(\partial_\nu B^\nu + m\phi)u \qquad [16]$$

is the free BRST current, which is conserved. (The interpretation of the integral in [15] requires some care.) Q_0 satisfies the assumptions of the (local) operator BRST formalism, in particular it is nilpotent and positive [4]. Distinguished representatives of the equivalence classes $[\phi] \in \mathrm{Ke}\, Q_0/\mathrm{Ra}\, Q_0$ are the states built up only from the three spatial (two transversal for $m = 0$, respectively) polarizations of A.

Perturbative Renormalization

The starting point for a perturbative construction of an interacting quantum field theory is Dyson's formula for the evolution operator in the interaction picture. To avoid conflicts with Haag's theorem on the nonexistence of the interaction picture in quantum field theory, one multiplies the interaction Lagrangian $\mathcal{L}$ with a test function g and studies the local S-matrix,

$$\begin{aligned} S(g\mathcal{L}) = 1 + \sum_{n=1}^{\infty} \frac{\mathrm{i}^n}{n!} \int \mathrm{d}x_1 \cdots \mathrm{d}x_n g(x_1) \cdots g(x_n) \\ \times T(\mathcal{L}(x_1) \cdots \mathcal{L}(x_n)) \end{aligned} \qquad [17]$$

where T denotes a time-ordering prescription. In the limit $g \to 1$ (adiabatic limit), $S(g\mathcal{L})$ tends to the scattering matrix. This limit, however, is plagued by infrared divergences and does not always exist. Interacting fields $F_{g\mathcal{L}}$ are obtained by the Bogoliubov formula:

$$F_{g\mathcal{L}}(x) = \frac{\delta}{\delta h(x)}|_{h=0} S(g\mathcal{L})^{-1} S(g\mathcal{L} + hF) \qquad [18]$$

The algebraic properties of the interacting fields within a region $\mathcal{O}$ depend only on the interaction within a slightly larger region (Brunetti and Fredenhagen 2000), hence the net of algebras in the sense of AQFT can be constructed in the adiabatic limit without the infrared problems (this is called the "algebraic adiabatic limit").

The construction of the interacting theory is thus reduced to a definition of time-ordered products of fields. This is the program of causal perturbation theory (CPT), which was developed by Epstein and Glaser (1973) on the basis of previous work by Stückelberg and Petermann (1953) and Bogoliubov

and Shirkov (1959). For simplicity, we describe CPT only for a real scalar field. Let φ be a classical real scalar field which is not restricted by any field equation. Let $\mathcal{P}$ denote the algebra of polynomials in φ and all its partial derivatives $\partial^a\varphi$ with multi-indices $a \in \mathbb{N}_0^4$. The time-ordered products $(T_n)_{n\in\mathbb{N}}$ are linear and symmetric maps $T_n : (\mathcal{P} \otimes \mathcal{D}(\mathbb{R}^4))^{\otimes n} \to L(\mathcal{D})$, where $L(\mathcal{D})$ is the space of operators on a dense invariant domain $\mathcal{D}$ in the Fock space of the scalar free field. One often uses the informal notation

$$\begin{aligned} T_n(g_1F_1 \otimes \cdots \otimes g_nF_n) \\ = \int \mathrm{d}x_1 \cdots \mathrm{d}x_n T_n(F_1(x_1), \ldots, F_n(x_n)) \\ \times g_1(x_1) \cdots g_n(x_n) \end{aligned} \quad [19]$$

where $F_j \in \mathcal{P}$, $g_j \in \mathcal{D}(\mathbb{R}^4)$.

The sequence (T_n) is constructed by induction on n, starting with the initial condition

$$T_1\left(\prod_j \partial^{a_j}\varphi(x)\right) =: \prod_j \partial^{a_j}\phi(x) : \quad [20]$$

where the right-hand side is a Wick polynomial of the free field ϕ. In the inductive step the requirement of causality plays the main role, that is, the condition that

$$\begin{aligned} T_n(f_1 \otimes \cdots \otimes f_n) = T_k(f_1 \otimes \cdots \otimes f_k) \\ \times T_{n-k}(f_{k+1} \otimes \cdots \otimes f_n) \end{aligned} \quad [21]$$

if

$$\begin{aligned} (\operatorname{supp} f_1 \cup \cdots \cup \operatorname{supp} f_k) \\ \cap ((\operatorname{supp} f_{k+1} \cup \cdots \cup \operatorname{supp} f_n) + \bar{V}_-) = \emptyset \end{aligned}$$

(where $\bar{V}_-$ is the closed backward light cone). This condition expresses the composition law for evolution operators in a relativistically invariant and local way. Causality determines T_n as an operator-valued distribution on $\mathbb{R}^{4n}$ in terms of the inductively known T_l, $l < n$, outside of the total diagonal $\Delta_n := \{(x_1, \ldots, x_n) \mid x_1 = \cdots = x_n\}$, that is, on test functions from $\mathcal{D}(\mathbb{R}^{4n} \setminus \Delta_n)$.

Perturbative renormalization is now the extension of T_n to the full test function space $\mathcal{D}(\mathbb{R}^{4n})$. Generally, this extension is nonunique. In contrast to other methods of renormalization, no divergences appear, but the ambiguities correspond to the finite renormalizations that persist after removal of divergences by infinite counter terms. The ambiguities can be reduced by (re-)normalization conditions, which means that one requires that certain properties which hold by induction on $\mathcal{D}(\mathbb{R}^{4n} \setminus \Delta_n)$ are maintained in the extension, namely:

- (N0) a bound on the degree of singularity near the total diagonal;
- (N1) Poincaré covariance;
- (N2) unitarity of the local S-matrix;
- (N3) a relation to the time-ordered products of subpolynomials;
- (N4) the field equation for the interacting field $\varphi_{g\mathcal{L}}$ [18];
- (AWI) the "action Ward identity" (Stora 2002, Dütsch and Fredenhagen 2003): $\partial^\mu T(\cdots F_l(x)\cdots) = T(\cdots \partial^\mu F_l(x)\cdots)$. This condition can be understood as the requirement that physics depends on the action only, so total derivatives in the interaction Lagrangian can be removed; and
- further symmetries, in particular in gauge theories, Ward identities expressing BRST invariance. A universal formulation of all symmetries which can be derived from the field equation in classical field theory is the "master Ward identity" (which presupposes (N3) and (N4)) (Boas and Dütsch 2002, Dütsch and Fredenhagen 2003); see next section.

The problem of perturbative renormalization is to construct a solution of all these normalization conditions. Epstein and Glaser have constructed the solutions of (N0)–(N3). Recently, the conditions (N4) and (AWI) have been included. The master Ward identity cannot always be fulfilled, the obstructions being the famous "anomalies" of perturbative quantum field theory.

Perturbative Construction of Gauge Theories

In the case of a purely massive theory, the adiabatic limit $S = \lim_{g\to 1} S(g\mathcal{L})$ exists (Epstein and Glaser 1976), and one may adopt a formalism due to Kugo and Ojima (1979), who use the fact that in these theories the BRST charge Q can be identified with the incoming (free) BRST charge Q_0 [15]. For the scattering matrix S to be a well-defined operator on the physical Hilbert space of the free theory, $\mathcal{H} = \operatorname{Ke} Q_0 / \operatorname{Ra} Q_0$, one then has to require

$$\lim_{g\to 1} [Q_0, T((g\mathcal{L})^{\otimes n})]|_{\ker Q_0} = 0 \quad [22]$$

This is the motivation for introducing the condition of "perturbative gauge invariance" (Dütsch *et al.* 1993, 1994); see Scharf (2001)): according to this condition, there should exist a Lorentz

vector $\mathcal{L}_1^\nu \in \mathcal{P}$ associated with the interaction $\mathcal{L}$, such that

$$[Q_0, T_n(\mathcal{L}(x_1)\cdots\mathcal{L}(x_n)] = \mathrm{i}\sum_{l=1}^{n}\partial_\nu^{x_l} T_n(\mathcal{L}(x_1)\cdots\mathcal{L}_1^\nu(x_l)\cdots\mathcal{L}(x_n)) \qquad [23]$$

This is a somewhat stronger condition than [22] but has the advantage that it can be formulated independently of the adiabatic limit. The condition [22] (or perturbative gauge invariance) can be satisfied for tree diagrams (i.e., the corresponding requirement in classical field theory can be fulfilled). In the massive case, this is impossible without a modification of the model; the inclusion of additional physical scalar fields (corresponding to Higgs fields) yields a solution. It is gratifying that, by making a polynomial ansatz for the interaction $\mathcal{L} \in \mathcal{P}$, perturbative gauge invariance [23] for tree diagrams, renormalizability (i.e., the mass dimension of $\mathcal{L}$ is ≤ 4), and some obvious requirements (e.g., the Lorentz invariance) determine $\mathcal{L}$ to a far extent. In particular, the Lie-algebraic structure needs not to be put in, as it can be derived in this way (Stora 1997, unpublished). Including loop diagrams (i.e., quantum effects), it has been proved that (N0)–(N2) and perturbative gauge invariance can be fulfilled to all orders for massless SU(N) Yang–Mills theories.

Unfortunately, in the massless case, it is unlikely that the adiabatic limit exists and, hence, an S-matrix formalism is problematic. One should better rely on the construction of local observables in terms of couplings with compact support. However, then the selection of the observables [1] has to be done in terms of the BRST transformation $\tilde{s}$ of the interacting fields.

For the corresponding BRST charge, one makes the ansatz

$$\tilde{Q} = \int \mathrm{d}^4x\, \tilde{j}_{g\mathcal{L}}^\mu(x) b_\mu(x), \quad \mathcal{L} = \sum_{n\geq 1}\mathcal{L}_n\lambda^n \qquad [24]$$

where (b_μ) is a smooth version of the δ-function characterizing a Cauchy surface and $\tilde{j}_{g\mathcal{L}}^\mu$ is the interacting BRST-current [18] (where $\tilde{j}_\mu = \sum_n j_\mu^{(n)}\lambda^n$ ($j_\mu^{(n)} \in \mathcal{P}$) is a formal power series with $j_\mu^{(0)}$ given by [16]). (Note that there is a volume divergence in this integral, which can be avoided by a spatial compactification. This does not change the abstract algebra $\mathcal{F}_\mathcal{L}(\mathcal{O})$.) A crucial requirement is that $\tilde{j}_{g\mathcal{L}}^\mu$ is conserved in a suitable sense. This condition is essentially equivalent to perturbative gauge invariance and hence its application to classical field theory determines the interaction $\mathcal{L}$ in the same way, and in addition the deformation $j^{(0)} \to \tilde{j}_{g\mathcal{L}}$. The latter also gives the interacting BRST charge and transformation, $\tilde{Q}$ and $\tilde{s}$, by [24] and [2]. The so-obtained $\tilde{Q}$ is often nilpotent in classical field theory (and hence this holds also for $\tilde{s}$). However, in QFT conservation of $\tilde{j}_{g\mathcal{L}}$ and $\tilde{Q}^2 = 0$ requires the validity of additional Ward identities, beyond the condition of perturbative gauge invariance [23]. All the necessary identities can be derived from the master Ward identity

$$T_{n+1}(A, F_1, \ldots, F_n) = -\sum_{k=1}^{n} T_n(F_1, \ldots, \delta_A F_k, \ldots, F_n) \qquad [25]$$

where $A = \delta_A S_0$ with a derivation δ_A. The master Ward identity is closely related to the quantum action principle which was formulated in the formalism of generating functionals of Green's functions. In the latter framework, the anomalies have been classified by cohomological methods. The vanishing of anomalies of the BRST symmetry is a selection criterion for physically acceptable models.

In the particular case of QED, the Ward identity

$$\partial_\mu^y T(j^\mu(y)F_1(x_1)\cdots F_n(x_n)) = \mathrm{i}\sum_{j=1}^{n}\delta(y-x_j) \times T\big(F_1(x_1)\cdots(\theta F_j)(x_j)\cdots F_n(x_n)\big) \qquad [26]$$

for the Dirac current $j^\mu := \bar{\psi}\gamma^\mu\psi$, is sufficient for the construction, where $(\theta F) := \mathrm{i}(r-s)F$ for $F = \psi^r\bar{\psi}^s B_1\cdots B_l$ ($B_1, \ldots, B_l$ are nonspinorial fields) and $F_1, \ldots, F_n$ run through all subpolynomials of $\mathcal{L} = j^\mu A_\mu$, (N0)–(N4) and [26] can be fulfilled to all orders (Dütsch and Fredenhagen, 1999).

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Batalin–Vilkovisky Quantization; BRST Quantization; Constrained Systems; Indefinite Metric; Perturbation Theory and its Techniques; Quantum Chromodynamics; Quantum Field Theory: A Brief Introduction; Quantum Fields with Indefinite Metric: Non-Trivial Models; Renormalization: General Theory; Renormalization: Statistical Mechanics and Condensed Matter; Standard Model of Particle Physics.

Further Reading

Becchi C, Rouet A, and Stora R (1975) Renormalization of the abelian Higgs–Kibble model. *Communications in Mathematical Physics* 42: 127.

Becchi C, Rouet A, and Stora R (1976) Renormalization of gauge theories. *Annals of Physics (NY)* 98: 287.

Bogoliubov NN and Shirkov DV (1959) *Introduction to the Theory of Quantized Fields*. New York: Interscience Publishers Inc.

Brunetti R and Fredenhagen K (2000) Microlocal analysis and interacting quantum field theories: renormalization on physical backgrounds. *Communications in Mathematical Physics* 208: 623.

Dütsch M, Hurth T, Krahe K, and Scharf G (1993) Causal construction of Yang–Mills theories. I. *N. Cimento A* 106: 1029.

Dütsch M, Hurth T, Krahe K, and Scharf G (1994) Causal construction of Yang–Mills theories. II. *N Cimento A* 107: 375.
Dütsch M and Fredenhagen K (1999) A local (perturbative) construction of observables in gauge theories: the example of QED. *Communications in Mathematical Physics* 203: 71.
Dütsch M and Boas F-M (2002) The Master Ward identity. *Reviews of Mathematical Physics* 14: 977–1049.
Dütsch M and Fredenhagen K (2003) The master Ward identity and generalized Schwinger–Dyson equation in classical field theory. *Communications in Mathematical Physics* 243: 275.
Dütsch M and Fredenhagen K (2004) Causal perturbation theory in terms of retarded products, and a proof of the action Ward identity. *Reviews in Mathematical Physics* 16: 1291–1348.
Epstein H and Glaser V (1973) *Annals Institut Henri Poincaré A* 19: 211.
Epstein H and Glaser V (1976) Adiabatic limit in perturbation theory. In: Velo G and Wightman AS (eds.) *Renormalization Theory*, pp. 193–254.
Henneaux M and Teitelboim C (1992) *Quantization of Gauge Systems*. Princeton: Princeton University Press.
Kugo T and Ojima I (1979) Local covariant operator formalism of nonabelian gauge theories and quark confinement problem. *Supplement of the Progress of Theoritical Physics* 66: 1.
Lam Y-MP (1972) Perturbation Lagrangian theory for scalar fields – Ward–Takahashi identity and current algebra. *Physics Reviews D* 6: 2145.
Lam Y-MP (1973) Equivalence theorem on Bogoliubov–Parasiuk–Hepp–Zimmermann – renormalized Lagrangian field theories. *Physics Reviews D* 7: 2943.
Lowenstein JH (1971) Differential vertex operations in Lagrangian field theory. *Communications in Mathematical Physics* 24: 1.
Piguet O and Sorella S (1995) Algebraic Renormalization: Perturbative Renormalization, Symmetries and Anomalies, Lecture Notes in Physics. Berlin: Springer.
Scharf G (1995) *Finite Quantum Electrodynamics. The Causal Approach*, 2nd edn. Berlin: Springer.
Scharf G (2001) *Quantum Gauge Theories – A True Ghost Story*. New York: Wiley.
Stora R (2002) Pedagogical experiments in renormalized perturbation theory. Contribution to the conference. *Theory of Renormalization and Regularization*. Germany: Hesselberg.
Stückelberg ECG and Petermann A (1953) La normalisation des constantes dans la theorie des quanta. *Helvetica Physica Acta* 26: 499–520.
Weinberg S (1996) *The Quantum Theory of Fields*. Cambridge: Cambridge University Press.

Phase Transition Dynamics

A Onuki, Kyoto University, Kyoto, Japan

Introduction

When an external parameter such as the temperature T is changed, physical systems in a homogeneous state often become unstable and tend to an ordered phase with broken symmetry. The growth of new order takes place with coarsening of domains or defect structures on mesoscopic spatial scales much longer than the microscopic molecular scale. Such ordering processes are ubiquitously observed in many systems such as ferromagnetic (spin) systems, solid alloys, and fluids. Historically, structural ordering and phase separation in solid alloys have been one of the central problems in metallurgy (Cahn 1961). These are highly nonlinear and far-from-equilibrium processes and have been studied as challenging subjects in condensed matter physics, polymer science, and metallurgy (Gunton *et al.* 1983, Binder 1991, Bray 1994, Onuki 2002). Here a short review on phase ordering is given on the basis of prototype mathematical models, which can be a starting point to understand the real complex problems.

Phase Ordering in Nonconserved Systems

Let us consider phase ordering in a system with a scalar spacetime-dependent variable $\psi(\boldsymbol{r}, t)$. If its space integral is not conserved in time, it is called the nonconserved order parameter, representing magnetization, electric polarization, etc. After appropriate scaling of time t, space $\boldsymbol{r}$, and ψ, the simplest dynamic equation reads

$$\frac{\partial}{\partial t}\psi = \nabla^2\psi - \tau\psi - \psi^3 + h + \theta \qquad [1]$$

The coefficient τ is related to the temperature by $\tau = A(T - T_c)$, where A is a constant and T_c is the critical temperature. The constant h is also an externally controllable parameter, proportional to the applied magnetic field for the ferromagnetic case. The last term is the Markovian Gaussian random noise needed when eqn [1] is treated as a Langevin (stochastic differential) equation. In physics its stochastic property is usually expressed as

$$\langle\theta(\boldsymbol{r}, t)\theta(\boldsymbol{r}', t')\rangle = 2\varepsilon\delta(\boldsymbol{r} - \boldsymbol{r}')\delta(t - t') \qquad [2]$$

where ε represents the strength of the noise (proportional to the temperature before the scaling). In the presence of θ, the variable ψ is a random variable, whose probability distribution $P(\{\psi\}, t)$

obeys the Fokker–Planck equation. The equilibrium (steady) distribution is given by

$$P_{\rm eq}\{\psi\} = \text{const.}\ \exp(-F\{\psi\}/\varepsilon) \qquad [3]$$

where

$$F = \int \mathrm{d}\boldsymbol{r}\left[\frac{\tau}{2}\psi^2 + \frac{1}{4}\psi^4 + \frac{1}{2}|\nabla\psi|^2 - h\psi\right] \qquad [4]$$

is the so-called Ginzburg–Landau free energy. Using F we rewrite eqn [1] in a standard form of the Langevin equation,

$$\frac{\partial}{\partial t}\psi = -\frac{\delta F}{\delta \psi} + \theta \qquad [5]$$

In equilibrium ψ consists of the average ψ_e and the deviation $\delta\psi$, where the latter is a Gaussian fluctuation in the limit of small ε. If $\tau > 0$ and $h = 0$, we obtain $\psi_e = 0$. If $\tau < 0$ and $h = 0$, there are two minima $\psi_e = \pm|\tau|^{1/2}$. These two states can coexist in equilibrium with a planar interface separating them at $h = 0$. If its normal is along the x-axis, the interface solution is of the form

$$\psi(x) = |\tau|^{1/2}\tanh(|\tau|^{1/2}x/\sqrt{2}) \qquad [6]$$

which tends to $\pm|\tau|^{1/2}$ as $x \pm \infty$ and satisfies

$$\delta F/\delta\psi = (\tau + \psi^2)\psi - \mathrm{d}^2\psi/\mathrm{d}x^2 = 0 \qquad [7]$$

It is well known that the fluctuations of ψ are increasingly enhanced near the critical point. The renormalization group theory shows how the equilibrium distribution $P_{\rm eq}\{\psi\}$ in eqn [3] depends on the upper cutoff wave number Λ of ψ, where we suppose that ψ consists of the Fourier components $\psi_{\boldsymbol{k}}$ with $k < \Lambda$ (Onuki 2002). In our phase-ordering problem the shortest relevant spatial scale is the interface width of the order of the thermal correlation length ξ at the final temperature. Therefore, near criticality, we may assume that the thermal fluctuations with wave numbers larger than ξ^{-1} have been eliminated in the model (or $\Lambda \sim \xi^{-1}$ at the starting point).

Domain Growth

Thermodynamic instability occurs when τ is changed from a positive value $\tau_{\rm i}$ to a negative value $\tau_{\rm f}$ at $t = 0$. We here assume $h = 0$. We set $\tau_{\rm f} = -1$ using the scaling. At long wavelengths $k < 1$, small plane wave fluctuations with wave vector $\boldsymbol{k}$ grow exponentially as

$$\psi_{\boldsymbol{k}}(t) \sim \exp[(1 - k^2)t] \qquad [8]$$

with the growth rate largest at $k = 0$. This suggests that the nonlinear term in eqn [1] becomes crucial after a transient time. Numerically obtained snapshots of the subsequent $\psi(\boldsymbol{r}, t)$ are shown in Figure 1

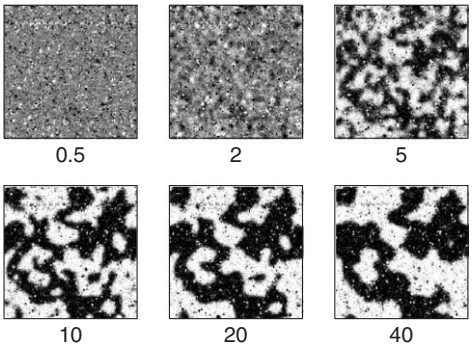


Figure 1 Time evolution of ψ in model [1] in 2D with system length = 128. The numbers are the times after quenching. Noise is added, but is not essential for large patterns or in the late stage. Reproduced with permission from Onuki A (2002) *Phase Transition Dynamics*. Cambridge, UK: Cambridge University Press.

in two dimensions (2D), where we can see the coarsening of the patterns. The characteristic domain size $\ell(t)$ grows algebraically as

$$\ell(t) \sim t^a \qquad [9]$$

where $a = 1/2$ is known for the model [1]. Scattering experiments detect the time-dependent correlation

$$g(r, t) = \langle \delta\psi(\boldsymbol{r} + \boldsymbol{r}_0, t)\delta\psi(\boldsymbol{r}_0, t)\rangle \qquad [10]$$

$$S(k, t) = \int \mathrm{d}\boldsymbol{r} g(r, t)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} \qquad [11]$$

where $S(k, t)$ is called the structure factor. We assume the translational invariance and the spatial isotropy after the thermal average $\langle\cdots\rangle$. If $\tau_i \gg 1$, the quartic term in F is negligible, leading to the initial structure factor

$$S(k, 0) \cong \varepsilon/(\tau_{\rm i} + k^2) \qquad [12]$$

which is produced by the thermal fluctuations. However, when the domain size $\ell(t)$ much exceeds the microscopic length (lattice constant), the following scaling behavior emerges:

$$g(r, t) = G(r/\ell(t)) \qquad [13]$$

$$S(k, t) = \ell(t)^d Q(\ell(t)k) \qquad [14]$$

where d is the space dimensionality and $G(x)$ and $Q(x)$ are the scaling functions of order unity for $x \sim 1$. The correlation on the scale of $\ell(t)$ in eqn [13] arises from large-scale domain structures, while eqn [14] is simply its Fourier transformation. The maximum of the structure factor grows as $\ell(t)^d$. When $\varepsilon \ll 1$, however, there can be a well-defined initial stage in which $S(k, t)$ grows exponentially at long wavelengths.

We may explain the roles of the terms on the right-hand side of eqn [1] in phase ordering in a simple manner.

1. The linear term $-\tau\psi$ triggers instability for $\tau < 0$.
2. The nonlinear term $-\psi^3$ gives rise to saturation of ψ into ± 1. To see this, we neglect $\nabla^2\psi$ and θ to have $\partial\psi/\partial t = (1-\psi^2)\psi$ for $\tau = -1$. This equation is solved to give

$$\psi(t) = \psi_0 / \sqrt{\psi_0^2 + (1-\psi_0^2)\mathrm{e}^{-2t}} \qquad [15]$$

   where $\psi_0 = \psi(0)$ is the initial value. Thus, $\psi \to 1$ for $\psi_0 > 0$ and $\psi \to -1$ for $\psi_0 < 0$ as $t \to \infty$.
3. The gradient term limits the instability only in the long wavelength region $k < 1$ in the initial stage (see eqn [8]) and creates the interfaces in the late stage (see eqn [7]).
4. The noise term θ is relevant only in the early stage where ψ is still on the order of the initial thermal fluctuations. The range of the early stage is of order 1 for $\varepsilon \gtrsim 1$, but weakly grows as $\ln(1/\varepsilon)$ for $\varepsilon \ll 1$. The noise term can be neglected once the fluctuations much exceed the thermal level.
5. If h is a small positive number, it favors growth of regions with $\psi \cong 1$.

Interface Dynamics

At long times $t \gg 1$ domains with typical size $\ell(t)$ are separated by sharp interfaces and the thermal noise is negligible. Allowing the presence of a small positive h, we may approximate the free energy F as

$$F = \sigma S(t) - 2hV_+(t) + \text{const.} \qquad [16]$$

where σ is a constant (surface tension), $S(t)$ is the surface area, and $V_+(t)$ is the volume of the regions with $\psi \cong 1$. In this stage the interface velocity $v_{\text{int}} = \boldsymbol{v}_{\text{int}} \cdot \boldsymbol{n}$ is given by the Allen–Cahn formula (Allen and Cahn 1979):

$$v_{\text{int}} = -\mathcal{K} + (2/\sigma)h \qquad [17]$$

The normal unit vector $\boldsymbol{n}$ is from a region with $\psi \cong 1$ to a region with $\psi \cong -1$. The $\mathcal{K}$ is the sum of the principal curvatures $1/R_1 + 1/R_2$ in 3D. This equation can be derived from eqn [1]. If the interface position $\boldsymbol{r}_a$ moves to $\boldsymbol{r}_a + \delta\zeta\boldsymbol{n}$ infinitesimally, the surface area changes by $\delta S = \int \mathrm{d}a\mathcal{K}\delta\zeta$, where $\int \mathrm{d}a \cdots$ denotes the surface integral. Therefore, F in eqn [16] changes in time as

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \int \mathrm{d}a(\sigma\mathcal{K} - 2h)v_{\text{int}} \le 0 \qquad [18]$$

which is non-negative-definite owing to eqn [17]. Furthermore, we may draw three results from eqn [17].

1. If we set $v_{\text{int}} \sim \ell(t)/t$ and $\mathcal{K} \sim 1/\ell(t)$, we obtain $a = 1/2$ in the growth law [9].
2. In phase ordering under very small positive h, the balance $1/\ell(t) \sim h/\sigma$ yields the crossover time $t_h \sim h^{-2}$. For $t < t_h$ the effect of h is small, while for $t > t_h$ the region with $\psi \cong 1$ becomes predominant.
3. A spherical droplet with $\psi \cong 1$ evolves as

$$\frac{\partial R}{\partial t} = -\frac{2}{R} + \frac{2h}{\sigma} \qquad [19]$$

   from which the critical radius is determined as

$$R_{\text{c}} = \sigma/h \qquad [20]$$

A droplet with $R > R_{\text{c}}(R < R_{\text{c}})$ grows (shrinks).

We mention a statistical theory of interface dynamics at $h = 0$ by Ohta (1982). There, a smooth subsidiary field $u(\boldsymbol{r}, t)$ is introduced to represent surfaces by $u = \text{const}$. The differential geometry is much simplified in terms of such a field. The two-phase boundaries are represented by $u = 0$. If all the surfaces follow $v_{\text{int}} = -\mathcal{K}$ in eqn [17] in the whole space, u obeys

$$\frac{\partial}{\partial t}u = \left[\nabla^2 - \sum_{ij} n_i n_j \nabla_i \nabla_j\right] u \qquad [21]$$

where $\nabla_i = \partial/\partial x_i$ and $n_i = \nabla_i u/|\nabla u|$. This equation becomes a linear diffusion equation if $n_i n_j \nabla_i \nabla_j$ is replaced by $d^{-1}\delta_{ij}\nabla^2$. Then u can be expressed in terms of its initial value and the correlation function of $\psi(\boldsymbol{r}, t) (\cong u(\boldsymbol{r}, t)/|u(\boldsymbol{r}, t)|$ in the late stage) is calculated in the form of eqn [13] with

$$G(x) = \frac{2}{\pi}\sin^{-1}\left[\exp\left(-\frac{1}{8(1-1/d)}x^2\right)\right] \qquad [22]$$

which excellently agrees with simulations.

Spinodal Decomposition in Conserved Systems

The order parameter ψ can be a conserved variable such as the density or composition in fluids or alloys. With the same F in eqn [4], a simple dynamic model in such cases reads

$$\frac{\partial}{\partial t}\psi = \nabla^2 \frac{\delta F}{\delta \psi} - \nabla \cdot \boldsymbol{j}^R \qquad [23]$$

Here $\boldsymbol{j}^R$ is the random current characterized by

$$\left\langle j_\alpha^R(\boldsymbol{r}, t) j_\beta^R(\boldsymbol{r}', t') \right\rangle = 2\varepsilon\delta_{\alpha\beta}\delta(\boldsymbol{r} - \boldsymbol{r}')\delta(t - t') \qquad [24]$$

which ensures the equilibrium distribution [3] of ψ. However, the noise $\boldsymbol{j}^R$ is negligible in late-stage phase separation as in the nonconserved case. Note that h in the conserved case is the chemical potential

conjugate to ψ and, if it is homogeneous, it vanishes in the dynamic equation [23]. In experiments the average order parameter

$$M = \langle \psi \rangle = \int d\boldsymbol{r}\psi(\boldsymbol{r})/V \qquad [25]$$

is used as a control parameter instead of h, where the integral is within the system with volume V. If there is no flux from outside, M is constant in time. Here the instability occurs below the so-called spinodal $M^2 < 1/3 (M^2 < |\tau|/3$ for general $\tau < 0)$. In fact, small fluctuations with wave vector $\boldsymbol{k}$ grow exponentially as

$$\psi_{\boldsymbol{k}}(t) \sim \exp[k^2(1 - 3M^2 - k^2)t] \qquad [26]$$

right after the quenching as in eqn [8]. The growth rate is largest at an intermediate wave number $k = k_m$ with

$$k_m = [(1 - 3M^2)/2]^{1/2} \qquad [27]$$

This behavior and the exponential growth of the structure factor have been observed in polymer mixtures where the parameter ε in eqn [3] or [12] is expected to be small (Onuki 2002). In late-stage coarsening the peak position of $S(k, t)$ decreases in time as

$$k_m(t) \sim 2\pi/\ell(t) \qquad [28]$$

in terms of the domain size $\ell(t)$. The growth exponent in eqn [9] is given by 1/3 for the simple model [23] (see eqn [33] below).

Figure 2 shows the patterns after quenching in 2D. For $M = 0$ the two phases are symmetric and the patterns are bicontinuous, while for $M \neq 0$ the minority phase eventually appears as droplets in the percolating region of the majority phase.

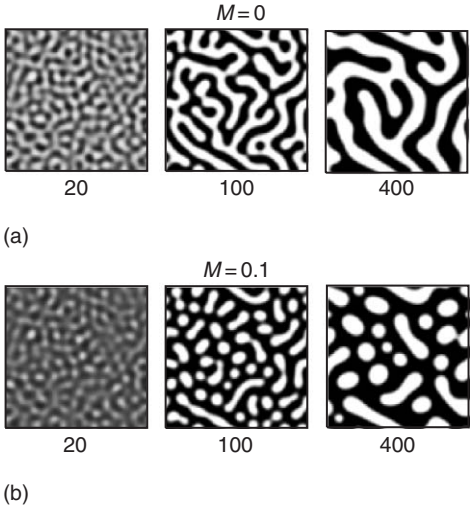


Figure 2 Time evolution of ψ in model [23] in 2D with system length = 128 without thermal noise: (a) $M = 0$ and (b) $M = 0.1$. The numbers are the times after quenching. Reproduced with permission from Onuki A (2002) *Phase Transition Dynamics*. Cambridge, UK: Cambridge University Press.

Interface Dynamics

Interface dynamics in the conserved case is much more complicated than in the nonconserved case, because the coarsening can proceed only through diffusion. Long-distance correlations arise among the domains and the interface velocity cannot be written in terms of the local quantities like the curvature. As a simple example, we give the counterpart of eqn [19]. In 3D a spherical droplet with $\psi \cong 1$ appears in a nearly homogeneous matrix with $\psi = M$ far from the droplet. The droplet radius R is then governed by (Lifshitz and Slyozov 1961)

$$\frac{\partial}{\partial t}R = D\left(\frac{\Delta}{R} - \frac{2d_0}{R^2}\right) \qquad [29]$$

where $\Delta = (M + 1)/2$ is called the supersaturation, while D and d_0 are constants (equal to 2 and $\sigma/8$, respectively, after the scaling). The critical radius is written as

$$R_c = 2d_0/\Delta \qquad [30]$$

The general definition of the supersaturation is

$$\Delta = \left(M - \psi_{cx}^{(2)}\right) \Big/ \left(\psi_{cx}^{(1)} - \psi_{cx}^{(2)}\right) \qquad [31]$$

Here the equilibrium values of ψ are written as $\psi_{cx}^{(1)}$ and $\psi_{cx}^{(2)}$ and M is supposed to be slightly different from $\psi_{cx}^{(2)}$.

Lifshitz and Slyozov (1961) analyzed domain coarsening in binary AB alloys when the volume fraction q of the A-rich domains is small. They noticed that the supersaturation Δ around each domain decreases in time with coarsening. That is, the A component atoms in the B-rich matrix are slowly absorbed onto the growing A-rich domains, while a certain fraction of the A-rich domains disappear. Thus, $q(t)$ and $\Delta(t)$ both depend on time, but satisfy the conservation law

$$q(t) + \Delta(t) = \Delta(0) = (M + 1)/2 \qquad [32]$$

With this overall constraint, they found the asymptotic late-stage behavior

$$\ell(t) \sim \Delta(t)^{-1} \sim t^{1/3} \qquad [33]$$

where $\ell(t)$ is the average droplet radius. Notice that this behavior is consistent with the droplet equation [29], where each term is of order $R/t \sim t^{-2/3}$.

Nucleation

In metastable states the free energy is at a local minimum but not at the true minimum. Such states

are stable for infinitesimal fluctuations, but rare spatially localized fluctuations, called critical nuclei, can continue to grow, leading to macroscopic phase ordering (Onuki 2002, Debenedetti 1996). The birth of a critical droplet is governed by the Boltzmann factor $\exp(-F_c/k_BT)$ at finite temperatures, where F_c is the free energy needed to create a critical droplet and k_BT is the thermal energy with k_B being the Boltzmann constant. In this section we explicitly write k_BT, but we may scale ψ and space such that $\tau = -1$ at the final temperature.

Droplet Free Energy and Experiments

In the nonconserved case we prepare a spin-down state with $\psi \cong -1$ in the time region $t < 0$ and then apply a small positive field h at $t = 0$. For $t > 0$ a spin-up droplet with radius R requires a free energy change

$$F(R) = 4\pi\sigma R^2 - \frac{8\pi}{3}hR^3 \qquad [34]$$

The first term is the surface free energy and the second term is the bulk decrease due to h. The critical radius R_c in eqn [20] gives the maximum of $F(R)$ given by

$$F_c = \frac{4\pi}{3}\sigma R_c^2 \qquad [35]$$

In fact, $F'(R) = \partial F(R)/\partial R$ is written as

$$F'(R) = 8\pi\sigma(R - R^2/R_c) \qquad [36]$$

In conserved systems such as fluids or alloys, we lower the temperature slightly below the coexistence curve with the average order parameter M held fixed. We again obtain the droplet free energy [34], but

$$h = (\sigma/2d_0)\Delta \qquad [37]$$

in terms of the (initial) supersaturation $\Delta = \Delta(0)$. Let the equilibrium values $\psi_{cx}^{(1)}$ and $\psi_{cx}^{(2)}$ in the two phases be written as $\pm A(T_c - T)^\beta$ with A and β being constants ($\beta \cong 1/3$ as $T \to T_c$). For each given M, we define the coexistence temperature T_{cx} by $M = \psi_{cx}^{(2)} = -A(T_c - T_{cx})^\beta$. In nucleation experiments the final temperature T is slightly below T_{cx} and $\delta T \equiv T_{cx} - T$ is a positive temperature increment. For small δT we find

$$\Delta \cong \frac{\beta}{2}\delta T/(T_c - T_{cx}) \qquad [38]$$

Droplet Size Distribution and Nucleation Rate

In a homogeneous metastable matrix, droplets of the new phase appear as rare thermal fluctuations. We describe this process by adding a thermal noise term to the droplet equation [19] or [29]. The droplet size distribution $n(R,t)$ then obeys the Fokker–Planck equation

$$\frac{\partial}{\partial t}n = \frac{\partial}{\partial R}\mathcal{L}(R)\left[\frac{\partial}{\partial R} + \frac{F'(R)}{k_BT}\right]n \qquad [39]$$

Here $n(R,t)dR$ denotes the droplet number density in the range $[R, R + dR]$. We determine the kinetic coefficient $\mathcal{L}(R)$ such that

$$v(R) \equiv -\mathcal{L}(R)F'(R)/k_BT \qquad [40]$$

is the right-hand side of eqn [19] or [29]. It is equal to $\partial R/\partial t$ when the thermal noise is neglected. Thus, $\mathcal{L}(R) \propto R^{-2}$ or R^{-3} for the nonconserved or conserved case. The second derivative $(\partial/\partial R)\mathcal{L}(R)(\partial/\partial R)$ in eqn [39] stems from the thermal noise and is negligible for $R - R_c \gtrsim 1$ in 3D (Onuki 2002). Hence, for $R - R_c \gtrsim 1$, the droplets follow the deterministic equation [19] or [29] and n obeys

$$\frac{\partial}{\partial t}n = -\frac{\partial}{\partial R}[v(R)n] \qquad [41]$$

In **Figure 3**, we plot the solution of eqn [39] for the conserved case with $F_c/k_BT = 17.4$ (Onuki 2002). The time is measured in units of $1/\Gamma_c$, which is the timescale of a critical droplet defined by

$$\Gamma_c = (\partial v(R)/\partial R)_{R=R_c} \qquad [42]$$

We notice $\Gamma_c \propto R_c^{-3}$ from eqn [29] so Γ_c is small. The initial distribution is given by

$$n(R,0) = n_0 \exp(-4\pi\sigma R^2/k_BT) \qquad [43]$$

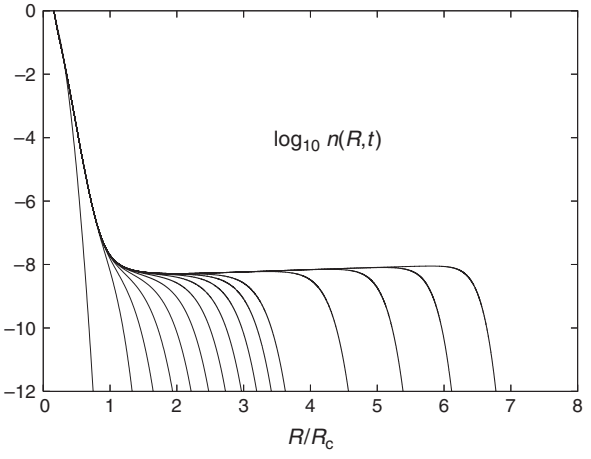


Figure 3 Time evolution of the droplet size distribution $n(R,t)$ on a semilogarithmic scale as a solution of eqn [39] in the 3D conserved case. The first 11 curves correspond to the times at $\Gamma_c t = 0, 1, \ldots$ and 10. The last four curves are those at $\Gamma_c t = 15, 20, 25$, and 30. Reproduced with permission from Onuki A (2002) *Phase Transition Dynamics*. Cambridge, UK: Cambridge University Press.

with n_0 being a constant number density. This form has been observed in computer simulations as the droplet size distribution on the coexistence curve ($h=0$). Figure 3 indicates that $n(R,t)$ tends to a steady solution $n_s(R)$ which satisfies

$$\mathcal{L}(R)\left[\frac{\partial}{\partial R}+\frac{F'(R)}{k_BT}\right]n_s=-I \qquad [44]$$

where I is a constant. Imposing the condition $n_s(R)\to 0$ as $R\to\infty$, we integrate the above equation as

$$n_s(R)=I\int_R^\infty dR_1\,\frac{1}{\mathcal{L}(R_1)}\exp\left[\frac{F(R_1)-F(R)}{k_BT}\right] \qquad [45]$$

For $R-R_c\gg 1$ we may replace $F(R_1)-F(R)$ by $F'(R)(R_1-R)$ in the integrand of eqn [45] to obtain

$$n_s(R)\cong I/v(R) \qquad [46]$$

which also follows from eqn [41]. Thus

$$n_s(R)dR=I\,dt \quad (dR=v(R)dt) \qquad [47]$$

This means that I is the nucleation rate of droplets with radii larger than R_c emerging per unit volume and per unit time. Furthermore, as $R\to 0$, we require $n_s(R)\to n_0=\text{const.}$ in eqn [43] so that

$$n_0=I\int_0^\infty dR_1\,\frac{1}{\mathcal{L}(R_1)}\exp\left[\frac{F(R_1)}{k_BT}\right] \qquad [48]$$

where the integrand becomes maximum around R_c. Using the expansion $F(R)=F_c+F''(R_c)\,(R-R_c)^2/2+\cdots$, we obtain the famous formula for the nucleation rate

$$I=I_0\,\exp(-F_c/k_BT) \qquad [49]$$

$$=I_0\,\exp(-C_0/\Delta^2) \qquad [50]$$

where the coefficient I_0 is of order $n_0\Gamma_c$. The second line holds in the 3D conserved case. Here, $C_0\sim 10^{-3}$ typically and I_0 is a very large number in units of $\text{cm}^{-3}\,\text{s}^{-1}$, say, 10^{30}. Then the exponential factor in I changes abruptly from a very small to a very large number with only a slight increase of Δ at small $\Delta\ll 1$. For example, if $C_0/\Delta^2=50$, I is increased by $\exp(100\delta\Delta/\Delta)$ with a small increase of Δ to $\Delta+\delta\Delta$. This factor can be of order 10^3 even for $\delta\Delta/\Delta=0.05$. Unless very close to criticality, simple metastable fluids become opaque suddenly with increasing Δ or δT at a rather definite cloud point. In near-critical fluids, however, I_0 itself becomes small ($\propto\xi^{-6}$) such that the cloud point considerably depends on the experimental timescale (observation time).

Remarks

The order parameter can be a scalar, a vector as in the Heisenberg spin system, a tensor as in liquid crystals, and a complex number as in superfluids and superconductors. In phase ordering a crucial role is played by topological singularities like interfaces in the scalar case and vortices in the complex number case. Furthermore, a rich variety of phase transition dynamics can be explained if the order parameter is coupled to other relevant variables in the free energy and/or in the dynamic equations. We mention couplings to velocity field in fluids, electrostatic field in charged systems, and elastic field in solids. Phase ordering can also be influenced profoundly by external fields such as electric field or shear flow.

See also: Reflection Positivity and Phase Transitions; Renormalization: Statistical Mechanics and Condensed Matter; Statistical Mechanics of Interfaces; Topological Defects and Their Homotopy Classification.

Further Reading

Allen SM and Cahn JW (1979) Microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta Metallurgica* 27: 1085.

Binder K (1991) Spinodal decomposition. In: Cohen RW, Haasen P, and Kramer EJ (eds.) *Material Sciences and Technology*, vol. 5. Weinheim: VCH.

Bray AJ (1994) Theory of phase-ordering kinetics. *Advances in Physics* 43: 357.

Cahn JW (1961) On spinodal decomposition. *Acta Metallurgica* 9: 795.

Debenedetti PG (1996) *Metastable Liquids*. Princeton: Princeton University.

Gunton JD, San Miguel M, and Sani PS (1983) The dynamics of first-order phase transitions. In: Domb C and Lebowitz JL (eds.) *Phase Transition and Critical Phenomena*, vol. 8. London: Academic Press.

Lifshitz IM and Slyozov VV (1961) The kinetics of precipitation from supersaturated solid solutions. *Journal of Physics and Chemistry of Solids* 19: 35.

Ohta T, Jasnow D, and Kawasaki K (1982) Universal scaling in the motion of random interfaces. *Physical Review Letters* 49: 1223.

Onuki A (2002) *Phase Transition Dynamics*. Cambridge: Cambridge University Press.

Phase Transitions in Continuous Systems

E Presutti, Università di Roma "Tor Vergata," Rome, Italy

Introduction

Many aspects of our everyday life, from weather to boiling water for a cup of coffee, involve heat exchanges and variations of pressure and, as a result, a phase transition. The general theory behind these phenomena is thermodynamics, which studies fluids and macroscopic bodies under these and more general transformations.

In the simple case of a one-component substance, the behavior under changes of temperature T and pressure P is described, according to the Gibbs phase rule, by a phase diagram such as the one in Figure 1. The curves in the (T, P) plane, distinguish regions where the substance is in its solid, liquid, and gas phases. Thus, in an experiment where we vary the pressure and temperature moving along a line which crosses a transition curve, we observe an abrupt and dramatic change at the crossing, when the system changes phase. As already stated, everyday life is an active source of examples of such phenomena.

The picture is "far from innocent", it states that air, liquid, and solid are not different elements of nature, as for long believed, but just different aspects of the same thing: substances are able to adapt to different external conditions in dramatically different ways. What properties of intermolecular forces are responsible for such astonishing behavior? The question has been extensively studied and it is the argument of the present article, where it will be discussed in the framework of statistical mechanics for continuous systems. Before entering into the matter, let us mention two basic motivations.

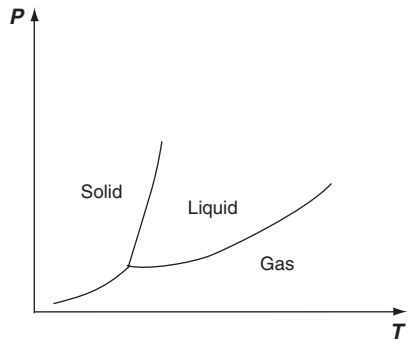


Figure 1 Phase diagram of a one-component substance.

As always, there is a "fundamental theory" aspect; in the specific case it is the attempt for an atomistic theory able to describe also macroscopic phenomena, thus ranging from the angstrom to the kilometer scales. From an engineering point of view, the target is, for instance, to understand why and when a substance is an insulator, or a conductor or, maybe, a superconductor, and, more importantly, how should we change its microscopic interactions to produce such effects: this opens the way to technologies which are indeed enormously affecting our life.

Phase Transitions and Statistical Mechanics

The modern theory of statistical mechanics is based upon the Gibbs hypothesis. In a classical (i.e., not quantum) framework, the macroscopic states are described by probability measures on a particle configuration phase space. The equilibrium states are then selected by the Gibbs prescription, which requires that the probability of observing a configuration which has energy E should be proportional to $e^{-\beta E}$, where $\beta = 1/kT$, k is the Boltzmann constant, and T the absolute temperature. These are the "Gibbs measures" and the purpose of statistical mechanics is to study their properties. A prerequisite for the success of the theory is compatibility with the principles of thermodynamics, the theory should then be able to explain the origin of the various phase diagrams and in particular to determine the circumstances under which phase transitions appear.

The theory, commonly called DLR, after Dobrushin, Lanford, and Ruelle, who, in the 1960s, contributed greatly to its foundations, has solid mathematical basis. Its main success is a rigorous proof of consistency with thermodynamics, which is derived under the only assumption that surface effects are negligible, a condition which is mathematically achieved by studying the system in a "thermodynamic limit," where the region containing the system invades the whole space.

In the thermodynamic limit, the equilibrium states can no longer be defined by the Gibbs prescription, because the energy of configurations in the whole space, being extensive, is typically infinite. The problem has been solved by first proving convergence of the finite-volume Gibbs measures in the thermodynamic limit. After defining the limit states, called "DLR states," as the equilibrium states of the

infinite systems, it is proved that the DLR states can be directly characterized (i.e., without using limit procedures) as the solutions of a set of equations, the "DLR equations," which generalize the finite-volume Gibbs prescription.

In terms of DLR states, the mathematical meaning of phase transitions becomes very clear and sharp. The starting point is the proof that the physical property that intensive variables in a pure phase have negligible fluctuations is verified by all the DLR measures which are in a special class, thus selected by this property, and which are therefore interpreted as "pure phases." All the other DLR measures are proved to be mixtures, that is, general convex combinations, of the pure DLR states. Thus, in the DLR theory, the system is in a single phase when there is only one DLR state, at the given values of the thermodynamic parameters (e.g., temperature and chemical potential), while the system is at a phase transition if there are several distinct DLR states.

While the theory beautifully clarifies the meaning of phase transitions, it does not say whether the phenomenon really occurs! This is maybe the main open problem in equilibrium statistical mechanics. A general proof of existence of phase diagrams is needed, which should at least capture the basic property behind the Gibbs phase rule, namely that in most of the space (of thermodynamic parameters) there is a single phase, with rare exceptions where several phases coexist. A more refined result should then indicate that coexistence occurs only on regular surfaces of positive codimension.

There is, however, a general result of existence of the gaseous phase, with a proof of uniqueness of DLR measures when temperature is large and density low. Coexistence of phases is much less understood at a general level, but results for particular classes of models exist, for instance, in lattice systems at low temperatures. The prototype is the ferromagnetic Ising model in two or more dimensions, where indeed the full diagram has been determined, see Figure 2. The transition curve is the segment $\{0 \leq T \leq T_c, h = 0\}$, in the (T, h) plane, h being the magnetic field. In the upper-half plane, there is a single phase with positive magnetization, in the lower one with a negative value; at $h = 0$, positive and negative magnetization states can coexist, if the temperature is lower than the critical value T_c. Correspondingly, there are, simultaneously, a positive and a distinctly negative DLR state, which describe the two phases.

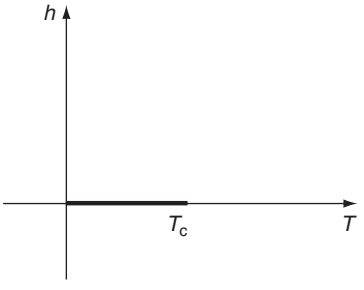


Figure 2 Phase diagram of the Ising ferromagnet.

An analogous result is missing for systems of particles in the continuum, but there has been recent progress on the analysis of the liquid–vapor branch of the phase diagram, and the issue will be the main focus of this article.

Sensitive Dependence on Boundary Conditions

Phase transitions describe exceptional regimes where the system is in a critical state; this is why they are so interesting and difficult to study. As in chaotic systems, criticality corresponds to a "butterfly effect," which, in a statistical-mechanics setting means changing far-away boundary conditions. Such changes affect the neighbors, which in turn influence their neighbors, and so on. In general, the effect decays with the distance but, at phase transition, it provokes an avalanche which propagates throughout the system reaching all its points. Its occurrence is not at all obvious, if we remember the stochastic nature of the theory. The domino effect described above can in fact, at each step, be subverted by stochastic fluctuations. The latter, in the end, may completely hide the effect of changing the boundary conditions. This is an instance of a competition between energy and entropy which is the ruling phenomenon behind phase transitions.

This intuitive picture also explains the relevance of space dimensionality. In a many-dimensional space, the influence of the boundary conditions has clearly many more ways to percolate, in contrast to the one-dimensional case, where in fact there is a general result on the uniqueness of DLR measures and therefore absence of phase transitions, for short-range interactions. For pair potentials, "short" means that the interaction energy between two molecules, respectively at r and r', decays as $|r - r'|^{-\alpha}$, $\alpha > 2$. There are results on the converse, namely on the presence of phase transitions when the above condition is not satisfied, mainly for lattice systems, but with partial extensions also to continuous systems. One-dimensional and long-range cases are not the main focus of this article, and the issue will not be discussed further here.

Ising Model

In order to make the previous ideas quantitative, let us first describe the simple case of the Ising model. Ising spin configurations are collections $\{\sigma(x), x \in \mathbb{Z}^d\}$ of $\sigma(x) \in \{\pm 1\}$ magnetic moments called spins. In the nearest-neighbor case, the interaction between two spins is $-J\sigma(x)\sigma(y)$, $J > 0$, if x and y are nearest neighbors on $\mathbb{Z}^d$, or is vanishing otherwise. There are, therefore, two ground states, one with all spins equal to $+1$ and the other one with all spins equal to -1. Since the Gibbs probability of higher energies vanishes as the temperature goes to zero, these are interpreted as the equilibrium states at temperature $T = 0$.

If $T > 0$, configurations with larger energy will appear, even though depressed by the Gibbs factor, but their occurrence is limited if T is small. In fact, in the ferromagnetic Ising model at zero magnetic field, dimensions $d \geq 2$, and low enough temperature, it has been proved that there are two distinct DLR measures, one called positive and the other negative. The typical configurations in the positive measure are mainly made by positive spins and, in such an "ocean of positive spins" there are rare and small islands of negative spins. The same situation, but with the positive and negative spins interchanged, occurs in the negative DLR state.

The selection of one of these two states can be made by choosing the positive or the negative boundary conditions, which shows how a surface effect, namely putting the boundary spins equal to 1 or -1, has a volume effect, as most of the spins in the system follow the value indicated by the boundary values. Again, this is more and more striking as we note that each spin is random, yet a strong, cooperative effect takes over and controls the system.

The original proof due to Peierls exploits the spin-flip symmetry of the Ising interaction, but it has subsequently been extended to a wider class of systems on the lattice, in the general framework of the "Pirogov–Sinai theory." This theory studies the low-temperature perturbations of ground states and it applies to many lattice systems, proving the existence of a phase transition and determining the structure of the phase diagram in the low-temperature region. The theory, however, does not cover continuous systems, where the low-temperature regime is essentially not understood, with the notable exception of the Widom and Rowlinson model.

Two Competing Species in the Continuum

The simplest version of the Widom and Rowlinson model has two types of particles, red and black, which are otherwise identical. Particles are massive points and the only interaction is a hard-core interaction among different colors, namely a red and a black particle cannot be closer than $2R_0, R_0 > 0$ being the hard-core radius.

The order parameter for the phase transition is the particle color. For large values of the chemical potential, and thus large densities, there are two states, one essentially red, the other black, while, if the density is low, the colors "are not separated" and there is a unique state. The proof of the statement starts by dividing the particles of a configuration into clusters, each cluster made by a maximal connected component, where two particles are called connected when their mutual distance is $< 2R_0$. Then, in each cluster, all particles have the same color (because of the hard-core exclusion between black and red), and the color is either black or red, with equal probability.

The question of phase transition is then related to cluster percolation, namely the existence of clusters which extend to infinity. If this occurs, then the influence of fixing the color of a particle may propagate infinitely far away, hence the characteristic "sensitive dependence phenomenon" of phase transitions. Percolation and hence phase transitions have been proved to exist in the positive and negative states, if the density is large and, respectively, small. The above argument is a more recent version of the original proof by Ruelle, which goes back to the 1970s.

The key element for the appearance of the phase transition is the competition between two different components, so that the analysis is not useful in explaining the mechanisms for coexistence in the case of identical particles, which are considered in the following.

Coarse Graining Transformations

The Peierls argument in Ising systems does not seem to extend to the continuum, certainly not in a trivial way. The ground states, in fact, will not be as simple as the constant configurations of a lattice system; they will instead be periodic or quasiperiodic configurations with a complicated dependence on the particle interactions. The typical fluctuations when we raise the temperature above zero have a much richer and complex structure and are correspondingly more difficult to control. Closeness to the ground states at nonzero temperature, as described in the Ising model, would prove the spontaneous breaking of the Euclidean symmetries and the existence of a crystalline phase. The question is, of course, of great interest, but it looks far beyond the reach of our present mathematical techniques.

The simpler Ising picture should instead reappear at the liquid–vapor coexistence line. Looking at the fluid on a proper spatial scale, we should in fact see a density that is essentially constant, except for small and rare fluctuations. Its value will differ in the liquid and in the gaseous states, $\rho_{\text{gas}} < \rho_{\text{liq}}$. Therefore, density is an order parameter for the transition and plays the role of the spin magnetization in the Ising picture.

There are general mathematical techniques developed to translate these ideas into proofs, they involve "coarse graining," "block spin transformations," and "renormalization group" procedures. The starting point is to ideally divide the space into cells. Their size should be chosen to be much larger than the typical microscopic distance between molecules, to depress fluctuations of the particle density in a cell. To study the probability distribution of the latter, we integrate out all the other degrees of freedom. After such a coarse graining, we are left with a system of spins on a lattice, the lattice sites labeling the cells (also called blocks) and each spin (also called block spin) giving the value of the density of particles in the corresponding cell. Translated into the language of block spins, the previous physical analysis of the state of the fluid suggests that most probably, in each block the density is approximately equal to either ρ_{liq} or ρ_{gas}, and the same in different blocks, except in the case of small and rare fluctuations. If we represent the probability distribution of the block spins in terms of a Gibbs measure (as always possible if the system is in a bounded region), the previous picture is compatible with a new Hamiltonian with a single spin (one-body) potential which favors the two values ρ_{liq} and ρ_{gas} and an attractive interaction between spins which suppresses changes from one to the other. A new effective low temperature should finally dampen the fluctuations.

Thus, after coarse graining, the system should be in the same universality class as of the low-temperature Ising model, and we may hope, in this way, to extend to the liquid–vapor branch of the phase diagram the Pirogov–Sinai theory of low-temperature lattice systems. In particular, as in the Ising model, we will then be able to select the liquid or the vapor phases by the introduction of suitable boundary conditions.

The conditional tense arises because the computation of the coarse graining transformation is in general very difficult, if not impossible, to carry out, but there is a class of systems where it has been accomplished. These are systems of identical point particles in $\mathbb{R}^d$, $d \geq 2$, which interact with "special" two- and four-body potentials, having finite range and which can be chosen to be rotation and translation invariant; their specific form will be described later. For such systems, the above coarse graining picture works and it has been proved that in a "small" region of the temperature–chemical potential plane, there is a part of the curve where two distinct phases coexist, while elsewhere in the neighborhood, the phase is unique.

The ideas behind the choice of the Hamiltonian go back to van der Waals, and the Ginzburg–Landau theory, which are milestones in the theory of phase transitions, while the mathematics of variational problems also enters here in an important way. These are briefly discussed in the next sections.

The van der Waals Liquid–Vapor Transition

Let us then do a step backwards and recall the van der Waals theory of the liquid–vapor transition. As typical intermolecular forces have a strong repulsive core and a rather long attractive tail, in a continuum, mesoscopic approximation of the system will be described by a free-energy functional of the type

$$F(\rho) = \int_{\Lambda} f^0_{\beta,\lambda}(\rho(r))\mathrm{d}r - \frac{1}{2}\int_{\Lambda\times\Lambda} J(r,r')\rho(r)\rho(r')\mathrm{d}r\,\mathrm{d}r' \qquad [1]$$

where $\rho = \{\rho(r), r \in \Lambda\}$ is the particles density and Λ the region where the system is confined, which, for simplicity, is taken here as a torus in $\mathbb{R}^d$, consisting of a cube with periodic boundary conditions. The term $-J(r,r')\rho(r)\rho(r')$, $J(r,r') \geq 0$, is the energy due to the attractive tail of the interaction, which is periodic in Λ; $f^0_{\beta,\lambda}(\rho) = f^0_{\beta,0}(\rho) - \lambda\rho$ is the free-energy density due to the short, repulsive part of the interaction, λ being the chemical potential.

As noted later, [1] can be rigorously derived by a coarse graining transformation; it will be used to build a bridge between the van der Waals theory and the previous block spin analysis of the liquid–vapor phase transition. Let us take for the moment [1] as a primitive notion. By invoking the second principle of thermodynamics, the equilibrium states can be found by minimizing the free-energy functional. Supposing J to be translation invariant, that is, $J(r,r') = J(r+a, r'+a)$, $r, r', a \in \mathbb{R}^d$, and calling $\alpha = \int J(r,r')\mathrm{d}r'$ the intensity of J, we can rewrite $F(\rho)$ as

$$F(\rho) = \int_{\Lambda}\left\{f^0_{\beta,\lambda}(\rho(r)) - \frac{\alpha\rho(r)^2}{2}\right\}\mathrm{d}r + \frac{1}{4}\int_{\Lambda\times\Lambda} J(r,r')[\rho(r)-\rho(r')]^2\mathrm{d}r\,\mathrm{d}r' \qquad [2]$$

This shows that the minimizer must have $\rho(r)$ constant (so that the second integral is minimized) and equal to any value which minimizes the function $\{f^0_{\beta,\lambda}(\rho) - \alpha\rho^2/2\}$. By thermodynamic principles, the free energy $f^0_{\beta,\lambda}(\rho)$ is convex in ρ, but, if α is large enough, the above expression is not convex and, by properly choosing the value of λ, the minimizers are no longer unique, hence the van der Waals phase transition.

Kac Potentials

The analogy between the above analysis of [2] and the previous heuristic study of the fluid based on coarse graining is striking. As customary in continuum theory, each mesoscopic point r should be regarded as representative of a cell containing many molecules. Then the functional $F(\rho)$ can be interpreted as the effective Hamiltonian after coarse graining. The role of the one-body term is played in [2] by the curly bracket, which selects two values of ρ (its minimizers, to be identified with ρ_{liq} and ρ_{gas}); the attractive two-body potential is then related to the last term in [2], as it suppresses the variations of ρ. The analogy clearly suggests a strategy for a rigorous proof of phase transitions in the continuum, an approach which has been and still is actively pursued. It will be discussed briefly in the sequel.

The first rigorous derivation of the van der Waals theory in a statistical-mechanics setting goes back to the 1960s and to Kac, who proposed a model where the particle pair interaction is

$$-\alpha\gamma^d \, \mathrm{e}^{-\gamma|q_i - q_j|} + \text{hard core}, \quad \gamma, \alpha > 0 \qquad [3]$$

The phase diagram of such systems, after the thermodynamic limit, can be quite explicitly determined in the limit $\gamma \to 0$, where it has been proved to converge to the van der Waals phase diagram, under a proper choice of $f^0_{\beta,\lambda}(\cdot)$ in [1].

The characteristic features of the first term in [3] are: (1) very long range, which scales as γ^{-1}, and (2) very small intensity, which scales as γ^d, so that the total intensity of the potential, defined as the integral over the second position, is independent of γ. The additional hard-core term (which imposes that any two particles cannot get closer than $2R_0, R_0 > 0$ being the hard-core radius) is to ensure stability of matter, that is, to avoid collapse of the whole system on an infinitesimally small region, as it would happen if only the attractive part of the interaction were present.

Derivation of the van der Waals theory has been proved for a general class of Kac potentials, where the exponential term in [3] is replaced by functions whose dependence on γ has the same scaling properties as mentioned above (in (1) and (2)), while the hard core can be replaced by suitably repulsive interactions.

The proof, in the version proposed by Lebowitz and Penrose, uses coarse graining and shows that the effective Hamiltonian is well approximated by the van der Waals functional [1], when γ is small, while the effective temperature scales as γ^d. The approximation becomes exact in the limit $\gamma \to 0$, where it reduces the computation of the partition function to the analysis of the minima and the ground states of an effective Hamiltonian which, in the limit $\gamma \to 0$, is exactly the van der Waals functional.

A true proof of phase transitions requires instead to keep $\gamma > 0$ fixed (instead of letting $\gamma \to 0$) and thus to control the difference of the effective Hamiltonian after coarse graining and the van der Waals functional, which is the effective Hamiltonian, but only in the actual limit $\gamma \to 0$. In general, there is no symmetry between the two ground states, unlike in the Ising case where they are related by spin flip, and the Pirogov–Sinai theory thus enters into play. The framework in fact is exactly similar, with the lattice Hamiltonian replaced by the functional and low temperatures by small γ (recall that the effective temperature scales as γ^d). The extension of the theory to such a setting, however, presents difficulties and success has so far been only partial.

A Model for Phase Transitions in the Continuum

The problem is twofold: to have a good control of (1) the limit theory and (2) the perturbations induced by a nonzero value of the Kac parameter γ. The former falls in the category of variational problems for integral functionals, whose prototype is the Ginzburg–Landau free energy

$$F^{\text{gl}}(\rho) = \int \{w(\rho) + |\nabla\rho|^2\} \, \mathrm{d}r \qquad [4]$$

which can be regarded as an approximation of [2] with w equal to the curly bracket in [2] and J replaced by a δ-function. Minimization problems for this and similar functionals have been widely analyzed in the context of general variational problems theory and partial differential equations (PDEs), and the study of the limit theory can benefit from a vast literature on the subject. The analysis of the corrections due to small γ is, however, so far quite limited. To implement the Pirogov–Sinai strategy, we need, in the case of the interaction [3],

a very detailed knowledge of the system without the Kac part of the interaction and with only hard cores. This, however, is so far not available when the particle density is near to close-packing (i.e., the maximal density allowed by the hard-core potential). Replacing hard cores by other short-range repulsive interactions does not help either, and this seems the biggest obstacle to the program.

The difficulty, however, can be avoided by replacing the hard-core potential by a repulsive many-body (more than two) Kac potential, which ensures stability as well. The class of systems covered by the approach is characterized by Hamiltonian of the form

$$H_{\gamma,\lambda}(q) = \int_{R^d} e_\lambda(\phi_\gamma(r))\mathrm{d}r \qquad [5]$$

where $e_\lambda(\phi)$ is a polynomial of the scalar field variable ϕ, a specific example being

$$e_\lambda(\phi) = \frac{\phi^4}{4!} - \frac{\phi^2}{2} - \lambda\phi \qquad [6]$$

This form of the Hamiltonian is familiar from Euclidean field theories. In these theories, the free distribution of the field is Gaussian; in our case, however, the field $\phi = \phi_\gamma(r)$ is a function of the particle configurations $q = (q_i, i = 1, \ldots, n)$:

$$\begin{aligned} \phi_\gamma(r) &= j_\gamma * q(r) = \sum_{i=1}^{n} j_\gamma(r, q_i) \\ j_\gamma(r, r') &= \gamma^d j(\gamma r, \gamma r') \end{aligned} \qquad [7]$$

where $j(r, r')$ is a translation-invariant, symmetric transition probability kernel. Thus, $\phi_\gamma(r)$ is a non-negative variable which has the meaning of a local density at r, weighted by the Kac kernel $j_\gamma(r, r')$.

Contours and Phase Indicators

The dependence on γ yields the scaling properties characteristic of the Kac potentials and [5] may be regarded as a generalized Kac Hamiltonian, which, in the polynomial case of [6], involves up to four-body Kac potentials. The phase diagram of the model, after taking first the thermodynamic limit and then the limit $\gamma \to 0$, is determined by the free-energy functional

$$F(\rho) = \int \left\{ e_\lambda(j * \rho(r)) - \frac{S(\rho(r))}{\beta} \right\} \mathrm{d}r \qquad [8]$$

$$S(\rho) = -\rho(\log\rho - 1) \qquad [9]$$

where [8] is taken to be defined on a torus (to avoid convergence problems of the integral), and $j = j_\gamma, \gamma = 1$.

Exploiting the concavity of the entropy $S(\rho)$, it is proved that the minimizers of $F(\cdot)$ are constant functions with the constants minimizing

$$f_{\lambda,\beta}(u) = e_\lambda(u) - \frac{S(u)}{\beta}, \quad u \geq 0 \qquad [10]$$

In the case of [6], to which we restrict in the sequel, for any $\beta > (3/2)^{3/2}$ there is λ_β so that $f_{\lambda_\beta,\beta}(u)$ is double-well with two minimizers, $\rho_{\mathrm{gas}} < \rho_{\mathrm{liq}}$ (dependence on β is omitted).

To "recognize" the densities ρ_{gas} and ρ_{liq} in a particle configuration, we use coarse graining and introduce two partitions of $\mathbb{R}^d$ into cubes $C^{(\ell_{\mp,\gamma})}$. The cubes $C^{(\ell_{-,\gamma})}$ of the first partition have side $\ell_{-,\gamma}$ proportional to $\gamma^{-1+\alpha}, \alpha > 0$ suitably small; those of the second one have length $\ell_{+,\gamma}$ proportional to $\gamma^{-1-\alpha}$; they are chosen so that each cube $C^{(\ell_{+,\gamma})}$ is union of cubes $C^{(\ell_{-,\gamma})}$. Notice that the small cubes have side much smaller than the interaction range (for small γ), while the opposite is true for the large cubes.

Given a particle configuration q, we say that a point r is in the liquid phase and write $\Theta(r; q) = 1$, if

$$\left| \frac{|q \sqcap C^{(\ell_{-,\gamma})}|}{\ell_{-\gamma}^d} - \rho_{\mathrm{liq}} \right| \leq \gamma^a, \quad a > 0 \text{ suitably small} \qquad [11]$$

for any small cube $C^{(\ell_{-,\gamma})}$ contained either in $C_r^{(\ell_{+,\gamma})}$ or in the cubes $C^{(\ell_{+,\gamma})}$ contiguous to $C_r^{(\ell_{+,\gamma})}$: $|q \sqcap C^{(\ell_{-,\gamma})}|$ is referred to as the number of particles of q in $C^{(\ell_{-,\gamma})}$, and $C_r^{(\ell_{+,\gamma})}$ as the large cube which contains r.

Thus, $\Theta(r; q) = 1$ if the local particle density is constantly close to ρ_{liq} in a large region around r. Defining $\Theta(r; q) = -1$ if the above holds with ρ_{gas} instead of ρ_{liq} and setting $\Theta(r; q) = 0$ in all the other cases, we then have a phase indicator $\Theta(r; q)$, which identifies, for all particle configurations, which spatial regions should be attributed to the liquid and gas phases. The connected components of the complementary region are called contours and the definition of $\Theta(r; q)$ has been structured in such a way that liquid and gas are always separated by a contour. The liquid phase will then be represented by a measure which gives large probability to configurations having mostly $\Theta = 1$, while the gas phase by configurations with mostly $\Theta = -1$.

This is quite similar to the Ising picture and, as in the Ising model, the existence of a phase transition follows from a Peierls estimate that contours have small probability. In fact, if there are few contours, the phase imposed on the boundaries of the region

where the system is observed percolates inside, invading most of the space. Thus, boundary conditions select the phase in the whole volume. The absence of the short-range potential, which was the hard-core interaction in [3], and hence the absence of all the difficulties which originate from it, allow one to carry through successfully the Pirogov–Sinai program and prove Peierls estimates on contours and, hence, the existence of a phase transition. In particular, the statistical weight of a contour is estimated by first relating the computation to one involving the functional [8] and then computing its value on density profiles compatible with the existence of the given contour. This part of the problem needs variational analysis for [8], with constraints and benefits of a vast literature on the subject.

The phase transition is very sharp, as shown by the following ideal experiment. Having fixed $\beta > (3/2)^{3/2}$, let λ vary in a (suitably) small interval $[\lambda_\beta - \delta, \lambda_\beta + \delta], \delta > 0$, centered around the mean-field critical value λ_β. We consider the system in a large region with, for instance, boundary conditions $\Theta = -1$ (i.e., forcing the gas phase) and fix γ small enough. At $\lambda = \lambda_\beta - \delta$, the system has $\Theta = -1$ in most of the domain, and this persists when we increase λ till a critical value, $\lambda_{\beta,\gamma}$, close to, but not the same as λ_β. For $\lambda > \lambda_{\beta,\gamma}, \Theta = 1$ in most of the domain, except for a small layer around the boundaries. The analogous picture holds if we choose boundary conditions $\Theta = 1$, and $\lambda = \lambda_{\beta,\gamma}$ is the only value of the chemical potential where the system is sensitive to the boundary conditions and both phases can be produced by the right boundary conditions. The fact that the actual value $\lambda_{\beta,\gamma}$ differs from λ_β, is characteristic of the Pirogov–Sinai approach and enlightens the delicate nature of the proofs.

Some Related Problems

In this concluding section, two important related problems, which have not been mentioned so far, are discussed.

A natural question, after proving a phase transition, is to describe how two phases coexist, once forced to be simultaneously present in the system. This can be achieved, for instance, by suitable boundary conditions (typically positive and negative on the top and bottom of the spatial domain) or by imposing a total density (or magnetization in the case of spins) intermediate between those of the pure phases. There will then be an interface separating the two phases with a corresponding surface tension and the geometry will be determined by the solution of a variational problem and given by the Wulff shape.

Can statistical mechanics explain and describe the phenomenon? Important progress has been made recently on the subject in the case of lattice systems at low temperatures. The question has also been widely studied at the mesoscopic level, in the context of variational problems for Ginzburg and Landau and many other functionals. Therefore, all the ingredients of further development of the theory in this direction are now present.

We have so far discussed only classical systems; a few words about extensions to the quantum case are now in order. In the range of values of temperatures and densities where the liquid–vapor transition occurs, the quantum effects are not expected to be relevant. Referring to the case of bosons, and away from the Bose condensation regime (and for system with Boltzmann statistics as well), the quantum delocalization of particles caused by the indeterminacy principle should essentially disappear after macroscopic coarse graining, and the block-spin variables should again behave classically, even though their underlying constituents are quantal. If this argument proves correct, then progress along these lines may be expected in near future.

See also: Cluster Expansion; Ergodic Theory; Finite Group Symmetry Breaking; Pirogov–Sinai Theory; Reflection Positivity and Phase Transitions; Statistical Mechanics and Combinatorial Problems; Statistical Mechanics of Interfaces; Symmetry Breaking in Field Theory; Two-Dimensional Ising Model.

Further Reading

Dal Maso G (1993) *An Introduction to Γ-Convergence*. Birkhäuser.

Lebowitz JL, Mazel A, and Presutti E (1999) Liquid–vapor phase transitions for systems with finite range interactions. *Journal of Statistical Physics* 94: 955–1025.

Ruelle D (1969) *Statistical Mechanics. Rigorous Results*. Benjamin.

Sinai YaG (1982) *Theory of Phase Transitions: Rigorous Results*. Pergamon Press (co-edition with Akademiai Kiadó, Budapest).

Pirogov–Sinai Theory

R Kotecký, Charles University, Prague, Czech Republic, and the University of Warwick, UK

Introduction

Pirogov–Sinai theory is a method developed to study the phase diagrams of lattice models at low temperatures. The general claim is that, under appropriate conditions, the phase diagram of a lattice model is, at low temperatures, a small perturbation of the zero-temperature phase diagram designed by ground states. The treatment can be generalized to cover temperature driven transitions with coexistence of ordered and disordered phases.

Formulation of the Main Result

Setting

Refraining first from full generality, we formulate the result for a standard class of lattice models with finite spin state and finite-range interaction. We will mention different generalizations later.

We consider *classical lattice models* on the d-dimensional hypercubic lattice $\mathbb{Z}^d$ with $d \geq 2$. A *spin configuration* $\sigma = (\sigma_x)_{x\in\mathbb{Z}^d}$ is an assignment of a spin with values in a finite set S to each lattice site $x \in \mathbb{Z}^d$; the configuration space is $\Omega = S^{\mathbb{Z}^d}$. For $\sigma \in \Omega$ and $\Lambda \subset \mathbb{Z}^d$, we use $\sigma_\Lambda \in \Omega_\Lambda = S^\Lambda$ to denote the restriction $\sigma_\Lambda = \{\sigma_x; x \in \Lambda\}$.

The Hamiltonian is given in terms of a collection of interaction potentials (Φ_A), where Φ_A are real functions on Ω, depending only on σ_x with $x \in A$, and A runs over all finite subsets of $\mathbb{Z}^d$. We assume that the potential is *periodic* with *finite range of interactions*. Namely, $\Phi_{A'}(\sigma') = \Phi_A(\sigma)$ whenever A and σ are related to A' and σ' by a translation from $(a\mathbb{Z})^d$ for some fixed integer a and there exists $R \geq 1$ such that $\Phi_A \equiv 0$ for all A with diameter exceeding R.

Without loss of generality (possibly multiplying the number a by an integer and increasing R), we may assume that $R = a$.

The Hamiltonian $H_\Lambda(\sigma|\eta)$ in Λ with boundary conditions $\eta \in \Omega$ is then given by

$$H_\Lambda(\sigma|\eta) = \sum_{A\cap\Lambda\neq\emptyset} \Phi_A(\sigma_\Lambda \vee \eta_{\Lambda^c}) \qquad [1]$$

where $\sigma_\Lambda \vee \eta_{\Lambda^c} \in \Omega$ is the configuration σ_Λ extended by η_{Λ^c} on Λ^c. The *Gibbs state in* Λ *under boundary conditions* $\eta \in \Omega$ (and with Hamiltonian H) is the probability $\mu_\Lambda(\cdot|\eta)$ on Ω_Λ defined by

$$\mu_\Lambda(\{\sigma_\Lambda\}|\eta) = \frac{\exp\{-\beta H_\Lambda(\sigma|\eta)\}}{Z(\Lambda|\eta)} \qquad [2]$$

with the *partition function*

$$Z(\Lambda|\eta) = \sum_{\sigma_\Lambda} \exp\{-\beta H_\Lambda(\sigma|\eta)\} \qquad [3]$$

We use $\mathcal{G}(H)$ to denote the set of all *periodic Gibbs states with Hamiltonian* H defined on Ω by means of the Dobrushin–Lanford–Ruelle (DLR) equations.

Ground-State Phase Diagram and the Removal of Degeneracy

A periodic configuration $\sigma \in \Omega$ is called a (*periodic*) *ground state* of a Hamiltonian $H = (\Phi_A)$ if

$$H(\tilde\sigma;\sigma) = \sum_A (\Phi_A(\tilde\sigma) - \Phi_A(\sigma)) \geq 0 \qquad [4]$$

for every finite perturbation $\tilde\sigma \neq \sigma$ of σ ($\tilde\sigma$ differs from σ at a finite number of lattice sites). We use $g(H)$ to denote the set of all periodic ground states of H. For every configuration $\sigma \in g(H)$, we define the *specific energy* $e_\sigma(H)$ by

$$e_\sigma(H) = \lim_{n\to\infty} \frac{1}{|V_n|} \sum_{A\cap V_n\neq\emptyset} \Phi_A(\sigma) \qquad [5]$$

(with V_n denoting a cube consisting of n^d lattice sites).

To investigate the phase diagram, we will consider a parametric class of Hamiltonians around a fixed Hamiltonian $H^{(0)}$ with a finite set of periodic ground states $g(H^{(0)}) = \{\sigma_1, \ldots, \sigma_r\}$. Namely, let $H^{(0)}$, $H^{(1)}, \ldots,$ and $H^{(r-1)}$ be Hamiltonians determined by potentials $\Phi^{(0)}, \Phi^{(1)}, \ldots,$ and $\Phi^{(r-1)}$, respectively, and consider the $(r-1)$-parametric set of Hamiltonians $H_{\boldsymbol{t}} = H^{(0)} + \sum_{\ell=1}^{r-1} t_\ell H^{(\ell)}$ with $\boldsymbol{t} = (t_1, \ldots, t_{r-1}) \in \mathbb{R}^{r-1}$. Using a shorthand $e_m(H) = e_{\sigma_m}(H)$, and introducing the vectors $\boldsymbol{e}(H) = (e_1(H), \ldots, e_r(H))$ and $\boldsymbol{h}(\boldsymbol{t}) = \boldsymbol{e}(H_{\boldsymbol{t}}) - \min_m e_m(H_{\boldsymbol{t}})$, we notice that for each $\boldsymbol{t} \in \mathbb{R}^{r-1}$, the vector $\boldsymbol{h}(\boldsymbol{t}) \in \partial Q_r$, the boundary of the positive octant in $\mathbb{R}^r$. A crucial assumption for such a parametrization $H_{\boldsymbol{t}}$ to yield a meaningful phase diagram is the *condition of removal of degeneracy*: we assume that $g(H^{(0)} + H^{(\ell)}) \subsetneq g(H^{(0)}), \ell = 1, \ldots, r-1$, and that the vectors $\boldsymbol{e}(H^{(\ell)}), \ell = 1, \ldots, r-1$, are linearly independent.

In particular, its immediate consequence is that the mapping $R^{r-1} \ni \boldsymbol{t} \mapsto \boldsymbol{h}(\boldsymbol{t}) \in \partial Q_r$ is a bijection. This fact has a straightforward interpretation in terms of *ground-state phase diagram*. Viewing the

phase diagram (at zero temperature) as a partition of the parameter space into regions K_g with a given set $g \subset g(H^{(0)})$ of ground states – "coexistence of zero-temperature phases from g" – the above bijection means that the region K_g is the preimage of the set

$$Q_g = \{\boldsymbol{h} \in \partial Q_r | h_m = 0 \text{ for } \sigma_m \in g \text{ and } h_m > 0 \text{ otherwise}\} \quad [6]$$

The partition of the set ∂Q_r has a natural hierarchical structure implied by the fact that $\overline{Q_{g_1}} \cap \overline{Q_{g_2}} = \overline{Q_{g_1 \cup g_2}}$ ($\overline{Q_g}$ is the closure of Q_g). Namely, the origin $\{0\} = Q_{g(H^{(0)})}$ is the intersection of r positive coordinate axes $\overline{Q_{\{\sigma_m, \bar{m} \neq m\}}}, m = 1, \ldots, r$; each of those half-lines is an intersection of $r-1$ two-dimensional quarter-planes with boundaries on positive coordinate axes, etc., up to $(r-1)$-dimensional planes $\overline{Q_{\{\sigma_m\}}}, m = 1, \ldots, r$. This hierarchical structure is thus inherited by the partition of the parameter space $\mathbb{R}^{r-1}$ into the regions K_g. The phase diagrams with such regular structure are sometimes said to satisfy the Gibbs phase rule.

We can thus summarize in a rather trivial conclusion that the condition of removal of degeneracy implies that the ground-state phase diagram obeys the Gibbs phase rule. The task of the Pirogov–Sinai theory is to provide means for proving that this remains true, at least in a neighborhood of the origin of parameter space, also for small nonzero temperatures. To achieve this, we need an effective control of excitation energies.

Peierls Condition

A crucial assumption for the validity of the Pirogov–Sinai theory is a lower bound on energy of excitations of ground states – the Peierls condition.

In spite of the fact that for a study of phase diagram we consider a parametric set of Hamiltonians whose set of ground states may differ, it is useful to introduce the Peierls condition with respect to a single fixed collection G of *reference configurations* (eventually, it will be identified with the ground states of the Hamiltonian $H^{(0)}$). Let thus a fixed set G of periodic configurations $\{\sigma_1, \ldots, \sigma_r\}$ be given. Again, without loss of generality, we may assume that the periodicity of all configurations $\sigma_m \in G$ is R.

Before formulating the Peierls condition, we have to introduce the notion of contours. Consider the set of all *sampling cubes* $C(x) = \{y \in \mathbb{Z}^d \| y_i - x_i| \leq R \text{ for } 1 \leq i \leq d\}, x \in \mathbb{Z}^d$. A *bad cube* of a configuration $\sigma \in \Omega$ is a sampling cube C for which σ_C differs from σ_m restricted to C for every $\sigma_m \in G$. The boundary $B(\sigma)$ of σ is the union of all bad cubes of σ. If $\sigma_m \in G$ and σ is its finite perturbation (differing from σ_m on a finite set of lattice sites), then, necessarily, $B(\sigma)$ is finite. A *contour* of σ is a pair $\gamma = (\Gamma, \sigma_\Gamma)$, where Γ (the *support* of the contour γ) is a connected component of $B(\sigma)$ (and σ_Γ is the restriction of σ on Γ). Here, the connectedness of Γ means that it cannot be split into two parts whose (Euclidean) distance is larger than 1. We use $\partial(\sigma)$ to denote the set of all contours of σ, $B(\sigma) = \bigcup_{\gamma \in \partial(\sigma)} \Gamma$.

Consider a configuration σ^γ such that γ is its unique contour. The set $\mathbb{Z}^d \backslash \Gamma$ has one infinite component to be denoted $\operatorname{Ext} \gamma$ and a finite number of finite components whose union will be denoted $\operatorname{Int} \gamma$. Observing that the configuration σ^γ coincides with one of the states $\sigma_m \in G$ on every component of $\mathbb{Z}^d \backslash B(\sigma)$, each of those components can be labeled by the corresponding m. Let q be the label of $\operatorname{Ext} \gamma$, we say that γ is a q-contour, and let $\operatorname{Int}_m \gamma$ be the union of all components of $\operatorname{Int} \gamma$ labeled by $m, m = 1, \ldots, r$.

Defining the "energy" $\Psi(\gamma)$ of a q-contour γ by the equation

$$\Psi(\gamma) = H(\sigma^\gamma; \sigma_q) + e_q(H)|\Gamma| - \sum_{m=1}^{r} (e_m(H) - e_q(H))|\operatorname{Int}_m \gamma| \quad [7]$$

the *Peierls condition* with respect to the set G of reference configurations is an assumption of the existence of $\rho > 0$ such that

$$\Psi(\gamma) \geq (\rho + \min_m e_m(H))|\Gamma| \quad [8]$$

for any contour of any configuration σ that is a finite perturbation of $\sigma_q \in G$.

Notice that if $G = g(H)$, the sum on the right-hand side of [7] vanishes.

Phase Diagram

The main claim of the Pirogov–Sinai theory provides, for β sufficiently large, a construction of regions $\mathcal{K}_g(\beta)$ of the parameter space characterized by the coexistence of phases labeled by configurations $\sigma_m \in g$. This is done similarly as for the ground-state phase diagram discussed earlier by constructing a homeomorphism $\boldsymbol{t} \mapsto \boldsymbol{a}(\boldsymbol{t})$ from a neighborhood of the origin of the parameter space to a neighborhood of the origin of ∂Q_r that provides the phase diagram (actually, the function $\boldsymbol{a}(\boldsymbol{t})$ will turn out to be just a perturbation of $\boldsymbol{h}(\boldsymbol{t})$ with errors of order $e^{-\beta}$).

Before stating the result, however, we have to clarify what exactly is meant by existence of phase m for a given Hamiltonian H. Roughly speaking, it is the existence of a periodic extremal Gibbs state $\mu_m \in \mathcal{G}(H)$, whose typical configurations do not differ too much from the ground-state configuration σ_m. In more technical terms, the existence of such a state is provided once we prove a

suitable bound, for the finite-volume Gibbs state $\mu_\Lambda(\{\sigma_\Lambda\}|\sigma_m)$ under the boundary conditions σ_m, on the probability that a fixed point in Λ is encircled by a contour from $\partial\sigma$. If this is the case, we say that the *phase m is stable*. It turns out that such a bound is actually an integral part of the construction of metastable free energies $f_m(\boldsymbol{t})$ yielding the homeomorphism $\boldsymbol{t} \mapsto \boldsymbol{a}(\boldsymbol{t})$. In this way, we get the main claim formulated as follows:

Theorem 1 *Consider a parametric set of Hamiltonians* $H_{\boldsymbol{t}} = H^{(0)} + \sum_{\ell=1}^{r-1} t_\ell H^{(\ell)}$ *with periodic finite-range interactions satisfying the condition of removal of degeneracy as well as the Peierls condition with respect to the reference set* $G = g(H^{(0)})$. *Let* $d \geq 2$ *and let* β *be sufficiently large. Then there exists a homeomorphism* $\boldsymbol{t} \mapsto \boldsymbol{a}(\boldsymbol{t})$ *of a neighborhood* V_β *of the origin of the parameter space* R^{r-1} *onto a neighborhood* U_β *of the origin of* ∂Q_r *such that, for any* $\boldsymbol{t} \in V_\beta$, *the set of all stable phases is* $\{m \in \{1, \ldots, r\} | a_m(\boldsymbol{t}) = 0\}$.

The Peierls condition can be actually assumed only for the Hamiltonian $H^{(0)}$ inferring its validity for $H_{\boldsymbol{t}}$ on a sufficiently small neighborhood V_β.

Notice also that the result can be actually stated not as a claim about phase diagram in a space of parameters, but as a statement about stable phases of a fixed Hamiltonian H. Namely, for a Hamiltonian H satisfying Peierls condition with respect to a reference set G, one can assure the existence of parameters a_m labeled by elements from G such that the set of extremal periodic Gibbs states of H consists of all those m-phases for which $a_m = 0$.

Construction of Metastable Free Energies

An important part of the Pirogov–Sinai theory is an actual construction of the *metastable free energies* – a set of functions $f_m(\boldsymbol{t}), m = 1, \ldots, r$, that provide the homeomorphism $\boldsymbol{a}(\boldsymbol{t})$ by taking $a_m(\boldsymbol{t}) = f_m(\boldsymbol{t}) - \min_{\bar{m}} f_{\bar{m}}(\boldsymbol{t})$.

We start with a *contour representation of partition function* $Z(\Lambda|\sigma_q)$. Considering, for each contributing configuration σ, the collection $\partial(\sigma)$ of its contours, we notice that, in addition to the fact that different contours $\gamma, \gamma' \in \partial(\sigma)$ have disjoint supports, $\Gamma \cap \Gamma' = \emptyset$, the contours from $\partial(\sigma)$ have to satisfy the *matching conditions*: if C is a connected component of $\mathbb{Z}^d \setminus \bigcup_{\gamma \in \partial} \Gamma$, then the restrictions of the spin configurations σ^γ to C are the same for all contours $\gamma \in \partial(\sigma)$ with $\mathrm{dist}(\Gamma, C) = 1$. In other words, the contours touching C induce the same label on C. Let us observe that there is actually one-to-one correspondence between configurations σ that coincide with σ_q on Λ^c and collections $\mathcal{M}(\Lambda, q)$ of contours ∂ in Λ satisfying the matching condition, and such that the external among them are q-contours. Here, a contour $\gamma \in \partial$ is called an *external contour* in ∂ if $\Gamma \subset \mathrm{Ext}\,\gamma'$ for all $\gamma' \in \partial$ different from γ.

With this observation and using $\Lambda_m(\partial)$ to denote the union of all components of $\Lambda \setminus \bigcup_{\gamma \in \partial} \Gamma$ with label m, we get

$$Z(\Lambda|\sigma_q) = \sum_{\partial \in \mathcal{M}(\Lambda, q)} \prod_m \mathrm{e}^{-\beta e_m(H)|\Lambda_m(\gamma)|} \prod_{\gamma \in \partial} \mathrm{e}^{-\beta \Psi(\gamma)} \quad [9]$$

Usefulness of such contour representations stems from an expectation that, for a stable phase q, contours should constitute a suppressed excitation and one should be able to use cluster expansions to evaluate the behavior of the Gibbs state μ_q. However, the direct use of the cluster expansion on [9] is trammeled by the presence of the energy terms $\mathrm{e}^{-\beta e_m(H)|\Lambda_m(\partial)|}$ and, more seriously, by the requirement that the contour labels match.

Nevertheless, one can rewrite the partition function in a form that does not involve any matching condition. Namely, considering first a sum over mutually external contours ∂^{ext} and resumming over collections of contours which are contained in their interiors without touching the boundary (being thus prevented to "glue" with external contours), we get

$$Z(\Lambda|\sigma_q) = \sum_{\partial^{\mathrm{ext}}} \mathrm{e}^{-\beta e_q(H)|\mathrm{Ext}|} \times \prod_{\gamma \in \partial^{\mathrm{ext}}} \left\{ \mathrm{e}^{-\beta \Psi(\gamma)} \prod_m Z^{\mathrm{dil}}(\mathrm{Int}_m \gamma | \sigma_m) \right\} \quad [10]$$

Here the sum goes over all collections of compatible external q-contours in Λ, $\mathrm{Ext} = \mathrm{Ext}_\Lambda(\partial^{\mathrm{ext}}) = \bigcap_{\gamma \in \partial^{\mathrm{ext}}} (\mathrm{Ext}\,\gamma \cap \Lambda)$, and the partition function $Z^{\mathrm{dil}}(\Lambda|\sigma_q)$ is defined by [9] with $\mathcal{M}(\Lambda, q)$ replaced by $\mathcal{M}^{\mathrm{dil}}(\Lambda, q) \subset \mathcal{M}(\Lambda, q)$, the set of all those collections whose external countours γ are such that $\mathrm{dist}\,(\Gamma, \Lambda^c) > 1$. Multiplying now each term by

$$1 = \prod_{\gamma \in \partial^{\mathrm{ext}}} \prod_m \frac{Z^{\mathrm{dil}}(\mathrm{Int}_m \gamma | \sigma_q)}{Z^{\mathrm{dil}}(\mathrm{Int}_m \gamma | \sigma_q)} \quad [11]$$

we get

$$Z(\Lambda|\sigma_q) = \sum_{\partial^{\mathrm{ext}}} \mathrm{e}^{-\beta e_q(H)|\mathrm{Ext}|} \times \prod_{\gamma \in \partial^{\mathrm{ext}}} \left(\mathrm{e}^{-\beta e_q(H)|\Gamma|} w_q(\gamma) Z^{\mathrm{dil}}(\mathrm{Int}\,\gamma | \sigma_q) \right) \quad [12]$$

where $w_q(\gamma)$ is given by

$$w_q(\gamma) = \mathrm{e}^{-\beta \Psi(\gamma)}\, \mathrm{e}^{\beta e_q(H)|\Gamma|} \prod_m \frac{Z^{\mathrm{dil}}(\mathrm{Int}_m \gamma | \sigma_m)}{Z^{\mathrm{dil}}(\mathrm{Int}_m \gamma | \sigma_q)} \quad [13]$$

Observing that a similar expression is valid for $Z^{\mathrm{dil}}(\Lambda|\sigma_q)$ (with an appropriate restriction on the sum over external contours ∂^{ext}) and proceeding by induction, we eventually get the representation

$$Z(\Lambda|\sigma_q) = \mathrm{e}^{-\beta e_q(H)|\Lambda|} \sum_{\partial \in \mathcal{C}(\Lambda,q)} \prod_{\gamma \in \partial} w_q(\gamma) \qquad [14]$$

where $\mathcal{C}(\Lambda, q)$ denotes the set of all collections of nonoverlapping q-contours in Λ. Clearly, the sum on the right-hand side is exactly of the form needed to apply cluster expansion, provided the contour weights satisfy the necessary convergence assumptions.

Even though this is not necessarily the case, there is a way to use this representation. Namely, one can artificially change the weights to satisfy the needed bound, for example, by modifying them to the form

$$w'_q(\gamma) = \min\left(w_q(\gamma), \mathrm{e}^{-\tau|\Gamma|}\right) \qquad [15]$$

with a suitable constant τ. The modified partition function

$$Z'(\Lambda|\sigma_q) = \mathrm{e}^{-\beta e_q(H)|\Lambda|} \sum_{\partial \in \mathcal{C}(\Lambda,q)} \prod_{\gamma \in \partial} w'_q(\gamma) \qquad [16]$$

can then be controlled by cluster expansion allowing to *define*

$$f_q(H) = -\frac{1}{\beta} \lim_{|\Lambda|\to\infty} \frac{1}{|\Lambda|} \log Z'(\Lambda|\sigma_q) \qquad [17]$$

This is the *metastable free energy* corresponding to the phase q. Applying the cluster expansion to the logarithm of the sum in [16], we get $|f_q(H) - e_q(H)| \leq \mathrm{e}^{-\tau/2}$. The metastable free energy corresponds to taking the ground state σ_q and its excitations as long as they are sufficiently suppressed. Once $w_q(\gamma)$ exceeds the weight $\mathrm{e}^{-\tau|\Gamma|}$ (and the contour would have been actually preferred), we suppress it "by hand." The point is that if the phase q is stable, this never happens and $w'_q(\gamma) = w_q(\gamma)$ for *all* q-contours γ. This is the idea behind the use of the function $f_q(H)$ as an indicator of the stability of the phase q by taking

$$a_q(\boldsymbol{t}) = f_q(H_{\boldsymbol{t}}) - \min_m f_m(H_{\boldsymbol{t}}) \qquad [18]$$

Of course, the difficult point is to actually prove that the stability of phase q (i.e., the fact that $a_q(\boldsymbol{t}) = 0$) indeed implies $w'_q(\gamma) = w_q(\gamma)$ for all γ. The crucial step is to prove, by induction on the diameter of Λ and γ, the following three claims (with $\epsilon = 2\mathrm{e}^{-\tau/2}$):

1. *If γ is a q-contour with $a_q(\boldsymbol{t})\,\mathrm{diam}\,\Gamma \leq \tau/4$, then* $w'_q(\gamma) = w_q(\gamma)$.
2. *If $a_q(\boldsymbol{t})\,\mathrm{diam}\,\Lambda \leq \tau/4$, then $Z(\Lambda|\sigma_q) = Z'(\Lambda|\sigma_q) \neq 0$ and*

$$\left|Z(\Lambda|\sigma_q)\right| \geq \mathrm{e}^{-f_q(H_t)|\Lambda| - \epsilon|\partial\Lambda|} \qquad [19]$$

3. *If $m \in G$, then*

$$|Z(\Lambda|\sigma_m)| \leq \mathrm{e}^{-\min_q f_q(H_t)|\Lambda|} \mathrm{e}^{\epsilon|\partial\Lambda|} \qquad [20]$$

A standard example illuminating the perturbative construction of the metastable free energies and showing the role of entropic contributions is the Blume–capel model. It is defined by the Hamiltonian

$$H_\Lambda(\sigma) = -J \sum_{\langle x,y \rangle} (\sigma_x - \sigma_y)^2 - \lambda \sum_{x\in\Lambda} \sigma_x^2 - h \sum_{x\in\Lambda} \sigma_x \qquad [21]$$

with spins $\sigma_x \in \{-1, 0, 1\}$. Taking into account only the lowest-order excitations, we get:

$$\tilde{f}_\pm(\lambda, h) = -\lambda \mp h - \frac{1}{\beta} \mathrm{e}^{-\beta(2d-\lambda\pm h)}$$

(sea of pluses or minuses with a single spin flip $\pm \to 0$) and

$$\tilde{f}_0(\lambda, h) = -\frac{1}{\beta} \mathrm{e}^{-\beta(2d+\lambda)} \left(\mathrm{e}^{\beta h} + \mathrm{e}^{-\beta h}\right)$$

(sea of zeros with a single spin flip either $0 \to +$ or $0 \to -$)

Since these functions differ from full metastable free energies $f_\pm(\lambda, h), f_0(\lambda, h)$ by terms of higher order ($\sim \mathrm{e}^{-(4d-2)\beta}$), the real phase diagram differs in this order from the one constructed by equating the functions $\tilde{f}_\pm(\lambda, h)$ and $\tilde{f}_0(\lambda, h)$. It is particularly interesting to inspect the origin, $\lambda = h = 0$. It is only the phase 0 that is stable there at all small temperatures since

$$f_0(0,0) \sim -\frac{2}{\beta} \mathrm{e}^{-\beta 2d} < f_\pm(0,0) \sim -\frac{1}{\beta} \mathrm{e}^{-\beta 2d} \qquad [22]$$

The only reason why the phase 0 is favored at this point with respect to phases $+$ and $-$ is that there are two excitations of order $\mathrm{e}^{-2d\beta}$ for the phase 0, while there is only *one* such excitation for $+$ or $-$. The entropy of the lowest-order contribution to $f_0(0,0)$ is overweighting the entropy of the contribution to $f_\pm(0,0)$ of the same order.

Applications

Several applications, stemming from the Pirogov–Sinai theory, are based on the fact that, due to the cluster expansion, we have quite accurate description of the model in finite volume.

One class of applications concerns various problems featuring interfaces between coexisting phases. To be able to transform the problem into a study of the random boundary line separating the two phases, one needs a precise cluster expansion formula for partition functions in volumes occupied by those phases. In the situation with no symmetry

between the phases, the use of the Pirogov–Sinai theory is indispensable.

Another interesting class of applications concerns the behavior of the system with periodic boundary conditions. It is based on the fact that the partition function Z_{T_N} on a torus T_N consisting of N^d sites can be, again with the help of the cluster expansions, explicitly and very accurately evaluated in terms of metastable free energies,

$$\left| Z_{T_N} - \sum_{q=1}^{r} \mathrm{e}^{-\beta f_q(H) N^d} \right| \leq \exp\{-\beta \min_m f_m(H) N^d - b\beta N\} \qquad [23]$$

with a fixed constant b. This formula (and its generalization to the case of complex parameters) allows us to obtain various results concerning the behavior of the model in finite volumes.

Finite-Size Effects

Considering as an illustration a perturbation of the Ising model, so that it does not have the $\pm$ symmetry any more (and the value $h_t(\beta)$ of external field at which the phase transition between plus and minus phase occurs is not known), we can pose a natural question that has an importance for correct interpretation of simulation data. Namely, what is the asymptotic behavior of the magnetization $m_N^{\mathrm{per}}(\beta, h) = \mu_{T_N}(1/\Lambda \sum_{x\in\Lambda} \sigma_x)$ on a torus? In the thermodynamic limit, the magnetization $m_\infty^{\mathrm{per}}(\beta, h)$ displays, as a function of h, a discontinuity at $h = h_t(\beta)$. For finite N, we get a *rounding* of the discontinuity – the jump is smoothed. What is the shift of a naturally chosen finite-volume transition point $h_t(N)$ with respect to the limiting value h_t? The answer can be obtained with the help of [23] once sufficient care is taken to use the freedom in the definition of the metastable free energies $f_+(h)$ and $f_-(h)$ to replace them with a sufficiently smooth version allowing an approximation of the functions $f_\pm(h)$ around limiting point h_t in terms of their Taylor expansion.

As a result, in spite of the asymmetry of the model, the finite-volume magnetization $m_N^{\mathrm{per}}(\beta, h)$ has a universal behavior in the neighborhood of the transition point h_t. With suitable constants m and m_0, we have

$$m_N^{\mathrm{per}}(\beta, h) \sim m_0 + m \tanh\{N^d \beta m (h - h_t)\} \qquad [24]$$

Choosing the inflection point $h_{\max}(N)$ of $m_N^{\mathrm{per}}(\beta, h)$ as a natural finite-volume indicator of the occurence of the transition, one can show that

$$h_{\max}(N) = h_t + \frac{3\chi}{2\beta^2 m^3} N^{-2d} + O(N^{-3d}) \qquad [25]$$

Zeros of Partition Functions

The full strength of the formula [23] is revealed when studying the zeros of the partition function $Z_{T_N}(z)$ as a polynomial in a complex parameter z entering the Hamiltonian of the model. To be able to use the theory in this case, one has to extend the definitions of the metastable free energies to complex values of z. Indeed, the construction still goes through, now yielding genuinely complex, contour models $w_\pm$ with the help of an inductive procedure. Notice that no analytic continuation is involved. An analog of [23] is still valid,

$$\left| Z_{T_N}(z) - \sum_{m=1}^{r} \mathrm{e}^{-\beta f_m(z) N^d} \right| \leq \exp\{-\beta \min_m \Re\mathrm{e} f_m(z) N^d - b\beta N\} \qquad [26]$$

Using [26], it is not difficult to convince oneself that the loci of zeros can be traced down to the phase coexistence lines. Indeed, on the line of the coexistence of two phases $\Re\mathrm{e} f_m = \Re\mathrm{e} f_q$, the partition function $Z_{T_N}(z)$ is approximated by $\mathrm{e}^{-\beta f N^d}(\mathrm{e}^{-\beta \Im\mathrm{m} f_m N^d} + \mathrm{e}^{-\beta \Im\mathrm{m} f_n N^d})$. The zeros of this approximation are thus given by the equations

$$\begin{aligned} &\Re\mathrm{e} f_m = \Re\mathrm{e} f_n < \Re\mathrm{e} f_\ell \quad \text{for all } \ell \neq m, n \\ &\beta N^d (\Im\mathrm{m} f_m - \Im\mathrm{m} f_n) = \pi \bmod 2\pi \end{aligned} \qquad [27]$$

The zeros of the full partition function $Z_{T_N}(z)$ can be proved to be exponentially close, up to a shift of order $\mathcal{O}(\mathrm{e}^{-\beta b N})$, to those of the discussed approximation.

Briefly, the zeros of $Z_{T_N}(z)$ asymptotically concentrate on the phase coexistence curves with the density $(1/2\pi)\beta N^d |(\mathrm{d}/\mathrm{d}z)(f_m - f_n)|$.

Bibliographical Remarks and Generalizations

The original works Pirogov and Sinai (1975, 1976) and Sinai (1982) introduced an analog of the weights $w'_q(\gamma)$ and parameters $a_q(H)$ as a fixed point of a suitable mapping on a Banach space. The inductive definition used here was introduced in Kotecký and Preiss (1983) and Zahradník (1984). The *completeness of phase diagram* – the fact that the stable phases exhaust the set of all periodic extremal Gibbs states was first proved in Zahradník (1984). Extension to complex parameters was first considered in Gawędzki *et al.* (1987) and Borgs and Imbrie (1989). For a review of the standard Pirogov–Sinai theory, see Sinai (1982) and Slawny (1987).

Application of Pirogov–Sinai theory for finite-size effects was studied in Borgs and Kotecký (1990) and

general theory of zeros of partition functions is presented in Biskup *et al.* (2004).

The basic statement of the Pirogov–Sinai theory yielding the construction of the full phase diagram has been extended to a large class of models. Let us mention just few of them (with rather incomplete references):

1. *Continuous spins*. The main difficulty in these models is that one has to deal with contours immersed in a sea of fluctuating spins (Dobrushin and Zahradník 1986, Borgs and Waxler 1989).
2. *Potts model*. An example of a system a transition in temperature with the coexistence of the low-temperature ordered and the high-temperature disordered phases. Contour reformulation is employing contours between ordered and disordered regions (Bricmont *et al.* 1985, Kotecký *et al.* 1990). The treatment is simplified with help of Fortuin–Kasteleyn representation (Laanait *et al.* 1991).
3. *Models with competing interactions*. ANNNI model, microemulsions. Systems with a rich phase structure (Dinaburg and Sinai 1985).
4. *Disordered systems*. An example is a proof of the existence of the phase transition for the three-dimensional random field Ising model (Bricmont and Kupiainen 1987, 1988) using a renormalization group version of the Pirogov–Sinai theory first formulated in Gawędzki *et al.* (1987).
5. *Quantum lattice models*. A class of quantum models that can be viewed as a quantum perturbation of a classical model. With the help of Feynman–Kac formula these are rewritten as a $(d+1)$-dimensional classical model that is, in its turn, treated by the standard Pirogov–Sinai theory (Datta *et al.* 1996, Borgs *et al.* 1996).
6. *Continuous systems*. Gas of particles in continuum interacting with a particular potential of Kac type. Pirogov–Sinai theory is used for a proof of the existence of the phase transitions after a suitable discretisation (Lebowitz *et al.* 1999).

See also: Cluster Expansion; Falicov–Kimball Model; Phase Transitions in Continuous Systems; Quantum Spin Systems.

Further Reading

Biskup M, Borgs C, Chayes JT, and Kotecký R (2004) Partition function zeros at first-order phase transitions: Pirogov–Sinai theory. *Journal of Statistical Physics* 116: 97–155.

Borgs C and Imbrie JZ (1989) A unified approach to phase diagrams in field theory and statistical mechanics. *Communications in Mathematical Physics* 123: 305–328.

Borgs C and Kotecký R (1990) A rigorous theory of finite-size scaling at first-order phase transitions. *Journal of Statistical Physics* 61: 79–119.

Borgs C, Kotecký R, and Ueltschi D (1996) Low temperature phase diagrams for quantum perturbations of classical spin systems. *Communications in Mathematical Physics* 181: 409–446.

Borgs C and Waxler R (1989) First order phase transitions in unbounded spin systems: construction of the phase diagram. *Communications in Mathematical Physics* 126: 291–324.

Bricmont J, Kuroda T, and Lebowitz J (1985) First order phase transitions in lattice and continuum systems: extension of Pirogov–Sinai theory. *Communications in Mathematical Physics* 101: 501–538.

Bricmont J and Kupiainen A (1987) Lower critical dimensions for the random field Ising model. *Physical Review Letters* 59: 1829–1832.

Bricmont J and Kupiainen A (1988) Phase transition in the 3D random field Ising model. *Communications in Mathematical Physics* 116: 539–572.

Datta N, Fernández R, and Fröhlich J (1996) Low-temperature phase diagrams of quantum lattice systems. I. Stability for quantum perturbations of classical systems with finitely-many ground states. *Helv. Phys. Acta* 69: 752–820.

Dinaburg EL and Sinaï YaG (1985) An analysis of ANNNI model by Peierls contour method. *Communications in Mathematical Physics* 98: 119–144.

Dobrushin RL and Zahradník M (1986) Phase diagrams of continuous lattice systems. In: Dobrushin RL (ed.) *Math. Problems of Stat. Physics and Dynamics*, pp. 1–123. Dordrecht: Reidel.

Gawędzki K, Kotecký R, and Kupiainen A (1987) Coarse-graining approach to first-order phase transitions. *Journal of Statistical Physics* 47: 701–724.

Kotecký R, Laanait L, Messager A, and Ruiz J (1990) The q-state Potts model in the standard Pirogov–Sinai theory: surface tensions and Wilson loops. *Journal of Statistical Physics* 58: 199–248.

Kotecký R and Preiss D (1983) An inductive approach to PS theory, *Proc. Winter School on Abstract Analysis, Suppl. ai Rend. del Mat. di Palermo*.

Lebowitz JL, Mazel A, and Presutti E (1999) Liquid–vapor phase transitions for systems with finite range interactions. *Journal of Statistical Physics* 94: 955–1025.

Laanait L, Messager A, Miracle-Solé S, Ruiz J, and Shlosman SB (1991) Interfaces in the Potts model I: Pirogov–Sinai theory of the Fortuin–Kasteleyn representation. *Communications in Mathematical Physics* 140: 81–91.

Pirogov SA and Sinai YaG (1975) Phase diagrams of classical lattice systems (Russian). *Theoretical and Mathematical Physics* 25(3): 358–369.

Pirogov SA and Sinai YaG (1976) Phase diagrams of classical lattice systems. Continuation (Russian). *Theoretical and Mathematical Physics* 26(1): 61–76.

Sinai YaG (1982) *Theory of Phase Transitions: Rigorous Results*. New York: Pergamon.

Slawny J (1987) Low temperature properties of classical lattice systems: phase transitions and phase diagrams. In: Domb C and Lebowitz JL (eds.) *Phase Transitions and Critical Phenomena*, vol. 11, pp. 127–205. New York: Academic Press.

Zahradník M (1984) An alternate version of Pirogov–Sinai theory. *Communications in Mathematical Physics* 93: 559–581.

Point-Vortex Dynamics

S Boatto, IMPA, Rio de Janeiro, Brazil
D Crowdy, Imperial College, London, UK

Introduction

Vortices have a long fascinating history. Descartes wrote in his *Le Monde*:

> ...que tous les mouvements qui se font au Monde sont en quelque façon circulaire: c'est à dire que, quand un corps quitte sa place, il entre toujours en celle d'un autre, et celui-ci en celle d'un autre, et ainsi de suite jusques au dernier, qui occupe au même instant le lieu délaissé par le premier.

In particular, Descartes thought of vortices to model the dynamics of the solar system, as reported by W W R Ball (1940):

> Descartes' physical theory of the universe, embodying most of the results contained in his earlier and unpublished *Le Monde*, is given in his *Principia*, 1644,... He assumes that the matter of the universe must be in motion, and that the motion must result in a number of vortices. He stated that the sun is the center of an immense whirlpool of this matter, in which the planets float and are swept round like straws in a whirlpool of water.

Descartes' theory was later on recused by Newton in his *Principia* in 1687. Few centuries later, W Thomson (1867) the later Lord Kelvin, made use of vortices to formulate his atomic theory: each atom was assumed to be made up of vortices in a sort of ideal fluid. In 1878–79 the American physicist A M Mayer conducted a few experiments with needle magnets placed on floating pieces of cork in an applied magnetic field, as toy models for studying atomic interactions and forms (Mayer 1878, Aref *et al.* 2003). In 1883 inspired by Mayer experiments, J J Thomson combined W Thomson's atomic theory with H von Helmholtz's point-vortex theory (Helmholtz 1858): he thought as the electrons were point vortices inside a positively charged shell (see Figure 1), the vortices being located at the vertices of regular parallelograms and investigated about the stability of such structures (see Thomson (1883, section 2.1)). The vortex-atomic theory survived for quite a few years up to Rutherford's experiments proved that atoms have quite a different structure! Before continuing this historical/modeling overview, let's address the following question:

> *what is a vortex and, more specifically, what is a point-vortex?*

Roughly speaking, following Descartes, a vortex is an entity which makes particles move along circular-like orbits. Examples are the cyclones and anticyclones in the atmosphere (see Figure 3). Mathematically speaking, let $\boldsymbol{u} = (u, v, w) \in \mathbb{R}^3$ be a velocity field, the associated vorticity field ω is defined to be

$$\omega = \nabla \wedge \boldsymbol{u} \qquad [1]$$

In this article we are considering exclusively inviscid flows which are also incompressible, that is,

$$\nabla \cdot \boldsymbol{u} = 0 \qquad [2]$$

and have constant density ρ, which we normalize to be equal to 1 ($\rho = 1$). In two dimensions, a point-vortex field is the simplest of all vorticity fields: it can be thought as an entity where the vorticity field is concentrated into a point. In other words, point vortices are singularities of the vorticity field! Then, in the plane the vorticity field associated to a system of N point vortices is

$$\omega(\boldsymbol{r}) = \sum_{\alpha=1}^{N} \Gamma_\alpha \delta(\boldsymbol{r} - \boldsymbol{r}_\alpha) \qquad [3]$$

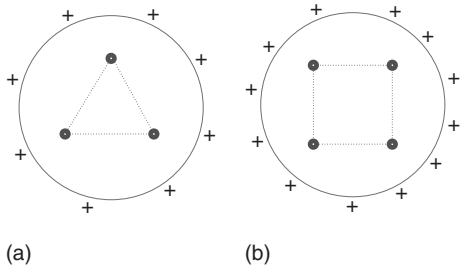


Figure 1 Thomson atomic model: (a) atom with three electrons and (b) atom with four electrons. From Thomson JJ (1883) *A Treatise on the Motion of Vortex Rings*. New York: Macmillan and Thomson JJ (1904) *Electricity and Matter*. Westminster: Archibald Constable.

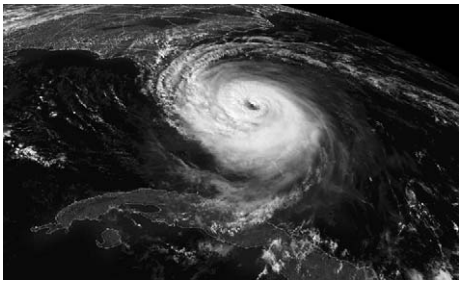


Figure 2 Hurricane Jeanne. Reproduced with permission from the National Oceanic and Atmospheric Administration (NOAA) (www.noaanews.noaa.gov).

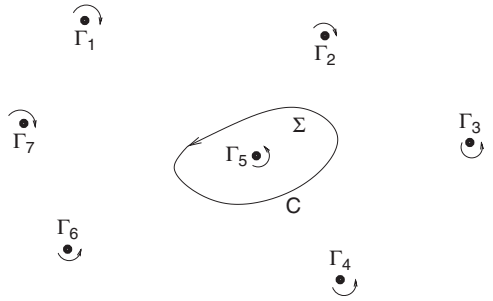


Figure 3 Cyclones and anticyclones in the atmosphere. Reproduced from Boatto S and Cabrel HE, *SIAM Journal of Applied Mathematics* 64:216–230 (2003). With the permission of SIAM.

where Γ_α, $\alpha = 1, \ldots, N$, is a constant and corresponds to the vorticity (or circulation) of the α-vortex, situated at $\boldsymbol{r}_\alpha$. In fact by definition, the circulation around a curve C delimiting a region Σ with boundary C,

$$\Gamma_C = \oint_C \boldsymbol{u} \cdot \mathrm{d}\boldsymbol{s} = \iint_\Sigma (\nabla \wedge \boldsymbol{u}) \cdot \boldsymbol{n}\, \mathrm{d}A = \iint_\Sigma \omega \quad [4]$$

where we have used Stokes' theorem to bring in the vorticity. Then if the region contains only the αth point vortex, we obtain

$$\Gamma_C = \iint_\Sigma \omega \cdot \mathrm{d}\boldsymbol{A} = \Gamma_\alpha \quad [5]$$

by eqn [3]. A positive (resp. negative) sign of Γ_α indicates that the corresponding point vortex induces an anticlockwise (resp. clockwise) particle motion, see **Figure 4a**)). Is there an analog of a point-vortex system for a three-dimensional flow?

Yes, and this brings in the analogy between vortex lines and magnetic field lines that Mayer used in his experiments with floating magnets. In fact, in three dimensions, the notion of a point vortex can be extended to that one of a straight vortex line (see **Figure 4b**), where, by definition, a vortex line is a curve that is tangent to the vorticity vector ω at each of its point. In this context we would like to mention the beautiful experiments of Yarmchuck–Gordon–Packard on vortices in superfluid helium. They observed the formation of stable polygonal configurations of identical vortices, quite similar to the ones observed by Mayer with his magnets (see **Figures 5** and **1**).

One would like to understand how such configurations form and to give a theoretical account about their stability. In order to answer these questions we have to first be able to describe the dynamics of a system of point vortices from a mathematical point of view.

Evolution Equations

Can point vortices be viewed as "discrete" (or localized) solutions of Euler equation in two dimensions? Let us consider the Euler equation

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} = -\nabla p + \boldsymbol{f} \quad [6]$$

where p is the pressure, $\boldsymbol{f} = -\nabla U$ is a conservative force, and restrict our attention to the two-dimensional setting, for example, vortex dynamics on the plane (or a sphere). Then it is immediate that by taking the curl of eqn [6] we obtain the evolution equation of the vorticity, that is,

$$\frac{\partial \omega}{\partial t} + \boldsymbol{u} \cdot \nabla \omega = 0, \quad \text{or} \quad \frac{D\omega}{Dt} = 0 \quad [7]$$

where the operator $D/Dt = \partial/\partial t + \boldsymbol{u} \cdot \nabla$ is called the material derivative and describes the evolution along the flow lines. It follows from eqn [7] that in two dimensions the vorticity is conserved as it is transported along the flow lines. Then a natural question arises: supposing the vorticity field ω is known, is it possible to deduce the velocity field $\boldsymbol{u}$ generating ω? Or in other words, is it possible to solve the system of eqns [1]–[2]? It is immediate to see that in general the solution is not unique, if some boundary conditions are not specified (see Marchioro and Pulvirenti (1993)). Furthermore, as already observed by Kirchhoff in 1876 (Boatto and Cabral 2003), in two dimensions we can recast the fluid equations [1]–[2] into a Hamiltonian formalism. In fact, notice that on the plane $\boldsymbol{u} = (\dot{x}, \dot{y})$ and eqn [2] is still satisfied if we represent the velocity components as

Figure 4 (a) Advected by the velocity field of one point vortex, a test particle follows a circular orbit, with a speed proportional to the absolute value of the vortex circulation and inversely proportional to the square of its distance from the vortex. (b) Straight vortex lines.

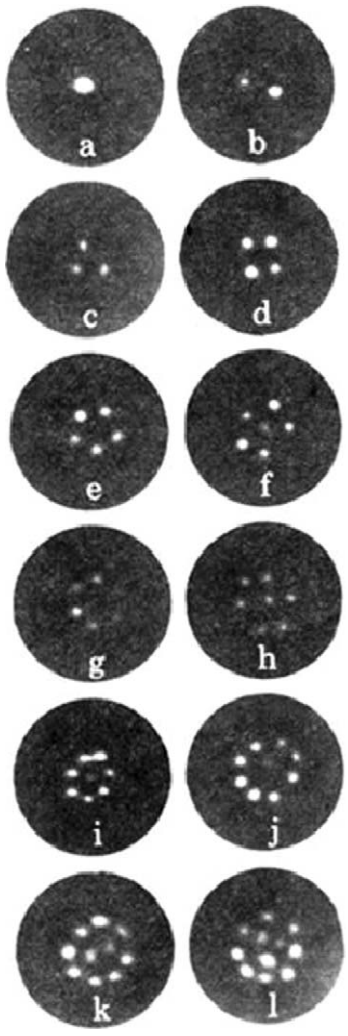


Figure 5 Photographs of vortex configurations in a rotated sample of superfluid helium with 1,...,11 vortices. Reprinted figure with permission from Yarmchuk EJ, Gordon MJV, and Packard RE (1979) Observation of stationary vortices arrays in rotating superfluid Helium. *Physical Review Letters* 43(3): 214–217. Copyright (1979) by the American Physical Society.

$$\dot{x} = \frac{\partial \Psi}{\partial y}, \qquad \dot{y} = -\frac{\partial \Psi}{\partial x} \tag{8}$$

that is, by means of Ψ, called the stream function. Formally, Ψ plays the rôle of a Hamiltonian for the pair of conjugate variables (x, y) and it is used to describe the dynamics of a test particle, located at (x, y) and advected by the flow. By substituting [8] into [1], we obtain

$$\Delta \Psi(\boldsymbol{r}) = \omega(\boldsymbol{r}) \tag{9}$$

that is, a Poisson equation with ω as a source term. Then, once we specify the vorticity field, by inverting [9] we obtain the stream function Ψ to be

$$\Psi(\boldsymbol{r}) = \int G(\boldsymbol{r}, \boldsymbol{r}')\omega(\boldsymbol{r}')\, \mathrm{d}\boldsymbol{r}' \tag{10}$$

where $G(\boldsymbol{r}, \boldsymbol{r}')$ is the Green's function, solution of the equation $\Delta G(x, y) = -\delta(x, y)$. The Green's function both for the plane and the sphere is (Marchioro and Pulvirenti 1993)

$$G(\boldsymbol{r}, \boldsymbol{r}') = -\frac{1}{4\pi} \log \|\boldsymbol{r} - \boldsymbol{r}'\|^2 \tag{11}$$

where $\|\boldsymbol{r} - \boldsymbol{r}'\|^2 = (x - x')^2 + (y - y')^2$. By [10], once we specify the vorticity field $\omega(\boldsymbol{r})$ we can compute Ψ, and by replacing it into [8] the velocity field becomes

$$\boldsymbol{u}(\boldsymbol{r}) = \int \boldsymbol{K}(\boldsymbol{r}, \boldsymbol{r}')\omega(\boldsymbol{r}')\, \mathrm{d}\boldsymbol{r}' \tag{12}$$

where $\boldsymbol{K}(\boldsymbol{r}, \boldsymbol{r}') = -(\boldsymbol{r} - \boldsymbol{r}')^{\perp}/[2\pi \|\boldsymbol{r} - \boldsymbol{r}'\|^2]$ and it represents the velocity field generated by a point vortex of intensity one, located at $\boldsymbol{r}'$. Then by considering the vorticity field generated by point vortices, eqn [3], together with eqn [11], eqn [10] becomes

$$\begin{aligned}\Psi(\boldsymbol{r}) &= -\frac{1}{4\pi} \int \log \|\boldsymbol{r} - \boldsymbol{r}'\|^2 \left(\sum_{\alpha=1}^{N} \Gamma_\alpha \delta(\boldsymbol{r}' - \boldsymbol{r}_\alpha) \right) \mathrm{d}\boldsymbol{r}' \\ &= -\frac{1}{4\pi} \sum_{\alpha=1}^{N} \Gamma_\alpha \log \|\boldsymbol{r} - \boldsymbol{r}_\alpha\|^2 \end{aligned} \tag{13}$$

Equation [13] describes together with [8], the dynamics of a test particle at a point $\boldsymbol{r} = (x, y)$ in the plane. Analogously, it can be shown that the dynamics of a systems of point vortices in the plane is given by the equations

$$\Gamma_\alpha \frac{\mathrm{d}x_\alpha}{\mathrm{d}t} = \frac{\partial H_\nu}{\partial y_\alpha}, \qquad \Gamma_\alpha \frac{\mathrm{d}y_\alpha}{\mathrm{d}t} = -\frac{\partial H_\nu}{\partial x_\alpha} \tag{14}$$

where $(q_\alpha, p_\alpha) = (x_\alpha, \Gamma_\alpha y_\alpha), \alpha = 1, \ldots, N$, is a pair of conjugate variables and H_ν is the generalization of the stream function Ψ (eqn [13]):

$$H_\nu = -\frac{1}{4\pi} \sum_{\substack{\alpha,\beta=1 \\ \alpha \neq \beta}}^{N} \Gamma_\alpha \Gamma_\beta \log \|\boldsymbol{r}_\alpha - \boldsymbol{r}_\beta\|^2 \tag{15}$$

Notice that the vortex Hamiltonian H_ν (eqn [15]) is an autonomous Hamiltonian and, as we will discuss in the first subsection, it provides a good Lyapunov-like function to study stability properties of some vortex configurations. Moreover, H_ν is invariant with respect to rotations and translations, then by the Noether theorem there are other first integrals of motion, that is,

$$L = \sum_{k=1}^{N} \Gamma_k \| \boldsymbol{x}_k \|^2, \qquad M_x = \sum_{k=1}^{N} \Gamma_k x_k,$$

$$M_y = \sum_{k=1}^{N} \Gamma_k y_k$$

expressing, respectively, the conservation of angular momentum, L, and linear momentum, $\boldsymbol{M} = (M_x, M_y)$, on the plane. We shall denote with M the magnitude of $\boldsymbol{M}$ (i.e., $M = \|\boldsymbol{M}\|$). Furthermore, by introducing the Poisson bracket

$$[f,g] = \sum_{\alpha=1}^{N}\left(\frac{\partial f}{\partial q_\alpha}\frac{\partial g}{\partial p_\alpha} - \frac{\partial f}{\partial p_\alpha}\frac{\partial g}{\partial q_\alpha}\right)$$
$$= \sum_{\alpha=1}^{N}\frac{1}{\Gamma_\alpha}\left(\frac{\partial f}{\partial x_\alpha}\frac{\partial g}{\partial y_\alpha} - \frac{\partial f}{\partial y_\alpha}\frac{\partial g}{\partial x_\alpha}\right)$$

we can construct three integrals in involution out of the four conserved quantities L, M_x, M_y, and H_ν. These are $L, M_x^2 + M_y^2$ and H_ν: in fact,

$$[H_\nu, L] = 0, \qquad \left[H_\nu, M_x^2 + M_y^2\right] = 0,$$
$$\left[L, M_x^2 + M_y^2\right] = 0$$

It is then possible to reduce the system of equations from N to $N-2$ degrees of freedom. A Hamiltonian system with N degrees of freedom is integrable whenever there are N independent integrals of motion in involution. It follows that a vortex system with $N \leq 3$ is integrable, whereas the system of equations of four identical vortices has been shown by Ziglin to be nonintegrable in the sense that there are no other first integrals analytically depending on the coordinates and circulations, and functionally independent of L, H_ν, M_x, M_y (see Ziglin (1982)). The following, however, has been shown:

1. Let $K = \sum_{\alpha=1}^{N} k_\alpha$ be the total vorticity, $\boldsymbol{M} = (M_x, M_y)$ the total momentum and $M = \|\boldsymbol{M}\|$. Then, as shown by Aref and Stremler (1999), if $K = 0$ and $M = 0$, N-vortex problem [16] is integrable.
2. A system of four identical vortices (i.e., $k_\alpha = k$ for $\alpha = 1, \ldots, 4$) can undergo periodic or quasiperiodic motion for special initial conditions (see Khanin (1981) *Russian Math. Surveys* 36: 231; Aref and Pomphrey (1982) *Proc. R. Soc. Lond. A* 380: 359–387). More specifically, the motion of a system of four identical vortices can be periodic, quasiperiodic, or chaotic depending on the symmetry of the initial configuration. In fact, every vortex configuration that belongs to the subspace of symmetric configurations – $x_\alpha = -x_{\alpha+2}$ and $y_\alpha = y_{\alpha+2}, \alpha = 1, 2$ – gives rise to an integrable vortex motion.

We have that up to two vortices, the motion is almost always periodic and the orbits are circles; the only exception being the case for which $k_2 = -k_1$, when the circles degenerate into straight lines. Thus, a configuration of two point vortices is always a relative equilibrium configuration, that is, there exists a specific reference frame in which the two vortices are at rest. If the vortices are identical ($\Gamma_1 = \Gamma_2 = \Gamma$), the motion is synchronous with frequency $\Omega = \Gamma/\pi$ and the vortices share the same circular orbit (see Figure 6a). If the vortices are not identical and have vorticities of different magnitudes (say $|\Gamma_1| > |\Gamma_2|$), their motion is still synchronous and periodic, with frequency $\Omega = (\Gamma_1 + \Gamma_2)/(2\pi)$, and the vortices move on different circular orbits (with $r_2 < r_1$) both centered at the center of vorticity. Note that for both cases, identical and nonidentical vortices, we can view the vortex dynamics in a co-rotating frame where the vortices are simply at rest.

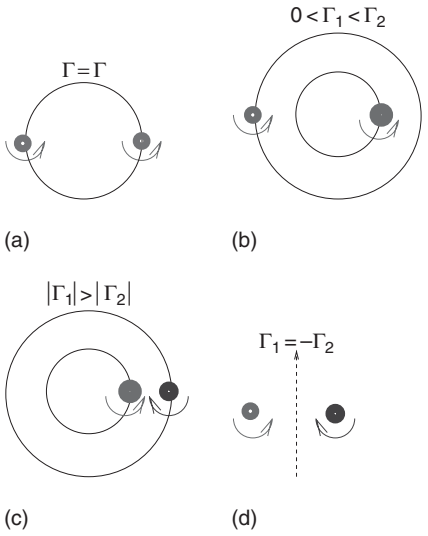


Figure 6(a–d) For $N = 2$ the vortex dipole exhibits a synchronous and the orbits are in general circular orbits, with the exception of the case (d) for which $\Gamma_1 = -\Gamma_2$ and the circular orbit degenerates into a line (or a circle of infinite radius).

For three vortices we can have periodic and quasiperiodic motion, depending on the initial conditions, and for four vortices we can have periodic, quasiperiodic, or weakly chaotic motion.

Remarks

(i) The nonintegrability of the 4-vortex system was also proved for configurations of nonidentical vortices. Koiller and Carvalho (1989) gave an analytical proof for $\Gamma_1 = -\Gamma_2$ and $\Gamma_3 = \Gamma_4 = \epsilon, 0 \ll \epsilon \ll 1$. Moreover, Castilla *et al.* (1993) considered the case: $\Gamma_1 = \Gamma_2 = \Gamma_3 = 1$ and $\Gamma_4 = \epsilon$.

(ii) Due to the translational and rotational symmetries of H_ν, there are some analogies between the N-vortex problem and the N-body problem, especially for what concerns configurations of relative equilibria (see Albouy (1996) and Glass (2000)). A relative equilibrium is a vortex (or mass) configuration that moves without change of shape or form, that is, a configuration which is steadily

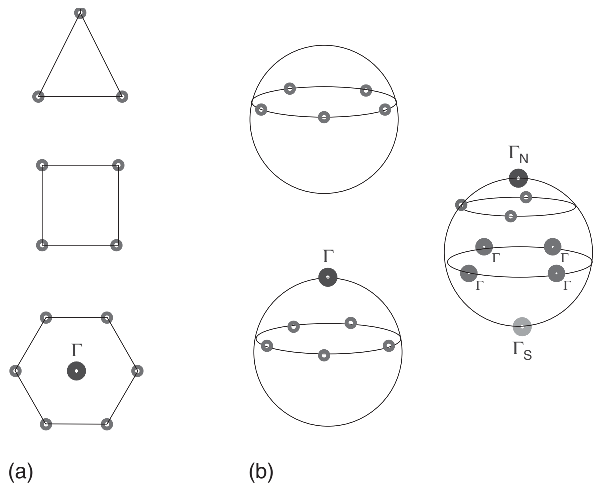


Figure 7 Polygonal configuration of vortices: (a) planar configurations and (b) configurations of vortex rings on a sphere, with and without polar vortices.

rotating or translating. A few examples are vortex polygons (see **Figure 7**) like the ones studied by Thomson, Mayer, Yarmchuk–Gordon–Packard, Boatto–Cabral (2003), Cabral–Schmidt (1999/2000), Dritschel–Polvani (1993), Lim–Montaldi–Roberts (2001), Sakajo (2004). For an exhaustive review on relative equilibria of vortices, see the article by Aref *et al.* (2003). We shall discuss stability of polygonal vortex configuration in the following subsection.

(iii) As shown by Kimura (1999) in a beautiful geometrical formalism, on the unit sphere (S^2) and on the Hyperbolic plane (H^2), the vortex Hamiltonians [15] are

$$H_v = -\frac{1}{4\pi}\sum_{\alpha\neq\beta}^{N}\Gamma_\alpha\Gamma_\beta\log(1-\cos\rho_{\alpha\beta}) \quad \text{on } S^2$$

$$H_v = -\frac{1}{4\pi}\sum_{\alpha\neq\beta}^{N}\Gamma_\alpha\Gamma_\beta\log\frac{\cosh\rho_{\alpha\beta}-1}{\cosh\rho_{\alpha\beta}+1} \quad \text{on } H^2$$

where

$$\cos\rho_{\alpha\beta} = \cos\theta_\alpha\cos\theta_\beta + \sin\theta_\alpha\sin\theta_\beta\cos(\phi_\alpha-\phi_\beta) \quad \text{on } S^2$$

$$\cosh\rho_{\alpha\beta} = \cosh\theta_\alpha\cosh\theta_\beta + \sinh\theta_\alpha\sinh\theta_\beta\cos(\phi_\alpha-\phi_\beta) \quad \text{on } H^2$$

On S^2, θ_α and ϕ_α are, respectively, the co-latitude and the longitude of the α-vortex, $\alpha = 1, \ldots, N$. We can define canonical variables q_α and p_α on S^2 and H^2, respectively, as

$$q_\alpha = \Gamma_\alpha\cos\theta_\alpha, \qquad p_\alpha = \phi_\alpha \quad \text{on } S^2$$

$$q_\alpha = \Gamma_\alpha\cosh\theta_\alpha, \qquad p_\alpha = \phi_\alpha \quad \text{on } H^2$$

Montaldi *et al.* (2002) studied vortex dynamics on a cylindrical surface, and Soulière and Tokieda (2002) considered vortex dynamics on surfaces with symmetries.

(iv) As we shall see in the section on point vortex motion, it is sometimes useful to employ the complex analysis formalism. Then the variables of interest are $z_\alpha = x_\alpha + iy_\alpha, \alpha = 1, \ldots, N$, and its conjugate $\bar{z}_\alpha$, the Hamiltonian [15] takes the form

$$H_v = -\frac{1}{2\pi}\sum_{\alpha\neq\beta}\Gamma_\alpha\Gamma_\beta\log|z_\alpha - z_\beta|$$

and the equations of motions become

$$\dot{z}_\alpha = \frac{i}{2\pi}\sum_{\beta\neq\alpha,\beta=1}^{N}\Gamma_\beta\frac{z_\alpha - z_\beta}{|z_\alpha - z_\beta|^2}, \quad \alpha = 1, \ldots, N \quad [16]$$

(v) Equation [14] can we rewritten in a more compact form as

$$\frac{dX}{dt} = J\nabla_X H_v \quad [17]$$

where

$$X = (q_1, \ldots, q_N, p_1, \ldots, p_N)$$

$$\nabla_X = \left(\frac{\partial}{\partial q_1}, \ldots, \frac{\partial}{\partial q_N}, \frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_N}\right)$$

$$J = \begin{pmatrix} O & \mathbb{I} \\ -\mathbb{I} & O \end{pmatrix}$$

$\mathbb{I}$ being the $N \times N$ identity matrix.

(vi) How close is the point-vortex model to the original Euler equation? Point-vortex systems represent discrete solutions of the Euler equation in a "weak" sense – see both the book and the article by Marchioro and Pulvirenti (1993, 1994). These authors proved that the Euler dynamics is "similar" to the vortex dynamics in which the vortices are localized in very small regions, and the vortex intensities are the total vorticities associated to such small regions. In particular, let us consider a vorticity field with compact support on a family of ϵ-balls, that is,

$$\omega^\epsilon = \sum_{i=1}^{N}\omega_i^\epsilon$$

with support of ω_i^ϵ contained in the ball of center x_i (independent of ϵ) and radius ϵ. Furthermore let us assume that

$$\int_{|\boldsymbol{r}-\boldsymbol{r}_i|\leq\epsilon}\omega_i^\epsilon\, d\boldsymbol{r} = \Gamma_i$$

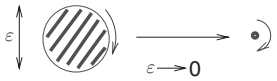


Figure 8 In the limit $\epsilon \to 0$, the dynamics of the center of vorticity of a vortex ϵ-ball is approximated by the dynamics of a point vortex.

with the γ_i independent of ϵ. Then in the limit $\epsilon \to 0$ the dynamics of the center of vorticity $B_\epsilon(t) = \int \boldsymbol{r}\omega_\epsilon(\boldsymbol{r}, t)\, d\boldsymbol{r}$, of a given ϵ-ball, "converges" to the motion of a single point vortex (see **Figure 8**). This result is important to illustrate as vortex systems provide both a useful heuristic tool in the analysis of the general properties of the solutions of Euler's equations (Poupaud 2002, Schochet 1995), and a useful starting point for the construction of practical algorithms for solving equations in specific situations. In particular, it provides a theoretical justification to the vortex method previously introduced by Carnevale *et al.* (1992). These authors constructed a numerical algorithm to study turbulence decaying in two dimensions. Their vortex method greatly simplifies fluid simulations as basically it relies on a discretization of the fluid into circular patches. The dynamics of patches is given by the centers of vorticity, which interact as a point-vortex system, endowed with a rule dictating how patches merge (see **Figure 9**).

Stability of a Vortex Ring

As mentioned in the Introduction section, the study of vortex relative equilibria has a long history. Kelvin showed that steadily rotating patterns of identical vortices arise as solutions of a variational problem in which the interaction energy (vortex Hamiltonian) is minimized subject to the constraint that the angular impulse be maintained (see Aref (2003). In 1883, while studying and modeling the atomic structure, J J Thomson investigated the linear stability of co-rotating point vortices in the plane. In particular, his interest was in configurations of identical vortices equally spaced along the circumference of a circle, that is, located at the vertices of a regular polygon (see **Figure 7**). He proved that for six or fewer vortices the polygonal configurations are stable, while for seven vortices – the Thomson heptagon – he erroneously concluded that the configuration is slightly unstable. It took more than a century to make some progresses on this problem. D G Dritschel (1985) succeeded in solving the heptagon mystery for what concerns its linear stability analysis, leaving open the nonlinear stability question: he proved that the Thomson heptagon is neutrally stable and that for eight or more vortices the corresponding polygonal configurations are linearly unstable. Later on in 1993, Polvani and Dritschel (1993) generalized the techniques used in Dritschel (1985) to study the linear stability of a "latitudinal" ring of point vortices on the sphere, as a function of the number N of vortices in the ring, and of the ring's co-latitude θ (see **Figure 10**). They proved that polygonal configurations are more unstable on the sphere than in the plane. In particular, they showed that at the pole, for $N < 7$ the configuration is stable, for $N = 7$ it is neutrally stable and for $N > 7$ it is unstable. By means of the energy momentum method (Marsden–Meyer–Weistein reduction), J E Marsden and S Pekarsky (1998) studied the nonlinear stability analysis for the integrable case of polygonal configurations of three vortices of arbitrary vorticities (Γ_1, Γ_2 and Γ_3) on the sphere, leaving open the stability analysis for nonintegrable vortex systems ($N > 3$). In 1999 H E Cabral and D S Schmidt completed the linear and nonlinear stability analysis at once for polygonal configurations in the plane. In 2003 Boatto and Cabral studied the nonlinear stability of a ring of vortices on the sphere, as a function of the number N and the ring colatitude θ.

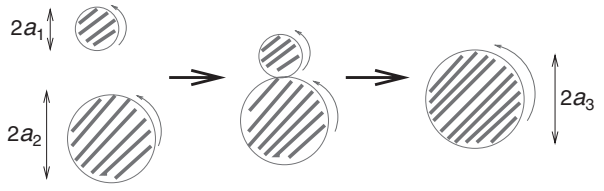


Figure 9 In Carnevale *et al.* (1992) the fluid is modeled by a dilute vortex gas with density ρ and typical radius a. The dynamics is governed by the point-vortex dynamics of the disk centers, each disk corresponding to a point vortex of intensity $\Gamma = \pi\xi_{ext}a^2$, where ξ_{ext} plays the role of a vorticity density. Two vortices or radius a_1 and a_2 merge when their center-to-center distance is less or equal to the sum of their radii, $a_1 + a_1$. Then a new vortex is created and its radius a_3 is given by $a_3 = (a_1^4 + a_2^4)^{1/4}$.

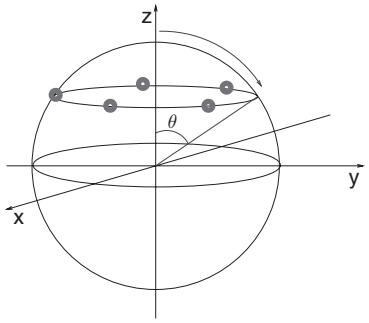


Figure 10 Latitudinal ring of vortices. Reproduced with permission from Boatto S and Cabral HE *SIAM Journal of Applied Mathematics* 64: 216–230 (2003).

Boatto and Simó (2004) generalized the stability analysis to the case of a ring with polar vortices and of multiple rings, the key idea being, as we shall discuss in this section, the structure of the Hessian of the Hamiltonian.

How to infer about linear and nonlinear stability of steadily rotating configurations?

Let us restrict the discussion to a polygonal ring of identical vortices on a sphere as illustrated in **Figure 7** (Boatto and Cabral 2003, Boatto and Simó 2004). The reasoning is easily generalized for the planar case. The case of multiple rings is discussed in great detail in Boatto and Simó (2004). A polygonal ring is a relative equilibrium of coordinates $X(t) = (q_1(t), \ldots, q_N(t), p_1(t), \ldots, p_N(t))$, where

$$\begin{aligned} q_\alpha(t) &= \phi_\alpha(t) = \omega t + \phi_{o\alpha} \\ p_\alpha(t) &= p_o = \Gamma \cos\theta_o \quad \alpha = 1, \ldots, N \end{aligned} \qquad [18]$$

$\omega = (N-1)p_o/r_o^2, r_o = \sqrt{1 - p_o^2/\Gamma^2}, \phi_{o\alpha}$ and $\theta_{o\alpha} = \theta_o$ being the initial longitude and co-latitude of the αth vortex.

Theorem 1 (Spherical case) (Boatto and Simó 2004). *The relative equilibrium* [18] *is* (*linearly and nonlinearly*) *stable if*

$$\begin{aligned} &-4(N-1)(11-N) + 24(N-1)r_o^2 \\ &\quad + 2N^2 + 1 + 3(-1)^N < 0 \end{aligned} \qquad [19]$$

and it is unstable if the inequality is reversed.

Remarks

(i) By Theorem 1 a vortex polygon, of N point vortices, is stable for $0^\circ \le \theta_o \le \theta_o^*$ and $(180^\circ - \theta_o^*) \le \theta_o \le 180^\circ$, where $\theta_o^* = \arcsin(r_o^*)$ and

$$\begin{aligned} r_o^{*2} &< \frac{7-N}{4} && \text{for } N \text{ odd} \\ r_o^{*2} &< -\frac{N^2 - 8N + 8}{4(N-1)} && \text{for } N \text{ even} \end{aligned}$$

where $r_o^* = \sin\theta_o^*$.

(ii) Theorem 1 includes at once the results of Thomson (1883), Dritschel (1985), and Polvani and Dritschel (1993) (and other authors who have been working in the area (Aref *et al.* 2003)). We recover the planar case by setting $r_o = 0$ in eqn [19], deducing that stability is guaranteed for $N < 7$.

To prove Theorem 1 it is useful to consider the Hamiltonian equations as in eqn [17]. The first step is to make a change of reference frame: view the dynamics in a frame co-rotating with the relative equilibrium configuration. In the co-rotating reference system, the Hamiltonian takes the form

$$\tilde{H} = H + \omega M$$

where M is the momentum of the system, and H and ω are, respectively, the Hamiltonian and the rotational frequency of the relative equilibrium in the original frame of reference. In the new reference frame, the relative equilibrium becomes an equilibrium, X^*, and the standard techniques can be used to study its stability.

To study linear stability, the relevant equation is

$$\frac{\mathrm{d}\Delta X}{\mathrm{d}t} = JS\Delta X \qquad [20]$$

where $X = X^* + \Delta X$, and S is the Hessian of $\tilde{H}$ evaluated at the equilibrium X^*. Then linear (or spectral) stability is deduced by studying the eigenvalues of the matrix JS (spectral stability). For nonlinear stability we make use of a sufficient stability criterion due to Dirichlet (1897) (see G Lejeune Dirichlet (1897). *Werke*, vol. 2, Berlin, pp. 5–8; Boatto and Cabral (2003) and references therein).

Theorem 2 *Let X^* be an equilibrium of an autonomous system of ordinary differential equations*

$$\frac{\mathrm{d}X}{\mathrm{d}t} = f(X), \qquad \Omega \subset \mathbb{R}^{2N} \qquad [21]$$

that is, $f(X^*) = 0$. If there exists a positive (*or negative*) *definite integral $\mathbf{F}$ of the system* [21] *in a neighborhood of the equilibrium X^*, then X^* is stable.*

In our case the Hamiltonian itself is an integral of motion. Then by studying definiteness of its Hessian, S, evaluated at X^*, we infer minimal stability intervals in θ and N. Details are given in Boatto and Cabral (2003) and Boatto and Simó (2004). The proof is mainly based on the following considerations:

1. Since S is a symmetric matrix it is diagonalizable, that is, there exists an orthogonal matrix C such that $C^{\mathrm{T}}SC = D$, where D is a diagonal matrix, $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$. Furthermore, the matrix C can be chosen to leave invariant the symplectic form (equivalently $J = C^{\mathrm{T}}JC$). Then by the canonical change of variables $Y = C^{\mathrm{T}}X$ eqn [20] becomes

$$\frac{\mathrm{d}\Delta Y}{\mathrm{d}t} = JD\Delta Y \qquad [22]$$

where $Y=(\tilde{q}_1,\ldots,\tilde{q}_N,\tilde{p}_1,\ldots,\tilde{p}_N)$ and $(\tilde{q}_j,\tilde{p}_j)$, $j=1,\ldots,N$, are pairs of conjugate variables. Equation [22] can be rewritten as

$$\frac{\mathrm{d}^2\Delta\tilde{q}_j}{\mathrm{d}t^2}=-\lambda_j\lambda_{j+N}\Delta\tilde{q}_j,\quad j=1,\ldots,N$$

2. When evaluated at the equilibrium X^*, the Hessian S takes the block structure

$$\tilde{S}=\left(\begin{array}{c|c}Q & O\\ \hline O & P\end{array}\right)$$

where the matrices Q and P are symmetric circulant matrices, that is, $(N\times N)$ matrices of the form

$$A=\begin{pmatrix}a_1 & a_2 & \ldots & a_N\\ a_N & a_1 & \ldots & a_{N-1}\\ \vdots & \vdots & \ddots & \vdots\\ a_2 & a_3 & \ldots & a_1\end{pmatrix} \qquad [23]$$

Circulant matrices are of special interest to us because we can easily compute their eigenvalues and eigenvectors for all N. In fact, it is immediate to show that:

Lemma 3 *All circulant matrices [23] have eigenvalues*

$$\lambda_j=\sum_{k=1}^{N}a_k r_j^{k-1},\quad j=1,\ldots,N$$

and corresponding eigenvectors $v_j=(1,r_j,\ldots,r_j^{N-1})^{\mathrm{T}}, j=1,\ldots,N$, *where* $r_j=\exp(2\pi(j-1)/N)$ *are solutions of* $r^N=1$.

Passive Tracers in the Velocity Fields of N Point Vortices: The Restricted $(N+1)$-Vortex Problem

The terminology "restricted $(N+1)$-vortex problem" is used in analogy with celestial mechanics literature, when one of the vorticities is taken to be zero. The zero-vorticity vortex does not affect the dynamics of the remaining N-vortices. For this reason, it is said to be passively advected by the flow of the remaining N-vortices and in the fluid mechanics literature the terminology "passive tracer" is also employed. The tracer dynamics is given by the Hamiltonian equations [8]. Notice that in general the Hamiltonian Ψ is time dependent, through the vortex variables $\boldsymbol{r}_j, j=1,\ldots,N$, that is,

$$\Psi(\boldsymbol{r},t)=\Psi(\boldsymbol{r},\boldsymbol{r}_1(t),\ldots,\boldsymbol{r}_N(t))$$

and $(q,p)=(x,y)$ play the role of conjugate canonical variables. There is an extensive literature on the subject both from theoretical (see, e.g., Boatto and Simó (2004) and Newton (2001)) and an experimental (van Heijst 1993, Ottino 1990) point of view. As discussed in the previous section, there are some vortex configurations, such as the polygonal ones, for which vortices undergo a periodic circular motion. Then by viewing the dynamics in a reference frame co-rotating with the vortices the tracer Hamiltonian is manifestly time independent and, therefore, integrable – since it reduces to a Hamiltonian of one degree of freedom. In such an occurrence, tracer trajectories form a web of homoclinic and heteroclinic orbits. An interesting theoretical problem is to study how the tracer transport properties (i.e., existence of barriers to transport, diffusion etc.) are affected by perturbing the polygonal vortex configuration, that is, by introducing in Ψ a "genuine" time dependence (periodic, quasi-periodic, or chaotic) (see, e.g., Boatto and Pierrehumbert (1999), Rom-Kedar, Leonard and Wiggins (1990), Kuznetsov and Zaslavsky (2000), and Newton (2001)). Furthermore, in the lab experiments, color dyes, which monitor the flow velocity field, are often used as the experimental equivalent of tracer particles. In this context we would like to stress the striking resemblance between theoretical particle trajectories, deduced from point vortex dynamics, and the actual dye visualizations observed by van Heijst and Flor for vortex dipoles in a stratified fluid (see **Figures 11 and 12**) (van Heijst 1993). Similarly, tripolar structures have been observed both in lab experiments (see **Figure 13**) and in nature (see **Figure 14**). Recently, the Danish group of Jansson–Haspang–Jensen–Hersen–Bohr has observed beautiful rotating polygons, such as squares and pentagons, on a fluid surface in the presence of a rotating cylinder (see **Figure 15**).

Point Vortex Motion with Boundaries

In comparison with the extensive literature on point vortex motion in unbounded domains, the study of point vortex motion in the presence of walls is modest. There is, however, a general theory for such problems, and some recent new developments in this area have resulted in a versatile tool for analyzing point vortex motion with boundaries. Newton (Newton 2001) contains a chapter on point vortex motion with boundaries and also features a detailed bibliography. The reader is referred there for standard treatments; here, we focus on more recent developments of the mathematical theory.

The Method of Images

When point vortices move around in bounded domains, it is clear that the motion is subject to the constraint that no fluid should penetrate any of

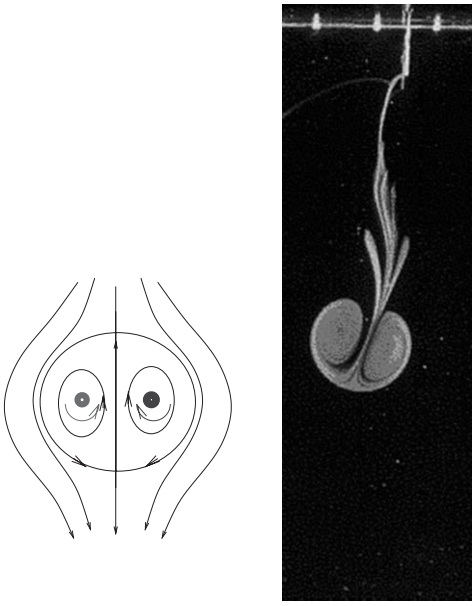


Figure 11 Test-particle trajectories: on the left, theoretical trajectories, from the point-vortex model; on the right, a top view of a laboratory experiment in stratified flows. Reproduced from van Heijst GJF and Flor JB (1989) Dipole formation and collisions in a stratified fluid. *Nature* 340: 212–215, with permission from Nature Publishing Group.

the boundary walls of the domain. If $\boldsymbol{n}$ denotes the local normal to the boundary walls, the boundary condition on the velocity field $\boldsymbol{u}$ is therefore $\boldsymbol{u} \cdot \boldsymbol{n} = 0$ everywhere on the walls. Another way to say the same thing is that all the walls must be streamlines so that the streamfunction, ψ say, must be constant on any boundary wall.

A classical approach to bounded vortex motion is the celebrated method of images – a rather special technique limited to cases where the domain of interest has certain geometrical symmetries so that an appropriate distribution of image vorticity can be ascertained, essentially by inspection. This image vorticity is placed in nonphysical regions of the plane in order to satisfy the boundary conditions that the walls act as impenetrable barriers for the flow.

The simplest example is the motion of a single vortex next to a straight plane wall of infinite extent. Suppose the wall is along $y = 0$ in an (x, y)-plane and that the fluid occupies the upper-half plane. If a circulation-Γ vortex is at the complex position $z_0 = x_0 + \mathrm{i}y_0$, the solution for the streamfunction is

$$\psi(z, \overline{z}) = -\frac{\Gamma}{2\pi} \log \left| \frac{z - z_0}{z - \overline{z_0}} \right| \qquad [24]$$

where $z = x + \mathrm{i}y$. This has a single logarithmic singularity in the upper-half plane at $z = z_0$

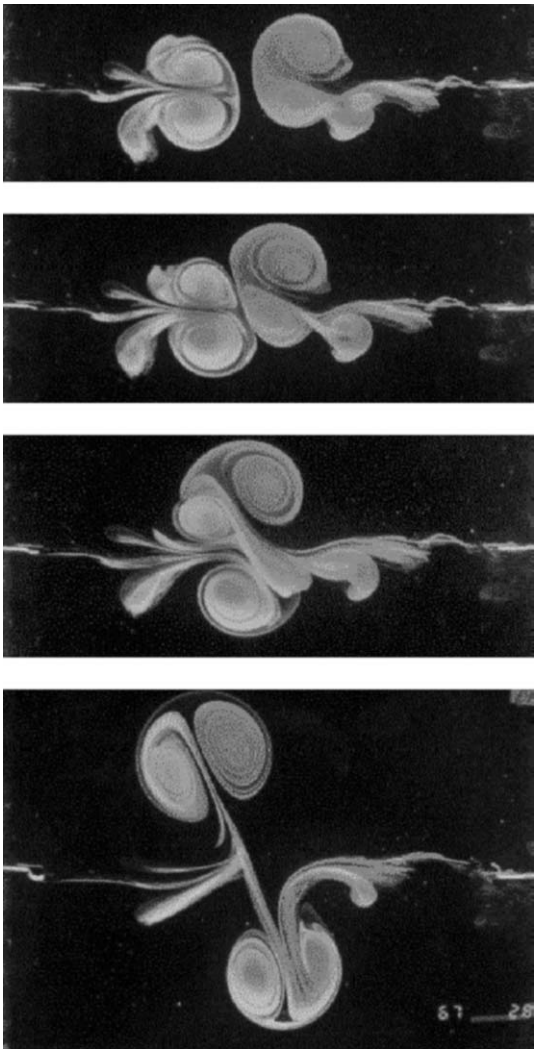


Figure 12 A frontal collision of two dipoles as observed in a stratified fluid: after a so called "partner-exchange" two new dipoles are formed. Reproduced from van Heijst GJF and Flor JB (1989) Dipole formation and collisions in a stratified fluid. *Nature* 340: 212–215, with permission from Nature Publishing Group.

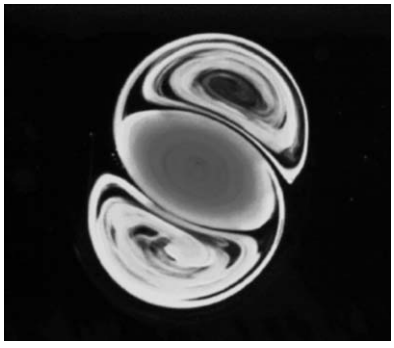


Figure 13 A tripolar vortex structure as observed in a rotating stratified fluid. Reproduced from van Heijst GJF, Kloosterziel RC, and Williams CWM (1991) Laboratory experiments on the tripolar vortex in a rotating fluid. *Journal of Fluid Mechanics* 225: 301–331, with permission from Cambridge University Press.

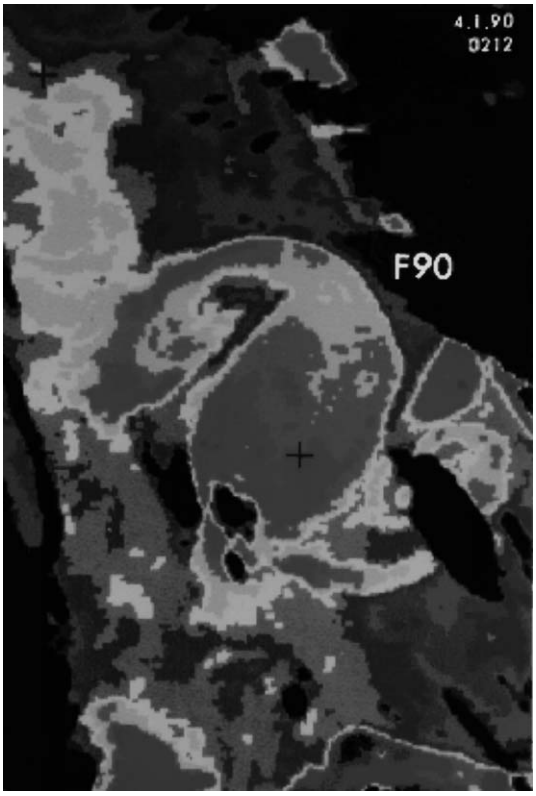


Figure 14 Infrared image taken by NOAA11 satellite on January 4 1990 (0212 UT) shows a tripolar structure in the Bay of Biscay. The central part of the tripole measures about 50–70 km and rotates clockwise, whereas the two satellite vortices rotate anticlockwise. The dipoles persisted for a few days before it fell apart. Reproduced from Pingree RD and Le Cann B, Anticyclonic Eddy X91 in the Southern Bay of Biscay, *Journal of Geophysical Research*, 97: 14353–14362, May 1991 to February 1992. Copyright (1992) American Geophysical Union. Reproduced/modified by permission of American Geophysical Union.

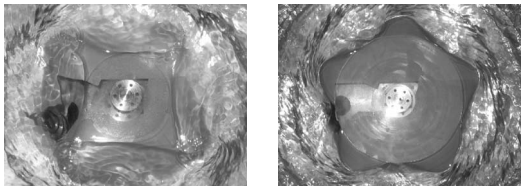


Figure 15 The free surface of a rotating fluid will, due to the centrifugal force, be pressed radially outward. If the flow is driven by rotating the bottom plate, the axial symmetry can break spontaneously and the surface can take the shape of a rigidly rotating polygon. With water Jansson–Haspang–Jensen–Hersen–Bohr have observed polygons with up to six corners. The rotation speed of the polygons does not coincide with that of the plate, but it is often mode-locked, such that the polygon rotates by one corner for each complete rotation of the plate. Reproduced from Jansson TRN, Haspang M, Jensen KH, Hersen P, and Bohr T (2005) Rotating polygons on a fluid surface. *Preprint*, with permission from T Bohr.

(corresponding to the point vortex) and it is easily checked that $\psi = 0$ on $y = 0$. Therefore, no fluid penetrates the wall. Equation [24] can be written as

$$\psi(z,\overline{z}) = -\frac{\Gamma}{2\pi}\log|z - z_0| + \frac{\Gamma}{2\pi}\log|z - \overline{z_0}| \qquad [25]$$

which is the sum of the streamfunction due to a point vortex of circulation Γ at $z_0 = x_0 + iy_0$ and another, one imagines, of circulation $-\Gamma$ at $\overline{z_0} = x_0 - iy_0$. In this case, the image vortex distribution is simple: it is just the second vortex sitting at the reflected point in the wall. The method of images can be applied to flows in other regions bounded by straight line segments (e.g., wedge regions of various angles (Newton 2001)).

A variant of the method of images is the Milne–Thomson circle theorem relevant to planar flow around a circular cylinder. Given a complex potential $w(z)$ with the required singularities in the fluid region exterior to the cylinder, but failing to satisfy the boundary condition that the surface of the cylinder is a streamline, this theorem says that the correct potential $W(z)$ is

$$W(z) = w(z) + \overline{w}(a^2/z) \qquad [26]$$

where a is the cylinder radius and $\overline{w}(z)$ is the conjugate function to $w(z)$. It is easy to verify that the imaginary part of $W(z)$, that is, the streamfunction, is zero on $|z| = a$. The second term, $\overline{w}(a^2/z)$, produces the required distribution of image vorticity inside the cylinder. A famous example is the Föppl vortex pair which is the simplest model of the trailing vortices shed in the wake of a circular aerofoil traveling at uniform speed.

Kirchhoff–Routh–Lin Theory

The most important general mathematical tool for point vortex motion in bounded planar regions is the Hamiltonian approach associated with the names of Kirchhoff (1876) and Routh (1881), who developed the early theory. It is now known that the problem of N-vortex motion in a simply connected domain is a Hamiltonian dynamical system. Moreover, the Hamiltonian has simple transformation properties when a given flow domain of interest is mapped conformally to another – a result originally due to Routh. A formula for the Hamiltonian can be built from knowledge of the instantaneous Green's function associated with motion of the point vortex in the simply connected domain D. In fact, [24] is precisely the relevant Green's function when D is the upper-half plane.

Much later, in 1941, Lin (1941a) extended these general results to the case of multiply connected fluid regions. To visualize such a region, think of a bounded region of the plane containing fluid but also a finite number of impenetrable islands whose boundaries act as barriers for the fluid motion. If the islands are infinitely thin, they can be thought of as straight wall segments immersed in the flow (see later examples). Lin (1941b) showed that both the Hamiltonian structure, and the transformation properties of the Hamiltonian under conformal mapping, are preserved in the multiply connected case.

Lin's Special Green's Function

Since Lin's result subsumes the earlier simply connected studies, we now outline the key results as presented in Lin (1941a). Consider a fluid region D, with outer boundary C_0 and M enclosed islands each having boundaries $\{C_j | j = 1, \ldots, M\}$. Lin introduced a special Green's function $G(x, y; x_0, y_0)$ satisfying the following properties:

1. the function

$$g(x, y; x_0, y_0) = -G(x, y; x_0, y_0) - \frac{1}{2\pi} \log r_0 \qquad [27]$$

is harmonic with respect to (x, y) throughout the region D including at the point (x_0, y_0). Here, $r_0 = \sqrt{(x - x_0)^2 + (y - y_0)^2}$;

2. if $\partial G / \partial n$ is the normal derivative of G on a curve then

$$G(x, y; x_0, y_0) = A_k, \quad \text{on } C_k,\ k = 1, \ldots, M$$
$$\oint_{C_k} \frac{\partial G}{\partial n}\, \mathrm{d}s = 0, \quad k = 1, \ldots, M \qquad [28]$$

where $\mathrm{d}s$ denotes an element of arc and $\{A_k\}$ are constants;

3. $G(x, y; x_0, y_0) = 0$ on C_0.

Flucher and Gustafsson (1997) refer to this G as the hydrodynamic Green's function. (In fact, it coincides with the modified Green's function arising in abstract potential theory – a function that is dual to the usual first-type Green's function that equals zero on all the domain boundaries.) On the use of G, Lin established the following two key results:

Theorem 4 *If N vortices of strengths $\{\Gamma_k | k = 1, \ldots, N\}$ are present in an incompressible fluid at the points $\{(x_k, y_k) | k = 1, \ldots, N\}$ in a general multiply connected region D bounded by fixed boundaries, the stream function of the fluid motion is given by*

$$\psi(x, y; x_k, y_k) = \psi_0(x, y) + \sum_{k=1}^{N} \Gamma_k G(x, y; x_k, y_k) \qquad [29]$$

where $\psi_0(x, y)$ is the streamfunction due to outside agencies and is independent of the point vortex positions.

Theorem 5 *For the motion of vortices of strengths $\{\Gamma_k | k = 1, \ldots, N\}$ in a general region D bounded by fixed boundaries, there exists a Kirchhoff–Routh function $H(\{x_k, y_k\})$, depending on the point vortex positions, such that*

$$\Gamma_k \frac{\mathrm{d}x_k}{\mathrm{d}t} = \frac{\partial H}{\partial y_k}, \qquad \Gamma_k \frac{\mathrm{d}y_k}{\mathrm{d}t} = -\frac{\partial H}{\partial x_k} \qquad [30]$$

where $H(\{x_k, y_k\})$ is given by

$$H(\{x_k, y_k\}) = \sum_{k=1}^{N} \Gamma_k \psi_0(x_k, y_k) + \sum_{\substack{k_1, k_2 = 1 \\ k_1 > k_2}}^{N} \Gamma_{k_1} \Gamma_{k_2} G(x_{k_1}, y_{k_1}; x_{k_2}, y_{k_2}) - \frac{1}{2} \sum_{k=1}^{N} \Gamma_k^2 g(x_k, y_k; x_k, y_k) \qquad [31]$$

In rescaled coordinates $(x_k, \Gamma_k y_k)$, [30] is a Hamiltonian system in canonical form. For historical reasons, H is often called the Kirchhoff–Routh path function. Analyzing the separate contributions to the path function [31] is instructive: the first term is the contribution from flows imposed from outside (e.g., background flows and round-island circulations), the second term is the "free-space" contribution (it is the relevant Hamiltonian when no boundaries are present) while the third term encodes the effect of the boundary walls (or, the effect of the "image vorticity" distribution discussed earlier).

Lin (1941a) went on to show that, with the Hamiltonian in some D given by H in [31], the Hamiltonian relevant to vortex motion in another domain obtained from D by a conformal mapping $z(\zeta)$ consists of [31] with some simple extra additive contributions dependent only on the derivative of the map $z(\zeta)$ evaluated at the point vortex positions.

Flucher and Gustafsson (1997) also introduce the Robin function $\mathcal{R}(x_0, y_0)$ defined as the regular part of the above hydrodynamic Green's function evaluated at the point vortex. Indeed, $\mathcal{R}(x_0, y_0) \equiv g(x_0, y_0; x_0, y_0)$, where g is defined in [27]. An interesting fact is that, for single-vortex motion in a simply connected domain, $\mathcal{R}(x_0, y_0)$ satisfies the quasilinear elliptic Liouville equation everywhere in

D with the boundary condition that it becomes infinite everywhere on the boundary of D.

By combining the Kirchhoff–Routh theory with conformal mapping theory, many interesting problems can be studied. What happens, for example, if there is a gap in the wall of **Figure 16**? In recent work, Johnson and McDonald (2005) show that if the vortex starts off, far from the gap, at a distance of less than half the gap width from the wall, then it will eventually penetrate the gap. Otherwise, it will dip towards the gap but not go through it. The trajectories are shown in **Figure 17**.

Unfortunately, Lin did not provide any explicit analytical expressions for G in the multiply connected case. This has limited the applicability of his theory beyond fluid regions that are anything other than simply and doubly connected. Recently, however, Lin's theory has recently been brought to implementational fruition by Crowdy and Marshall (2005a), who, up to conformal mapping, have derived explicit formulas for the hydrodynamic Green's function in multiply connected fluid regions of arbitrary finite connectivity. Their approach makes use of elements of classical function theory dating back to the work of Poincaré, Schottky, and Klein (among others). This allows new problems involving bounded vortex motion to be tackled. For example, the motion of a single vortex around multiple circular islands has been studied in Crowdy and Marshall (2005b), thereby extending recent work on the two-island problem (Johnson and McDonald 2005). If the wall in **Figure 17** happens to have two (or more) gaps, then the fluid region is multiply connected. The two-gap (doubly connected) case was recently solved by Johnson and McDonald (2005) using Schwarz–Christoffel maps combined with elements of elliptic function theory (see **Figure 18**). Crowdy and Marshall have solved the problem of an arbitrary number of gaps in a wall by exploiting the new general theory presented in Crowdy and Marshall (2005a,b) (and related works by the authors). The case of a wall with three gaps represents a triply connected fluid region and the critical vortex trajectory is plotted in **Figure 19**.

Point vortex motion in bounded domains on the surface of a sphere has received scant attention in

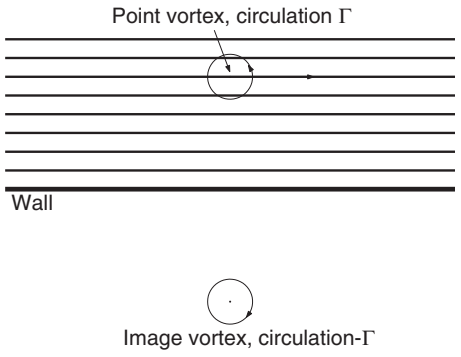


Figure 16 The motion of a point vortex near an infinite straight wall. The vortex moves, at constant speed, maintaining a constant distance from the wall. Other possible trajectories are shown; they are all straight lines parallel to the wall. The motion can be thought of as being induced by an opposite-circulation "image" vortex at the reflected point in the wall.

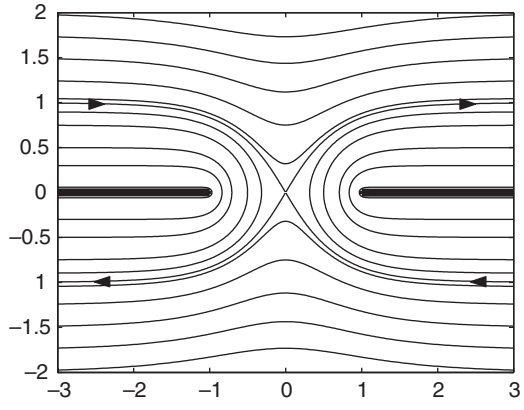


Figure 17 Distribution of point vortex trajectories near a wall with a single gap of length 2. There is a critical trajectory which, far from the gap, is unit distance from the wall.

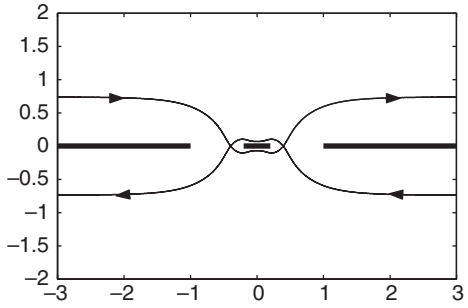


Figure 18 The critical trajectory when there are two symmetric gaps in a wall. The fluid region is now doubly connected. This problem is solved in Johnson and McDonald (2005) and Crowdy and Marshall (2005).

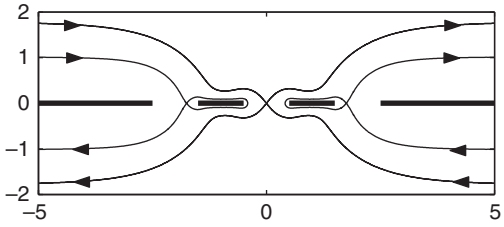


Figure 19 The critical vortex trajectories when there are three gaps in the wall. This time the fluid region is triply connected. This problem is solved in Crowdy and Marshall (2005) using the general methods in Crowdy and Marshall (2005).

the literature, although Kidambi and Newton (2000) and Newton (2001) have recently made a contribution. Such paradigms are clearly relevant to planetary-scale oceanographic flows in which oceanic eddies interact with topography such as ridges and land masses and deserve further study.

Further Reading

Albouy A (1996) The symmetric central configurations of four equal masses. *Contemporary Mathematics* 198: 131–135.

Aref H and Stremler MA (1999) Four-vortex motion with zero total circulation and impulse. *Physics of Fluids* 11(12): 3704–3715.

Aref H, Newton PK, Stremler MA, Tokieda T, and Vainchtein DL (2003) Vortex crystals. *Advances in Applied Mathematics* 39: 1–79.

Ball WWR (1940) *A Short Account of the History of Mathematics*, 12th edn. London: MacMillan.

Boatto S and Pierrehumbert RT (1999) Dynamics of a passive tracer in the velocity field of four identical point-vortices. *Journal of Fluid Mechanics* 394: 137–174.

Boatto S and Cabral HE (2003) Nonlinear stability of a latitudinal ring of point vortices on a non-rotating sphere. *SIAM Journal of Applied Mathematics* 64: 216–230.

Boatto S and Simó C (2004) Stability of latitudinal vortex rings with polar vortices. Mathematical Physics Preprint Archive (mp_arc) 04-67.

Cabral HE and Schmidt DS (1999/00) Stability of relative equilibria in the problem of $N+1$ vortices. *SIAM Journal of Mathematical Analysis* 31(2): 231–250.

Carnevale GF, McWilliams JC, Pomeau Y, Weiss JB, and Young WR (1991) Evolution of vortex statistics in two-dimensional turbulence. *Physical Review Letters* 66(21): 2735–2737.

Carnevale GF, McWilliams JC, Pomeau Y, Weiss JB, and Young WR (1992) Rates, pathways, and end states of nonlinear evolution in decaying two-dimensional. *Physics of Fluids A* 4(6): 1314–1316.

Castilla MSAC, Moauro V, Negrini P, and Oliva WM (1993) The four positive vortices problem – regions of chaotic behavior and non-integrability. *Annales de l'Institut Henri Poincaré. Section A. Physique Theorique* 59(1): 99–115.

Crowdy DG and Marshall JS (2005a) Analytical formulae for the Kirchhoff–Routh path function in multiply connected domains. *Proceedings of the Royal Society A* 461: 2477–2501.

Crowdy DG and Marshall JS (2005b) The motion of a point vortex around multiple circular islands. *Physics of Fluids* 17: 560–602.

Dritschel DG (1985) The stability and energetics of co-rotating uniform vortices. *Journal of Fluid Mechanics* 157: 95–134.

Flucher M and Gustafsson B (1997) *Vortex Motion in Two Dimensional Hydrodynamics*. Royal Institute of Technology Report No. TRITA-MAT-1997-MA-02.

Glass K (2000) Symmetry and bifurcations of planar configurations of the *N*-body and other problems. *Dynamics and Stability of Systems* 15(2): 59–73.

van Heijst GJF and Flor JB (1989) Dipole formation and collisions in a stratified fluid. *Nature* 340: 212–215.

van Heijst GJF, Kloosterziel RC, and Williams CWM (1991) Laboratory experiments on the tripolar vortex in a rotating fluid. *Journal of Fluid Mechanics* 225: 301–331.

van Heijst GJF (1993) Self-organization of two-dimensional flows. *Nederlands Tijdschrift voor Natuurkunde* 59: 321–325 (http://www.fluid.tue.nl).

von Helmholtz H (1858) On the integrals of the hydrodynamical equations which express vortex motion. *Philosophical Magazine* 4(33): 485–512.

Jansson TRN, Haspang M, Jensen KH, Hersen P, and Bohr T (2005) Rotating polygons on a fluid surface. *Preprint*.

Johnson ER and McDonald NR (2005) Vortices near barriers with multiple gaps. *Journal of Fluid Mechanics* 531: 335–358.

Kidambi R and Newton PK (2000) Point vortex motion on a sphere with solid boundaries. *Physical Fluids* 12: 581–588.

Kimura Y (1999) Vortex motion on surfaces with constant curvature. *Proceedings of the Royal Society of London A* 455: 245–259.

Kirchhoff G (1876) *Vorlesunger über mathematische Physik, Mechanik*. Leipzig.

Koiller J and Carvalho SP (1989) Non-integrability of the 4-vortex system: analytical proof. *Communications in Mathematical Physics* 120(4): 643–652.

Kuznetsov L and Zavlasky GM (2000) Passive tracer transport in three-vortex flow. *Physical Review A* 61(4): 3777–3792.

Lim C, Montaldi J, and Roberts M (2001) Relative equilibria of point vortices on the sphere. *Physica D* 148: 97–135.

Lin CC (1941a) On the motion of vortices in two dimensions. I. Existence of the Kirchhoff–Routh function. *Proceedings of the National Academy of Sciences* 27(12): 570–575.

Lin CC (1941b) On the motion of vortices in two dimensions. II. Some further investigations on the Kirchhoff–Routh function. *Proceedings of the National Academy Sciences* 27(12): 575–577.

Marchioro C and Pulvirenti M (1993) Vortices and localization in Euler flows. *Communications in Mathematical Physics* 154: 49–61.

Marchioro C and Pulvirenti M (1994) *Mathematical Theory of Incompressible Non-viscous Fluids*. vol. 96, AMS. New York: Springer.

Mayer AM (1878) Floating magnetics. *Nature* 17: 487–488.

Mayer AM (1878) *Scientific American* 2045–2047.

Mayer AM (1878) On the morphological laws of the configurations formed by magnets floating vertically and subjected to the attraction of a superposed magnetic. *American Journal of Science* 16: 247–256.

Montaldi J, Soulière A, and Tokieda T (2002) Vortex dyanmics on a cylinder. *SIAM Journal of Applied Dynamical Systems* 2(3): 417–430.

Newton PK (2001) *The N-Vortex Problem. Analytical Techniques*. New York: Springer.

Ottino JM (1990) *The Kinematics of Mixing: Stretching, Chaos and Transport*. Cambridge: Cambridge University Press.

Pingree RD and Le Cann B (1992) Anticyclonic Eddy X91 in the Southern Bay of Biscay. *Journal of Geophysical Research* 97: 14353–14362.

Polvani LM and Dritschel DG (1993) Wave and vortex dynamics on the surface of a sphere. *Journal of Fluid Mechanics* 255: 35–64.

Poupaud F (2002) Diagonal defect measures, adhesion dynamics and Euler equation. *Methods and Applications of Analysis* 9(4): 533–562.

Rom-Kedar V, Leonard A, and Wiggins S (1990) An analytical study of transport, mixing and chaos in an unsteady vortical flow. *Journal of Fluid Mechanics* 214: 347–394.

Routh E (1881) Some applications of conjugate functions. *Proceedings of the London Mathematical Society* 12: 73–89.

Sakajo T (2004) Transition of global dynamics of a polygonal vortex ring on a sphere with pole vortices. *Physica D* 196: 243–264.

Schochet S (1995) The weak vorticity formulation of the 2-D Euler equations and concentration-cancellation. *Communications in Partial Differential Equations* 20(5&6): 1077–1104.

Soulière A and Tokieda T (2002) Periodic motion of vortices on surfaces with symmetries. *Journal of Fluid Mechanics* 460: 83–92.

Thomson JJ (1883) *A Treatise on the Motion of Vortex Rings*. New York: Macmillan.

Thomson JJ (1904) *Electricity and Matter*. Westmister Archibald Constable.

Thomson W (1867) On vortex atoms. *Proceedings of the Royal Society of Edinburgh* 6: 94–105.

Yarmchuk EJ, Gordon MJV, and Packard RE (1979) Observation of stationary vortices arrays in rotating superfluid helium. *Physical Review Letters* 43(3): 214–217.

Ziglin SL (1982) Quasi-periodic motions of vortex systems. *Physica D* 4: 261–269 (addendum to K M Khanin).

Poisson Lie Groups *see* Classical *r*-Matrices, Lie Bialgebras, and Poisson Lie Groups

Poisson Reduction

J-P Ortega, Université de Franche-Comté, Besançon, France
T S Ratiu, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland

Introduction

The Poisson reduction techniques allow the construction of new Poisson structures out of a given one by combination of two operations: "restriction" to submanifolds that satisfy certain compatibility assumptions and passage to a "quotient space" where certain degeneracies have been eliminated. For certain kinds of reduction, it is necessary to pass first to a submanifold and then take a quotient. Before making this more explicit, we introduce the notations that will be used in this article. All manifolds in this article are finite dimensional.

Poisson Manifolds

A "Poisson manifold" is a pair $(M, \{\cdot\,,\cdot\})$, where M is a manifold and $\{\cdot\,,\cdot\}$ is a bilinear operation on $C^\infty(M)$ such that $(C^\infty(M), \{\cdot\,,\cdot\})$ is a Lie algebra and $\{\cdot\,,\cdot\}$ is a derivation (i.e., the Leibniz identity holds) in each argument. The pair $(C^\infty(M), \{\cdot\,,\cdot\})$ is also called a "Poisson algebra." The functions in the center $\mathcal{C}(M)$ of the Lie algebra $(C^\infty(M), \{\cdot\,,\cdot\})$ are called "Casimir functions." From the natural isomorphism between derivations on $C^\infty(M)$ and vector fields on M, it follows that each $h \in C^\infty(M)$ induces a vector field on M via the expression $X_h = \{\cdot\,, h\}$, called the "Hamiltonian vector field" associated to the "Hamiltonian function" h. The triplet $(M, \{\cdot\,,\cdot\}, h)$ is called a "Poisson dynamical system." Any Hamiltonian system on a symplectic manifold is a Poisson dynamical system relative to the Poisson bracket induced by the symplectic structure. Given a Poisson dynamical system $(M, \{\cdot\,,\cdot\}, h)$, its "integrals of motion" or "conserved quantities" are defined as the centralizer of h in $(C^\infty(M), \{\cdot\,,\cdot\})$ that is, the subalgebra of $(C^\infty(M), \{\cdot\,,\cdot\})$ consisting of the functions $f \in C^\infty(M)$ such that $\{f, h\} = 0$. Note that the terminology is justified since, by Hamilton's equations in Poisson bracket form, we have $\dot{f} = X_h[f] = \{f, h\} = 0$, that is, f is constant on the flow of X_h. A smooth mapping $\varphi : M_1 \to M_2$, between the two Poisson manifolds $(M_1, \{\cdot\,,\cdot\}_1)$ and $(M_2, \{\cdot\,,\cdot\}_2)$, is called "canonical" or "Poisson" if for all g, $h \in C^\infty(M_2)$ we have $\varphi^*\{g, h\}_2 = \{\varphi^* g, \varphi^* g\}_1$. If $\varphi : M_1 \to M_2$ is a smooth map between two Poisson manifolds $(M_1, \{\cdot\,,\cdot\}_1)$ and $(M_2, \{\cdot\,,\cdot\}_2)$, then φ is a Poisson map if and only if $T\varphi \circ X_{h\circ\varphi} = X_h \circ \varphi$ for any $h \in C^\infty(M_2)$, where $T\varphi : TM_1 \to TM_2$ denotes the tangent map (or derivative) of φ.

Let $(S, \{\cdot\,,\cdot\}^S)$ and $(M, \{\cdot\,,\cdot\}^M)$ be two Poisson manifolds such that $S \subset M$ and the inclusion $i_S : S \hookrightarrow M$ is an immersion. The Poisson manifold $(S, \{\cdot\,,\cdot\}^S)$ is called a "Poisson submanifold" of $(M, \{\cdot\,,\cdot\}^M)$ if i_S is a canonical map. An immersed submanifold Q of M is called a "quasi-Poisson submanifold" of $(M, \{\cdot\,,\cdot\}^M)$ if for any $q \in Q$, any open neighborhood U of q in M, and any $f \in C^\infty(U)$ we have $X_f(i_Q(q)) \in T_q i_Q(T_q Q)$, where $i_Q : Q \hookrightarrow M$ is the inclusion and X_f is the Hamiltonian vector field of f on U with respect to the Poisson bracket of M restricted to U. If $(S, \{\cdot\,,\cdot\}^S)$ is a Poisson submanifold of $(M, \{\cdot\,,\cdot\}^M)$, then there is no other bracket $\{\cdot\,,\cdot\}'$ on S making the inclusion $i : S \hookrightarrow M$ into a canonical map. If Q is a quasi-Poisson submanifold of $(M, \{\cdot\,,\cdot\})$, then there exists a unique Poisson structure $\{\cdot\,,\cdot\}^Q$ on Q that makes it into a Poisson submanifold of $(M, \{\cdot\,,\cdot\})$ but this Poisson structure may be different from the given one on Q. Any Poisson submanifold is quasi-Poisson but the converse is not true in general.

The Poisson Tensor and Symplectic Leaves

The derivation property of the Poisson bracket implies that for any two functions $f, g \in C^\infty(M)$, the value of the bracket $\{f,g\}(z)$ at an arbitrary point $z \in M$ (and therefore $X_f(z)$ as well) depends on f only through $\boldsymbol{d}f(z)$ which allows us to define a contravariant antisymmetric 2-tensor $B \in \Lambda^2(T^*M)$, called the "Poisson tensor," by $B(z)(\alpha_z, \beta_z) = \{f,g\}(z)$, where $\boldsymbol{d}f(z) = \alpha_z \in T_z^*M$ and $\boldsymbol{d}g(z) = \beta_z \in T_z^*M$. The vector bundle map $B^\sharp : T^*M \to TM$ over the identity naturally associated to B is defined by $B(z)(\alpha_z, \beta_z) = \langle \alpha_z, B^\sharp(\beta_z)\rangle$. Its range $D := B^\sharp(T^*M) \subset TM$ is called the "characteristic distribution" of $(M, \{\cdot\,,\cdot\})$ since D is a generalized smooth integrable distribution. Its maximal integral leaves are called the "symplectic leaves" of M for they carry a symplectic structure that makes them into Poisson submanifolds. As integral leaves of an integrable distribution, the symplectic leaves $\mathcal{L}$ are "initial submanifolds" of M, that is, the inclusion $i : \mathcal{L} \hookrightarrow M$ is an injective immersion such that for any smooth manifold P, an arbitrary map $g : P \to \mathcal{L}$ is smooth if and only if $i \circ g : P \to M$ is smooth.

Poisson Reduction

Canonical Lie Group Actions

Let $(M, \{\cdot\,,\cdot\})$ be a Poisson manifold and let G be a Lie group acting canonically on M via the map $\Phi: G \times M \to M$. An action is called "canonical" if for any $h \in G$ and $f, g \in C^\infty(M)$, one has

$$\{f \circ \Phi_h, g \circ \Phi_h\} = \{f,g\} \circ \Phi_h$$

If the G-action is free and proper, then the orbit space M/G is a smooth regular quotient manifold. Moreover, it is also a Poisson manifold with the Poisson bracket $\{\cdot\,,\cdot\}^{M/G}$, uniquely characterized by the relation

$$\{f,g\}^{M/G}(\pi(m)) = \{f \circ \pi, g \circ \pi\}(m) \qquad [1]$$

for any $m \in M$ and where $f, g : M/G \to \mathbb{R}$ are two arbitrary smooth functions. This bracket is appropriate for the reduction of Hamiltonian dynamics in the sense that if $h \in C^\infty(M)^G$ is a G-invariant smooth function on M, then the Hamiltonian flow F_t of X_h commutes with the G-action, so it induces a flow $F_t^{M/G}$ on M/G that is Hamiltonian on $(M/G, \{\cdot\,,\cdot\}^{M/G})$ for the reduced Hamiltonian function $[h] \in C^\infty(M/G)$ defined by $[h] \circ \pi = h$.

If the Poisson manifold $(M, \{\cdot\,,\cdot\})$ is actually symplectic with form ω and the G-action has an associated momentum map $\boldsymbol{J} : M \to \mathfrak{g}^*$, then the symplectic leaves of $(M/G, \{\cdot\,,\cdot\}^{M/G})$ are given by the spaces $(M^c_{\mathcal{O}_\mu} := G \cdot \boldsymbol{J}^{-1}(\mu)^c / G, \omega^c_{\mathcal{O}_\mu})$, where $\boldsymbol{J}^{-1}(\mu)^c$ is a connected component of the fiber $\boldsymbol{J}^{-1}(\mu)$ and $\omega^c_{\mathcal{O}_\mu}$ is the restriction to $M^c_{\mathcal{O}_\mu}$ of the symplectic form $\omega_{\mathcal{O}_\mu}$ of the symplectic orbit reduced space $M_{\mathcal{O}_\mu}$ (*see* Symmetry and Symplectic Reduction). If, additionally, G is compact, M is connected, and the momentum map $\boldsymbol{J}$ is proper, then $M^c_{\mathcal{O}_\mu} = M_{\mathcal{O}_\mu}$.

In the remainder of this section, we characterize the situations in which new Poisson manifolds can be obtained out of a given one by a combination of restriction to a submanifold and passage to the quotient with respect to an equivalence relation that encodes the symmetries of the bracket.

Definition 1 Let $(M, \{\cdot\,,\cdot\})$ be a Poisson manifold and $D \subset TM$ a smooth distribution on M. The distribution D is called "Poisson" or "canonical," if the condition $\boldsymbol{d}f|_D = \boldsymbol{d}g|_D = 0$, for any $f, g \in C^\infty(U)$ and any open subset $U \subset P$, implies that $\boldsymbol{d}\{f,g\}|_D = 0$.

Unless strong regularity assumptions are invoked, the passage to the leaf space of a canonical distribution destroys the smoothness of the quotient topological space. In such situations, the Poisson algebra of functions is too small and the notion of presheaf of Poisson algebras is needed. *See* Singularity and Bifurcation Theory for more information on singularity theory.

Definition 2 Let M be a topological space with a presheaf $\mathcal{F}$ of smooth functions. A presheaf of Poisson algebras on $(M, \mathcal{F})$ is a map $\{\cdot\,,\cdot\}$ that assigns to each open set $U \subset M$ a bilinear operation $\{\cdot\,,\cdot\}_U : \mathcal{F}(U) \times \mathcal{F}(U) \to \mathcal{F}(U)$ such that the pair $(\mathcal{F}(U), \{\cdot\,,\cdot\}_U)$ is a Poisson algebra. A presheaf of Poisson algebras is denoted as a triple $(M, \mathcal{F}, \{\cdot\,,\cdot\})$. The presheaf of Poisson algebras $(M, \mathcal{F}, \{\cdot\,,\cdot\})$ is said to be "nondegenerate" if the following condition holds: if $f \in \mathcal{F}(U)$ is such that $\{f,g\}_{U\cap V} = 0$, for any $g \in \mathcal{F}(V)$ and any open set of V, then f is constant on the connected components of U.

Any Poisson manifold $(M, \{\cdot\,,\cdot\})$ has a natural presheaf of Poisson algebras on its presheaf of smooth functions that associates to any open subset U of M the restriction $\{\cdot\,,\cdot\}|_U$ of $\{\cdot\,,\cdot\}$ to $C^\infty(U) \times C^\infty(U)$.

Definition 3 Let P be a topological space and $\mathcal{Z} = \{S_i\}_{i\in I}$ a locally finite partition of P into smooth manifolds $S_i \subset P, i \in I$, that are locally closed topological subspaces of P (hence their manifold topology is the relative one induced by P). The pair $(P, \mathcal{Z})$ is called a "decomposition" of P with "pieces" in $\mathcal{Z}$, or a "decomposed space," if the following "frontier condition" holds:

Condition (DS) If $R, S \in \mathcal{Z}$ are such that $R \cap \bar{S} \neq \emptyset$, then $R \subset \bar{S}$. In this case, we write $R \preceq S$. If, in addition, $R \neq S$ we say that R is incident to S or that it is a boundary piece of S and write $R \prec S$.

Definition 4 Let M be a differentiable manifold and $S \subset M$ a decomposed subset of M. Let $\{S_i\}_{i\in I}$

be the pieces of this decomposition. The topology of S is not necessarily the relative topology as a subset of M. Then $D \subset TM|_S$ is called a "smooth distribution" on S adapted to the decomposition $\{S_i\}_{i\in I}$, if $D \cap TS_i$ is a smooth distribution on S_i for all $i \in I$. The distribution D is said to be "integrable" if $D \cap TS_i$ is integrable for each $i \in I$.

In the situation described by the previous definition and if D is integrable, the integrability of the distributions $D_{S_i} := D \cap TS_i$ on S_i allows us to partition each S_i into the corresponding maximal integral manifolds. Thus, there is an equivalence relation on S_i whose equivalence classes are precisely these maximal integral manifolds. Doing this on each S_i, we obtain an equivalence relation D_S on the whole set S by taking the union of the different equivalence classes corresponding to all the D_{S_i}. Define the quotient space S/D_S by

$$S/D_S := \bigcup_{i\in I} S_i/D_{S_i}$$

and let $\pi_{D_S}: S \to S/D_S$ be the natural projection.

The Presheaf of Smooth Functions on S/D_S

Define the presheaf of smooth functions C^∞_{S/D_S} on S/D_S as the map that associates to any open subset V of S/D_S the set of functions $C^\infty_{S/D_S}(V)$ characterized by the following property: $f \in C^\infty_{S/D_S}(V)$ if and only if for any $z \in V$ there exists $m \in \pi^{-1}_{D_S}(V), U_m$ open neighborhood of m in M, and $F \in C^\infty(U_m)$ such that

$$f \circ \pi_{D_S}|_{\pi^{-1}_{D_S}(V)\cap U_m} = F|_{\pi^{-1}_{D_S}(V)\cap U_m} \qquad [2]$$

F is called a "local extension" of $f \circ \pi_{D_S}$ at the point $m \in \pi^{-1}_{D_S}(V)$. When the distribution D is trivial, the presheaf C^∞_{S/D_S} coincides with the presheaf of Whitney smooth functions $C^\infty_{S,M}$ on S induced by the smooth functions on M.

The presheaf C^∞_{S/D_S} is said to have the (D, D_S)-local extension property when the topology of S is stronger than the relative topology and, at the same time, the local extensions of $f \circ \pi_{D_S}$ defined in [2] can always be chosen to satisfy

$$dF(n)|_{D(n)} = 0 \quad \text{for any } n \in \pi^{-1}_{D_S}(V) \cap U_m$$

F is called a "local D-invariant extension" of $f \circ \pi_{D_S}$ at the point $m \in \pi^{-1}_{D_S}(V)$. If S is a smooth embedded submanifold of M and D_S is a smooth, integrable, and regular distribution on S, then the presheaf C^∞_{S/D_S} coincides with the presheaf of smooth functions on S/D_S when considered as a regular quotient manifold.

The following definition spells out what we mean by obtaining a bracket via reduction.

Definition 5 Let $(M, \{\cdot,\cdot\})$ be a Poisson manifold, S a decomposed subset of M, and $D \subset TM|_S$ a Poisson-integrable generalized distribution adapted to the decomposition of S. Assume that C^∞_{S/D_S} has the (D, D_S)-local extension property. Then $(M, \{\cdot,\cdot\}, D, S)$ is said to be "Poisson reducible" if $(S/D_S, C^\infty_{S/D_S}, \{\cdot,\cdot\}^{S/D_S})$ is a well-defined presheaf of Poisson algebras where, for any open set $V \subset S/D_S$, the bracket $\{\cdot,\cdot\}^{S/D_S}_V : C^\infty_{S/D_S}(V) \times C^\infty_{S/D_S}(V) \to C^\infty_{S/D_S}(V)$ is given by

$$\{f, g\}^{S/D_S}_V(\pi_{D_S}(m)) := \{F, G\}(m)$$

for any $m \in \pi^{-1}_{D_S}(V)$ for local D-invariant extensions F,G at m of $f \circ \pi_{D_S}$ and $g \circ \pi_{D_S}$, respectively.

Theorem 1 *Let $(M, \{\cdot,\cdot\})$ be a Poisson manifold with associated Poisson tensor $B \in \Lambda^2(T^*M)$, S a decomposed space, and $D \subset TM|_S$ a Poisson-integrable generalized distribution adapted to the decomposition of S (see Definitions 4 and 1). Assume that C^∞_{S/D_S} has the (D, D_S)-local extension property. Then $(M, \{\cdot,\cdot\}, D, S)$ is Poisson reducible if for any $m \in S$*

$$B^\sharp(\Delta_m) \subset \left[\Delta^S_m\right]^\circ \qquad [3]$$

where $\Delta_m := \{dF(m) | F \in C^\infty(U_m), dF(z)|_{D(z)} = 0$, for all $z \in U_m \cap S$, and for any open neighborhood U_m of m in $M\}$ and $\Delta^S_m := \{dF(m) \in \Delta_m | F|_{U_m\cap V_m}$ is constant for an open neighborhood U_m of m in M and an open neighborhood V_m of m in $S\}$.

If S is endowed with the relative topology, then $\Delta^S_m := \{dF(m) \in \Delta_m | F|_{U_m\cap V_m}$ is constant for an open neighborhood U_m of m in $M\}$.

Reduction by Regular Canonical Distributions

Let $(M, \{\cdot,\cdot\})$ be a Poisson manifold and S an embedded submanifold of M. Let $D \subset TM|_S$ be a sub-bundle of the tangent bundle of M restricted to S such that $D_S := D \cap TS$ is a smooth, integrable, regular distribution on S and D is canonical.

Theorem 2 *With the above hypotheses, $(M, \{\cdot,\cdot\}, D, S)$ is Poisson reducible if and only if*

$$B^\sharp(D^\circ) \subset TS + D \qquad [4]$$

Applications of the Poisson Reduction Theorem

Reduction of Coisotropic Submanifolds

Let $(M, \{\cdot,\cdot\})$ be a Poisson manifold with associated Poisson tensor $B \in \Lambda^2(T^*M)$ and S an immersed smooth submanifold of M. Denote by $(TS)^\circ := \{\alpha_s \in T^*_sM | \langle \alpha_s, v_s\rangle = 0$, for all $s \in S$, $v_s \in T_sS\} \subset T^*M$ the

conormal bundle of the manifold S; it is a vector sub-bundle of $T^*M|_S$. The manifold S is called "coisotropic" if $B^\sharp((TS)^\circ) \subset TS$. In the physics literature, coisotropic submanifolds appear sometimes under the name of "first-class constraints." The following are equivalent:

1. S is coisotropic;
2. if $f \in C^\infty(M)$ satisfies $f|_S \equiv 0$, then $X_f|_S \in \mathfrak{X}(S)$;
3. for any $s \in S$, any open neighborhood U_s of s in M, and any function $g \in C^\infty(U_s)$ such that $X_g(s) \in T_sS$, if $f \in C^\infty(U_s)$ satisfies $\{f, g\}(s) = 0$, it follows that $X_f(s) \in T_sS$;
4. the subalgebra $\{f \in C^\infty(M) \mid f|_S \equiv 0\}$ is a Poisson subalgebra of $(C^\infty(M), \{\cdot\,,\cdot\})$.

The following proposition shows how to endow the coisotropic submanifolds of a Poisson manifold with a Poisson structure by using the reduction theorem 1.

Proposition 1 *Let $(M, \{\cdot\,,\cdot\})$ be a Poisson manifold with associated Poisson tensor $B \in \Lambda^2(T^*M)$. Let S be an embedded coisotropic submanifold of M and $D := B^\sharp((TS)^\circ)$. Then*

(i) *$D = D \cap TS = D_S$ is a smooth generalized distribution on S.*
(ii) *D is integrable.*
(iii) *If C^∞_{S/D_S} has the (D, D_S)-local extension property, then $(M, \{\cdot\,,\cdot\}, D, S)$ is Poisson reducible.*

Coisotropic submanifolds usually appear as the level sets of integrals in involution. Let $(M, \{\cdot\,,\cdot\})$ be a Poisson manifold with Poisson tensor B and let $f_1, \ldots, f_k \in C^\infty(M)$ be k smooth functions in involution, that is, $\{f_i, f_j\} = 0$, for any $i, j \in \{1, \ldots, k\}$. Assume that $\mathbf{0} \in \mathbb{R}^k$ is a regular value of the function $F := (f_1, \ldots, f_k) : M \to \mathbb{R}^k$ and let $S := F^{-1}(0)$. Since for any $s \in S$, $\operatorname{span}\{df_1(s), \ldots, df_k(s)\} \subset (T_sS)^\circ$ and the dimensions of both sides of this inclusion are equal, it follows that $\operatorname{span}\{df_1(s), \ldots, df_k(s)\} = (T_sS)^\circ$. Hence, $B^\sharp(s)((T_sS)^\circ) = \operatorname{span}\{X_{f_1}(s), \ldots, X_{f_k}(s)\}$ and $B^\sharp(s)\ ((T_sS)^\circ) \subset T_sS$ by the involutivity of the components of F. Consequently, S is a coisotropic submanifold of $(M, \{\cdot\,,\cdot\})$.

Cosymplectic Submanifolds and Dirac's Constraints Formula

The Poisson reduction theorem 2 allows us to define Poisson structures on certain embedded submanifolds that are not Poisson submanifolds.

Definition 6 Let $(M, \{\cdot\,,\cdot\})$ be a Poisson manifold and let $B \in \Lambda^2(T^*M)$ be the corresponding Poisson tensor. An embedded submanifold $S \subset M$ is called cosymplectic if

(i) $B^\sharp((TS)^\circ) \cap TS = \{0\}$,
(ii) $T_sS + T_s\mathcal{L}_s = T_sM$,

for any $s \in S$ and $\mathcal{L}_s$ the symplectic leaf of $(M, \{\cdot\,,\cdot\})$ containing $s \in S$.

The cosymplectic submanifolds of a symplectic manifold (M, ω) are its symplectic submanifolds. Cosymplectic submanifolds appear in the physics literature under the name of "second-class constraints."

Proposition 2 *Let $(M, \{\cdot\,,\cdot\})$ be a Poisson manifold, $B \in \Lambda^2(T^*M)$ the corresponding Poisson tensor, and S a cosymplectic submanifold of M. then, for any $s \in S$,*

(i) *$T_s\mathcal{L}_s = (T_sS \cap T_s\mathcal{L}_s) \oplus B^\sharp(s)((T_sS)^\circ)$, where $\mathcal{L}_s$ is the symplectic leaf of $(M, \{\cdot\,,\cdot\})$ that contains $s \in S$.*
(ii) *$(T_sS)^\circ \cap \ker B^\sharp(s) = \{0\}$.*
(iii) *$T_sM = B^\sharp(s)((T_sS)^\circ) \oplus T_sS$.*
(iv) *$B^\sharp((TS)^\circ)$ is a sub-bundle of $TM|_S$ and hence $TM|_S = B^\sharp((TS)^\circ) \oplus TS$.*
(v) *The symplectic leaves of $(M, \{\cdot\,,\cdot\})$ intersect S transversely and hence $S \cap \mathcal{L}$ is an initial submanifold of S, for any symplectic leaf $\mathcal{L}$ of $(M, \{\cdot\,,\cdot\})$.*

Theorem 3 (The Poisson structure of a cosymplectic submanifold). *Let $(M, \{\cdot\,,\cdot\})$ be a Poisson manifold, $B \in \Lambda^2(T^*M)$ the corresponding Poisson tensor, and S a cosymplectic submanifold of M. Let $D := B^\sharp\ ((TS)^\circ) \subset TM|_S$. Then,*

(i) *$(M, \{\cdot\,,\cdot\}, D, S)$ is Poisson reducible.*
(ii) *The corresponding quotient manifold equals S and the reduced bracket $\{\cdot\,,\cdot\}^S$ is given by*

$$\{f, g\}^S(s) = \{F, G\}(s) \qquad [5]$$

where $f, g \in C^\infty_{S,M}(V)$ are arbitrary and $F, G \in C^\infty(U)$ are local D-invariant extensions of f and g around $s \in S$, respectively.
(iii) *The Hamiltonian vector field X_f of an arbitrary function $f \in C^\infty_{S,M}(V)$ is given either by*

$$Ti \circ X_f = X_F \circ i \qquad [6]$$

where $F \in C^\infty(U)$ is a local D-invariant extension of f and $i : S \hookrightarrow M$ is the inclusion, or by

$$Ti \circ X_f = \pi_S \circ X_{\overline{F}} \circ i \qquad [7]$$

where $\overline{F} \in C^\infty(U)$ is an arbitrary local extension of f and $\pi_S : TM|_S \to TS$ is the projection induced by the Whitney sum decomposition $TM|_S = B^\sharp((TS)^\circ) \oplus TS$ of $TM|_S$.
(iv) *The symplectic leaves of $(S, \{\cdot\,,\cdot\}^S)$ are the connected components of the intersections $S \cap \mathcal{L}$, where $\mathcal{L}$ is a symplectic leaf of $(M, \{\cdot\,,\cdot\})$. Any*

symplectic leaf of $(S, \{\cdot,\cdot\}^S)$ *is a symplectic submanifold of the symplectic leaf of* $(M,\{\cdot,\cdot\})$ *that contains it.*

(v) *Let* $\mathcal{L}_s$ *and* $\mathcal{L}_s^S$ *be the symplectic leaves of* $(M, \{\cdot,\cdot\})$ *and* $(S, \{\cdot,\cdot\}^S)$, *respectively, that contain the point* $s \in S$. *Let* $\omega_{\mathcal{L}_s}$ *and* $\omega_{\mathcal{L}_s^S}$ *be the corresponding symplectic forms. Then* $B^\sharp(s)((T_sS)^\circ)$ *is a symplectic subspace of* $T_s\mathcal{L}_s$ *and*

$$B^\sharp(s)((T_sS)^\circ) = \left(T_s\mathcal{L}_s^S\right)^{\omega_{\mathcal{L}_s}(s)} \qquad [8]$$

where $(T_s\mathcal{L}_s^S)^{\omega_{\mathcal{L}_s}(s)}$ *denotes the* $\omega_{\mathcal{L}_s}(s)$*-orthogonal complement of* $T_s\mathcal{L}_s^S$ *in* $T_s\mathcal{L}_s$.

(vi) *Let* $B_S \in \Lambda^2(T^*S)$ *be the Poisson tensor associated to* $(S, \{\cdot,\cdot\}^S)$. *Then*

$$B_S^\sharp = \pi_S \circ B^\sharp|_S \circ \pi_S^* \qquad [9]$$

where $\pi_S^* : T^*S \to T^*M|_S$ *is the dual of* $\pi_S : TM|_S \to TS$.

The "Dirac constraints formula" is the expression in coordinates for the bracket of a cosymplectic submanifold. Let $(M, \{\cdot,\cdot\})$ be an n-dimensional Poisson manifold and let S be a k-dimensional cosymplectic submanifold of M. Let z_0 be an arbitrary point in S and $(U, \overline{\kappa})$ a submanifold chart around z_0 such that $\overline{\kappa} = (\overline{\varphi}, \overline{\psi}) : U \to V_1 \times V_2$, where V_1 and V_2 are two open neighborhoods of the origin in two Euclidean spaces such that $\overline{\kappa}(z_0) = (\overline{\varphi}(z_0), \overline{\psi}(z_0)) = (0,0)$ and

$$\overline{\kappa}(U \cap S) = V_1 \times \{0\} \qquad [10]$$

Let $\overline{\varphi} =: (\overline{\varphi}^1, \ldots, \overline{\varphi}^k)$ be the components of $\overline{\varphi}$ and define $\widehat{\varphi}^1 := \overline{\varphi}^1|_{U\cap S}, \ldots, \widehat{\varphi}^k := \overline{\varphi}^k|_{U\cap S}$. Extend $\widehat{\varphi}^1, \ldots, \widehat{\varphi}^k$ to D-invariant functions $\varphi^1, \ldots, \varphi^k$ on U. Since the differentials $\boldsymbol{d}\widehat{\varphi}^1(s), \ldots, \boldsymbol{d}\widehat{\varphi}^k(s)$ are linearly independent for any $s \in U \cap S$, we can assume (by shrinking U if necessary) that $\boldsymbol{d}\varphi^1(z), \ldots, \boldsymbol{d}\varphi^k(z)$ are also linearly independent for any $z \in U$. Consequently, (U,κ) with $\kappa := (\varphi^1, \ldots, \varphi^k, \psi^1, \ldots, \psi^{n-k})$ is a submanifold chart for M around z_0 with respect to S such that, by construction,

$$\boldsymbol{d}\varphi^1(s)|_{B^\#(s)((T_sS)^\circ)} = \cdots = \boldsymbol{d}\varphi^k(s)|_{B^\#(s)((T_sS)^\circ)} = 0$$

for any $s \in U \cap S$. This implies that for any $i \in \{1, \ldots, k\}$, $j \in \{1, \ldots, n-k\}$, and $s \in S$

$$\{\varphi^i, \psi^j\}(s) = \boldsymbol{d}\varphi^i(s)\left(X_{\psi^j}(s)\right) = 0$$

since $\boldsymbol{d}\psi^j(s) \in (T_sS)^\circ$ by [10] and hence

$$X_{\psi^j}(s) \in B^\#(s)((T_sS)^\circ) \qquad [11]$$

Additionally, since the functions $\varphi^1, \ldots, \varphi^k$ are D-invariant, by [6], it follows that

$$X_{\varphi^1}(s) = X_{\widehat{\varphi}^1}(s) \in T_sS, \ldots, X_{\varphi^k}(s) = X_{\widehat{\varphi}^k}(s) \in T_sS$$

for any $s \in S$. Consequently, $\{X_{\varphi^1}(s), \ldots, X_{\varphi^k}(s), X_{\psi^1}(s), \ldots, X_{\psi^{n-k}}(s)\}$ spans $T_s\mathcal{L}_s$ with

$$\{X_{\varphi^1}(s), \ldots, X_{\varphi^k}(s)\} \subset T_sS \cap T_s\mathcal{L}_s$$

and

$$\{X_{\psi^1}(s), \ldots, X_{\psi^{n-k}}(s)\} \subset B^\#(s)((T_sS)^\circ)$$

By Proposition 2(i),

$$\operatorname{span}\{X_{\varphi^1}(s), \ldots, X_{\varphi^k}(s)\} = T_sS \cap T_s\mathcal{L}_s$$

and

$$\operatorname{span}\{X_{\psi^1}(s), \ldots, X_{\psi^{n-k}}(s)\} = B^\#(s)((T_sS)^\circ)$$

Since $\dim(B^\#(s)((T_sS)^\circ)) = n-k$ by Proposition 2(iii), it follows that $\{X_{\psi^1}(s), \ldots, X_{\psi^{n-k}}(s)\}$ is a basis of $B^\#(s)((T_sS)^\circ)$.

Since $B^\#(s)((T_sS)^\circ)$ is a symplectic subspace of $T_s\mathcal{L}_s$ by Theorem 3(v), there exists some $r \in \mathbb{N}$ such that $n-k = 2r$ and, additionally, the matrix $C(s)$ with entries

$$C^{ij}(s) := \{\psi^i, \psi^j\}(s), \quad i,j \in \{1, \ldots, n-k\}$$

is invertible. Therefore, in the coordinates $(\varphi^1, \ldots, \varphi^k, \psi^1, \ldots, \psi^{n-k})$, the matrix associated to the Poisson tensor $B(s)$ is

$$B(s) = \begin{pmatrix} B_S(s) & 0 \\ 0 & C(s) \end{pmatrix}$$

where $B_S \in \Lambda^2(T^*S)$ is the Poisson tensor associated to $(S, \{\cdot,\cdot\}^S)$. Let $C_{ij}(s)$ be the entries of the matrix $C(s)^{-1}$.

Proposition 3 (Dirac formulas). *In the coordinate neighborhood* $(\varphi^1, \ldots, \varphi^k, \psi^1, \ldots, \psi^{n-k})$ *constructed above and for* $s \in S$ *we have, for any* $f, g \in C^\infty_{S,M}(V)$:

$$X_f(s) = X_F(s) - \sum_{i,j=1}^{n-k} \{F, \psi^i\}(s) C_{ij}(s) X_{\psi^j}(s) \qquad [12]$$

and

$$\{f,g\}^S(s) = \{F,G\}(s) - \sum_{i,j=1}^{n-k} \{F, \psi^i\}(s) C_{ij}(s) \{\psi^j, G\}(s) \qquad [13]$$

where $F, G \in C^\infty(U)$ *are arbitrary local extensions of* f *and* g, *respectively, around* $s \in S$.

See also: Classical *r*-Matrices, Lie Bialgebras, and Poisson Lie Groups; Cotangent Bundle Reduction; Graded Poisson Algebras; Symmetry and Symplectic Reduction; Hamiltonian Group Actions; Lie, Symplectic, and Poisson Groupoids and their Lie Algebroids; Singularity and Bifurcation Theory.

Further Reading

Abraham R and Marsden JE (1978) *Foundations of Mechanics*, 2nd edn. Reading, MA: Addison–Wesley.

Casati P and Pedroni M (1992) Drinfeld–Sokolov reduction on a simple Lie algebra from the bi-Hamiltonian point of view. *Letters in Mathematical Physics* 25(2): 89–101.

Castrillón-López M and Marsden JE (2003) Some remarks on Lagrangian and Poisson reduction for field theories. *Journal of Geometry and Physics* 48: 52–83.

Cendra H, Marsden JE, and Ratiu TS (2003) Cocycles, compatibility, and Poisson brackets for complex fluids. In: Capriz G and Mariano P (eds.) *Advances in Multifield Theories of Continua with Substructures*, Memoirs, pp. 51–73. Aarhus: Aarhus Univ.

Faybusovich L (1991) Hamiltonian structure of dynamical systems which solve linear programming problems. *Physica D* 53: 217–232.

Faybusovich L (1995) A Hamiltonian structure for generalized affine-scaling vector fields. *Journal of Nonlinear Science* 5(1): 11–28.

Gotay MJ, Nester MJ, and Hinds G (1978) Presymplectic manifolds and the Dirac–Bergmann theory of constraints. *Journal of Mathematical Physics* 19: 2388–2399.

Krishnaprasad PS and Marsden JE (1987) Hamiltonian structure and stability for rigid bodies with flexible attachments. *Archives for Rational and Mechanical Analysis* 98: 137–158.

Lewis D, Marsden JE, Montgomery R, and Ratiu TS (1986) The Hamiltonian structure for dynamic free boundary problems. *Physica D* 18: 391–404.

Lu J-H and Weinstein A (1990) Poisson Lie groups, dressing transformations and Bruhat decompositions. *Journal of Differential Geometry* 31: 510–526.

Marsden JE and Ratiu TS (1986) Reduction of Poisson manifolds. *Letters in Mathematical Physics* 11: 161–169.

Marsden JE and Ratiu TS (2003) *Introduction to Mechanics and Symmetry*, second printing; 1st edn. (1994), Texts in Applied Mathematics, 2nd edn., vol. 17. New York: Springer.

Ortega J-P and Ratiu TS (1998) Singular reduction of Poisson manifolds. *Letters in Mathematical Physics* 46: 359–372.

Ortega J-P and Ratiu TS (2003) *Momentum Maps and Hamiltonian Reduction*. Progress in Math. vol. 222. Boston: Birkhäuser.

Pedroni M (1995) Equivalence of the Drinfeld–Sokolov reduction to a bi-Hamiltonian reduction. *Letters in Mathematical Physics*, 35(4): 291–302.

Sundermeyer K (1982) *Constrained Dynamics*. Lecture Notes in Physics, vol. 169. New York: Springer.

Weinstein A (1983) The local structure of Poisson manifolds. *Journal of Differential Geometry* 18: 523–557.

Weinstein A (1985) The local structure of Poisson manifolds – errata and addenda. *Journal of Differential Geometry* 22(2): 255.

Zaalani N (1999) Phase space reduction and Poisson structure. *Journal of Mathematical Physics* 40: 3431–3438.

Polygonal Billiards

S Tabachnikov, Pennsylvania State University, University Park, PA, USA

Mechanical Examples. Unfolding Billiard Trajectories

The billiard system inside a polygon P has a very simple description: a point moves rectilinearly with the unit speed until it hits a side of P; there it instantaneously changes its velocity according to the rule "the angle of incidence equals the angle of reflection," and continues the rectilinear motion. If the point hits a corner, its further motion is not defined. (*see* Billiards in Bounded Convex Domains). From the point of view of the theory of dynamical systems, polygonal billiards provide an example of parabolic dynamics in which nearby trajectories diverge with subexponential rate.

One of the motivations for the study of polygonal billiards comes from the mechanics of elastic particles in dimension 1. For example, consider the system of two point-masses m_1 and m_2 on the positive half-line $x \geq 0$. The collision between the points is elastic, that is, the energy and momentum are conserved. The reflection off the left endpoint of the half-line is also elastic: if a point hits the "wall" $x = 0$, its velocity changes sign. The configuration space of this system is the wedge $0 \leq x_1 \leq x_2$. After the rescaling $\bar{x}_i = \sqrt{m_i}x_i$, $i = 1, 2$, this system identifies with the billiard inside a wedge with the angle measure $\arctan\sqrt{m_1/m_2}$.

Likewise, the system of two elastic point-masses on a segment is the billiard system in a right triangle; a system of a number of elastic point-masses on the positive half-line or a segment is the billiard inside a multidimensional polyhedral cone or a polyhedron, respectively. The system of three elastic point-masses on a circle has three degrees of freedom; one can reduce one by assuming that the center of mass of the system is fixed. The resulting two-dimensional system is the billiard inside an acute triangle with the angles

$$\arctan\left(m_i\sqrt{\frac{m_1+m_2+m_3}{m_1m_2m_3}}\right), \quad i = 1, 2, 3$$

For comparison, the more realistic system of elastic balls identifies with the billiard system in a domain with nonflat boundary components.

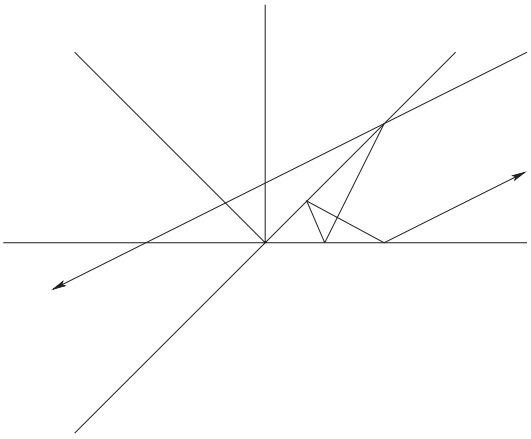


Figure 1 Unfolding a billiard trajectory in a wedge.

A useful elementary method of study is unfolding: instead of reflecting the billiard trajectory in the sides of the polygon, reflect the polygon in the respective side and unfold the billiard trajectory to a straight line. This method yields an upper bound

$$\left[\frac{\pi}{\arctan\sqrt{m_1/m_2}}\right]$$

for the number of collisions in the system of two point-masses m_1 and m_2 on the positive half-line. Likewise, the number of collisions for any number of elastic point-masses on the positive half-line is bounded above by a constant depending on the masses only. Similar results are known for systems of elastic balls (Figure 1).

Similarly, one studies the billiard inside the unit square. Unfolding the square yields a square grid in the plane, acted upon by the group of parallel translations $2\mathbf{Z} \oplus 2\mathbf{Z}$. Factorizing by this group action yields a torus, and the billiard flow in a given direction becomes a constant flow on the torus. If the slope is rational, then all orbits are periodic, and if the slope is irrational, then all orbits are dense and the billiard flow is ergodic. Its metric entropy is equal to zero. Periodic trajectories of the billiard in a square come in bands of parallel ones. Let $f(\ell)$ be the number of such bands of length not greater than ℓ. Then, $f(\ell)$ equals the number of coprime lattice points inside the circle of radius ℓ, that is, $f(\ell)$ has quadratic growth in ℓ.

Periodic Trajectories

The simplest example of a periodic orbit in a polygonal billiard is the 3-periodic Fangano trajectory in an acute triangle: it connects the bases of the three altitudes of the triangle and has minimal perimeter among inscribed triangles. The Fagnano trajectory belongs to a band of 6-periodic ones. It is not known whether every acute triangle has other periodic trajectories.

For a right triangle, one has the following result: almost every (in the sense of the Lebesgue measure) billiard trajectory that leaves a leg in the perpendicular direction returns to the same leg in the same direction and is therefore periodic. A similar existence result holds for polygons whose sides have only two directions.

In general, not much is known about the existence of periodic billiard trajectories in polygons. Conjecturally, every polygon has one, but this is not known even for all obtuse triangles. Recently, R Schwartz proved that every obtuse triangle with the angles not exceeding 100° has a periodic billiard path. This work substantially relies on a computer program, McBilliards, written by Schwartz and Hooper.

If an arbitrary small perturbation of the vertices of a billiard polygon leads to a perturbation of a periodic billiard trajectory, but not to its destruction, then this trajectory is called stable. Label the sides of the polygon $1, 2, \ldots, k$. Then a periodic trajectory is coded by the word consisting of the labels of the consecutively visited sides. An even-periodic trajectory is stable if and only if the numbers in the respective word can be partitioned in pairs of equal numbers, so that the number from each pair appears once at an even position, and once at an odd one. As a consequence, if the angles of a polygon are independent over the rational numbers, then every periodic billiard trajectory in it is stable.

Complexity of Billiard Trajectories

The encoding of billiard trajectories by the consecutively visited sides of the billiard polygon provides a link between billiard and symbolic dynamics. For a billiard k-gon P, denote by Σ the set of words in letters $1, 2, \ldots, k$ corresponding to billiard trajectories in P, and let Σ_n be the set of such words of length n.

One has a general theorem: *the topological entropy of the billiard flow is zero*. This implies that a number of quantities, associated with a polygonal billiard, grow slower than exponentially, as functions of n: the cardinality $|\Sigma_n|$, the number of strips of n-periodic trajectories, the number of generalized diagonals with n links (i.e., billiard trajectories that start and end at corners of the billiard polygon), etc. Conjecturally, all these quantities have polynomial growth in n.

The complexity of the billiard in a polygon is defined as the function $p(n) = |\Sigma_n|$. Likewise, one may consider the billiard trajectories in a given direction θ and define the corresponding complexity $p_\theta(n)$.

In the case of a square, one modifies the encoding using only two symbols, say, 0 and 1, to indicate that a trajectory reflects in a horizontal or a vertical side, respectively. If θ is a direction with an irrational slope, then $p_\theta(n) = n + 1$. This is a classical result by Hedlund and Morse. The sequences with complexity $p(n) = n + 1$ are called Sturmian; this is the smallest complexity of aperiodic sequences. A generalization for multidimensional cubes and parallelepipeds, due to Yu Baryshnikov, is known.

For a k-gon P, let N be the least common denominator of its π-rational angles and s be the number of its distinct π-irrational angles. Then,

$$p_\theta(n) \le kNn\left(1 + \frac{n}{2}\right)^s$$

Concerning billiard trajectories in all directions, one has a lower bound for complexity: $p(n) \ge cn^2$ for a constant c depending on the polygon. A similar estimate holds for a d-dimensional polyhedron with the exponent 2 replaced by d.

Rational Polygons and Flat Surfaces

The only class of polygons for which the billiard dynamics is well understood are rational one, the polygons satisfying the property that the angles between all pairs of sides are rational multiples of π.

Let P be a simply connected (without holes) rational k-gon with angles $\pi m_i/n_i$, where m_i and n_i are coprime integers. The reflections in the sides of P generate a subgroup of the group of isometries of the plane. Let $G(P) \subset \mathrm{O}(2)$ consist of the linear parts of the elements of this group. Then, $G(P)$ is the dihedral group D_N consisting of $2N$ elements. When a billiard trajectory reflects in a side of P, its direction changes by the action of the group $G(P)$, and the orbit of a generic direction $\theta \ne k\pi/N$ on the unit circle consists of $2N$ points.

The phase space of the billiard flow is the unit tangent bundle $P \times S^1$. Let M_θ be the subset of points whose projection to S^1 belongs to the orbit of θ under $G(P) = D_N$. Then, M_θ is an invariant surface of the billiard flow in P. The surface M_θ is obtained from $2N$ copies of P by gluing their sides according to the action of D_N. This oriented compact surface depends only on the polygon P, but not on the choice of θ, and may be denoted by M. The directional billiard flows F_θ on M in directions θ are obtained, one from another, by rotations. The genus of M is given by the formula

$$1 + \frac{N}{2}\left(k - 2 - \sum \frac{1}{n_i}\right)$$

For example, if P is a right triangle with an acute angle $\pi/8$, then M is a surface of genus 2 (Figure 2).

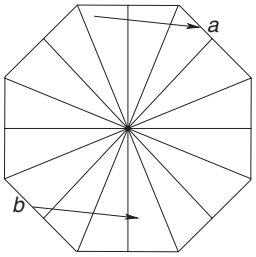


Figure 2 The invariant surface for a right triangle with acute angle $\pi/8$ has genus 2.

The cases when M is a torus are as follows: the angles of P are all of the form π/n_i, where n_i are equal, up to permutations, to

$$(3,3,3), (2,4,4), (2,3,6), (2,2,2,2)$$

and the respective polygons are an equilateral triangle, an isosceles right triangle, a right triangle with an acute angle $\pi/6$, and a square. All these polygons tile the plane.

The billiard flow on the surface M has saddle singularities at the points obtained from the vertices of P. The surface M inherits a flat metric from P with a finite number of cone-type singularities, corresponding to the vertices of P, with cone angles multiples of 2π (Figure 3).

A flat surface M is a compact smooth surface with a distinguished finite set of points Σ. On $M \setminus \Sigma$, one has coordinate charts $v = (x, y)$ such that the transition functions on the overlaps are of the form

$$v \to v + c \quad \text{or} \quad v \to -v + c$$

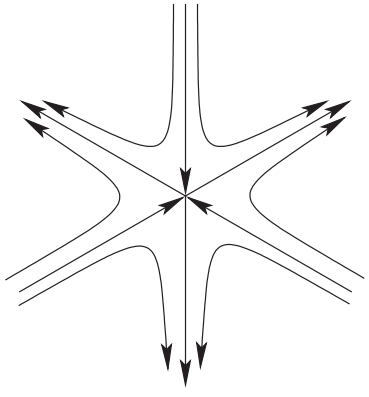


Figure 3 A cone singularity for the flow on an invariant surface.

In particular, one may talk about directions on a flat surface.

The group $\mathrm{PSL}(2, \boldsymbol{R})$ acts on the space of flat structures. From the point of view of complex analysis, a flat surface is a Riemann surface with a holomorphic quadratic differential; the set of cone points Σ corresponds to the zeros of the quadratic differential. Not every flat surface is associated with a polygonal billiard.

Concerning ergodicity, one has the theorem of Kerckhoff, Masur, and Smillie: given a flat surface of genus not less than 2, for almost all directions θ (in the sense of the Lebesgue measure), the flow F_θ is uniquely ergodic. Furthermore, the Hausdorff dimension of the set of angles θ for which ergodicity fails does not exceed 1/2, and this bound is sharp. As a consequence, the billiard flow on the invariant surface is uniquely ergodic for almost all directions. Another corollary: *there is a dense G_δ subset in the space of polygons consisting of polygons for which the billiard flow is ergodic. If a billiard polygon admits approximation by rational polygons at a superexponentially fast rate, then the billiard flow in it is ergodic.*

Concerning periodic orbits, one has the following theorem due to H Masur: given a flat surface of genus not less than 2, there exists a dense set of angles θ such that F_θ has a closed trajectory. As a consequence, for any rational billiard polygon, there is a dense set of directions each with a periodic orbit. Furthermore, periodic points are dense in the phase space of the billiard flow in a rational polygon.

Similarly to the case of a square, let $f(\ell)$ be the number of strips of periodic trajectories of length not greater than ℓ in a rational polygon P. By a theorem of H Masur, there exist constants c and C such that for sufficiently large ℓ one has: $c\ell^2 < f(\ell) < C\ell^2$, and likewise for flat surfaces.

There is a class of flat surfaces, called Veech (or lattice) surfaces, for which more refined results are available. The groups of affine transformations of a flat surface determine a subgroup in $SL(2, \boldsymbol{R})$. If this subgroup is a lattice in $SL(2, \boldsymbol{R})$, then the flat surface is called a Veech surface. Similarly, one defines a Veech rational polygon. For example, regular polygons and isosceles triangles with equal angles π/n are Veech. All acute Veech triangles are described.

For a Veech surface, one has the following Veech dichotomy: for any direction θ, either the flow F_θ is minimal or its every leaf is closed (unless it is a saddle connection, i.e., a segment connecting cone points). For a Veech surface (and polygon), the quadratic bounds for the counting function $f(\ell)$ become quadratic asymptotics: $f(\ell)/\ell^2$ has a limit as $\ell \to \infty$. The value of this limit is expressed in arithmetical terms.

A generic flat surface also has quadratic asymptotics. The value of the limit depends only on the stratum of the Teichmuller space that contains this surface. These values are known, due to Eskin, Masur, Okunkov, and Zorich. Since a generic flat surface does not correspond to a rational polygon, this result does not immediately apply to polygonal billiards. However, quadratic asymptotics are established for rectangular billiards with barriers.

Note, in conclusion, a close relation of billiards in rational polygons and interval exchange transformations; the reduction of the former to the latter is a particular case of the reduction of the billiard flow to the billiard ball map. On an invariant surface M of the billiard flow, consider a segment I, perpendicular to the directional flow. Since "the width of a beam" is an invariant transversal measure for the constant flow, the first return map to I is a piecewise orientation preserving isometry, that is, an interval exchange transformation.

Acknowledgment

This work was partially supported by NSF.

See also: Billiards in Bounded Convex Domains; Ergodic Theory; Fractal Dimensions in Dynamics; Generic Properties of Dynamical Systems; Holomorphic Dynamics; Hyperbolic Billiards; Riemann Surfaces.

Further Reading

Burago D, Ferleger S, and Kononenko A (2000) *A Geometric Approach to Semi-Dispersing Billiards. Hard Ball Systems and the Lorentz Gas*, pp. 9–27. Berlin: Springer.

Chernov N and Markarian R *Theory of Chaotic Billiards* (to appear).

Galperin G, Stepin A, and Vorobets Ya (1992) Periodic billiard trajectories in polygons: generating mechanisms. *Russian Mathematical Surveys* 47(3): 5–80.

Gutkin E (1986) Billiards in polygons. *Physica D* 19: 311–333.

Gutkin E (1996) Billiards in polygons: survey of recent results. *Journal of Statistical Physics* 83: 7–26.

Gutkin E (2003) Billiard dynamics: a survey with the emphasis on open problems. *Regular and Chaotic Dynamics* 8: 1–13.

Katok A and Hasselblatt B (1995) *Introduction to the Modern Theory of Dynamical Systems*. Cambridge: Cambridge University Press.

Kozlov V and Treshchev D (1991) *Billiards. A Genetic Introduction to the Dynamics of Systems with Impacts*. Providence: American Mathematical Society.

Masur H and Tabachnikov S (2002) *Rational Billiards and Flat Structures. Handbook of Dynamical Systems*, vol. 1A, pp. 1015–1089. Amsterdam: North-Holland.

Sinai Ya (1976) *Introduction to Ergodic Theory*. Princeton: Princeton University Press.

Smillie J (2000) *The Dynamics of Billiard Flows in Rational Polygons*, Encyclopaedia of Mathematical Sciences, vol. 100, pp. 360–382. Berlin: Springer.

Tabachnikov S (1995) *Billiards, Société Math.* de France, Panoramas et Syntheses, No 1.

Tabachnikov S (2005) *Geometry and Billiards*. American Mathematical Society.

Positive Maps on C*-Algebras

F Cipriani, Politecnico di Milano, Milan, Italy

Introduction

The theme of positive maps on *-algebras and other ordered vector spaces, dates back to the Perron–Frobenius theory of matrices with positive entries, the Shur's product of matrices, the study of doubly stochastic matrices describing discrete-time random walks and the behavior of limits of powers of positive matrices in ergodic theory.

A long experience proved that far-reaching generalizations of the above situations have to be considered in various fields of mathematical physics and that C^*-algebras, their positive cones, and other associated ordered vector spaces provide a rich unifying framework of functional analysis to treat them.

It is the scope of this note to review some of the basic aspects both of the general theory and of the applications.

In the next section we briefly recall the definitions of C^*-algebras and their positive cones. However, throughout this article we refer to C^*-Algebras and their Classification and von Neumann Algebras: Introduction, Modular Theory and Classification Theory as sources of the definitions and general properties of the objects of these operator algebras. We then introduce positive maps, illustrate their general properties, and discuss some relevant classes of them. The correspondence between states and representations is described next, as well as the appearance of vector, normal and non-normal states in applications. We then illustrate the structure of completely positive maps and their relevance in mathematical physics. Finally, we describe the relevance of the class of completely positive maps to understand the structure of nuclear C^*-algebras.

Positive Cones in C^*-Algebras

A C^*-algebra A is a complex Banach algebra with a conjugate-linear involution $a \mapsto a^*$ such that $\|a^*a\| = \|a\|^2$ for all $a \in A$.

When A has a unit 1_A, the spectrum $\mathrm{Sp}(a)$ of an element a is the subset of all complex numbers λ such that $a - \lambda \cdot 1_A$ is not invertible in A. When A is realized as a subalgebra of some $\mathcal{B}(\mathcal{H})$, and this is always possible, the set $\mathrm{Sp}(a)$ coincides with the spectrum of the bounded operator a on the Hilbert space $\mathcal{H}$.

The involution determines the self-adjoint part $A_h := \{a \in A\colon a = a^*\}$ of A, a real subspace such that $A = A_h + \mathrm{i}A_h$. A self-adjoint element a of A satisfies $\mathrm{Sp}(a) \subseteq \mathbb{R}$ and, if $k \geq 0$, one has $\|a\| \leq k$ if and only if $\mathrm{Sp}(a) \subseteq [-k, k]$.

The involution determines another important subset of A: $A_+ := \{a^*a\colon a \in A\}$. This subset of A_h is closed in the norm topology of A and contains the sums of its elements as well as their multiples by positive scalars: in other words, it is a closed convex cone. From a spectral point of view, one has the following characterization: a self-adjoint element a belongs to A_+ if and only if its spectrum is positive $\mathrm{Sp}(a) \subset [0, +\infty)$. It is this property that allows us to call A_+ the positive cone of A and its elements positive. If it exists, a unit 1_A in A is always positive and a Hermitian element a is positive if and only if $\|1_A - a/\|a\|\| \leq 1$.

The continuous functional calculus in A allows to write any self-adjoint element of A_h as a difference of elements of A_+: $A_h = A_+ - A_+$. Moreover, $A_+ \cap (-A_+) = \{0\}$ and the decomposition $a = b - c$ of a self-adjoint element a as difference of positive elements b and c is unique provided one requires that $bc = cb = 0$. In this case, it is called the orthogonal decomposition.

The cone A_+ determines an underlying structure of order space on A: for $a, b \in A$ one says that a is less than or equal to b, in symbols $a \leq b$, if and only if $b - a \in A_+$. In particular, $a \geq 0$ just means that a is positive.

Another fundamental characterization of the positive cone is the following: a self-adjoint element $a = a^*$ is positive if and only if there exists an element b in A such that $a = b^2$. Moreover, among the elements b with this property, there exists one and only one which is positive, the square root of a. Some examples of positive cones are provided in the following.

Example 1 By a fundamental result of I M Gelfand, a commutative C^*-algebra A is isomorphic to the C^*-algebra $C_0(X)$ of all complex continuous functions vanishing at infinity on a locally compact Hausdorff topological space X. The algebraic operations have the usual pointwise meaning and the norm is the uniform one. The constant function 1 represents the unit precisely when X is compact. The positive cone $C_0(X)_+$ coincides with that of the positive continuous functions in $C_0(X)$.

Example 2 Finite dimensional C^*-algebras A are classified as finite sums $M_{n_1}(\mathbb{C}) \oplus M_{n_2}(\mathbb{C}) \oplus \cdots \oplus M_{n_k}(\mathbb{C})$ of full matrix algebras $M_{n_i}(\mathbb{C})$. An element

$a_1 \oplus a_2 \oplus \cdots \oplus a_k$ is positive if and only if the matrices a_i have positive eigenvalues.

Example 3 When a C^*-algebra $A \subseteq \mathcal{B}(\mathcal{H})$ is represented as a self-adjoint closed algebra of operators on a Hilbert space $\mathcal{H}$, its positive elements are those which have non-negative spectrum.

Positive Maps on C^*-Algebras

Among the various relevant classes of maps between C^*-algebras, we are going to consider the following ones, whose properties are connected with the underlying structures of ordered vector spaces.

Definition 1 Given two C^*-algebras A and B, a map $\phi: A \longrightarrow B$ is called positive if $\phi(A_+) \subseteq B_+$. In other words, a map is positive if and only if it transforms the positive elements of A into positive elements of B:

$$a \in A \Rightarrow \phi(a^*a) \in B_+ \qquad [1]$$

If A and B have units, the map is called unital provided $\phi(1_A) = 1_B$.

Morphisms and Jordan Morphisms

A *-morphism between C^*-algebras $\phi: A \longrightarrow B$ is positive; in fact, $\phi(a^*a) = \phi(a)^*\phi(a) \geq 0$.

This also the case for Jordan *-morphism, the linear maps satisfying $\phi(a^*) = \phi(a)^*$ and $\phi(\{a,b\}) = \{\phi(a), \phi(b)\}$, where $\{a,b\} = ab + ba$ denotes the Jordan product. In fact, if $a = a^*$ then $\phi(a^2) = \phi(a)^2$ is positive.

Shur's Product of Matrices

Let $A \in M_n(\mathbb{C})$ be a positive matrix and define a linear map $\phi: M_n(\mathbb{C}) \longrightarrow M_n(\mathbb{C})$ through the Shur's product of matrices: $\phi_A(B) := [A_{ij}B_{ij}]_{i,j=1}^n$. Since the Shur's product of positive matrices is positive too (i.e., the positive cone of $M_n(\mathbb{C})$ is a semigroup under matrix product), the above map is positive.

Positive-Definite Function on Groups

Positive maps also arise naturally in harmonic analysis. Let G be a locally compact topological group with identity e and left Haar's measure m. Let $p: G \longrightarrow \mathbb{C}$ be a continuous positive-definite function on G. This just means that for all $n \geq 1$ and all $s_1, \ldots, s_n \in G$, the matrix $\{p(s_i^{-1}s_j)\}_{i,j=1}^n$ belongs to the positive cone of $M_n(\mathbb{C})$: $\sum_{i,j=1}^n p(s_i^{-1}s_j)\overline{\alpha}_i\alpha_j \geq 0$ for all $\alpha_1, \ldots, \alpha_n$. Such functions are necessarily bounded with $\|p\|_\infty \leq p(e)$, so that an operator $\phi: L^1(G,m) \longrightarrow L^1(G,m)$ is well defined by pointwise multiplication: $\phi(f)(s) := p(s)f(s)$. This map extends to a positive map $\phi: C^*(G) \longrightarrow C^*(G)$, which is unital when $p(e) = 1$, on the full group C^*-algebra $C^*(G)$. When G is amenable, this algebra coincides with reduced C^*-algebra $C_r(G)$ so that, if G is also unimodular (as is the case if G is compact), the positive elements can be approximated by positive-definite functions in $L^1(G,m)$ and the positivity of ϕ follows exactly as in the previous example.

Positive Maps in Commutative C^*-Algebras

Positive maps $\phi: C_0(Y) \longrightarrow C_0(X)$ between commutative C^*-algebras have the following structure: $\phi(a)(x) = \int_Y k(x, \mathrm{d}y)a(y), a \in C_0(Y)$. Here the kernel $x \mapsto k(x, \cdot)$ is a continuous map from X to the space of positive Radon measures on Y. In case X and Y are compact, the map is unital provided $k(x, \cdot)$ is a probability measure for each $x \in X$. In fact, for a fixed $x \in X$, the map $a \mapsto \phi(a)(x)$ is a positive linear functional from $C_0(Y)$ to $\mathbb{C}$ and Riesz's theorem guarantees that it can be represented by a positive Radon measure on Y.

In probability theory, one-parameter semigroups $\phi_t \circ \phi_s = \phi_{t+s}$ of positive maps $\phi_t: C_0(X) \longrightarrow C_0(X)$ such that $\phi_t(1) \leq 1$ for all $t \geq 0$, are called Markovian semigroups (conservative, if the maps are unital). They represent the expectation at time $t > 0$ of Markovian stochastic processes on X. In this case, the time-dependent kernel $k(t, x, \cdot)$ represents the distribution probability at time t of a particle starting in $x \in X$ at time $t = 0$.

These kinds of maps arise also in potential theory, where the dependence of the solution $\phi(a)$ of a Dirichlet problem on a bounded domain Ω, with nice boundary $\partial\Omega$, upon the continuous boundary data $a \in C(\partial\Omega)$ gives rise to a linear unital map $\phi: C(\partial\Omega) \longrightarrow C(\Omega \cup \partial\Omega)$, whose positivity and unitality translates the "maximum principle" for harmonic functions. When Ω is the unit disk, k is the familiar Poisson's kernel.

Continuity and Algebraic Properties of Positive Maps

Since the order structure of a C^*-algebra A is defined by its positive cone A_+, positive maps are

1. real: $\phi(a^*) = \phi(a)^*$ and
2. order preserving: $\phi(a) \leq \phi(b)$ whenever $a \leq b$.

From this follows an important interplay between positivity and continuity:

> a positive map $\phi: A \longrightarrow B$
> between C^*-algebras is continuous

In case A has a unit, this follows by the fact that ϕ is order preserving and that, for self-adjoint a, one has

$-\|a\|1_A \le a \le +\|a\|1_A$, so that $-\|a\|\phi(1_A) \le \phi(a) \le +\|a\|\phi(1_A)$ and then $\|\phi(a)\| \le \|\phi(1_A)\| \cdot \|a\|$. In general, splitting $a = b + ic$ as a combination of Hermitian elements b and c, as $\|b\| \le \|a\|$ and $\|c\| \le \|a\|$, one obtains

$$\begin{aligned}\|\phi(a)\| &\le \|\phi(b)\| + \|\phi(c)\| \\ &\le \|\phi(1_A)\|(\|b\| + \|c\|) \\ &\le 2\|\phi(1_A)\| \cdot \|a\|\end{aligned}$$

The second general result concerning positivity and continuity is the following:

Let $\phi : A \longrightarrow B$ be a linear map between C*-algebras with unit such that $\phi(1_A) = 1_B$; then ϕ is positive if and only if $\|\phi\| = 1$.

The result relies, among other things, on the generalized Schwarz inequality for unital positive maps on normal elements,

$$\phi(a^*)\phi(a) \le \phi(a^*a), \quad a^*a = aa^*$$

These results may be used to reveal the strong interplay between the algebraic, continuity and positivity properties of maps:

Let $\phi : A \longrightarrow B$ be an invertible linear map between unital C*-algebras such that $\phi(1_A) = 1_B$. The following properties are equivalent:

1. ϕ is Jordan isomorphism,
2. ϕ is an isometry, and
3. ϕ is an order isomorphism (ϕ and ϕ^{-1} are order preserving).

The above conclusions can be strengthened if, instead of individual maps, continuous groups of maps are considered.

Let $t \mapsto \alpha_t$ be a strongly continuous, one-parameter group of maps of a unital C*-algebra A and assume that $\alpha_t(1_A) = 1_A$ for all $t \in \mathbb{R}$. The following properties are equivalent:

1. α_t is a *-automorphism of A for all $t \in \mathbb{R}$,
2. $\|\alpha_t\| \le 1$ for all $t \in \mathbb{R}$, and
3. α_t is positive for all $t \in \mathbb{R}$.

An analogous result holds true for w^*-continuous groups on abelian, or factors, von Neumann algebras.

States on C*-Algebras

A state on a C*-algebra A is a positive functional $\phi : A \longrightarrow \mathbb{C}$ of norm 1:

- $\phi(a^*a) \ge 0$ for all $a \in A$, and
- $\|\phi\| = 1$.

As $\mathbb{C}$ is a C*-algebra, when A is unital, a state on it is just a unital positive map:

- $\phi(a^*a) \ge 0$ for all $a \in A$, and
- $\phi(1_A) = 1$.

States for which $\phi(ab) = \phi(ba)$ are called tracial states.

States constitute a distinguished class of positive maps, both from a mathematical viewpoint and for application to mathematical physics. We will see below that states are deeply connected to representations of C*-algebras (*see* C*-Algebras and their Classification).

States on Commutative C*-Algebras

Since this is a subcase of positive maps in commutative C*-algebras we only add a comment. As far as a C*-algebra represents observable quantities of a physical system, states carry our actual knowledge about the system itself. The smallest C*-sub-algebra $\{f(a): f \in C_0(\mathbb{R})\}$ of A containing a given self-adjoint element $a \in A$, representing a certain observable quantity, is isomorphic to the algebra $C(\mathrm{Sp}(a))$ of continuous functions on the spectrum of a. A state on A induces, by restriction, a state on $C(\mathrm{Sp}(a))$, which, by the Riesz representation theorem, is associated to a probability measure μ_a on $\mathrm{Sp}(a)$ through the formula

$$\phi(f(a)) = \int_{\mathrm{Sp}(a)} f(x)\mu_a(\mathrm{d}x)$$

Since $\mathrm{Sp}(a)$ represents the possible values of the observable associated to a, μ_a represents the distribution of these values when the physical state of the system is represented by ϕ.

Vector States and Density Matrices

In case A is acting on a Hilbert space h, $A \subseteq \mathcal{B}(h)$, each unit vector $\xi \in h$ gives rise to a vector state $\phi_\xi(a) = (\xi | a\,\xi)$. In the quantum-mechanical description of a finite system, as far as observables with discrete spectrum are concerned, one can assume A to be the C*-algebra $\mathcal{K}(h)$ of compact operators on the Hilbert space h. In this case every state is a convex superposition of vector states, in the sense that it can be represented by the formula

$$\phi(a) = \mathrm{tr}(\rho a)/\mathrm{tr}(\rho), \quad a \in \mathcal{K}(h)$$

for a suitable density matrix ρ, that is, a positive, compact operator with finite trace. In quantum statistical mechanics, the grand canonical Gibbs equilibrium state of a finite system at inverse temperature β and chemical potential μ, with Hamiltonian H and number operator N, is of the above type

$$\phi_{\beta,\mu}(a) = \mathrm{tr}(\mathrm{e}^{-\beta K}a)/\mathrm{tr}(\mathrm{e}^{-\beta K})$$

where $K = H - \mu N$, and the spectrum of H is assumed to be discrete and such that $e^{-\beta K}$ is trace-class. For infinite systems, A is a quasilocal C^*-algebra generated by a net $\{A_\alpha\}_\alpha$ of C^*-subalgebras describing observables referred to finite-volume regions. Infinite-volume equilibrium states on A can then be obtained as thermodynamic limits of finite-volume Gibbs equilibrium states of the above type.

Normal and Singular States

When observables with continuous spectrum have to be considered and one chooses the algebra $\mathcal{B}(h)$ of all bounded operators, the above formula, although still meaningful, does not describe all states on $\mathcal{B}(h)$ but only the important subclass of the normal ones. To this class, which can be considered on any von Neumann algebra $\mathcal{M}$, belong states ϕ which are σ-weakly continuous functionals. Equivalently, these are the states such that for all increasing net $a_\alpha \in \mathcal{M}_+$ with least upper bound $a \in \mathcal{M}_+$, $\phi(a)$ is least upper bound of the net $\phi(a_\alpha)$.

In general, each state ϕ on a von Neumann algebra $\mathcal{M}$ splits as a sum of a maximal normal piece and a singular one. Singular traces appear in noncommutative geometry as very useful tools to get back local objects from spectral ones via the familiar principle that local properties of functions depend on the asymptotics of their Fourier coefficients.

This is best illustrated on a compact, Riemannian n-manifold M by the formula

$$\int_M f \, \mathrm{d}m = c_n \cdot \tau_\omega(M_f |D|^{-n})$$

which expresses the Riemannian integral of a nice function f in terms of the Dirac operator D acting on the Hilbert space of square-integrable spinors, the multiplication operator M_f by f, and the singular Dixmier tracial state τ_ω on $\mathcal{B}(H)$. Here the compactness of M implies the compactness of the operator $M_f|D|^{-n}$ and τ_ω is a limiting procedure depending only on the asymptotic behavior of the eigenvalues of $M_f|D|^{-n}$. Similar formulas are valid on self-similar fractals as well as on quasiconformal manifolds. Local index formulas represent cyclic cocycles in Connes' spectral geometry (*see* Noncommutative Geometry and the Standard Model; Noncommutative Geometry from Strings; Path-Integrals in Noncommutative Geometry).

States and Representations: The GNS Construction

A fundamental tool in studying a C^*-algebra A are its representations. These are morphisms of C^*-algebras $\pi : A \longrightarrow \mathcal{B}(\mathcal{H})$ from A to the algebra of all bounded operators on some Hilbert space $\mathcal{H}$. There is a symbiotic appearance of states and representations on C^*-algebras. In fact, given a representation $\pi : A \longrightarrow \mathcal{B}(\mathcal{H})$, one easily constructs states on A by unit vectors $\xi \in H$ by

$$\phi_\xi(a) = (\xi | \pi(a) \xi)$$

In fact, one checks that $\phi_\xi(a^*a) = (\xi|\pi(a^*a)\,\xi) = (\xi|\pi(a^*)\pi(a)\,\xi) = \|\pi(a)\,\xi\|^2 \geq 0$ and, at least if a unit exists, that $\phi_\xi(1_A) = \|\xi\|^2 = 1$.

A fundamental construction due to Gelfand, Naimark, and Segal allows to associate a representation to each state in such a way that each state is a vector state for a suitable representation.

"Let ω be a state over the C^*-algebra A. It follows that there exists cyclic representation $(\pi_\omega, \mathcal{H}_\omega, \xi_\omega)$ of A such that

$$\omega(a) = (\xi_\omega | \pi_\omega(a) \xi_\omega)$$

Moreover, the representation is unique up to unitary equivalence. It is called the canonical cyclic representation of A associated with ω."

The positivity property of the state allows to introduce the positive-semidefinite scalar product $\langle a|b\rangle = \omega(a^*b)$ on the vector space A. Moreover, its kernel $\mathcal{I}_\omega = \{a \in A : \omega(a^*a) = 0\}$ is a left-ideal of A: in fact, if $a \in A$ and $b \in \mathcal{I}_\omega$ then $\omega((ba)^*(ba)) \leq \|a\|^2 \omega(b^*b) = 0$. This allows to define, on the quotient pre-Hilbert space $A/\mathcal{I}_\omega$, an action of the elements $a \in A$: $\pi_\omega(a)(b + \mathcal{I}_\omega) := ab + \mathcal{I}_\omega$. It is the extension of this action to the Hilbert space completion $\mathcal{H}_\omega$ of $A/\mathcal{I}_\omega$ that gives the representation associated to ω. When A has a unit, the cyclic vector ξ_ω with the stated properties is precisely the image of $1_A + \mathcal{I}_\omega$. By definition, the cyclicity of the representation amounts to check that $\pi_\omega(A)\xi_\omega$ is dense in $\mathcal{H}_\omega$.

Completely Positive Maps

In a sense, the order structure of a C^*-algebra A is better understood through the sequence of C^*-algebras $A \otimes M_n(\mathbb{C}) \cong M_n(A)$, obtained as tensor products of A and full matrix algebras $M_n(\mathbb{C})$. For example, C^*-algebras are matrix-ordered vector spaces as $\alpha^*(M_m(A))_+\alpha \subseteq (M_n(A))_+$ for all matrices $\alpha \in M_{m\times n}(\mathbb{C})$.

In this respect, one is naturally led to consider stronger notion of positivity:

"A map $\phi : A \longrightarrow B$ is called n-positive if its extension

$$\phi \otimes 1_n : A \otimes M_n(\mathbb{C}) \longrightarrow B \otimes M_n(\mathbb{C})$$
$$(\phi \otimes 1_n)[a_{i,j}]_{i,j} = [\phi(a_{i,j})]_{i,j}$$

is positive and completely positive (CP map for short) if this happens for all n."

Equivalently, n-positive means that $\sum_{i,j=1}^{n} b_i^* \phi \times (a_i^* a_j) b_j \geq 0$ for all $a_1, \ldots, a_n \in A$ and $b_1, \ldots, b_n \in b$. In particular, if ϕ is n-positive then it is k-positive for all $k \leq n$. Many positive maps we considered are in fact CP maps:

1. morphisms of C^*-algebras are CP maps;
2. positive maps $\phi : A \longrightarrow B$ are automatically CP maps provided A, B or both are commutative and states are, in particular, CP maps; and
3. an important class of CP maps is the following. A norm one projection $\varepsilon : A \longrightarrow B$, from a C^*-algebra A onto a C^*-subalgebra B, is a contraction such that $\varepsilon(b) = b$ for all $b \in B$. It can be proved that these maps satisfy $\varepsilon(bac) = b\varepsilon(a)c$ for all $a \in A$ and $b, c \in B$ and for this reason they are called conditional expectations. This property then implies that they are CP maps.

However, the identity map from a C^*-algebra A into its opposite A° is positive but not 2-positive unless A is commutative, the transposition $a \mapsto a^t$ in $M_n(\mathbb{C})$ is positive and not 2-positive if $n \geq 2$ and, for all n, there exist n-positive maps which are not $(n+1)$-positive.

CP Maps in Mathematical Physics

In several fields of application, the transition of a state of a system into another state can be described by a completely positive map $\phi : A \longrightarrow B$ between C^*-algebras: for any given state ω of B, $\omega \circ \phi$ is then a state of A.

1. In the theory of quantum communication processes (*see* Channels in Quantum Information Theory; Optimal Cloning of Quantum States; Source Coding in Quantum Information Theory; Capacity for Quantum Information), for example, B and A represent the input and output systems, respectively, ω the signal to be transmitted, $\omega \circ \phi$ the received signal, and ϕ the system of transmission, called the channel.
2. In quantum probability and in the theory of quantum open systems, continuous semigroups of CP maps (*see* Quantum Dynamical Semigroups) describe dissipative time evolutions of a system due to interaction with an external one (heat bath).
3. In the theory of measurement in quantum mechanics, an observable can be described by a positive-operator-valued (POV) measure M which assigns a positive element $m(E)$ in a C^*-algebra A to each Borel subset E of a topological space X. For each $a \in C_0(X)$, one can define its integral $\phi(f) := \int_X f \, \mathrm{d}E$ as an element of A. The map $\phi : C_0(X) \longrightarrow A$, called the observation channel, is then a CP map.
4. Another field of mathematical physics in which CP maps play a distinguished role is in the construction and application of the quantum dynamical entropy, an extension of the Kolmogorov–Sinai entropy of measure preserving transformations (*see* Quantum Entropy). When dealing with a noncommutative dynamical system $(\mathcal{M}, \alpha, \tau)$ in which τ is a normal trace state on a finite von Neumann algebra $\mathcal{M}$, the Connes–Størmer entropy $h_\tau(\alpha)$ is defined through the consideration of an entropy functional $H_\tau(N_1, \ldots, N_k)$ of finite-dimensional von Neumann subalgebras $N_1, \ldots, N_k \subset \mathcal{M}$. To extend the definition to more general C^*-algebras and states on them, one has to face the fact that C^*-algebras may have no nontrivial C^*-subalgebras. To circumvent the problem A Connes, H Narnhofer, and W Thirring (CNT) introduced an entropy functional $H(\gamma_1, \ldots, \gamma_k)$ associated to a set $\gamma_i : A_i \longrightarrow A$ of CP maps (finite channels) from finite-dimensional C^*-algebras A_i into A. This led to the CNT entropy $h_\omega(\alpha)$ of a noncommutative dynamical system (A, α, ω), where ω is a state on A and α is an automorphism or a CP map preserving it: $\omega \circ \alpha = \omega$.

CP Maps and Continuity

Since for an element $a \in A$ of a unital C^*-algebra, one has $\|a\| \leq 1$ precisely when

$$\begin{pmatrix} 1 & a \\ a^* & 1 \end{pmatrix}$$

is positive in $M_2(A)$, it follows that

2-positive unital maps are contractive

Unital 2-positive maps satisfy, in particular, the generalized Schwarz inequality for all $a \in A$,

$$\phi(a^*)\phi(a) \leq \phi(a^* a)$$

In particular,

"CP maps are completely bounded as $\sup_n \|\phi \otimes 1_n\| = \|\phi(1_A)\|$ and completely contractive if they are unital. Conversely unital, completely contractive maps are CP maps."

CP Maps and Matrix Algebras

When the domain or the target space of a map are matrix algebras, one has the following equivalences concerning positivity. Let $[e_{i,j}]_{i,j}$ denote the standard

matrix units in $M_n(\mathbb{C})$ and $\phi : M_n(\mathbb{C}) \longrightarrow B$ into a C^*-algebra B. The following conditions are equivalent:

1. ϕ is a CP map,
2. ϕ is n-positive, and
3. $[\phi(e_{i,j})]_{i,j}$ is positive in $M_n(B)$.

Associating to a linear map $\phi : A \longrightarrow M_n(\mathbb{C})$, the linear functional $s_\phi : M_n(A) \longrightarrow \mathbb{C}$ by $s_\phi([a_{i,j}]) := \sum_{i,j} \phi(a_{i,j})_{i,j}$, one has the following equivalent properties:

1. ϕ is a CP map,
2. ϕ is n-positive,
3. s_ϕ is positive, and
4. s_ϕ is positive on $A_+ \otimes M_n(\mathbb{C})_+$.

Stinspring Representation of CP Maps

CP maps are relatively easy to handle, thanks to the following dilation result due to W F Stinspring. It describes a CP map as the compression of a morphism of C^*-algebras.

Let A be a unital C^*-algebra and $\phi : A \longrightarrow \mathcal{B}(\mathcal{H})$ a linear map. Then ϕ is a CP map if and only if it has the form

$$\phi(a) = V^* \pi(a) V$$

for some representation $\pi : A \longrightarrow \mathcal{B}(\mathcal{K})$ on a Hilbert space $\mathcal{K}$, and some bounded linear map $V : \mathcal{H} \longrightarrow \mathcal{K}$. If A is a von Neumann algebra and ϕ is normal then π can be taken to be normal. When $A = \mathcal{B}(\mathcal{H})$ and $\mathcal{H}$ is separable, one has, for some $b_n \in \mathcal{B}(\mathcal{H})$,

$$\phi(a) = \sum_{n=1}^{\infty} b_n^* a b_n$$

The proof of this result is reminiscent of the GNS construction for states and its extension, by G Kasparov, to C^*-modules is central in bivariant K-homology theory.

Despite the above satisfactory result, one should be aware that positive but not CP maps are much less understood and only for maps on very low dimensional matrix algebras do we have a definitive classification. To have an idea of the intricacies of the matter, one may consult Størmer (1963).

Positive Semigroups on Standard Forms of von Neumann Algebras and Ground State for Physical Hamiltonians

The above result allows one to derive the structure of generators of norm-continuous dynamical semigroups in terms of dissipative operators.

Strongly continuous positive semigroups, which are KMS symmetric with respect to a KMS state ω of a given automorphism group of a C^*-algebra A, can be analyzed as positive semigroups in the standard representation $(\mathcal{M}, \mathcal{H}, \mathcal{P}, J)$ (*see* Tomita–Takesaki Modular Theory) of the von Neumann algebra $\mathcal{M} := \pi_\omega(A)''$. A semigroup on A gives rise to a corresponding w^*-continuous positive semigroup on $\mathcal{M}$ and to a strongly continuous positive semigroup on the ordered Hilbert space $(\mathcal{H}, \mathcal{P})$ of the standard form. In the latter framework, one can develop an infinite-dimensional, noncommutative extension of the classical Perron–Frobenius theory for matrices with positive entries. This applies, in particular, to semigroups generated by physical Hamiltonians and has been used to prove existence and uniqueness of the ground state for bosons and fermions systems in quantum field theory (one may consult Gross (1972)).

Nuclear C^*-Algebras and Injective von Neumann Algebras

The nonabelian character of the product in C^*-algebras may prevent the existence of nontrivial morphisms between them, while one may have an abundance of CP maps. For example, there are no nontrivial morphisms from the algebra of compact operators to $\mathbb{C}$, but there exist sufficiently many states to separate its elements. A much more well-behaved category of C^*-algebras is obtained by considering CP maps as morphisms. This is true, in particular, for nuclear C^*-algebras: those for which any tensor product $A \otimes B$ with any other C^*-algebra B admits a unique C^*-cross norm (*see* C^*-Algebras and their Classification). The intimate relation between this class of algebras and CP maps is illustrated by the following characterization:

1. A is nuclear;
2. the identity map of A is a pointwise limit of CP maps of finite rank;
3. the identity map of A can be approximately factorized, $\lim_\alpha (T_\alpha \circ S_\alpha)a \to a$ for all $a \in A$, through matrix algebras and nets of CP maps $S_\alpha : A \to M_n(\mathbb{C}), T_\alpha : M_n(\mathbb{C}) \to A$.

A second important relation between nuclear C^*-algebras and CP maps emerges in connection to the lifting problem.

"Let A be a nuclear C^*-algebra and J a closed two-sided ideal in a C^*-algebra B. Then every CP map $\phi : A \to B/J$ can be lifted to a CP map $\phi' : A \to B$. In other words, ϕ factors through B by the quotient map $q : B \to B/J$: $\phi = q \circ \phi$."

This and related results are used to prove that the Brown–Douglas–Fillmore K-homology invariant $\mathrm{Ext}(A)$ is a group for separable, nuclear C^*-algebras.

Our last basic result, due to W Arveson, about CP maps concerns the extension problem.

> "Let A be a unital C^*-algebra and N a self-adjoint closed subspace of A containing the identity. Then every CP map $\phi: N \to \mathcal{B}(\mathcal{H})$ from N into a type I factor $\mathcal{B}(\mathcal{H})$ can be extended to a CP map $\phi: A \to \mathcal{B}(\mathcal{H})$."

This result can be restated by saying that type I factors are injective von Neumann algebras. It may suggest how the notion of a completely positive map plays a fundamental role along Connes' proof of one culminating result of the theory of von Neumann algebras, namely the fact that the class of injective von Neumann algebras coincides with the class of approximately finite-dimensional ones (*see* von Neumann Algebras: Introduction, Modular Theory and Classification Theory).

See also: Capacity for Quantum Information; C^*-Algebras and Their Classification; Channels in Quantum Information Theory; Noncommutative Geometry and the Standard Model; Noncommutative Geometry from Strings; Optimal Cloning of Quantum States; Path Integrals in Noncommutative Geometry; Quantum Dynamical Semigroups; Quantum Entropy; Source Coding in Quantum Information Theory; Tomita–Takesaki Modular Theory; von Neumann Algebras: Introduction, Modular Theory, and Classification Theory.

Further Reading

Bratteli O and Robinson DW (1987) *Operator Algebras and Quantum Statistical Mechanics 1*, 2nd edn., 505 pp. Berlin: Springer; New York: Heidelberg.

Bratteli O and Robinson DW (1997) *Operator Algebras and Quantum Statistical Mechanics 2*, 2nd edn., 518 pp. Berlin: Springer; New York: Heidelberg.

Connes A (1994) *Noncommutative Geometry*, 661 pp. San Diego, CA: Academic Press.

Davies EB (1976) *Quantum Theory of Open Systems*, 171 pp. London: Academic Press.

Gross L (1972) Existence and uniqueness of physical ground states. *Journal of Functional Analysis* 10: 52–109.

Lance EC (1995) *Hilbert C^*-Modules*, 130 pp. London: Cambridge University Press.

Ohya M and Petz D (1993) *Quantum Entropy and Its Use*, 335 pp. New York: Academic Press.

Paulsen VI (1996) *Completely Bounded Maps and Dilations*, 187 pp. Harlow: Longman Scientific-Technical.

Pisier G (2003) *Introduction to Operator Space Theory*, 479 pp. London: Cambridge University Press.

Størmer E (1963) Positive linear maps of operator algebras. *Acta Mathematica* 110: 233–278.

Takesaki M (2004a) Theory of Operator Algebras I, Second printing of the 1st edn., 415 pp. Berlin: Springer.

Takesaki M (2004b) *Theory of Operator Algebras II*, 384 pp. Berlin: Springer.

Takesaki M (2004c) Theory of Operator Algebras III, 548 pp. Berlin: Springer.

Pseudo-Riemannian Nilpotent Lie Groups

P E Parker, Wichita State University, Wichita, KS, USA

Nilpotent Lie Groups

While not much had been published on the geometry of nilpotent Lie groups with a left-invariant Riemannian metric till around 1990, the situation is certainly better now; see the references in Eberlein (2004). However, there is still very little that is conspicuous about the more general pseudo-Riemannian case. In particular, the two-step nilpotent groups are nonabelian and as close as possible to being abelian, but display a rich variety of new and interesting geometric phenomena (Cordero and Parker 1999). As in the Riemannian case, one of many places where they arise naturally is as groups of isometries acting on horospheres in certain (pseudo-Riemannian) symmetric spaces. Another is in the Iwasawa decomposition $G = KAN$ of semisimple groups with the Killing metric tensor, which need not be (positive or negative) definite even on N. Here, K is compact and A is abelian.

An early motivation for this study was the observation that there are two nonisometric pseudo-Riemannian metrics on the Heisenberg group H_3, one of which is flat. This is a strong contrast to the Riemannian case in which there is only one (up to positive homothety) and it is *not* flat. This is not an anomaly, as we now well know.

While the idea of more than one timelike dimension has appeared a few times in the physics literature, both in string/M-theory and in brane-world scenarios, essentially all work to date assumes only one. Thus, all applications so far are of Lorentzian or definite nilpotent groups. Guediri and co-workers led the Lorentzian studies, and most of their results stated near the end of the section "Lorentzian groups" concern a major, perennial interest in relativity: the (non)existence of closed timelike geodesics in compact Lorentzian manifolds.

Others have made use of nilpotent Lie groups with left-invariant (positive or negative) definite metric tensors, such as Hervig's (2004) constructions of black hole spacetimes from solvmanifolds (related to solvable groups: those with Iwasawa decomposition $G = AN$), including the so-called BTZ constructions. Definite groups and their applications, already having received thorough surveys elsewhere, most notably those of Eberlein, are not included here.

Although the geometric properties of Lie groups with left-invariant definite metric tensors have been studied extensively, the same has not occurred for indefinite metric tensors. For example, while the paper of Milnor (1976) has already become a classic reference, in particular for the classification of positive-definite (Riemannian) metrics on three-dimensional Lie groups, a classification of the left-invariant Lorentzian metric tensors on these groups became available only in 1997. Similarly, only a few partial results in the line of Milnor's study of definite metrics were previously known for indefinite metrics. Moreover, in dimension 3, there are only two types of metric tensors: Riemannian (definite) and Lorentzian (indefinite). But in higher dimensions, there are many distinct types of indefinite metrics while there is still essentially only one type of definite metric. This is another reason why this area has special interest now.

The list in "Further reading" at the end of this article consists of general survey articles and a select few of the more historically important papers. Precise bibliographical information for references merely mentioned or alluded to in this article may be found in those. The main, general reference on pseudo-Riemannian geometry is O'Neill's (1983) book. Eberlein's (2004) article covers the Riemannian case. At this time, there is no other comprehensive survey of the pseudo-Riemannian case. One may use Cordero and Parker (1999) and Guediri (2003) and their reference lists to good advantage, however.

Inner Product and Signature

By an inner product on a vector space V we shall mean a nondegenerate, symmetric bilinear form on V, generally denoted by $\langle , \rangle$. In particular, we *do not* assume that it is positive definite. It has become customary to refer to an ordered pair of nonnegative integers (p, q) as the signature of the inner product, where p denotes the number of positive eigenvalues and q the number of negative eigenvalues. Then nondegeneracy means that $p + q = \dim V$. Note that there is no real geometric difference between (p, q) and (q, p); indeed, O'Neill gives handy conversion procedures for this and for the other major sign variant (e.g., curvature) (see O'Neill (1983, pp. 92 and 89, respectively)).

A Riemannian inner product has signature $(p, 0)$. In view of the preceding remark, one might as well regard signature $(0, q)$ as also being Riemannian, so that "Riemannian geometry is that of definite metric tensors." Similarly, a Lorentzian inner product has either $p = 1$ or $q = 1$. In this case, both sign conventions are used in relativistic theories with the proviso that the "1" axis is always timelike.

If neither p nor q is 1, there is no physical convention. We shall say that $v \in V$ is *timelike* if $\langle v, v \rangle > 0$, *null* if $\langle v, v \rangle = 0$, and *spacelike* if $\langle v, v \rangle < 0$. (In a Lorentzian example, one may wish to revert to one's preferred relativistic convention.) We shall refer to these collectively as the *causal type* of a vector (or of a curve to which a vector is tangent).

Considering indefinite inner products (and metric tensors) thus greatly expands one's purview, from one type of geometry (Riemannian), or possibly two (Riemannian and Lorentzian), to a total of $\lfloor (p + q)/2 \rfloor + 1$ distinctly different types of geometries on the same underlying differential manifolds.

Rise of 2-Step Groups

Throughout, N will denote a connected (and simply connected, usually), nilpotent Lie group with Lie algebra $\mathfrak{n}$ having center $\mathfrak{z}$. We shall use $\langle , \rangle$ to denote either an inner product on $\mathfrak{n}$ or the induced left-invariant pseudo-Riemannian (indefinite) metric tensor on N.

For all nilpotent Lie groups, the exponential map $\exp : \mathfrak{n} \to N$ is surjective. Indeed, it is a diffeomorphism for simply connected N; in this case, we shall denote the inverse by log.

One of the earliest papers on the Riemannian geometry of nilpotent Lie groups was Wolf (1964). Since then, a few other papers about general nilpotent Lie groups have appeared, including Karidi (1994) and Pauls (2001), but the area has not seen a lot of progress.

However, everything changed with Kaplan's (1981) publication. Following this paper and its successor (Kaplan 1983), almost all subsequent work on the left-invariant geometry of nilpotent groups has been on two-step groups.

Briefly, Kaplan defined a new class of nilpotent Lie groups, calling them *of Heisenberg type*. This was soon abbreviated to H-type, and has since been called also as Heisenberg-like and (unfortunately) "generalized Heisenberg." (Unfortunate, because that term was already in use for another class, not all of which are of H-type.) What made them so

compelling was that (almost) everything was explicitly calculable, thus making them the next great test bed after symmetric spaces.

Definition 1 We say that N (or $\mathfrak{n}$) is 2-step nilpotent when $[\mathfrak{n},\mathfrak{n}] \subseteq \mathfrak{z}$. Then $[[\mathfrak{n},\mathfrak{n}],\mathfrak{n}] = 0$ and the generalization to k-step nilpotent is clear:

$$[[\cdots[[[\mathfrak{n},\mathfrak{n}],\mathfrak{n}],\mathfrak{n}]\cdots],\mathfrak{n}] = 0$$

with $k+1$ copies of $\mathfrak{n}$ (or k nested brackets, if you prefer).

It soon became apparent that H-type groups comprised a subclass of 2-step groups; for a nice, modern proof see Berndt *et al.* (1995). By around 1990, they had also attracted the attention of the spectral geometry community, and Eberlein produced the seminal survey (with important new results) from which the modern era began. (It was published in 1994 (Eberlein 1994), but the preprint had circulated widely since 1990.) Since then, activity around 2-step nilpotent Lie groups has mushroomed; see the references in Eberlein (2004).

Finally, turning to pseudo-Riemannian nilpotent Lie groups, with perhaps one or two exceptions, all results so far have been obtained only for 2-step groups. Thus, the remaining sections of this article will be devoted almost exclusively to them.

The Baker–Campbell–Hausdorff formula takes on a particularly simple form in these groups:

$$\exp(x)\exp(y) = \exp\left(x + y + \tfrac{1}{2}[x,y]\right) \qquad [1]$$

Proposition 1 *In a pseudo-Riemannian 2-step nilpotent Lie group, the exponential map preserves causal character. Alternatively, one-parameter subgroups are curves of constant causal character.*

Of course, one-parameter subgroups need not be geodesics.

Lattices and Completeness

We shall need some basic facts about lattices in N. In nilpotent Lie groups, a lattice is a discrete subgroup Γ such that the homogeneous space $M = \Gamma\backslash N$ is compact. Here we follow the convention that a lattice acts on the left, so that the coset space consists of left cosets and this is indicated by the notation. Other subgroups will generally act on the right, allowing better separation of the effects of two simultaneous actions.

Lattices do not always exist in nilpotent Lie groups.

Theorem 1 *The simply connected, nilpotent Lie group N admits a lattice if and only if there exists a basis of its Lie algebra $\mathfrak{n}$ for which the structure constants are rational.*

Such a group is said to have a rational structure, or simply to be rational.

A *nilmanifold* is a (compact) homogeneous space of the form $\Gamma\backslash N$, where N is a connected, simply connected (rational) nilpotent Lie group and Γ is a lattice in N. An *infranilmanifold* has a nilmanifold as a finite covering space. They are commonly regarded as a noncommutative generalization of tori, the Klein bottle being the simplest example of an infranilmanifold that is not a nilmanifold.

We recall the result of Marsden from O'Neill (1983).

Theorem 2 *A compact, homogeneous pseudo-Riemannian space is geodesically complete.*

Thus, if a rational N is provided with a bi-invariant metric tensor $\langle\,,\rangle$, then M becomes a compact, homogeneous pseudo-Riemannian space which is therefore complete. It follows that $(N,\langle\,,\rangle)$ is itself complete. In general, however, the metric tensor is not bi-invariant and N need not be complete.

For 2-step nilpotent Lie groups, things work nicely as shown by this result first published by Guediri.

Theorem 3 *On a 2-step nilpotent Lie group, all left-invariant pseudo-Riemannian metrics are geodesically complete.*

No such general result holds for 3- and higher-step groups, however.

2-Step Groups

In the Riemannian (positive-definite) case, one splits $\mathfrak{n} = \mathfrak{z} \oplus \mathfrak{v} = \mathfrak{z} \oplus \mathfrak{z}^{\perp}$, where the superscript denotes the orthogonal complement with respect to the inner product $\langle\,,\rangle$. In the general pseudo-Riemannian case, however, $\mathfrak{z} \oplus \mathfrak{z}^{\perp} \neq \mathfrak{n}$. The problem is that $\mathfrak{z}$ might be a degenerate subspace; that is, it might contain a null subspace $\mathfrak{U}$ for which $\mathfrak{U} \subseteq \mathfrak{U}^{\perp}$.

It turns out that this possible degeneracy of the center causes the essential differences between the Riemannian and pseudo-Riemannian cases. So far, the only general success in studying groups with degenerate centers was in Cordero and Parker (1999) where an adapted Witt decomposition of $\mathfrak{n}$ was used together with an involution ι exchanging the two null parts.

Observe that if $\mathfrak{z}$ is degenerate, the null subspace $\mathfrak{U}$ is well defined invariantly. We shall use a decomposition

$$\mathfrak{n} = \mathfrak{z} \oplus \mathfrak{v} = \mathfrak{U} \oplus \mathfrak{Z} \oplus \mathfrak{V} \oplus \mathfrak{E} \qquad [2]$$

in which $\mathfrak{z} = \mathfrak{U} \oplus \mathfrak{Z}$ and $\mathfrak{v} = \mathfrak{V} \oplus \mathfrak{E}$, $\mathfrak{U}$ and $\mathfrak{V}$ are complementary null subspaces, and $\mathfrak{U}^\perp \cap \mathfrak{V}^\perp = \mathfrak{Z} \oplus \mathfrak{E}$. Although the choice of $\mathfrak{V}$ is not well defined invariantly, once a $\mathfrak{V}$ has been chosen then $\mathfrak{Z}$ and $\mathfrak{E}$ are well defined invariantly. Indeed, $\mathfrak{Z}$ is the portion of the center $\mathfrak{z}$ in $\mathfrak{U}^\perp \cap \mathfrak{V}^\perp$, and $\mathfrak{E}$ is its orthocomplement in $\mathfrak{U}^\perp \cap \mathfrak{V}^\perp$. This is a Witt decomposition of $\mathfrak{n}$ given $\mathfrak{U}$, easily seen by noting that $(\mathfrak{U} \oplus \mathfrak{V})^\perp = \mathfrak{Z} \oplus \mathfrak{E}$, adapted to the special role of the center in $\mathfrak{n}$.

We shall also need to use an involution ι that interchanges $\mathfrak{U}$ and $\mathfrak{V}$ and which reduces to the identity on $\mathfrak{Z} \oplus \mathfrak{E}$ in the Riemannian (positive-definite) case. (The particular choice of such an involution is not significant.) It turns out that ι is an isometry of $\mathfrak{n}$ which does not integrate to an isometry of N. The adjoint with respect to $\langle\,,\rangle$ of the adjoint representation of the Lie algebra $\mathfrak{n}$ on itself is denoted by $\mathrm{ad}^\dagger$.

Definition 2 The linear mapping

$$j : \mathfrak{U} \oplus \mathfrak{Z} \to \mathrm{End}(\mathfrak{V} \oplus \mathfrak{E})$$

is given by

$$j(a)x = \iota\, \mathrm{ad}_x^\dagger\, \iota a$$

Formulas for the connection and curvatures, and explicit forms for many examples, may be found in Cordero and Parker (1999). It turns out there is a relatively large class of flat spaces, a clear distinction from the Riemannian case in which there are none.

Let $x, y \in \mathfrak{n}$. Recall that homaloidal planes are those for which the numerator $\langle R(x,y)y, x\rangle$ of the sectional curvature formula vanishes. This notion is useful for degenerate planes tangent to spaces that are not of constant curvature.

Definition 3 A submanifold of a pseudo-Riemannian manifold is flat if and only if every plane tangent to the submanifold is homaloidal.

Theorem 4 *The center Z of N is flat.*

Corollary 1 *The only N of constant curvature are flat.*

The degenerate part of the center can have a profound effect on the geometry of the whole group.

Theorem 5 *If $[\mathfrak{n}, \mathfrak{n}] \subseteq \mathfrak{U}$ and $\mathfrak{E} = \{0\}$, then N is flat.*

Among these spaces, those that also have $\mathfrak{Z} = \{0\}$ (which condition itself implies $[\mathfrak{n}, \mathfrak{n}] \subseteq \mathfrak{U}$) are fundamental, with the more general ones obtained by making nondegenerate central extensions. It is also easy to see that the product of any flat group with a nondegenerate abelian factor is still flat.

This is the best possible result in general. Using weaker hypotheses in place of $\mathfrak{E} = \{0\}$, such as $[\mathfrak{V}, \mathfrak{V}] = \{0\} = [\mathfrak{E}, \mathfrak{E}]$, it is easy to construct examples which are not flat.

Corollary 2 *If $\dim Z \geq \lceil n/2 \rceil$, then there exists a flat metric on N.*

Here $\lceil r \rceil$ denotes the least integer greater than or equal to r and $n = \dim N$.

Before continuing, we pause to collect some facts about the condition $[\mathfrak{n}, \mathfrak{n}] \subseteq \mathfrak{U}$ and its consequences.

Remark 1 Since it implies $j(z) = 0$ for all $z \in \mathfrak{Z}$, this latter is possible with no pseudo-Euclidean de Rham factor, unlike the Riemannian case. (On the other hand, a pseudo-Euclidean de Rham factor is characterized in terms of the Kaplan-Eberlein map j whenever the center is nondegenerate.)

Also, it implies $j(u)$ interchanges $\mathfrak{V}$ and $\mathfrak{E}$ for all $u \in \mathfrak{U}$ if and only if $[\mathfrak{V}, \mathfrak{V}] = [\mathfrak{E}, \mathfrak{E}] = \{0\}$. Examples are the Heisenberg group and the groups $H(p, 1)$ for $p \geq 2$ with null centers.

Finally, we note that it implies that, for every $u \in \mathfrak{U}$, $j(u)$ maps $\mathfrak{V}$ to $\mathfrak{V}$ if and only if $j(u)$ maps $\mathfrak{E}$ to $\mathfrak{E}$ if and only if $[\mathfrak{V}, \mathfrak{E}] = \{0\}$.

Proposition 2 *If $j(z) = 0$ for all $z \in \mathfrak{Z}$ and $j(u)$ interchanges $\mathfrak{V}$ and $\mathfrak{E}$ for all $u \in \mathfrak{U}$, then N is Ricci flat.*

Proposition 3 *If $j(z) = 0$ for all $z \in \mathfrak{Z}$, then N is scalar flat. In particular, this occurs when $[\mathfrak{n}, \mathfrak{n}] \subseteq \mathfrak{U}$.*

Much like the Riemannian case, we would expect that $(N, \langle\,,\rangle)$ should in some sense be similar to flat pseudo-Euclidean space. This is seen, for example, via the existence of totally geodesic subgroups (Cordero and Parker 1999). (O'Neill (1983, ex. 9, p. 125) has extended the definition of totally geodesic to degenerate submanifolds of pseudo-Riemannian manifolds.)

Example 1 For any $x \in \mathfrak{n}$ the one-parameter subgroup $\exp(tx)$ is a geodesic if and only if $x \in \mathfrak{z}$ or $x \in \mathfrak{U} \oplus \mathfrak{E}$. This is essentially the same as the Riemannian case, but with some additional geodesic one-parameter subgroups coming from $\mathfrak{U}$.

Example 2 Abelian subspaces of $\mathfrak{V} \oplus \mathfrak{E}$ are Lie subalgebras of $\mathfrak{n}$, and give rise to complete, flat, totally geodesic abelian subgroups of N, just as in the Riemannian case. Eberlein's construction is valid in general, and shows that if $\dim \mathfrak{V} \oplus \mathfrak{E} \geq 1 + k + k \dim \mathfrak{z}$, then every nonzero element of $\mathfrak{V} \oplus \mathfrak{E}$ lies in an abelian subspace of dimension $k + 1$.

Example 3 The center Z of N is a complete, flat, totally geodesic submanifold. Moreover, it determines a foliation of N by its left translates, so each leaf is flat and totally geodesic, as in the Riemannian

case. In the pseudo-Riemannian case, this foliation in turn is the orthogonal direct sum of two foliations determined by $\mathfrak{U}$ and $\mathfrak{Z}$, and the leaves of the $\mathfrak{U}$-foliation are also null. All these leaves are complete.

There is also the existence of $\dim \mathfrak{z}$ independent first integrals, a familiar result in pseudo-Euclidean space, and the geodesic equations are completely integrable; in certain cases (mostly when the center is nondegenerate), one can obtain explicit formulas. Unlike the Riemannian case, there are flat groups (nonabelian) which are isometric to pseudo-Euclidean spaces (abelian).

Theorem 6 *If* $[\mathfrak{n},\mathfrak{n}] \subseteq \mathfrak{U}$ *and* $\mathfrak{E} = \{0\}$, *then* N *is geodesically connected. Consequently, so is any nilmanifold with such a universal covering space.*

Thus, these compact nilmanifolds are much like tori. This is also illustrated by the computation of their period spectrum.

Isometry Group

The main new feature is that when the center is degenerate, the isometry group can be strictly larger in a significant way than when the center is nondegenerate (which includes the Riemannian case).

Letting $\mathrm{Aut}(N)$ denote the automorphism group of N and $I(N)$ the isometry group of N, set $O(N) = \mathrm{Aut}(N) \cap I(N)$. In the Riemannian case, $I(N) = O(N) \ltimes N$, the semidirect product where N acts as left translations. We have chosen the notation $O(N)$ to suggest an analogy with the pseudo-Euclidean case in which this subgroup is precisely the (general, including reflections) pseudo-orthogonal group. According to Wilson (1982), this analogy is good for any nilmanifold (not necessarily 2-step).

To see what is true about the isometry group in general, first consider the (left-invariant) splitting of the tangent bundle $TN = \mathfrak{z}N \oplus \mathfrak{v}N$.

Definition 4 Denote by $I^{\mathrm{spl}}(N)$ the subgroup of the isometry group $I(N)$ which preserves the splitting $TN = \mathfrak{z}N \oplus \mathfrak{v}N$. Further, let $I^{\mathrm{aut}}(N) = O(N) \ltimes N$, where N acts by left translations.

Proposition 4 *If* N *is a simply connected, 2-step nilpotent Lie group with left-invariant metric tensor, then* $I^{\mathrm{spl}}(N) \leq I^{\mathrm{aut}}(N)$.

There are examples to show that $I^{\mathrm{spl}} < I^{\mathrm{aut}}$ is possible when $\mathfrak{U} \neq \{0\}$.

When the center is degenerate, the relevant group analogous to a pseudo-orthogonal group may be larger.

Proposition 5 *Let* $\tilde{O}(N)$ *denote the subgroup of* $I(N)$ *which fixes* $1 \in N$. *Then* $I(N) \cong \tilde{O}(N) \ltimes N$, *where* N *acts by left translations.*

The proof is obvious from the definition of $\tilde{O}$. It is also obvious that $O \leq \tilde{O}$. Examples show that $O < \tilde{O}$, hence $I^{\mathrm{aut}} < I$, is possible when the center is degenerate.

Thus, we have three groups of isometries, not necessarily equal in general: $I^{\mathrm{spl}} \leq I^{\mathrm{aut}} \leq I$. When the center is nondegenerate ($\mathfrak{U} = \{0\}$), the Ricci transformation is block-diagonalizable and the rest of Kaplan's proof using it now also works.

Corollary 3 *If the center is nondegenerate, then* $I(N) = I^{\mathrm{spl}}(N)$ *whence* $\tilde{O}(N) \cong O(N)$.

In the next few results, we use the phrase "a subgroup isometric to" a group to mean that the isometry is also an isomorphism of groups.

Proposition 6 *For any* N *containing a subgroup isometric to the flat three-dimensional Heisenberg group,*

$$I^{\mathrm{spl}}(N) < I^{\mathrm{aut}}(N) < I(N)$$

Unfortunately, this class does not include our flat groups in which $[\mathfrak{n},\mathfrak{n}] \subseteq \mathfrak{U}$ and $\mathfrak{E} = \{0\}$. However, it does include many groups that do not satisfy $[\mathfrak{n},\mathfrak{n}] \subseteq \mathfrak{U}$, such as the simplest quaternionic Heisenberg group.

Remark 2 A direct computation shows that on this flat H_3 with null center, the only Killing fields with geodesic integral curves are the nonzero scalar multiples of a vector field tangent to the center.

Proposition 7 *For any* N *containing a subgroup isometric to the flat* $H_3 \times \mathbb{R}$ *with null center,*

$$I^{\mathrm{spl}}(N) < I^{\mathrm{aut}}(N) < I(N)$$

Many of our flat groups in which $[\mathfrak{n},\mathfrak{n}] \subseteq \mathfrak{U}$ and $\mathfrak{E} = \{0\}$ have such a subgroup isometrically embedded, as in fact do many others which are not flat.

Lattices and Periodic Geodesics

In this subsection, we assume that N is rational and let Γ be a lattice in N.

Certain tori T_F and T_B provide the model fiber and the base for a submersion of the coset space $\Gamma \backslash N$. This submersion may not be pseudo-Riemannian in the usual sense, because the tori may be degenerate. We began the study of periodic geodesics in these compact nilmanifolds, and obtained a complete calculation of the period spectrum for certain flat spaces.

To the compact nilmanifold $\Gamma\backslash N$ we may associate two flat (possibly degenerate) tori.

Definition 5 Let N be a simply connected, two-step nilpotent Lie group with lattice Γ and let $\pi:\mathfrak{n}\to\mathfrak{v}$ denote the projection. Define

$$T_{\mathfrak{z}} = \mathfrak{z}/(\log\Gamma\cap\mathfrak{z})$$
$$T_{\mathfrak{v}} = \mathfrak{v}/\pi(\log\Gamma)$$

Observe that $\dim T_{\mathfrak{z}} + \dim T_{\mathfrak{v}} = \dim\mathfrak{z} + \dim\mathfrak{v} = \dim\mathfrak{n}$.

Let $m = \dim\mathfrak{z}$ and $n = \dim\mathfrak{v}$. It is a consequence of a theorem of Palais and Stewart that $\Gamma\backslash N$ is a principal T^m-bundle over T^n. The model fiber T^m can be given a geometric structure from its closed embedding in $\Gamma\backslash N$; we denote this geometric m-torus by T_F. Similarly, we wish to provide the base n-torus with a geometric structure so that the projection $p_B:\Gamma\backslash N\twoheadrightarrow T_B$ is the appropriate generalization of a pseudo-Riemannian submersion (O'Neill 1983) to (possibly) degenerate spaces. Observe that the splitting $\mathfrak{n}=\mathfrak{z}\oplus\mathfrak{v}$ induces splittings $TN=\mathfrak{z}N\oplus\mathfrak{v}N$ and $T(\Gamma\backslash N)=\mathfrak{z}(\Gamma\backslash N)\oplus\mathfrak{v}(\Gamma\backslash N)$, and that p_{B*} just mods out $\mathfrak{z}(\Gamma\backslash N)$. Examining O'Neill's definition, we see that the key is to construct the geometry of T_B by defining

$$p_{B*}:\mathfrak{v}_\eta(\Gamma\backslash N)\to T_{p_B(\eta)}(T_B)$$
$$\text{for each } \eta\in\Gamma\backslash N \text{ is an isometry} \qquad [3]$$

and

$$\nabla^{T_B}_{p_{B*}x} p_{B*}y = p_{B*}(\pi\nabla_x y)$$
$$\text{for all } x,y\in\mathfrak{v}=\mathfrak{V}\oplus\mathfrak{E} \qquad [4]$$

where $\pi:\mathfrak{n}\to\mathfrak{v}$ is the projection. Then the rest of the usual results will continue to hold, provided that sectional curvature is replaced by the numerator of the sectional curvature formula at least when elements of $\mathfrak{V}$ are involved:

$$\langle R_{T_B}(p_{B*}x,p_{B*}y)\,p_{B*}y,p_{B*}x\rangle$$
$$= \langle R_{\Gamma\backslash N}(x,y)y,x\rangle + \tfrac{3}{4}\langle[x,y],[x,y]\rangle \qquad [5]$$

Now p_B will be a pseudo-Riemannian submersion in the usual sense if and only if $\mathfrak{U}=\mathfrak{V}=\{0\}$, as is always the case for Riemannian spaces.

In the Riemannian case, Eberlein showed that $T_F\cong T_{\mathfrak{z}}$ and $T_B\cong T_{\mathfrak{v}}$. In general, T_B is flat only if N has a nondegenerate center or is flat.

Remark 3 Observe that the torus T_B may be decomposed into a topological product $T_E\times T_V$ in the obvious way. It is easy to check that T_E is flat and isometric to $(\log\Gamma\cap\mathfrak{E})\backslash\mathfrak{E}$, and that T_V has a linear connection not coming from a metric and not flat in general. Moreover, the geometry of the product is "twisted" in a certain way. It would be interesting to determine which tori could appear as such a T_V and how.

Theorem 7 *Let N be a simply connected, 2-step nilpotent Lie group with lattice Γ, a left-invariant metric tensor, and tori as above. The fibers T_F of the (generalized) pseudo-Riemannian submersion $\Gamma\backslash N\twoheadrightarrow T_B$ are isometric to $T_{\mathfrak{z}}$. If in addition the center Z of N is nondegenerate, then the base T_B is isometric to $T_{\mathfrak{v}}$.*

We recall that elements of N can be identified with elements of the isometry group $I(N)$: namely, $n\in N$ is identified with the isometry $\phi=L_n$ of left translation by n. We shall abbreviate this by writing $\phi\in N$.

Definition 6 We say that $\phi\in N$ translates the geodesic γ by ω if and only if $\phi\gamma(t)=\gamma(t+\omega)$ for all t. If γ is a unit-speed geodesic, we say that ω is a period of ϕ.

Recall that unit speed means that $|\dot\gamma| = |\langle\dot\gamma,\dot\gamma\rangle|^{1/2}=1$. Since there is no natural normalization for null geodesics, we do not define periods for them. In the Riemannian case and in the timelike Lorentzian case in strongly causal spacetimes, unit-speed geodesics are parametrized by arclength and this period is a translation distance. If ϕ belongs to a lattice Γ, it is the length of a closed geodesic in $\Gamma\backslash N$.

In general, recall that if γ is a geodesic in N and if $p_N:N\twoheadrightarrow\Gamma\backslash N$ denotes the natural projection, then $p_N\gamma$ is a periodic geodesic in $\Gamma\backslash N$ if and only if some $\phi\in\Gamma$ translates γ. We say *periodic* rather than *closed* here because in pseudo-Riemannian spaces it is possible for a null geodesic to be closed but not periodic. If the space is geodesically complete or Riemannian, however, then this does not occur; the former is in fact the case for our 2-step nilpotent Lie groups. Further, recall that free homotopy classes of closed curves in $\Gamma\backslash N$ correspond bijectively with conjugacy classes in Γ.

Definition 7 Let $\mathcal{C}$ denote either a nontrivial, free homotopy class of closed curves in $\Gamma\backslash N$ or the corresponding conjugacy class in Γ. We define $\wp(\mathcal{C})$ to be the set of all periods of periodic unit-speed geodesics that belong to $\mathcal{C}$.

In the Riemannian case, this is the set of lengths of closed geodesics in $\mathcal{C}$, frequently denoted by $\ell(\mathcal{C})$.

Definition 8 The *period spectrum* of $\Gamma\backslash N$ is the set

$$\mathrm{spec}_\wp(\Gamma\backslash N) = \bigcup_{\mathcal{C}}\wp(\mathcal{C})$$

where the union is taken over all nontrivial, free homotopy classes of closed curves in $\Gamma\backslash N$.

In the Riemannian case, this is the length spectrum $\text{spec}_\ell(\Gamma\backslash N)$.

Example 4 Similar to the Riemannian case, we can compute the period spectrum of a flat torus $\Gamma\backslash\mathbb{R}^m$, where Γ is a lattice (of maximal rank, isomorphic to $\mathbb{Z}^m$). Using calculations in an analogous way as for finding the length spectrum of a Riemannian flat torus, we easily obtain

$$\text{spec}_\wp(\Gamma\backslash\mathbb{R}^m) = \{|g| \neq 0 \mid g \in \Gamma\}$$

It is also easy to see that the nonzero d'Alembertian spectrum is related to the analogous set produced from the dual lattice Γ^*, multiplied by factors of $\pm 4\pi^2$, almost as in the Riemannian case.

As in this example, simple determinacy of periods of unit-speed geodesics helps make calculation of the period spectrum possible purely in terms of $\log\Gamma \subseteq \mathfrak{n}$.

For the rest of this subsection, we assume that N is a simply connected, two-step nilpotent Lie group with left-invariant pseudo-Riemannian metric tensor $\langle\,,\rangle$. Note that non-null geodesics may be taken to be of unit speed. Most non-identity elements of N translate some geodesic, but not necessarily one of unit speed.

For our special class of flat 2-step nilmanifolds, we can calculate the period spectrum completely.

Theorem 8 *If* $[\mathfrak{n},\mathfrak{n}] \subseteq \mathfrak{U}$ *and* $\mathfrak{E} = \{0\}$, *then* $\text{spec}_\wp(M)$ *can be completely calculated from* $\log\Gamma$ *for any* $M = \Gamma\backslash N$.

Thus, we see again just how much these flat, two-step nilmanifolds are like tori. All periods can be calculated purely from $\log\Gamma \subseteq \mathfrak{n}$, although some will not show up from the tori in the fibration.

Corollary 4 $\text{spec}_\wp(T_B)$ (*respectively*, T_F) *is* $\cup_{\mathcal{C}}\wp(\mathcal{C})$ *where the union is taken over all those free homotopy classes* $\mathcal{C}$ *of closed curves in* $M = \Gamma\backslash N$ *that* do not (*respectively*, do) *contain an element in the center of* $\Gamma \cong \pi_1(M)$, *except for those periods arising only from unit-speed geodesics in* M *that project to null geodesics in both* T_B *and* T_F.

We note that one might consider using this to assign periods to some null geodesics in the tori T_B and T_F.

When the center is nondegenerate, we obtain results similar to Eberlein's. Here is part of them.

Theorem 9 *Assume* $\mathfrak{U} = \{0\}$. *Let* $\phi \in N$ *and write* $\log\phi = z^* + e^*$. *Assume* ϕ *translates the unit-speed geodesic* γ *by* $\omega > 0$. *Let* z' *denote the component of* z^* *orthogonal to* $[e^*,\mathfrak{n}]$ *and set* $\omega^* = |z' + e^*|$. *Let* $\dot\gamma(0) = z_0 + e_0$. *Then*

(i) $|e^*| \leq \omega$. *In addition,* $\omega < \omega^*$ *for timelike* (*spacelike*) *geodesics with* $\omega z_0 - z'$ *timelike* (*spacelike*), *and* $\omega > \omega^*$ *for timelike* (*spacelike*) *geodesics with* $\omega z_0 - z'$ *spacelike* (*timelike*);

(ii) $\omega = |e^*|$ *if and only if* $\gamma(t) = \exp(te^*/|e^*|)$ *for all* $t \in \mathbb{R}$; *and*

(iii) $\omega = \omega^*$ *if and only if* $\omega z_0 - z'$ *is null*.

Although ω^* need not be an upper bound for periods as in the Riemannian case, it nonetheless plays a special role among all periods, as seen in (iii) above, and we shall refer to it as the distinguished period associated with $\phi \in N$. When the center is definite, for example, we do have $\omega \leq \omega^*$.

Now the following definitions make sense at least for N with a nondegenerate center.

Definition 9 Let $\mathcal{C}$ denote either a nontrivial, free homotopy class of closed curves in $\Gamma\backslash N$ or the corresponding conjugacy class in Γ. We define $\wp^*(\mathcal{C})$ to be the distinguished periods of periodic unit-speed geodesics that belong to $\mathcal{C}$.

Definition 10 The *distinguished period spectrum* of $\Gamma\backslash N$ is the set

$$\mathcal{D}\text{spec}_\wp(\Gamma\backslash N) = \bigcup_{\mathcal{C}} \wp^*(\mathcal{C})$$

where the union is taken over all nontrivial, free homotopy classes of closed curves in $\Gamma\backslash N$.

Then we get this result:

Corollary 5 *Assume the center is nondegenerate. If* $\mathfrak{n}$ *is nonsingular, then* $\text{spec}_\wp(T_B)$ (*respectively*, T_F) *is precisely the period spectrum* (*respectively*, *the distinguished period spectrum*) *of those free homotopy classes* $\mathcal{C}$ *of closed curves in* $M = \Gamma\backslash N$ *that do not* (*respectively*, *do*) *contain an element in the center of* $\Gamma \cong \pi_1(M)$, *except for those periods arising only from unit-speed geodesics in* M *that project to null geodesics in both* T_B *and* T_F.

Conjugate Loci

This is the only general result on conjugate points.

Proposition 8 *Let* N *be a simply connected, 2-step nilpotent Lie group with left-invariant metric tensor* $\langle\,,\rangle$, *and let* γ *be a geodesic with* $\dot\gamma(0) = a \in \mathfrak{z}$. *If* $\text{ad}^\dagger_\bullet a = 0$, *then there are no conjugate points along* γ.

In the rest of this subsection, we assume that the center of N is nondegenerate.

For convenience, we shall use the notation $J_z = \text{ad}^\dagger_\bullet z$ for any $z \in \mathfrak{z}$. (Since the center is

nondegenerate, the involution ι may be omitted.) We follow Ciatti (2000) for this next definition. As in the Riemannian case, one might as well make 2-step nilpotency part of the definition since it effectively is so anyway.

Definition 11 N is said to be of *pseudoH-type* if and only if

$$J_z^2 = -\langle z, z\rangle I$$

for any $z \in \mathfrak{z}$.

Complete results on conjugate loci have been obtained only for these groups (Jang *et al.* 2005). For example, using standard results from analytic function theory, one can show that the conjugate locus is an analytic variety in N. This is probably true for general two-step groups, but the proof we know works only for pseudo$H$-type.

Definition 12 Let γ denote a geodesic and assume that $\gamma(t_0)$ is conjugate to $\gamma(0)$ along γ. To indicate that the multiplicity of $\gamma(t_0)$ is m, we shall write $\text{mult}_{\text{cp}}(t_0) = m$. To distinguish the notions clearly, we shall denote the multiplicity of λ as an eigenvalue of a specified linear transformation by $\text{mult}_{\text{ev}}\lambda$.

Let γ be a geodesic with $\gamma(0) = 1$ and $\dot\gamma(0) = z_0 + x_0 \in \mathfrak{z} \oplus \mathfrak{v}$, respectively, and let $J = J_{z_0}$. If γ is not null, we may assume that γ is normalized so that $\langle\dot\gamma, \dot\gamma\rangle = \pm 1$. As usual, $\mathbb{Z}^*$ denotes the set of all integers with 0 removed.

Theorem 10 *Under these assumptions, if N is of pseudoH-type, then:*

(i) *if $z_0 = 0$ and $x_0 \neq 0$, then $\gamma(t)$ is conjugate to $\gamma(0)$ along γ if and only if $\langle x_0, x_0\rangle < 0$ and*

$$-\frac{12}{t^2} = \langle x_0, x_0\rangle$$

in which case $\text{mult}_{\text{cp}}(t) = \dim \mathfrak{z}$;

(ii) *if $z_0 \neq 0$ and $x_0 = 0$, then $\gamma(t)$ is conjugate to $\gamma(0)$ along γ if and only if $\langle z_0, z_0\rangle > 0$ and*

$$t \in \frac{2\pi}{|z_0|}\mathbb{Z}^*$$

in which case $\text{mult}_{\text{cp}}(t) = \dim \mathfrak{v}$.

Theorem 11 *Let γ be such a geodesic in a pseudoH-type group N with $z_0 \neq 0 \neq x_0$.*

(i) *If $\langle z_0, z_0\rangle = \alpha^2$ with $\alpha > 0$, then $\gamma(t_0)$ is conjugate to $\gamma(0)$ along γ if and only if*

$$t_0 \in \frac{2\pi}{\alpha}\mathbb{Z}^* \cup \mathbb{A}_1 \cup \mathbb{A}_2$$

where

$$\mathbb{A}_1 = \left\{ t \in \mathbb{R} \middle| \langle x_0, x_0\rangle \frac{\alpha t}{2} \cot\frac{\alpha t}{2} = \langle\dot\gamma, \dot\gamma\rangle \right\}$$

and

$$\mathbb{A}_2 = \left\{ t \in \mathbb{R} \middle| \alpha t = \frac{\langle x_0, x_0\rangle}{\langle\dot\gamma, \dot\gamma\rangle + \langle z_0, z_0\rangle} \sin\alpha t \right\}$$

when $\dim \mathfrak{z} \geq 2$

If $t_0 \in (2\pi/\alpha)\mathbb{Z}^*$, then

$$\text{mult}_{\text{cp}}(t_0) = \begin{cases} \dim \mathfrak{v} - 1 & \text{if } \langle\dot\gamma, \dot\gamma\rangle + \langle z_0, z_0\rangle \neq 0 \\ \dim \mathfrak{n} - 2 & \text{if } \langle\dot\gamma, \dot\gamma\rangle + \langle z_0, z_0\rangle = 0 \end{cases}$$

If $t_0 \notin (2\pi/\alpha)\mathbb{Z}^*$, then

$$\text{mult}_{\text{cp}}(t_0) = \begin{cases} 1 & \text{if } t_0 \in \mathbb{A}_1 - \mathbb{A}_2 \\ \dim \mathfrak{z} - 1 & \text{if } t_0 \in \mathbb{A}_2 - \mathbb{A}_1 \\ \dim \mathfrak{z} & \text{if } t_0 \in \mathbb{A}_1 \cap \mathbb{A}_2 \end{cases}$$

(ii) *If $\langle z_0, z_0\rangle = -\beta^2$ with $\beta > 0$, then $\gamma(t_0)$ is a conjugate point along γ if and only if $t_0 \in \mathbb{B}_1 \cup \mathbb{B}_2$ where*

$$\mathbb{B}_1 = \left\{ t \in \mathbb{R} \middle| \langle x_0, x_0\rangle \frac{\beta t}{2} \coth\frac{\beta t}{2} = \langle\dot\gamma, \dot\gamma\rangle \right\}$$

and

$$\mathbb{B}_2 = \left\{ t \in \mathbb{R} \middle| \beta t = \frac{\langle x_0, x_0\rangle}{\langle\dot\gamma, \dot\gamma\rangle + \langle z_0, z_0\rangle} \sinh\beta t \right\}$$

when $\dim \mathfrak{z} \geq 2$

The multiplicity is

$$\text{mult}_{\text{cp}}(t_0) = \begin{cases} 1 & \text{if } t_0 \in \mathbb{B}_1 - \mathbb{B}_2 \\ \dim \mathfrak{z} - 1 & \text{if } t_0 \in \mathbb{B}_2 - \mathbb{B}_1 \\ \dim \mathfrak{z} & \text{if } t_0 \in \mathbb{B}_1 \cap \mathbb{B}_2 \end{cases}$$

(iii) *If $\langle z_0, z_0\rangle = 0$, then $\gamma(t_0)$ is a conjugate point along γ if and only if*

$$t_0^2 = -\frac{12}{\langle x_0, x_0\rangle}$$

and $\text{mult}_{\text{cp}}(t_0) = \dim \mathfrak{z} - 1$.

This covers all cases for a pseudo$H$-type group with a center of any dimension.

Some results on other two-step groups and examples (including pictures in dimension 3) may be found in the references cited in Jang *et al.* (2005). When the groups are not pseudo$H$-type, however, complete results are available only when the center is one dimensional. Guediri (2004) has results in the timelike Lorentzian case.

Lorentzian Groups

Not too long ago, only a few partial results in the line of Milnor's study of definite metrics were known for indefinite metrics (Barnet 1989, Nomizu 1979), and they were Lorentzian.

Guediri (2003) and others have made special study of Lorentzian two-step groups, partly because of their relevance to general relativity, where they can be used to provide interesting and important (counter)examples. Special features of Lorentzian geometry frequently enable them to obtain much more complete and explicit results than are possible in general.

For example, Guediri (2003) was able to provide a complete and explicit integration of the geodesic equations for Lorentzian 2-step groups. This includes the case of a degenerate center, which only required extremely careful handling through a number of cases. He also paid special attention to the existence of closed timelike geodesics, reflecting the relativistic concerns.

As usual, N denotes a connected and simply connected 2-step nilpotent Lie group. For the rest of this section, we assume that the left-invariant metric tensor is Lorentzian. Whenever a lattice is mentioned, we also assume that the group is rational.

Proposition 9 *If the center is degenerate, then no timelike geodesic can be translated by a central element.*

Thus, there can be no closed timelike geodesics parallel to the center in any nilmanifold obtained from such an N.

Theorem 12 *If the center is Lorentzian, then* $\Gamma\backslash N$ *contains no timelike or null closed geodesics for any lattice* Γ.

To handle degenerate centers, three refined notions for nonsingular are used: *almost*, *weakly*, and *strongly* nonsingular. The precise definitions involve an adapted Witt decomposition (as in the general pseudo-Riemannian case, but a rather different one here) and are quite technical, as is typical. We refer to Guediri (2003) for details.

Theorem 13 *If N is weakly nonsingular, then no timelike geodesic can be translated by an element of N.*

Corollary 6 *If N is flat, then no timelike geodesic can be translated by a non-identity element.*

Corollary 7 *If N is flat, then $\Gamma\backslash N$ contains no closed timelike geodesics for any lattice Γ.*

Corollary 8 *If N is weakly nonsingular, then $\Gamma\backslash N$ contains no closed timelike geodesic.*

Corollary 9 *If $N = H_{2k+1}$ is a Lorentzian Heisenberg group with degenerate center, then $\Gamma\backslash N$ contains no closed timelike geodesic.*

Guediri also has the only non-Riemannian results so far about the phenomenon Eberlein called "in resonance." Roughly speaking, this occurs when the eigenvalues of the map j have rational ratios. (The Lorentzian case actually requires a slightly more complicated condition when the center is degenerate.)

Theorem 14 *If N is almost nonsingular, then N is in resonance if and only if every geodesic of N is translated by some element of N.*

See also: Classical Groups and Homogeneous Spaces; Einstein Equations: Exact Solutions; Lorentzian Geometry.

Further Reading

Barnet F (1989) On Lie groups that admit left-invariant Lorentz metrics of constant sectional curvature. *Illinois Journal of Mathematics* 33: 631–642.

Berndt J, Tricerri F, and Vanhecke L (1995) *Generalized Heisenberg Groups and Damek-Ricci Harmonic Spaces*, LNM 1598. Berlin: Springer.

Ciatti P (2000) Scalar products on Clifford modules and pseudo-H-type Lie algebras. *Annali di Matematica Pura ed Applicata* 178: 1–31.

Cordero LA and Parker PE (1999) Pseudoriemannian 2-step nilpotent Lie groups, Santiago-Wichita. *Preprint* DGS/CP4, (arXiv:math.DG/9905188).

Eberlein P (1994) Geometry of 2-step nilpotent groups with a left-invariant metric. *Annales Scientifique de l'École Normale Supérieure* 27: 611–660.

Eberlein P (2004) Left-invariant geometry of Lie groups. *Cubo* 6: 427–510. (See also http://www.math.unc.edu/faculty/pbe.)

Guediri M (2003) Lorentz geometry of 2-step nilpotent Lie groups. *Geometriae Dedicata* 100: 11–51.

Guediri M (2004) The timelike cut locus and conjugate points in Lorentz 2-step nilpotent Lie groups. *Manuscripta Mathematica* 114: 9–35.

Hervig S (2004) Einstein metrics: homogeneous solvmanifolds, generalised Heisenberg groups and black holes. *Journal of Geometry and Physics* 52: 298–312.

Jang C, Parker PE, and Park K (2005) Pseudo$H$-type 2-step nilpotent Lie groups. *Houston Journal of Mathematics* 31: 765–786 (arXiv:math.DG/0307368).

Kaplan A (1981) Riemannian nilmanifolds attached to Clifford modules. *Geometriae Dedicata* 11: 127–136.

Kaplan A (1983) On the geometry of Lie groups of Heisenberg type. *Bulletin of the London Mathematical Society* 15: 35–42.

Karidi R (1994) Geometry of balls in nilpotent Lie groups. *Duke Mathematical Journal* 74: 301–317.

Milnor J (1976) Curvatures of left-invariant metrics on Lie groups. *Advances in Mathematics* 21: 293–329.

Nomizu K (1979) Left-invariant Lorentz metrics on Lie groups. *Osaka Mathematical Journal* 16: 143–150.

O'Neill B (1983) *Semi-Riemannian Geometry*. New York: Academic Press.

Pauls SD (2001) The large scale geometry of nilpotent Lie groups. *Communications in Analysis and Geometry* 9: 951–982.

Wilson EN (1982) Isometry groups on homogeneous manifolds. *Geometriae Dedicata* 12: 337–346.

Wolf JA (1964) Curvature in nilpotent Lie groups. *Proceedings of the American Mathematical Society* 15: 271–274.

Q

q-Special Functions

T H Koornwinder, University o
f Amsterdam, Amsterdam, The Netherlands

Introduction

In this article we give a brief introduction to q-special functions, that is, q-analogs of the classical special functions. Here q is a deformation parameter, usually $0 < q < 1$, where $q = 1$ is the classical case. The deformation is such that the calculus simultaneously deforms to a q-calculus involving q-derivatives and q-integrals. The main topics to be treated are q-hypergeometric series, with some selected evaluation and transformation formulas, and some q-hypergeometric orthogonal polynomials, most notably the Askey–Wilson polynomials. In several variables, we discuss Macdonald polynomials associated with root systems, with most emphasis on the A_n case. The rather new theory of elliptic hypergeometric series gets some attention. While much of the theory of q-special functions keeps q fixed, some of the deeper aspects with number-theoretic and combinatorial flavor emphasize expansion in q. Finally, we indicate applications and interpretations in quantum groups, Chevalley groups, affine Lie algebras, combinatorics, and statistical mechanics.

Conventions

$q \in \mathbb{C}\backslash\{1\}$ in general, but $0 < q < 1$ in all infinite sums and products.
n, m, N will be non-negative integers unless mentioned otherwise.

q-Hypergeometric Series

Definitions

For $a, q \in \mathbb{C}$ the q-shifted factorial $(a;q)_k$ is defined as a product of k factors:

$$(a;q)_k := (1-a)(1-aq)\cdots(1-aq^{k-1}) \quad (k \in \mathbb{Z}_{>0}); \quad (a;q)_0 := 1 \qquad [1]$$

If $|q| < 1$ this definition remains meaningful for $k = \infty$ as a convergent infinite product:

$$(a;q)_\infty := \prod_{j=0}^{\infty}(1-aq^j) \qquad [2]$$

We also write $(a_1, \ldots, a_r; q)_k$ for the product of r q-shifted factorials:

$$(a_1, \ldots, a_r; q)_k := (a_1;q)_k \ldots (a_r;q)_k \quad (k \in \mathbb{Z}_{\geq 0} \text{ or } k = \infty) \qquad [3]$$

A q-hypergeometric series is a power series (for the moment still formal) in one complex variable z with power series coefficients which depend, apart from q, on r complex upper parameters $a_1, \ldots, a_r$ and s complex lower parameters $b_1, \ldots, b_s$ as follows:

$$ {}_r\phi_s\left[\begin{matrix}a_1,\ldots,a_r\\ b_1,\ldots,b_s\end{matrix};q,z\right] = {}_r\phi_s(a_1,\ldots,a_r;b_1,\ldots,b_s;q,z) := \sum_{k=0}^{\infty}\frac{(a_1,\ldots,a_r;q)_k}{(b_1,\ldots,b_s;q)_k(q;q)_k} \times \left((-1)^k q^{(1/2)k(k-1)}\right)^{s-r+1} z^k \quad (r,s \in \mathbb{Z}_{\geq 0}) \qquad [4]$$

Clearly the above expression is symmetric in $a_1, \ldots, a_r$ and symmetric in $b_1, \ldots, b_s$. On the right-hand side of [4], we have that

$$\frac{(k+1)\text{th term}}{k\text{th term}} = \frac{(1-a_1q^k)\cdots(1-a_rq^k)(-q^k)^{s-r+1}z}{(1-b_1q^k)\cdots(1-b_sq^k)(1-q^{k+1})} \qquad [5]$$

is rational in q^k. Conversely, any rational function in q^k can be written in the form of the right-hand side of [5]. Hence, any series $\sum_{k=0}^{\infty} c_k$ with $c_0 = 1$ and c_{k+1}/c_k rational in q^k is of the form of a q-hypergeometric series [4].

In order to avoid singularities in the terms of [4], we assume that $b_1, \ldots, b_s \neq 1, q^{-1}, q^{-2}, \ldots$. If, for some $i, a_i = q^{-n}$, then all terms in the series [4] with $k > n$ will vanish. If none of the a_i is equal to q^{-n}

and if $|q| < 1$, then the radius of convergence of the power series [4] equals ∞ if $r < s+1$, 1 if $r = s+1$, and 0 if $r > s+1$.

We can view the q-shifted factorial as a q-analog of the shifted factorial (or Pochhammer symbol) by the limit formula

$$\lim_{q\to 1} \frac{(q^a;q)_k}{(1-q)^k} = (a)_k := a(a+1)\cdots(a+k-1) \quad [6]$$

Hence the q-binomial coefficient

$$\begin{bmatrix} n \\ k \end{bmatrix}_q := \frac{(q;q)_n}{(q;q)_k (q;q)_{n-k}} \quad (n,k \in \mathbb{Z}, n \geq k \geq 0) \quad [7]$$

tends to the binomial coefficient for $q \to 1$:

$$\lim_{q\to 1} \begin{bmatrix} n \\ k \end{bmatrix}_q = \binom{n}{k} \quad [8]$$

and a suitably renormalized q-hypergeometric series tends (at least formally) to a hypergeometric series as $q \uparrow 1$:

$$\lim_{q\uparrow 1} {}_{r+r'}\phi_{s+s'} \left[\begin{matrix} q^{a_1}, \ldots, q^{a_r}, c_1, \ldots, c_{r'} \\ q^{b_1}, \ldots, q^{b_s}, d_1, \ldots, d_{s'} \end{matrix} ; q, (q-1)^{1+s-r} z \right]$$
$$= {}_rF_s \left(\begin{matrix} a_1, \ldots, a_r \\ b_1, \ldots, b_s \end{matrix} ; \frac{(c_1-1)\cdots(c_{r'}-1)z}{(d_1-1)\cdots(d_{s'}-1)} \right) \quad [9]$$

At least formally, there are limit relations between q-hypergeometric series with neighboring r, s:

$$\lim_{a_r\to\infty} {}_r\phi_s \left[\begin{matrix} a_1, \ldots, a_r \\ b_1, \ldots, b_s \end{matrix} ; q, \frac{z}{a_r} \right] = {}_{r-1}\phi_s \left[\begin{matrix} a_1, \ldots, a_{r-1} \\ b_1, \ldots, b_s \end{matrix} ; q, z \right] \quad [10]$$

$$\lim_{b_s\to\infty} {}_r\phi_s \left[\begin{matrix} a_1, \ldots, a_r \\ b_1, \ldots, b_s \end{matrix} ; q, b_s z \right] = {}_r\phi_{s-1} \left[\begin{matrix} a_1, \ldots, a_r \\ b_1, \ldots, b_{s-1} \end{matrix} ; q, z \right] \quad [11]$$

A terminating q-hypergeometric series $\sum_{k=0}^{n} c_k z^k$ rewritten as $z^n \sum_{k=0}^{n} c_{n-k} z^{-k}$ yields another terminating q-hypergeometric series, for instance:

$${}_{s+1}\phi_s \left[\begin{matrix} q^{-n}, a_1, \ldots, a_s \\ b_1, \ldots, b_s \end{matrix} ; q, z \right]$$
$$= (-1)^n q^{-(1/2)n(n+1)} \frac{(a_1, \ldots, a_n; q)_n}{(b_1, \ldots, b_s; q)_n} z^n$$
$$\times {}_{s+1}\phi_s \left[\begin{matrix} q^{-n}, q^{-n+1}b_1^{-1}, \ldots, q^{-n+1}b_s^{-1} \\ q^{-n+1}a_1^{-1}, \ldots, q^{-n+1}a_s^{-1} \end{matrix} ; q, \frac{q^{n+1}b_1\cdots b_s}{a_1\cdots a_s z} \right] \quad [12]$$

Often, in physics and quantum groups related literature, the following notation is used for q-number, q-factorial, and q-Pochhammer symbol:

$$[a]_q := \frac{q^{(1/2)a} - q^{-(1/2)a}}{q^{1/2} - q^{-1/2}} \quad [k]_q! := \prod_{j=1}^{k} [j]_q$$
$$([a]_q)_k := \prod_{j=0}^{k-1} [a+j]_q \quad (k \in \mathbb{Z}_{\geq 0}) \quad [13]$$

For $q \to 1$, these symbols tend to their classical counterparts without the need for renormalization. They are expressed in terms of the standard notation [1] as follows:

$$[k]_q! = q^{-(1/4)k(k-1)} \frac{(q;q)_k}{(1-q)^k}$$
$$([a]_q)_k = q^{-(1/2)k(a-1)} q^{-(1/4)k(k-1)} \frac{(q^a;q)_k}{(1-q)^k} \quad [14]$$

Special Cases

For $s = r-1$, formula [4] simplifies to

$${}_r\phi_{r-1} \left[\begin{matrix} a_1, \ldots, a_r \\ b_1, \ldots, b_{r-1} \end{matrix} ; q, z \right] = \sum_{k=0}^{\infty} \frac{(a_1, \ldots, a_r; q)_k}{(b_1, \ldots, b_{r-1}; q)_k (q;q)_k} z^k \quad [15]$$

which has radius of convergence 1 in the nonterminating case. The case $r = 2$ of [15] is the q-analog of the Gauss hypergeometric series.

q-Binomial series

$${}_1\phi_0(a; -; q, z) = \sum_{k=0}^{\infty} \frac{(a;q)_k z^k}{(q;q)_k} = \frac{(az;q)_\infty}{(z;q)_\infty}$$
(if series is not terminating, then $|z| < 1$) [16]

q-Exponential series

$$e_q(z) := {}_1\phi_0(0; -; q, z) = \sum_{k=0}^{\infty} \frac{z^k}{(q;q)_k} = \frac{1}{(z;q)_\infty} \quad (|z| < 1) \quad [17]$$

$$E_q(z) := {}_0\phi_0(-; -; q, -z) = \sum_{k=0}^{\infty} \frac{q^{(1/2)k(k-1)} z^k}{(q;q)_k}$$
$$= (-z;q)_\infty = \big(e_q(-z)\big)^{-1} \quad (z \in \mathbb{C}) \quad [18]$$

$$\varepsilon_q(z) := {}_1\phi_1(0; -q^{1/2}; q^{1/2}, -z) = \sum_{k=0}^{\infty} \frac{q^{(1/4)k(k-1)}}{(q;q)_k} z^k \quad (z \in \mathbb{C}) \quad [19]$$

Jackson's *q*-Bessel functions

$$J_\nu^{(1)}(x;q) := \frac{(q^{\nu+1};q)_\infty}{(q;q)_\infty}\left(\frac{1}{2}x\right)^\nu \times {}_2\phi_1\left[\begin{matrix}0,0\\ q^{\nu+1}\end{matrix};q,-\frac{1}{4}x^2\right] \quad (0<x<2) \quad [20]$$

$$J_\nu^{(2)}(x;q) := \frac{(q^{\nu+1};q)_\infty}{(q;q)_\infty}\left(\frac{1}{2}x\right)^\nu {}_0\phi_1\left[\begin{matrix}-\\ q^{\nu+1}\end{matrix};q,-\frac{1}{4}q^{\nu+1}x^2\right] = \left(-\frac{1}{4}x;q\right)_\infty J_\nu^{(1)}(x;q) \quad (x>0) \quad [21]$$

$$J_\nu^{(3)}(x;q) := \frac{(q^{\nu+1};q)_\infty}{(q;q)_\infty}\left(\frac{1}{2}x\right)^\nu \times {}_1\phi_1\left[\begin{matrix}0\\ q^{\nu+1}\end{matrix};q,\frac{1}{4}qx^2\right] \quad (x>0) \quad [22]$$

See [90] for the orthogonality relation for $J_\nu^{(3)}(x;q)$.

If $\exp_q(z)$ denotes one of the three q-exponentials [17]–[19], then $(1/2)(\exp_q(\mathrm{i}x)+\exp_q(-\mathrm{i}x))$ is a q-analog of the cosine and $-(1/2)\mathrm{i}(\exp_q(\mathrm{i}x)-\exp_q(-\mathrm{i}x))$ is a q-analog of the sine. The three q-cosines are essentially the case $\nu=-1/2$ of the corresponding q-Bessel functions [20]–[22], and the three q-sines are essentially the case $\nu=1/2$ of x times the corresponding q-Bessel functions.

q-Derivative and *q*-Integral

The q-derivative of a function f given on a subset of $\mathbb{R}$ or $\mathbb{C}$ is defined by

$$(D_q f)(x) := \frac{f(x)-f(qx)}{(1-q)x} \quad (x\neq 0,\ q\neq 1) \quad [23]$$

where x and qx should be in the domain of f. By continuity, we set $(D_q f)(0) := f'(0)$, provided $f'(0)$ exists. If f is differentiable on an open interval I, then

$$\lim_{q\uparrow 1}(D_q f)(x) = f'(x) \quad (x\in I) \quad [24]$$

For $a\in\mathbb{R}\setminus\{0\}$ and a function f given on $(0,a]$ or $[a,0)$, we define the q-integral by

$$\int_0^a f(x)\,\mathrm{d}_q x := a(1-q)\sum_{k=0}^\infty f(aq^k)\,q^k = \sum_{k=0}^\infty f(aq^k)\,(aq^k-aq^{k+1}) \quad [25]$$

provided the infinite sum converges absolutely (e.g., if f is bounded). If $F(a)$ is given by the left-hand side of [25], then $D_q F = f$. The right-hand side of [25] is an infinite Riemann sum. For $q\uparrow 1$ it converges, at least formally, to $\int_0^a f(x)\,\mathrm{d}x$.

For nonzero $a,b\in\mathbb{R}$ we define

$$\int_a^b f(x)\,\mathrm{d}_q x := \int_0^b f(x)\,\mathrm{d}_q x - \int_0^a f(x)\,\mathrm{d}_q x \quad [26]$$

For a q-integral over $(0,\infty)$, we have to specify a q-lattice $\{aq^k\}_{k\in\mathbb{Z}}$ for some $a>0$ (up to multiplication by an integer power of q):

$$\int_0^{a.\infty} f(x)\,\mathrm{d}_q x := a(1-q)\sum_{k=-\infty}^\infty f(aq^k)\,q^k = \lim_{n\to\infty}\int_0^{q^{-n}a} f(x)\,\mathrm{d}_q x \quad [27]$$

The *q*-Gamma and *q*-Beta Functions

The q-gamma function is defined by

$$\Gamma_q(z) := \frac{(q;q)_\infty\,(1-q)^{1-z}}{(q^z;q)_\infty} \quad (z\neq 0,-1,-2,\ldots) \quad [28]$$

$$= \int_0^{(1-q)^{-1}} t^{z-1}\,E_q(-(1-q)qt)\,\mathrm{d}_q t \quad (\Re z>0) \quad [29]$$

Then

$$\Gamma_q(z+1) = \frac{1-q^z}{1-q}\,\Gamma_q(z) \quad [30]$$

$$\Gamma_q(n+1) = \frac{(q;q)_n}{(1-q)^n} \quad [31]$$

$$\lim_{q\uparrow 1}\Gamma_q(z) = \Gamma(z) \quad [32]$$

The q-beta function is defined by

$$B_q(a,b) := \frac{\Gamma_q(a)\Gamma_q(b)}{\Gamma_q(a+b)} = \frac{(1-q)\,(q,q^{a+b};q)_\infty}{(q^a,q^b;q)_\infty} \quad (a,b\neq 0,-1,-2,\ldots) \quad [33]$$

$$= \int_0^1 t^{b-1}\,\frac{(qt;q)_\infty}{(q^a t;q)_\infty}\,\mathrm{d}_q t \quad (\Re b>0,\ a\neq 0,-1,-2,\ldots) \quad [34]$$

The *q*-Gauss Hypergeometric Series

q-Analog of Euler's integral representation

$${}_2\phi_1(q^a,q^b;q^c;q,z) = \frac{\Gamma_q(c)}{\Gamma_q(a)\Gamma_q(c-b)}\int_0^1 t^{b-1}\,\frac{(tq;q)_\infty}{(tq^{c-b};q)_\infty}\times\frac{(tzq^a;q)_\infty}{tz;q)_\infty}\,\mathrm{d}_q t \quad (\Re b>0,\ |z|<1) \quad [35]$$

By substitution of [25], formula [35] becomes a transformation formula:

$$ {}_2\phi_1(a,b;c;q,z) = \frac{(az;q)_\infty}{(z;q)_\infty}\frac{(b;q)_\infty}{(c;q)_\infty}\,{}_2\phi_1(c/b,z;az;q,b) \quad [36] $$

Note the mixing of argument z and parameters a,b,c on the right-hand side.

Evaluation formulas in special points

$$ {}_2\phi_1(a,b;c;q,c/(ab)) = \frac{(c/a,c/b;q)_\infty}{(c,c/(ab);q)_\infty} \quad (|c/(ab)|<1) \quad [37] $$

$$ {}_2\phi_1(q^{-n},b;c;q,cq^n/b) = \frac{(c/b;q)_n}{(c;q)_n} \quad [38] $$

$$ {}_2\phi_1(q^{-n},b;c;q,q) = \frac{(c/b;q)_n\,b^n}{(c;q)_n} \quad [39] $$

Two general transformation formulas

$$ {}_2\phi_1\left[\begin{matrix}a,b\\c\end{matrix};q,z\right] = \frac{(az;q)_\infty}{(z;q)_\infty}\,{}_2\phi_2\left[\begin{matrix}a,c/b\\c,az\end{matrix};q,bz\right] \quad [40] $$

$$ = \frac{(abz/c;q)_\infty}{(z;q)_\infty}\,{}_2\phi_1\left[\begin{matrix}c/a,c/b\\c\end{matrix};q,\frac{abz}{c}\right] \quad [41] $$

Transformation formulas in the terminating case

$$ {}_2\phi_1\left[\begin{matrix}q^{-n},b\\c\end{matrix};q,z\right] = \frac{(c/b;q)_n}{(c;q)_n}\,{}_3\phi_2\left[\begin{matrix}q^{-n},b,q^{-n}bc^{-1}z\\q^{1-n}bc^{-1},0\end{matrix};q,q\right] \quad [42] $$

$$ = (q^{-n}bc^{-1}z;q)_n\,{}_3\phi_2\left[\begin{matrix}q^{-n},cb^{-1},0\\c,qcb^{-1}z^{-1}\end{matrix};q,q\right] \quad [43] $$

$$ = \frac{(c/b;q)_n}{(c;q)_n}b^n\,{}_3\phi_1\left[\begin{matrix}q^{-n},b,qz^{-1}\\q^{1-n}bc^{-1}\end{matrix};q,\frac{z}{c}\right] \quad [44] $$

Second order q-difference equation

$$ z(q^c-q^{a+b+1}z)(D_q^2u)(z) + \left(\frac{1-q^c}{1-q}-\left(q^b\frac{1-q^a}{1-q}+q^a\frac{1-q^{b+1}}{1-q}\right)z\right)(D_qu)(z) - \frac{1-q^a}{1-q}\frac{1-q^b}{1-q}u(z)=0 \quad [45] $$

Some special solutions of [45] are:

$$ u_1(z) := {}_2\phi_1(q^a,q^b;q^c;q,z) \quad [46] $$

$$ u_2(z) := z^{1-c}\,{}_2\phi_1(q^{1+a-c},q^{1+b-c};q^{2-c};q,z) \quad [47] $$

$$ u_3(z) := z^{-a}\,{}_2\phi_1(q^a,q^{a-c+1};q^{a-b+1};q,q^{-a-b+c+1}z^{-1}) \quad [48] $$

They are related by:

$$ u_1(z) + \frac{(q^a,q^{1-c},q^{c-b};q)_\infty}{(q^{c-1},q^{a-c+1},q^{1-b};q)_\infty} \times \frac{(q^{b-1}z,q^{2-b}z^{-1};q)_\infty}{(q^{b-c}z,q^{c-b+1}z^{-1};q)_\infty}u_2(z) = \frac{(q^{1-c},q^{a-b+1};q)_\infty}{(q^{1-b},q^{a-c+1};q)_\infty} \times \frac{(q^{a+b-c}z,q^{c-a-b+1}z^{-1};q)_\infty z^a}{(q^{b-c}z,q^{c-b+1}z^{-1};q)_\infty}u_3(z) \quad [49] $$

Summation and Transformation Formulas for ${}_r\phi_{r-1}$ Series

An ${}_r\phi_{r-1}$ series [15] is called "balanced" if $b_1\dots b_{r-1} = qa_1\dots a_r$ and $z=q$, and the series is called "very well-poised" if $qa_1=a_2b_1=a_3b_2=\cdots=a_rb_{r-1}$ and $qa_1^{1/2}=a_2=-a_3$. The following more compact notation is used for very well-poised series:

$$ {}_rW_{r-1}(a_1;a_4,a_5,\dots,a_r;q,z) := {}_r\phi_{r-1}\left[\begin{matrix}a_1,qa_1^{1/2},-qa_1^{1/2},a_4,\dots,a_r\\a_1^{1/2},-a_1^{1/2},qa_1/a_4,\dots,qa_1/a_r\end{matrix};q,z\right] \quad [50] $$

Below only a few of the most important identities are given. See Gasper and Rahman (2004) for many more. An important tool for obtaining complicated identities from more simple ones is Bailey's Lemma, which can moreover be iterated (Bailey chain), see Andrews (1986, ch.3).

The q-Saalschütz sum for a terminating balanced ${}_3\phi_2$

$$ {}_3\phi_2\left[\begin{matrix}a,b,q^{-n}\\c,q^{1-n}abc^{-1}\end{matrix};q,q\right] = \frac{(c/a,c/b;q)_n}{(c,c/(ab);q)_n} \quad [51] $$

Jackson's sum for a terminating balanced ${}_8W_7$

$$ {}_8W_7(a;b,c,d,q^{n+1}a^2/(bcd),q^{-n};q,q) = \frac{(qa,qa/(bc),qa/(bd),qa/(cd);q)_n}{(qa/b,qa/c,qa/d,qa/(bcd);q)_n} \quad [52] $$

Watson's transformation of a terminating ${}_8W_7$ into a terminating balanced ${}_4\phi_3$

$$
\begin{aligned}
&{}_8W_7\left(a;b,c,d,e,q^{-n};q,\frac{q^{n+2}a^2}{bcde}\right)\\
&\quad=\frac{(qa,qa/(de);q)_n}{(qa/d,qa/e;q)_n}\\
&\qquad\times{}_4\phi_3\left[\begin{matrix}q^{-n},d,e,qa/(bc)\\ qa/b,qa/c,q^{-n}de/a\end{matrix};q,q\right]
\end{aligned}
\qquad[53]
$$

Sears' transformation of a terminating balanced ${}_4\phi_3$

$$
\begin{aligned}
&{}_4\phi_3\left[\begin{matrix}q^{-n},a,b,c\\ d,e,f\end{matrix};q,q\right]\\
&\quad=\frac{(e/a,f/a;q)_n}{(e,f;q)_n}a^n\,{}_4\phi_3\left[\begin{matrix}q^{-n},a,d/b,d/c\\ d,q^{1-n}a/e,q^{1-n}a/f\end{matrix};q,q\right]
\end{aligned}
\qquad[54]
$$

By iteration and by symmetries in the upper and in the lower parameters, many other versions of this identity can be found. An elegant comprehensive formulation of all these versions is as follows.

Let $x_1x_2x_3x_4x_5x_6=q^{1-n}$. Then the following expression is symmetric in x_1,x_2,x_3,x_4,x_5,x_6:

$$
\begin{aligned}
&\frac{q^{(1/2)n(n-1)}(x_1x_2x_3x_4,x_1x_2x_3x_5,x_1x_2x_3x_6;q)_n}{(x_1x_2x_3)^n}\\
&\quad\times{}_4\phi_3\left[\begin{matrix}q^{-n},x_2x_3,x_1x_3,x_1x_2\\ x_1x_2x_3x_4,x_1x_2x_3x_5,x_1x_2x_3x_6\end{matrix};q,q\right]
\end{aligned}
\qquad[55]
$$

Similar formulations involving symmetry groups can be given for other transformations, see Van der Jeugt and Srinivasa Rao (1999).

Bailey's transformation of a terminating balanced ${}_{10}W_9$

$$
\begin{aligned}
&{}_{10}W_9\left(a;b,c,d,e,f,\frac{q^{n+2}a^3}{bcdef},q^{-n};q,q\right)\\
&\quad=\frac{(qa,qa/(ef),(qa)^2/(bcde),(qa)^2/(bcdf);q)_n}{(qa/e,qa/f,(qa)^2/(bcdef),(qa)^2/(bcd);q)_n}\\
&\qquad\times{}_{10}W_9\left(\frac{qa^2}{bcd};\frac{qa}{cd},\frac{qa}{bd},\frac{qa}{bc},e,f,\frac{q^{n+2}a^3}{bcdef},q^{-n};q,q\right)
\end{aligned}
\qquad[56]
$$

Rogers–Ramanujan Identities

$$
{}_0\phi_1(-;0;q,q)=\sum_{k=0}^{\infty}\frac{q^{k^2}}{(q;q)_k}=\frac{1}{(q,q^4;q^5)_\infty}
\qquad[57]
$$

$$
{}_0\phi_1(-;0;q,q^2)=\sum_{k=0}^{\infty}\frac{q^{k(k+1)}}{(q;q)_k}=\frac{1}{(q^2,q^3;q^5)_\infty}
\qquad[58]
$$

Bilateral Series

Definition [1] can be extended by

$$
(a;q)_k:=\frac{(a;q)_\infty}{(aq^k;q)_\infty}\quad(k\in\mathbb{Z})
\qquad[59]
$$

Define a bilateral q-hypergeometric series by the Laurent series

$$
\begin{aligned}
&{}_r\psi_s\left[\begin{matrix}a_1,\ldots,a_r\\ b_1,\ldots,b_s\end{matrix};q,z\right]={}_r\psi_s(a_1,\ldots,a_r;b_1,\ldots,b_s;q,z)\\
&:=\sum_{k=-\infty}^{\infty}\frac{(a_1,\ldots,a_r;q)_k}{(b_1,\ldots,b_s;q)_k}\left((-1)^kq^{(1/2)k(k-1)}\right)^{s-r}z^k\\
&\quad(a_1,\ldots,a_r,b_1,\ldots,b_s\neq0,\ s\geq r)
\end{aligned}
\qquad[60]
$$

The Laurent series is convergent if $|b_1\ldots b_s/(a_1\ldots a_r)|<|z|$ and moreover, for $s=r$, $|z|<1$.

Ramanujan's ${}_1\psi_1$ summation formula

$$
\begin{aligned}
&{}_1\psi_1(b;c;q,z)\\
&\quad=\frac{(q,c/b,bz,q/(bz);q)_\infty}{(c,q/b,z,c/(bz);q)_\infty}\quad(|c/b|<|z<1)
\end{aligned}
\qquad[61]
$$

This has as a limit case

$${}_0\psi_1(-;c;q,z)=\frac{(q,z,q/z;q)_\infty}{(c,c/z;q)_\infty}\quad(|z|>|c|)
\qquad[62]$$

and as a further specialization the Jacobi triple product identity

$$
\begin{aligned}
&\sum_{k=-\infty}^{\infty}(-1)^k\,q^{(1/2)k(k-1)}\,z^k\\
&\qquad=(q,z,q/z;q)_\infty\quad(z\neq0)
\end{aligned}
\qquad[63]
$$

which can be rewritten as a product formula for a theta function:

$$
\begin{aligned}
\theta_4(x;q)&:=\sum_{k=-\infty}^{\infty}(-1)^k\,q^{k^2}\,\mathrm{e}^{2\pi ikx}\\
&=\prod_{k=1}^{\infty}(1-q^{2k})\\
&\quad\times\left(1-2q^{k-1}\cos(2\pi x)+q^{4k-2}\right)
\end{aligned}
\qquad[64]
$$

q-Hypergeometric Orthogonal Polynomials

Here we discuss families of orthogonal polynomials $\{p_n(x)\}$ which are expressible as terminating q-hypergeometric series $(0<q<1)$ and for which either (1) $P_n(x):=p_n(x)$ or (2) $P_n(x):=p_n$

$((1/2)(x+x^{-1}))$ are eigenfunctions of a second-order q-difference operator, that is,

$$A(x)\,P_n(qx)+B(x)\,P_n(x)+C(x)\,P_n(q^{-1}x) = \lambda_n\,P_n(x) \qquad [65]$$

where $A(x), B(x)$, and $C(x)$ are independent of n, and where the λ_n are the eigenvalues. The generic cases are the four-parameter classes of "Askey–Wilson polynomials" (continuous weight function) and q-Racah polynomials (discrete weights on finitely many points). They are of type (2) (quadratic q-lattice). All other cases can be obtained from the generic cases by specialization or limit transition. In particular, one thus obtains the generic three-parameter classes of type (1) (linear q-lattice). These are the big q-Jacobi polynomials (orthogonality by q-integral) and the q-Hahn polynomials (discrete weights on finitely many points).

Askey–Wilson Polynomials

Definition as q-hypergeometric series

$$p_n(\cos\theta)=p_n(\cos\theta;a,b,c,d\,|\,q) := \frac{(ab,ac,ad;q)_n}{a^n}\,{}_4\phi_3 \times\left[\begin{matrix}q^{-n},q^{n-1}abcd,ae^{i\theta},ae^{-i\theta}\\ ab,ac,ad\end{matrix};q,q\right] \qquad [66]$$

This is symmetric in a,b,c,d.

Orthogonality relation Assume that a,b,c,d are four reals, or two reals and one pair of complex conjugates, or two pairs of complex conjugates. Also assume that $|ab|,|ac|,|ad|,|bc|,|bd|,|cd|<1$. Then

$$\int_{-1}^{1} p_n(x)p_m(x)w(x)\,dx + \sum_k p_n(x_k)\,p_m(x_k)\omega_k = h_n\,\delta_{n,m} \qquad [67]$$

where

$$2\pi\sin\theta\,w(\cos\theta) = \left|\frac{(e^{2i\theta};q)_\infty}{(ae^{i\theta},be^{i\theta},ce^{i\theta},de^{i\theta};q)_\infty}\right|^2 \qquad [68]$$

$$h_0 = \frac{(abcd;q)_\infty}{(q,ab,ac,ad,bc,bd,cd;q)_\infty}$$

$$\frac{h_n}{h_0} = \frac{1-abcdq^{n-1}}{1-abcdq^{2n-1}} \times\frac{(q,ab,ac,ad,bc,bd,cd;q)_n}{(abcd;q)_n} \qquad [69]$$

and the x_k are the points $(1/2)(eq^k+e^{-1}q^{-k})$ with e any of the a,b,c,d of absolute value >1; the sum is over the $k\in\mathbb{Z}_{\geq 0}$ with $|eq^k|>1$. The ω_k are certain weights which can be given explicitly. The sum in [67] does not occur if moreover $|a|,|b|,|c|,|d|<1$.

A more uniform way of writing the orthogonality relation [67] is by the contour integral

$$\frac{1}{2\pi i}\oint_C p_n\left(\tfrac{1}{2}(z+z^{-1})\right)p_m\left(\tfrac{1}{2}(z+z^{-1})\right) \times\frac{(z^2,z^{-2};q)_\infty}{(az,az^{-1},bz,bz^{-1},cz,cz^{-1},dz,dz^{-1};q)_\infty}\,\frac{dz}{z} = 2h_n\delta_{n,m} \qquad [70]$$

where C is the unit circle traversed in positive direction with suitable deformations to separate the sequences of poles converging to zero from the sequences of poles diverging to ∞.

The case $n=m=0$ of [70] or [67] is known as the Askey–Wilson integral.

q-Difference equation

$$A(z)P_n(qz)-\big(A(z)+A(z^{-1})\big)P_n(z)+A(z^{-1})P_n(q^{-1}z) = (q^{-n}-1)\,(1-q^{n-1}abcd)P_n(z) \qquad [71]$$

where $P_n(z)=p_n(\frac{1}{2}(z+z^{-1}))$ and $A(z)=(1-az)(1-bz)(1-cz)\,(1-dz)/((1-z^2)(1-qz^2))$

Special cases These include the continuous q-Jacobi polynomials (two parameters), the continuous q-ultraspherical polynomials (symmetric one-parameter case of continuous q-Jacobi), the Al-Salam-Chihara polynomials (Askey–Wilson with $c=d=0$), and the continuous q-Hermite polynomials (Askey–Wilson with $a=b=c=d=0$).

Continuous q-Ultraspherical Polynomials

Definitions as finite Fourier series and as special Askey–Wilson polynomial

$$C_n(\cos\theta;\beta\,|\,q) := \sum_{k=0}^{n}\frac{(\beta;q)_k(\beta;q)_{n-k}}{(q;q)_k(q;q)_{n-k}}\,e^{i(n-2k)\theta} \qquad [72]$$

$$= \frac{(\beta;q)_n}{(q;q)_n}\,p_n(\cos\theta;\beta^{1/2},q^{1/2}\beta^{1/2},-\beta^{1/2}, -q^{1/2}\beta^{1/2}\,|\,q) \qquad [73]$$

Orthogonality relation $(-1 < \beta < 1)$

$$\frac{1}{2\pi}\int_0^{\pi} C_n(\cos\theta;\beta,q)C_m(\cos\theta;\beta,q)\left|\frac{(e^{2i\theta};q)_\infty}{(\beta e^{2i\theta};q)_\infty}\right|^2 d\theta = \frac{(\beta,q\beta;q)_\infty}{(\beta^2,q;q)_\infty}\frac{1-\beta}{1-\beta q^n}\frac{(\beta^2;q)_n}{(q;q)_n}\delta_{n,m} \quad [74]$$

q-Difference equation

$$A(z)P_n(qz) - \big(A(z)+A(z^{-1})\big)P_n(z) + A(z^{-1})P_n(q^{-1}z) = (q^{-n}-1)(1-q^n\beta^2)P_n(z) \quad [75]$$

where $P_n(z) = C_n(\frac{1}{2}(z+z^{-1});\beta\,|\,q)$ and $A(z) = (1-\beta z^2)(1-q\beta z^2)/((1-z^2)(1-qz^2))$.

Generating function

$$\frac{(\beta e^{i\theta}z, \beta e^{-i\theta}z;q)_\infty}{(e^{i\theta}z, e^{-i\theta}z;q)_\infty} = \sum_{n=0}^{\infty} C_n(\cos\theta;\beta\,|\,q)z^n \quad (|z|<1, 0\le\theta\le\pi, -1<\beta<1) \quad [76]$$

Special case: the continuous q-Hermite polynomials

$$H_n(x\,|\,q) = (q;q)_n\, C_n(x;0\,|\,q) \quad [77]$$

Special cases: the Chebyshev polynomials

$$C_n(\cos\theta;q\,|\,q) = U_n(\cos\theta) := \frac{\sin((n+1)\theta)}{\sin\theta} \quad [78]$$

$$\lim_{\beta\uparrow 1}\frac{(q;q)_n}{(\beta;q)_n}C_n(\cos\theta;\beta\,|\,q) = T_n(\cos\theta) := \cos(n\theta)\ (n>0) \quad [79]$$

q-Racah Polynomials

Definition as q-hypergeometric series $(n = 0, 1, \ldots, N)$

$$R_n(q^{-y}+\gamma\delta q^{y+1};\alpha,\beta,\gamma,\delta\,|\,q) := {}_4\phi_3\left[\begin{matrix} q^{-n}, \alpha\beta q^{n+1}, q^{-y}, \gamma\delta q^{y+1} \\ q\alpha, q\beta\delta, q\gamma \end{matrix}; q, q\right] \quad (\alpha, \beta\delta \text{ or } \gamma = q^{-N-1}) \quad [80]$$

Orthogonality relation

$$\sum_{y=0}^{N} R_n(q^{-y}+\gamma\delta q^{y+1})R_m(q^{-y}+\gamma\delta q^{y+1})\omega_y = h_n\delta_{n,m} \quad [81]$$

where ω_y and h_n can be explicitly given.

Big *q*-Jacobi Polynomials

Definition as q-hypergeometric series

$$P_n(x) = P_n(x;a,b,c;q) := {}_3\phi_2\left[\begin{matrix} q^{-n}, q^{n+1}ab, x \\ qa, qc \end{matrix}; q, q\right] \quad [82]$$

Orthogonality relation

$$\int_{qc}^{qa} P_n(x)P_m(x)\frac{(a^{-1}x, c^{-1}x;q)_\infty}{(x, bc^{-1}x;q)_\infty}d_q x = h_n\,\delta_{n,m}, \quad (0<a<q^{-1}, 0<b<q^{-1}, c<0) \quad [83]$$

where h_n can be explicitly given.

q-Difference equation

$$A(x)P_n(qx) - (A(x)+C(x))P_n(x) + C(x)P_n(q^{-1}x) = (q^{-n}-1)(1-abq^{n+1})P_n(x) \quad [84]$$

where $A(x) = aq(x-1)(bx-c)/x^2$ and $C(x) = (x-qa)(x-qc)/x^2$

Limit case: Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$

$$\lim_{q\uparrow 1} P_n(x;q^\alpha,q^\beta,-q^{-1}d;q) = \frac{n!}{(\alpha+1)_n}P_n^{(\alpha,\beta)}\left(\frac{2x+d-1}{d+1}\right) \quad [85]$$

Special case: the little q-Jacobi polynomials

$$p_n(x;a,b;q) = (-b)^{-n}q^{-(1/2)n(n+1)} \times \frac{(qb;q)_n}{(qa;q)_n}P_n(qbx;b,a,0;q) \quad [86]$$

$$= {}_2\phi_1(q^{-n}, q^{n+1}ab;qa;q,qx) \quad [87]$$

which satisfy orthogonality relation (for $0<a<q^{-1}$ and $b<q^{-1}$)

$$\int_0^1 p_n(x;a,b;q)p_m(x;a,b;q)\frac{(qx;q)_\infty}{(qbx;q)_\infty}x^{\log_q a}\,d_q x = \frac{(q,qab;q)_\infty}{(qa,qb;q)_\infty}\frac{(1-q)(qa)^n}{1-abq^{2n+1}}\frac{(q,qb;q)_n}{(qa,qab;q)_n}\delta_{n,m} \quad [88]$$

Limit case: Jackson's third q-Bessel function (see [22])

$$\lim_{N\to\infty} p_{N-n}(q^{N+k};q^\nu,b;q) = \frac{(q;q)_\infty}{(q^{\nu+1};q)_\infty}q^{-\nu(n+k)} \times J_\nu^{(3)}(2q^{(1/2)(n+k)};q) \quad (\nu>-1) \quad [89]$$

by which [88] tends to the orthogonality relation for $J_\nu^{(3)}(x;q)$:

$$\sum_{k=-\infty}^{\infty} J_\nu^{(3)}(2q^{(1/2)(n+k)};q)J_\nu^{(3)}(2q^{(1/2)(m+k)};q)q^k = \delta_{n,m}q^{-n} \quad (n,m\in\mathbb{Z}) \qquad [90]$$

q-Hahn Polynomials

Definition as *q*-hypergeometric series

$$Q_n(x;\alpha,\beta,N;q) := {}_3\phi_2\left[\begin{matrix}q^{-n},q^{n+1}\alpha\beta,x\\ q\alpha,q^{-N}\end{matrix};q,q\right] \quad (n=0,1,\ldots,N) \qquad [91]$$

Orthogonality relation

$$\sum_{y=0}^{N} Q_n(q^{-y})Q_m(q^{-y})\frac{(q\alpha,q^{-N};q)_y(q\alpha\beta)^{-y}}{(q^{-N}\beta^{-1},q;q)_y} = h_n\delta_{n,m} \qquad [92]$$

where h_n can be explicitly given.

Stieltjes–Wigert Polynomials

Definition as *q*-hypergeometric series

$$S_n(x;q) = \frac{1}{(q;q)_n}{}_1\phi_1\left[\begin{matrix}q^{-n}\\0\end{matrix};q,-q^{n+1}x\right] \qquad [93]$$

The orthogonality measure is not uniquely determined:

$$\int_0^\infty S_n(q^{1/2}x;q)S_m(q^{1/2}x;q)w(x)\,\mathrm{d}x = \frac{1}{q^n(q;q)_n}\delta_{n,m},$$

where, for instance

$$w(x) = \frac{q^{1/2}}{\log(q^{-1})(q,-q^{1/2}x,-q^{1/2}x^{-1};q)_\infty} \quad \text{or} \quad \frac{q^{1/2}}{\sqrt{2\pi\log(q^{-1})}}\exp\left(-\frac{\log^2 x}{2\log(q^{-1})}\right) \qquad [94]$$

Rahman–Wilson Biorthogonal Rational Functions

The following functions are rational in their first argument:

$$R_n(\tfrac{1}{2}(z+z^{-1});a,b,c,d,e) := {}_{10}W_9(a/e;q/(be),q/(ce),q/(de),az,a/z,q^{n-1}abcd,q^{-n};q,q) \qquad [95]$$

They satisfy the biorthogonality relation

$$\frac{1}{2\pi\mathrm{i}}\oint_C R_n\left(\frac{1}{2}(z+z^{-1});a,b,c,d,e\right) \times R_m\left(\frac{1}{2}(z+z^{-1});a,b,c,d,\frac{q}{abcde}\right)w(z)\frac{\mathrm{d}z}{z} = 2h_n\delta_{n,m} \qquad [96]$$

where the contour C is as in [70], and where

$$w(z) = \frac{(z^2,z^{-2},abcdez,abcde/z;q)_\infty}{(az,a/z,bz,b/z,cz,c/z,dz,d/z,ez,e/z;q)_\infty} \qquad [97]$$

$$h_0 = \frac{(bcde,acde,abde,abce,abcd;q)_\infty}{(q,ab,ac,ad,ae,bc,bd,be,cd,ce,de;q)_\infty} \qquad [98]$$

and h_n/h_0 can also be given explicitly. For $ab=q^{-N}, n,m\in\{0,1,\ldots,N\}$, there is a related discrete biorthogonality of the form

$$\sum_{k=0}^{N} R_n\left(\frac{1}{2}(aq^k+a^{-1}q^{-k});a,b,c,d,e\right) \times R_m\left(\frac{1}{2}(aq^k+a^{-1}q^{-k});a,b,c,d,\frac{q}{abcde}\right)w_k = 0 \quad (n\neq m) \qquad [99]$$

Identities and Functions Associated with Root Systems

η-Function Identities

Let R be a root system on a Euclidean space of dimension l. Then Macdonald (1972) generalizes Weyl's denominator formula to the case of an affine root system. The resulting formula can be written as an explicit expansion in powers of q of

$$\prod_{n=1}^{\infty}\left((1-q^n)^l\prod_{\alpha\in R}(1-q^ne^\alpha)\right)$$

which expansion takes the form of a sum over a lattice related to the root system. For root system A_1 this reduces to Jacobi's triple product identity [63]. Macdonald's formula implies a similar expansion in powers of q of $\eta(q)^{l+|R|}$, where $\eta(q)$ is "Dedekind's η-function" $\eta(q) := q^{1/24}(q;q)_\infty$.

Constant Term Identities

Let R be a reduced root system, R^+ the positive roots, and $k\in\mathbb{Z}_{>0}$. Macdonald conjectured the second equality in

$$\frac{\int_T\prod_{\alpha\in R^+}(e^{-\alpha};q)_k(qe^\alpha;q)_k\,\mathrm{d}x}{\int_T\mathrm{d}x} = \mathrm{CT}\left(\prod_{\alpha\in R^+}\prod_{i=1}^{k}(1-q^{i-1}e^{-\alpha})(1-q^ie^\alpha)\right) = \prod_{i=1}^{l}\left[\begin{matrix}kd_i\\k\end{matrix}\right]_q \qquad [100]$$

where T is a torus determined by R, CT means the constant term in the Laurent expansion in e^{α}, and the d_i are the degrees of the fundamental invariants of the Weyl group of R. The conjecture was extended for real $k > 0$, for several parameters k (one for each root length), and for root system BC_n, where Gustafson's five-parameter n-variable analog of the Askey–Wilson integral ([70] for $n = 0$) settles:

$$\int_{[0,2\pi]^n} |\Delta(e^{i\theta_1},\ldots,e^{i\theta_n})|^2 \frac{d\theta_1\ldots\theta_n}{(2\pi)^n} = 2^n n! \times \prod_{j=1}^{n} \frac{(t,t^{n+j-2}abcd;q)_\infty}{(t^j,q,abt^{j-1},act^{j-1},\ldots,cdt^{j-1};q)_\infty} \quad [101]$$

where

$$\Delta(z) := \prod_{1\le i<j\le n} \frac{(z_iz_j, z_i/z_j;q)_\infty}{(tz_iz_j, tz_i/z_j;q)_\infty} \times \prod_{j=1}^{n} \frac{(z_j^2;q)_\infty}{(az_j,bz_j,cz_j,dz_j;q)_\infty} \quad [102]$$

Further extensions were in Macdonald's conjectures for the quadratic norms of Macdonald polynomials associated with root systems (see the subsection "Macdonald–Koornwinder polynomials"), and finally proved by Cherednik.

Macdonald Polynomials for Root System A_{n-1}

Let $n \in \mathbb{Z}_{>0}$. We work with partitions $\lambda = (\lambda_1,\ldots,\lambda_n)$ of length $\le n$, where $\lambda_1 \ge \cdots \ge \lambda_n \ge 0$ are integers. On the set of such partitions, we take the partial order $\lambda \le \mu \Rightarrow \lambda_1 + \cdots + \lambda_n = \mu_1 + \cdots + \mu_n$ and $\lambda_1 + \cdots + \lambda_i \le \mu_1 + \cdots + \mu_i$ $(i = 1,\ldots,n-1)$. Write $\lambda < \mu$ iff $\lambda \le \mu$ and $\lambda \ne \mu$. The monomials are $z^{\alpha} = z_1^{\alpha_1}\ldots z_n^{\alpha_n}$ $(\alpha_1,\ldots,\alpha_n \in \mathbb{Z}_{\ge 0})$. For λ a partition the symmetrized monomials $m_\lambda(z)$ and the Schur functions $s_\lambda(z)$ are defined by:

$$m_\lambda(z) := \sum_{\alpha} z^{\alpha} \quad \text{(sum over all distinct permutations } \alpha \text{ of } (\lambda_1,\ldots,\lambda_n)) \quad [103]$$

$$s_\lambda(z) := \frac{\det(z_i^{\lambda_j+n-j})_{i,j=1,\ldots,n}}{\det(z_i^{n-j})_{i,j=1,\ldots,n}} \quad [104]$$

We integrate a function over the torus $T := \{z \in \mathbb{C}^n \mid |z_1| = \cdots = |z_n| = 1\}$ as

$$\int_T f(z)\,dz := \frac{1}{(2\pi)^n} \times \int_0^{2\pi}\cdots\int_0^{2\pi} f(e^{i\theta_1},\ldots,e^{i\theta_n})\,d\theta_1\ldots d\theta_n \quad [105]$$

Definition For λ a partition and for $0 \le t \le 1$, the (analytically defined) Macdonald polynomial $P_\lambda(z) = P_\lambda(z;q,t)$ is of the form

$$P_\lambda(z) = P_\lambda(z;q,t) = m_\lambda(z) + \sum_{\mu<\lambda} u_{\lambda,\mu} m_\mu(z) \quad (u_{\lambda,\mu} \in \mathbb{C})$$

such that for all $\mu < \lambda$

$$\int_T P_\lambda(z)\,\overline{m_\mu(z)}\,\Delta(z)\,dz = 0$$

where

$$\Delta(z) = \Delta(z;q,t) := \prod_{i\ne j} \frac{(z_iz_j^{-1};q)_\infty}{(tz_iz_j^{-1};q)_\infty} \quad [106]$$

Orthogonality relation

$$\frac{1}{n!}\int_T P_\lambda(z)\,\overline{P_\mu(z)}\,\Delta(z)\,dz = \prod_{i<j} \frac{(q^{\lambda_i-\lambda_j}t^{j-i}, q^{\lambda_i-\lambda_j+1}t^{j-i};q)_\infty}{(q^{\lambda_i-\lambda_j}t^{j-i+1}, q^{\lambda_i-\lambda_j+1}t^{j-i-1};q)_\infty}\delta_{\lambda,\mu} \quad [107]$$

q-Difference equation

$$\sum_{i=1}^{n}\prod_{j\ne i}\frac{tz_i - z_j}{z_i - z_j}\tau_{q,z_i}P_\lambda(z;q,t) = \left(\sum_{i=1}^{n} q^{\lambda_i}t^{n-i}\right)P_\lambda(z;q,t) \quad [108]$$

where τ_{q,z_i} is the q-shift operator: $\tau_{q,z_i}f(z_1,\ldots,z_n) := f(z_1,\ldots,qz_i,\ldots,z_n)$. See (Macdonald 1995, ch. VI, §3) for the full system of q-difference equations.

Special value

$$P_\lambda(1,t,\ldots,t^{n-1};q,t) = \prod_{i=1}^{n} t^{(i-1)\lambda_i} \times \prod_{i<j}\frac{(tq^{j-i};q)_{\lambda_i-\lambda_j}}{(q^{j-i};q)_{\lambda_i-\lambda_j}} \quad [109]$$

Restriction of number of variables

$$P_{\lambda_1,\lambda_2,\ldots,\lambda_{n-1},0}(z_1,\ldots,z_{n-1},0;q,t) = P_{\lambda_1,\lambda_2,\ldots,\lambda_{n-1}}(z_1,\ldots,z_{n-1};q,t) \quad [110]$$

Homogeneity

$$P_{\lambda_1,\ldots,\lambda_n}(z;q,t) = z_1\ldots z_n P_{\lambda_1-1,\ldots,\lambda_n-1}(z;q,t) \quad (\lambda_n > 0) \quad [111]$$

Self-duality Let λ, μ be partitions.

$$\frac{P_\lambda(q^{\mu_1}t^{n-1}, q^{\mu_2}t^{n-2}, \ldots, q^{\mu_n}; q, t)}{P_\lambda(t^{n-1}, t^{n-2}, \ldots, 1; q, t)} = \frac{P_\mu(q^{\lambda_1}t^{n-1}, q^{\lambda_2}t^{n-2}, \ldots, q^{\lambda_n}; q, t)}{P_\mu(t^{n-1}, t^{n-2}, \ldots, 1; q, t)} \quad [112]$$

Special cases and limit relations

Continuous q-ultraspherical polynomials (see [72]):

$$P_{m,n}(re^{i\theta}, re^{-i\theta}; q, t) = \frac{(q;q)_{m-n}}{(t;q)_{m-n}} r^{m+n} \times C_{m-n}(\cos\theta; t \mid q) \quad [113]$$

Symmetrized monomials (see [103]):

$$P_\lambda(z; q, 1) = m_\lambda(z) \quad [114]$$

Schur functions (see [104]):

$$P_\lambda(z; q, q) = s_\lambda(z) \quad [115]$$

Hall–Littlewood polynomials (see Macdonald (1995), ch. III):

$$P_\lambda(z; 0, t) = P_\lambda(z; t) \quad [116]$$

Jack polynomials (see Macdonald (1995), §VI.10):

$$\lim_{q\uparrow 1} P_\lambda(z; q, q^a) = P_\lambda^{(1/a)}(z) \quad [117]$$

Algebraic definition of Macdonald polynomials Macdonald polynomials can also be defined algebraically. We work now with partitions λ ($\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$) of arbitrary length $l(\lambda)$, and with symmetric polynomials in arbitrarily many variables $x_1, x_2, \ldots$, which can be canonically extended to symmetric functions in infinitely many variables $x_1, x_2, \ldots$. The rth power sum p_r and the symmetric functions p_λ are formally defined by

$$p_r = \sum_{i\geq 1} x_i^r, \quad p_\lambda = p_{\lambda_1} p_{\lambda_2} \cdots \quad [118]$$

Put

$$z_\lambda := \prod_{i\geq 1} i^{m_i} m_i! \quad \text{where } m_i = m_i(\lambda) \text{ is the number of parts of } \lambda \text{ equal to } i. \quad [119]$$

Define an inner product $\langle , \rangle_{q,t}$ on the space of symmetric functions such that

$$\langle p_\lambda, p_\mu \rangle_{q,t} = \delta_{\lambda,\mu} z_\lambda \prod_{i=1}^{l(\lambda)} \frac{1-q^{\lambda_i}}{1-t^{\lambda_i}} \quad [120]$$

For partitions λ, μ the partial ordering $\lambda \geq \mu$ means now that $\sum_{j\geq 1} \lambda_j = \sum_{j\geq 1} \mu_j$ and $\lambda_1 + \cdots + \lambda_i \geq \mu_1 + \cdots + \mu_i$ for all i. The Macdonald polynomial $P_\lambda(x; q, t)$ can now be algebraically defined as the unique symmetric function P_λ of the form $P_\lambda = \sum_{\mu\leq\lambda} u_{\lambda,\mu} m_\mu$ ($u_{\lambda,\mu} \in \mathbb{C}, u_{\lambda,\lambda} = 1$) such that

$$\langle P_\lambda, P_\mu \rangle_{q.t} = 0 \quad \text{if } \lambda \neq \mu \quad [121]$$

If $l(\lambda) \leq n$, then the newly defined $P_\lambda(x)$ with $x_{n+1} = x_{n+2} = \cdots = 0$ coincides with $P_\lambda(x; q, t)$ defined analytically, and the new inner product is a constant multiple (depending on n) of the old inner product.

Bilinear sum

$$\sum_\lambda \frac{1}{\langle P_\lambda, P_\lambda \rangle_{q,t}} P_\lambda(x; q, t) P_\lambda(y; q, t) = \prod_{i,j\geq 1} \frac{(tx_i y_j; q)_\infty}{(x_i y_j; q)_\infty} \quad [122]$$

Generalized Kostka numbers The Kostka numbers $K_{\lambda,\mu}$ occurring as expansion coefficients in $s_\lambda = \sum_\mu K_{\lambda,\mu} m_\mu$ were generalized by Macdonald to coefficients $K_{\lambda,\mu}(q, t)$ occurring in connection with Macdonald polynomials, see Macdonald (1995, §VI.8). Macdonald's conjecture that $K_{\lambda,\mu}(q, t)$ is a polynomial in q and t with coefficients in $\mathbb{Z}_{\geq 0}$ was fully proved in Haiman (2001).

Macdonald–Koornwinder Polynomials

Macdonald (2000, 2001) also introduced Macdonald polynomials associated with an arbitrary root system. For root system BC_n this yields a three-parameter family which can be extended to the five-parameter Macdonald–Koornwinder (M–K) polynomials (Koornwinder 1992). They are orthogonal with respect to the measure occurring in [101] with $\Delta(z)$ given by [102]. The M–K polynomials are n-variable analogs of the Askey–Wilson polynomials. All polynomials just discussed tend, for $q \uparrow 1$, to Jacobi polynomials associated with root systems.

Macdonald conjectured explicit expressions for the quadratic norms of the Macdonald polynomials associated with root systems and of the M–K polynomials. These were proved by Cherednik by considering these polynomials as Weyl group symmetrizations of non-invariant polynomials

which are related to double affine Hecke algebras (see Macdonald (2003)).

Elliptic Hypergeometric Series

Let $p, q \in \mathbb{C}, |p|, |q| < 1$. Define a modified Jacobi theta function by

$$\theta(x;p) := (x, p/x; p)_\infty \quad (x \neq 0) \qquad [123]$$

and the elliptic shifted factorial by

$$(a;q,p)_k := \theta(a;p)\theta(aq;p)\dots\theta(aq^{k-1};p)$$
$$(k \in \mathbb{Z}_{>0}), (a;q,p)_0 := 1 \qquad [124]$$

$$(a_1,\dots,a_r;q,p)_k := (a_1;q,p)_k \dots (a_r;q,p)_k \qquad [125]$$

where $a, a_1, \dots, a_r \neq 0$. For $q = e^{2\pi i\sigma}$, $p = e^{2\pi i\tau}$ ($\Im\tau > 0$), and $a \in \mathbb{C}$ we have

$$\frac{\theta(ae^{2\pi i\sigma(x+\sigma^{-1})}; e^{2\pi i\tau})}{\theta(ae^{2\pi i\sigma x}; e^{2\pi i\tau})} = 1$$
$$\frac{\theta(ae^{2\pi i\sigma(x+\tau\sigma^{-1})}; e^{2\pi i\tau})}{\theta(ae^{2\pi i\sigma x}; e^{2\pi i\tau})} = -a^{-1}q^{-x} \qquad [126]$$

A series $\sum_{k=0}^{\infty} c_k$ with c_{k+1}/c_k being an elliptic (i.e., doubly periodic meromorphic) function of k considered as a complex variable is called an elliptic hypergeometric series. In particular, define the ${}_rE_{r-1}$ theta hypergeometric series as the formal series

$${}_rE_{r-1}(a_1,\dots,a_r;b_1,\dots,b_{r-1};q,p;z)$$
$$:= \sum_{k=0}^{\infty} \frac{(a_1,\dots,a_r;q,p)_k}{(b_1,\dots,b_{r-1};q,p)_k} \frac{z^k}{(q;q,p)_k} \qquad [127]$$

It has $g(k) := c_{k+1}/c_k$ with

$$g(x) = \frac{z\theta(a_1q^x;p)\dots\theta(a_rq^x;p)}{\theta(q^{x+1};p)\,\theta(b_1q^x;p)\dots\theta(b_{r-1}q^x;p)}$$

By [126], $g(x)$ is an elliptic function with periods σ^{-1} and $\tau\sigma^{-1}$ ($q = e^{2\pi i\sigma}, p = e^{2\pi i\tau}$) if the balancing condition $a_1\dots a_r = qb_1\dots b_{r-1}$ is satisfied.

The ${}_rV_{r-1}$ very well-poised theta hypergeometric series (a special ${}_rE_{r-1}$) is defined, in case of argument 1, as:

$${}_rV_{r-1}(a_1;a_6,\dots,a_r;q,p)$$
$$:= \sum_{k=0}^{\infty} \frac{\theta(a_1q^{2k};p)}{\theta(a_1;p)} \frac{(a_1,a_6,\dots,a_r;q,p)_k}{(qa_1/a_6,\dots,qa_1/a_r;q,p)_k}$$
$$\times \frac{q^k}{(q;q,p)_k} \qquad [128]$$

The series is called balanced if $a_6^2\dots a_r^2 = a_1^{r-6}q^{r-4}$. The series terminates if, for instance, $a_r = q^{-n}$.

Elliptic Analog of Jackson's ${}_8W_7$ Summation

$${}_{10}V_9(a;b,c,d,q^{n+1}a^2/(bcd),q^{-n};q,p)$$
$$= \frac{(qa,qa/(bc),qa/(bd),qa/(cd);q,p)_n}{(qa/b,qa/c,qa/d,qa/(bcd);q,p)_n} \qquad [129]$$

Elliptic Analog of Bailey's ${}_{10}W_9$ Transformation

$${}_{12}V_{11}\left(a;b,c,d,e,f,\frac{q^{n+2}a^3}{bcdef},q^{-n};q,p\right)$$
$$= \frac{(qa,qa/(ef),(qa)^2/(bcde),(qa)^2/(bcdf);q,p)_n}{(qa/e,qa/f,(qa)^2/(bcdef),(qa)^2/(bcd);q,p)_n}$$
$$\times {}_{12}V_{11}\left(\frac{qa^2}{bcd};\frac{qa}{cd},\frac{qa}{bd},\frac{qa}{bc},e,f,\frac{q^{n+2}a^3}{bcdef},q^{-n};q,p\right) \qquad [130]$$

Suitable ${}_{12}V_{11}$ functions satisfy a discrete biorthogonality relation which is an elliptic analog of [99].

Ruijsenaars' elliptic gamma function

$$\Gamma(z;q,p) := \prod_{j,k=0}^{\infty} \frac{1 - z^{-1}q^{j+1}p^{k+1}}{1 - zq^jp^k} \qquad [131]$$

which is symmetric in p and q. Then

$$\Gamma(qz;q,p) = \theta(z;p)\Gamma(z;q,p)$$
$$\Gamma(q^nz;q,p) = (z;q,p)_n\Gamma(z;q,p) \qquad [132]$$

Applications

Quantum Groups

A specific quantum group is usually a Hopf algebra which is a q-deformation of the Hopf algebra of functions on a specific Lie group or, dually, of a universal enveloping algebra (viewed as Hopf algebra) of a Lie algebra. The general philosophy is that representations of the Lie group or Lie algebra also deform to representations of the quantum group, and that special functions associated with the representations in the classical case deform to q-special functions associated with the representations in the quantum case. Sometimes this is straightforward, but often new subtle phenomena occur.

The representation-theoretic objects which may be explicitly written in terms of q-special functions include matrix elements of representations with respect to specific bases (in particular spherical elements), Clebsch–Gordan coefficients and Racah coefficients. Many one-variable q-hypergeometric functions have found interpretation in some way in connection with a quantum analog of a three-dimensional Lie group (generically the Lie group

$SL(2,\mathbb{C})$ and its real forms). Classical by now are: little q-Jacobi polynomials interpreted as matrix elements of irreducible representations of $SU_q(2)$ with respect to the standard basis; Askey–Wilson polynomials similarly interpreted with respect to a certain basis not coming from a quantum subgroup; Jackson's third q-Bessel functions as matrix elements of irreducible representations of $E_q(2)$; q-Hahn polynomials and q-Racah polynomials interpreted as Clebsch–Gordan coefficients and Racah coefficients, respectively, for $SU_q(2)$.

Further developments include: Macdonald polynomials as spherical elements on quantum analogs of compact Riemannian symmetric spaces; q-analogs of Jacobi functions as matrix elements of irreducible unitary representations of $SU_q(1,1)$; Askey–Wilson polynomials as matrix elements of representations of the SU(2) dynamical quantum group; an interpretation of discrete ${}_{12}V_{11}$ biorthogonality relations on the elliptic U(2) quantum group.

Since the q-deformed Hopf algebras are usually presented by generators and relations, identities for q-special functions involving noncommuting variables satisfying simple relations are important for further interpretations of q-special functions in quantum groups, for instance:

q-Binomial formula with q-commuting variables

$$(x+y)^n = \sum_{k=0}^{n} \begin{bmatrix} n \\ k \end{bmatrix}_q y^{n-k} x^k \quad (xy = qyx) \qquad [133]$$

Functional equations for q-exponentials with $xy = qyx$

$$\begin{aligned} e_q(x+y) &= e_q(y)e_q(x) \\ E_q(x+y) &= E_q(x)E_q(y) \end{aligned} \qquad [134]$$

$$\begin{aligned} e_q(x+y-yx) &= e_q(x)e_q(y) \\ E_q(x+y+yx) &= E_q(y)E_q(x) \end{aligned} \qquad [135]$$

Various Algebraic Settings

Classical groups over finite fields (Chevalley groups) q-Hahn polynomials and various kinds of q-Krawtchouk polynomials have interpretations as spherical and intertwining functions on classical groups (GL_n, SO_n, Sp_n) over a finite field F_q with respect to suitable subgroups, see Stanton (1984).

Affine Kac–Moody algebras (see Lepowsky (1982)) The Rogers–Ramanujan identities [57], [58] and some of their generalizations were interpreted in the context of characters of representations of the simplest affine Kac–Moody algebra $A_1^{(1)}$. Macdonald's generalization of Weyl's denominator formula to affine root systems has an interpretation as an identity for the denominator of the character of a representation of an affine Kac–Moody algebra.

Partitions of Positive Integers

Let n be a positive integer, $p(n)$ the number of partitions of n, $p_N(n)$ the number of partitions of n into parts $\leq N$, $p_{\text{dist}}(n)$ the number of partitions of n into distinct parts, and $p_{\text{odd}}(n)$ the number of partitions of n into odd parts. Then, Euler observed:

$$\frac{1}{(q;q)_\infty} = \sum_{n=0}^{\infty} p(n)q^n \quad \frac{1}{(q;q)_N} = \sum_{n=0}^{\infty} p_N(n)q^n \qquad [136]$$

$$\begin{aligned} (-q;q)_\infty &= \sum_{n=0}^{\infty} p_{\text{dist}}(n)q^n \\ \frac{1}{(q;q^2)_\infty} &= \sum_{n=0}^{\infty} p_{\text{odd}}(n)q^n \end{aligned} \qquad [137]$$

and

$$(-q;q)_\infty = \frac{1}{(q;q^2)_\infty}, \qquad p_{\text{dist}}(n) = p_{\text{odd}}(n) \qquad [138]$$

The Rogers–Ramanujan identity [57] has the following partition-theoretic interpretation: the number of partitions of n with parts differing at least 2 equals the number of partitions of n into parts congruent to 1 or 4 (mod 5). Similarly, [58] yields: the number of partitions of n with parts larger than 1 and differing at least 2 equals the number of partitions of n into parts congruent to 2 or 3 (mod 5).

The left-hand sides of the Rogers–Ramanujan identities [57] and [58] have interpretations in the "hard hexagon model," see Baxter (1982). Much further work has been done on Rogers–Ramanujan-type identities in connection with more general models in statistical mechanics. The so-called "fermionic expressions" do occur.

See also: Combinatorics: Overview; Eight Vertex and Hard Hexagon Models; Hopf Algebras and q-Deformation Quantum Groups; Integrable Systems: Overview; Ordinary Special Functions; Solitons and Kac–Moody Lie Algebras.

Further Reading

Andrews GE (1986) q-Series: Their Development and Application in Analysis, Number Theory, Combinatorics, Physics, and Computer Algebra, *CBMS Regional Conference Series in Mathematics*, vol. 66. Providence, RI: American Mathematical Society.

Andrews GE, Askey R, and Roy R (1999) *Special Functions*. Cambridge: Cambridge University Press.
Andrews GE and Eriksson K (2004) *Integer Partitions*. Cambridge: Cambridge University Press.
Baxter RJ (1982) *Exactly Solved Models in Statistical Mechanics*. London: Academic Press.
Gasper G and Rahman M (2004) *Basic Hypergeometric Series*, 2nd edn. Cambridge: Cambridge University Press.
Haiman M (2001) Hilbert schemes, polygraphs and the Macdonald positivity conjecture. *Journal of the American Mathematical Society* 14: 941–1006.
Koekoek R and Swarttouw RF (1998) *The Askey-Scheme of Hypergeometric Orthogonal Polynomials and Its q-Analogue*. *Report 98-17*, Faculty of Technical Mathematics and Informatics, Delft University of Technology.
Koornwinder TH (1992) Askey–Wilson polynomials for root systems of type *BC*. In: Richards DStP (ed.) *Hypergeometric Functions on Domains of Positivity, Jack Polynomials, and Applications*, Contemp. Math., vol. 138, pp. 189–204. Providence, RI: American Mathematical Society.
Koornwinder TH (1994) Compact quantum groups and *q*-special functions. In: Baldoni V and Picardello MA (eds.) *Representations of Lie Groups and Quantum Groups*, Pitman Research Notes in Mathematics Series, vol. 311, pp. 46–128. Harlow: Longman Scientific & Technical.
Koornwinder TH (1997) Special functions and *q*-commuting variables. In: Ismail MEH, Masson DR, and Rahman M (eds.) *Special Functions, q-Series and Related Topics*, Fields Institute Communications, vol. 14, pp. 131–166. Providence, RI: American Mathematical Society.
Lepowsky J (1982) Affine Lie algebras and combinatorial identities. In: Winter DJ (ed.) *Lie Algebras and Related Topics*, Lecture Notes in Math., vol. 933, pp. 130–156. Berlin: Springer.
Macdonald IG (1972) Affine root systems and Dedekind's η-function. *Inventiones Mathematicae* 15: 91–143.
Macdonald IG (1995) *Symmetric Functions and Hall Polynomials*, 2nd edn. Oxford: Clarendon.
Macdonald IG (2000, 2001) Orthogonal polynomials associated with root systems. *Séminaire Lotharingien de Combinatoire* 45: Art. B45a.
Macdonald IG (2003) *Affine Hecke Algebras and Orthogonal Polynomials*. Cambridge: Cambridge University Press.
Stanton D (1984) Orthogonal polynomials and Chevalley groups. In: Askey RA, Koornwinder TH, and Schempp W (eds.) *Special Functions: Group Theoretical Aspects and Applications*, pp. 87–128. Dordrecht: Reidel.
Suslov SK (2003) *An Introduction to Basic Fourier Series*. Dordrecht: Kluwer Academic Publishers.
Van der Jeugt J and Srinivasa Rao K (1999) Invariance groups of transformations of basic hypergeometric series. *Journal of Mathematical Physics* 40: 6692–6700.
Vilenkin NJ and Klimyk AU (1992) *Representation of Lie Groups and Special Functions*, vol. 3. Dordrecht: Kluwer Academic Publishers.

Quantum 3-Manifold Invariants

C Blanchet, Université de Bretagne-Sud, Vannes, France
V Turaev, IRMA, Strasbourg, France

Introduction

The idea to derive topological invariants of smooth manifolds from partition functions of certain action functionals was suggested by A Schwarz (1978) and highlighted by E Witten (1988). Witten interpreted the Jones polynomial of links in the 3-sphere S^3 as a partition function of the Chern–Simons field theory. Witten conjectured the existence of mathematically defined topological invariants of 3-manifolds, generalizing the Jones polynomial (or rather its values in complex roots of unity) to links in arbitrary closed oriented 3-manifolds. A rigorous construction of such invariants was given by N Reshetikhin and V Turaev (1989) using the theory of quantum groups. The Witten–Reshetikhin–Turaev invariants of 3-manifolds, also called the "quantum invariants," extend to a topological quantum field theory (TQFT) in dimension 3.

Ribbon and Modular Categories

The Reshetikhin–Turaev approach begins with fixing suitable algebraic data, which are best described in terms of monoidal categories. Let $\mathcal{C}$ be a monoidal category (i.e., a category with an associative tensor product and unit object $\mathbb{1}$). A "braiding" in $\mathcal{C}$ assigns to any objects $V, W \in \mathcal{C}$ an invertible morphism $c_{V,W} : V \otimes W \to W \otimes V$ such that, for any $U, V, W \in \mathcal{C}$,

$$c_{U,V\otimes W} = (\mathrm{id}_V \otimes c_{U,W})(c_{U,V} \otimes \mathrm{id}_W)$$
$$c_{U\otimes V,W} = (c_{U,W} \otimes \mathrm{id}_V)(\mathrm{id}_U \otimes c_{V,W})$$

A "twist" in $\mathcal{C}$ assigns to any object $V \in \mathcal{C}$ an invertible morphism $\theta_V : V \to V$ such that, for any $V, W \in \mathcal{C}$,

$$\theta_{V\otimes W} = c_{W,V}\, c_{V,W}(\theta_V \otimes \theta_W)$$

A "duality" in $\mathcal{C}$ assigns to any object $V \in \mathcal{C}$ a "dual" object $V^* \in \mathcal{C}$, and evaluation and co-evaluation morphisms $d_V : V^* \otimes V \to \mathbb{1}$, $b_V : \mathbb{1} \to V \otimes V^*$ such that

$$(\mathrm{id}_V \otimes d_V)(b_V \otimes \mathrm{id}_V) = \mathrm{id}_V$$
$$(d_V \otimes \mathrm{id}_{V^*})(\mathrm{id}_{V^*} \otimes b_V) = \mathrm{id}_{V^*}$$

The category $\mathcal{C}$ with duality, braiding, and twist is ribbon, if for any $V \in \mathcal{C}$,

$$(\theta_V \otimes \mathrm{id}_{V^*})b_V = (\mathrm{id}_V \otimes \theta_{V^*})b_V$$

For an endomorphism $f : V \to V$ of an object $V \in \mathcal{C}$, its trace "$\mathrm{tr}(f) \in \mathrm{End}_{\mathcal{C}}(\mathbb{1})$" is defined as

$$\mathrm{tr}(f) = d_V c_{V,V^*}((\theta_V f) \otimes \mathrm{id}_{V^*})b_V : \mathbb{1} \to \mathbb{1}$$

This trace shares a number of properties of the standard trace of matrices, in particular, $\mathrm{tr}(fg) = \mathrm{tr}(gf)$ and $\mathrm{tr}(f \otimes g) = \mathrm{tr}(f)\mathrm{tr}(g)$. For an object $V \in \mathcal{C}$, set

$$\dim(V) = \mathrm{tr}(\mathrm{id}_V) = d_V c_{V,V^*}(\theta_V \otimes \mathrm{id}_{V^*})b_V$$

Ribbon categories nicely fit the theory of knots and links in S^3. A link $L \subset S^3$ is a closed one-dimensional submanifold of S^3. (A manifold is closed if it is compact and has no boundary.) A link is oriented (resp. framed) if all its components are oriented (resp. provided with a homotopy class of nonsingular normal vector fields). Given a framed oriented link $L \subset S^3$ whose components are labeled with objects of a ribbon category $\mathcal{C}$, one defines a tensor $\langle L \rangle \in \mathrm{End}_{\mathcal{C}}(\mathbb{1})$. To compute $\langle L \rangle$, present L by a plane diagram with only double transversal crossings such that the framing of L is orthogonal to the plane. Each double point of the diagram is an intersection of two branches of L, going over and under, respectively. Associate with such a crossing the tensor $(c_{V,W})^{\pm 1}$ where $V, W \in \mathcal{C}$ are the labels of these two branches and ± 1 is the sign of the crossing determined by the orientation of L. We also associate certain tensors with the points of the diagram where the tangent line is parallel to a fixed axis on the plane. These tensors are derived from the evaluation and co-evaluation morphisms and the twists. Finally, all these tensors are contracted into a single element $\langle L \rangle \in \mathrm{End}_{\mathcal{C}}(\mathbb{1})$. It does not depend on the intermediate choices and is preserved under isotopy of L in S^3. For the trivial knot $O(V)$ with framing 0 and label $V \in \mathcal{C}$, we have $\langle O(V) \rangle = \dim(V)$.

Further constructions need the notion of a tangle. An (oriented) tangle is a compact (oriented) one-dimensional submanifold of $\boldsymbol{R}^2 \times [0,1]$ with endpoints on $\boldsymbol{R} \times 0 \times \{0,1\}$. Near each of its endpoints, an oriented tangle T is directed either down or up, and thus acquires a sign ± 1. One can view T as a morphism from the sequence of ± 1's associated with its bottom ends to the sequence of ± 1's associated with its top ends. Tangles can be composed by putting one on top of the other. This defines a category of tangles $\mathcal{T}$ whose objects are finite sequences of ± 1's and whose morphisms are isotopy classes of framed oriented tangles. Given a ribbon category $\mathcal{C}$, we can consider $\mathcal{C}$-labeled tangles, that is, (framed oriented) tangles whose components are labeled with objects of $\mathcal{C}$. They form a category $\mathcal{T}_{\mathcal{C}}$. Links appear here as tangles without endpoints, that is, as morphisms $\emptyset \to \emptyset$. The link invariant $\langle L \rangle$ generalizes to a functor $\langle \cdot \rangle : \mathcal{T}_{\mathcal{C}} \to \mathcal{C}$.

To define 3-manifold invariants, we need modular categories (Turaev 1994). Let $\boldsymbol{k}$ be a field. A monoidal category $\mathcal{C}$ is $\boldsymbol{k}$-additive if its Hom sets are $\boldsymbol{k}$-vector spaces, the composition and tensor product of the morphisms are bilinear, and $\mathrm{End}_{\mathcal{C}}(\mathbb{1}) = \boldsymbol{k}$. An object $V \in \mathcal{C}$ is simple if $\mathrm{End}_{\mathcal{C}}(V) = \boldsymbol{k}$. A modular category is a $\boldsymbol{k}$-additive ribbon category $\mathcal{C}$ with a finite family of simple objects $\{V_\lambda\}_\lambda$ such that (1) for any object $V \in \mathcal{C}$ there is a finite expansion $\mathrm{id}_V = \sum_i f_i g_i$ for certain morphisms $g_i : V \to V_{\lambda_i}$, $f_i : V_{\lambda_i} \to V$ and (2) the S-matrix $(S_{\lambda,\mu})$ is invertible over $\boldsymbol{k}$ where $S_{\lambda,\mu} = \mathrm{tr}(c_{V_\lambda, V_\mu} c_{V_\mu, V_\lambda})$. Note that $S_{\lambda,\mu} = \langle H(\lambda,\mu) \rangle$ where $H(\lambda,\mu)$ is the oriented Hopf link with framing 0, linking number $+1$, and labels V_λ, V_μ.

Axiom (1) implies that every simple object in $\mathcal{C}$ is isomorphic to exactly one of V_λ. In most interesting cases (when there is a well-defined direct summation in $\mathcal{C}$), this axiom may be rephrased by saying that $\mathcal{C}$ is finite semisimple, that is, $\mathcal{C}$ has a finite set of isomorphism classes of simple objects and all objects of $\mathcal{C}$ are direct sums of simple objects. A weaker version of the axiom (2) yields premodular categories.

The invariant $\langle \cdot \rangle$ of links and tangles extends by linearity to the case where labels are finite linear combinations of objects of $\mathcal{C}$ with coefficients in $\boldsymbol{k}$. Such a linear combination $\Omega = \sum_\lambda \dim(V_\lambda) V_\lambda$ is called the Kirby color. It has the following sliding property: for any object $V \in \mathcal{C}$, the two tangles in Figure 1 yield the same morphism $V \to V$. Here, the dashed line represents an arc on the closed component labeled by Ω. This arc can be knotted or linked with other components of the tangle (not shown in the figure).

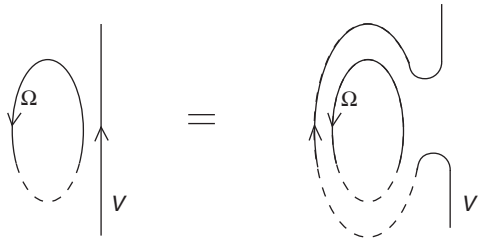


Figure 1 Sliding property.

Invariants of Closed 3-Manifolds

Given an embedded solid torus $g: S^1 \times D^2 \hookrightarrow S^3$, where D^2 is a 2-disk and $S^1 = \partial D^2$, a 3-manifold can be built as follows. Remove from S^3 the interior of $g(S^1 \times D^2)$ and glue back the solid torus $D^2 \times S^1$ along $g|_{S^1 \times S^1}$. This process is known as "surgery." The resulting 3-manifold depends only on the isotopy class of the framed knot represented by g. More generally, a surgery on a framed link $L = \cup_{i=1}^m L_i$ in S^3 with m components yields a closed oriented 3-manifold M_L. A theorem of W Lickorish and A Wallace asserts that any closed connected oriented 3-manifold is homeomorphic to M_L for some L. R Kirby proved that two framed links give rise to homeomorphic 3-manifolds if and only if these links are related by isotopy and a finite sequence of geometric transformations called Kirby moves. There are two Kirby moves: adjoining a distant unknot O^ε with framing $\varepsilon = \pm 1$, and sliding a link component over another one as in Figure 1.

Let $L = \cup_{i=1}^m L_i \subset S^3$ be a framed link and let $(b_{i,j})_{i,j=1,\ldots,m}$ be its linking matrix: for $i \neq j$, $b_{i,j}$ is the linking number of L_i, L_j, and $b_{i,i}$ is the framing number of L_i. Denote by e_+ (resp. e_-) the number of positive (resp. negative) eigenvalues of this matrix. The sliding property of modular categories implies the following theorem. In its statement, a knot K with label Ω is denoted by $K(\Omega)$.

Theorem 1 *Let $\mathcal{C}$ be a modular category with Kirby color Ω. Then $\langle O^1(\Omega)\rangle \neq 0, \langle O^{-1}(\Omega)\rangle \neq 0$ and the expression*

$$\tau_{\mathcal{C}}(M_L) = \langle O^1(\Omega)\rangle^{-e_+} \langle O^{-1}(\Omega)\rangle^{-e_-} \langle L_1(\Omega), \ldots, L_m(\Omega)\rangle$$

is invariant under the Kirby moves on L. This expression yields, therefore, a well-defined topological invariant $\tau_{\mathcal{C}}$ of closed connected oriented 3-manifolds.

Several competing normalizations of $\tau_{\mathcal{C}}$ exist in the literature. Here, the normalization used is such that $\tau_{\mathcal{C}}(S^3) = 1$ and $\tau_{\mathcal{C}}(S^1 \times S^2) = \sum_\lambda (\dim(V_\lambda))^2$. The invariant $\tau_{\mathcal{C}}$ extends to 3-manifolds with a framed oriented $\mathcal{C}$-labeled link K inside by

$$\tau_{\mathcal{C}}(M_L, K) = \langle O^1(\Omega)\rangle^{-e_+} \langle O^{-1}(\Omega)\rangle^{-e_-} \langle L_1(\Omega), \ldots, L_m(\Omega), K\rangle$$

Three-Dimensional TQFTs

A three-dimensional TQFT $\boldsymbol{V}$ assigns to every closed oriented surface X a finite-dimensional vector space $\boldsymbol{V}(X)$ over a field $\boldsymbol{k}$ and assigns to every cobordism (M, X, Y) a linear map $\boldsymbol{V}(M) = \boldsymbol{V}(M, X, Y): \boldsymbol{V}(X) \to \boldsymbol{V}(Y)$. Here, a "cobordism" (M, X, Y) between surfaces X and Y is a compact oriented 3-manifold M with $\partial M = (-X) \amalg Y$ (the minus sign indicates the orientation reversal). A TQFT has to satisfy axioms which can be expressed by saying that $\boldsymbol{V}$ is a monoidal functor from the category of surfaces and cobordisms to the category of vector spaces over $\boldsymbol{k}$. Homeomorphisms of surfaces should induce isomorphisms of the corresponding vector spaces compatible with the action of cobordisms. From the definition, $\boldsymbol{V}(\emptyset) = \boldsymbol{k}$. Every compact oriented 3-manifold M is a cobordism between $\emptyset$ and ∂M so that $\boldsymbol{V}$ yields a "vacuum" vector $\boldsymbol{V}(M) \in \mathrm{Hom}(\boldsymbol{V}(\emptyset), \boldsymbol{V}(\partial M)) = \boldsymbol{V}(\partial M)$. If $\partial M = \emptyset$, then this gives a numerical invariant $\boldsymbol{V}(M) \in \boldsymbol{V}(\emptyset) = \boldsymbol{k}$.

Interestingly, TQFTs are often defined for surfaces and 3-cobordisms with additional structure. The surfaces X are normally endowed with Lagrangians, that is, with maximal isotropic subspaces in $H_1(X; \boldsymbol{R})$. For 3-cobordisms, several additional structures are considered in the literature: for example, 2-framings, p_1-structures, and numerical weights. All these choices are equivalent. The TQFTs requiring such additional structures are said to be "projective" since they provide projective linear representations of the mapping class groups of surfaces.

Every modular category $\mathcal{C}$ with ground field $\boldsymbol{k}$ and simple objects $\{V_\lambda\}_\lambda$ gives rise to a projective three-dimensional TQFT $\boldsymbol{V}_{\mathcal{C}}$. It depends on the choice of a square root $\mathcal{D}$ of $\sum_\lambda (\dim(V_\lambda))^2 \in \boldsymbol{k}$. For a connected surface X of genus g,

$$\boldsymbol{V}_{\mathcal{C}}(X) = \mathrm{Hom}_{\mathcal{C}}\left(\mathbb{1}, \bigoplus_{\lambda_1, \ldots, \lambda_g} \bigotimes_{r=1}^{g} (V_{\lambda_r} \otimes V_{\lambda_r}^*)\right)$$

The dimension of this vector space enters the Verlinde formula

$$\dim_{\boldsymbol{k}}(\boldsymbol{V}_{\mathcal{C}}(X)) \cdot 1_{\boldsymbol{k}} = \mathcal{D}^{2g-2} \sum_\lambda (\dim(V_\lambda))^{2-2g}$$

where $1_{\boldsymbol{k}} \in \boldsymbol{k}$ is the unit of the field $\boldsymbol{k}$. If $\mathrm{char}(\boldsymbol{k}) = 0$, then this formula computes $\dim_{\boldsymbol{k}}(\boldsymbol{V}_{\mathcal{C}}(X))$. For a closed connected oriented 3-manifold M with numerical weight zero, $\boldsymbol{V}_{\mathcal{C}}(M) = \mathcal{D}^{-b_1(M)-1} \tau_{\mathcal{C}}(M)$, where $b_1(M)$ is the first Betti number of M.

The TQFT $\boldsymbol{V}_{\mathcal{C}}$ extends to a vaster class of surfaces and cobordisms. Surfaces may be enriched with a finite set of marked points, each labeled with an object of $\mathcal{C}$ and endowed with a tangent direction. Cobordisms may be enriched with ribbon (or fat) graphs whose edges are labeled with objects of $\mathcal{C}$ and whose vertices are labeled with appropriate intertwiners. The resulting TQFT, also denoted $\boldsymbol{V}_{\mathcal{C}}$, is nondegenerate in the sense that, for any surface X, the vacuum vectors in $\boldsymbol{V}(X)$ determined by all M

with $\partial M = X$ span $V(X)$. A detailed construction of $V_{\mathcal{C}}$ is given in Turaev (1994).

The two-dimensional part of $V_{\mathcal{C}}$ determines a "modular functor" in the sense of G Segal, G Moore, and N Seiberg.

Constructions of Modular Categories

The universal enveloping algebra $U\mathfrak{g}$ of a (finite-dimensional complex) simple Lie algebra $\mathfrak{g}$ admits a deformation $U_q\mathfrak{g}$, which is a quasitriangular Hopf algebra. The representation category $\mathrm{Rep}(U_q\mathfrak{g})$ is $\boldsymbol{C}$-linear and ribbon. For generic $q \in \boldsymbol{C}$, this category is semisimple. (The irreducible representations of $\mathfrak{g}$ can be deformed to irreducible representations of $U_q\mathfrak{g}$.) For q, an appropriate root of unity, a certain subquotient of $\mathrm{Rep}(U_q\mathfrak{g})$ is a modular category with ground field $\boldsymbol{k} = \boldsymbol{C}$. For $\mathfrak{g} = \mathrm{sl}_2(\boldsymbol{C})$, it was pointed out by Reshetikhin and Turaev; the general case involves the theory of tilting modules. The corresponding 3-manifold invariant τ is denoted $\tau_q^{\mathfrak{g}}$. For example, if $\mathfrak{g} = \mathrm{sl}_2(\boldsymbol{C})$ and M is the Poincaré homology sphere (obtained by surgery on a left-hand trefoil with framing –1), then (Le 2003)

$$\tau_q^{\mathfrak{g}}(M) = (1-q)^{-1} \sum_{n \geq 0} q^n (1-q^{n+1}) \times (1-q^{n+2}) \cdots (1-q^{2n+1})$$

The sum here is finite since q is a root of unity.

There is another construction (Le 2003) of a modular category associated with a simple Lie algebra $\mathfrak{g}$ and certain roots of unity q. The corresponding quantum invariant of 3-manifolds is denoted $\tau_q^{P\mathfrak{g}}$. (Here, it is normalized so that $\tau_q^{P\mathfrak{g}}(S^3) = 1$.) Under mild assumptions on the order of q, we have $\tau_q^{\mathfrak{g}}(M) = \tau_q^{\mathfrak{g}}(M)\tau'(M)$ for all M, where $\tau'(M)$ is a certain Gauss sum determined by $\mathfrak{g}$, the homology group $H = H_1(M)$ and the linking form $\mathrm{Tors}\, H \times \mathrm{Tors}\, H \to \boldsymbol{Q}/\boldsymbol{Z}$.

A different construction derives modular categories from the category of framed oriented tangles $\mathcal{T}$. Given a ring K, a bigger category $K[\mathcal{T}]$ can be considered whose morphisms are linear combinations of tangles with coefficients in K. Both $\mathcal{T}$ and $K[\mathcal{T}]$ have a natural structure of a ribbon monoidal category.

The skein method builds ribbon categories by quotienting $K[\mathcal{T}]$ using local "skein" relations, which appear in the theory of knot polynomials (the Alexander–Conway polynomial, the Homfly polynomial, and the Kauffman polynomial). In order to obtain a semisimple category, one completes the quotient category with idempotents as objects (the Karoubi completion). Choosing appropriate skein relations, one can recover the modular categories derived from quantum groups of series A, B, C, D. In particular, the categories determined by the series A arise from the Homfly skein relation shown in Figure 2 where $a, s \in K$. The categories determined by the series B, C, D arise from the Kauffman skein relation.

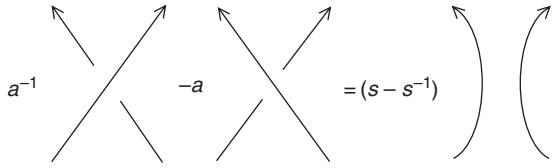


Figure 2 The Homfly relation.

The quantum invariants of 3-manifolds and the TQFTs associated with sl_N can be directly described in terms of the Homfly skein theory, avoiding the language of ribbon categories (W Lickorish, C Blanchet, N Habegger, G Masbaum, P Vogel for sl_2 and Y Yokota for all sl_N).

Unitarity

From both physical and topological viewpoints, one is mainly interested in Hermitian and unitary TQFTs (over $\boldsymbol{k} = \boldsymbol{C}$). A TQFT $\boldsymbol{V}$ is Hermitian if the vector space $\boldsymbol{V}(X)$ is endowed with a nondegenerate Hermitian form $\langle .,. \rangle_X : \boldsymbol{V}(X) \otimes_{\mathbb{C}} \boldsymbol{V}(X) \to \mathbb{C}$ such that:

1. the form $\langle .,. \rangle_X$ is natural with respect to homeomorphisms and multiplicative with respect to disjoint union and
2. for any cobordism (M, X, Y) and any $x \in \boldsymbol{V}(X)$, $y \in \boldsymbol{V}(Y)$,

$$\langle \boldsymbol{V}(M,X,Y)(x), y \rangle_Y = \langle x, \boldsymbol{V}(-M,Y,X)(y) \rangle_X$$

If $\langle .,. \rangle_X$ is positive definite for every X, then the Hermitian TQFT is "unitary." Note two features of Hermitian TQFTs. If $\partial M = \emptyset$, then $\boldsymbol{V}(-M) = \overline{\boldsymbol{V}(M)}$. The group of self-homeomorphisms of any X acts in $\boldsymbol{V}(X)$ preserving the form $\langle .,. \rangle_X$. For a unitary TQFT, this gives an action by unitary matrices.

The three-dimensional TQFT derived from a modular category $\mathcal{V}$ is Hermitian (resp. unitary) under additional assumptions on $\mathcal{V}$ which are discussed briefly. A "conjugation" in $\mathcal{V}$ assigns to each morphism $f : V \to W$ in $\mathcal{V}$ a morphism $\bar{f} : W \to V$ so that

$$\bar{\bar{f}} = f, \quad \overline{f+g} = \bar{f} + \bar{g} \quad \text{for any } f, g : V \to W$$
$$\overline{f \otimes g} = \bar{f} \otimes \bar{g} \quad \text{for any morphisms } f, g \text{ in } \mathcal{C}$$
$$\overline{f \circ g} = \bar{g} \circ \bar{f} \quad \text{for any morphisms } f : V \to W,\ g : W \to V$$

One calls $\mathcal{V}$ Hermitian if it is endowed with conjugation such that

$$\overline{\theta_V} = (\theta_V)^{-1}, \quad \overline{c_{V,W}} = (c_{V,W})^{-1}$$
$$\overline{b_V} = d_V c_{V,V^*}(\theta_V \otimes \mathbf{1}_{V^*})$$
$$\overline{d_V} = (\mathbf{1}_{V^*} \otimes \theta_V^{-1}) c_{V^*,V}^{-1} b_V$$

for any objects V, W of $\mathcal{V}$. A Hermitian modular category $\mathcal{V}$ is unitary if $\mathrm{tr}(f\bar{f}) \geq 0$ for any morphism f in $\mathcal{V}$. The three-dimensional TQFT, derived from a Hermitian (resp. unitary) modular category, has a natural structure of a Hermitian (resp. unitary) TQFT.

The modular category derived from a simple Lie algebra $\mathfrak{g}$ and a root of unity q is always Hermitian. It may be unitary for some q. For simply laced $\mathfrak{g}$, there are always such roots of unity q of any given sufficiently big order. For non-simply-laced $\mathfrak{g}$, this holds under certain divisibility conditions on the order of q.

Integral Structures in TQFTs

The quantum invariants of 3-manifolds have one fundamental property: up to an appropriate rescaling, they are algebraic integers. This was first observed by H Murakami, who proved that $\tau_q^{\mathrm{sl}_2}(M)$ is an algebraic integer, provided the order of q is an odd prime and M is a homology sphere. This extends to an arbitrary closed connected oriented 3-manifold M and an arbitrary simple Lie algebra $\mathfrak{g}$ as follows (Le 2003): for any sufficiently big prime integer r and any primitive rth root of unity q,

$$\tau_q^{P\mathfrak{g}}(M) \in \mathbf{Z}[q] = \mathbf{Z}[\exp(2\pi \mathrm{i}/r)] \qquad [1]$$

This inclusion allows one to expand $\tau_q^{P\mathfrak{g}}(M)$ as a polynomial in q. A study of its coefficients leads to the Ohtsuki invariants of rational homology spheres and further to perturbative invariants of 3-manifolds due to T Le, J Murakami, and T Ohtsuki (see Ohtsuki (2002)). Conjecturally, the inclusion [1] holds for nonprime (sufficiently big) r as well. Connections with the algebraic number theory (specifically modular forms) were studied by D Zagier and R Lawrence.

It is important to obtain similar integrality results for TQFTs. Following P Gilmer, fix a Dedekind domain $D \subset \mathbf{C}$ and call a TQFT V almost D-integral if it is nondegenerate and there is $d \in \mathbf{C}$ such that $dV(M) \in D$ for all M with $\partial M = \emptyset$. Given an almost-integral TQFT V and a surface X, we define $S(\hat{X})$ to be the D-submodule of $V(X)$, generated by all vacuum vectors for X. This module is preserved under the action of self-homeomorphisms of X. It turns out that $S(X)$ is a finitely generated projective D-module and $V(X) = S(X) \otimes_D \mathbf{C}$. A cobordism (M, X, Y) is targeted if all its connected components meet Y along a nonempty set. In this case, $V(M)(S(X)) \subset S(Y)$. Thus, applying S to surfaces and restricting τ to targetet cobordisms, we obtain an "integral version" of V. In many interesting cases, the D-module $S(X)$ is free and its basis may be described explicitly. A simple Lie algebra $\mathfrak{g}$ and a primitive rth (in some cases $4r$th) root of unity q with sufficiently big prime r give rise to an almost D-integral TQFT for $D = \mathbf{Z}[q]$.

State-Sum Invariants

Another approach to three-dimensional TQFTs is based on the theory of $6j$-symbols and state sums on triangulations of 3-manifolds. This approach introduced by V Turaev and O Viro is a quantum deformation of the Ponzano–Regge model for the three-dimensional lattice gravity. The quantum $6j$-symbols derived from representations of $U_q(\mathrm{sl}_2\mathbf{C})$ are $\mathbf{C}$-valued rational functions of the variable $q_0 = q^{1/2}$

$$\begin{vmatrix} i & j & k \\ l & m & n \end{vmatrix} \qquad [2]$$

numerated by 6-tuples of non-negative integers i, j, k, l, m, n. One can think of these integers as labels sitting on the edges of a tetrahedron (see Figure 3). The $6j$-symbol admits various equivalent normalizations and we choose the one which has full tetrahedral symmetry. Now, let $q_0 \in \mathbf{C}$ be a primitive $2r$th root of unity with $r \geq 2$. Set $I = \{0, 1, \ldots, r-2\}$. Given a labeled tetrahedron T as in Figure 3 with $i, j, k, l, m, n \in I$, the $6j$-symbol [2] can be evaluated at q_0 and we can obtain a complex number denoted $|T|$. Consider a closed three-dimensional manifold M with triangulation t. (Note that all 3-manifolds can be triangulated.) A coloring of M is a mapping φ from the set $\mathrm{Edg}(t)$ of the edges of t to I. Set

$$|M| = (\sqrt{2r}/(q_0 - q_0^{-1}))^{-2a} \sum_{\varphi} \prod_{e \in \mathrm{Edg}(t)} \langle \varphi(e) \rangle \prod_{T} |T^{\varphi}|$$

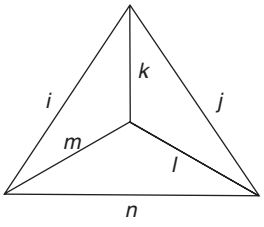


Figure 3 Labeled tetrahedron.

where a is the number of vertices of t, $\langle n\rangle = (-1)^n (q_0^n - q_0^{-n})/(q_0 - q_0^{-1})$ for any integer n, T runs over all tetrahedra of t, and T^φ is T with the labeling induced by φ. It is important to note that $|M|$ *does not depend on the choice of* t *and thus yields a topological invariant of* M.

The invariant $|M|$ is closely related to the quantum invariant $\tau_q^{\mathfrak{g}}(M)$ for $\mathfrak{g} = \mathrm{sl}_2(\mathbf{C})$. Namely, $|M|$ is the square of the absolute value of $\tau_q^{\mathfrak{g}}(M)$, that is, $|M| = |\tau_q^{\mathfrak{g}}(M)|^2$. This computes $|\tau_q^{\mathfrak{g}}(M)|$ inside M without appeal to surgery. No such computation of the phase of $\tau_q^{\mathfrak{g}}(M)$ is known.

These constructions generalize in two directions. First, they extend to manifolds with boundary. Second, instead of the representation category of $U_q(\mathrm{sl}_2\mathbf{C})$, one can use an arbitrary modular category $\mathcal{C}$. This yields a three-dimensional TQFT, which associates to a surface X a vector space $|X|_{\mathcal{C}}$, and to a 3-cobordism (M, X, Y) a homomorphism $|M|_{\mathcal{C}} : |X|_{\mathcal{C}} \to |Y|_{\mathcal{C}}$, (see Turaev (1994)). When $X = Y = \emptyset$, this homomorphism is multiplication $\mathbf{C} \to \mathbf{C}$ by a topological invariant $|M|_{\mathcal{C}} \in \mathbf{C}$. The latter is computed as a state sum on a triangulation of M involving the $6j$-symbols associated with $\mathcal{C}$. In general, these $6j$-symbols are not numbers but tensors so that, instead of their product, one should use an appropriate contraction of tensors. The vectors in $V(X)$ are geometrically represented by trivalent graphs on X such that every edge is labeled with a simple object of $\mathcal{C}$ and every vertex is labeled with an intertwiner between the three objects labeling the incident edges. The TQFT $|\cdot|_{\mathcal{C}}$ is related to the TQFT $V = V_{\mathcal{C}}$ by $|M|_{\mathcal{C}} = |V(M)|^2$. Moreover, for any closed oriented surface X,

$$\begin{aligned} |X|_{\mathcal{C}} &= \mathrm{End}(V(X)) = V(X) \otimes (V(X))^* \\ &= V(X) \otimes V(-X) \end{aligned}$$

and for any three-dimensional cobordism (M, X, Y),

$$\begin{aligned} |M|_{\mathcal{C}} = V(M) \otimes V(-M) &: V(X) \otimes V(-X) \\ &\to V(Y) \otimes V(-Y) \end{aligned}$$

J Barrett and B Westbury introduced a generalization of $|M|_{\mathcal{C}}$ derived from the so-called spherical monoidal categories (which are assumed to be semisimple with a finite set of isomorphism classes of simple objects). This class includes modular categories and a most interesting family of (unitary monoidal) categories arising in the theory of subfactors (see Evans and Kawahigashi (1998) and Kodiyalam and Sunder (2001)). Every spherical category $\mathcal{C}$ gives rise to a topological invariant $|M|_{\mathcal{C}}$ of a closed oriented 3-manifold M. (It seems that this approach has not yet been extended to cobordisms.)

Every monoidal category $\mathcal{C}$ gives rise to a double (or a center) $Z(\mathcal{C})$, which is a braided monoidal category (see Majid (1995)). If $\mathcal{C}$ is spherical, then $Z(\mathcal{C})$ is modular. Conjecturally, $|M|_{\mathcal{C}} = \tau_{Z(\mathcal{C})}(M)$. In the case where $\mathcal{C}$ arises from a subfactor, this has been recently proved by Y Kawahigashi, N Sato, and M Wakui.

The state sum invariants above are closely related to spin networks, spin foam models, and other models of quantum gravity in dimension $2 + 1$ (see Baez (2000) and Carlip (1998)).

See also: Axiomatic Approach to Topological Quantum Field Theory; Braided and Modular Tensor Categories; Chern–Simons Models: Rigorous Results; Finite-type Invariants of 3-Manifolds; Large-*N* and Topological Strings; Schwarz-Type Topological Quantum Field Theory; Topological Quantum Field Theory: Overview; von Neumann Algebras: Subfactor Theory.

Further Reading

Baez JC (2000) *An Introduction to Spin Foam Models of BF Theory and Quantum Gravity, Geometry and Quantum Physics*, Lecture Notes in Physics, No. 543, pp. 25–93. Berlin: Springer.

Bakalov B and Kirillov A Jr. (2001) *Lectures on Tensor Categories and Modular Functors*. University Lecture Series, vol. 21. Providence, RI: American Mathematical Society.

Blanchet C, Habegger N, Masbaum G, and Vogel P (1995) Topological quantum field theories derived from the Kauffman bracket. *Topology* 34: 883–927.

Carlip S (1998) *Quantum Gravity in* $2 + 1$ *Dimensions*,, Cambridge Monographs on Mathematical Physics Cambridge: Cambridge University Press

Carter JS, Flath DE, and Saito M (1995) *The Classical and Quantum 6j-Symbols*. Mathematical Notes, vol. 43. Princeton: Princeton University Press.

Evans D and Kawahigashi Y (1998) *Quantum Symmetries on Operator Algebras*, Oxford Mathematical Monographs, Oxford Science Publications. New York: The Clarendon Press, Oxford University Press.

Kauffman LH (2001) *Knots and Physics*, 3rd edn., Series on Knots and Everything, vol. 1. River Edge, NJ: World Scientific.

Kerler T and Lyubashenko V (2001) *Non-Semisimple Topological Quantum Field Theories for 3-Manifolds with Corners*. Lecture Notes in Mathematics, vol. 1765. Berlin: Springer.

Kodiyalam V and Sunder VS (2001) *Topological Quantum Field Theories from Subfactors*. Research Notes in Mathematics, vol. 423. Boca Raton, FL: Chapman and Hall/CRC Press.

Le T (2003) Quantum invariants of 3-manifolds: integrality, splitting, and perturbative expansion. *Topology and Its Applications* 127: 125–152.

Lickorish WBR (2002) *Quantum Invariants of 3-Manifolds. Handbook of Geometric Topology*, pp. 707–734. Amsterdam: North-Holland.

Majid S (1995) *Foundations of Quantum Group Theory*. Cambridge: Cambridge University Press

Ohtsuki T (2002) *Quantum Invariants. A Study of Knots, 3-Manifolds, and Their Sets*, Series on Knots and Everything, vol. 29. River Edge, NJ: World Scientific.

Turaev V (1994) *Quantum Invariants of Knots and 3- Manifolds*. de Gruyter Studies in Mathematics, vol. 18. Berlin: Walter de Gruyter.

Quantum Calogero–Moser Systems

R Sasaki, Kyoto University, Kyoto, Japan

Introduction

Calogero–Moser (C–M) systems are multiparticle (i.e., finite degrees of freedom) dynamical systems with long-range interactions. They are integrable and solvable at both classical and quantum levels. These systems offer an ideal arena for interplay of many important concepts in mathematical/theoretical physics: to name a few, classical and quantum mechanics, classical and quantum integrability, exact and quasi-exact solvability, addition of discrete (spin) degrees of freedom, quantum Lax pair formalism, supersymmetric quantum mechanics, crystallographic root systems and associated Weyl groups and Lie algebras, noncrystallographic root systems, and Coxeter groups or finite reflection groups. The quantum integrability or solvability of C–M systems does not depend on such known solution mechanisms as Yang–Baxter equations, quantum R-matrix or Bethe ansatz for the quantum systems. In fact, quantum C–M systems provide a good material for pondering about quantum integrability.

Quantum (Liouville) Integrability

The classical Liouville theorem for an integrable system consists of two parts. Let us consider Hamiltonian dynamics of finite degrees of freedom N with coordinates $q=(q_1,\ldots,q_N)$ and conjugate momenta $p=(p_1,\ldots,p_N)$ equipped with Poisson brackets $\{q_j,p_k\}=\delta_{jk},\{q_j,q_k\}=\{p_j,p_k\}=0$. The first part is the existence of a set of independent and involutive $\{K_j,K_k\}=0$ conserved quantities $\{K_j\}$ as many as the degrees of freedom ($j=1,\ldots,N$). The second part asserts that the generating function of the canonical transformation for the action-angle variables can be constructed from the conserved quantities via quadrature. In other words, the second part, that is, the reducibility to the action-angle variables is the integrability. The quantum counterpart of the first half is readily formulated: that is, the existence of a set of independent and mutually commuting (involutive) $[K_j,K_k]=0$ conserved quantities $\{K_j\}$ as many as the degrees of freedom. (This does not necessary imply, however, that they are well defined in a proper Hilbert space.) The definition of the quantum integrability should come as a second part, which is yet to be formulated. It is clear that the quantum Liouville integrability does not imply the complete determination of the eigenvalues and eigenfunctions. Such systems would be called exactly solvable. This can be readily understood by considering any (autonomous) degree-1 Hamiltonian system, which, by definition, is Liouville integrable at the classical and quantum levels. However, it is known that the number of excatly solvable degree-1 Hamiltonians are very limited. What would be the quantum counterpart of the "transformation to action-angle variables by quadrature"? Could it be better formulated in terms of a path integral? Many questions remain to be answered. The quantum C–M systems, an infinite family of exactly solvable multiparticle Hamiltonians, would shed some light on the problem of quantum integrability, in addition to their own beautiful structure explored below.

Throughout this article, the dependence on Planck's constant, $\hbar$, is shown explicitly to distinguish the quantum effects.

Simplest Cases (Based on A_{r-1} Root System)

The simplest example of a C–M system consists of r particles of equal mass (normalized to unity) on a line with pairwise $1/(\text{distance})^2$ interactions described by the following Hamiltonian:

$$\hat{\mathcal{H}}=\frac{1}{2}\sum_{j=1}^{r}p_j^2+g(g-\hbar)\sum_{j<k}^{r}\frac{1}{(q_j-q_k)^2} \qquad [1]$$

in which g is a real positive coupling constant. Here $q=(q_1,\ldots,q_r)$ are the coordinates and $p=(p_1,\ldots,p_r)$ are the conjugate canonical momenta obeying the canonical commutation relations: $[q_j,p_k]=\mathrm{i}\hbar\delta_{jk},[q_j,q_k]=[p_j,p_k]=0,\quad j,k=1,\ldots,r$. The Heisenberg equations of motion are $\dot{q}_j=(\mathrm{i}/\hbar)[\hat{\mathcal{H}},q_j]=p_j$, $\ddot{q}_j=\dot{p}_j=(\mathrm{i}/\hbar)[\hat{\mathcal{H}},p_j]=2g(g-\hbar)\sum_{k\neq j}1/(q_j-q_k)^3$. The repulsive $1/(\text{distance})^2$ potential cannot be surmounted classically or quantum mechanically, and the relative position of the particles on the line is not changed during the time evolution. Classically, it means that if a motion starts at a configuration $q_1>q_2>\cdots>q_r$, then the inequalities remain valid throughout the time evolution. At the quantum level, the wave functions vanish at the boundaries, and the configuration space can be naturally limited to $q_1>q_2>\cdots>q_r$ (the principal Weyl chamber).

Similar integrable quantum many-particle dynamics are obtained by replacing the inverse square potential in [1] by the trigonometric

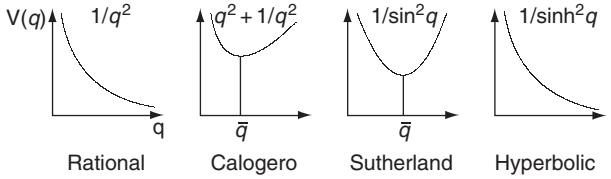


Figure 1 Four different types of quantum C–M potentials.

(hyperbolic) counterpart (see **Figure 1**) $1/(q_j - q_k)^2 \to a^2/\sinh^2 a(q_j - q_k)$, in which $a > 0$ is a real parameter. The $1/\sin^2 q$ potential case (the Sutherland system) corresponds to the $1/(\text{distance})^2$ interaction on a circle of radius $1/2a$, see **Figure 2**. A harmonic confining potential $\omega^2 \sum_{j=1}^{r} q_j^2/2$ can be added to the rational Hamiltonian [1] without breaking the integrability (the Calogero system, see **Figure 1**). At the classical level, the trigonometric (hyperbolic) and rational C–M systems are obtained from the elliptic potential systems (with the Weierstrass $\wp$ function) as the degenerate limits: $\wp(q_1 - q_2) \to a^2/\sinh^2 a(q_1 - q_2) \to 1/(q_1 - q_2)^2$, namely as one (two) period(s) of the $\wp$ function tends to infinity.

It is remarkable that these equations of motion can be expressed in a matrix form (Lax pair): $\mathrm{i}/\hbar[\hat{\mathcal{H}}, L] = \mathrm{d}L/\mathrm{d}t = LM - ML = [L, M] \Leftrightarrow$ Heisenberg equation of motion, in which L and M are given by

$$L = \begin{pmatrix} p_1 & \frac{\mathrm{ig}}{q_1-q_2} & \cdots & \frac{\mathrm{ig}}{q_1-q_r} \\ \frac{\mathrm{ig}}{q_2-q_1} & p_2 & \cdots & \frac{\mathrm{ig}}{q_2-q_r} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{ig}}{q_r-q_1} & \frac{\mathrm{ig}}{q_r-q_2} & \cdots & p_r \end{pmatrix}$$

$$M = \begin{pmatrix} m_1 & -\frac{\mathrm{ig}}{(q_1-q_2)^2} & \cdots & -\frac{\mathrm{ig}}{(q_1-q_r)^2} \\ -\frac{\mathrm{ig}}{(q_2-q_1)^2} & m_2 & \cdots & -\frac{\mathrm{ig}}{(q_2-q_r)^2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\mathrm{ig}}{(q_r-q_1)^2} & -\frac{\mathrm{ig}}{(q_r-q_2)^2} & \cdots & m_r \end{pmatrix} \qquad [2]$$

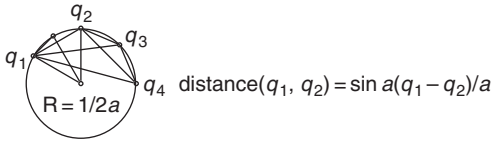


Figure 2 Sutherland potential is $1/(\text{distance})^2$ interaction on a circle. The large-radius limit, $a \to 0$, gives the rational potential.

The diagonal element m_j of M is given by $m_j = \mathrm{ig} \sum_{k \neq j} 1/(q_j - q_k)^2$. The matrix M has a special property $\sum_{j=1}^{r} M_{jk} = \sum_{k=1}^{r} M_{jk} = 0$, which ensures the quantum conserved quantities as the total sum of powers of Lax matrix L: $[\hat{\mathcal{H}}, K_n] = 0$, $K_n \equiv \mathrm{Ts}(L^n) = \sum_{j,k} (L^n)_{jk}$, $(n = 1, 2, 3, \ldots)$, $[K_n, K_m] = 0$. It should be stressed that the trace of L^n is not conserved because of the noncommutativity of q and p. The Hamiltonian is equivalent to $K_2, \hat{\mathcal{H}} \propto K_2 +$ const. In other words, the Lax matrix L is like a "square root" of the Hamiltonian. The quantum equations of motion for the Sutherland and hyperbolic potentials are again expressed by Lax pairs if the following replacements are made: $1/(q_j - q_k) \to a \coth a(q_j - q_k)$ in L and $1/(q_j - q_k)^2 \to a^2/\sinh^2 a(q_j - q_k)$ in M. The quantum conserved quantities are obtained in the same manner as above for the systems with the trigonometric and hyperbolic interactions.

The main goal here is to find all the eigenvalues $\{\mathcal{E}\}$ and eigenfunctions $\{\psi(q)\}$ of the Hamiltonians with the rational, Calogero, Sutherland, and hyperbolic potentials: $\hat{\mathcal{H}}\psi(q) = \mathcal{E}\psi(q)$. The momentum operator p_j acts as differential operators $p_j = -\mathrm{i}\hbar\partial/\partial q_j$. For example, for the rational model Hamiltonian [1], the eigenvalue equation reads

$$\left[-\frac{\hbar^2}{2} \sum_{j=1}^{r} \frac{\partial^2}{\partial q_j^2} + g(g - \hbar) \sum_{j<k}^{r} \frac{1}{(q_j - q_k)^2} \right] \psi(q) = \mathcal{E}\psi(q) \qquad [3]$$

which is a second-order Fuchsian differential equation for each variable $\{q_j\}$ with a regular singularity at each hyperplane $q_j = q_k$ whose exponents are $g/\hbar, 1 - g/\hbar$. Any solution ψ of [3] is regular at all points, except for those on the union of hyperplanes $q_j = q_k$. Since the structure of the singularity is the same for the other three types of potentials, the same assertion for the regularity and singularity of the solution ψ holds for these cases, too. For the trigonometric (Sutherland) case, there are other singularities at $q_j - q_k = l\pi/a$, $l \in \mathbf{Z}$, due to the periodicity of the potential. As is clear from the shape of the potentials, see **Figure 1**, the rational and hyperbolic Hamiltonians have only continuous spectra, whereas the Calogero and Sutherland Hamiltonians have only discrete spectra.

The integrability or more precisely the triangularity of the quantum C–M Hamiltonian was first discovered by Calogero for particles on a line with inverse square potential plus a confining harmonic force and by Sutherland for the particles on a circle

with the trigonometric potential. Later, classical integrability of the models in terms of Lax pairs was proved by Moser. Olshanetsky and Perelomov showed that these systems were based on A_{r-1} root systems, that is, $q_j - q_k = \alpha \cdot q$, and α is one of the root vectors of A_{r-1} root system [13]. They also introduced generalizations of the C–M systems based on any root system including the noncrystallographic ones.

As shown by Heckman–Opdam and Sasaki and collaborators, quantum C–M systems with degenerate potentials (i.e., the rational potentials with/without harmonic force, the hyperbolic, and the trigonometric potentials), based on any root system can be formulated and solved universally. To be more precise, the rational and Calogero systems are integrable for all root systems, the crystallographic and noncrystallographic. The hyperbolic and trigonometric (Sutherland) systems are integrable for any crystallographic root system. The universal formulas for the Hamiltonians, Lax pairs, ground state wave functions, conserved quantities, the triangularity, the discrete spectra for the Calogero and Sutherland systems, the creation and annihilation operators, etc., are equally valid for any root system. This will be shown in the next section. Some rudimentary facts of the root systems and reflections are summarized in the appendix.

Universal Formalism

A C–M system is a Hamiltonian dynamical systems associated with a root system Δ of rank r, which is a set of vectors in $\mathbf{R}^r$ with its standard inner product. A brief review of the properties of the root systems and the associated reflections together with explicit realizations of all the classical root systems will be found in the appendix.

Factorized Hamiltonian

The Hamiltonian for the quantum C–M system can be written in terms of a pre-potential $W(q)$ in a "factorized form":

$$\mathcal{H} = \frac{1}{2}\sum_{j=1}^{r}\left(p_j - \mathrm{i}\frac{\partial W(q)}{\partial q_j}\right)\left(p_j + \mathrm{i}\frac{\partial W(q)}{\partial q_j}\right) \quad [4]$$

The pre-potential is a sum over positive roots:

$$W(q) = \sum_{\alpha\in\Delta_+} g_\alpha \ln|w(\alpha\cdot q)| + \left(-\frac{\omega}{2}q^2\right) \quad [5]$$

The real positive coupling constants g_α are defined on orbits of the corresponding Coxeter group, that is, they are identical for roots in the same orbit. That is, for the simple Lie algebra cases, one coupling constant, $g_\alpha = g$, for all roots in simply laced models and two independent coupling constants, $g_\alpha = g_{\mathrm{L}}$ for long roots and $g_\alpha = g_{\mathrm{S}}$ for short roots, in non-simply laced models. The function $w(u)$ and the other functions $x(u)$ and $y(u)$ appearing in the Lax pair [10], [11] are listed in **Table 1** for each type of degenerate potentials. The dynamics of the prepotentials $W(q)$ (eqn [5]) has been discussed by Dyson from a different point of view (random-matrix model). The above factorized Hamiltonian [4] consists of an operator part $\hat{\mathcal{H}}$, which is the Hamiltonian in the usual definition (see the Hamiltonians in the previous section, e.g., [1]), and a constant $\mathcal{E}_0$ which is the ground-state energy, $\mathcal{H} = \hat{\mathcal{H}} - \mathcal{E}_0$. The factorized Hamiltonian [4] also arises within the context of supersymmetric quantum mechanics.

Table 1 Functions appearing in the prepotential and Lax pair

Potential	$w(u)$	$x(u)$	$y(u)$
Rational	u	$1/u$	$-1/u^2$
Hyperbolic	$\sinh au$	$a\coth au$	$-a^2/\sinh^2 au$
Trigonometric	$\sin au$	$a\cot au$	$-a^2/\sin^2 au$

The pre-potential and the Hamiltonian are invariant under reflection of the phase space variables in the hyperplane perpendicular to any root $W(s_\alpha(q)) = W(q), \mathcal{H}(s_\alpha(p), s_\alpha(q)) = \mathcal{H}(p,q), \forall\alpha \in \Delta$, with s_α defined by [12]. The above Coxeter (Weyl) invariance is the only (discrete) symmetry of the C–M systems. The main problem is, as in the A_{r-1} case, to find all the eigenvalues $\{\mathcal{E}\}$ and eigenfunctions $\{\psi(q)\}$ of the above Hamiltonian $\mathcal{H}\psi(q) = \mathcal{E}\psi(q)$.

For any root system and for any choice of potential, the C–M system has a hard repulsive potential $\sim 1/(\alpha\cdot q)^2$ near the reflection hyperplane $H_\alpha = \{q \in \mathbf{R}^r, \alpha\cdot q = 0\}$. The C–M eigenvalue equation is a second-order Fuchsian differential equation with regular singularities at each reflection hyperplane H_α and those arising from the periodicity in the case of the Sutherland potential. Near the reflection hyperplane H_α, the solution behaves as follows:

$$\psi \sim (\alpha\cdot q)^{g_\alpha/\hbar}(1 + \text{regular terms}), \quad \text{or}$$
$$\psi \sim (\alpha\cdot q)^{1-g_\alpha/\hbar}(1 + \text{regular terms})$$

The former solution is chosen for the square integrability. Because of the singularities, the configuration space is restricted to the principal Weyl chamber PW or the principal Weyl alcove PW_{T} for the trigonometric potential (see **Figure 3**): $\mathrm{PW} = \{q \in \mathbf{R}^r \mid \alpha\cdot q > 0, \alpha \in \Pi\}$, $\mathrm{PW}_{\mathrm{T}} = \{q \in \mathbf{R}^r \mid \alpha\cdot q > 0,$

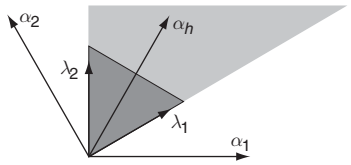


Figure 3 Simple roots, the highest root, fundamental weights, and the principal Weyl alcove (grey) and the principal Weyl chamber (light grey, extending to infinity) in a two-dimensional root system.

$\alpha \in \Pi, \alpha_h \cdot q < \pi/a\}$, ($\Pi$: set of simple roots, see the appendix). Here α_h is the highest root.

Ground-State Wave Function and Energy

One straightforward outcome of the factorized Hamiltonian [4] is the universal ground-state wave function, which is given by

$$\begin{aligned}\Phi_0(q) &= \mathrm{e}^{W(q)/\hbar} \\ &= \prod_{\alpha\in\Delta_+} |w(\alpha\cdot q)|^{g_\alpha/\hbar}\left(\times \mathrm{e}^{-(\omega/2\hbar)q^2}\right) \\ \mathcal{H}\Phi_0(q) &= 0\end{aligned} \qquad [6]$$

The exponential factor $\mathrm{e}^{-(\omega/2\hbar)q^2}$ exists only for the Calogero systems. The ground-state energy, that is, the constant part of $\mathcal{H} = \hat{\mathcal{H}} - \mathcal{E}_0$, has a universal expression for each potential:

$$\mathcal{E}_0 = \begin{cases} 0 & \text{rational} \\ \omega\left(\hbar r/2 + \sum_{\alpha\in\Delta_+} g_\alpha\right) & \text{Calogero} \end{cases} \qquad [7]$$

$$\mathcal{E}_0 = 2a^2\rho^2 \times \begin{cases} -1 & \text{hyperbolic} \\ 1 & \text{Sutherland} \end{cases}$$

where $\rho = 1/2\sum_{\alpha\in\Delta_+} g_\alpha\alpha$ is called a "deformed Weyl vector." Obviously, $\Phi_0(q)$ is square integrable in the configuration spaces for the Calogero and Sutherland systems and not square integrable for the rational and hyperbolic potentials.

Excited States, Triangularity, and Spectrum

Excited states of the C–M systems can be easily obtained as eigenfunctions of a differential operator $\tilde{\mathcal{H}}$ obtained from $\mathcal{H}$ by a similarity transformation:

$$\begin{aligned}\tilde{\mathcal{H}} &= \mathrm{e}^{-W/\hbar}\mathcal{H}\mathrm{e}^{W/\hbar} \\ &= \frac{-1}{2}\sum_{j=1}^{r}(\hbar^2\partial^2/\partial q_j^2 + 2\hbar\partial W/\partial q_j\partial/\partial q_j)\end{aligned}$$

The eigenvalue equation for $\tilde{\mathcal{H}}$, $\tilde{\mathcal{H}}\Psi_\mathcal{E} = \mathcal{E}\Psi_\mathcal{E}$, is then equivalent to that of the original Hamiltonian, $\mathcal{H}\Psi_\mathcal{E}\mathrm{e}^W = \mathcal{E}\Psi_\mathcal{E}\mathrm{e}^W$. Since all the singularities of the Fuchsian differential equation $\mathcal{H}\psi(q) = \mathcal{E}\psi(q)$ are contained in the ground-state wave function e^W, $\Psi_\mathcal{E}$ must be regular at finite q, including all the reflection boundaries. As for the rational and hyperbolic potentials, the energy eigenvalues are only continuous. For the rational case, the eigenfunctions are multivariable generalization of Bessel functions.

Calogero systems The similarity-transformed Hamiltonian $\tilde{\mathcal{H}}$ reads

$$\begin{aligned}\tilde{\mathcal{H}} = \hbar\omega q\cdot\frac{\partial}{\partial q} - \frac{\hbar^2}{2}\sum_{j=1}^{r}\frac{\partial^2}{\partial q_j^2} \\ -\hbar\sum_{\alpha\in\Delta_+}\frac{g_\alpha}{\alpha\cdot q}\alpha\cdot\frac{\partial}{\partial q}\end{aligned} \qquad [8]$$

which maps a Coxeter-invariant polynomial in q of degree d to another of degree d. Thus, the Hamiltonian $\tilde{\mathcal{H}}$ (8) is lower-triangular in the basis of Coxeter-invariant polynomials and the diagonal elements have values as $\hbar\omega \times$ degree, as given by the first term. Independent Coxeter-invariant polynomials exist at the degrees f_j listed in Table 2: $f_j = 1 + e_j$, $j = 1, \ldots, r$, where $\{e_j\}$, $j = 1, \ldots, r$, are the exponents of Δ.

The eigenvalues of the Hamiltonian $\mathcal{H}$ are $\hbar\omega N$ with N a non-negative integer. N can be expressed as $N = \sum_{j=1}^{r} n_j f_j$, $n_j \in \mathbf{Z}_+$, and the degeneracy of the eigenvalue $\hbar\omega N$ is the number of partitions of N. It is remarkable that the coupling constant dependence appears only in the ground-state energy $\mathcal{E}_0$. This is a deformation of the isotropic harmonic oscillator confined in the principal Weyl chamber. The eigenpolynomials are generalization of multivariable Laguerre (Hermite) polynomials. One immediate consequence of this spectrum is the periodicity of the quantum motion. If a system has a wave function $\psi(0)$ at $t = 0$, then at $t = T = 2\pi/\omega$ the system has physically the same wave function as $\psi(0)$, that is, $\psi(T) = \mathrm{e}^{-\mathrm{i}\mathcal{E}_0 T/\hbar}\psi(0)$. The same assertion holds at the classical level, too.

Table 2 The degrees f_j in which independent Coxeter-invariant polynomials exist

Δ	$f_j = 1 + \mathrm{e}_j$	Δ	$f_j = 1 + \mathrm{e}_j$
A_r	2, 3, 4, ..., $r+1$	E_8	2, 8, 12, 14, 18, 20, 24, 30
B_r	2, 4, 6, ..., $2r$	F_4	2, 6, 8, 12
C_r	2, 4, 6, ..., $2r$	G_2	2, 6
D_r	2, 4, ..., $2r-2$, r	$I_2(m)$	2, m
E_6	2, 5, 6, 8, 9, 12	H_3	2, 6, 10
E_7	2, 6, 8, 10, 12, 14, 18	H_4	2, 12, 20, 30

Sutherland Systems The periodicity of the trigonometric potential dictates that the wave function should be a Bloch factor $e^{2ia\mu\cdot q}$ (where μ is a weight) multiplied by a Fourier series in terms of simple roots. The basis of the Weyl invariant wave functions is specified by a dominant weight $\lambda=\sum_{j=1}^{r} m_j\lambda_j$, $m_j\in\mathbf{Z}_+$, $\phi_\lambda(q)\equiv\sum_{\mu\in O_\lambda} e^{2ia\mu\cdot q}$, where O_λ is the orbit of λ by the action of the Weyl group: $O_\lambda=\{g(\lambda)\,|\,g\in G_\Delta\}$. The set of functions $\{\phi_\lambda\}$ has an order $\succ$, $|\lambda|^2>|\lambda'|^2\Rightarrow\phi_\lambda\succ\phi_{\lambda'}$. The similarity-transformed Hamiltonian $\tilde{\mathcal{H}}$ given by

$$\tilde{\mathcal{H}}=-\frac{\hbar^2}{2}\sum_{j=1}^{r}\frac{\partial^2}{\partial q_j^2}-a\hbar\sum_{\alpha\in\Delta_+} g_\alpha\cot(a\alpha\cdot q)\alpha\cdot\frac{\partial}{\partial q} \quad [9]$$

is lower-triangular in this basis: $\tilde{\mathcal{H}}\phi_\lambda=2a^2(\hbar^2\lambda^2+2\hbar\rho\cdot\lambda)\phi_\lambda+\sum_{|\lambda'|<|\lambda|} c_{\lambda'}\phi_{\lambda'}$. That is, the eigenvalue is $\mathcal{E}=2a^2(\hbar^2\lambda^2+2\hbar\rho\cdot\lambda)$ or $\mathcal{E}+\mathcal{E}_0=2a^2(\hbar\lambda+\rho)^2$. Again, the coupling constant dependence comes solely from the deformed Weyl vector ρ. This spectrum is a deformation of the spectrum corresponding to the free motion with momentum $2\hbar a\lambda$ in the principal Weyl alcove. The corresponding eigenfunction is called a generalized Jack polynomial or Heckman–Opdam's Jacobi polynomial. For the rank-2 ($r=2$) root systems, A_2, $B_2\cong C_2$ and $I_2(m)$ (the dihedral group), the complete set of eigenfunctions are known explicitly.

Quantum Lax Pair and Quantum Conserved Quantities

The universal Lax pair for C–M systems is given in terms of the representations of the Coxeter (Weyl) group in stead of the Lie algebra. The Lax operators without spectral parameter for the rational, trigonometric, and hyperbolic potentials are

$$\begin{aligned} L(p,q)&=p\cdot\hat{H}+X(q)\\ X(q)&=\mathrm{i}\sum_{\alpha\in\Delta_+} g_\alpha(\alpha\cdot\hat{H})x(\alpha\cdot q)\hat{s}_\alpha \end{aligned} \quad [10]$$

$$M(q)=\frac{\mathrm{i}}{2}\sum_{\alpha\in\Delta_+} g_\alpha\alpha^2 y(\alpha\cdot q)(\hat{s}_\alpha-I) \quad [11]$$

where I is the identity operator and $\{\hat{s}_\alpha|\alpha\in\Delta\}$ are the reflection operators of the root system. They act on a set of $\mathbf{R}^r$ vectors, $\mathcal{R}=\{\mu^{(k)}\in\mathbf{R}^r\,|\,k=1,\ldots,d\}$, permuting them under the action of the reflection group. The vectors in $\mathcal{R}$ form a basis for the representation space $\mathbf{V}$ of dimension d. The matrix elements of the operators $\{\hat{s}_\alpha\,|\,\alpha\in\Delta\}$ and $\{\hat{H}_j\,|\,j=1,\ldots,r\}$ are defined as follows: $(\hat{s}_\alpha)_{\mu\nu}=\delta_{\mu,s_\alpha(\nu)}=\delta_{\nu,s_\alpha(\mu)}$, $(\hat{H}_j)_{\mu\nu}=\mu_j\delta_{\mu\nu}$, $\alpha\in\Delta$, μ, $\nu\in\mathcal{R}$. The form of the functions x,y depends on the chosen potential as given in Table 1. Then the equations of motion can be expressed in a matrix form $\mathrm{d}L/\mathrm{d}t=\mathrm{i}/\hbar[\mathcal{H},L]=[L,M]$. The operator M satisfies the relation $\sum_{\mu\in\mathcal{R}} M_{\mu\nu}=\sum_{\nu\in\mathcal{R}} M_{\mu\nu}=0$, which is essential for deriving quantum conserved quantities as the total sum (Ts) of all the matrix elements of L^n: $K_n=\mathrm{Ts}(L^n)\equiv\sum_{\mu,\nu\in\mathcal{R}}(L^n)_{\mu\nu}$, $[\mathcal{H},K_n]=0$, $[K_m,K_n]=0$, $n,m=1,\ldots$ In particular, the power 2 is universal to all the root systems, and the quantum Hamiltonian is given by $\mathcal{H}\propto K_2+\mathrm{const}$. As in the affine Toda molecule systems, a Lax pair with a spectral parameter can also be introduced universally for all the above potentials. The Dunkl operators, or the commuting differential–difference operators are also used to construct quantum conserved quantities for some root systems. This method is essentially equivalent to the universal Lax operator formalism. As the Lax operators do not contain the Planck's constant, the quantum Lax pair is essentially of the same form as the classical Lax pair. The difference between the trace (tr) and the total sum (Ts) vanishes as $\hbar\to 0$.

Lax pair for Calogero systems The quantum Lax pair for the Calogero systems is obtained from the universal Lax pair [10] by replacement $L\to L^\pm=L\pm\mathrm{i}\omega Q$, $Q\equiv q\cdot\hat{H}$, which correspond to the creation and annihilation operators of a harmonic oscillator. The equations of motion are rewritten as $\mathrm{d}L^\pm/\mathrm{d}t=\mathrm{i}/\hbar[\mathcal{H},L^\pm]=[L^\pm,M]\pm\mathrm{i}\omega L^\pm$. Then $\mathcal{L}^\pm=L^\pm L^\mp$ satisfy the Lax type equation $\mathrm{d}\mathcal{L}^\pm/\mathrm{d}t=\mathrm{i}/\hbar[\mathcal{H},\mathcal{L}^\pm]$, giving rise to conserved quantities $\mathrm{Ts}(\mathcal{L}^\pm)^n$, $n=1,2,\ldots$ The Calogero Hamiltonian is given by $\mathcal{H}\propto\mathrm{Ts}(\mathcal{L}^\pm)$.

All the eigenstates of the Calogero Hamiltonian $\mathcal{H}$ with eigenvalues $\hbar\omega N$, $N=\sum_{j=1}^{r} n_jf_j$, $n_j\in\mathbf{Z}_+$, are simply constructed in terms of $L^\pm$: $\prod_{j=1}^{r}(B_{f_j}^+)^{n_j}e^{\mathrm{W}}$. Here the integers $\{f_j\}$, $j=1,\ldots,r$, are listed in Table 2. The creation operators $B_{f_j}^+$ and the corresponding annihilation operators $B_{f_j}^-$ are defined by $B_{f_j}^\pm=\mathrm{Ts}(L^\pm)^{f_j}$, $j=1,\ldots,r$. They are Hermitian conjugate to each other $(B_{f_j}^\pm)^\dagger=B_{f_j}^\mp$ with respect to the standard Hermitian inner product of the states defined in PW. They satisfy commutation relations $[\mathcal{H},B_k^\pm]=\pm\hbar k\omega B_k^\pm$, $[B_k^+,B_l^+]=[B_k^-,B_l^-]=0$, $k,l\in\{f_j\,|\,j=1,\ldots,r\}$. The ground state is annihilated by all the annihilation operators $B_{f_j}^-e^{\mathrm{W}}=0, j=1,\ldots,r$.

Further Developments

Rational Potentials: Superintegrability

The systems with the rational potential have a remarkable property: superintegrability. A rational C–M system based on a rank-r root system has $2r-1$

independent conserved quantities. Roughly speaking, they are of the form $K_n = \mathrm{Ts}(L^n), J_m = \mathrm{Ts}(QL^m), Q \equiv q \cdot \hat{H}$, among which only r are involutive. At the classical level, superintegrability can be characterized as algebraic linearizability. Since a commutator of any conserved quantities is again a conserved quantity, these conserved quantities form a nonlinear algebra called a quadratic algebra. It can be considered as a finite-dimensional analog of the W-algebra appearing in certain conformal field theory.

Quantum vs Classical Integrability

In C–M systems, the classical and quantum integrability are very closely related. The quantum discrete spectra of the Calogero and the Sutherland systems are, as shown above, expressed in terms of the coupling constant (ω, g) and the exponents or the weights of the corresponding root systems. Namely, they are integral multiples of coupling constants. The corresponding classical systems with the potential $V(q) = (1/2)\sum_{j=1}^{r} (\partial W(q)/\partial q_j)^2$ share many remarkable properties. As is clear from Figure 1, they always have an equilibrium position. The equilibrium positions $(\bar{q})$ are described by the zeros of a classical orthogonal polynomial; the Hermite polynomial (A-type Calogero), the Laguerre polynomial (B, C, D-type Calogero), the Chebyshev polynomial (A-type Sutherland) and the Jacobi polynomial (B, C, D-type Sutherland). For the exceptional root systems, the corresponding polynomials were not known for a long time. The minimum energy of the classical potential $V(q)$ at the equilibrium is the quantum ground-state energy $\lim_{\hbar \to 0} \mathcal{E}_0$ itself. It is also an integral multiple of coupling constants for both Calogero and Sutherland cases. Near a classical equilibrium, a multiparticle dynamical system is always reduced to a system of coupled harmonic oscillators. For Calogero systems, the eigenfrequencies of these small oscillations are, in fact, exactly the same as the quantum eigenfrequencies, $\omega f_j = \omega(1 + e_j)$. For Sutherland systems, the classical eigenfrequencies are the same as the $o(\hbar)$ part of the quantum spectra corresponding to all the fundamental weights λ_j: $2a^2 \lambda_j \cdot \rho$. Moreover, the eigenvalues of various Lax matrices L and M at the equilibrium take many "interesting values." These results provide ample explicit examples of the general theorem on the quantum–classical correspondence formulated by Loris–Sasaki.

Spin Models

For any root system Δ and an irreducible representation $\mathcal{R}$ of the Coxeter (Weyl) group G_Δ, a spin C–M system can be defined for each of the potentials: rational, Calogero, hyperbolic and Sutherland. For each member μ of $\mathcal{R}$, to be called a "site," a vector space $\mathbf{V}_\mu$ is associated whose element is called a "spin." The dynamical variables are those of the particles $\{q_j, p_j\}$ and the spin exchange operators $\{\hat{\mathcal{P}}_\alpha\}$ $(\alpha \in \Delta)$ which exchange the spins at the sites μ and $s_\alpha(\mu)$. For each Δ and $\mathcal{R}$ a spin exchange model can be defined by "freezing" the particle degrees of freedom at the equilibrium point of the corresponding classical potential $\{q, p\} \to \{\bar{q}, 0\}$. These are generalization of Haldane–Shastry model for Sutherland potentials and that of Polychronakos for the Calogero potentials. Universal Lax pair operators for both spin C–M systems and spin exchange models are known and conserved quantities are constructed.

Integrable Deformations

C–M systems allow various integrable deformations at the classical and/or quantum levels. One of the well-known deformations is the so-called "relativistic" C–M system or the Ruijsenaars–Schneider (R–S) system. For degenerate potentials, they are integrable both at the classical and quantum levels. The classical quantities of the R–S systems at equilibrium exhibit many interesting properties, too. The equilibrium positions are described by the zeros of certain deformation of the above-mentioned classical polynomials. The frequencies of small oscillations are also related to the exact quantum spectrum, and they can be expressed as coupling constant times the (q-) integers.

Inozemtsev models are classically integrable multiparticle dynamical systems related to C–M systems based on classical root systems (A, B, C, D) with additional q^6 (rational) or $\sin^2 2q$ (trigonometric) potentials. Their quantum versions are not exactly solvable in contrast to the C–M or R–S systems, although there is some evidence of their Liouville integrability (without a proper Hilbert space). Quantum Inozemtsev systems can be deformed to be a widest class of quasi-exactly solvable multiparticle dynamical systems. They possess a form of higher-order supersymmetry for which the method of prepotential is also useful.

Appendix: Root Systems

Some rudimentary facts of the root systems and reflections are recapitulated here. The set of roots Δ is invariant under reflections in the hyperplane perpendicular to each vector in Δ. In other words, $s_\alpha(\beta) \in \Delta, \forall \alpha, \beta \in \Delta$, where

$$s_\alpha(\beta) = \beta - (\alpha^\vee \cdot \beta)\alpha, \quad \alpha^\vee \equiv 2\alpha/|\alpha|^2 \qquad [12]$$

The set of reflections $\{s_\alpha \mid \alpha \in \Delta\}$ generates a group G_Δ, known as a Coxeter group, or finite reflection group. The orbit of $\beta \in \Delta$ is the set of root vectors resulting from the action of the Coxeter group on it. The set of positive roots Δ_+ may be defined in terms of a vector $U \in \mathbf{R}^r$, with $\alpha \cdot U \neq 0, \forall \alpha \in \Delta$, as the roots $\alpha \in \Delta$ such that $\alpha \cdot U > 0$. Given Δ_+, there is a unique set of r simple roots $\Pi = \{\alpha_j \mid j = 1, \ldots, r\}$ defined such that they span the root space and the coefficients $\{a_j\}$ in $\beta = \sum_{j=1}^r a_j \alpha_j$ for $\beta \in \Delta_+$ are all non-negative. The highest root α_h, for which $\sum_{j=1}^r a_j$ is maximal, is then also determined uniquely. The subset of reflections $\{s_\alpha \mid \alpha \in \Pi\}$ in fact generates the Coxeter group G_Δ. The products of s_α, with $\alpha \in \Pi$, are subject solely to the relations $(s_\alpha s_\beta)^{m(\alpha,\beta)} = 1$, $\alpha, \beta \in \Pi$. The interpretation is that $s_\alpha s_\beta$ is a rotation in some plane by $2\pi/m(\alpha, \beta)$. The set of positive integers $m(\alpha, \beta)$ (with $m(\alpha, \alpha) = 1, \forall \alpha \in \Pi$) uniquely specifies the Coxeter group. The weight lattice $P(\Delta)$ is defined as the $\mathbf{Z}$-span of the fundamental weights $\{\lambda_j\}$, defined by $\alpha_j^\vee \cdot \lambda_k = \delta_{jk}, \forall \alpha_j \in \Pi$.

The root systems for finite reflection groups may be divided into two types: crystallographic and noncrystallographic. Crystallographic root systems satisfy the additional condition $\alpha^\vee \cdot \beta \in \mathbf{Z}, \forall \alpha, \beta \in \Delta$. The remaining noncrystallographic root systems are H_3, H_4, whose Coxeter groups are the symmetry groups of the icosahedron and four-dimensional 600-cell, respectively, and the dihedral group of order $2m, \{I_2(m) | m \geq 4\}$.

The explicit examples of the classical root systems, that is, A, B, C, and D are given below. For the exceptional and noncrystallographic root systems, the reader is referred to Humphrey's book. In all cases, $\{\boldsymbol{e}_j\}$ denotes an orthonormal basis in $\mathbf{R}^r$.

1. A_{r-1}: This root system is related with the Lie algebra $su(r)$.

$$\Delta = \bigcup_{1 \leq j \leq k \leq r} \{\pm(\boldsymbol{e}_j - \boldsymbol{e}_k)\},$$
$$\prod = \bigcup_{j=1}^{r-1} \{\boldsymbol{e}_j - \boldsymbol{e}_{j+1}\} \qquad [13]$$

2. B_r: This root system is associated with Lie algebra $so(2r+1)$. The long roots have $(\text{length})^2 = 2$ and short roots have $(\text{length})^2 = 1$:

$$\Delta = \bigcup_{1 \leq j \leq k \leq r} \{\pm \boldsymbol{e}_j \pm \boldsymbol{e}_k\} \cup_{j=1}^r \{\pm \boldsymbol{e}_j\}$$
$$\prod = \bigcup_{j=1}^{r-1} \{\boldsymbol{e}_j - \boldsymbol{e}_{j+1}\} \cup \{\boldsymbol{e}_r\} \qquad [14]$$

3. C_r: This root system is associated with Lie algebra $sp(2r)$. The long roots have $(\text{length})^2 = 4$ and short roots have $(\text{length})^2 = 2$:

$$\Delta = \bigcup_{1 \leq j \leq k \leq r} \{\pm \boldsymbol{e}_j \pm \boldsymbol{e}_k\} \cup_{j=1}^r \{\pm 2\boldsymbol{e}_j\}$$
$$\prod = \bigcup_{j=1}^{r-1} \{\boldsymbol{e}_j - \boldsymbol{e}_{j+1}\} \cup \{2\boldsymbol{e}_r\} \qquad [15]$$

4. D_r: This root system is associated with Lie algebra $so(2r)$:

$$\Delta = \bigcup_{1 \leq j \leq k \leq r} \{\pm \boldsymbol{e}_j \pm \boldsymbol{e}_k\}$$
$$\prod = \bigcup_{j=1}^{r-1} \{\boldsymbol{e}_j - \boldsymbol{e}_{j+1}\} \cup \{\boldsymbol{e}_{r-1} + \boldsymbol{e}_r\} \qquad [16]$$

See also: Calogero–Moser–Sutherland Systems of Nonrelativistic and Relativistic Type; Dynamical Systems in Mathematical Physics: An Illustration from Water Waves; Functional Equations and Integrable Systems; Integrable Discrete Systems; Integrable Systems in Random Matrix Theory; Integrable Systems: Overview; Isochronous Systems; Toda Lattices.

Further Reading

Calogero F (1971) Solution of the one-dimensional N-body problem with quadratic and/or inversely quadratic pair potentials. *Journal of Mathematical Physics* 12: 419–436.

Calogero F (2001) *Classical Many-Body Problems Amenable to Exact Treatments*. New York: Springer.

Dunkl C (2001) *Orthogonal Polynomials of Several Variables*. Cambridge: Cambridge University Press.

Humphreys JE (1990) *Reflection Groups and Coxeter Groups*. Cambridge: Cambridge University Press.

Macdonald IG (1995) *Symmetric Functions and Hall Polynomials*, 2nd edn. Oxford University Press.

Moser J (1975) Three integrable Hamiltonian systems connected with isospectral deformations. *Advances in Mathematics* 16: 197–220.

Olshanetsky MA and Perelomov AM (1983) Quantum integrable systems related to Lie algebras. *Physics Reports* 94: 313–404.

Ruijsenaars SNM (1999) Systems of Calogero–Moser Type. *CRM Series in Mathematical Physics* 1: 251–352. Springer.

Sasaki R (2001) *Quantum Calogero–Moser Models: Complete Integrability for All the Root Systems*. Proceedings Quantum Integrable Models and Their Applications, 195–240. World Scientific.

Sasaki R (2002) *Quantum vs Classical Calogero–Moser Systems*. NATO ARW Proceedings, Elba, Italy.

Stanley R (1989) Some combinatorial properties of Jack symmetric function. *Adv. Math.* 77: 76–115.

Sutherland B (1972) Exact results for a quantum many-body problem in one dimension. II. *Physical Review A* 5: 1372–1376.

Quantum Central-Limit Theorems

A F Verbeure, Institute for Theoretical Physics, KU Leuven, Belgium

Introduction

Statistical physics deals with systems with many degrees of freedom and the problems concern finding procedures for the extraction of relevant physical quantities for these extremely complex systems. The idea is to find relevant reduction procedures which map the complex systems onto simpler, tractable models at the price of introducing elements of uncertainty. Therefore, probability theory is a natural mathematical tool in statistical physics. Since the early days of statistical physics, in classical (Newtonian) physical systems, it is natural to model the observables by a collection of random variables acting on a probability space. Kolmogorovian probability techniques and results are the main tools in the development of classical statistical physics. A random variable is usually considered as a measurable function with expectation given as its integral with respect to a probability measure. Alternatively, a random variable can also be viewed as a multiplication operator by the associated function. Different random variables commute as multiplication operators, and one speaks of a commutative probabilistic model.

Now, looking at genuine quantum systems, in many cases the procedure mentioned above leads to commutative probabilistic models, but there exist the realms of physics where quantum noncommutative probabilistic concepts are unavoidable. Typical examples of such areas are quantum optics, low-temperature solid-state physics and ground-state physics such as quantum field theory. During the last 50 years physicists have developed more or less heuristic methods to deal with, for example, manifestations of fluctuations of typical quantum nature. In the last 30 years, mathematical foundations of such theories were also formulated, and a notion of quantum probability was launched as a branch of mathematical physics and mathematics (Cushen and Hudson 1971, Fannes and Quaegebeur 1983, Quaegebeur 1984, Hudson 1973, Giri and von Waldenfels 1978).

The aim of this article is to review briefly a few selected rigorous results concerning noncommutative limit theorems. This choice is made not only because of the author's interest but also for its close relation to concrete problems in statistical physics where one aims at understanding the macroscopic phenomena on the basis of the microscopic structure. A precise definition or formulation of a microscopic and a macroscopic system is of prime importance. The so-called algebraic approach of dynamical systems (Brattelli and Robinson 1979 and 2002) offers the necessary generality and mathematical framework to deal with classical and quantum, microscopic and macroscopic, finite and infinite systems. The observables of any system are assumed to be elements of an (C^*- or von Neumann) algebra $\mathcal{A}$, and the physical states are given by positive linear normalized functionals ω of $\mathcal{A}$, mapping the observables on their expectation values.

A common physicist's belief is that the macroscopic behavior of an idealized infinite system is described by a reduced set of macroscopic quantities (Sewell 1986). Some examples of these are the average densities of particles, energy, momentum, magnetic moment, etc. Analogously as the microscopic quantities, the macroscopic observables should be elements of an algebra, and macroscopic states of the system should be states on this algebra. The main problem is to construct the precise mathematical procedures to go from a given microscopic system to its macroscopic systems.

A well-known macroscopic system is the one given by the algebra of the observables at infinity (Lanford and Ruelle 1969) containing the spacial averages of local micro-observables, that is, for any local observable A one considers the observable

$$A_\omega = \omega\_\lim_{V\to\infty} \frac{1}{V}\int_V \mathrm{d}x\, \tau_x A$$

where V is any finite volume in $\mathbb{R}^\nu$ and τ_x the translation over $x \in \mathbb{R}^\nu$, and where $\omega\_\lim$ is the weak operator limit in the microstate ω. The limits A_ω obtained correspond to the law of large numbers in probability. The algebra generated by these limit observables $\mathcal{A}_\omega = \{A_\omega \mid A \in \mathcal{A}\}$ is an abelian algebra of observables of a macroscopic system. This algebra can be identified with an algebra with pointwise product of measurable functions for some measure or macroscopic state.

The content of this review is to describe an analogous mapping from micro to macro but for a different type of scaling, namely the scaling of fluctuations. For any local observable $A \in \mathcal{A}$, one considers the limit

$$\lim_V \frac{1}{V^{1/2}}\int_V \mathrm{d}x(\tau_x A - \omega(\tau_x A)) \equiv F(A)$$

The problem consists in characterizing the $F(A)$ as an operator on a Hilbert space, called fluctuation

operator, and to specify the algebraic character of the set of all of these.

Based on this quantum central-limit theorem, one notes that not all locally different microscopic observables always yield different fluctuation operators. Hence the central-limit theorem realizes a well-defined procedure of coarse graining or reduction procedure which is handled by the mathematical notion of an equivalence relation on the microscopic observables yielding the same fluctuation operator.

In the following sections we discuss the preliminaries, the basic results about normal and abnormal fluctuations. Three model-independent applications are also discussed. In this review, we omit the properties of the so-called modulated fluctuations.

One should remark that we discuss only fluctuations in space. One can also consider timelike fluctuations. The theory of fluctuation operators for these has not been explicitly worked out so far. However, it is clear that for normal fluctuations the clustering properties of the time correlation functions will play a crucial role. On the other hand, typical properties of the structure of this fluctuation algebra may come up.

Another point which one has to stress is that all systems, which are treated in this review, are quasilocal systems. Other systems, for example, fermion systems, are note treated. But, in particular, fermion systems share many properties of quasilocality, and many of the results mentioned hold true also for fermion systems.

Preliminaries

Quantum Lattice Systems

Although all results we review can be extented to continuous or more general systems, modulo some technicalities, we limit ourself to quasilocal quantum dynamical lattice systems.

We consider the quasilocal algebra built on a ν-dimensional lattice $\mathbb{Z}^\nu$. Let $\mathcal{D}(\mathbb{Z}^\nu)$ be the directed set of finite subsets of $\mathbb{Z}^\nu$ where the direction is the inclusion. With each point $x \in \mathbb{Z}^\nu$ we associate an algebra (C^*- or von Neumann algebra) $\mathcal{A}_x$, all copies of an algebra $\mathcal{A}$. For all $\Lambda \in \mathcal{D}(\mathbb{Z}^\nu)$, the tensor product $\otimes_{x\in\Lambda}\mathcal{A}_x$ is denoted by $\mathcal{A}_\Lambda$. We take $\mathcal{A}$ to be nuclear, then there exists a unique C^*-norm on $\mathcal{A}_\Lambda$. Every copy $\mathcal{A}_x$ is naturally embedded in $\mathcal{A}_\Lambda$. The family $\{\mathcal{A}_\Lambda\}_{\Lambda\in\mathcal{D}(\mathbb{Z}^\nu)}$ has the usual relations of locality and isotony:

$$[\mathcal{A}_{\Lambda_1}, \mathcal{A}_{\Lambda_2}] = 0 \quad \text{if } \Lambda_1 \cap \Lambda_2 = \emptyset \qquad [1]$$

$$\mathcal{A}_{\Lambda_1} \subseteq \mathcal{A}_{\Lambda_2} \quad \text{if } \Lambda_1 \subseteq \Lambda_2 \qquad [2]$$

Denote by $\mathcal{A}_\mathrm{L}$ all local observables, that is,

$$\mathcal{A}_\mathrm{L} = \bigcup_\Lambda \mathcal{A}_\Lambda$$

This algebra is naturally equipped with a C^*-norm $\|\cdot\|$ and its closure

$$\mathcal{B} = \overline{\mathcal{A}}_\mathrm{L}$$

is called a quasilocal C^*-algebra and considered as the microscopic algebra of observables of the system. Typical examples are spin systems where $\mathcal{A} = M_n$ is the $n \times n$ complex matrix algebra. In this case, every state ω of $\mathcal{B}$ is then locally normal, that is, there exists a family of density matrices $\{\rho_\Lambda \mid \Lambda \in \mathcal{D}(\mathbb{Z}^\nu)\}$ such that

$$\omega(A) = \mathrm{tr}\,\rho_\Lambda A \quad \text{for all } A \in \mathcal{A}_\Lambda$$

An important group of $*$-automorphisms of $\mathcal{B}$ is the group of space translations $\{\tau_x, x \in \mathbb{Z}^\nu\}$:

$$\tau_x : A_y \in \mathcal{A}_y \to \tau_x A_y = A_{x+y} \in \mathcal{A}_{y+x}$$

for all $A \in \mathcal{A}$.

Note that the quasilocal algebra $\mathcal{B}$ is asymptotically abelian for space translations: that is, for all $A, B \in \mathcal{B}$

$$\lim_{|x|\to\infty} \|[A, \tau_x B]\| = 0$$

A state ω of $\mathcal{B}$ represents a physical state of the system, assigning to every observable A its expectation value $\omega(A)$. Therefore, this setting can be viewed as the quantum analog of the classical probabilistic setting. Sequences of random variables or observables can be constructed by considering an observable and its translates, that is, $\tau_x(A)_{x\in\mathbb{Z}^\nu}$ is a noncommutative random field. If a state ω is translation invariant, that is, $\omega \circ \tau_x = \omega$ for all x, then all $\tau_x(A)$ are identically distributed random variables. The mixing property of the random field is then expressed by the spatial correlations tending to zero:

$$\omega\big(\tau_x(A)\tau_y(B)\big) - \omega(\tau_x(A))\omega\big(\tau_y(B)\big) \to 0 \qquad [3]$$

if $|x - y| \to \infty$.

One of the basic limit theorems of probability theory is the weak law of large numbers. In this noncommutative setting the law of large numbers is translated into the problem of the convergence of space averages of an observable $A \in \mathcal{B}$. A first result was given by the mean ergodic theorem of von Neumann (1929). In Brattelli and Robinson (1979, 2002) one finds the following theorem: if the state ω is space translation invariant and mixing (see [3]) then for all A, B, and C in $\mathcal{B}$

$$\lim_{\Lambda\to\mathbb{Z}^\nu} \omega\left(A\frac{1}{|\Lambda|}\left(\sum_{x\in\Lambda}\tau_x(B)\right)C\right) = \omega(AC)\omega(B) \qquad [4]$$

That is, in the GNS (Gelfand–Naimark–Segal) representation of the state ω, the sequence $S_\Lambda(B) = 1/|\Lambda| \sum_{x\in\Lambda} \tau_x B$ converges weakly to a multiple of the identity: $S(B) \equiv \omega(B)\mathbb{1}$. This theorem, called the mean ergodic theorem, characterizes the class of states yielding a weak law of large numbers. Clearly, these limits $\{S(A)|A \in \mathcal{B}\}$ form a trivial abelian algebra of macroscopic observables.

Now we go a step further and consider space fluctuations. Define the local fluctuation of an observable A in a homogeneous (spatial invariant) state ω by

$$F_\Lambda(A) = \frac{1}{|\Lambda|^{1/2}} \sum_{x\in\Lambda} (\tau_x A - \omega(A)) \qquad [5]$$

The problem is to give a rigorous meaning to $\lim F_\Lambda(A)$ for Λ tending to $\mathbb{Z}^\nu$ in the sense of extending boxes. When does such a limit exist? What are the properties of the fluctuations or the limits $F(A) = \lim F_\Lambda(A)$, etc.? Again, the $F(A)$ are macroscopic variables of the microsystem.

Already we remark the following: if A, B are strictly local elements, $A, B \in \mathcal{A}_{\mathrm{L}}$, then

$$\sum_{y\in\mathbb{Z}^\nu} [A, \tau_y B] \in \mathcal{A}_{\mathrm{L}}$$

and an easy computation yields, by [4],

$$\begin{aligned}
&\operatorname{weak}\lim_\Lambda [F_\Lambda(A), F_\Lambda(B)] \\
&= \operatorname{weak}\lim_\Lambda \frac{1}{|\Lambda|} \sum_{x\in\Lambda} \tau_x \left(\sum_{y\in\Lambda} [A, \tau_{y-x} B] \right) \\
&= \operatorname{weak}\lim_\Lambda \frac{1}{|\Lambda|} \sum_{x\in\Lambda} \tau_x \left(\sum_{y\in\mathbb{Z}^\nu} [A, \tau_y B] \right) \\
&= \sum_{y\in\mathbb{Z}^\nu} \omega\big([A, \tau_y B]\big) \equiv \mathrm{i}\sigma(A, B)\mathbb{1}
\end{aligned}$$

that is, if the $F(A)$ and $F(B)$ limits do exist, then

$$[F(A), F(B)] = \mathrm{i}\sigma(A, B)\mathbb{1} \qquad [6]$$

This property indicates that fluctuations should have the same commutation relations as boson fields. If fluctuations can be characterized as macroscopic observables, they must satisfy the canonical commutation relations (CCRs). Therefore, in the next section we introduce the essentials on CCR representations.

CCR Representations

We present the abstract Weyl CCR C^*-algebra. More details can be found in Brattelli and Robinson (1979, 2002) and in particular in Manuceau *et al.* (1973), where the case of a real test function space (H, σ) with a possibly degenerate symplectic form σ is treated. Hence, H is a real vector space and σ a bilinear, antisymmetric form on H.

Denote by $W(H, \sigma)$ the complex vector space generated by the functions $W(f)$, $f \in H$, defined by

$$\begin{aligned}
W(f) : H \to \mathbb{C} : g \to W(f)g \\
= \begin{cases} 0 & \text{if } f \neq g \\ 1 & \text{if } f = g \end{cases}
\end{aligned}$$

$W(H, \sigma)$ becomes an algebra with unit $W(0)$ for the product

$$W(f)W(g) = W(f+g)\mathrm{e}^{-(\mathrm{i}/2)\sigma(f,g)}; \quad f, g \in H$$

and a *-algebra for the involution

$$W(f) \to W(f)^* = W(-f)$$

It becomes a C^*-algebra $C^*(H, \sigma)$ following the construction of Verbeure and Zagrebnov (1992). A linear functional ω of a C^*-algebra $C^*(H, \sigma)$ is called a state if $\omega(I) = 1$ and $\omega(A^*A) \geq 0$ for all $A \in C^*(H, \sigma)$ and $I = W(0)$. Every state gives rise to a representation through the GNS construction (Brattelli and Robinson 1979, 2002). In particular, ω is a state if for any choice of $A = \sum_j c_j W(f_j)$ we have

$$\begin{aligned}
&\sum_{j^k} c_j \bar{c}_k \omega\big(W(f_j - f_j)\big)\mathrm{e}^{-\mathrm{i}\sigma(f_j, f_k)} \geq 0 \\
&\omega(W(0)) = 1
\end{aligned}$$

A remark about the special case that σ is degenerate is in order. Denote by H_0 the kernel of σ:

$$H_0 = \{f \in H | \, \sigma(f, g) = 0 \text{ for all } g \in H\}$$

If $H = H_0 \oplus H_1$ with σ_1 a nondegenerate symplectic form on H_1 and σ_1 equal to the restriction of σ to H_1, we have that $C^*(H, \sigma)$ is a tensor product:

$$C^*(H, \sigma) = C^*(H_0, 0) \otimes C^*(H_1, \sigma_1)$$

Note that $C^*(H_0, 0)$ is abelian and that each positive-definite normalized functional φ,

$$\varphi : h \in H_0 \to \varphi(W(h))$$

defines a state $\omega(W(h)) = \varphi(W(h))$ on $C^*(H_0, 0)$.

Let ξ be any character of the abelian additive group H, then the map τ_ξ,

$$\tau_\xi W(f) = \xi(f) W(f)$$

extends to a *-automorphism of $C^*(H, \sigma)$. Let s be a positive symmetric bilinear form on H such that for all $f, g \in H$:

$$\tfrac{1}{4}|\sigma(f, g)|^2 \leq s(f, f)\, s(g, g) \qquad [7]$$

and let $\omega_{s,\xi}$ be the linear functional on $C^*(H,\sigma)$ given by

$$\omega_{s,\xi}(W(h)) = \xi(h)e^{-(1/2)s(h,h)} \qquad [8]$$

then it is straightforward (Brattelli and Robinson 1979, 2002) to check that $\omega_{s,\xi}$ is a state on $C^*(H,\sigma)$. All states of the type [8] are called quasifree states on the CCR algebra $C^*(H,\sigma)$.

A state ω of $C^*(H,\sigma)$ is called a regular state if, for all $f,g\in H$, the map $\lambda\in\mathbb{R}\to\omega(W(\lambda f+g))$ is continuous. The regularity property of a state yields the existence of a Bose field as follows. Let $(\mathcal{H},\pi,\Omega)$ be the GNS representation (Brattelli and Robinson 1979, 2002) of the state w, then the regularity of w implies that there exists a real linear map $b:H\to\mathcal{L}(\mathcal{H})$ (linear operators on $\mathcal{H}$) such that $\forall f\in H$: $b(f)^*=b(f)$ and

$$\pi(W(f)) = \exp(ib(x))$$

The map b is called the Bose field satisfying the Bose field commutation relations:

$$[b(f),b(g)] = i\sigma(f,g) \qquad [9]$$

Note that the Bose fields are state dependent. Note also already that if ξ is a continuous character of H, then any quasifree state [8] is a regular state guaranteeing the existence of a Bose field.

Normal Fluctuations

In this section we develop the theory of normal fluctuations for ν-dimensional quantum lattice systems with a quasilocal structure (see the section "Quantum lattice systems") and for technical simplicity we assume that the local C^*-algebra $\mathcal{A}_x, x\in\mathbb{Z}^\nu$, are copies of the matrix algebra $M_n(\mathbb{C})$ of $n\times n$ complex matrices. Most of the results stated can be extended to the case where $\mathcal{A}_x$ is a general C^*-algebra (Goderis *et al.* 1989, 1990, Goderis and Vets 1989).

We consider a physical system $(\mathcal{B},\omega)$ where ω is a translation-invariant state of $\mathcal{B}$, that is, $\omega\circ\tau_x=\omega$ for all $x\in\mathbb{Z}^\nu$. Later on we extend the situation to a C^*-dynamical system $(\mathcal{B},\omega,\alpha_t)$ and analyze the properties of the dynamics α_t under the central limit.

For any local A we introduced its local fluctuation in the state ω of the system:

$$F_\Lambda(A) = \frac{1}{|\Lambda|^{1/2}}\sum_{x\in\Lambda}(\tau_x A-\omega(A)) \qquad [10]$$

The main problem is to give a rigorous mathematical meaning to the limits

$$\lim_{\Lambda\to\infty} F_\Lambda(A)\equiv F(A)$$

where the limit is taken for any increasing $\mathbb{Z}^\nu$-absorbing sequence $\{\Lambda\}_\Lambda$ of finite volumes of $\mathbb{Z}^\nu$. The limits $F(A)$ are called the macroscopic fluctuation operators of the system $(\mathcal{B},\omega)$.

Already earlier work (Cushen and Hudson 1971, Sewell 1986) suggested that the fluctuations behave like bosons. We complete this idea by proving that one gets a well-defined representation of a CCR C^*-algebra of fluctuations uniquely defined by the original system $(\mathcal{B},\omega)$.

Denote by $\mathcal{A}_{\mathrm{L,sa}}$ and $\mathcal{B}_{\mathrm{sa}}$ the real vector space of the self-adjoint elements of $\mathcal{A}_{\mathrm{L}}$, respectively, $\mathcal{B}$.

Definition 1 An observable $A\in\mathcal{B}_{\mathrm{sa}}$ satisfies the central-limit theorem if

(i) $\lim_\Lambda \omega(F_\Lambda(A)^2)\equiv s_\omega(A,A)$ exists and is finite, and
(ii) $\lim_\Lambda \omega(e^{itF_\Lambda(A)}) = e^{(-t/2)^2 s_\omega(A,A)}$ for all $t\in\mathbb{R}$.

Clearly, our definition coincides with the notion in terms of characteristic functions, for classical systems ($\mathcal{A}$ abelian) equivalent with the notion of convergence in distribution. For quantum systems, there does not exist a standard notion of "convergence in distribution." Only the concept of expectations is relevant. This does not exclude the notion of central-limit theorem in terms of the moments, which is the analog of the moment problem (Giri and von Waldenfels 1978).

Definition 2 The system $(\mathcal{B},\omega)$ is said to have normal fluctuations if ω is translation invariant and if

(i) $\forall A,B\in\mathcal{A}_{\mathrm{L}}$

$$\sum_{x\in\mathbb{Z}^\nu}|\omega(A\tau_x B)-\omega(A)\omega(B)|<\infty$$

(ii) the central-limit theorem holds for all $A\in\mathcal{A}_{\mathrm{L,sa}}$.

Note that (i) implies that the state ω is mixing for space translations. Also by (i), one can define a sesquilinear form on $\mathcal{A}_{\mathrm{L}}$:

$$\begin{aligned}\langle A,B\rangle_\omega &= \lim_\Lambda \omega(F_\Lambda(A^*)F_\Lambda(B))\\ &= \sum(\omega(A^*\tau_x B)-\omega(A^*)\omega(B))\end{aligned}$$

and denote

$$\begin{aligned}s_\omega(A,B) &= \mathrm{Re}\langle A,B\rangle_\omega\\ \sigma_\omega(A,B) &= 2\,\mathrm{Im}\langle A,B\rangle_\omega\end{aligned}$$

For $A,B\in\mathcal{A}_{\mathrm{L,sa}}$ one has

$$\sigma_\omega(A,B) = -i\sum_{x\in\mathbb{Z}^\nu}\omega([A,\tau_x B]) \qquad [11]$$

$$s_\omega(A,A) = \langle A,A\rangle_\omega \qquad [12]$$

Clearly, $(\mathcal{A}_{\mathrm{L,sa}},\sigma_\omega)$ is a symplectic space and s_ω a non-negative symmetric bilinear form on $\mathcal{A}_{\mathrm{L,sa}}$.

Following the discussion in the section "CCR representations" we get a natural CCR C^*-algebra $C^*(\mathcal{A}_{\mathrm{L,sa}}, \sigma_\omega)$ defined on this symplectic space. The following theorem is an essential step in the construction of a macroscopic physical system of fluctuations of the microsystem $(\mathcal{B}, \omega)$.

Theorem 1 *If the system $(\mathcal{B}, \omega)$ has normal fluctuations, then the limits $\{\lim_\Lambda \omega(e^{iF_\Lambda(A)}) = \exp((-1/2)s_\omega(A,A)), A \in \mathcal{A}_L\}$ define a quasifree state $\tilde{\omega}$ on the CCR C^*-algebra $C^*(\mathcal{A}_{\mathrm{L,sa}}, \sigma_\omega)$ by*

$$\tilde{\omega}(W(A)) = \exp\left(-\tfrac{1}{2}s_\omega(A,A)\right)$$

Proof The proof is clear from the definition [8] if one can prove that the positivity condition [7] holds. But the latter follows readily from

$$\begin{aligned}\tfrac{1}{4}|\sigma_\omega(A,B)|^2 &= \lim_\Lambda |\mathrm{Im}\,\omega(F_\Lambda(A)F_\Lambda(B))|^2 \\ &\le \lim_\Lambda \omega(F_\Lambda(A)^2)\omega(F_\Lambda(B)^2) \\ &= s_\omega(A,A)s_\omega(B,B)\end{aligned}$$

by Schwarz inequality. □

This theorem indicates that the quantum-mechanical alternative for (classical) Gaussian measures are quasifree states on CCR algebras. However, the following basic question arises: is it possible to take the limits of products of the form

$$\lim_\Lambda \omega\left(e^{iF_\Lambda(A)}e^{iF_\Lambda(B)}\cdots\right)$$

and, if they exist, do they preserve the CCR structure? Clearly, this is a typical noncommutative problem.

Using the following general bounds: for $C^* = \mathbb{C}$ and $D^* = D$ norm-bounded operators one has

$$\begin{aligned}&\left\|e^{i(C+D)} - e^{iC}\right\| \le \|D\| \\ &\left\|[e^{iC}, e^{iD}]\right\| \le \|[C,D]\| \\ &\left\|e^{i(C+D)} - e^{iC}e^{iD}\right\| \le \tfrac{1}{2}\|[C,D]\|\end{aligned}$$

and by the expansion of the exponential function one proves easily that

$$\lim_\Lambda \left\|e^{iF_\Lambda(A)}e^{iF_\Lambda(B)} - e^{i(F_\Lambda(A)+F_\Lambda(B))} \times e^{-(1/2)[F_\Lambda(A),F_\Lambda(B)]}\right\| = 0 \qquad [13]$$

if A and B are one-point observables, that is, if $A, B \in \mathcal{A}_{\{0\}}$. For general local elements the proof is somewhat more technical and can be based on a Bernstein-like argument (for details see Goderis and Vets (1989)). The property [13] can be seen as a Baker–Campbell–Hausdorff formula for fluctuations.

From [13], the mean ergodic theorem, and Theorem 1 we get:

Theorem 2 *If the system $(\mathcal{B}, \omega)$ has normal fluctuations then for $A, B \in \mathcal{A}_{\mathrm{L,sa}}$:*

$$\begin{aligned}&\lim_\Lambda \omega\left(e^{iF_\Lambda(A)}e^{iF_\Lambda(B)}\right) \\ &\quad = \exp\left\{-\frac{1}{2}s_\omega(A+B, A+B) - \frac{i}{2}\sigma_\omega(A,B)\right\} \\ &\quad = \tilde{\omega}(W(A)W(B))\end{aligned}$$

with $\tilde{\omega}$ a quasifree state on the CCR algebra $C^*(\mathcal{A}_{\mathrm{L,sa}}^{\sigma_\omega})$.

Theorems 1 and 2 describe completely the topological and analytical aspects of the quantum central-limit theorem under the condition of normal fluctuations (Definition 2). In fact, the quantum central limit yields, for every microphysical system $(\mathcal{B}, \omega)$, a macrophysical system $(C^*(\mathcal{A}_{\mathrm{L,sa}}, \sigma_\omega), \tilde{\omega})$ defined by the CCR C^*-algebra of fluctuation observables $C^*(\mathcal{A}_{\mathrm{L,sa}}, \sigma_\omega)$ in the representation defined by the quasifree state $\tilde{\omega}$. As the state $\tilde{\omega}$ is a quasifree state, it is a regular state, that is, the map $\lambda \in \mathbb{R} \to \tilde{\omega}(W(\lambda A + B))$ is continuous. From in section "CCR representations" we know that this regularity property yields the existence of a Bose field, that is, there exists a real linear map

$$F : A \in \mathcal{A}_{\mathrm{L,sa}} \to F(A)$$

where $F(A)$ is a self-adjoint operator on the GNS representation space $\tilde{\mathcal{H}}$ of $\tilde{\omega}$, such that for all $A, B \in \mathcal{A}_{\mathrm{L,sa}}$:

$$[F(A), F(B)] = i\sigma_\omega(A,B)$$

Moreover, if one has a complex structure J on $(\mathcal{A}_{\mathrm{L,sa}}, \sigma_\omega)$ such that $J^2 = -\mathbb{1}$ and for all $A, B \in \mathcal{A}_{\mathrm{L,sa}}$:

$$\begin{aligned}&\sigma_\omega(JA, B) = -\sigma_\omega(A, JB) \\ &\sigma_\omega(A, JB) > 0\end{aligned}$$

then one defines the boson creation and annihilation operators

$$F^\pm(A) = \frac{1}{\sqrt{2}}(F(A) \mp iF(JA))$$

satisfying the usual boson commutation relations

$$[F^-(A), F^+(B)] = \sigma_\omega(A, JB) + i\sigma_\omega(A,B)$$

Finally, it is straightforward, nevertheless important, to remark that Theorems 1 and 2 hold true if the linear space of local observables $\mathcal{A}_{\mathrm{L,sa}}$ is replaced by any of its subspaces. Some of them can have greater physical importance than others. This means that the quantum central-limit theorems can realize several macrophysical systems of fluctuations. But all of them are Bose field systems.

It is also important to remark that these results end up in giving a probabilistic canonical basis of the canonical commutation relations.

Now we analyze the notion of coarse graining due to the quantum central limit. Consider on $\mathcal{A}_{\mathrm{L}}$ the sesquilinear form (see [11], [12]) again

$$\begin{aligned}\langle A,B\rangle_\omega &= \sum_{x\in\mathbb{Z}^\nu}(\omega(A^*\tau_x B)-\omega(A)\omega(B))\\ &= s_\omega(A,B)+\mathrm{i}\sigma_\omega(A,B)\end{aligned}\qquad[14]$$

This form defines a topology on $\mathcal{A}_{\mathrm{L}}$ which is not comparable with the operator topologies induced by ω. In fact, this form is not closable in the weak, strong, ultraweak, or ultrastrong operator topologies.

We call A and B in $\mathcal{A}_{\mathrm{L}}$ equivalent, denoted by $A\sim B$ if $\langle A-B,A-B\rangle_\omega=0$. Clearly, this defines an equivalence relation on $\mathcal{A}_{\mathrm{L}}$. The property of coarse graining is mathematically characterized by the following: for all $A,B\in\mathcal{A}_{\mathrm{L,sa}}$ the relation $A\sim B$ is equivalent with $F(A)=F(B)$. Suppose first that $F(A)=F(B)$, then

$$[W(A),W(B)]=0$$

hence $\sigma_\omega(A,B)=0$. Therefore, from Theorem 1:

$$\begin{aligned}1&=\tilde{\omega}(W(A)W(B)^*)=\tilde{\omega}(W(A)W(-B))\\ &=\tilde{\omega}(W(A-B))=\exp\left(-\tfrac{1}{2}s_\omega(A-B,A-B)\right)\end{aligned}$$

and from [12] and [14]: $\langle A-B,A-B\rangle_\omega=0$. The converse is equally straightforward.

From this property, it follows immediately that, for example, the action of the translation group is trivial or that $F(\tau_x A)=F(A)$ for all $x\in\mathbb{Z}^\nu$. Therefore, the map $F:\mathcal{A}_{\mathrm{L,sa}}\to C^*(\mathcal{A}_{\mathrm{L,sa}},\sigma_\omega)$ is not injective. This expresses the physical phenomenon of coarse graining and gives a mathematical signification of the fluctuations being macroscopic observables.

In the above, we have constructed the new macroscopic physical system of quantum fluctuations for any microsystem with the property of normal fluctuations (see Definition 2). The main problem remains: when the microsystem does have normal fluctuations. We end this section with the formulation of a general sufficient clustering condition for the microstate ω in order that the microsystem $(\mathcal{B},\omega)$ has normal fluctuations.

Let $\Lambda,\Lambda'\in\mathcal{D}(\mathbb{Z}^\nu)$ and ω a translation invariant state, denote

$$\alpha^\omega(\Lambda,\Lambda')=\sup_{\substack{A\in\mathcal{A}_\Lambda;\|A\|=1\\ B\in\mathcal{A}_{\Lambda'};\|B\|=1}}|\omega(AB)-\omega(A)\omega(B)|$$

The cluster function $\alpha_N^\omega(d)$ is defined by

$$\begin{aligned}\alpha_N^\omega(d)=\sup\{\alpha^\omega(\Lambda,\Lambda'):\ &d(\Lambda,\Lambda')\geq d \text{ and}\\ &\max(|\Lambda|,|\Lambda'|)\leq N\}\end{aligned}$$

where $N,d\in\mathbb{R}^+$ and $d(\Lambda,\Lambda')$ is the Euclidean distance between Λ and Λ'. It is obvious that

$$\begin{aligned}\alpha_N^\omega(d)&\leq\alpha_N^\omega(d') \quad \text{if } d\geq d'\\ \alpha_N^\omega(d)&\leq\alpha_{N'}^\omega(d) \quad \text{if } N\leq N'\end{aligned}$$

The clustering condition is expressed by the following scaling law:

$$\exists\delta>0:\quad\lim_{N\to\infty}N^{1/2}\alpha_N^\omega\left(N^{1/2\nu-\delta}\right)=0\qquad[15]$$

or, equivalently,

$$\exists\delta>0:\quad\lim_{N\to\infty}N^{\nu+\delta}\alpha^\omega_{N^{2(\nu+\delta)}}(N)=0\qquad[16]$$

Note that this condition implies that

$$\sum_{x\in\mathbb{Z}^\nu}\alpha_N^\omega(|x|)<\infty$$

that is, that the function $\alpha_N^\omega(\,\cdot\,)$ is an $L^1(\mathbb{Z}^\nu)$-function for all N. In fact, this condition corresponds to the uniform mixing condition in the commutative (classical) central-limit theorem (see, e.g., Ibragimov and Linnick (1971)). This condition can also be called the modulus of decoupling. Product states, for example, equilibrium states of mean-field systems are uniformly clustering with $\alpha^\omega(d)=0$ for $d>0$.

The normality of the fluctuations of the microsystem $(\mathcal{B},\omega)$ for product states is proved and extensively studied in Goderis *et al.* (1989), and for states satisfying the condition [15] or [16] in Goderis and Vets (1989). In the latter case, the proofs are very technical and based on a generalization of the well-known Bernstein argument (Ibragimov and Linnick 1971) of the classical central-limit theorem to the noncommutative situation. A refinement of these arguments can be found in Goderis *et al.* (1990). For the sake of formal self-consistency we formulate the theorem:

Theorem 3 (Central-limit theorem) *Take the microsystem $(\mathcal{B},\omega)$ such that ω is lattice translation invariant and satisfies the clustering condition* [15]; *then the system has normal fluctuations for all elements of the vector space of local observables $\mathcal{A}_{\mathrm{L,sa}}$.* □

In Goldshtein (1982) a noncommutative central-limit theorem is derived using similar techniques. The main difference, however, is its strictly local character, namely for one local operator separately. The conditions depend on the spectral properties of the operator. It excludes a global approach resulting in a CCR algebra structure.

Even for quantum lattice systems, it is not straightforward to check whether a state satisfies

the degree of mixing as expressed in conditions [15]–[16]. Clearly, one expects the condition to hold for equilibrium states at high enough temperatures. For quantum spin chains, a theorem analogous with Theorem 3 under weaker conditions than [15] is proved for example, in Matsui (2003).

So far we have reviewed the quantum central-limit theorem for physical C^*-spin systems $(\mathcal{B}, \omega)$ with normal fluctuations.

Now we extend the physical system to a C^*-dynamical system $(\mathcal{B}, \omega, \alpha_t)$ (Brattelli and Robinson 1979, 2002) and we investigate the properties of the dynamics α_t under the central limit. As usual, the dynamics is supposed to be of the short-range type in order to guarantee the norm limit:

$$\alpha_t(\cdot) = n - \lim_{\Lambda} e^{itH_\Lambda} \cdot e^{-itH_\Lambda}$$

and space homogeneous $\alpha_t \cdot \tau_x = \tau_x \cdot \alpha_t, \forall t \in \mathbb{R}, \forall x \in \mathbb{Z}^\nu$. We suppose that the state ω is both space as time translation invariant. Moreover, we assume that the state ω satisfies the mixing condition [15] for normal fluctuations.

In [10] we defined, for every local $A \in \mathcal{A}_{\mathrm{L,sa}}$, the local fluctuation $F_\Lambda(A)$ and obtained a clear meaning of $F(A) = \lim_\Lambda F_\Lambda(A)$ from the central-limit theorem. Now we are interested in the dynamics of the fluctuations $F(A)$. Clearly, for all $A \in \mathcal{A}_{\mathrm{L,sa}}$ and all finite Λ:

$$\alpha_t F_\Lambda(A) = F_\Lambda(\alpha_t A) \qquad [17]$$

and one is tempted to define the dynamics $\tilde{\alpha}_t$ of the fluctuations in the Λ-limit by the formula

$$\tilde{\alpha}_t F(A) = F(\alpha_t A) \qquad [18]$$

Note, however, that in general $\alpha_t A$ is not a local element of $\mathcal{A}_{\mathrm{L,sa}}$. It is unclear whether the central limit of elements of the type $\alpha_t A$, with $A \in \mathcal{A}_{\mathrm{L,sa}}$ exists or not and hence whether one can give a meaning to $F(\alpha_t A)$. Moreover, if $F(\alpha_t A)$ exists, it remains to prove that $(\tilde{\alpha}_t)_t$ defines a weakly continuous group of $*$-automorphisms on the fluctuation CCR algebra $\tilde{\mathcal{M}} = C^*(\mathcal{A}_{\mathrm{L,sa}}, \sigma_\omega)''$ (the von Neumann algebra generated by the $\tilde{\omega}$-representation of $C^*(\mathcal{A}_{\mathrm{L,sa}}, \sigma_\omega)$). All this needs a proof. In Goderis *et al.* (1990), one finds the proof of the following basic theorem about the dynamics.

Theorem 4 *Under the conditions on the dynamics α_t and on the state ω expressed above, the limit $F(\alpha_t A) = \lim_\Lambda F_\Lambda(\alpha_t A)$ exists as a central limit as in Theorem 2, and the maps $\tilde{\alpha}_t$ defined by [18] extend to a weakly continuous one-parameter group of $*$-automorphisms of the von Neumann algebra $\tilde{\mathcal{M}}$. The quasifree state $\tilde{\omega}$ is $\tilde{\alpha}_t$-invariant (time invariant).*

This theorem yields the existence of a dynamics $\tilde{\alpha}_t$ on the fluctuations algebra and shows that it is of the quasifree type

$$\tilde{\alpha}_t F(A) = F(\alpha_t A)$$

where $F(A)$ is a representation of a Bose field in a quasifree state $\tilde{\omega}$, the noncommutative version of a Gaussian distribution. In physical terms, it also means that any microdynamics α_t induces a linear process on the level of its fluctuations.

We can conclude that on the basis of the Theorems 3 and 4 the quantum central-limit theorem realized a map from the microdynamical system $(\mathcal{B}, \omega, \alpha_t)$ to a macrodynamical system $(C^*(\mathcal{A}_{\mathrm{L,sa}}, \sigma_\omega), \tilde{\omega}, \tilde{\alpha}_t)$ of the quantum fluctuations. The latter system is a quasifree Boson system.

Note that, contrary to the central-limit theorem, the law of large numbers [4] maps local observables to their averages forming a trivial commutative algebra of macro-observables. The macrodynamics is mapped to a trivial dynamics as well. Therefore, the consideration of law of large numbers does not allow one to observe genuine quantum phenomena. On the other hand, on the level of the fluctuations, macroscopic quantum phenomena are observable.

Abnormal Fluctuations

The results about normal fluctuations in the last section contain two essential elements. On the one hand, the central limit has to exist. The condition in order that this occurs is the validity of the cluster condition ([15] or [16]) guaranteeing the normality of the fluctuations. On the other hand, there is the reconstruction theorem, identifying the CCR algebra representation of the fluctuation observables or operators in the quasifree state, which is denoted by $\tilde{\omega}$.

The cluster condition is in general not satisfied for systems with long-range correlations, for example, for equilibrium states at low temperatures with phase transitions. It is a challenging question to also study in this case the existence of fluctuations operators and, if they exist, to study their mathematical structure. Here we detect structures other than the CCR structure, other states or distributions different from quasifree states, etc.

Progress in the elucidation of all these questions started with a detailed study of abnormal fluctuations in the harmonic and anharmonic crystal models (Verbeure and Zagrebnov 1992, Momont *et al.* 1997). More general Lie algebras are obtained than the Heisenberg Lie algebra of the CCR algebra, and more general states $\tilde{\omega}$ or quantum distributions

are computed beyond quasifree states, which is the case for normal fluctuations.

Abnormal fluctuations turn up, if one has an ergodic state ω with long-range correlations. We have in mind continuous (second-order) phase transitions, then typically, for example, the heat capacity or some more general susceptibilities diverge at critical points or lines. This means that normally scaled (with the factor $|\Lambda|^{-1/2}$) fluctuations of some observables diverge. This is equivalent with the divergence of sums of the type

$$\sum_{x\in\mathbb{Z}^\nu}(\omega(A\tau_x A)-\omega(A)^2)$$

for some local observable A.

In order to deal with these situations, we rescale the local fluctuations. One determines a scaling index $\delta_A\in(-1/2,1/2)$, depending on the observable A, such that the abnormally scaled local fluctuations

$$F_\Lambda^{\delta_A}=|\Lambda|^{-\delta_A}F_\Lambda(A)$$

with $F_\Lambda(A)$ as in [10], yield a nontrivial characteristic function: $\forall t\in\mathbb{R}$,

$$\lim_\Lambda \omega_\Lambda(\mathrm{e}^{\mathrm{i}tF_\Lambda^{\delta_A}(A)})\equiv\phi_A(t) \qquad [19]$$

where we limit our discussion to states ω_Λ local Gibbs states. The index δ_A is a measure for the abnormality of the fluctuation of A. Note that $\delta_A=-1/2$ yields a triviality and that $\delta_A=1/2$ would lead to a law of large numbers (theory of averages). Observe also that in general the characteristic function ϕ_A or the corresponding state $\tilde{\omega}$ need not be Gaussian or quasifree.

In the physics literature, one describes the long-range order by means of the asymptotic form of the connected two-point function in terms of the critical exponent η

$$\omega_\Lambda(a\tau_x A)-\omega_\Lambda(A)^2\simeq 0\left(\frac{1}{|x|^{\nu-2+\eta}}\right);\quad |x|\to\infty \qquad [20]$$

Our scaling index δ_A is related to the critical exponent η by the straightforward relation

$$\eta=2-2\nu\delta_A$$

As stated above, the index δ_A is determined by the existence of the central limit and explicitly computed in several model calculations, for example, Verbeure and Zagrebnov (1992), and for equilibrium states. Apart from the strong model dependence, the indices also depend strongly on the chosen boundary conditions. This fact draws a new light on the universality of the critical exponents.

Suppose now that the indices δ_A are determined by the existence of the central limit [19]. The next problem is to find out whether also in these cases a reconstruction theorem, comparable to, for example, Theorem 2, can be proved giving again a mathematical meaning to the limits

$$\lim_\Lambda F_\Lambda^{\delta_A}(A)\equiv F^{\delta_A}(A) \qquad [21]$$

as operators, in general unbounded, on a Hilbert space.

Here we develop a proof of the Lie algebra character of the abnormal fluctuations under the conditions: (1) the δ-indices are determined by the existence of the variances (second moments), and (2) the existence of the third moments (for more details see, e.g., Momont *et al.* (1997)).

Consider a local algebra, namely an n-dimensional vector space $\mathcal{G}$ with basis $\{v_i\}_{i=1,\ldots,n}$ and product

$$v_j\cdot v_k\equiv[v_j,v_k]=\sum_{\ell=1}^n c_{jk}^\ell v_\ell \qquad [22]$$

with structure constants c_{jk}^ℓ satisfying

$$c_{jk}^\ell+c_{kj}^\ell=0$$

$$\sum_r(c_{ij}^r c_{rk}^s+c_{jk}^r c_{ri}^s+c_{ki}^r c_{rj}^s)=0$$

Consider the concrete Lie algebra basis of operators in $\mathcal{A}_{\{0\}}$

$$\{L_0=i\mathbb{1},L_1,\ldots,L_m\},\quad m<\infty$$

such that $L_j^*=-L_j$, $j=0,1,\ldots,m$ and $\omega(L_j)=\lim_\Lambda\omega_\Lambda(L_j)=0$ for $j>0$. Clearly, $\omega_\Lambda(L_0)=i$ for all Λ, and the $\{L_i\}$ satisfy eqn [22]. Because of the special choices of L_0 one has $c_{ok}^\ell=c_{ko}^\ell=0$ and $c_{jk}^o=-\mathrm{i}\lim_\Lambda\omega_\Lambda([L_j,L_k])$. We consider now the fluctuations of these generators and we are looking for a characterization of the Lie algebra of the fluctuations if any.

For a translation-invariant local state $\omega_\Lambda,\Lambda\subseteq\mathbb{Z}^\nu$, such that $\omega=\lim_\Lambda\omega_\Lambda$ is mixing, define the local fluctuations, for $j=1,\ldots,m$,

$$F_j^{\delta_j},\Lambda=\frac{1}{|\Lambda|^{1/2+\delta_j}}\sum_{x\in\Lambda}(\tau_x L_j-\omega_\Lambda(L_j)) \qquad [23]$$

and for notational convenience, take

$$F_{0,\Lambda}=\mathrm{i}\mathbb{1}$$

Now we formulate the conditions for our purposes.

Condition A We assume that the parameters δ_j are determined by the existence of the finite and nontrivial variances: for all $j=1,\ldots,m$,

$$0<\lim_\Lambda\omega_\Lambda\left((F_{j,\Lambda}^{\delta_j})^2\right)<\infty \qquad [24]$$

After reordering, take $1/2 > \delta_1 \geq \delta_2 \geq \cdots \geq \delta_m > -1/2$.

Condition B Assume that all third moments are finite, that is,

$$\lim_{\Lambda} \left| \omega_\Lambda \left(F_{j,\Lambda}^{\delta_j} F_{k,\Lambda}^{\delta_k} F_{\ell,\Lambda}^{\delta_\ell} \right) \right| < \infty$$

We have in mind, that the ω_Λ's are Gibbs states for some local Hamiltonians with some specific boundary conditions. The limit $\Lambda \to \mathbb{Z}^\nu$ may depend very strongly on these boundary conditions, in the sense that they are visible in the values of the indices δ_j (see, e.g., Verbeure and Zagrebnov (1992)). If for some $j \geq 1$, the corresponding $\delta_j = 0$ then the operator L_j has a normal fluctuation operator

$$F_j^{\delta_j} = \lim_{\Lambda} F_{j,\Lambda}^{\delta_j} \qquad [25]$$

where the limit is understood in the sense of Condition A, namely a finite nontrivial variance. If, for some $j \geq 1$, the corresponding $\delta_j \neq 0$, then the fluctuation [25] is called an abnormal fluctuation operator. In order to satisfy Condition A, it happens sometimes that δ_j has to be chosen negative (see, e.g., Verbeure and Zagrebnov (1992)). In this case, it is reasonable to limit our discussion to the situation that all $\delta_j > -1/2$.

On the basis of Condition A, the limit set $\{F_j^{\delta_j}\}_{j=0,\ldots,m}$ of fluctuation operators generates a Hilbert space $\mathcal{H}$ with scalar product

$$\left(F_j^{\delta_j}, F_k^{\delta_k} \right) = \lim_{\Lambda} \omega_\Lambda \left((F_{j,\Lambda}^{\delta_j})^* F_{k,\Lambda}^{\delta_k} \right) \qquad [26]$$

On the basis of Condition B, the fluctuation operators are defined as multiplication operators of the Hilbert space $\mathcal{H}$. Note that the Conditions A and B are not sufficient to obtain a characteristic function. However, they are sufficient to obtain the notion of fluctuation operator. Now we proceed to clarify the Lie algebra character of these fluctuation operators on $\mathcal{H}$.

Consider the Lie product of two local fluctuations for a finite Λ, one gets

$$\left[F_{j,\Lambda}^{\delta_j}, F_{k,\Lambda}^{\delta_k} \right] = \sum_{\ell=0}^{m} c_{jk}^{\ell}(\Lambda) F_{\ell,\Lambda}^{\delta_\ell} \qquad [27]$$

with

$$c_{jk}^{\ell}(\Lambda) = \frac{c_{jk}^{\ell}}{|\Lambda|^{1/2+\delta_j+\delta_k-\delta_\ell}}, \quad \ell = 1, \ldots, m$$

$$c_{jk}^{0}(\Lambda) = |\Lambda|^{-\delta_j-\delta_k} \sum_{\ell=0}^{m} c_{jk}^{\ell} \omega_\Lambda(F_{\ell,\Lambda}^{\delta_\Lambda})$$

It is an easy exercise to check that the $\{c_{jk}^{\ell}(\Lambda)\}$ are the structure coefficients of a Lie algebra $\mathcal{G}(\Lambda)$. Hence, by considering local fluctuations, one constructs a map from the Lie algebra $\mathcal{G}$ onto the Lie algebra $\mathcal{G}(\Lambda)$ by a nontrivial change of the structure constants. When the transformed structure constants approach a well-defined limit, a new nonisomorphic Lie algebra might appear. The limit algebra $\mathcal{G}(\mathbb{Z}^\nu)$, called the contracted one of the original one $\mathcal{G}$ is always nonsemisimple. This contraction is a typical Inönü–Wigner contraction (Inönü and Wigner 1953). About the limit algebra $\mathcal{G}(\mathbb{Z}^\nu)$, the following results are obtained (see Momont *et al.* (1997)):

$$\lim_{\Lambda} c_{jk}^{\ell}(\Lambda) = \begin{cases} 0 & \text{if } \frac{1}{2} + \delta_j + \delta_k - \delta_\ell > 0 \\ c_{jk}^{\ell} & \text{if } \ldots\ldots\ldots\ldots\ldots = 0 \\ 0 & \text{if } \ldots\ldots\ldots\ldots\ldots < 0 \end{cases} \qquad [28]$$

It is interesting to distinguish a number of special cases:

1. If all fluctuations are normal, one recovers the Heisenberg algebra of the canonical commutation relations with the right symplectic form σ_ω.
2. If $1/2 + \delta_j + \delta_k - \delta_\ell > 0$ for all j, k, ℓ one obtains an abelian Lie algebra of fluctuations.
3. One gets the richest structure if $1/2 + \delta_j + \delta_k - \delta_\ell = 0$ for all j, k, ℓ or for some of them. One notes a phenomenon of scale invariance, the $c_{jk}^{\ell}(\Lambda)$ are Λ-independent. Algebras different from the CCR algebra are observed. A particularly interesting case turns up if $\delta_j = -\delta_k \neq 0$, that is, one of the indices is negative, for example, $\delta_j < 0$, the corresponding fluctuation $F_j^{\delta_j}$ shows a property of space squeezing, and then $\delta_k > 0$, the fluctuation $F_k^{\delta_k}$ expresses the property of space dilation. These phenomena are observed and computed in several models (see, e.g., Verbeure and Zagrebnov (1992)). This yields in particular a microscopic explanation of the phenomenon of squeezing (squeezed states and all that) in quantum optics. We refer also to the section "Spontaneous symmetry breaking" for this phenomenon as being the basis of the construction of the Goldstone normal modes of the Goldstone particle appearing in systems showing spontaneous symmetry breakdown.

Some Applications

The notion of fluctuation operator as presented above, and the mathematical structure of the algebra of fluctuations have been tested in several soluble models. Many applications of this theory of quantum fluctuations can be found in the list of references. Here we are not entering into the details

of any model, but we limit ourselves to mention three applications which are of a general nature and totally model independent.

Conservation of the KMS Property under the Transition from Micro to Macro

Suppose that we start with a micro-dynamical system $(\mathcal{B}, \omega, \alpha_t)$ with normal fluctuations, that is, we are in the situation as treated in the section "Normal fluctuations." Hence, we know that the quantum central-limit theorem maps the system $(\mathcal{B}, \omega, \alpha_t)$ onto the macrodynamical system $(C^*(\mathcal{A}_{\mathrm{L,sa}}, \sigma_\omega), \tilde{\omega}, \tilde{\alpha}_t)$ of quantum fluctuations.

If the microstate ω is α_t-time invariant ($\omega \cdot \alpha_t = \omega$ for all $t \in \mathbb{R}$), then it also follows readily that the macrostate $\tilde{\omega}$ is $\tilde{\alpha}_t$-time invariant (see Theorem 4, i.e., $\tilde{\omega} \cdot \tilde{\alpha}_t = \tilde{\omega}$ for all $t \in \mathbb{R}$).

A less trivial question to pose is: suppose that the microstate ω is an equilibrium state for the micro-dynamics α_t, is then the macrostate $\tilde{\omega}$ also an equilibrium state for the macrodynamics $\tilde{\alpha}_t$ of the fluctuations? In Goderis *et al.* (1990) this question is answered positively in the following more technical sense: if ω is an α_t-KMS state of $\mathcal{B}$ at inverse temperature β, then $\tilde{\omega}$ is an $\tilde{\alpha}_t$-KMS state at the same temperature.

This property proves that the notion of equilibrium is preserved under the operation of coarse graining induced by the central-limit theorem. This statement constitutes a proof of one of the basic assumptions of the phenomenological theory of Onsager about small oscillations around equilibrium.

This result also yields a contribution to the discussion whether or not quantum systems should be described at a macroscopic level by classical observables. The result above states that the macroscopic fluctuation observables behave classically if and only if they are time invariant. In other words, it can only be expected *a priori* that conserved quantities behave classically. In principle, other observables follow a quantum dynamics.

Linear Response Theory

In particular, in the study of equilibrium states (KMS states) a standard procedure is to perturb the system and to study the response of the system as a function of the perturbation. The response elucidates many, if not all, of the properties of the equilibrium state.

Technically, one considers a perturbation of the dynamics by adding a term to the Hamiltonian. One expands the perturbed dynamics in terms of the perturbation and the unperturbed dynamics. It is often argued that when the perturbation is small, one can limit the study of the response to the first-order term in the perturbation in the corresponding Dyson expansion. This is the basis of what is called the "linear response theory of Kubo."

A long-term debate is going on about the validity of the linear response theory. The question is how to understand from a microscopic point of view the validity of the response theory being linear or not. One must realize that the linear response theory actually observed in macroscopic systems seems to have a significant range of validity beyond the criticism being expressed about it.

Here we discuss the main result of the paper (Goderis *et al.* 1991) in which contours are sketched for the exactness of the response being linear.

We assume:

1. that the microdynamics α_t is the norm-limit of the local dynamics $\alpha_t^\Lambda = \mathrm{e}^{\mathrm{i}tH_\Lambda} \cdot \mathrm{e}^{-\mathrm{i}tH_\Lambda}$, where H_Λ contains only standard finite-range interactions (as in the section "Normal fluctuations");
2. that the ω_Λ are states such that $\omega = \lim_\Lambda \omega_\Lambda$ is a state which is time and space translation invariant; and
3. that ω satisfies the cluster condition [15] or [16].

From the time invariance of the state, one has a Hamiltonian GNS representation of the dynamics: $\alpha_t = \mathrm{e}^{\mathrm{i}tH} \cdot \mathrm{e}^{-\mathrm{i}tH}$. On the basis of Theorem 4, one has the dynamics $\tilde{\alpha}_t$ of the fluctuation algebra $C^*(\mathcal{A}_{\mathrm{L,sa}}, \sigma_\omega)$ in the state $\tilde{\omega}$. This GNS representation yields a Hamiltonian representation for $\tilde{\alpha}_t$:

$$\tilde{\alpha}_t = \mathrm{e}^{\mathrm{i}t\tilde{H}} \cdot \mathrm{e}^{-\mathrm{i}t\tilde{H}}$$

Now take any local perturbation $P \in \mathcal{A}_{\mathrm{L,sa}}$ of α_t, namely

$$\alpha_{t,\Lambda}^P = \mathrm{e}^{\mathrm{i}t(H+F_\Lambda(P))} \cdot \mathrm{e}^{-\mathrm{i}t(H+F_\Lambda(P))}$$

where $F_\Lambda(P)$ is the local fluctuation of P in ω. Then one proves the following central-limit theorem (Goderis *et al.* 1991): for all A and B in $\mathcal{A}_{\mathrm{L,sa}}$, one has the perturbed dynamics

$$\tilde{\alpha}_t^P = \mathrm{e}^{\mathrm{i}t(\tilde{H}+F(P))} \cdot \mathrm{e}^{-\mathrm{i}t(\tilde{H}+F(P))}$$

of the fluctuation algebra in the sense of [18]:

$$\tilde{\alpha}_t^P F(A) = \lim_\Lambda F(\alpha_{t,\Lambda}^P(A))$$

This proves the existence and the explicit form of the perturbed dynamics lifted to the level of the fluctuations. In particular, one has

$$\lim_\Lambda \omega_\Lambda \left(\alpha_{t,\Lambda}^P(F_\Lambda(A)) \right) = \tilde{\omega}(\tilde{\alpha}_t^P F(A))$$

This is nothing but the existence of the relaxation function of Kubo but lifted to the level of the fluctuations and instead of dealing with strictly local observables here one considers fluctuations.

Assume, furthermore, that the state ω is an (α_t, β)-KMS state; then one derives readily Kubo's famous formula of his linear response theory:

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\omega}(\tilde{\alpha}_t^P F(A)) = \mathrm{i}\tilde{\omega}([F(P), \tilde{\alpha}_t F(A)])$$

which shows full linearity in the perturbation observable P. Kubo's formula arises as the central limit of the microscopic response to the dynamics perturbed by a fluctuation observable. We remark that if ω is an equilibrium state, then the right-hand side of the formula above can be expressed in terms of the Duhamel two-point function, which is the common way of doing in linear response theory.

Spontaneous Symmetry Breaking

SSB is one of the basic phenomena accompanying collective phenomena, such as phase transitions in statistical mechanics, or specific ground states in field theory. SSB goes back to the Goldstone theorem. There are many different situations to consider, for example, in the case of short-range interactions, it is typical that SSB yields a dynamics which remains symmetric, whereas for long-range interactions SSB also breaks the symmetry of the dynamics. However, in all cases the physics literature predicts the appearance of a particular particle, namely the Goldstone boson, to appear as a result of SSB. The theory of fluctuation operators allows the construction of the canonical coordinates of this particle. The most general result can be found in Michoel and Verbeure (2001). We sketch the essentials in two cases, namely for systems of long-range interactions (mean fields) and for systems with short-range interactions.

Long-range (mean-field) interactions Here we give explicitly the example of the strong-coupling BCS model in one dimension ($\nu = 1$). The microscopic algebra of observables is $\mathcal{B} = \otimes_i (M_2)_i$, where M_2 is the algebra of 2×2 complex matrices. The local Hamiltonian of the models is given by

$$H_N = \epsilon \sum_{i=-N}^{N} \sigma_i^z - \frac{1}{2N+1} \sum_{i,j=-N}^{N} \sigma_i^+ \sigma_j^-$$
$$0 < \epsilon < \tfrac{1}{2}$$

where $\sigma^z, \sigma^\pm$ are the usual 2×2 Pauli matrices. In the thermodynamic limit, the KMS equation has the following product state solutions: $\omega_\lambda = \otimes_i \mathrm{tr}\rho_\lambda$, where

$$\rho_\lambda = \frac{\mathrm{e}^{-\beta h_\lambda}}{\mathrm{tr}\,\mathrm{e}^{-\beta h_\lambda}}, \quad \lambda = \mathrm{tr}\,\rho_\lambda \sigma^- = \omega_\lambda(\sigma^-)$$
$$h_\lambda = \epsilon\sigma^z - \lambda\sigma^+ - \lambda\sigma^-$$

Note that $\lambda = \mathrm{tr}\rho_\lambda\sigma^-$ is a nonlinear equation for λ whose solutions determine the density matrix ρ_λ. This equation always has the solution $\lambda = 0$, describing the so-called normal phase. For $\beta > \beta_c$, with $th\beta_c\epsilon = 2\epsilon$, one has a solution $\lambda \neq 0$, describing the superconducting phase. Remark that if λ is a solution, then also $\lambda e^{i\phi}$ for all ϕ is a solution as well. It is clear that H_N is invariant under the continuous gauge transformation automorphism group $\mathcal{G} = \{\gamma_\varphi \mid \varphi \in [0, 2\pi]\}$ of $\mathcal{B}$:

$$\gamma_\varphi(\sigma_i^+) = e^{-i\varphi}\sigma_i^+$$

Hence $\mathcal{G}$ is a symmetry group. On the other hand: $\omega_\lambda(\gamma_\varphi(\sigma_i^+)) = e^{-i\varphi}\omega_\lambda(\sigma_i^+) \neq \omega_\lambda(\sigma_i^+)$. The gauge group $\mathcal{G}$ is spontaneously broken. Remark also that the gauge transformations are implemented locally by the charges

$$Q_N = \sum_{j=-N}^{N} \sigma_i^z, \quad \text{i.e.,}\ \gamma_\varphi(\sigma_i^+) = \mathrm{e}^{-\mathrm{i}\varphi Q_N}\sigma_i^+\mathrm{e}^{\mathrm{i}\varphi Q_N}$$

and σ^z is the symmetry generator density. As the states ω_λ are product states, all fluctuations are normal (see the section "Normal fluctuations"). One considers the local operators

$$Q = \frac{|\lambda|^2}{\mu^2}\sigma^z + \frac{\epsilon}{\mu^2}(\lambda\sigma^+ + \overline{\lambda}\sigma^-)$$
$$P = \frac{\mathrm{i}}{\mu}(\lambda\sigma^+ - \overline{\lambda}\sigma^-)$$

where $\mu = (\epsilon^2 + |\lambda|^2)^{1/2}$. Note that P is essentially the order parameter operator, that is, the operator P is breaking the symmetry:

$$\frac{\mathrm{d}}{\mathrm{d}\varphi}\omega_\lambda(\gamma_\varphi(A)) \neq 0, \quad \omega_\lambda(A) = 0$$

On the other hand, Q is essentially the generator of the symmetry σ^z normalized to zero, that is, $\omega_\lambda(Q) = 0$.

Michoel and Verbeure (2001) proved in detail that the fluctuations $F(Q)$ and $F(P)$ form a canonical pair

$$[F(Q), F(P)] = \mathrm{i}\frac{4|\lambda|^2}{\mu}$$

and that they behave, under the time evolution, as harmonic oscillator coordinates oscillating with a

frequency equal to 2μ. This frequency is called a plasmon frequency. Moreover, the variances are

$$\tilde{\omega}_\lambda(F(Q)^2) = \frac{|\lambda|^2}{\mu^2} = \tilde{\omega}_\lambda(F(P)^2)$$

This means that these coordinates vanish or disappear if $\lambda = 0$. The coordinates $F(Q)$ and $F(P)$ are the canonical coordinates of a particle appearing only if there is spontaneous symmetry breakdown. They are the canonical coordinates of the Goldstone boson, which arise if SSB occurs.

Short-range interactions An analogous result, as for long-range interactions, can be derived for systems with short-range interactions. However, in this case we have equilibrium states with poor cluster properties. We are now in the situation as described in the "Abnormal Fluctuations" section. Also in this case we have the phenomenon of SSB, which shows the appearance of a Goldstone particle. Also in this case one is able to construct its canonical coordinates. The details of this construction can be found in Michoel and Verbeure (2001). Here we give a heuristic picture of this construction.

Consider again a microsystem $(\mathcal{B}, \omega, \alpha_t)$ and let γ_s be a strongly continuous one-parameter symmetry group of α_t which is locally generated by $Q_\Lambda = \sum_{x\in\Lambda} q_x$. SSB amounts to find an equilibrium (KMS) or ground state ω which breaks the symmetry, that is, there exists a local observable $A \in \mathcal{A}_{\mathrm{L,sa}}$ such that for $s \neq 0$ holds: $\omega(\gamma_s(A)) \neq \omega(A)$ and $\alpha_t\gamma_s = \gamma_s\alpha_t$. This is equivalent to

$$\left.\frac{\mathrm{d}}{\mathrm{d}s}\omega(\gamma_s(A))\right|_{s=0} = \lim_\Lambda \omega([Q_\Lambda, A]) = c \neq 0$$

with c a constant.

Now we turn this equation into a relation for fluctuations. Using space translation invariance of the state, one gets

$$\lim_\Lambda \frac{1}{|\Lambda|}\omega\left(\left[\sum_{x\in\Lambda}(q_x - \omega(q))\sum_{y\in\Lambda}(\tau_x A - \omega(A))\right]\right) = c$$

We now use another consequence of the Goldstone theorem, namely that SSB implies poor clustering properties for the order parameter A, that is, in the line of what is done in the last section, we assume that the lack of clustering is expressed by the existence of a positive index δ such that

$$\lim_\Lambda \omega\left(\frac{1}{|\Lambda|^{1+2\delta}}\left(\sum_{x\in\Lambda}(\tau_x A - \omega(A))\right)^2\right)$$

is nontrivial and finite. This means that the fluctuation $F^\delta(A)$ exists. Then we get

$$\lim_\Lambda \omega\left(\left[\frac{1}{|\Lambda|^{1/2-\delta}}\sum_{x\in\Lambda}(q_x - \omega(q)), \frac{1}{|\Lambda|^{1/2+\delta}}\sum_{y\in\Lambda}(\tau_x A - \omega(A))\right]\right) = c$$

Hence

$$\tilde{\omega}\left(\left[F^{-\delta}(q), F^\delta(A)\right]\right) = c$$

which for equilibrium states ω, turns into the operator equation for fluctuations

$$[F^{-\delta}(q), F^\delta(A)] = c\mathbb{1}$$

In other words, one obtains a canonical pair $(F^{-\delta}(q), F^\delta(A))$ of normal coordinates of the collective Goldstone mode.

Note that the long-range correlation of the order-parameter operator (positive δ) is exactly compensated by a squeezing, described by the negative index $-\delta$, for the fluctuation operator of the local generator of the broken symmetry. This result can also be expressed as typical for SSB, namely that the symmetry is not completely broken, but only partially. More detailed information about all this is found in Michoel and Verbeure (2001).

See also: Algebraic Approach to Quantum Field Theory; Large Deviations in Equilibrium Statistical Mechanics; Macroscopic Fluctuations and Thermodynamic Functionals; Quantum Phase Transitions; Quantum Spin Systems; Symmetry Breaking in Field Theory; Tomita–Takesaki Modular Theory.

Further Reading

Brattelli O and Robinson DW (1979) *Operator Algebras and Quantum Statistical Mechanics*, vol. I. New York–Heidelberg–Berlin: Springer.

Brattelli O and Robinson D (2002) *Operator Algebras and Quantum Statistical Mechanics*, vol. II. New York–Heidelberg–Berlin: Springer.

Cushen CD and Hudson RL (1971) A quantum mechanical central limit theorem. *Journal of Applied Probability* 8: 454–469.

Fannes M and Quaegebeur J (1983) Central limits of product mappings between CAR-algebras. *Publications of the Research Institute for Mathematical Studies Kyoto* 19: 469–491.

Giri N and von Waldenfels W (1978) An algebraic version of the central limit theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte gebiete* 42: 129–134.

Goderis D, Verbeure A, and Vets P (1989) Non-commutative central limits. *Probability and Related Fields* 82: 527–544.

Goderis D, Verbeure A, and Vets P (1990) Dynamics of fluctuations for quantum lattice systems. *Communications in Mathematical Physics* 128: 533–549.

Goderis D, Verbeure A, and Vets P (1991) About the exactness of the linear response theory. *Communications in Mathematical Physics* 136: 265–583.

Goderis D and Vets P (1989) Central limit theorem for mixing quantum systems and the CCR-algebra of fluctuations. *Communications in Mathematical Physics* 122: 249.

Goldshtein BG (1982) A central limit theorem of non-commutative probability theory. *Theory of Probability and its Applications* 27: 703.

Hudson RL (1973) A quantum mechanical central limit theorem for anti-commuting observables. *Journal of Applied Probability* 10: 502–509.

Ibragimov IA and Linnick YuV (1971) *Independent and stationary sequences of random variables*. Groningen: Wolters-Noordhoff.

Inönü E and Wigner EP (1953) On the contraction groups and their representations. *Proceedings of the National Academy of Sciences, USA* 39: 510–524.

Lanford DE and Ruelle D (1969) Observables at infinity and states with short-range correlations in statistical mechanics. *Journal of Mathematical Physics* 13: 194.

Manuceau J, Sirugue M, Testard D, and Verbeure A (1973) The smallest C^*-algebra for canonical commutation relations. *Communications in Mathematical Physics* 32: 231.

Matsui T (2003) On the algebra of fluctuations in quantum spin chains. *Annales Henri Poincaré* 4: 63–83.

Michoel T and Verbeure A (2001) Goldstone boson normal coordinates. *Communications in Mathematical Physics* 216: 461–490.

Momont B, Verbeure A, and Zagrebnov VA (1997) Algebraic structure of quantum fluctuations. *Journal of Statistical Physics* 89: 633–653.

Quaegebeur J (1984) A non-commutative central limit theorem for CCR-algebras. *Journal of Functional Analysis* 57: 1–20.

Sewell GL (1986) *Quantum theory of collective phenomena*. Oxford: Oxford University Press.

Verbeure A and Zagrebnov VA (1992) Phase transitions and algebra of fluctuation operators in an exactly soluble model of a quantum anharmonic crystal. *J. Stat. Phys.* 69: 329–359.

Quantum Channels: Classical Capacity

A S Holevo, Steklov Mathematical Institute, Moscow, Russia

The Definition

A numerical measure of the ability of a classical or quantum information processing system (for definiteness, one speaks of a *communication channel*) to transmit information expressible as a text message (called "classical information" as distinct from quantum information). It is equal to the least upper bound for rates of the asymptotically perfect transmission of classical information through the system, when the transmission time tends to infinity, and arbitrary pre- and post-processing (encoding and decoding) are allowed at the input and the output of the system. Typically, for rates exceeding the capacity, not only the asymptotically perfect transmission is impossible, but the error probability with arbitrary encoding–decoding scheme tends to 1, so that the capacity has a nature of a threshold parameter.

From Classical to Quantum Information Theory

A central result of the classical information theory is the *Shannon coding theorem*, giving an explicit expression to the capacity in terms of the maximal mutual information between the input and the output of the channel. The issue of the information capacity of quantum communication channels arose soon after the publication of the pioneering papers by Shannon and goes back to the classical works of Gabor, Brillouin, and Gordon, asking for fundamental physical limits on the rate and quality of information transmission. This work laid a physical foundation and raised the question of consistent quantum treatment of the problem. Important steps in this direction were made in the early 1970s when a quantum probabilistic framework for this type of problem was created and the conjectured upper bound for the classical capacity of quantum channel was proved. A long journey to the quantum coding theorem culminated in 1996 with the proof of achievability of the upper bound (the Holevo–Schumacher–Westmoreland theorem; see Holevo (1998) for a detailed historical survey). Moreover, it was realized that quantum channel is characterized by the whole spectrum of capacities depending on the nature of the information resources and the specific protocols used for the transmission. To a great extent, this progress was stimulated by an interplay between the quantum communication theory and quantum information ideas related to more recent development in quantum computing. This new age of quantum information science is characterized by emphasis on the new possibilities (rather than restrictions) opened by the quantum nature of the information processing agent. On the other hand, the question of information capacity is important for the theory of quantum computer, particularly in connection with quantum error-correcting codes, communication and algorithmic complexity, and a number of other important issues.

The Quantum Coding Theorem

In the simplest and most basic memoryless case, the information processing system is described by the sequence of block channels,

$$\Phi^{\otimes n} = \underbrace{\Phi \otimes \cdots \otimes \Phi}_{n}, \quad n = 1, \ldots$$

of n parallel and independent uses of a channel Φ, n playing the role of transmission time (Holevo 1998). More generally, one can consider memory channels given by open dynamical systems with a kind of ergodic behavior and the limit where the transmission time goes to infinity (Kretschmann and Werner 2005).

Restricting to the memoryless case, encoding is given by a mapping of classical messages x from a given codebook of size N into states (density operators) $\rho_x^{(n)}$ in the input space $\mathcal{H}_1^{\otimes n}$ of the block channel $\Phi^{\otimes n}$, and decoding – by an observable $M^{(n)}$ in the output space $\mathcal{H}_2^{\otimes n}$, that is, a family $\{M_y^{(n)}\}$ of operators constituting a resolution of the identity in $\mathcal{H}_2^{\otimes n}$:

$$M_y^{(n)} \geq 0, \qquad \sum_y M_y^{(n)} = I$$

Here y plays the role of outcomes of the whole decoding procedure involving both the quantum measurement at the output and the possible classical information post-processing. Then the diagram for the classical information transmission is

$$x \to \underbrace{\rho_i^{(n)}}_{\substack{\text{input}\\ \text{state}}} \to \underbrace{\Phi^{\otimes n}[\rho_x^{(n)}]}_{\substack{\text{output}\\ \text{state}}} \overset{M^{(n)}}{\to} y$$

The such-described encoding and decoding constitute a quantum block code of length n and size N for the memoryless channel. The conditional probability of obtaining an outcome y provided the message x was sent for a chosen block code is given by the statistical formula

$$p^{(n)}(y|x) = \operatorname{tr} \Phi^{\otimes n}[\rho_x^{(n)}] M_y^{(n)}$$

and the error probability for the code is just $\max_x (1 - p^{(n)}(x\,|\,x))$.

Denoting by $p_e(n, N)$ the infimum of the error probability over all codes of length n and size N, the classical capacity $C(\Phi)$ of the memoryless channel is defined as the least upper bound of the rates R for which $\lim_{n\to\infty} p_e(n, 2^{nR}) = 0$.

Let Φ be a quantum channel from the input to the output quantum systems, assumed to be finite dimensional. The *coding theorem* for the classical capacity says that

$$C(\Phi) = \lim_{n\to\infty} \frac{1}{n} C_\chi(\Phi^{\otimes n}) \qquad [1]$$

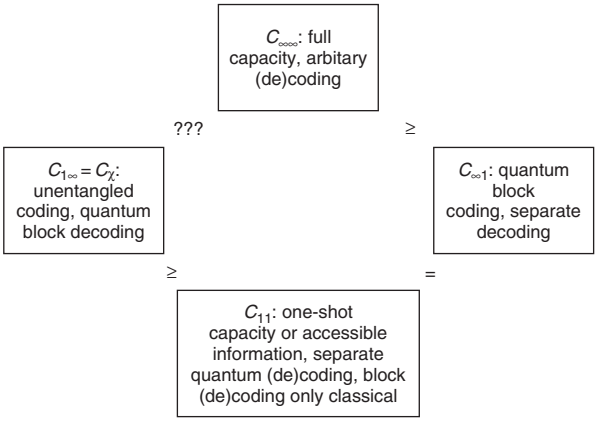
where

$$C_\chi(\Phi) = \max\left\{ H\left(\sum_x p_x \Phi[\rho_x]\right) - \sum_x p_x H(\Phi[\rho_x]) \right\} \qquad [2]$$

$H(\rho) = -\operatorname{tr} \rho \log_2 \rho$ is the binary von Neumann entropy, and the maximum is taken over all probability distributions $\{p_x\}$ and collections of density operators $\{\rho_x\}$ in $\mathcal{H}_1$.

The Variety of Capacities

This basic definition and the formulas [1], [2] generalize the definition of the Shannon capacity and the coding theorem for classical memoryless channels. For quantum channel, there are several different capacities because one may consider sending different kinds (classical or quantum) of information, restrict the admissible coding and decoding operations, and/or allow the use of additional resources, such as shared entanglement, forward or backward communication, leading to really different quantities (Bennett *et al.* 2004). Few of these resources (such as feedback) also exist for classical channels but usually influence the capacity less dramatically (at least for memoryless channels). Restricting to the transmission of classical information with no additional resources, one can distinguish at least four capacities (Bennett and Shor 1998), according to whether, for each block length n, one is allowed to use arbitrary entangled quantum operations on the full block of input (resp. output) systems, or if, for each of the parallel channels, one has to use a separate quantum encoding (resp. decoding), and combine these only by classical pre- (resp. post-) processing:



The full capacity $C_{\infty\infty}$ is just the classical capacity $C(\Phi)$ given by [1]. That $C_{1\infty}$ coincides with the

quantity $C_\chi(\Phi)$ given by [2] is the essential content of the HSW theorem, from which [1] is obtained by additional blocking. Since C_χ is apparently superadditive, $C_\chi(\Phi_1 \otimes \Phi_2) \geq C_\chi(\Phi_1) + C_\chi(\Phi_2)$, one has $C_{\infty\infty} \geq C_\chi$. It is still not known whether the quantity $C_\chi(\Phi)$ is in fact additive for all channels, which would imply the equalities here. Additivity of $C_\chi(\Phi)$ would have the important physical consequence – it would mean that using entangled input states does not increase the classical capacity of quantum channel. While such a result would be very much welcome, giving a single-letter expression for the classical capacity, it would call for a physical explanation of asymmetry between the effects of entanglement in encoding and decoding procedures. Indeed, the inequality in the lower left is known to be strict sometimes (Holevo 1998), which means that entangled decodings *can* increase the classical capacity. There is even an intermediate capacity between C_{11} and $C_{1\infty}$ obtained by restricting the quantum block decodings to adaptive ones (Shor 2002). The additivity of the quantity C_χ for *all* channels is one of the central open problems in quantum information theory; it was shown to be equivalent to several other important open problems, notably (super)additivity of the entanglement of formation and additivity of the minimal output entropy (Shor 2004).

For infinite-dimensional quantum processing systems, one needs to consider the input constraints such as the power constraint for bosonic Gaussian channels. The definition of the classical capacity and the capacity formula are then modified by introducing the constraint in a way similar to the classical theory (Holevo 1998, Holevo and Werner 2001). Another important extension concerns multiuser quantum information processing systems and their capacity regions (Devetak and Shor 2003).

See also: Capacities Enhanced by Entanglement; Capacity for Quantum Information; Channels in Quantum Information Theory; Entanglement Measures.

Further Reading

Bennett CH and Shor PW (1998) Quantum information theory. *IEEE Transactions on Information Theory* 44: 2724–2742.

Bennett CH, Devetak I, Shor PW, and Smolin JA Inequality and separation between assisted capacities of quantum channel, e-print quant-ph/0406086.

Devetak I and Shor PW The capacity of quantum channel for simultaneous transmission of classical and quantum information, e-print quant-ph/0311131.

Holevo AS (1998) Quantum Coding Theorems. *Russian Math. Surveys* vol. 53. pp. 1295–1331, e-print quant-ph/9808023.

Holevo AS (2000) Coding theorems of quantum information theory. In: Grigoryan A, Fokas A, Kibble T, and Zegarlinski B (eds.) *Proc. XIII ICMP*, pp. 415–422. London: International Press of Boston.

Holevo AS and Werner RF (2001) Evaluating capacities of bosonic Gaussian channels. *Physical Review A* 63: 032312 (e-print quant-ph/9912067).

Kretschmann D and Werner RF Quantum channels with memory, e-print quant-ph/0502106.

Shor PW The adaptive classical capacity of a quantum channel, or information capacity of 3 symmetric pure states in three dimensions, e-print quant-ph/0206058.

Shor PW (2004) Equivalence of additivity questions in quantum information theory. *Communications in Mathematical Physics* 246: 4334–4340 (e-print quant-ph/0305035).

Quantum Chromodynamics

G Sterman, Stony Brook University, Stony Brook, NY, USA

Introduction

Quantum chromodynamics, or QCD, as it is normally called in high-energy physics, is the quantum field theory that describes the strong interactions. It is the SU(3) gauge theory of the current standard model for elementary particles and forces, $SU(3) \times SU(2)_L \times U(1)$, which encompasses the strong, electromagnetic, and weak interactions. The symmetry group of QCD, with its eight conserved charges, is referred to as color SU(3). As is characteristic of quantum field theories, each field may be described in terms of quantum waves or particles.

Because it is a gauge field theory, the fields that carry the forces of QCD transform as vectors under the Lorentz group. Corresponding to these vector fields are the particles called "gluons," which carry an intrinsic angular momentum, or spin, of 1 in units of $\hbar$. The strong interactions are understood as the cumulative effects of gluons, interacting among themselves and with the quarks, the spin-1/2 particles of the Dirac quark fields.

There are six quark fields of varying masses in QCD. Of these, three are called "light" quarks, in a sense to be defined below, and three "heavy." The light quarks are the up (u), down (d), and strange (s), while the heavy quarks are the charm (c), bottom (b),

and top (t). Their well-known electric charges are $e_f = 2e/3(u, c, t)$ and $e_f = -e/3(d, s, b)$, with e the positron charge. The gluons interact with each quark field in an identical fashion, and the relatively light masses of three of the quarks provide the theory with a number of approximate global symmetries that profoundly influence the manner in which QCD manifests itself in the standard model.

These quark and gluon fields and their corresponding particles are enumerated with complete confidence by the community of high-energy physicists. Yet, none of these particles has ever been observed in isolation, as one might observe a photon or an electron. Rather, all known strongly interacting particles are colorless; most are "mesons," combinations with the quantum numbers of a quark q and a antiquark $\bar{q}'$, or "baryons" with the quantum numbers of (possibly distinct) combinations of three quarks $qq'q''$. This feature of QCD, that its underlying fields never appear as asymptotic states, is called "confinement." The very existence of confinement required new ways of thinking about field theory, and only with these was the discovery and development of QCD possible.

The Background of QCD

The strong interactions have been recognized as a separate force of nature since the discovery of the neutron as a constituent of atomic nuclei, along with the proton. Neutrons and protons (collectively, nucleons) possess a force, attractive at intermediate distances and so strong that it overcomes the electric repulsion of the protons, each with charge e. A sense of the relative strengths of the electromagnetic and strong interactions may be inferred from the typical distance between mutually repulsive electrons in an atom, $\sim 10^{-8}$ cm, and the typical distance between protons in a nucleus, of order 10^{-13} cm.

The history that led up to the discovery of QCD is a fascinating one, beginning with Yukawa's 1935 theory of pion exchange as the source of the forces that bind nuclei, still a useful tool for low-energy scattering. Other turning points include the creation of nonabelian gauge theories by Yang and Mills in 1954, the discovery of the quantum number known as strangeness, the consequent development of the quark model, and then the proposal of color as a global symmetry. The role of pointlike constituents in hadrons was foreshadowed by the identification of electromagnetic and weak currents and the analysis of their quantum-mechanical algebras. Finally, the observation of "scaling" in deep-inelastic scattering, which we will describe below, made QCD, with color as a local symmetry, the unique explanation of the strong interactions, through its property of asymptotic freedom.

The Lagrangian and Its Symmetries

The QCD Lagrangian may be written as

$$\mathcal{L} = \sum_{f=1}^{n_f} \bar{q}_f \left(\mathrm{i}\, \not{D}[A] - m_f\right) q_f - \frac{1}{2}\,\mathrm{tr}\left[F_{\mu\nu}^2(A)\right] - \frac{\lambda}{2}\left(B_a(A)\right)^2 + \bar{c}_b \left[\frac{\delta B_b(A)}{\delta \alpha_a}\right] c_a \qquad [1]$$

with $\not{D}[A] = \gamma \cdot \partial + \mathrm{i} g_s \gamma \cdot A$ the covariant derivative in QCD. The γ^μ are the Dirac matrices, satisfying the anticommutation relations, $[\gamma^\mu, \gamma^\nu]_+ = 2g^{\mu\nu}$. The SU(3) gluon fields are $A^\mu = \sum_{a=1}^{8} A_a^\mu T_a$, where T_a are the generators of SU(3) in the fundamental representation. The field strengths $F_{\mu\nu}[A] = \partial_\mu A_\nu - \partial_\nu A_\mu + \mathrm{i} g_s [A_\mu, A_\nu]$ specify the three- and four-point gluon couplings of nonabelian gauge theory. In QCD, there are $n_f = 6$ flavors of quark fields, q_f, with conjugate $\bar{q}_f = q_f^\dagger \gamma^0$.

The first two terms in the expression [1] make up the classical Lagrangian, followed by the gauge-fixing term, specified by a (usually, but not necessarily linear) function $B_a(A)$, and the ghost Lagrangian. The ghost (anti-ghost) fields $c_a(\bar{c}_a)$ carry the same adjoint index as the gauge fields.

The classical QCD Lagrangian before gauge fixing is invariant under the local gauge transformations

$$\begin{aligned} A'_\mu(x) &= \frac{\mathrm{i}}{g_s} \partial_\mu \Omega(x) \Omega^{-1}(x) + \Omega(x) A'_\mu(x) \Omega^{-1}(x) \\ &= A_\mu(x) - \partial_\mu \delta\alpha(x) \\ &\quad + \mathrm{i} g_s \left[\delta\alpha(x), A_\mu(x)\right] + \cdots \\ \psi'_i(x) &= \Omega(x)_{ij} \psi_j(x) = \psi_i(x) \\ &\quad + \mathrm{i} g_s \delta\alpha(x)_{ij} \psi_j(x) + \cdots \\ \delta\alpha(x) &= \sum_{a=1}^{8} \delta\alpha_a(x) T_a \end{aligned} \qquad [2]$$

The full QCD action including gauge-fixing and ghost terms is also invariant under the Bechi, Rouet, Stora, Tyutin (BRST) transformations with $\delta\xi$ an anticommuting variable.

$$\begin{aligned} \delta A_{\mu,a} &= \left(\delta_{ab}\partial_\mu + g A_{\mu c} f_{abc}\right) c_b \delta\xi \\ \delta c_a &= -\tfrac{1}{2} g C_{abc}\, c_b\, c_c \delta\xi, \quad \delta\bar{c} = \lambda B_a \delta\xi \\ \delta\psi_i &= \mathrm{i} g [T_b]_{ij}\, c_b \psi_j \end{aligned} \qquad [3]$$

with f_{abc} the SU(3) structure constants. The Jacobian of these transformations is unity.

In addition, neglecting masses of the light quarks, u, d, and s, the QCD Lagranian has a class of global flavor and chiral symmetries, the latter connecting left- and right-handed components of the quark fields, $\psi_{\mathrm{L,R}} \equiv (1/2)(1 \mp \gamma_5)\psi$,

$$\psi'(x) = \mathrm{e}^{\mathrm{i}\alpha\gamma_5^P} \psi(x), \quad P = 0, 1 \qquad [4]$$

Here, power $P=0$ describes phase, and $P=1$ chiral, transformations. Both transformations can be extended to transformations among the light flavors, by letting ψ become a vector, and α an element in the Lie algebra of SU(M), with $M=2$ if we take only the u and d quarks, and $M=3$ if we include the somewhat heavier strange quark. These symmetries, not to be confused with the local symmetries of the standard model, are strong isospin and its extension to the "eightfold way," which evolved into the (3-)quark model of Gell–Mann and Zweig. The many successes of these formalisms are automatically incorporated into QCD.

Green Functions, Phases, and Gauge Invariance

In large part, the business of quantum field theory is to calculate Green functions,

$$G_n(x_1 \ldots x_n) = \langle 0|T(\Phi_1(x_1)\ldots\Phi_i(x_i)\ldots\Phi_n(x_n))|0\rangle \quad [5]$$

where T denotes time ordering. The $\Phi_i(x)$ are elementary fields, such as A or q_f, or composite fields, such as currents like $J^\mu=\bar{q}_f\gamma^\mu q_f$. Such a Green function generates amplitudes for the scattering of particles of definite momenta and spin, when in the limit of large times the x_i-dependence of the Green function is that of a plane wave. For example, we may have in the limit $x_i^0 \to \infty$,

$$G_n(x_1\ldots x_n) \to \phi_i(p,\lambda)\,\mathrm{e}^{\mathrm{i}p\cdot x_i}\langle(p,\lambda)|T(\Phi_1(x_1)\ldots \Phi_{i-1}(x_{i-1})\Phi_{i+1}(x_{i+1})\ldots\Phi_n(x_n))|0\rangle \quad [6]$$

where $\phi_i(p,\lambda)$ is a solution to the free-field equation for field Φ_i, characterized by momentum p and spin λ. (An inegral over possible momenta p is understood.) When this happens for field i, the vacuum state is replaced by $|(p,\lambda)\rangle$, a particle state with precisely this momentum and spin; when it occurs for all fields, we derive a scattering (S)-matrix amplitude. In essence, the statement of confinement is that Green functions with fields $q_f(x)$ never behave as plane waves at large times in the past or future. Only Green functions of color singlet composite fields, invariant under gauge transformations, are associated with plane wave behavior at large times.

Green functions remain invariant under the BRST transformations [3], and this invariance implies a set of Ward identities

$$\frac{\delta}{\delta\xi(z)}\sum_{i=1}^{n}\langle 0|T(\Phi_1(x_1)\ldots\delta_{\mathrm{BRS}}\Phi_i(x_i)\ldots \Phi_n(x_n))|0\rangle = 0 \quad [7]$$

The variation of the anti-ghost as in [3] is equivalent to an infinitesimal change in the gauge-fixing term; variations in the remaining fields all cancel single-particle plane wave behavior in the corresponding Green functions. These identities then ensure the gauge invariance of the perturbative S-matrix, a result that turns out to be useful despite confinement.

To go beyond a purely perturbative description of QCD, it is useful to introduce a set of nonlocal operators that are variously called nonabelian phases, ordered exponentials, and Wilson lines,

$$U_C(z,y) = P\,\exp\left[-\mathrm{i}g_{\mathrm{s}}\int_y^z \mathrm{d}x^\mu A_\mu(x)\right] \quad [8]$$

where C is some self-avoiding curve between y and z. The U's transform at each end linearly in nonabelian gauge transformations $\Omega(x)$ at that point,

$$U'_C(z,y) = \Omega(z)U_C(z,y)\Omega^{-1}(y) \quad [9]$$

Especially interesting are closed curves C, for which $z=y$. The phases about such closed loops are, like their abelian counterparts, sensitive to the magnetic flux that they enclose, even when the field strengths vanish on the curve.

QCD at the Shortest and Longest Distances

Much of the fascination of QCD is its extraordinary variation of behavior at differing distance scales. Its discovery is linked to asymptotic freedom, which characterizes the theory at the shortest scales. Asymptotic freedom also suggests (and in part provides) a bridge to longer distances.

Most analyses in QCD begin with a path-integral formulation in terms of the elementary fields $\Phi_a = q_f\ldots$,

$$G_n(x_i,(z_j,y_j)) = \int\left[\prod_{a=q,\bar{q},G,c,\bar{c}}\mathcal{D}\Phi_a\right]\prod_i\Phi_i(x_i) \times\prod_j U_{C_j}(z_j,y_j)\,\mathrm{e}^{\mathrm{i}S_{\mathrm{QCD}}} \quad [10]$$

with S_{QCD} the action. Perturbation theory keeps only the kinetic Lagrangian, quadratic in fields, in the exponent, and expands the potential terms in the coupling. This procedure produces Feynman diagrams, with vertices corresponding to the cubic and quartic terms in the QCD Lagrangian [1].

Most nonperturbative analyses of QCD require studying the theory on a Euclean, rather than Minkowski space, related by an analytic continuation in the times x^0, y^0, z^0 in G_n from real to imaginary values. In Euclidean space, we find, for example,

classical solutions to the equations of motion, known as instantons, that provide nonperturbative contributions to the path integral. Perhaps the most flexible nonperturbative approach approximates the action and the measure at a lattice of points in four-dimensional space. For this purpose, integrals over the gauge fields are replaced by averages over "gauge links," of the form of eqn [8] between neighboring points.

Perturbation theory is most useful for processes that occur over short timescales and at high relative energies. Lattice QCD, on the other hand, can simulate processes that take much longer times, but is less useful when large momentum transfers are involved. The gap between the two methods remains quite wide, but between the two they have covered enormous ground, enough to more than confirm QCD as the theory of strong interactions.

Asymptotic Freedom

QCD is a renormalizable field theory, which implies that the coupling constant g must be defined by its value at a "renormalization scale," and is denoted $g(\mu)$. Usually, the magnitude of $\alpha_s(\mu) \equiv g^2/4\pi$, is quoted at $\mu = m_Z$, where it is ~ 0.12. In effect, $g(\mu)$ controls the amplitude that connects any state to another state with one more or one fewer gluon, including quantum corrections that occur over timescales from zero up to $\hbar/\mu$ (if we measure μ in units of energy). The QCD Ward identities mentioned above ensure that the coupling is the same for both quarks and gluons, and indeed remains the same in all terms in the Lagrangian, ensuring that the symmetries of QCD are not destroyed by renormalization.

Quantum corrections to gluon emission are not generally computable directly in renormalizable theories, but their dependence on μ is computable, and is a power series in $\alpha_s(\mu)$ itself,

$$\mu^2 \frac{d\alpha_s(\mu)}{d\mu^2} = -b_0 \frac{\alpha_s^2(\mu)}{4\pi} - b_1 \frac{\alpha_s^3(\mu)}{(4\pi)^2} + \cdots \equiv \beta(\alpha_s) \quad [11]$$

where $b_0 = 11 - 2n_f/3$ and $b_1 = 2(31 - 19n_f/3)$. The celebrated minus signs on the right-hand side are associated with both the spin and self-interactions of the gluons.

The solution to this equation provides an expression for α_s at any scale μ_1 in terms of its value at any other scale μ_0. Keeping only the lowest-order, b_0, term, we have

$$\alpha_s(\mu_1) = \frac{\alpha_s(\mu_0)}{1 + (b_0/4\pi)\ln(\mu_1^2/\mu_0^2)} = \frac{4\pi}{b_0 \ln\left(\mu_1^2/\Lambda_{QCD}^2\right)} \quad [12]$$

where in the second form, we have introduced Λ_{QCD}, the scale parameter of the theory, which embodies the condition that we get the same coupling at scale μ_1 no matter which scale μ_0 we start from. Asymptotic freedom consists of the observation that at larger renormalization masses μ, or correspondingly shorter timescales, the coupling weakens, and indeed vanishes in the limit $\mu \to \infty$. The other side of the coin is that over longer times or lower momenta, the coupling grows. Eventually, near the pole at $\mu_1 = \Lambda_{QCD}$, the lowest-order approximation to the running fails, and the theory becomes essentially nonperturbative. Thus, the discovery of asymptotic freedom suggested, although it certainly does not prove, that QCD is capable of producing very strong forces, and confinement at long distances. Current estimates of Λ_{QCD} are ~ 200 MeV.

Spontaneous Breaking of Chiral Symmetry

The number of quarks and their masses is an external input to QCD. In the standard model masses are provided by the Higgs mechanism, but in QCD they are simply parameters. Because the standard model has chosen several of the quarks to be especially light, QCD incorporates the chiral symmetries implied by eqn [4] (with $P = 1$). In the limit of zero quark masses, these symmetries becomes exact, respected to all orders of perturbation theory, that is, for any finite number of gluons emitted or absorbed.

At distances on the order to $1/\Lambda_{QCD}$, however, QCD cannot respect chiral symmetry, which would require each state to have a degenerate partner with the opposite parity, something not seen in nature. Rather, QCD produces, nonperturbatively, nonzero values for matrix elements that mix right- and left-handed fields, such as $\langle 0|\bar{u}_L u_R|0\rangle$, with u the up-quark field. Pions are the Goldstone bosons of this symmetry, and may be thought of as ripples in the chiral condensate, rotating it locally as they pass along. The observation that these Goldstone bosons are not exactly massless is due to the "current" masses of the quarks, their values in $\mathcal{L}_{QCD}$. The (chiral perturbation theory) expansion in these light-quark masses also enables us to estimate them quantitatively: $1.5 \le m_u \le 4$ MeV, $4 \le m_d \le 8$ MeV, and $80 \le m_s \le 155$ MeV. These are the light quarks, with masses smaller than Λ_{QCD}. (Like α_s, the masses are renormalized; these are quoted from Eidelman (2004) with $\mu = 2$ GeV.) For comparison, the heavy quarks have masses $m_c \sim 1$–1.5 GeV, $m_b \sim 4$–4.5 GeV, and $m_t \sim 180$ GeV (the giant among the known elementary particles).

Although the mechanism of the chiral condensate (and in general other nonperturbative aspects of

QCD) has not yet been demonstrated from first principles, a very satisfactory description of the origin of the condensate, and indeed of much hadronic structure, has been given in terms of the attractive forces between quarks provided by instantons. The actions of instanton solutions provide a dependence $\exp[-8\pi^2/g_s^2]$ in Euclidean path integrals, and so are characteristically nonperturbative.

Mechanisms of Confinement

As described above, confinement is the absence of asymptotic states that transform nontrivially under color transformations. The full spectrum of QCD, however, is a complex thing to study, and so the problem has been approached somewhat indirectly. A difficulty is the same light-quark masses associated with approximate chiral symmetry. Because the masses of the light quarks are far below the scale Λ_{QCD} at which the perturbative coupling blows up, light quarks are created freely from the vacuum and the process of "hadronization," by which quarks and gluons form mesons and baryons, is both nonperturbative and relativistic. It is therefore difficult to approach in both perturbation theory and lattice simulations.

Tests and studies of confinement are thus normally formulated in truncations of QCD, typically with no light quarks. The question is then reformulated in a way that is somewhat more tractable, without relativistic light quarks popping in and out of the vacuum all the time. In the limit that its mass becomes infinite compared to the natural scale of fluctuations in the QCD vacuum, the propagator of a quark becomes identical to a phase operator, [8], with a path C corresponding to a constant velocity. This observation suggests a number of tests for confinement that can be implemented in the lattice theory. The most intuitive is the vacuum expectation value of a "Wilson loop," consisting of a rectangular path, with sides along the time direction, corresponding to a heavy quark and antiquark at rest a distance R apart, and closed at some starting and ending times with straight lines. The vacuum expectation value of the loop then turns out to be the exponential of the potential energy between the quark pair, multiplied by the elapsed time,

$$\left\langle 0 \left| P \exp\left[-ig_s \oint_C A_\mu(x)\, dx^\mu \right] \right| 0 \right\rangle = \exp(-V(R)T/\hbar) \qquad [13]$$

When $V(R) \propto R$ ("area law" behavior), there is a linearly rising, confining potential. This behavior, not yet proven analytically yet well confirmed on the lattice, has an appealing interpretation as the energy of a "string," connecting the quark and antiquark, whose energy is proportional to its length.

Motivation for such a string picture was also found from the hadron spectrum itself, before any of the heavy quarks were known, and even before the discovery of QCD, from the observation that many mesonic ($\bar{q}q'$) states lie along "Regge trajectories," which consist of sets of states of spin J and mass m_J^2 that obey a relation

$$J = \alpha' m_J^2 \qquad [14]$$

for some constant α'. Such a relation can be modeled by two light particles ("quarks") revolving around each other at some constant (for simplicity, fixed nonrelativistic) velocity v_0 and distance $2R$, connected by a "string" whose energy per unit length is a constant ρ.

Suppose the center of the string is stationary, so the overall system is at rest. Then neglecting the masses, the total energy of the system is $M = 2R\rho$. Meanwhile, the momentum density per unit length at distance r from the center is $v(r) = (r/R)v_0$, and the total angular momentum of the system is

$$J = 2\rho v_0 \int_0^R dr\, r^2 = \frac{2\rho v_0}{3} R^2 = \frac{v_0}{6\rho} M^2 \qquad [15]$$

and for such a system, [14] is indeed satisfied. Quantized values of angular momentum J give quantized masses m_J, and we might take this as a sort of "Bohr model" for a meson. Indeed, string theory has its origin in related consideration in the strong interactions.

Lattice data are unequivocal on the linearly rising potential, but it requires further analysis to take a lattice result and determine what field configurations, stringlike or not, gave that result. Probably the most widely accepted explanation is in terms of an analogy to the Meissner effect in superconductivity, in which type II superconductors isolate magnetic flux in quantized tubes, the result of the formation of a condensate of Cooper pairs of electrons. If the strings of QCD are to be made of the gauge field, they must be electric ($F^{\mu 0}$) in nature to couple to quarks, so the analogy postulates a "dual" Meissner effect, in which electric flux is isolated as the result of a condensate of objects with magnetic charge (producing nonzero F^{ij}). Although no proof of this mechanism has been provided yet, the role of magnetic fluctuations in confinement has been widely investigated in lattice simulations, with encouraging results. Of special interest are magnetic field configurations, monopoles or vortices, in the Z_3 center of SU(3), $\exp[i\pi k/3] I_{3\times 3}$, $k = 0, 1, 2$. Such configurations, even when localized, influence closed gauge loops [13] through the nonabelian Aharonov–Bohm effect. Eventually, of course, the role of light quarks must be crucial for any complete

description of confinement in the real world, as emphasized by Gribov.

Another related choice of closed loop is the "Polyakov loop," implemented at finite temperature, for which the path integral is taken over periodic field configurations with period $1/T$, where T is the temperature. In this case, the curve C extends from times $t=0$ to $t=1/T$ at a fixed point in space. In this formulation it is possible to observe a phase transition from a confined phase, where the expectation is zero, to a deconfined phase, where it is nonzero. This phase transition is currently under intense experimental study in nuclear collisions.

Using Asymptotic Freedom: Perturbative QCD

It is not entirely obvious how to use asymptotic freedom in a theory that should (must) have confinement. Such applications of asymptotic freedom go by the term perturbative QCD, which has many applications, not the least as a window to extensions of the standard model.

Lepton Annihilation and Infrared Safety

The electromagnetic current, $J_\mu = \sum_f e_f \bar{q}_f \gamma_\mu q_f$, is a gauge-invariant operator, and its correlation functions are not limited by confinement. Perhaps, the simplest application of asymptotic freedom, yet of great physical relevance, is the scalar two-point function,

$$\pi(Q) = \frac{-\mathrm{i}}{3}\int \mathrm{d}^4x\, \mathrm{e}^{-\mathrm{i}Q\cdot x}\langle 0|\mathrm{T}\big(J^\mu(0)J_\mu(x)\big)|0\rangle \qquad [16]$$

The imaginary part of this function is related to the total cross section for the annihilation process $\mathrm{e}^+\mathrm{e}^- \to$ hadrons in the approximation that only one photon takes part in the reaction. The specific relation is $\sigma_{\mathrm{QCD}} = (e^4/Q^2)\,\mathrm{Im}\,\pi(Q^2)$, which follows from the optical theorem, illustrated in Figure 1. The perturbative expansion of the function $\pi(Q)$ depends, in general, on the mass scales Q and the quark masses m_f as well as on the strong coupling $\alpha_s(\mu)$ and on the renormalization scale μ. We may also worry about the influence of other, truly nonperturbative scales, proportional to powers of Λ_{QCD}. At large values of Q^2, however, the situation simplifies greatly, and dependence on all scales below Q is suppressed by powers of Q. This may be expressed in terms of the operator product expansion,

$$\begin{aligned}\langle 0|T\big(J^\mu(0)J_\mu(x)\big)|0\rangle &= \sum_{O_I}(x^2)^{-3+d_I/2}C_I(x^2\mu^2,\alpha_s(\mu)) \\ &\quad\times\langle 0|O_I(0)|0\rangle \end{aligned} \qquad [17]$$

where d_I is the mass dimension of operator O_I, and where the dimensionless coefficient functions C_I incorporate quantum corrections. The sum over operators begins with the identity ($d_I=0$), whose coefficient function is identified with the sum of quantum corrections in the approximation of zero masses. The sum continues with quark mass corrections, which are suppressed by powers of at least m_f^2/Q^2, for those flavors with masses below Q. Any QCD quantity that has this property, remaining finite in perturbation theory when all particle masses are set to zero, is said to be "infrared safe."

The effects of quarks whose masses are above Q are included indirectly, through the couplings and masses observed at the lower scales. In summary, the leading power behavior of $\pi(Q)$, and hence of the cross section, is a function of Q, μ, and $\alpha_s(\mu)$ only. Higher-order operators whose vacuum matrix elements receive nonperturbative corrections include the "gluon condensate," identified as the product $\alpha_s(\mu)G_{\alpha\beta}G^{\alpha\beta} \propto \Lambda^4_{\mathrm{QCD}}$.

Once we have concluded that Q is the only physical scale in π, we may expect that the right choice of the renormalization scale is $\mu = Q$. Any observable quantity is independent of the choice of renormalization scale, μ, and neglecting quark masses, the chain rule gives

$$\mu\frac{\mathrm{d}\sigma(Q/\mu,\alpha_s(\mu))}{\mathrm{d}\mu} = \mu\frac{\partial\sigma}{\partial\mu} + 2\beta(\alpha_s)\frac{\partial\sigma}{\partial\alpha_s} = 0 \qquad [18]$$

which shows that we can determine the beta function directly from the perturbative expansion of the cross section. Defining $a \equiv \alpha_s(\mu)/\pi$, such a perturbative calculation gives

$$\begin{aligned}\mathrm{Im}\pi(Q^2) = \frac{3}{4\pi}\sum_f e_f^2\Bigg(1 + a + a^2\Bigg(1.986 \\ - 0.115n_f - (b_0/4\pi)\ln\frac{Q^2}{\mu^2}\Bigg)\Bigg)\end{aligned} \qquad [19]$$

with b_0 as above. Now, choosing $\mu = Q$, we see that asymptotic freedom implies that when Q is large, the total cross section is given by the lowest order,

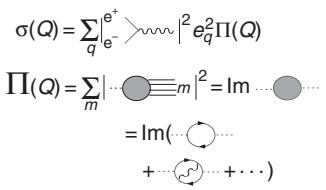


Figure 1 First line: schematic relation of lowest order $\mathrm{e}^+\mathrm{e}^-$ annihilation to sum over quarks q, each with electric charge e_q. Second line: perturbative unitarity for the current correlation function $\pi(Q)$.

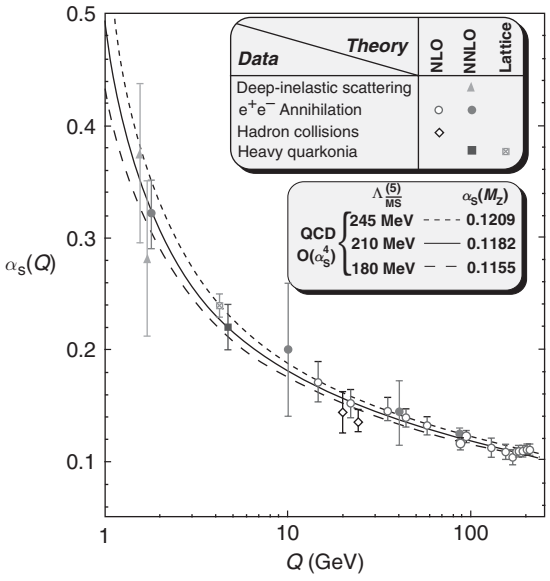


Figure 2 Experimental variation of the strong coupling with scales. Reproduced from Bethke S (2004) Alpha(s) at Zinnowitz. *Nuclear Physics Proceedings Supplements* 135: 345–352, with permission from Elsevier.

plus small and calculable QCD corrections, a result that is borne out in experiment. Comparing experiment to an expression like [19], one can measure the value of $\alpha_s(Q)$, and hence, with eqn [12], $\alpha_s(\mu)$ for any $\mu \gg \Lambda_{QCD}$. Figure 2 shows a recent compilation of values of α_s from this kind of analysis in different experiments at different scales, clearly demonstrating asymptotic freedom.

Factorization, Scaling, and Parton distributions

One step beyond vacuum matrix elements of currents are their expectation values in single-particle states, and here we make contact with the discovery of QCD, through scaling. Such expectations are relevant to the class of experiments known as deep-inelastic scattering, in which a high-energy electron exchanges a photon with a nucleon target. All QCD information is contained in the tensor matrix element

$$W_N^{\mu\nu}(p,q) \equiv \frac{1}{8\pi}\sum_\sigma \int d^4x\, e^{-iq\cdot x}\langle p,\sigma|J^\mu(0)J^\nu(x)|p,\sigma\rangle \quad [20]$$

with q the momentum transfer carried by the photon, and p,σ the momentum and spin of the target nucleon, N. This matrix element is not infrared safe, since it depends in principle on the entire history of the nucleon state. Thus, it is not accessible to direct perturbative calculation.

Nevertheless, when the scattering involves a large momentum transfer compared to Λ_{QCD}, we may

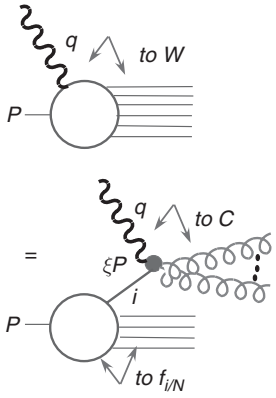


Figure 3 Schematic depiction of factorization in deep-inelastic scattering.

expect a quantum-mechanical incoherence between the scattering reaction, which occurs (by the uncertainty principle) at short distances, and the forces that stabilize the nucleon. After all, we have seen that the latter, strong forces, should be associated with long distances. Such a separation of dynamics, called factorization, can be implemented in perturbation theory, and is assumed to be a property of full QCD. Factorization is illustrated schematically in Figure 3. Of course, short and long distances are relative concepts, and the separation requires the introduction of a so-called factorization scale, μ_F, not dissimilar to the renormalization scale described above. For many purposes, it is convenient to choose the two equal, although this is not required.

The expression of factorization for deep-inelastic scattering is

$$W_N^{\mu\nu}(p,q) = \sum_{i=q_f,\bar{q}_f,G}\int_x^1 d\xi C_i^{\mu\nu}(\xi p,q,\mu_F,\alpha_s(\mu_F)) \times f_{i/N}(\xi,\mu_F) \quad [21]$$

where the functions $C_i^{\mu\nu}$ (the coefficient functions) can be computed as an expansion in $\alpha_s(\mu_F)$, and describe the scattering of the "partons," quarks, and gluons, of which the target is made. The variable ξ ranges from unity down to $x \equiv -q^2/2p\cdot q > 0$, and has the interpretation of the fractional momentum of the proton carried by parton i. (Here $-q^2 = Q^2$ is positive.) The parton distributions $f_{i/N}$ can be defined in terms of matrix elements in the nucleon, in which the currents are replaced by quark (or antiquark or gluon) fields, as

$$f_{q/N}(x,\mu) = \frac{1}{4\pi}\int_{-\infty}^{\infty} d\lambda\, e^{-i\lambda x p^+} \times \langle p,\sigma|\bar{q}(\lambda n)U_n(n\lambda,0)n\cdot\gamma q(0)|p,\sigma\rangle \quad [22]$$

n^μ is a light-like vector, and U_n a phase operator whose path C is in the n-direction. The dependence of the parton distribution on the factorization scale is through the renormalization of the composite operator consisting of the quark fields, separated along the light cone, and the nonabelian phase operator $U_n(n\lambda, 0)$, which renders the matrix element gauge invariant by eqn [9]. By combining the calculations of the C's and data for $W_N^{\mu\nu}$, we can infer the parton distributions, $f_{i/N}$. Important factorizations of a similar sort also apply to some exclusive processes, including amplitudes for elastic pion or nucelon scattering at large momentum transfer.

Equation [21] has a number of extraordinary consequences. First, because the coefficient function is an expansion in α_s, it is natural to choose $\mu_F^2 \sim Q^2 \sim p \cdot q$ (when x is of order unity). When Q is large, we may approximate $C_i^{\mu\nu}$ by its lowest order, which is first order in the electromagnetic coupling of quarks to photons, and zeroth order in α_s. In this approximation, dependence on Q is entirely in the parton distributions. But such dependence is of necessity weak (again for x not so small as to produce another scale), because the μ_F dependence of $f_{i/N}(\xi, \mu_F)$ must be compensated by the μ_F dependence of $C_i^{\mu\nu}$, which is order α_s. This means that the overall Q dependence of the tensor $W_N^{\mu\nu}$ is weak for Q large when x is moderate. This is the scaling phenomenon that played such an important role in the discovery of QCD.

Evolution: Beyond Scaling

Another consequence of the factorization [21], or equivalently of the operator definition [22], is that the μ_F-dependence of the coefficient functions and the parton distributions are linked. As in the lepton annihilation cross section, this may be thought of as due to the independence of the physically observable tensor $W_N^{\mu\nu}$ from the choice of factorization and renormalization scales. This implies that the μ_F-dependence of $f_{i/N}$ may be calculated perturbatively since it must cancel the corresponding dependence in C_i. The resulting relation is coventionally expressed in terms of the "evolution equations,"

$$\mu \frac{\mathrm{d}f_{a/N}(x,\mu)}{\mathrm{d}\mu} = \sum_c \int_x^1 \mathrm{d}\xi \; P_{ac}(x/\xi, \alpha_s(\mu)) f_{c/N}(\xi, \mu) \qquad [23]$$

where $P_{ac}(\xi)$ are calculable as power series, now known up to α_s^3. This relation expands the applicability of QCD from scales where parton distributions can be inferred directly from experiment, to arbitrarily high scales, reachable in accelerators under construction or in the imagination, or even on the cosmic level.

At very high energy, however, the effective values of the variable x can become very small and introduce new scales, so that eventually the evolution of eqn [23] fails. The study of nuclear collisions may provide a new high-density regime for QCD, which blurs the distinction between perturbative and nonperturbative dynamics.

Inclusive Production

Once we have evolution at our disposal, we can take yet another step, and replace electroweak currents with any operator from any extension of QCD, in the standard model or beyond, that couples quarks and gluons to the particles of as-yet unseen fields. Factorization can be extended to these situations as well, providing predictions for the production of new particles, F of mass M, in the form of factorized inclusive cross sections,

$$\sigma_{AB\to F(M)}(M, p_A, p_B) = \sum_{i,j=q_f\bar{q}_f,G} \int \mathrm{d}\xi_a \, \mathrm{d}\xi_b f_{i/A}(\xi_a, \mu) f_{j/B}(\xi_b, \mu) \times H_{ij\to F(M)}(x_a p_A, x_b p_B, M, \mu, \alpha_s(\mu)) \qquad [24]$$

where the functions $H_{ij\to F}$ may be calculated perturbatively, while the $f_{i/A}$ and $f_{j/B}$ parton distributions are known from a combination of lower-energy observation and evolution. In this context, they are said to be "universal," in that they are the same functions in hadron–hadron collisions as in the electron–hadron collisions of deep-inelastic scattering. In general, the calculation of hard-scattering functions H_{ij} is quite nontrivial beyond lowest order in α_s. The exploration of methods to compute higher orders, currently as far as α_s^2, has required extraordinary insight into the properties of multidimensional integrals.

The factorization method helped predict the observation of the W and Z bosons of electroweak theory, and the discovery of the top quark. The extension of factorization from deep-inelastic scattering to hadron production is nontrivial; indeed, it only holds in the limit that the velocities, β_i, of the colliding particles approach the speed of light in the center-of-momentum frame of the produced particle. Corrections to the relation [24] are then at the level of powers of $\beta_i - 1$, which translates into inverse powers of the invariant mass(es) of the produced particle(s) M. Factorizations of this sort do not apply to low-velocity collisions. Arguments for this

result rely on relativistic causality and the uncertainty principle. The creation of the new state happens over timescales of order $1/M$. Before that well-defined event, the colliding particles are approaching at nearly the speed of light, and hence cannot affect the distributions of each others' partons. After the new particle is created, the fragments of the hadrons recede from each other, and the subsequent time development, when summed over all possible final states that include the heavy particle, is finite in perturbation theory as a direct result of the unitarity of QCD.

Structure of Hadronic Final States

A wide range of semi-inclusive cross sections are defined by measuring properties of final states that depend only on the flow of energy, and which bring QCD perturbation theory to the threshold of nonperturbative dynamics. Schematically, for a state $N = |k_1 \dots k_N\rangle$, we define $\mathcal{S}(N) = \sum_i s(\Omega_i) k_i^0$, where $s(\Omega)$ is some smooth function of directions. We generalize the e^+e^- annihilation case above, and define a cross section in terms of a related, but highly nonlocal, matrix element,

$$\frac{d\sigma(Q)}{d\mathcal{S}} \equiv \sigma_0 \int d^4x\, e^{-iQ\cdot x} \left\langle 0 \middle| J^\mu(0) \times \delta\left(\int d^2\Omega s(\Omega)\mathcal{E}(\Omega) - \mathcal{S} \right) J_\mu(x) \middle| 0 \right\rangle \qquad [25]$$

where σ_0 is a zeroth-order cross section, and where $\mathcal{E}$ is an operator at spatial infinity, which measures the energy flow of any state in direction Ω: $\mathcal{E}(\Omega)|k_1 \dots k_N\rangle = (1/Q)\sum_i k_i^0 \delta^2(\Omega - \Omega_i)$. This may seem a little complicated, but like the total annihilation cross section, the only dimensional scale on which it depends is Q. The operator $\mathcal{E}$ can be defined in a gauge-invariant manner, through the energy–momentum tensor for example, and has a meaning independent of partonic final states. At the same time, this sort of cross section may be implemented easily in perturbation theory, and like the total annihilation cross section, it is infrared safe. To see why, notice that when a massless ($k^2 = 0$) particle decays into two particles of momenta xk and $(1-x)k$ ($0 \le x \le 1$), the quantity $\mathcal{S}$ is unchanged, since the sum of the new energies is the same as the old. This makes the observable $\mathcal{S}(N)$ insensitive to processes at low momentum transfer.

For the case of leptonic annihilation, the lowest-order perturbative contribution to energy flow requires no powers of α_s, and consists of an oppositely moving quark and antiquark pair. Any measure of energy flow that includes these configurations will dominate over correlations that require α_s corrections. As a result, QCD predicts that in most leptonic annihilation events, energy will flow in two back-to-back collimated sets of particles, known as "jets." In this way, quarks and gluons are observed clearly, albeit indirectly.

With varying choices of $\mathcal{S}$, many properties of jets, such as their distributions in invariant mass, and the probabilities and angular distributions of multijet events, and even the energy dependence of their particle multiplicities, can be computed in QCD. This is in part because hadronization is dominated by the production of light quarks, whose production from the vacuum requires very little momentum transfer. Paradoxically, the very lightness of quarks is a boon to the use of perturbative methods. All these considerations can be extended to hadronic scattering, and jet and other semi-inclusive properties of final states also computed and compared to experiment.

Conclusions

QCD is an extremely broad field, and this article has hardly scratched the surface. The relation of QCD-like theories to supersymmetric and string theories, and implications of the latter for confinement and the computation of higher-order perturbative amplitudes, have been some of the most exciting developments of recent years. As another example, we note that the reduction of the heavy-quark propagator to a nonabelian phase, noted in our discussion of confinement, is related to additional symmetries of heavy quarks in QCD, with many consequences for the analysis of their bound states. Of the bibliography given below, one may mention the four volumes of Shifman (2001, 2002), which communicate in one place a sense of the sweep of work in QCD.

Our confidence in QCD as the correct description of the strong interactions is based on a wide variety of experimental and observational results. At each stage in the discovery, confirmation, and exploration of QCD, the mathematical analysis of relativistic quantum field theory entered new territory. As is the case for gravity or electromagnetism, this period of exploration is far from complete, and perhaps never will be.

See also: AdS/CFT Correspondence; Aharonov–Bohm Effect; BRST Quantization; Current Algebra; Dirac Operator and Dirac Field; Euclidean Field Theory; Effective Field Theories; Electroweak Theory; Lattice Gauge Theory; Operator Product Expansion in Quantum Field Theory; Perturbation Theory and its Techniques; Perturbative Renormalization Theory and BRST; Quantum Field Theory: A Brief Introduction; Random

Matrix Theory in Physics; Renormalization: General Theory; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools; Scattering, Asymptotic Completeness and Bound States; Seiberg–Witten Theory; Standard Model of Particle Physics.

Further Reading

Bethke S (2004) α_s at Zinnowitz, 2004. *Nuclear Physics Proceeding Supplements* 135: 345–352.

Brodsky SJ and Lepage P (1989) Exclusive processes in quantum chromodynamics. In: Mueller AH (ed.) *Perturbative Quantum Chromodynamics*. Singapore: World Scientific.

Collins JC, Soper DE, and Sterman G (1989) Factorization. In: Mueller AH (ed.) *Perturbative Quantum Chromodynamics*. Singapore: World Scientific.

Dokshitzer Yu L and Kharzeev DE (2004) Gribov's conception of quantum chromodynamics. *Annual Review of Nuclear and Particle Science* 54: 487–524.

Dokshitzer Yu L, Khoze V, Troian SI, and Mueller AH (1988) QCD coherence in high-energy reactions. *Reviews of Modern Physics* 60: 373.

Eidelman S *et al.* (2004) Review of particle physics. *Physics Letters B* 592: 1–1109.

Ellis RK, Stirling WJ, and Webber BR (1996) QCD and Collider Physics, *Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology*, vol. 8. Cambridge: Cambridge University Press.

Greensite J (2003) The confinement problem in lattice gauge theory. *Progress in Particle and Nuclear Physics* 51: 1.

Mandelstam S (1976) Vortices and quark confinement in nonabelian gauge theories. *Physics Reports* 23: 245–249.

Muta T (1986) *Foundations of Quantum Chromodynamics*. Singapore: World Scientific.

Neubert H (1994) Heavy quark symmetry. *Physics Reports* 245: 259–396.

Polyakov AM (1977) Quark confinement and topology of gauge groups. *Nuclear Physics B* 120: 429–458.

Schafer T and Shuryak EV (1998) Instantons and QCD. *Reviews of Modern Physics* 70: 323–426.

Shifman M (ed.) (2001) *At the Frontier of Particle Physics: Handbook of QCD*, vols. 1–3. River Edge, NJ: World Scientific.

Shifman M (ed.) (2002) *At the Frontier of Particle Physics: Hand book of QCD*, vol. 4. River Edge, NJ: World Scientific.

Sterman G (1993) *An Introduction to Quantum Field Theory*. Cambridge: Cambridge University Press.

't Hooft G (1977) On the phase transition towards permanent quark confinement (1978). *Nuclear Physics B* 138: 1.

't Hooft G (ed.) (2005) *Fifty Years of Yang–Mills Theories*. Hackensack: World Scientific.

Weinberg S (1977) The problem of mass. *Transactions of the New York Academy of Science* 38: 185–201.

Wilson KG (1974) Confinement of quarks. *Physical Review D* 10: 2445–2459.

Quantum Cosmology

M Bojowald, The Pennsylvania State University, University Park, PA, USA

Introduction

Classical gravity, through its attractive nature, leads to a high curvature in important situations. In particular, this is realized in the very early universe where in the backward evolution energy densities are growing until the theory breaks down. Mathematically, this point appears as a singularity where curvature and physical quantities diverge and the evolution breaks down. It is not possible to set up an initial-value formulation at this place in order to determine the further evolution.

In such a regime, quantum effects are expected to play an important role and to modify the classical behavior such as the attractive nature of gravity or the underlying spacetime structure. Any candidate for quantum gravity thus allows us to reanalyze the singularity problem in a new light which implies the tests of the characteristic properties of the respective candidate. Moreover, close to the classical singularity, in the very early universe, quantum modifications will give rise to new equations of motion which turn into Einstein's equations only on larger scales. The analysis of these equations of motion leads to new classes of early universe phenomenology.

The application of quantum theory to cosmology presents a unique problem with not only mathematical but also many conceptual and philosophical ramifications. Since by definition there is only one universe which contains everything accessible, there is no place for an outside observer separate from the quantum system. This eliminates the most straightforward interpretations of quantum mechanics and requires more elaborate, and sometimes also more realistic, constructions such as decoherence. From the mathematical point of view, this situation is often expected to be mirrored by a new type of theory which does not allow one to choose initial or boundary conditions separately from the dynamical laws. Initial or boundary conditions, after all, are meant to specify the physical system prepared for observations which is impossible in cosmology. Since we observe only one universe, the expectation goes, our theories should finally present us with only

one, unique solution without any freedom for further conditions. This solution then contains all the information about observations as well as observers. Mathematically, this is an extremely complicated problem which has received only scant attention. Equations of motion for quantum cosmology are usually of the type of partial differential or difference equations such that new ingredients from quantum gravity are needed to restrict the large freedom of solutions.

Minisuperspace approximation

In most investigations, the problem of applying full quantum gravity to cosmology is simplified by a symmetry reduction to homogeneous or isotropic geometries. Originally, the reduction was performed at the classical level, leaving in the isotropic case only one gravitational degree of freedom given by the scale factor a. Together with homogeneous matter fields, such as a scalar ϕ, there are then only finitely many degrees of freedom which one can quantize using quantum mechanics. The classical Friedmann equation for the evolution of the scale factor, depending on the spatial curvature $k = 0$ or ± 1, is then quantized to the Wheeler–DeWitt equation, commonly written as

$$\left(\frac{1}{9}\ell_{\mathrm{P}}^4 a^{-x}\frac{\partial}{\partial a}a^x\frac{\partial}{\partial a} - ka^2\right)\psi(a,\phi) = -\frac{8\pi G}{3}a\hat{H}_{\mathrm{matter}}(a)\psi(a,\phi) \qquad [1]$$

for the wave function $\psi(a,\phi)$. The matter Hamiltonian $\hat{H}_{\mathrm{matter}}(a)$, such as

$$\hat{H}_{\mathrm{matter}}(a) = -\frac{1}{2}\hbar^2 a^{-3}\frac{\partial^2}{\partial\phi^2} + a^3 V(\phi) \qquad [2]$$

is left unspecified here, and x parametrizes factor ordering ambiguities (but not completely). The Planck length $\ell_{\mathrm{P}} = \sqrt{8\pi G\hbar}$ is defined in terms of the gravitational constant G and the Planck constant $\hbar$.

The central conceptual issue then is the generality of effects seen in such a symmetric model and its relation to the full theory of quantum gravity. This is completely open in the Wheeler–DeWitt form since the full theory itself is not even known. On the other hand, such relations are necessary to value any potential physical statement about the origin and early history of the universe. In this context, symmetric situations thus present models, and the degree to which they approximate full quantum gravity remains mostly unknown. There are examples, for instance, of isotropic models in anisotropic but still homogeneous models, where a minisuperspace quantization does not agree at all with the information obtained from the less symmetric model. However, often those effects already have a classical analog such as instability of the more symmetric solutions. A wider investigation of the reliability of models and when correction terms from ignored degrees of freedom have to be included has not been done yet.

With candidates for quantum gravity being available, the current situation has changed to some degree. It is then not only possible to reduce classically and then simply use quantum mechanics, but also perform at least some of the reduction steps at the quantum level. The relation to models is then much clearer, and consistency conditions which arise in the full theory can be made certain to be observed. Moreover, relations between models and the full theory can be studied to elucidate the degree of approximation. Even though new techniques are now available, a detailed investigation of the degree of approximation given by a minisuperspace model has not been completed due to its complexity.

This program has mostly been developed in the context of loop quantum gravity, where the specialization to homogeneous models is known as loop quantum cosmology. More specifically, symmetries can be introduced at the level of states and basic operators, where symmetric states of a model are distributions in the full theory, and basic operators are obtained by the dual action on those distributions. In such a way, the basic representation of models is not assumed but derived from the full theory where it is subject to much stronger consistency conditions. This has implications even in homogeneous models with finitely many degrees of freedom, despite the fact that quantum mechanics is usually based on a unique representation if the Weyl operators e^{isq} and e^{itp} for the variables q and p are represented weakly continuously in the real parameters s and t.

The continuity condition, however, is not necessary in general, and so inequivalent representations are possible. In quantum cosmology this is indeed realized, where the Wheeler–DeWitt representation assumes that the conjugate to the scale factor, corresponding to extrinsic curvature of an isotropic slice, is represented through a continuous Weyl operator, while the representation derived for loop quantum cosmology shows that the resulting operator is not weakly continuous. Furthermore, the scale factor has a continuous spectrum in the Wheeler–DeWitt representation but a discrete spectrum in the loop representation. Thus, the underlying geometry

of space is very different, and also evolution takes a new form, now given by a difference equation of the type

$$\begin{aligned}&(V_{\mu+5}-V_{\mu+3})\mathrm{e}^{\mathrm{i}k}\psi_{\mu+4}(\phi)\\&\quad-(2+k^2)(V_{\mu+1}-V_{\mu-1})\psi_{\mu}(\phi)\\&\quad+(V_{\mu-3}-V_{\mu-5})\mathrm{e}^{-\mathrm{i}k}\psi_{\mu-4}(\phi)\\&\quad=-\tfrac{4}{3}\pi G\ell_{\mathrm{P}}^2\hat{H}_{\mathrm{matter}}(\mu)\psi_{\mu}(\phi)\end{aligned}\qquad[3]$$

in terms of volume eigenvalues $V_\mu=(\ell_{\mathrm{P}}^2|\mu|/6)^{3/2}$. For large μ and smooth wave functions, one can see that the difference equation reduces to the Wheeler–DeWitt equation with $|\mu|\propto a^2$ to leading order in derivatives of ψ. At small μ, close to the classical singularity, however, both equations have very different properties and lead to different conclusions. Moreover, the prominent role of difference equations leads to new mathematical problems.

This difference equation is not simply obtained through a discretization of [1], but derived from a constraint operator constructed with methods from full loop quantum gravity. It is, thus, to be regarded as more fundamental, with [1] emerging in a continuum limit. The structure of [3] depends on the properties of the full theory such that its qualitative analysis allows conclusions for full quantum gravity.

Applications

Traditionally, quantum cosmology has focused on three main conceptual issues:

- the fate of classical singularities,
- initial conditions and the "prediction" of inflation (or other early universe scenarios), and
- arrow of time and the emergence of a classical world.

The first issue consists of several subproblems since there are different aspects to a classical singularity. Often, curvature or energy densities diverge and one can expect quantum gravity to provide a natural cutoff. More importantly, however, the classical evolution breaks down at a singularity, and quantum gravity, if it is to cure the singularity problem, has to provide a well-defined evolution which does not stop. Initial conditions are often seen in relation to the singularity problem since early attempts tried to replace the singularity by choosing appropriate conditions for the wave function at $a=0$. Different proposals then lead to different solutions for the wave function, whose dependence on the scalar ϕ can be used to determine its probability distribution such as that for an inflaton. Since initial conditions often provide special properties early on, the combination of evolution and initial conditions has been used to find a possible origin of an arrow of time.

Singularities

While classical gravity is based on spacetime geometry and thus metric tensors, this structure is viewed as emergent only at large scales in canonical quantum gravity. A gravitational system, such as a whole universe, is instead described by a wave function which, at best, yields expectation values for a metric. The singularity problem thus takes a different form since it is not metrics which need to be continued as solutions to Einstein's field equations but the wave function describing the quantum system. In the strong curvature regime around a classical singularity, one does not expect classical geometry to be applicable, such that classical singularities may just be a reflection of the breakdown of this picture, rather than a breakdown of physical evolution. Nevertheless, the basic feature of a singularity as presenting a boundary to the evolution of a system equally applies to the quantum equations. One can thus analyze this issue, using new properties provided by the quantum evolution.

The singularity issue is not resolved in the Wheeler–DeWitt formulation since energy densities, with a being a multiplication operator, diverge and the evolution does not continue anywhere beyond the classical singularity at $a=0$. In some cases one can formally extend the evolution to negative a, but this possibility is not generic and leaves open what negative a means geometrically. This is different in the loop quantization: here, the theory is based on triad rather than metric variables. There is thus a new sign factor corresponding to spatial orientation, which implies the possibility of negative μ in the difference equation. The equation is then defined on the full real line with the classical singularity $\mu=0$ in the interior. Outside $\mu=0$, we have positive volume at both sides, and opposite orientations. Using the difference equation, one can then see that the evolution does not break down at $\mu=0$, showing that the quantum evolution is singularity free.

For the example [3] shown here, one can follow the evolution, for instance, backward in internal time μ, starting from initial values for ψ at large positive μ. By successively solving for $\psi_{\mu-4}$, the wave function at lower μ is determined. This goes on in this manner only until the coefficient $V_{\mu-3}-V_{\mu-5}$ of $\psi_{\mu-4}$ vanishes, which is the case if and only if $\mu=4$.

The value ψ_0 of the wave function exactly at the classical singularity is thus not determined by initial data, but one can easily see that it completely drops out of the evolution. In fact, the wave function at all negative μ is uniquely determined by initial values at positive μ. Equation [3] corresponds to one particular ordering, which in the Wheeler–DeWitt case is usually parametrized by the parameter x (although the particular ordering obtained from the continuum limit of [3] is not contained in the special family [1]). Other nonsingular orderings exist, such as that after symmetrizing the constraint operator, in which case the coefficients never become 0.

In more complicated systems, this behavior is highly nontrivial but still known to be realized in a similar manner. It is not automatic that the internal time evolution does not continue since even in isotropic models one can easily write difference equations for which the evolution breaks down. That the most natural orderings imply nonsingular evolution can be taken as a support of the general framework of loop quantum gravity. It should also be noted that the mechanism described here, providing essentially a new region beyond a classical singularity, presents one mechanism for quantum gravity to remove classical singularities, and so far the only known one. Nevertheless, there is no claim that the ingredients have to be realized in any nonsingular scenario in the same manner. Different scenarios can be imagined, depending on how quantum evolution is understood and what the interpretation of nonsingular behavior is. It is also not claimed that the new region is semiclassical in any sense when one looks at it at large volume. If the initial values for the wave function describe a semiclassical wave packet, its evolution beyond the classical singularity can be deformed and develop many peaks. What this means for the re-emergence of a semiclassical spacetime has to be investigated in particular models, and also in the context of decoherence.

Initial Conditions

Traditional initial conditions in quantum cosmology have been introduced by physical intuition. The main mathematical problem, once such a condition is specified in sufficient detail, then is to study well-posedness, for instance, for the Wheeler–DeWitt equation. Even formulating initial conditions generally, and not just for isotropic models, is complicated, and systematic investigations of the well-posedness have rarely been undertaken. An exception is the historically first such condition, due to DeWitt, that the wave function vanishes at parts of minisuperspace, such as $a=0$ in the isotropic case, corresponding to classical singularities. This condition, unfortunately, can easily be seen to be ill posed in anisotropic models where in general the only solution vanishes identically. In other models, $\lim_{a\to 0}\psi(a)$ does not even exist. Similar problems of the generality of conditions arise in other scenarios. Most well known are the no-boundary and tunneling proposal where initial conditions are still imposed at $a=0$, but with a nonvanishing wave function there.

This issue is quite different for difference equations since at first the setup is less restrictive: there are no continuity or differentiability conditions for a solution. Moreover, oscillations that become arbitrarily rapid, which can be responsible for the nonexistence of $\lim_{a\to 0}\psi(a)$, cannot be supported on a discrete lattice. It can then easily happen that a difference equation is well posed, while its continuum limit with an analogous initial condition is ill posed. One example are the dynamical initial conditions of loop quantum cosmology which arise from the dynamical law in the following way: the coefficients in [3] are not always nonzero but vanish if and only if they are multiplied with the value of the wave function at the classical singularity $\mu=0$. This value thus decouples and plays no role in the evolution. The instance of the difference equation that would determine ψ_0, for example, the equation for $\mu=4$ in the backward evolution, instead implies a condition on the previous two values, ψ_4 and ψ_8, in the example. Since they have already been determined in previous iteration steps, this translates to a linear condition on the initial values chosen. We thus have one example where indeed initial conditions and the evolution follow from only one dynamical law, which also extends to anisotropic models. Without further conditions, the initial-value problem is always well posed, but may not be complete, in the sense that it results in a unique solution up to norm. Most of the solutions, however, will be rapidly oscillating. In order to guarantee the existence of a continuum approximation, one has to add a condition that these oscillations are suppressed in large volume regimes. Such a condition can be very restrictive, such that the issue of well-posedness appears in a new guise: nonzero solutions do exist, but in some cases all of them may be too strongly oscillating.

In simple cases, one can use generating function techniques advantageously to study oscillating solutions, at least if oscillations are of alternating nature between two subsequent levels of the difference equation. The idea is that a generating function $G(x)=\sum_n \psi_n x^n$ has a stronger pole at $x=-1$ if ψ_n

is alternating compared to a solution of constant sign. Choosing initial conditions which reduce the pole order thus implies solutions with suppressed oscillations. As an example, we can look at the difference equation

$$\psi_{n+1} + \frac{2}{n}\psi_n - \psi_{n-1} = 0 \qquad [4]$$

whose generating function is

$$G(x) = \frac{\psi_1 x + \psi_0(1 + 2x(1 - \log(1-x)))}{(1+x)^2} \qquad [5]$$

The pole at $x = -1$ is removed for initial values $\psi_1 = \psi_0(2\log 2 - 1)$ which corresponds to nonoscillating solutions. In this way, analytical expressions can be used instead of numerical attempts which would be sensitive to rounding errors. Similarly, the issue of finding bounded solutions can be studied by continued fraction methods. This illustrates how an underlying discrete structure leads to new questions and the application of new techniques compared to the analysis of partial differential equations which appear more commonly.

More General Models

Most of the time, homogeneous models have been studied in quantum cosmology since even formulating the Wheeler–DeWitt equation in inhomogeneous cases, the so-called midisuperspace models, is complicated. Of particular interest among homogeneous models is the Bianchi IX model since it has a complicated classical dynamics of chaotic behavior. Moreover, through the Belinskii–Khalatnikov–Lifschitz (BKL) picture, the Bianchi IX mixmaster behavior is expected to play an important role even for general inhomogeneous singularities. The classical chaos then indicates a very complicated approach to classical singularities, with structure on arbitrarily small scales.

On the other hand, the classical chaos relies on a curvature potential with infinitely high walls, which can be mapped to a chaotic billiard motion. The walls arise from the classical divergence of curvature, and so quantum effects have been expected to change the picture, and shown to do so in several cases.

Inhomogeneous models (e.g., the polarized Gowdy models) have mostly been studied in cases where one can reformulate the problem as that of a massless free scalar on flat Minkowski space. The scalar can then be quantized with familiar techniques in a Fock space representation, and is related to metric components of the original model in rather complicated ways. Quantization can thus be performed, but transforming back to the metric at the operator level and drawing conclusions is quite involved. The main issue of interest in the recent literature has been the investigation of field theory aspects of quantum gravity in a tractable model. In particular, it turns out that self-adjoint Hamiltonians, and thus unitary evolution, do not exist in general.

Loop quantizations of inhomogeneous models are available even in cases where a reformulation such as a field theory on flat space does not exist, or is not being made use of to avoid special gauges. This is quite valuable in order to see if specific features exploited in reformulations lead to artifacts in the results. So far, the dynamics has not been investigated in detail, even though conclusions for the singularity issue can already be drawn.

From a physical perspective, it is most important to introduce inhomogeneities at a perturbative level in order to study implications for cosmological structure formation. On a homogeneous background, one can perform a mode decomposition of metric and matter fields and quantize the homogeneous modes as well as amplitudes of higher modes. Alternatively, one can first quantize the inhomogeneous system and then introduce the mode decomposition at the quantum level. This gives rise to a system of infinitely many coupled equations of infinitely many variables, which needs to be truncated, for example, for numerical investigations. At this level, one can then study the question to which degree a given minisuperspace model presents a good approximation to the full theory, and where additional correction terms should be introduced. It also allows one to develop concrete models of decoherence, which requires a "bath" of many weakly interacting degrees of freedom usually thought of as being provided by inhomogeneities in cosmology, and an understanding of the semiclassical limit.

Interpretations

Due to the complexity of full gravity, investigations without symmetry assumptions or perturbative approximations usually focus on conceptual issues. As already discussed, cosmology presents a unique situation for physics since there cannot be any outside observer. While this fact has already implications on the interpretation of observations at the classical level, its full force is noticed only in quantum cosmology. Since some traditional interpretations of quantum mechanics require the role of

observers outside the quantum system, they do not apply to quantum cosmology.

Sometimes, alternative interpretations such as Bohm theory or many-world scenarios are championed in this situation, but more conventional relational pictures are most widely adopted. In such an interpretation, the wave function yields relational probabilities between degrees of freedom rather than absolute probabilities for measurements done by an outside observer. This has been used, for instance, to determine the probability of the right initial conditions for inflation, but it is marred by unresolved interpretational issues and still disputed. These problems can be avoided by using effective equations, in analogy to an effective action, which modify classical equations on small scales. Since the new equations are still of classical type, that is, differential equations in coordinate time, no interpretational issues arise at least if one stays in semiclassical regimes. In this manner, new inflationary scenarios motivated from quantum cosmology have been developed.

In general, a relational interpretation, though preferable conceptually, leads to technical complications since the situation is much more involved and evolution is not easy to disentangle. In cosmology, one often tries to single out one degree of freedom as internal time with respect to which evolution of other degrees of freedom is measured. In homogeneous models, one can simply take the volume as internal time, such as a or μ earlier, but in full no candidate is known. Even in homogeneous models, the volume is not suitable as internal time to describe a possible recollapse. One can use extrinsic curvature around such a point, but then one has to understand what changing the internal time in quantum cosmology implies, that is, whether evolution pictures obtained in different internal time formulations are equivalent to each other.

There are thus many open issues at different levels, which, strictly speaking, do not apply only to quantum cosmology but to all of physics. After all, every physical system is part of the universe, and thus a potential ingredient of quantum cosmology. Obviously, physics works well in most situations without taking into account its being part of one universe. Similarly, much can be learned about a quantum universe if only some degrees of freedom of gravity are considered as in mini- or midisuperspace models. In addition, complicated interpretational issues, as important as they are for a deep understanding of quantum physics, do not prevent the development of physical applications in quantum cosmology, just as they did not do so in the early stages of quantum mechanics.

See also: Canonical General Relativity; Cosmology: Mathematical Aspects; Loop Quantum Gravity; Quantum Geometry and its Applications; Spacetime Topology, Causal Structure and Singularities; Wheeler–De Witt Theory.

Further Reading

Bojowald M (2001a) Absence of a singularity in loop quantum cosmology. *Physical Review Letters* 86: 5227–5230 (gr-qc/0102069).

Bojowald M (2001b) Dynamical initial conditions in quantum cosmology. *Physical Review Letters* 87: 121301 (gr-qc/0104072).

Bojowald M (2003) Initial conditions for a universe. *General Relativity and Gravitation* 35: 1877–1883 (gr-qc/0305069).

Bojowald M (2005) Loop quantum cosmology. *Living Reviews in Relativity* (to appear).

Bojowald M and Morales-Técotl HA (2004) Cosmological applications of loop quantum gravity. In: *Proceedings of the Fifth Mexican School (DGFM): The Early Universe and Observational Cosmology*, Lecture Notes in Physics, vol. 646, pp. 421–462. Berlin: Springer (gr-qc/0306008).

DeWitt BS (1967) Quantum theory of gravity. I. The canonical theory. *Physical Review* 160: 1113–1148.

Giulini D, Kiefer C, Joos E, Kupsch J, Stamatescu IO *et al.* (1996) *Decoherence and the Appearance of a Classical World in Quantum Theory*. Berlin: Springer.

Hartle JB (2003) What connects different interpretations of quantum mechanics? quant-ph/0305089.

Hartle JB and Hawking SW (1983) Wave function of the universe. *Physical Review D* 28: 2960–2975.

Kiefer C (2005) Quantum cosmology and the arrow of time. In: *Proceedings of the Conference DICE2004*, Piombino, Italy, September 2004 (gr-qc/0502016).

Kuchař KV (1992) Time and interpretations of quantum gravity. In: Kunstatter G, Vincent DE, and Williams JG (eds.) *Proceedings of the 4th Canadian Conference on General Relativity and Relativistic Astrophysics*. Singapore: World Scientific.

McCabe G (2005) The structure and interpretation of cosmology: Part II. The concept of creation in inflation and quantum cosmology. *Studies in History and Philosophy of Modern Physics* 36: 67–102 (gr-qc/0503029).

Vilenkin A (1984) Quantum creation of universes. *Physical Review D* 30: 509–511.

Wiltshire DL (1996) An introduction to quantum cosmology. In: Robson B, Visvanathan N, and Woolcock WS (eds.) *Cosmology: The Physics of the Universe*, pp. 473–531. Singapore: World Scientific (gr-qc/0101003).

Quantum Dynamical Semigroups

R Alicki, University of Gdańsk, Gdańsk, Poland

Introduction

With a given quantum system we associate a Hilbert space $\mathcal{H}$ such that pure states of the system are represented by normalized vectors ψ in $\mathcal{H}$ or equivalently by one-dimensional projections $|\psi\rangle\langle\psi|$, whereas mixed states are given by density matrices $\rho=\sum_j p_j|\psi_j\rangle\langle\psi_j|, p_j>0, \sum_j p_j=1$, that is, positive trace-1 operators and observables are identified with self-adjoint operators A acting on $\mathcal{H}$. The mean value of an observable A at a state ρ is given by the following expression:

$$<A>_\rho=\mathrm{tr}(\rho A) \qquad [1]$$

The time evolution of the isolated system is determined by the self-adjoint operator H (Hamiltonian) corresponding to the energy of the system. The infinitesimal change of state of the isolated system can be written as

$$\begin{aligned}\psi(t+dt)&=\psi(t)-\mathrm{i}H\mathrm{d}t\psi(t), \quad \text{or}\\ \rho(t+\mathrm{d}t)&=\rho(t)-\mathrm{i}\mathrm{d}t[H,\rho]\end{aligned} \qquad [2]$$

what leads to a reversible purity preserving unitary dynamics $\psi(t)=\mathrm{e}^{-\mathrm{i}tH}\psi, \rho(t)=\mathrm{e}^{-\mathrm{i}tH}\rho\mathrm{e}^{\mathrm{i}tH}$. We use the notation $[A,B]\equiv AB-BA, \{A,B\}=AB+BA$ and put $\hbar\equiv 1$. An interaction with environment leads to irreversible changes of the density matrix transforming, in general, pure states into mixed ones. Such a process can be modeled phenomenologically by a transition map $V:\mathcal{H}\mapsto\mathcal{H}$ leading to

$$\rho(t+\mathrm{d}t)=\rho(t)+\mathrm{d}tV\rho V^*-\mathrm{d}t\frac{1}{2}\{V^*V,\rho\} \qquad [3]$$

Combining Hamiltonian dynamics with several irreversible processes governed by a family of transition operators $\{V_j\}$ we obtain the following formal evolution equation in the Schrödinger picture (quantum Markovian master equation)

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\rho(t)&=\mathcal{L}\rho(t)\\ &=-\mathrm{i}[H,\rho(t)]+\frac{1}{2}\sum_{j\in I}\left([V_j,\rho(t)V_j^*]+[V_j\rho(t),V_j^*]\right)\\ &=D\rho(t)+\rho(t)D^*+\Phi\rho(t)\end{aligned} \qquad [4]$$

with the initial condition $\rho(0)=\rho$. Here $D=-\mathrm{i}H-(1/2)\sum_j V_j^*V_j, \Phi\rho=\sum_{j\in I}V_j\rho V_j^*$, and I is a certain countable set of indices. Assume for the moment that the Hilbert space $\mathcal{H}=\mathbb{C}^n$. Then the eqn [4] is always meaningful and its solution is given in terms of the exponential $\rho(t)=\Lambda(t)\rho\equiv\mathrm{e}^{t\mathcal{L}}\rho$. The linear map Φ is a general completely positive map on matrices, which preserves the positivity of ρ and $\Phi\otimes\mathbb{I}_d$ preserves positivity of $nd\times nd$ matrices for arbitrary $d=1,2,3,\ldots$ A useful Dyson-type expansion

$$\begin{aligned}\mathrm{e}^{t\mathcal{L}}\rho=\mathcal{W}(t)\rho+\sum_{k=1}^{\infty}\int_0^t dt_k\int_0^{t_k}dt_{k-1}\cdots\\ \times\int_0^{t_2}dt_1\mathcal{W}(t-t_k)\Phi\mathcal{W}(t_k-t_{k-1})\\ \times\Phi\cdots\Phi\mathcal{W}(t_1)\rho\end{aligned} \qquad [5]$$

with $\mathcal{W}(t)\rho\equiv W(t)\rho W(t)^*, W(t)=\mathrm{e}^{tD}$ shows that $\Lambda(t)$ is also completely positive. It is often convenient to describe quantum evolution in terms of observables (Heisenberg picture)

$$\begin{aligned}<A>_{\rho(t)}&=\mathrm{tr}\left(\left(\mathrm{e}^{t\mathcal{L}}\rho\right)A\right)\\ &=\mathrm{tr}\left(\rho\,\mathrm{e}^{t\mathcal{L}^*}A\right)=<A(t)>_\rho\end{aligned} \qquad [6]$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}A(t)&=\mathcal{L}^*A(t)\\ &=\mathrm{i}[H,A(t)]+\frac{1}{2}\sum_{j\in I}\left(V_j^*[A(t),V_j]+[V_j^*,A(t)]V_j\right)\\ &=D^*\rho(t)+\rho(t)D+\Phi^*\rho(t)\end{aligned} \qquad [7]$$

with the initial condition $A(0)=A$, completely positive $\Phi^*A=\sum_{j\in I}V_j^*AV_j$ and the corresponding Dyson expansion.

The solutions of eqns [4] and [7] are given in terms of dynamical semigroups. Their general mathematical properties and particular examples will be reviewed in this article. Various methods of derivation of master equations for open quantum systems from the underlying Hamiltonian dynamics of composed systems will also be presented.

Semigroups and Their Generators

For standard quantum-mechanical models it is convenient to define quantum dynamical semigroup in the Schrödinger picture as a one-parameter family $\{\Lambda(t);t\geq 0\}$ of linear and bounded maps acting on the Banach space of trace-class operators $\mathcal{T}(\mathcal{H})$ equipped with the norm $\|\sigma\|_1=\mathrm{tr}(\sigma\sigma^*)^{1/2}$ and satisfying the following conditions:

1. *Composition (semigroup) law*

$$\Lambda(t)\Lambda(s)=\Lambda(t+s), \quad \text{for all } t,s\geq 0 \qquad [8]$$

2. *Complete positivity*

$$\Lambda(t)\otimes\mathbb{I}_d \text{ is positive on } \mathcal{T}(\mathcal{H}\otimes\mathbb{C}^d) \quad \text{for all } d=1,2,3,\ldots \text{ and } t\geq 0 \qquad [9]$$

3. *Conservativity* (*trace preservation*)

$$\mathrm{tr}(\Lambda(t)\rho)=\mathrm{tr}(\rho), \quad \text{for all } \rho\in\mathcal{T}(\mathcal{H}) \qquad [10]$$

4. *Continuity* (*in a weak sense*)

$$\lim_{t\to 0}\mathrm{tr}(A\Lambda(t)\rho)=\mathrm{tr}(A\rho) \quad \text{for all } \rho\in\mathcal{T}(\mathcal{H}),\ A\in\mathcal{B}(\mathcal{H}) \qquad [11]$$

From a general theory of one-parameter semigroups on Banach spaces it follows that under the conditions (1)–(4) $\Lambda(t)$ is a one-parameter strongly continuous semigroup of contractions on $\mathcal{T}(\mathcal{H})$ uniquely characterized by a generally unbounded but densely defined semigroup generator $\mathcal{L}$ with the domain $\mathrm{dom}(\mathcal{L})\subset\mathcal{T}(\mathcal{H})$ such that for any $\rho\in\mathrm{dom}(\mathcal{L})$

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t)=\mathcal{L}\rho(t), \quad \rho(t)=\Lambda(t)\rho \qquad [12]$$

One can show that for $\lambda>0$ the resolvent $R(\lambda)=(\lambda\mathbb{I}-\mathcal{L})^{-1}$ can be extended to a bounded operator satisfying $\|R(\lambda)\|\leq\lambda^{-1}$ and, therefore, the following formula makes sense:

$$\lim_{n\to\infty}\left(\mathbb{I}-\frac{t}{n}\mathcal{L}\right)^{-n}\rho=\Lambda(t)\rho, \quad \text{for all } \rho\in\mathcal{T}(\mathcal{H}) \qquad [13]$$

Under the additional assumption that the generator $\mathcal{L}$ is bounded (and hence everywhere defined) Gorini, *et al.* (1976) and Lindblad (1976) proved that eqns [4] and [7] with bounded H, V_j and $\sum_j V_jV_j^*$ provide the most general form of $\mathcal{L}$. The choice of H and V_j is not unique and the sum over j can be replaced by an integral. In the case of n-dimensional Hilbert space we can always choose the form of eqn [4] with at most n^2-1 V_j's. Sometimes the structure [4] is hidden as for the following useful example of the relaxation process to a fixed density matrix ρ_0 with the rate $\mu>0$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t)=\mu(\rho_0-\rho(t)) \qquad [14]$$

The general structure of an unbounded $\mathcal{L}$ is not known. However, the formal expressions [4] and [7] with possibly unbounded D and V_j are meaningful under the following conditions:

- the operator D generates a strongly continuous contracting semigroup $\{\mathrm{e}^{tD};t\geq 0\}$ on $\mathcal{H}$;
- $\mathrm{dom}(V_j)\supseteq\mathrm{dom}(D)$, for all j;
- $<\phi,D\psi>+<D\phi,\psi>+\sum_j<V_j\phi,V_j\psi>=0$, for all $\phi,\psi\in\mathrm{dom}(D)$.

We can solve eqn [4] in terms of a minimal solution. Defining by Z the generator of the contracting semigroup $\rho\mapsto\mathrm{e}^{tD}\rho\mathrm{e}^{tD^*}$ and denoting by J the completely positive (unbounded) map $\rho\mapsto\sum_{j\in I}V_j\rho V_j^*$, one can show that for any $\lambda>0$, $J(\lambda\mathbb{I}-Z)^{-1}$ possesses a unique bounded completely positive extension denoted by A_λ with $\|A_\lambda\|\leq 1$. Hence, for any $0\leq r<1$ there exists a strongly continuous, completely positive and contracting semigroup $\Lambda^{(r)}(t)$ with the resolvent explicitly given by

$$R^{(r)}(\lambda)=(\lambda\mathbb{I}-Z)^{-1}\sum_{k=0}^{\infty}r^kA_\lambda^k \qquad [15]$$

As $\|R^{(r)}(\lambda)\|\leq 1$ the limit $\lim_{r\to 1}R^{(r)}(\lambda)=R(\lambda)$, where $R(\lambda)$ is the resolvent of the semigroup $\Lambda(t)$ satisfying (1), (2), and (4) and called the minimal solution of the eqn [4]. The minimal solution need not be a unique solution or conservative (generally $\mathrm{tr}\,\rho(t)\leq\mathrm{tr}\,\rho(0)$ and for any other solution $\rho'(t)\geq\rho(t)$). There exist useful sufficient conditions for conservativity, an example of a sufficient and necessary condition is the following: $A_\lambda^n\to 0$ strongly as $n\to\infty$ for all $\lambda>0$ (Chebotarev and Fagnola 1988).

Examples

Bloch equation The simplest two-level system can be described in terms of spin operators $S_k=(1/2)\sigma_k, k=1,2,3$, where σ_k are Pauli matrices. The most general master equation of the form [4] can be written as (Alicki and Fannes 2001, Ingarden *et al.* 1997)

$$\frac{\mathrm{d}\rho}{\mathrm{d}t}=-\mathrm{i}\sum_{k=1}^{3}h_k[S_k,\rho]+\frac{1}{2}\sum_{k,l=1}^{3}a_{kl}\{[S_k\rho,S_l]+[S_k,\rho S_l]\} \qquad [16]$$

where $h_k\in\mathbb{R}$ and $[a_{kl}]$ is a 3×3 complex, positively defined matrix. Introducing the magnetization vector $M_k(t)=\mathrm{tr}(\rho(t)S_k)$, we obtain the following Bloch equation used in the magnetic resonance theory:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{M}(t)=\boldsymbol{h}\times(\boldsymbol{M}(t)-\boldsymbol{M}_0)-\mathbb{F}(\boldsymbol{M}(t)-\boldsymbol{M}_0) \qquad [17]$$

where the tensor $\mathbb{F}$ (real, symmetric, and positive 3×3 matrix) and the vector $\boldsymbol{M}_0$ are functions of $[a_{kl}]$. In particular, complete positivity implies the following inequalities for the inverse relaxation times $\gamma_1,\gamma_2,\gamma_3$ (eigenvalues of $\mathbb{F}$):

$$\gamma_k\geq 0, \quad \gamma_1+\gamma_2\geq\gamma_3, \quad \gamma_3+\gamma_1\geq\gamma_2, \quad \gamma_2+\gamma_3\geq\gamma_1 \qquad [18]$$

Damped and pumped harmonic oscillator The quantum master equation for a linearly damped and pumped harmonic oscillator with frequency ω and the damping (pumping) coefficient $\gamma_\downarrow(\gamma_\uparrow)$ has form

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\mathrm{i}\omega[a^*a,\rho] + \frac{\gamma_\downarrow}{2}([a\rho,a^*]+[a,\rho a^*]) + \frac{\gamma_\uparrow}{2}([a^*\rho,a]+[a^*,\rho a]) \qquad [19]$$

where a^*, a are creation and annihilation operators satisfying $[a,a^*]=1$. Taking diagonal elements $p_n = <n,\rho n>$ in the "particle number" basis $a^*a|n>=n|n>$, $n=0,1,2,\ldots$, which evolve independently of the off-diagonal elements, one obtains the birth and death process,

$$\frac{\mathrm{d}p_n}{\mathrm{d}t} = \gamma_\downarrow(n+1)p_{n+1} + \gamma_\uparrow n p_{n-1} - \big(\gamma_\downarrow n + \gamma_\uparrow(n+1)\big)p_n \qquad [20]$$

It is convenient to use the Heisenberg picture and find an explicit solution in terms of Weyl unitary operators $W(z)=\exp[(\mathrm{i}/\sqrt{2})(za+\bar{z}a^*)]$,

$$\Lambda^*(t)W(z) = \exp\left\{-\frac{|z|^2}{4}\frac{\gamma_\downarrow}{\gamma_\downarrow-\gamma_\uparrow}\left(1-\mathrm{e}^{-(\gamma_\downarrow-\gamma_\uparrow)t}\right)\right\}W(z(t)) \qquad [21]$$

where $z(t)=\exp\{-(\mathrm{i}\omega+\frac{1}{2}(\gamma_\downarrow-\gamma_\uparrow))t\}, t\geq 0$. For $\gamma_\downarrow>\gamma_\uparrow$ the solution of eqn [19] always tends to the stationary Gibbs state

$$\rho_\beta = Z^{-1}\mathrm{e}^{-\beta\omega a^*a}, \qquad Z=\mathrm{tr}\mathrm{e}^{-\beta\omega a^*a}$$
$$\beta = \frac{1}{\omega}\ln(\gamma_\downarrow/\gamma_\uparrow) \qquad [22]$$

Quasifree semigroups The previous example is the simplest instance of the dynamical semigroups for noninteracting bosons and fermions which are completely determined on the single-particle level. Such systems are defined by a single-particle Hilbert space $\mathcal{H}_1$ and a linear map $\mathcal{H}_1 \ni \phi \mapsto a^*(\phi)$ into creation operators satisfying canonical commutation or anticommutation relations (CCRs or CARs, respectively) for bosons and fermions, respectively

$$[a(\psi),a^*(\phi)]_\pm = <\psi,\phi>$$
$$[A,B]_\pm \equiv AB-(\pm 1)BA \qquad [23]$$

In all expressions containing $(\pm)$, sign $(+)$ refers to bosons and $(-)$ to fermions.

Consider a nonhomogeneous evolution equation on the trace-class operators $\sigma\in\mathcal{T}(\mathcal{H}_1)$:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}t} = -\mathrm{i}[H_1,\sigma] - \frac{1}{2}\{(\Gamma_\downarrow-(\pm)\Gamma_\uparrow),\sigma\} + \Gamma_\uparrow \qquad [24]$$

with a single-particle Hamiltonian H_1 and a damping (pumping) positive operator $\Gamma_\downarrow(\Gamma_\uparrow)\geq 0$. The operators $H_1,\Gamma_\downarrow$, and $\Gamma_\uparrow$ need not be bounded provided $-\mathrm{i}H_1-(1/2)\{(\Gamma_\downarrow-(\pm)\Gamma_\uparrow)$ generates a (contracting in the fermionic case) semigroup $\{T(t);t\geq 0\}$ on $\mathcal{H}_1$ and the formal solution of eqn [24]

$$\sigma(t) = T(t)\sigma(0)T^*(t)+Q(t)$$
$$\text{where } Q(t)=\int_0^t T(s)\Gamma_\uparrow T^*(s)\mathrm{d}s \qquad [25]$$

is meaningful. We can now define the quasifree dynamical semigroup for the many-particle system described by the Fock space $\mathcal{F}_\pm(\mathcal{H}_1)$ (Alicki and Lendi 1987, Alicki and Fannes 2001). The simplest definition involves Heisenberg evolution of the ordered monomials in $a^*(\psi_j)$ and $a(\phi_j)$:

$$\Lambda^*(t)a^*(\psi_1)\cdots a^*(\psi_m)a(\phi_1)\cdots a(\phi_n) = \sum_P \epsilon^\pm \mathrm{Det}_\pm\big[<\psi_{j_k},Q(t)\phi_{i_l}>\big]_{k,l=1,2,\ldots,r} \times a^*(T^*(t)\psi_{\alpha_1})\cdots a^*(T^*(t)\psi_{\alpha_{m-r}}) \times a\big(T^*(t)\phi_{\beta_1}\big)\cdots a\big(T^*(t)\phi_{\beta_{n-r}}\big) \qquad [26]$$

The sum is taken over all partitions $\{(j_1,\ldots,j_r)(\alpha_1,\ldots,\alpha_{m-r})\},\{(i_1,\ldots,i_r)(\beta_1,\ldots,\beta_{n-r}\}$ such that $j_1<j_2<\cdots<j_r$, $\alpha_1<\alpha_2,\cdots<\cdots<\alpha_{m-r}$, $i_1<i_2<\cdots<r_r$, $\beta_1<\beta_2\cdots<\alpha_{n-r}$; $\epsilon^+\equiv 1$, ϵ^- is a product of signatures of the permutations $\{1,2,\ldots,m\}\mapsto\{j_1,\ldots,j_r,\alpha_1,\ldots,\alpha_{m-r}\},\{1,2,\ldots,n\}\mapsto\{i_1,\ldots,i_r,\beta_1,\ldots,\beta_{n-r}\}$; a permanent Det_+ is taken for bosons, a determinant Det_- for fermions.

Introducing an orthonormal basis $\{e_k\}$ in $\mathcal{H}_1$ and using the notation $a^*(e_k)\equiv a_k^*$, we can write a formal master equation for density matrices on the Fock space corresponding to eqn [26]:

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\mathrm{i}[H_F,\rho] + \frac{1}{2}\sum_{k,l}\Big(\Gamma_\downarrow^{kl}([a_k\rho,a_l^*] + [a_k,\rho a_l^*]) + \Gamma_\uparrow^{kl}([a_k^*\rho,a_l]+[a_k^*,\rho a_l])\Big) \qquad [27]$$

Again, formally,

$$H_F = \sum_{k,l} <e_k,H_1e_l> a_k^*a_l$$
$$\Gamma_\downarrow^{kl} = <e_k,\Gamma_\downarrow e_l>, \qquad \Gamma_\uparrow^{kl} = <e_k,\Gamma_\uparrow e_l> \qquad [28]$$

Often the formulas [27], [28] are not well-defined, but replacing the (infinite) matrices by (distribution-valued) integral kernels, sums by integrals, and a_k^*, a_l by quantum fields, we can obtain meaningful objects.

Quasifree dynamical semigroups find applications in the theory of unstable particles, quantum linear

optics, solid-state physics, quantum information theory, etc. (Alicki and Lendi 1987, Sewell 2002).

Ergodic Properties

Dynamical semigroups which possess stationary states satisfying $\mathcal{L}\rho_0 = 0$ are of particular interest, for example, in the description of relaxation processes toward equilibrium states (Frigerio 1977, Spohn 1980, Alicki and Lendi 1987). The dynamical semigroup $\{\Lambda(t)\}$ with a stationary state ρ_0 is called ergodic if

$$\lim_{t\to\infty} \Lambda(t)\rho = \rho_0, \quad \text{for any initial } \rho \qquad [29]$$

For the case of finite-dimensional $\mathcal{H}$ at least one stationary state always exists. If, moreover, it is strictly positive, $\rho_0 > 0$, then we have the following sufficient condition of ergodicity:

$$\{V_j; j \in I\}' \equiv \{A; A \in \mathcal{B}(\mathcal{H}), [A, V_j] = 0, j \in I\} = \mathbb{C}\mathbf{1} \qquad [30]$$

Open systems interacting with heat baths at the temperature T are described by the semigroups with generators [4] of the special form

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t) = -\,\mathrm{i}[H, \rho(t)] + \frac{1}{2}\sum_{\omega_j \geq 0}\Big\{\Big([V_j, \rho(t)V_j^*] + [V_j\rho(t), V_j^*]\Big) + \mathrm{e}^{-\beta\omega_j}\Big([V_j^*, \rho(t)V_j] + [V_j^*\rho(t), V_j]\Big)\Big\} \qquad [31]$$

where

$$\beta = \frac{1}{k_{\mathrm{B}}T}, \qquad [H, V_j] = -\omega_j V_j \qquad [32]$$

The Gibbs state $\rho_\beta = Z^{-1}\,\mathrm{e}^{-\beta H}$ is a stationary state for eqn [31] and the condition $\{V_j, V_j^*; j \in I\}' = \mathbb{C}\mathbf{1}$ implies ergodicity (return to equilibrium). Moreover, the matrix elements of ρ diagonal in H-eigenbasis transform independently of the off-diagonal ones and satisfy the Pauli master equation

$$\frac{\mathrm{d}p_k}{\mathrm{d}t} = \sum_l (a_{kl}p_l - a_{lk}p_k) \qquad [33]$$

with the detailed balance condition $a_{kl}\,\mathrm{e}^{-\beta E_l} = a_{lk}\,\mathrm{e}^{-\beta E_k}$, where E_k are eigenvalues of H.

Define the new Hilbert space $\mathcal{L}^2(\mathcal{H}, \rho_\beta)$ as a completion of $\mathcal{B}(\mathcal{H})$ with respect to the scalar product $(A, B)_{\rho_\beta} \equiv \mathrm{tr}(\rho_\beta A^* B)$. The semigroup's generators in the Heisenberg picture corresponding to eqn [31] are normal operators in $\mathcal{L}^2(\mathcal{H}, \rho_\beta)$ with the Hamiltonian part $\mathrm{i}[H, \cdot\,]$ being the anti-Hermitian one (automatically for bounded $\mathcal{L}^*$, and for unbounded one under technical conditions concerning domains). This allows spectral decomposition of $\mathcal{L}^*$ and a proper definition of damping rates for the obtained eigenvectors. The normality condition is one of the possible definitions of quantum detailed balance. The other, based on the time-reversal operation, often coincides with the previous one for important examples.

Interesting examples of nonergodic dynamical semigroups are given for open systems consisting of N identical particles with Hamiltonians $H^{(N)}$ and operators $V_j^{(N)}$ invariant with respect to particles permutations. Then the commutant $\{H^{(N)}, V_j^{(N)}, j \in I\}'$ contains an abelian algebra generated by projections on irreducible tensors corresponding to Young tables.

From Hamiltonian Dynamics to Semigroups

One of the main tasks in the quantum theory of open systems is to derive master equations [4] from the model of a "small" open system S interacting with a "large" reservoir R at a certain reference state ω_R (Davies 1976, Spohn 1980, Alicki and Lendi 1987, Breuer and Petruccione 2002, Garbaczewski and Olkiewicz 2002). Starting with the total Hamiltonian $H_\lambda = H_S \otimes \mathbf{1}_R + \mathbf{1}_S \otimes H_R + \lambda\sum_\alpha S_\alpha \otimes R_\alpha$, where $S_\alpha = S_\alpha^*, R_\alpha = R_\alpha^*, \mathrm{tr}(\omega_R R_\alpha) = 0$, and λ is a coupling constant, we define the reduced dynamics of S by

$$\rho(t) = \Lambda^{(\lambda)}(t)\rho = \mathrm{tr}_R\big(U_\lambda(t)\rho \otimes \omega_R U_\lambda^*(t)\big) \qquad [34]$$

with $U_\lambda(t) = \exp(-\mathrm{i}tH_\lambda)$. Here tr_R denotes a partial trace over R defined in terms of an arbitrary basis $\{e_k\}$ of R by the formula $<\phi, (\mathrm{tr}_R A)\phi> = \sum_k <\phi\otimes e_k, A\phi\otimes e_k>$. Generally, $\Lambda^{(\lambda)}(t+s) \neq \Lambda^{(\lambda)}(t)\Lambda^{(\lambda)}(s)$, but dynamical semigroups can provide good approximations in important cases.

Weak-Coupling Limit

Under the conditions of sufficiently fast decay of multitime correlation functions constructed from the observables R_α at the state ω_R, one can prove that for small coupling constant λ the exact dynamical map $\Lambda^{(\lambda)}(t)$ can be approximated by the dynamical semigroup corresponding to the following master equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t) = -\mathrm{i}[H, \rho(t)] + \frac{\lambda^2}{2}\sum_{\alpha\beta}\sum_{\omega\in Sp} C_{\alpha\beta}(\omega) \times \Big([V_\omega^\alpha, \rho(t)V_\omega^{\beta *}] + [V_\omega^\alpha\rho(t), V_\omega^{\beta *}]\Big) \qquad [35]$$

where $H = H_S + \lambda^2 \sum_{\alpha\beta} \sum_{\omega \in Sp} K_{\alpha\beta}(\omega) V_\omega^{\alpha *} V_\omega^\beta$ is a renormalized Hamiltonian, $\sum_{\omega \in Sp}$ denotes the sum over eigenfrequencies of $-[H, \cdot\,]$, $e^{itH} S_\alpha e^{-itH} = \sum_{\omega \in Sp} V_\omega^\alpha e^{-i\omega t}$ and

$$\int_0^\infty e^{i\omega t} \mathrm{tr}\left(\omega_R e^{itH_R} R_\alpha e^{-itH_R} R_\beta\right) dt = \tfrac{1}{2} C_{\alpha\beta}(\omega) + iK_{\alpha\beta}(\omega) \quad [36]$$

The rigorous derivation involves van Hove or weak coupling limit, $\lambda \to 0$, with $\tau = \lambda^2 t$ kept fixed.

It follows from the Bochner theorem that the matrix $[C_{\alpha\beta}(\omega)]$ is positively defined and therefore by its diagonalization we can convert eqn [35] into the standard form [4]. If the reservoir's state ω_R is an equilibrium state (Kubo–Martin–Schwinger state) then $C_{\alpha\beta}(-\omega) = e^{-\omega/k_B T}\, C_{\beta\alpha}(\omega)$ and therefore eqn [35] can be written in a form [31]. Moreover, transition probabilities a_{kl} from eqn [33] coincide with those obtained using the "Fermi golden rule."

Low-Density Limit

If the reservoir can be modeled by a gas of noninteracting particles (bosons or fermions) at low density ν, we can derive the following master equation which approximates an exact dynamics [34] in the low-density limit ($\nu \to 0$, with $\tau = \nu t$ kept fixed)

$$\frac{d}{dt}\rho(t) = -i[H, \rho(t)] + \nu\pi \sum_{\omega \in S} \int_{\mathbb{R}^6} d^3p\, d^3p'\, G(p) \times \delta(E_{p'} - E_p + \omega)\left([T_\omega(p,p')\rho(t), T_\omega(p,p')^*] + [T_\omega(p,p')\rho(t), T_\omega(p,p')^*]\right) \quad [37]$$

Here H is a renormalized Hamiltonian of the system S, $e^{itH} T e^{-itH} = \sum_{\omega \in S} T_\omega e^{-i\omega t}$, T is a T-matrix describing the scattering process involving S and a single particle, $T = V\Omega_+$, where V is a particle-system potential and Ω_+ is a Møller operator. $T_\omega(p,p')$ denotes the integral kernel corresponding to T_ω expressed in terms of momenta of the bath particle, E_p the kinetic energy of a particle, and $G(p)$ its probability distribution in the momentum space. If $G(p) \sim \exp(-E_p/k_B T)$ and microreversibility conditions, $E_p = E_{-p}$ and $T_\omega(-p, -p') = T_\omega(p', p)$, hold, then eqn [37] satisfies the quantum detailed-balance condition with the stationary Gibbs state ρ_β, $\beta = 1/k_B T$.

Entropy and Purity

The relative entropy $S(\rho\,|\,\sigma) = \mathrm{tr}(\rho \ln \rho - \rho \ln \sigma)$ is monotone with respect to any trace-preserving completely positive map Λ, that is, $S(\Lambda\rho|\Lambda\sigma) \le S(\rho|\sigma)$. Hence, for the quantum dynamical semigroup $\Lambda(t)$ with the stationary state ρ_0 we obtain the following relation for the von Neumann entropy $S(\rho) = -\mathrm{tr}(\rho \ln \rho)$:

$$\frac{d}{dt} S(\rho(t)) = -\frac{d}{dt} S(\rho(t)|\rho_0) - \frac{d}{dt} \mathrm{tr}(\rho(t) \ln \rho_0) \quad [38]$$

where $-(d/dt)S(\rho(t)\,|\,\rho_0) \ge 0$ is an entropy production and the second term describes entropy exchange with environment (Spohn 1980, Alicki and Lendi 1987).

Bistochastic dynamical semigroups preserve the maximally mixed state, that is, $\mathcal{L}(\mathbf{1}) = 0$. For them, the von Neumann entropy does not decrease and the purity $\mathrm{tr}\,\rho^2$ never increases (Streater 1995). Two important classes of master equations, used to describe decoherence, yield bistochastic dynamical semigroups:

$$\frac{d}{dt}\rho(t) = -i[H, \rho(t)] - \sum_j [A_j, [A_j, \rho(t)]], \quad A_j = A_j^* \quad [39]$$

$$\frac{d}{dt}\rho(t) = -i[H, \rho(t)] + \int_M \mu(d\alpha)(U(\alpha)\rho(t)U^*(\alpha) - \rho(t)) \quad [40]$$

where $U(\alpha)$ are unitary and $\mu(\cdot)$ is a (positive) measure on M.

Itô–Schrödinger Equations

Up to technical problems in the case of unbounded operators, the master equation [4] is completely equivalent to the following stochastic differential equation (in Itô form):

$$d\psi(t) = -iH\psi(t)\,dt - \frac{1}{2}\sum_{j \in I} V_j^* V_j \psi(t)\,dt - i\sum_{j \in I} V_j \psi(t) dX_j(t) \quad [41]$$

where $X_j(t)$ are arbitrary statistically independent stochastic processes with independent increments (continuous or jump processes) such that the expectation $E(dX_j(t)\,dX_k(t)) = \delta_{jk}\,dt$. Equation [41] should be understood as an integral equation involving stochastic Itô integrals with respect to $\{X_j(t)\}$ computed according to the Itô rule: $dX_j(t)\,dX_k(t) = \delta_{jk}\,dt$. Taking the average $\rho(t) = E(|\psi(t)><\psi(t)|)$ one can show, using the Itô rule, that $\rho(t)$ satisfies eqn [4]. For numerical

applications, it is convenient to use the nonlinear version of eqn [41] for the normalized stochastic vector $\phi(t) = \psi(t)/\|\psi(t)\|$, which can be easily derived from eqn [41] (Breurer and Petruccione 2002).

Introducing quantum noises, for example, quantum Brownian motions defined in terms of bosonic or fermionic fields and satisfying suitable quantum Itô rules one can develop the theory of noncommutative stochastic differential equations (NSDE) (Hudson and Parthasarathy 1984). Both, eqn [41] and NSDE, provide examples of unitary dilations – (physically singular) mathematical constructions of the environment R and the R–S coupling which exactly reproduce dynamical semigroups as reduced dynamics [34].

Algebraic Formalism

In order to describe open systems in thermodynamical limit (e.g., infinite spin systems) or systems in the quantum field theory one needs the formalism based on C^* or von Neumann algebras. In the C^*-algebraic language, by dynamical semigroup (in the Heisenberg picture) we mean a family $\{T(t); t \geq 0\}$ of linear maps on the unital C^*-algebra $\mathcal{A}$ satisfying the following conditions: (1) complete positivity, (2) $T(t)T(s) = T(t+s)$, (3) weak (or strong) continuity, and (4) $T(t)\mathbf{1} = \mathbf{1}$. Assuming the existence of a faithful stationary state $\omega = \omega \circ T(t)$ on $\mathcal{A}$, one can use a Gelfand–Naimark–Segal (GNS) representation $\pi_\omega(\mathcal{A})$ of $\mathcal{A}$ in terms of bounded operators on the suitable Hilbert space $\mathcal{H}_\omega$ with the cyclic and separating vector Ω satisfying $\omega(A) = <\Omega, \pi_\omega(A)\Omega>$ for all $A \in \mathcal{A}$. Then the dynamical semigroup can be defined on the von Neumann algebra $\mathcal{M}$ (obtained by a weak closure of $\pi_\omega(\mathcal{A})$) as $\hat{T}(t)\pi_\omega(A) \equiv \pi_\omega(T(t)A)$. The Kadison inequality valid even for 2-positive bounded maps Λ on $\mathcal{A}$

$$\Lambda(AA^*) \geq \Lambda(A)\Lambda(\mathbf{1})\Lambda(A^*) \qquad [42]$$

implies that $\omega([T(t)A]^*T(t)A) \leq \omega(A^*A)$, which allows one to extend the dynamical semigroup to the contracting semigroup $\tilde{T}(t)[\pi_\omega(A)\Omega] \equiv [\pi_\omega(T(t)A)]\Omega$ on the GNS Hilbert space $\mathcal{H}_\omega$. Typically, one tries to define the semigroup in terms of the proper limiting procedures $T(t) = \lim_{n\to\infty} T_n(t)$, where $T_n(t)$ is well defined on $\mathcal{A}$. However, the limit may not exist as an operator on $\mathcal{A}$ but can be well defined on the von Neumann algebra $\mathcal{M}$. If not, the contracting semigroup on $\mathcal{H}_\omega$ may still be a useful object.

Although there exists a rich ergodic theory of dynamical semigroups for the special types of von Neumann algebras, the most difficult problem of constructing physically relevant semigroups for generic infinite systems remains unsolved (Majewski and Zegarliński 1996, Garbaczewski and Olkiewicz 2002).

Nonlinear Dynamical Semigroups

The reduced description of many-body classical or quantum systems in terms of single-particle states (probability distributions, wave functions, or density matrices) leads to nonlinear dynamics (e.g., Boltzmann, Vlasov, Hartree, or Hartree–Fock equations) (Spohn 1980, Garbaczewski and Olkiewicz 2002). A large class of nonlinear evolution equations for single-particle density matrices ρ can be written as Alicki and Lendi (1987)

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \mathcal{L}[\rho]\rho \qquad [43]$$

where $\sigma \mapsto \mathcal{L}[\sigma]$ is a map from density matrices to semigroup generators of the type [4]. Under certain technical conditions the solution of eqn [43] exists and defines a nonlinear dynamical semigroup – a family $\{\Gamma(t); t \geq 0\}$ of maps on the set of density matrices satisfying the composition law $\Gamma(t+s) = \Gamma(t)\Gamma(s)$.

A simple example is provided by an open N-particle system with the total Hamiltonian invariant with respect to particle permutations. The Markovian approximation combined with the mean-field method leads to a nonlinear dynamical semigroup which preserves purity and for initial pure states is governed by the nonlinear Schrödinger equation with the following structure:

$$\begin{aligned}\frac{\mathrm{d}\psi}{\mathrm{d}t} &= -\mathrm{i}(h + NU(\psi))\psi \\ &\quad + \frac{N}{2}\sum_j \Big(<\psi, V_j^*\psi> V_j\psi \\ &\quad - <\psi, V_j\psi> V_j^*\psi \Big)\end{aligned} \qquad [44]$$

Here h is a single-particle Hamiltonian, $U(\psi)$ a Hartree potential, and V_j are single-particle operators describing collective dissipation.

See also: Boltzmann Equation (Classical and Quantum); Channels in Quantum Information Theory; Evolution Equations: Linear and Nonlinear; Kinetic Equations; Nonequilibrium Statistical Mechanics (Stationary): Overview; Positive Maps on C^*-Algebras; Quantum Error Correction and Fault Tolerance; Quantum

Mechanical Scattering Theory; Stochastic Differential Equations.

Further Reading

Alicki R and Fannes M (2001) *Quantum Dynamical Systems*. Oxford: Oxford University Press.

Alicki R and Lendi K (1987) *Quantum Dynamical Semigroups and Applications*. LNP, vol. 286. Berlin: Springer.

Breurer H-P and Petruccione F (2002) *Theory of Open Quantum Systems*. Oxford: Oxford University Press.

Chebotarev AM and Fagnola F (1998) Sufficient conditions for conservativity of quantum dynamical semigroups. *Journal of Functional Analysis* 153: 382–404.

Davies EB (1976) *Quantum Theory of Open Systems*. London: Academic Press.

Frigerio A (1977) Quantum dynamical semigroups and approach to equilibrium. *Letters in Mathematical Physics* 2: 79–87.

Garbaczewski P and Olkiewicz R (eds.) (2002) *Dynamics of Dissipation*. LNP, vol. 597. Berlin: Springer.

Gorini V, Kossakowski A, and Sudarshan ECG (1976) Completely positive dynamical semigroups of *n*-level systems. *Journal of Mathematical Physics* 17: 821–825.

Hudson R and Parthasarathy KR (1984) Quantum Itô's formula and stochastic evolutions. *Communications in Mathematical Physics* 93: 301–323.

Ingarden RS, Kossakowski A, and Ohya M (1997) *Information Dynamics and Open Systems*. Dordrecht: Kluwer.

Lindblad G (1976) On the generators of quantum dynamical semigroups. *Communications in Mathematical Physics* 48: 119–130.

Majewski WA and Zegarliński B (1996) Quantum stochastic dynamics II. *Reviews in Mathematical Physics* 8: 689–713.

Sewell G (2002) *Quantum Mechanics and Its Emergent Macrophysics*. Princeton: Princeton University Press.

Spohn H (1980) Kinetic equations from Hamiltonian dynamics. *Review of Modern Physics* 52: 569–616.

Streater RF (1995) *Statistical Dynamics*. London: Imperial College Press.

Quantum Dynamics in Loop Quantum Gravity

H Sahlmann, Universiteit Utrecht, Utrecht, The Netherlands

Introduction

In general relativity, the metric is a dynamic entity, there is no preferred notion of time, and the theory is invariant under diffeomorphisms. Therefore, one expects the concept of dynamics to be very different from that in mechanical or special relativistic systems. Indeed, in a canonical formulation, the diffeomorphism symmetry manifests itself through the appearance of constraints (*see* Constrained Systems). In particular, in the absence of boundaries, the Hamiltonian turns out to be a linear combination of them. Thus, the dynamics is completely encoded in the constraints.

To quantize such a system following Dirac, one has to define operators corresponding to the constraints on an auxiliary Hilbert space. Solutions to the quantum dynamics are then vectors that are annihilated by all the constraint operators. Technical complications can arise, and the solutions might not lie in the auxiliary Hilbert space but in an appropriately chosen dual.

Physical observables on the other hand are associated with operators on the auxiliary space that commute with the constraints or, equivalently, operators that act within the space of solutions.

Since the solutions of the quantum dynamics will not depend on any sort of time parameter in an explicit way, they cannot be readily interpreted as a (quantum) spacetime history. The conceptual questions related to this are known as the "problem of time" in quantum gravity.

We should mention that there is a proposal – consistent discretizations – that allows us to eliminate constraints, at the expense of a discretization of the classical theory and dynamical specification of Lagrange multipliers. Application of this technique to gravity is currently under study.

Loop quantum gravity (LQG) (*see* Loop Quantum Gravity) is based on the choice of a canonical pair (A_a, E^b) of an SU(2) connection and an su(2)-valued vector density. The constraints come in three classes:

$$G_i[A,E](x) = 0, \qquad V_a[A,E](x) = 0, \qquad C[A,E](x) = 0$$

the Gauss, vector, and scalar constraints, respectively.

Before giving some detail about the quantization of the constraints and their solutions, we should mention that there exists an analogous classical formulation in terms of complex (self-dual) variables. The quantization in that formulation faces serious technical obstacles, but in the case of positive cosmological constant an elegant formal solution to all the constraints – the Kodama state – is known. It is related to the Chern–Simons action on the spatial slice.

As said before, strictly speaking, implementing the dynamics comprises quantizing and satisfying all the constraints. Here we will however focus on C since it is the most challenging, and most closely related to standard dynamics in that it generates changes under timelike deformations of the Cauchy surface Σ on which the canonical formulation is based.

The quantum solutions of the other constraints, linear combinations of s-knots, lie in a Hilbert space $\mathcal{K}_{\text{diff}}$ which is part of the dual of the kinematical Hilbert space $\mathcal{K}$ of the theory. For details on these solutions as well as some basic definitions that will be used without comment below (*see* Loop Quantum Gravity). Since s-knots are labeled, among other things, by a diffeomorphism equivalence class of a graph, relations to knot theory are emerging at this level (*see* Knot Invariants and Quantum Gravity).

It is important to note that C does not Poisson-commute with the diffeomorphism constraints. Therefore, in the quantum theory it does matter in which order the constraints are solved. It turns out that on the quantum solutions to the other constraints, the scalar constraint can be defined by introducing a regulator, and stays well defined even when the regulator is removed. This ultraviolet finiteness on $\mathcal{K}_{\text{diff}}$ can be intuitively understood from the diffeomorphism invariance of its elements: There is no problematic short-distance regime since the states do not contain any scale at all.

In the following we will briefly review the implementation of the scalar constraint in LQG and comment on some ramifications and open questions.

The Scalar Constraint Operator

In the Lorentzian theory the scalar constraint C is the sum of the scalar constraint C^{E} of the Euclidean theory:

$$C^{\text{E}} = (\det q)^{-1/2} \operatorname{tr}(F_{ab}[E^a, E^b])$$

a second term of a similar form, but with the curvature F of the connection A replaced by the curvature associated to a certain triad e, and possibly matter terms. In the following we will just discuss C^{E}, the other terms can be handled in a similar fashion.

There appear to be a number of obstacles to the quantization of C^{E}: for one, the inverse of the determinant would likely be ill defined, as the volume operator – essentially a quantization of $\int(\det q)^{1/2}$ – has a large kernel. In addition, there are no well-defined operators corresponding to F and E evaluated at points. Rather, only holonomies $h_e[A]$ of A along curves e and certain functionals of E are well defined as operators. These issues can however be dealt with in an elegant way as follows.

The first step is to absorb the determinant factor into a Poisson bracket,

$$C^{\text{E}} = \frac{2}{\kappa}\epsilon^{abc} \operatorname{tr}(F_{ab}\{A_c, V\})$$

where V is the volume of the spatial slice Σ. Then one approximates the curvature by (identity minus) the holonomy around a small loop. In the present case one finds that for a small tetrahedron Δ with base point v, one can approximate

$$\begin{aligned} C^{\text{E}}_{\Delta}(N) &:= 2\kappa^{-1}\int_{\Delta} N \operatorname{tr}(F \wedge \{A, V\}) \\ &\approx -\frac{2}{3\kappa} N(v)\epsilon^{ijk} \operatorname{tr}(h_{\alpha_{ij}} h_{s_k}\{h_{s_k}^{-1}, V\}) \quad [1] \end{aligned}$$

where (see **Figure 1a**)) the s_i are edges of Δ incident at v and the α_{ij} loops around the faces of Δ incident at v.

This suggests how to define an operator $\widehat{C}^{\text{E}}_{\Gamma}$ that acts on cylindrical functions on a given graph Γ: one chooses a triangulation adapted to the graph and quantizes the $C^{\text{E}}_{\Delta}(N)$ (where Δ is a tetrahedron of this triangulation) using the right-hand side of [1] – holonomies are quantized by the holonomy operators of the quantum theory, V by the volume operator $\widehat{V}$, and the Poisson bracket by the corresponding commutator divided by $i\hbar$. To be more precise, the triangulation is chosen such that the s_k in [1] are part of Γ, and the operators corresponding to the h_α are creating new edges that connect the endpoints of the s_k (see **Figure 1b**).

Still this is not sufficient, since the definition of $\widehat{C}^{\text{E}}_{\Gamma}$ depends quite heavily on the choice of the triangulation, and there is no natural way to choose one. Furthermore, there is no choice that would guarantee

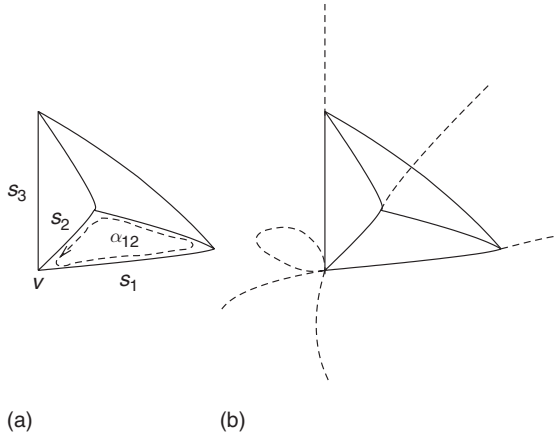


Figure 1 (a) A tetrahedron Δ and its labeling of edges and loops. (b) A tetrahedron Δ adapted to the edges (dashed lines) of a graph Γ.

that the $\widehat{C}^{\mathrm{E}}_{\Gamma}$ for different Γ are consistent in the sense that they correspond to the action of the same operator $\widehat{C}^{\mathrm{E}}$ on two different cylindrical subspaces. Here, the diffeomorphism invariance of the theory comes to the rescue: a well-defined operator largely free of ambiguities can be obtained by letting the operators above act (by duality) on $\mathcal{K}_{\mathrm{diff}}$ to give elements in $\mathcal{K}^*$. When acting on diffeomorphism-invariant states, the ambiguities in the definition of the triangulations can be eliminated, and the operators $\widehat{C}^{\mathrm{E}}_{\Gamma}$ for different Γ are consistent and together define an operator $\widehat{C}^{\mathrm{E}}(N)$. Roughly speaking, for a diffeomorphism-invariant state, it does not matter anymore where on the graph the endpoints of the s_k lie and how they are connected to form the loops α. The final picture looks as follows: for each s-knot s, the operator gives a sum of contributions, one for each vertex of s, that is, $\widehat{C}^{\mathrm{E}}(N)s = \sum_v \widehat{C_v}(N)s$. The terms in this sum are not diffeomorphism invariant. Their evaluation on a spin network S is of the form

$$(\widehat{C_v}s)[S] = \sum_{s'} c(s')N(x(v))s'[S] \qquad [2]$$

where the s' are s-knots that differ from s by the addition or deletion of certain edges, and corresponding changes in coloring (by $\pm 1/2$) and intertwiners. As an example, Figure 2 schematically depicts the action on a trivalent vertex. The point $x(v)$ on which N is evaluated in the above formula gets determined as follows: the evaluation $s'[S]$ is zero unless the graph Γ on which S is based is an element in the diffeomorphism equivalence class on which s' is based. $x(v)$ is the position of the vertex v in this element of the equivalence class. Because of this $x(v)$, the action of $\widehat{C}^{\mathrm{E}}(N)$ is not diffeomorphism invariant.

Similar techniques give a quantization $\widehat{C}$ of the full constraint. The solutions to the constraint can be determined as the vectors $\psi \in \mathcal{K}_{\mathrm{diff}}$ that are annihilated by $\widehat{C}$ in the sense that $(\widehat{C}(N)\psi)[f] = 0$ for all functions N and elements f of $\mathcal{K}$. The solutions are more or less explicitly known; however, the task of interpreting them is a hard one and remains an object of current research.

It should be mentioned that, strictly speaking, one can arrive at several slightly different versions of the

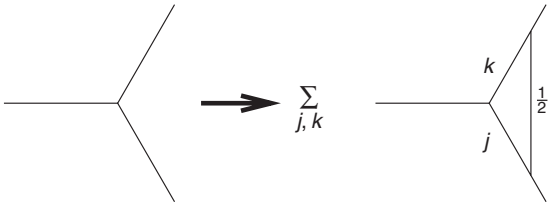


Figure 2 A schematic rendering of the action of the operator $\widehat{C}_v$ for a trivalent vertex.

constraint operator along the lines sketched above. The quantization ambiguities include changes in the power of the volume operator and the spin quantum number that the constraint creates or annihilates. An interesting check on these quantizations would be to inspect the algebra of constraint operators for anomalies. In the present situation, this can only be carried out to a certain extent, because $\widehat{C}$ is defined on diffeomorphism-invariant states. The Poisson bracket between two scalar constraints is proportional to a diffeomorphism constraint, and indeed it turns out that in the quantum theory the commutator of two scalar constraint operators vanishes for quantizations as described above. In that sense they are ambiguity free; however, this criterion is not strong enough to distinguish between the candidates.

Recently, a slightly different strategy has been proposed, which, if successfully implemented, would eliminate some of the questions regarding the constraint algebra. The idea is to combine the constraints $C(N)$ for different lapse functions N into one master constraint

$$M = \int_{\Sigma} (\det q)^{-1/2} C^2 \, \mathrm{d}^3x$$

M is manifestly diffeomorphism invariant and could replace all the noncommuting constraints $C(N)$, hence simplifying the constraint algebra considerably.

The interpretation of the solutions of all the constraints hinges on the construction of observables for the theory. This is already a difficult task in the classical theory, and thus even more so after quantization. Though there is no general solution to this problem available, interesting proposals are being studied.

Finally, it should be said that the quantization of the scalar constraint can be used to obtain a picture that resembles more the standard time evolution in quantum field theory. The (formal) power series expansion of the projector

$$P = \prod_{x\in\Sigma} \delta(\widehat{C}(x)) = \int D[N] \, \exp\left[\mathrm{i} \int_{\Sigma} N(x)\widehat{C}(x)\right]$$

onto the kernel of $\widehat{C}$ can be described by a spin foam model (*see* Spin Foams).

For further information on the subject of this article see the references: Thiemann (to appear), Rovelli (2004), and Ashtekar and Lewandowski (2004) for general reviews on LQG (with a systematic exposition of a large class of quantizations of the scalar constraint and their solutions in Ashtekar and Lewandowski (2004)); Thiemann (1998) for a seminal work on the quantization of the scalar constraint; Rovelli (1999) and Reisenberger and Rovelli (1997) on the connection to spin foam models; Di Bartolo *et al.* (2002) on

consistent discretizations; Kodama (1990) and Freidel and Smolin (2004) on the Kodama state; and Thiemann (2003) on the master constraint program.

See also: Constrained Systems; Knot Invariants and Quantum Gravity; Loop Quantum Gravity; Quantum Geometry and its Applications; Spin Foams; Wheeler–De Witt Theory.

Further Reading

Ashtekar A and Lewandowski J (2004) Background independent quantum gravity: a status report. *Classical and Quantum Gravity* 21: R53.

Di Bartolo C, Gambini R, and Pullin J (2002) Canonical quantization of constrained theories on discrete space-time lattices. *Classical Quantum Gravity* 19: 5275.

Freidel L and Smolin L (2004) The linearization of the Kodama state. *Classical and Quantum Gravity* 21: 3831.

Kodama H (1990) Holomorphic wave function of the universe. *Physical Review D* 42: 2548.

Reisenberger MP and Rovelli C (1997) "Sum over surfaces" form of loop quantum gravity. *Physical Review D* 56: 3490.

Rovelli C (1999) The projector on physical states in loop quantum gravity. *Physical Review D* 59: 104015.

Rovelli C (2004) *Quantum Gravity*, Cambridge Monographs in Mathematical Physics. Cambridge: Cambridge University Press.

Thiemann T (1998) Quantum spin dynamics (QSD). *Classical and Quantum Gravity* 15: 839.

Thiemann T (2003) The Phoenix project: master constraint programme for loop quantum gravity, arXiv:gr-qc/0305080.

Thiemann T (2006) *Modern Canonical Quantum General Relativity*, Cambridge University Press (to appear).

Quantum Electrodynamics and Its Precision Tests

S Laporta, Università di Parma, Parma, Italy
E Remiddi, Università di Bologna, Bologna, Italy

Introduction

Quantum electrodynamics (QED) describes the interaction of the electromagnetic field (EMF) with charged particles. Any physical particle interacts, directly or indirectly, with any other particle (including itself); in the case of the electron, however, at low and medium energy (say, up to a few GeV) the interaction with the EMF is by and far the most important, so that QED describes with great precision the dynamics of the electron, and at the same time the electron provides with the most stringent tests of QED currently available.

In the various sections of this article we will discuss, in the following order, the origin of QED, the structure of the radiative corrections, the application of QED to various bound states problems (the hydrogen-like atoms, the muonium, and positronium) and the anomalous magnetic moments of the leptons (the muon and the electron).

Origin of QED

The origin of QED can ideally be traced back to the very beginning of quantum mechanics, the black-body formula by M Planck (1900), which was soon understood as pointing to a discretization of the energy and momentum associated to the EMF into quanta of light or photons (Einstein 1905).

The quantization of the EMF was first worked out by P Jordan, within the article (1926) by M Born, W Heisenberg, and P Jordan (usually referred to as the *Dreimännerarbeit*) and then in the paper "The quantum theory of emission and absorption of radiation" by PAM Dirac, commonly considered the beginning of the so-called second quantization formalism.

In the subsequent year (1928) Dirac published the famous equation for the relativistic electron, from which it was immediately deduced, on a firmer basis, that the electron has spin 1/2, that its spin gyromagnetic ratio (the ratio between spin and associated magnetic moment in suitable dimensionless units; see below for more details) is twice the value predicted by classical physics (a result expressed as $g_e = 2$) and that the levels of atomic hydrogen with the same principal quantum number n are not fully degenerate, as in the nonrelativistic limit, but do possess the so-called fine structure splitting. In particular, the energy of the $n = 2$ levels splits into two values, one value for $2P_{3/2}$ states with total angular momentum $J = 3/2$ and another value for the states $2S_{1/2}$ and $2P_{1/2}$, which have $J = 1/2$; note that the $2S_{1/2}$ and $2P_{1/2}$ states are still degenerate.

Very soon it was realized that Dirac's equation also requires that each particle must be accompanied by its antiparticle, with exactly the same mass and opposite charge. The antiparticle of the electron, the positron, was indeed discovered by C Anderson

(1932), establishing Dirac's equation as one of the cornerstones of theoretical physics.

All the ingredients needed for the evaluation of the perturbative corrections to the QED theory (usually called radiative corrections) were already present at that moment, but radiative corrections were not systematically investigated for several years, due perhaps to the length and difficulty of the calculations and the absence of important disagreements between theoretical predictions and experimental results.

The situation changed in 1947, when two experiments were carried out, measuring the energy difference between the $2^2S_{1/2}$ and $2^2P_{1/2}$ levels of the hydrogen atom and the gyromagnetic ratios of the electron.

Lamb and Retherford (1947), by using the "great wartime advances in microwaves techniques," succeeded in establishing that in the hydrogen atom "the $2^2S_{1/2}$ state is higher than the $2^2P_{1/2}$ by about 1000 Mc/sec.," while (as observed above) according to the Dirac theory the two states are expected to have exactly the same energy. Subsequent refinements of the experiment (Triebwasser *et al.* 1953) gave for the difference (now referred to as Lamb shift) the value 1057.77 ± 0.10 MHz, with a relative error 1×10^{-4}.

The authors of the second 1947 experiment (Kusch and Foley 1947) measured the frequencies associated with the Zeeman splitting of two different states of gallium, finding an inconsistency with the theoretical values of the gyromagnetic ratios of the electron. More exactly, write the magnetic moments $\boldsymbol{\mu}_L, \boldsymbol{\mu}_S$ associated to the (dimensionless) orbital and spin angular momenta $\boldsymbol{L}, \boldsymbol{S}$ of the electron as

$$\boldsymbol{\mu}_L = -g_L \frac{e\hbar}{2m_e c}\boldsymbol{L}, \quad \boldsymbol{\mu}_S = -g_S \frac{e\hbar}{2m_e c}\boldsymbol{S} \qquad [1]$$

where $(-e)$ is the charge of the electron $(e > 0)$, m_e its mass, c the speed of light and g_L, g_S, respectively, the orbital and spin gyromagnetic ratios; the Dirac theory then predicts $g_L = 1$ and $g_S = 2$, while the results of Kusch and Foley (1947) gave a discrepancy which could be accounted for by taking $g_S = 2.00229 \pm 0.00008$ and $g_L = 1$, or alternatively $g_S = 2$ and $g_L = 0.99886 \pm 0.00004$. In modern notation the first conjecture can be rewritten as

$$g_S = g_e = 2(1 + a_e), \quad a_e = 0.001145 \pm 0.00004 \qquad [2]$$

where a_e is the anomalous magnetic moment (or magnetic anomaly) of the electron.

The need of explaining the two experimental results gave rise to a rapid development of covariant perturbation theory (which replaced the previous noncovariant "old fashioned" perturbation theory) and of the renormalization theory, which liberated the perturbative expansion from the divergences plaguing the older approach, opening the path to the evaluation of radiative corrections and to the great success of precision predictions of QED.

The formalism improved quickly, evolving in the more general quantum field theory (QFT) approach; three of the main contributors were Sin-Itiro Tomonaga, Julian Schwinger, and Richard P Feynman, awarded a few years later (1965) the Nobel price "for their fundamental work in quantum electrodynamics, with deep-ploughing consequences for the physics of elementary particles." QFT was then successfully used for describing the weak interactions in the electroweak model and later on also for the strong interactions theory, dubbed quantum chromodynamics (or QCD, in analogy with the popular QED acronym). For more details and references to original works, the reader is invited to look at any treatise on QED or QFT, such as, for instance, Weinberg (1995).

Initially, the Lamb shift was perhaps more important than the electron magnetic anomaly both for the establishment of renormalization theory and as a test of QED, but in the following years it was supplanted by the latter as a precision test of QED.

In 1947 the "best values" for some fundamental constants were indeed

$$\begin{aligned} c &= (2.99776 \pm 0.00004) \times 10^{10}\ \text{cm s}^{-1} \\ R_\infty &= \frac{m_e c^2 \alpha^2}{2hc} = 109737.303 \pm 0.017\ \text{cm}^{-1} \\ 1/\alpha &= 137.030 \pm 0.016 \end{aligned} \qquad [3]$$

where R_∞ is the Rydberg constant for infinite mass, h the Planck constant, and α the fine structure constant (let us observe here in passing that R_∞ was and is still known much better than the separate values of m_e, α, and h entering in its definition); for comparison, the current (2005) values for c and R_∞ are

$$\begin{aligned} c &= 299792458\ \text{m s}^{-1} \\ R_\infty &= 109737.31568525(73)\ \text{cm}^{-1} \end{aligned} \qquad [4]$$

where the value of c is exact (it is in fact the definition of the meter), and the relative error in R_∞ is 6.6×10^{-12} (the value of α will be discussed later). The measurement of the Lamb shift, repeated several times, gave results in nice agreement with the original value, and for several years it was providing either a test of QED or a precise value for α. But the Lamb shift is the energy difference between the metastable level $2S_{1/2}$ (whose lifetime

is about 1/7 s) and the $2P_{1/2}$ level, which has a lifetime of about 1.596 ns or a natural linewidth of 99.7 MHz. Such a large linewidth poses a strong intrinsic limitation to the precision attainable in the measure of the Lamb shift, which is just ten times larger; as a matter of fact, that precision could never reach the 1×10^{-6} relative error level, while in the meantime the relative precision in a_e reached the 10^{-9} range, replacing the Lamb shift in the role of the leading quantity in high-precision QED.

The Structure of Radiative Corrections

For obvious space problems we can only superficially sketch here the lines along which the perturbative expansion of QED leading to the evaluation of radiative corrections can be built, considering for simplicity only the photon and the electron. One can start from a QED Lagrangian, formally similar to the classical Lagrangian, involving the electron field and the vector potentials of the electromagnetic (or photon) field. The theory is a gauge theory (its physical content should not change if a gradient is added to the vector potentials); it is further an abelian gauge theory as the EMF does not interact directly with itself.

The QED Lagrangian is separated into a free part and an interaction part. From the free part, one derives the wave functions of the free-particle states and the corresponding time-evolution operators (free Green's functions or propagators; let us just recall here that to obtain a convenient photon propagator one has to break the gauge invariance by adding to the Lagrangian a suitable gauge-breaking term), while the interaction part of the Lagrangian gives the "interaction vertices" of the theory.

Aim of the theory is to build the Green's function for the various processes in the presence of the interaction; from these Green's functions, one then derives all the physical quantities of interest.

With the free propagators and the interaction vertices, one generates the perturbative expansion of the Green's functions. The result, namely the contributions to the perturbative expansion (or radiative corrections), can be depicted in terms of Feynman graphs: they consist of various particle lines joined in the interaction vertices, with external lines corresponding to the initial and final particles and internal lines corresponding to intermediate or virtual particle states. Each graph stands for an integral on the momenta of all the intermediate states, each vertex implying among other things an interaction constant, which is $(-e)$ in the case of electron QED, and a δ-function imposing the conservation of the momenta at that vertex. For each process, the Feynman graphs are naturally classified by the total number of the interaction vertices they contain. In the simplest graphs for a given process (the so-called tree graphs) the δ-functions at the vertices make the integrations trivial; but when the number of vertices increases, closed loops of virtual particle states appear, whose evaluation quickly becomes extremely demanding. In QED, each loop gives an extra factor $(-e)^2$ with respect to the tree graph; it is customary to express it in terms of $(\alpha/\pi) = (e/2\pi)^2$, so that the resulting power of (α/π) corresponds to the number of internal loops. The typical QED prediction for a physical quantity is then expressed as a series of powers of the fine structure constant α (and of its logarithm in bound-state problems). As α is small ($\alpha \simeq 1/137$), and the first coefficients of the expansions are usually of the order of 1, a small number of terms in the expansion is in general sufficient to match the precision of the available experimental data.

But the number of different graphs for a given number of loops grows quickly with the number of the loops; in turn, each graph consists in general of a great number of terms and the loop integrations become prohibitively difficult when the number of loops increases, so that the evaluation of radiative corrections proved to be one of the major computational challenges of theoretical physics. As a matter of fact, it prompted the development of computer programs (Veltman 1999) for processing the huge algebraic expressions usually encountered, and of many sophisticated numerical and analytical techniques for performing the loop integrations.

It should be further mentioned here that Feynman graphs written by naively following the above sketched rules are often mathematically ill-defined, taking the form of nonconvergent integrals on the loop momenta. A regularization procedure is needed to give an unambiguous meaning to all the integrals; currently the most powerful regularization is the continuous dimensional regularization scheme, in which the loop integrations are carried out in d continuous dimensions, with d unspecified; renormalization counter-terms are also evaluated in the same scheme, and the physical quantities are recovered in the $d \to 4$ limit (unrenormalized loop integrals and renormalization counterterms are usually singular as powers of $1/(d-4)$ in the $d \to 4$ limit, but all those divergences cancel out in the physical combinations of interest).

QED describes the main interaction of the charged leptons (e, μ, and τ) which have, however, weak interactions as well. Strictly speaking, pure

QED processes do not exist; it is an essential feature of QFT that any existing particle can contribute to the Feynman graphs for any process, when the approximation is pushed to a sufficiently high degree. In particular the photon, which is the main carrier of the QED interaction, is directly coupled also to the strongly interacting particles (the resulting contributions are referred to as "hadronic vacuum polarization" effects).

The precision tests of QED are then to be necessarily searched for in those phenomena where non-QED contributions are presumably small and which involve quantities already well known independently of QED itself. But such high-precision quantities are not always available, and as QED is known better than the rest of physics, very often it is taken to be correct by assumption, and used as a tool for extracting or measuring some of the non-QED quantities relevant to various physical processes.

In any case, as QED predictions are expressed in terms of the fine structure constant α, a determination of α independent of QED is needed; without it, the most precise predictions of QED would simply become measures of α and not tests of the theory.

Finally, it is to be recalled that, ironically, the problem of the convergence of the expansion in powers of α is still open, even if it is commonly accepted that convergence problems will matter only for precisions and corresponding perturbative orders (say at order $1/\alpha \simeq 137$) absolutely out of reach of present experimental and computational possibilities, involving further extremely high energies, where the other fundamental interactions are expected to be as important as QED, so that it would be meaningless to consider only QED.

In the following we will discuss only the QED predictions for bound states and the anomalous magnetic moments of μ and e.

The Bound States

A very good review of the current status of the theory of hydrogen-like atoms can be found in Eides *et al.* (2001), to which we refer for more details and citation of the original papers. The starting point for studying the bound-state problem in QED is the scattering amplitude of two charged particles, predicted by perturbative QED (pQED) as a (formal) series expansion in powers of α. In the static limit $v \to 0$, where v is the relative velocity of the two particles, some of the pQED terms behave as α/v, so that the naive expansion in α becomes meaningless. Fortunately, it is relatively easy to identify the origin of those terms (which are essentially due to the Coulomb interaction between the two charges) and to devise techniques for their resummation. Among them, one can quote the Bethe–Salpeter equation, formally very elegant and complete but difficult to use in practice. A great progress has been achieved by the NRQED (nonrelativistic QED) approach, which is a nonrelativistic theory designed to reproduce the full QED scattering amplitude in the nonrelativistic limit by the *ad hoc* definition, *a posteriori*, of a suitable effective Hamiltonian. The Hamiltonian is then divided into a part containing the Coulomb interaction, which is treated exactly and which gives rise to the bound states, and all the rest, to be treated perturbatively. The power of the NRQED approach was further boosted by the continuous dimensional regularization technique of Feynman graph integrals.

Traditionally, the results are expressed in terms of the energies of the bound states, but as in practice the precise measurements concern the transition frequencies between various levels, it is customary to express any energy contribution to some level, say ΔE, also in terms of the associated frequency $\nu = (\Delta E)/h$, where h is the Planck constant.

The Hydrogen-Like Atoms

Quite in general, a hydrogen-like atom consists of a single electron bound to a positively charge particle, which is a proton for the hydrogen atom, a deuteron nucleus for deuterium, a Helium nucleus for an He^+ ion, a μ^+ meson for muonium, or a positron for positronium. Even if QED alone is not sufficient to treat the dynamical properties of the nuclei, their strong interactions can be described by introducing suitable form factors and a few phenomenological parameters; weak interactions could be treated perturbatively, but are not yet required at the precision levels achieved so far.

The QED results for the hydrogen-like atoms can be expressed in terms of the mass M of the positive particle and of its charge Ze (of course $Z = 1$ for hydrogen). When the electron mass m_e is smaller then M (which is always the case, except the positronium case) one can take as a starting point the QED electron moving in the external field of the positive particle, and treat all the other aspects of the relativistic two-body problem (the so-called recoil effects) perturbatively in m_e/M.

Neglecting the spin of the positive particle, the energy levels of the hydrogen-like atom are identified by the usual principal quantum number n, the orbital angular momentum l (with the convention of writing $S, P, D, \ldots$ instead of $l = 0, l = 1, l = 2, \ldots$) and j, the total angular momentum including the spin of the electron. It turns out that the bound

levels consist of very many contributions of different kinds; dropping quantum number indices for simplicity, the energy levels can be written as an expression of the form

$$E = -\frac{m_e c^2 (Z\alpha)^2}{2}\left(\frac{m_r}{m_e}\right) \times \left[\frac{1}{n^2} + (Z\alpha)^2 f_4 + (Z\alpha)^4 f_6 + \cdots\right] + \Delta E_{rad} + \Delta E_{rec} + \Delta E_{nucl} + \cdots \quad [5]$$

Let us observe that it is convenient to write explicitly the Z factors even when $Z=1$ for a better bookkeeping of the various corrections. As usual, m_r is the reduced mass of the electron, $m_r = m_e M/(m_e + M)$ the mass of the nucleus being M; the first term in the square bracket, $1/n^2$, the familiar Balmer term, is by and far the dominant one, giving for the $n=1$ level in the $Z=1$ case an energy of about 13.6 eV or a corresponding frequency of 3.3×10^{15} Hz. The other terms in the square bracket, f_4 and f_6, are known coefficients (depending also on the small parameter m_e/M; f_4 is essentially the fine structure).

The term ΔE_{rad}, is the bulk of the radiative QED corrections; it can be written as a multiple expansion on $(Z\alpha), \alpha$ and $L = \ln[1/(Z\alpha)^2]$, which turns out to have the following explicit form:

$$\Delta E_{rad} = m_e c^2 (Z\alpha)^4 \frac{1}{n^3}\left\{\left(\frac{\alpha}{\pi}\right)\left[A_{41}L + A_{40} + (Z\alpha)A_{50} + (Z\alpha)^2\left(A_{62}L^2 + A_{61}L + A_{60}\right) + \cdots\right] + \left(\frac{\alpha}{\pi}\right)^2\left[B_{40} + (Z\alpha)B_{50} + (Z\alpha)^2\left(B_{63}L^3 + B_{62}L^2 + B_{61}L + B_{60}\right) + \cdots\right] + \left(\frac{\alpha}{\pi}\right)^3\left[C_{40} + (Z\alpha)C_{50} + \cdots\right] + \cdots\right\} \quad [6]$$

The first index of the coefficients refers to the power of $(Z\alpha)$, the second to the power of L; as a rule, there are three powers of $(Z\alpha)$ due to the normalization of the wave function and one power of $(Z\alpha)$ for each interaction with the nucleus (in the leading term of eqn [5] one must subtract two powers of $(Z\alpha)$ due to the long-range nature of the Coulomb interaction), while the terms in $L = \ln[1/(Z\alpha)^2]$ are related to the infrared divergences of the scattering amplitude, with the binding energy acting as infrared cutoff. The A-coefficients refers to order (α/π) or one-loop virtual correction (we do not distinguish here between one-loop self-mass and vacuum-polarization contribution, as usually done in the literature), the B-coefficients to two loops, etc. The coefficients are pure numbers, entirely determined within QED, even if their actual calculation is an extremely demanding task. One of the first results obtained in 1947 was $A_{41} = (4/3)\delta_{l0}$, contributing to the $2S$ but not to the $2P$ states (quite in general, most corrections are much bigger for $l=0$ states than for higher-angular-momentum states), which is sufficient to give the right order of magnitude of the $(2S_{1/2}$–$2P_{1/2})$ Lamb shift (about 1000 MHz). The other coefficients are now known, thanks to the strenuous and continued efforts (Eides *et al.* 2001) since then, which is impossible to refer properly here in any detail. The current frontier of the theoretical calculation (around the dots in the previous formula) corresponds to 8–9 total powers of (α/π) and $(Z\alpha)$ or some kHz for the $1S$ state.

The next term in eqn [5], ΔE_{rec} contains contributions of order $m_e c^2 (Z\alpha)^5 (m_e/M)$ or smaller (some care must be done for classifying the contributions of order m_e/M, which can be accounted for by proper use of m_r rather than m_e and genuine m_e/M contributions), and are sufficiently known for practical purposes; the same is true for many other contributions discussed in Eides *et al.* (2001) and skipped in eqn [5]. A troublesome contribution comes however from ΔE_{nucl}; at leading order, one has

$$\Delta E_{nucl} = \frac{2(Z\alpha)^4 mc^2}{3n^3}\left(\frac{mcR_p}{\hbar}\right)^2 \delta_{l0}$$

where R_p is the so-called root-mean-square charge radius of the proton, which is not well known experimentally (in the literature, there are indeed two direct measurements, $R_p = 0.805(11)$ fm and $R_p = 0.862(12)$ fm, in poor agreement with each other; a new independent measurement is strongly needed).

The hyperfine splitting The effect of the interaction of the electron with the spin of the positive particle introduces the so-called hyperfine splitting of all the levels. The order of magnitude of the hyperfine splitting of the $1S$ state is given by the Fermi energy

$$E_F = \frac{4}{3} m_e c^2 (Z\alpha)^4 g_p \frac{m_e}{m_p}$$

where $g_p \simeq 5.586$ is the g-factor of the proton, which gives $\simeq 1.42$ GHz. It was dubbed hyperfine because it is smaller than the fine structure terms by the factor m_e/m_p. Many classes of corrections can be worked out, with patterns similar to those of the previous subsection, and also in this case the nuclear contributions (this time mainly due to the theoretically unknown magnetic form factor and the

so-called polarizability of the proton) prevent from obtaining predictions with an error less than 1 kHz (or a relative precision better than 1×10^{-6}).

The comparison with the experiments Experimen-Experimentally, one measures transition frequencies among the various levels. For many years the precision record was given by the hyperfine splitting of the ground states of hydrogen $\nu_{\rm hfs}(1S)$ was measured long ago (see Hellwig *et al.* (1970) and Essen *et al.* (1971)),

$$\nu_{\rm hfs}(1S) = 1\,420\,405.751\,766\,7(9)\,\text{kHz} \qquad [7]$$

with a relative error 6×10^{-13}. The current record in the optical range is the value of the $(1S–2S)$ hydrogen transition frequency, obtained by means of two-photon Doppler-free spectroscopy Niering *et al.* (2000),

$$\nu(1S–2S) = 2\,466\,061\,413\,187.103(46)\,\text{kHz} \qquad [8]$$

with a relative precision 1.9×10^{-14}; other optical transitions, such as $(2S–8D)$, $(2S–12D)$ are measured with precision of about 1×10^{-11}.

The measurement of the Lamb shift was repeated several times, with results in nice agreement with the original value, such as Lundeen and Pipkin (1986), 1057.845(9) MHz. The most precise value, 1057.8514 ± 0.0019 MHz was given in Palchikov *et al.* (1985) (the result depends, however, on the theoretical value of the lifetime, and should be changed into 1057.8576 ± 0.0021 according to subsequent analysis (see Karshenboim (1996)). The experimental $(2S_{1/2}–2P_{1/2})$ Lamb shift was also obtained as the difference between the measured fine structure separation $(2P_{3/2}–2S_{1/2})$ and the theoretical value of the $(2P_{3/2}–2P_{1/2})$ frequency, and the radiative corrections $\Delta E_{\rm rad}$ to any level are now referred to as the Lamb shift of that level.

As a somewhat deceiving conclusion, the wonderful experimental results of eqns [7] and [8] cannot be used as a high-precision test of the theory or to obtain precise values of many fundamental constants, as the theoretical calculations depend, unfortunately, on hadronic quantities which are not known accurately. Combining theoretical predictions, the above transitions and Lamb shift data, and the available values of α and $m_{\rm e}/m_{\rm p}$, one can indeed obtain a measure of $R_{\rm p}$ ($R_{\rm p} = 0.883 \pm 0.014$, according to Melnikov and van Ritbergen (2000)) and the value of R_∞ already quoted above.

Muonium

The muonium is the bound state of a positive μ^+ meson and an electron. At variance with the proton, the μ^+ lepton has no strong interactions, the μ^+e^- system can be studied theoretically within pure QED, with the weak interactions giving a known and small perturbation. Further, the ratio of the masses $m_{\rm e}/m_\mu \simeq 4.8 \times 10^{-3}$ is small, so that the external field approximation holds. However, the μ is unstable (lifetime $\simeq 2.2\,\mu$s), which makes experiments more difficult to carry out. The best measured quantity is the hyperfine splitting of the $1S$ ground state (see Liu *et al.* (1999))

$$\nu_{\rm hfs}(\mu e, 1S) = 4\,463\,302\,765(53)\,\text{Hz}$$

with a relative precision of 12×10^{-9}. The theoretical treatment is similar to the case of hydrogen, with the important advantage that nuclear interactions are absent and everything can be evaluated within QED, so that the bulk of the contribution is given by a formula with the structure of eqn [6]. But the prediction depends, in any case, on the $m_{\rm e}/m_\mu$ mass, which is not known with the required precision. Indeed, a recent theoretical calculation (Czarnecki *et al.* 2002) (which includes also a contribution of 0.233(3) kHz from hadronic vacuum polarization) gives 4 463 302 680(510)(30)(220) Hz, where the first (and biggest) error comes from $m_{\rm e}/m_\mu$, the second from α, and the third is the theoretical error (an estimation of higher-order contributions not yet evaluated).

Positronium

The positronium is the bound state of an electron and a positron. Theoretically, it is an ideal system to study, as it can be described entirely within QED, without any unknown parameter of non-QED origin. As the masses of the two constituents, positron and electron, are strictly equal, the reduced mass of the system is exactly equal to half of the electron mass, $m_{\rm r} = m_{\rm e}/2$, and the energy scale of the bound states is half of R_∞.

At variance with the muonium case, the external field approximation is not valid, so that positronium must be treated with the full two-body bound-state machinery of QFT, of which it provides an excellent test (Karshenboim 2004).

Experimentally, radioactive positron sources are available, so that positronium is easier to produce than muonium. It is, however, unstable; states with total spin S equal 0 (also called parapositronium states) annihilate into an even number (mainly two) of gammas, and states with $S = 1$ (orthopositronium) into an odd number (mainly three) of gammas, with short lifetimes (which make precise measurements difficult). Further, as positronium is the lightest

atom, Doppler-broadening effects are very important, reducing the precision of spectroscopical measurements.

Positronium decay rates There has been a long-time discrepancy between theory and experiment in decay rate of ground-state orthopositronium, which prompted thorough theoretical investigations looking for errors in the calculations or flaws in the formalism, but it turned out that the flaw was on the experimental side. The current theoretical prediction for the ground state $S=1$ decay is (Adkins *et al.* 2002)

$$\begin{aligned}\Gamma(1S,\text{ortho}) &= \Gamma_0\left\{1 + A\left(\frac{\alpha}{\pi}\right) + \frac{1}{3}\alpha^2\ln\alpha \right.\\ &\quad \left. + B\left(\frac{\alpha}{\pi}\right)^2 - \frac{3\alpha^3}{2\pi}\ln^2\alpha + C\frac{\alpha^3}{\pi}\ln\alpha + \cdots\right\}\\ &= 7.039979(11)\,\mu\text{s}^{-1}\end{aligned}$$

where $\Gamma_0 = 2(\pi^2-9)m_e\alpha^6/(9\pi) = 7.2111670(1)$, $A=-10.286606(10), B=45.06(26), C=-5.517$, in nice agreement with the less precise experimental result of Karshenboim (2004, ref. 38) 7.0404 (10)(8) μs^{-1}. As a curiosity, the coefficients A, B above are among the greatest coefficients so far appeared in QED radiative corrections.

The agreement between theory and experiment for the ground-state parapositronium decay rate has always been good; the current status of Karshenboim (2004, ref. 41) is 7990.9(1.7) μs^{-1} for the experimental result and of Karshenboim (2004, ref. 43) 7989.64(2) μs^{-1} for the theoretical prediction.

Positronium levels The quantum number structure of the levels is similar to muonium, with the important difference, however, that the hyperfine splitting (which in hydrogen or muonium is small because it is proportional to the ratio of the masses of the two components) is in fact of the same order as the fine structure. The theoretical evaluation of the energy levels provides a very stringent check of QED and of the overall treatment of the bound-state problem. Corrections have been evaluated, typically, up to order $mc^2\alpha^7$. The best-known quantities are the ground state (hyper)fine splitting, experimental value (Ritter *et al.* 1984) 203.38910(74) GHz (3.6×10^{-6} relative error), theoretical (Karshenboim 2004) 203.3917(6), and the 1*S*–2*S* transition for orthopositronium, experiment (Fee *et al.* 1993) 1 233 607 216.4 (3.2) MHz, theory 1 233 607 222.2(6). The general agreement is good; the precisions achieved are, however, not yet sufficient to allow a determination of R_∞ or α competitive with other measurements.

The Anomalous Magnetic Moments of Leptons

The precision of the measurements requires, for both the e and μ leptons, to also take into account graphs with contributions from the other leptons as virtual intermediate states and those of hadronic and weak origin. Quite in general, if the mass of the virtual particle, say m_v, is smaller than the mass of the external lepton, say m_l, one can have an $\ln(m_l/m_v)$ behavior of the contributions; that is the case of the virtual electron contributions to the muon magnetic anomaly a_μ, which can be enhanced by powers of $\ln(m_\mu/m_e)$. In the opposite case, $m_v > m_l$, the contribution has the behavior $(m_l/m_v)^2$; that is the case of the $(m_\mu/m_\tau)^2$ contributions to a_μ from τ loops and of the $(m_e/m_\mu)^2$ contributions from μ loops to th electron magnetic anomaly, a_e. As strong and weak interactions are in general associated with heavy-mass particles, they are expected to be more important for a_μ than a_e; further, a given heavy particle contribution to a_e is smaller by a factor $(m_e/m_\mu)^2$ than the corresponding contribution to a_μ.

The Magnetic Anomaly a_μ of the μ

The a_μ has been reviewed in Passera (2005). The present (2005) world average experimental value is

$$a_\mu(\text{exp}) = 116\,592\,080(60)\times10^{-11}$$

with a relative error 0.5×10^{-6}.

Theoretically, one can write

$$a_\mu = a_\mu(\text{QED}) + a_\mu(\text{had}) + a_\mu(\text{EW}) \qquad [9]$$

where the three terms stand for the contributions from pure QED, strong interacting hadrons and electroweak interactions. In turn, one can expand $a_\mu(\text{QED})$ in powers of α as

$$\begin{aligned}a_\mu(\text{QED}) &= \sum_l C_l\left(\frac{\alpha}{\pi}\right)^l\\ &= \sum_l\left[A_1^{(l)} + A_2^{(l)}\left(\frac{m_\mu}{m_e}\right) + A_2^{(l)}\left(\frac{m_\mu}{m_\tau}\right)\right.\\ &\quad \left. + A_3^{(l)}\left(\frac{m_\mu}{m_e},\frac{m_\mu}{m_\tau}\right)\right]\left(\frac{\alpha}{\pi}\right)^l \qquad [10]\end{aligned}$$

The coefficients $A_1^{(l)}$ involve only the photon and the external lepton as virtual states, are identically the same as in a_e; they are known up to $l=4$ included (but, strictly speaking, the contribution of $A_1^{(4)}$ is smaller than the experimental error of a_μ) and will be discussed later for the electron. The $A_2^{(l)}(m_\mu/m_e)$ are very large, being enhanced by powers of $\ln(m_\mu/m_e)$, and are required and known up to $l=5$; $A_2^{(l)}(m_\mu/m_\tau)$ starts with $A_2^{(2)}(m_\mu/m_\tau)\simeq$

$1/45(m_\mu/m_\tau)^2$, contributing 4.2×10^{-11} to a_μ, so that the $A_2^{(l)}(m_\mu/m_\tau)$ with higher values of l are not needed. $A_3^{(l)}(m_\mu/m_e,\ m_\mu/m_\tau)$, finally, starts from $l=3$, and gives a negligible contribution 0.7×10^{-11}. Summing up, one finds $C_1=1/2$, $C_2=0.765\,857\,410(27)$ (the error is from the experimental errors in the lepton masses) $C_3=24.050\,509\,64(43)$, $C_4=131.011(8)$, and $C_5=677(40)$. As already observed, the coefficients are large due to the presence of $\ln(m_\mu/m_e)$ factors. The last term C_5 contributes $4.6(0.3)\times10^{-11}$ to a_μ, and the total QED contribution is

$$a_\mu(\text{QED})=116\,584\,718.8(0.3)(0.4)\times10^{-11}$$

where the first error is due to the uncertainties in the coefficients C_2, C_3, and C_5 and the second from the value of α coming from atom interferometry measurements (see below).

The hadronic contributions are of two kinds, those due to vacuum polarization, a_μ(vac.pol), which can be evaluated by sound theoretical methods by using existing experimental data, and those due to light-by-light hadronic scattering, a_μ(lbl), whose evaluation relies on much less firmer grounds and are entirely model-dependent. The value of a_μ(vac.pol) varies slightly among the various authors (see Passera (2005) for reference to original work), let us take as a typical value $a_\mu(\text{vac.pol})=6834(92)\times10^{-11}$ (based on e^+e^- scattering data and including also first-order radiative corrections). The model-dependent value of the light-by-light contribution changed several times in the years (also in sign!) but now there is a general consensus that it should be positive; let us take, somewhat arbitrarily, $a_\mu(\text{lbl})=136(25)\times10^{-11}$, so that the total hadronic contribution becomes

$$a_\mu(\text{had})=6970(92)\times10^{-11}$$

The electroweak contribution, finally, is

$$a_\mu(\text{EW})=154(2)\times10^{-11}$$

which accounts for a one-loop purely weak contribution and a two-loop electromagnetic and weak contribution, which turns out to be very large (-42×10^{-11}) for the presence of logarithms in the masses (the error is due to the uncertainty in the Higgs boson mass).

Summing up, eqn [9] gives $a_\mu=116\,591\,842(92)\times10^{-11}$, so that

$$a_\mu(\text{exp})-a_\mu=138(60)(90)\times10^{-11}$$

The substantial agreement can be considered to be a good overall check of QED and electroweak interactions. But another attitude is often adopted in the scientific community: the validity of QED and electroweak models is taken for granted, and a disagreement, if any, is considered to be an indication of new physics. To obtain significant information in that direction, however, the experimental and the theoretical errors (dominated in turn by the experimental error in e^+e^- scattering data) should be significantly reduced.

The Magnetic Anomaly a_e of the Electron

Experimentally, one has the 1987 value (Kinoshita 2005, ref. 1).

$$a_e(\text{exp})=1\,159\,652\,188.4(4.3)\times10^{-12} \qquad [11]$$

with a relative error 3.7×10^{-9} and the preliminary Harvard (2004) measurement (Kinoshita 2005, ref. 3).

$$a_e(\text{Harvard})=1\,159\,652\,180.86(0.57)\times10^{-12} \qquad [12]$$

with 0.5×10^{-9} relative error, that is, an increase in precision by a factor 7.

Theoretically, eqns [9] and [10] apply also to the electron; given the smallness of the electron mass, the relevant terms up to the precision of the experimental data are

$$\begin{aligned}a_e=&A_1^{(1)}\left(\frac{\alpha}{\pi}\right)+A_1^{(2)}\left(\frac{\alpha}{\pi}\right)^2+A_1^{(3)}\left(\frac{\alpha}{\pi}\right)^3\\&+A_1^{(4)}\left(\frac{\alpha}{\pi}\right)^4+\cdots+A_2^{(2)}\left(\frac{m_e}{m_\mu}\right)\left(\frac{\alpha}{\pi}\right)^2\\&+a_e(\text{had})+a_e(\text{EW})\end{aligned} \qquad [13]$$

The explicit calculation gives

$$A_1^{(1)}=\frac{1}{2}\quad\text{(Passera 2005, ref. 1)}$$

$$\begin{aligned}A_1^{(2)}&=\frac{197}{144}+\frac{1}{12}\pi^2-\frac{1}{2}\pi^2\ln2+\frac{3}{4}\zeta(3)\\&=-0.328\,478\,965\,579\ldots\end{aligned}$$

(Passera 2005, ref. 17)

$$\begin{aligned}A_1^{(3)}=&\frac{83}{72}\pi^2\zeta(3)-\frac{215}{24}\zeta(5)\\&+\frac{100}{3}\left[a_4+\frac{1}{24}\ln^4 2-\frac{1}{24}\pi^2\ln^2 2\right]\\&-\frac{239}{2160}\pi^4+\frac{139}{18}\zeta(3)-\frac{298}{9}\pi^2\ln2\\&+\frac{17101}{810}\pi^2+\frac{28259}{5184}\\=&1.181\,241\,456\ldots\end{aligned}$$

(Laporta and Remiddi 1996)

$$A_1^{(4)}=-1.7283(35)\quad\text{(Kinoshita 2005)}$$

and

$$A_2^{(2)}\left(\frac{m_e}{m_\mu}\right)\left(\frac{\alpha}{\pi}\right)^2 \simeq \frac{1}{45}\left(\frac{m_e}{m_\mu}\right)^2\left(\frac{\alpha}{\pi}\right)^2 \simeq 2.72\times 10^{-12}$$

$$a_e(\text{had}) = 1.67(0.02)\times 10^{-12}$$

$$a_e(\text{EW}) = 0.03\times 10^{-12} \qquad [14]$$

For obtaining a meaningful prediction, one needs now a precise value of α. The most precise value available at present is that of Passera (2005, ref. 49)

$$\alpha^{-1}(\text{aif}) = 137.036\,000\,3(10)$$

with relative error 7×10^{-9}, obtained by the atom interferometry method (which is independent of QED, depending only on the kinematics of the Doppler effect). With that value of α, the theoretical prediction for a_e becomes

$$a_e = 1\,159\,652\,175.9(8.5)(0.1)10^{-12}$$

where the first error comes from α and the second from C_4; conversely, one can use the QED prediction for a_e and a_e(Harvard) for obtaining α; one obtains in that way

$$\alpha^{-1}(\text{QED}, a_e) = 137.035\,999\,708(12)(67)$$

where the first uncertainty is from C_4 and the second from the experiment. We see that theory and experiment are in good agreement.

As a concluding remark, another independent and more precise (or analytic!) evaluation of C_4 contribution would be welcome. The five-loop term is not known; but as $(\alpha/\pi)^5 \approx 0.07\times 10^{-12}$, if C_5 is, say, not greater than 2, its contribution to a_e becomes equal to the contribution of the error ΔC_4 of C_4 and is not yet required to match the current precision of $a_e(\text{exp})$. The ultimate theoretical limit, the error of the hadronic contribution, $\Delta a_e(\text{had}) = 0.02\times 10^{-12}$, is still smaller, corresponding to a change $\Delta C_4 = 0.0007$ of C_4 or $\Delta C_5 = 0.3$ of C_5.

See also: Abelian and Nonabelian Gauge Theories Using Differential Forms; Anomalies; Effective Field Theories; Electroweak Theory; Quantum Field Theory: A Brief Introduction; Standard Model of Particle Physics.

Further Reading

Adkins GS, Fell RN, and Sapirstein J (2002) Two-loop correction to the orthopositronium decay rate. *Annals of Physics (NY)* 295: 136.

Czarnecki A, Eidelman SI, and Karshenboim SG (2002) Muonium hyperfine structure and hadronic effects. *Physical Review D* 65: 053004.

Dirac PAM (1927) The quantum theory of emission and absorption of radiation. *Proceedings of the Royal Society A* 114: 243.

Eides MI, Grotch H, and Shelyuto VA (2001) Theory of light hydrogenlike atoms. *Physics Reports* 342: 63 (arXiv:hep-ph/0002158).

Essen L, Donaldson RW, Bangman MJ, and Hope EG (1971) Frequency of the hydrogen maser. *Nature* 229: 110.

Fee MS *et al.* Measurement of the positronium 1^3s_1–2^3s_1 interval by continuous-wave two-photon excitation. *Physical Review Letters* 70: 1397.

Hellwig H, Vessot RFC, Levine MW, Zitzewitz PW, Allan DW, and Glaze DJ (1970) Measurement of the unperturbed hydrogen hyperfine transition frequency. *IEEE Transactions on Instrument Measurement* IM19: 200.

Karshenboim SG (1996) Leading logarithmic corrections and uncertainty of muonium hyperfine splitting calculations. *Z. Physics D* 36: 11.

Karshenboim SG (2004) Precision study of positronium: testing bound state QED theory. *International Journal of Modern Physics A* 19: 3879 (e-print hep-ph/0310099).

Kinoshita T (2005) Improved α^4 term of the electron anomalous magnetic moment (hep-ph/0507249).

Kusch P and Foley HM (1947) Precision measurements of the ratio of the atomic 'g values' in the $^2P_{3/2}$ and $^2P_{1/2}$ states of gallium. *Physical Review* 72: 1256(L).

Lamb WE Jr. and Retherford RC (1947) Fine structure of the hydrogen atom by a microwave method. *Physical Review* 72: 241.

Laporta S and Remiddi E (1996) The analytical value of the electron $(g-2)$ at order α^3 in QED. *Physics Letters B* 379: 283.

Liu W *et al.* (1999) High precision measurements of the ground state hyperfine structure interval of muonium and of the muon magnetic moment. *Physical Review Letters* 82: 711.

Lundeen SR and Pipkin FM (1986) Separated oscillatory field measurement of Lamb shift in atomic hydrogen, $n=2$. *Metrologia* 22(1): 9.

Melnikov K and van Ritbergen T (2000) Three-loop slope of the Dirac form factor and the 1S Lamb shift in hydrogen. *Physical Review Letters* 84: 1673 (hep-ph/9911277).

Niering M *et al.* (2000) Measurement of the hydrogen 1s–2s transition frequency by phase coherent comparison with a microwave cesium fountain clock. *Physics Review Letters* 84: 5496.

Pal'chikov VG, Sokolov YL, and Yakovlev VP (1985) Measurement of the Lamb shift in H, $n=2$. *Metrologia* 21: 99.

Passera M (2005) The standard model prediction of the muon anomalous magnetic moment. *Journal of Physics G* 31: R75 (hep-ph/0411168).

Ritter MW, Egan PO, Hughes VW, and Woodle KA (1984) Precision determination of the hyperfine-structure interval in the ground state of positronium, V. *Physical Review A* 30: 1331.

Triebwasser S, Dayhoff ES, and Lamb WE Jr. (1953) Fine structure of the hydrogen atom, V. *Physical Review* 89: 98.

Veltman MJC (1999) *Nobel Lecture* (http://nobelprize.org/physics/laureates/1999/veltman-lecture.pdf).

Weinberg S (1995) *The Quantum Theory of Fields*. Cambridge: Cambridge University Press.

Quantum Entropy

D Petz, Budapest University of Technology and Economics, Budapest, Hungary

In the past 50 years, entropy has broken out of thermodynamics and statistical mechanics and invaded communication theory, ergodic theory mathematical statistics, and even the social and life sciences. The favorite subjects of entropy concern macroscopic phenomena, irreversibility, and incomplete knowledge. In the strictly mathematical sense entropy is related to the asymptotics of probabilities or concerns the asymptotic behavior of probabilities.

This review is organized as follows. First the history of entropy is discussed generally and then we concentrate on the von Neumann entropy again somewhat historically following the work of von Neumann. Umegaki's quantum relative entropy is discussed both in case of finite systems and in the setting of C^*-algebras. An axiomatization is presented. To show physical applications of the concept of entropy, the statistical thermodynamics is reviewed in the setting of spin chains. The relative entropy shows up in the asymptotic theory of hypothesis testing and data compression.

General Introduction to Entropy: From Clausius to von Neumann

The word "entropy" was created by Rudolf Clausius and it appeared in his work *Abhandlungen über die mechanische Wärmetheorie* published in 1864. The word has a Greek origin, its first part reminds us of "energy" and the second part is from "tropos," which means "turning point." Clausius' work is the foundation stone of classical thermodynamics. According to Clausius, the change of entropy of a system is obtained by adding the small portions of heat quantity received by the system divided by the absolute temperature during the heat absorption. This definition is satisfactory from a mathematical point of view and gives nothing other than an integral in precise mathematical terms. Clausius postulated that the entropy of a closed system cannot decrease, which is generally referred to as the second law of thermodynamics.

The concept of entropy was really clarified by Ludwig Boltzmann. His scientific program was to deal with the mechanical theory of heat in connection with probabilities. Assume that a macroscopic system consists of a large number of microscopic ones, we simply call them particles. Since we have ideas of quantum mechanics in mind, we assume that each of the particles is in one of the energy levels $E_1 < E_2 < \cdots < E_m$. The number of particles in the level E_i is N_i, so $\sum_i N_i = N$ is the total number of particles. A macrostate of our system is given by the occupation numbers $N_1, N_2, \ldots, N_m$. The energy of a macrostate is $E = \sum_i N_i E_i$. A given macrostate can be realized by many configurations of the N particles, each of them at a certain energy level E_i. These configurations are called microstates. Many microstates realize the same macrostate. We count the number of ways of arranging N particles in m boxes (i.e., energy levels) such that each box has $N_1, N_2, \ldots, N_m$ particles. There are

$$\binom{N}{N_1, N_2, \ldots, N_m} := \frac{N!}{N_1! N_2! \ldots N_m!} \qquad [1]$$

such ways. This multinomial coefficient is the number of microstates realizing the macrostate $(N_1, N_2, \ldots, N_m)$ and it is proportional to the probability of the macrostate if all configurations are assumed to be equally likely. Boltzmann called [1] the thermodynamical probability of the macrostate, in German "thermodynamische Wahrscheinlichkeit," hence the letter W was used. Of course, Boltzmann argued in the framework of classical mechanics and the discrete values of energy came from an approximation procedure with "energy cells."

If we are interested in the thermodynamic limit N increasing to infinity, we use the relative numbers $p_i := N_i/N$ to label a macrostate and, instead of the total energy $E = \sum_i N_i E_i$, we consider the average energy pro particle $E/N = \sum_i p_i E_i$. To find the most probable macrostate, we wish to maximize [1] under a certain constraint. The Stirling approximation of the factorials gives

$$\frac{1}{N} \log \binom{N}{N_1, N_2, \ldots, N_m} = H(p_1, p_2, \ldots, p_m) + O(N^{-1} \log N) \qquad [2]$$

where

$$H(p_1, p_2, \ldots, p_m) := \sum_i -p_i \log p_i \qquad [3]$$

If N is large then the approximation [2] yields that instead of maximizing the quantity [1] we can maximize [3]. For example, maximizing [3] under the constraint $\sum_i p_i E_i = e$, we get

$$p_i = \frac{\mathrm{e}^{-\lambda E_i}}{\sum_j \mathrm{e}^{-\lambda E_j}} \qquad [4]$$

where the constant λ is the solution of the equation

$$\sum_i E_i \frac{e^{-\lambda E_i}}{\sum_j e^{-\lambda E_j}} = e$$

Note that the last equation has a unique solution if $E_1 < e < E_m$, and the distribution [4] is now known as the discrete Maxwell–Boltzmann law.

Let $p_1, p_2, \ldots, p_n$ be the probabilities of different outcomes of a random experiment. According to Shannon, the expression [1] is a measure of our ignorance prior to the experiment. Hence it is also the amount of information gained by performing the experiment. The quantity [1] is maximum when all the p_i's are equal. In information theory, logarithms with base 2 are used and the unit of information is called bit (from binary digit). As will be seen below, an extra factor equal to Boltzmann's constant is included in the physical definition of entropy.

The comprehensive mathematical formalism of quantum mechanics was first presented in the famous book *Mathematische Grundlagen der Quantenmechanik* published in 1932 by Johann von Neumann. In the traditional approach to quantum mechanics, a physical system is described in a Hilbert space: observables correspond to self-adjoint operators and statistical operators are associated with the states. In fact, a statistical operator describes a mixture of pure states. Pure states are really the physical states and they are given by rank-1 statistical operators, or equivalently by rays of the Hilbert space.

von Neumann associated an entropy quantity to a statistical operator in 1927 and the discussion was extended in his book (von Neumann 1932). His argument was a gedanken experiment on the grounds of phenomenological thermodynamics. Let us consider a gas of $N(\gg 1)$ molecules in a box. Suppose that the gas behaves like a quantum system and is described by a statistical operator ω which is a mixture $\sum_i \lambda_i |\varphi_i\rangle\langle\varphi_i|$, where $|\varphi_i\rangle \equiv \varphi_i$ are orthogonal state vectors. We may take $\lambda_i N$ molecules in the pure state φ_i for every i. The gedanken experiment gave

$$S\left(\sum_i \lambda_i |\varphi_i\rangle\langle\varphi_i|\right) = \sum_i \lambda_i S(|\varphi_i\rangle\langle\varphi_i|) - \kappa \sum_i \lambda_i \log \lambda_i \qquad [5]$$

where κ is Boltzmann's constant and S is certain thermodynamical entropy quantity (relative to the fixed temperature and molecule density).

After this, von Neumann showed that $S(|\varphi\rangle\langle\varphi|)$ is independent of the state vector $|\varphi\rangle$, so that

$$S\left(\sum_i \lambda_i |\varphi_i\rangle\langle\varphi_i|\right) = -\kappa \sum_i \lambda_i \log \lambda_i \qquad [6]$$

up to an additive constant, which could be chosen to be 0 as a matter of normalization. Equation [6] is von Neumann's celebrated entropy formula; it has a more elegant form

$$S(\omega) = \kappa \operatorname{tr} \eta(\omega) \qquad [7]$$

where the state ω is identified with the corresponding statistical operator, and $\eta : \mathbb{R}^+ \to \mathbb{R}$ is the continuous function $\eta(t) = -t \log t$.

von Neumann solved the maximization problem for $S(\omega)$ under the constraint $\operatorname{tr} \omega H = e$. This means the determination of the ensemble of maximal entropy when the expectation of the energy operator H is a prescribed value e. It is convenient to rephrase his argument in terms of conditional expectations. $H = H^*$ is assumed to have a discrete spectrum and we have a conditional expectation E determined by the eigenbasis of H. If we pass from an arbitrary statistical operator ω with $\operatorname{tr} \omega H = e$ to $E(\omega)$, then the entropy is increasing, on the one hand, and the expectation of the energy does not change, on the other, so the maximizer should be searched among the operators commuting with H. In this way we are (and von Neumann was) back to the classical problem of statistical mechanics treated at the beginning of this article. In terms of operators, the solution is in the form

$$\frac{\exp(-\beta H)}{\operatorname{tr} \exp(-\beta H)} \qquad [8]$$

which is called Gibbs state today.

The von Neumann Entropy

von Neumann was aware of the fact that statistical operators form a convex set whose extreme points are exactly the pure states. He also knew that entropy is a concave functional, so

$$S\left(\sum_i \lambda_i \omega_i\right) \geq \sum_i \lambda S(\omega_i) \qquad [9]$$

for any convex combination. To determine the entropy of a statistical operator, he used the Schatten decomposition, which is an orthogonal extremal decomposition in our present language. For a statistical operator ω there are many ways to write it in the form

$$\omega = \sum_i \lambda_i |\psi_i\rangle\langle\psi_i|$$

if we do not require the state vectors to be orthogonal. The geometry of the statistical operators, that is, the state space, allows many extremal decompositions and among them there is a unique orthogonal one if the spectrum of ω is not

degenerate. Nonorthogonal pure states are essentially nonclassical. They are between identical and completely different. Jaynes recognized in 1956 that from the point of view of information the Schatten decomposition is optimal. He proved that

$$S(\omega) = \sup\left\{-\sum\nolimits_i \lambda_i \log \lambda_i \colon\ \omega = \sum\nolimits_i \lambda_i \omega_i\right\} \qquad [10]$$

where the supremum is over all convex combinations $\omega = \sum_i \lambda_i \omega_i$ statistical operators. This is Jaynes contribution to the von Neumann entropy. By the way, formula [10] may be used to define von Neumann entropy for states of an arbitrary C*-algebra whose states cannot be described by statistical operators.

Certainly the highlight of quantum entropy theory in the 1970s was the discovery of subadditivity. This property is formulated in a tripartite system whose Hilbert space $\mathcal{H}$ is a tensor product $\mathcal{H}_A \otimes \mathcal{H}_B \otimes \mathcal{H}_C$. A statistical operator ω_{ABC} admits several reduced densities, ω_{AB}, ω_B, ω_{BC}, and others. The strong subadditivity is the inequality due to Lieb and Ruskai in 1973:

Theorem 1

$$S(\omega_{ABC}) + S(\omega_B) \leq S(\omega_{AB}) + S(\omega_{BC}) \qquad [11]$$

The strong subadditivity inequality [11] is conveniently rewritten in terms of the relative entropy. For statistical operators ρ and ω,

$$S(\rho\|\omega) = \operatorname{tr} \rho(\log \rho - \log \omega) \qquad [12]$$

if $\operatorname{supp} \rho \leq \operatorname{supp} \omega$, otherwise $S(\rho\|\omega) = +\infty$. The relative entropy expresses statistical distinguishability and therefore it decreases under stochastic mappings:

$$S(\rho\|\omega) \geq S(\mathcal{E}(\rho)\|\mathcal{E}(\omega)) \qquad [13]$$

for a completely positive trace-preserving mapping $\mathcal{E}$.

The strong subadditivity is equivalent to

$$S(\omega_{AB}, \varphi \otimes \omega_B) \leq S(\omega_{ABC}, \varphi \otimes \omega_{BC}) \qquad [14]$$

where φ is any state on $B(\mathcal{H}_A)$ of finite entropy. This inequality is a consequence of monotonicity of the relative entropy, since $\omega_{AB} = \mathcal{E}(\omega_{ABC})$ and $\varphi \otimes \omega_B = \mathcal{E}(\varphi \otimes \omega_{BC})$, where $\mathcal{E}$ is the partial trace over $\mathcal{H}_C$. Clearly, the equality in [11] is equivalent to equality in [14].

Theorem 2 *The equality holds in* [11] *if and only if there is an orthogonal decomposition* $p_B \mathcal{H}_B = \bigoplus_n \mathcal{H}_{nB}^{\mathrm{L}} \otimes \mathcal{H}_{nB}^{\mathrm{R}}, p_B = \operatorname{supp} \omega_B$, *such that the density operator of* ω_{ABC} *satisfies*

$$\omega_{ABC} = \sum_n \omega_B(p_n) \omega_n^{\mathrm{L}} \otimes \omega_n^{\mathrm{R}} \qquad [15]$$

where $\omega_n^{\mathrm{L}} \in B(\mathcal{H}_A) \otimes B(\mathcal{H}_{nB}^{\mathrm{L}})$ *and* $\omega_n^{\mathrm{R}} \in B(\mathcal{H}_{nB}^{\mathrm{R}}) \otimes B(\mathcal{H}_C)$ *are density operators and* $p_n \in B(\mathcal{H}_B)$ *are the orthogonal projections* $\mathcal{H}_B \to \mathcal{H}_{nB}^{\mathrm{L}} \otimes \mathcal{H}_{nB}^{\mathrm{R}}$.

Quantum Relative Entropy

The quantum relative entropy is an information measure representing the uncertainty of a state with respect to another state. Hence it indicates a kind of distance between the two states. The formal definition [12] is due to Umegaki.

Now we approach quantum relative entropy axiomatically. Our crucial postulate includes the notion of conditional expectation. Let us recall that in the setting of operator algebras conditional expectation (or projection of norm 1) is defined as a positive unital idempotent linear mapping onto a subalgebra.

Now we list the properties of the relative entropy functional which will be used in an axiomatic characterization:

1. *Conditional expectation property*. Assume that $\mathcal{A}$ is a subalgebra of $\mathcal{B}$ and there exists a projection of norm 1 E of $\mathcal{B}$ onto $\mathcal{A}$, such that $\varphi \circ E = \varphi$. Then for every state ω of $\mathcal{B}$ $S(\omega, \varphi) = S(\omega|\mathcal{A}, \varphi|\mathcal{A}) + S(\omega, \omega \circ E)$ holds.
2. *Invariance property*. For every automorphism α of $\mathcal{B}$ we have $S(\omega, \varphi) = S(\omega \circ \alpha, \varphi \circ \alpha)$.
3. *Direct sum property*. Assume that $\mathcal{B} = \mathcal{B}_1 \oplus \mathcal{B}_2$. Let $\varphi_{12}(a \oplus b) = \lambda\varphi_1(a) + (1-\lambda)\varphi_2(b)$ and $\omega_{12}(a \oplus b) = \lambda\omega_1(a) + (1-\lambda)\omega_2(b)$ for every $a \in \mathcal{B}_1$, $b \in \mathcal{B}_2$ and some $0 < \lambda < 1$. Then $S(\omega_{12}, \varphi_{12}) = \lambda S\ (\omega_1, \varphi_1) + (1-\lambda)S(\omega_2, \varphi_2)$.
4. *Nilpotence property*. $S(\varphi, \varphi) = 0$.
5. *Measurability property*. The function $(\omega, \varphi) \mapsto S(\omega, \varphi)$ is measurable on the state space of the finite dimensional C*-algebra $\mathcal{B}$ (when φ is assumed to be faithful).

Theorem 3 *If a real valued functional* $R(\omega, \varphi)$ *defined for faithful states* φ *and arbitrary states* ω *of finite quantum systems shares the properties* [1]–[5], *then there exists a constant* $c \in \mathbb{R}$ *such that*

$$R(\omega, \varphi) = c \operatorname{Tr} D_\omega (\log D_\omega - \log D_\varphi)$$

The relative entropy may be defined for linear functionals of an arbitrary C*-algebra. The general definition may go through von Neumann algebras, normal states and the relative modular operator. Another possibility is based on the monotonicity. Let ω and φ be states of a C*-algebra $\mathcal{A}$. Consider finite-dimensional algebras $\mathcal{B}$ and completely positive unital mappings $\alpha : \mathcal{B} \to \mathcal{A}$. Then the supremum

of the relative entropies $S(\omega\circ\alpha\|\varphi\circ\alpha)$ (over all α) can be defined as $S(\omega\|\varphi)$.

Theorem 4 *The relative entropy of states of C*-algebras shares the following properties.*

(i) $(\omega,\varphi)\mapsto S(\omega\|\varphi)$ *is convex and weakly lower-semicontinuous.*
(ii) $\|\varphi-\omega\|^2\le 2S(\omega,\varphi)$.
(iii) *For a unital Schwarz map* $\alpha:\mathcal{A}_0\to\mathcal{A}_1$ *the relation* $S(\omega\circ\alpha\|\varphi\circ\alpha)\le S(\omega\|\varphi)$ *holds.*

Property (iii) is Uhlmann's monotonicity theorem, which we have already applied above.

The relative entropy appears in many concepts and problems in the area of quantum information theory (Nielsen and Chuang 2000, Schumacher and Westmoreland 2002).

Statistical Thermodynamics

Let an infinitely extended system of quantum spins be considered in the simple cubic lattice $L=\mathbb{Z}^\nu$, where ν is a positive integer. The observables confined to a lattice site $x\in\mathbb{Z}^\nu$ form the self-adjoint part of a finite-dimensional C*-algebra $\mathcal{A}_x$ which is a copy of the matrix algebra $M_d(\mathbb{C})$. It is assumed that the local observables in any bounded region $\Lambda\subset\mathbb{Z}^\nu$ are those of the finite quantum system

$$\mathcal{A}_\Lambda=\bigotimes_{x\in\Lambda}\mathcal{A}_x$$

It follows from the definition that for $\Lambda\subset\Lambda'$ we have $\mathcal{A}_{\Lambda'}=\mathcal{A}_\Lambda\otimes\mathcal{A}_{\Lambda'\setminus\Lambda}$, where $\Lambda'\setminus\Lambda$ is the complement of Λ in Λ'. The algebra $\mathcal{A}_\Lambda$ and the subalgebra $\mathcal{A}_\Lambda\otimes\mathbb{C}I_{\Lambda'\setminus\Lambda}$ of $\mathcal{A}_{\Lambda'}$ have identical structure and we identify the element $A\in\mathcal{A}_\Lambda$ with $A\otimes I_{\Lambda'\setminus\Lambda}$ in $\mathcal{A}_{\Lambda'}$. If $\Lambda\subset\Lambda'$ then $\mathcal{A}_\Lambda\subset\mathcal{A}_{\Lambda'}$ and it is said that $\mathcal{A}_\Lambda$ is isotonic with respect to Λ. The definition also implies that if Λ_1 and Λ_2 are disjoint then elements of $\mathcal{A}_{\Lambda_1}$ commute with those of $\mathcal{A}_{\Lambda_2}$. The quasilocal C*-algebra $\mathcal{A}$ is the norm completion of the normed algebra $\mathcal{A}_\infty=\cup_\Lambda\mathcal{A}_\Lambda$, the union of all local algebras $\mathcal{A}_\Lambda$ associated with bounded (finite) regions $\Lambda\subset\mathbb{Z}^\nu$.

We denote by a_x the element of $\mathcal{A}_x$ corresponding to $a\in\mathcal{A}_0(x\in\mathbb{Z}^\nu)$. It follows from the definition that the algebra $\mathcal{A}_\infty$ consists of linear combinations of terms $a^{(1)}_{x_1}\cdots a^{(k)}_{x_k}$ where $x_1,\ldots,x_k$ and $a^{(1)},\ldots,a^{(k)}$ run through $\mathbb{Z}^\nu$ and $\mathcal{A}_0$, respectively. We define γ_x to be the linear transformation

$$a^{(1)}_{x_1}\cdots a^{(k)}_{x_k}\longmapsto a^{(1)}_{x_1+x}\cdots a^{(k)}_{x_k+x}$$

γ_x corresponds to the space translation by $x\in\mathbb{Z}^\nu$ and it extends to an automorphism of $\mathcal{A}$. Hence γ is a representation of the abelian group $\mathbb{Z}^\nu$ by automorphisms of the quasilocal algebra $\mathcal{A}$. Clearly, the covariance condition

$$\gamma_x(\mathcal{A}_\Lambda)=\mathcal{A}_{\Lambda+x}$$

holds, where $\Lambda+x$ is the space-translate of the region Λ by the displacement x.

Having described the kinematical structure of lattice systems, we turn to the dynamics. The local Hamiltonian $H(\Lambda)$ is taken to be the total potential energy between the particles confined to Λ. This energy may come from many-body interactions of various orders. Most generally, we assume that there exists a global function Φ such that for any finite subsystem Λ the local Hamiltonian takes the form

$$H(\Lambda)=\sum_{X\subset\Lambda}\Phi(X) \qquad [16]$$

Each $\Phi(X)$ represents the interaction energy of the particles in X. Mathematically, $\Phi(X)$ is a self-adjoint element of $\mathcal{A}_X$ and $H(\Lambda)$ will be a self-adjoint operator in $\mathcal{A}_\Lambda$. We restrict our discussion to translation-invariant interactions, which satisfy the additional requirement

$$\gamma_x(\Phi(X))=\Phi(X+x)$$

for every $x\in\mathbb{Z}^\nu$ and every region $X\subset\mathbb{Z}^\nu$. An interaction Φ is said to be of finite range if $\Phi(\Lambda)=0$ when the cardinality (or diameter) of Λ is large enough, $d(\Lambda)\ge d_\Phi$. The infimum of such numbers is called the range of Φ.

If φ is a state of the quasilocal algebra $\mathcal{A}$ then it will induce a state φ_Λ on $\mathcal{A}(\Lambda)$, the finite system comprising the spin in the bounded region Λ of $\mathbb{Z}^\nu$. The (local) energy, entropy, and free energy of this finite system are given by the following formulas:

$$\begin{aligned}E_\Lambda(\varphi)&:=\mathrm{tr}_\Lambda\omega_\Lambda H(\Lambda)\\ S_\Lambda(\varphi)&:=-\mathrm{tr}_\Lambda\omega_\Lambda\log\omega_\Lambda\\ F^\beta_\Lambda(\varphi)&:=E_\Lambda(\varphi)-\frac{1}{\beta}S_\Lambda(\varphi)\end{aligned} \qquad [17]$$

Here ω_Λ denotes the density of φ_Λ with respect to the trace tr_Λ of $\mathcal{A}_\Lambda$, and β denotes the inverse temperature. The functionals E_Λ, S_Λ, and F^β_Λ are termed local. It is rather obvious that all three local functionals are continuous if the weak* topology is considered on the state space of the quasilocal algebra. The energy is affine, the entropy is concave and consequently, the free energy is a convex functional.

The free energy functional F^β_Λ is minimized by the Gibbs state (see [8] with $H=H(\Lambda)$), and the minimum value is given by

$$-\frac{1}{\beta}\log\,\mathrm{tr}_\Lambda e^{-\beta H(\Lambda)} \qquad [18]$$

Our aim is to explain this variational principle after the thermodynamic limit is performed.

The thermodynamic limit "Λ tends to infinity" may be taken along lattice parallelepipeds. Let $a \in \mathbb{Z}^\nu$ with positive coordinates and define

$$\Lambda(a) = \{x \in \mathbb{Z}^\nu:\ 0 \le x_i < a_i,\ i = 1, 2, \ldots, \nu\} \quad [19]$$

When $a \to \infty$, $\Lambda(a)$ tends to infinity in a manner suitable for the study of thermodynamic limit: the boundary of the parallelepipeds is getting more and more negligible compared with the volume. The notion of limit in the sense of van Hove makes this idea more precise and physically more satisfactory. For the sake of simplicity, we restrict ourselves to thermodynamic limit along parallelepipeds.

Denoting by $|\Lambda|$ the volume of Λ (or the number of points in Λ), we may define the global energy, entropy, and free energy functionals of translationally invariant states to be

$$e(\varphi) := \lim_{\Lambda\to\infty} E_\Lambda(\varphi)/|\Lambda| \quad [20]$$

$$s(\varphi) := \lim_{\Lambda\to\infty} S_\Lambda(\varphi)/|\Lambda| \quad [21]$$

$$f^\beta(\varphi) := \lim_{\Lambda\to\infty} F^\beta_\Lambda(\varphi)/|\Lambda| \quad [22]$$

The existence of the limit in [21] is guaranteed by the strong subadditivity of entropy, while that of the limits in [20] and [22] is assumed if the interaction is suitably tempered, as it certainly does if the interaction is of finite range.

Theorem 5 *If φ is a translationally invariant state of the quasilocal algebra $\mathcal{A}$, then the limit* [21] *exists and*

$$s(\varphi) = \inf\{S_{\Lambda(a)}(\varphi)/|\Lambda(a)|:\ a \in \mathbb{Z}^\nu_+\} \quad [23]$$

Moreover, the von Neumann entropy density functional $\varphi \mapsto s(\varphi)$ is affine and upper-semicontinuous when the state space is endowed with the weak topology.*

Let Φ be an interaction of finite range. Then the thermodynamic limit [20] *exists and the energy density is given by*

$$e(\varphi) = \varphi(E_\Phi) \quad \text{and} \quad E_\Phi = \sum_{0\in\Lambda} \frac{\Phi(\Lambda)}{|\Lambda|}$$

Furthermore, $e(\varphi)$ is an affine weak continuous functional of φ.*

It follows that the free energy density $f(\varphi)$ exists and it is an affine lower-semicontinuous function of the translation-invariant state φ.

For $0 < \beta < \infty$ the thermodynamic limit

$$\lim_{\Lambda\to\infty} \frac{1}{|\Lambda|} \log \mathrm{tr}_\Lambda e^{-\beta H(\Lambda)} \equiv p(\beta, \Phi)$$

exists.

In accordance with the lattice-gas interpretation of our model, the global quantity p is termed pressure.

In the treatment of quantum spin systems, the set $\mathfrak{S}_\gamma$ of all translation-invariant states is essential. The global entropy functional s is a continuous affine function on $\mathfrak{S}_\gamma$ and physically it is a macroscopic quantity which does not have microscopic (i.e., local) counterpart. Indeed, the local entropy functional is not an observable because it is not affine on the (local) state space. The local internal energy $E_\Lambda(\varphi)$ is microscopic observable and the energy density functional e of $\mathfrak{S}_\gamma$ is the corresponding global extensive quantity.

As an analog of the variational principle for finite quantum systems, the global free-energy functional f_β attains an absolute minimum at a translationally invariant state, and the minimum value of f^β is equal to the thermodynamic limit of the canonical free-energy densities of the local finite systems. In the next theorem, this global variational principle will be formulated in a slightly different but equivalent way.

Theorem 6 *When Φ is an interaction of finite range, then*

$$p(\beta, \Phi) = \sup\{s(\omega) - \beta e(\omega)\}$$

holds, when the supremum is over all translationally invariant states ω on $\mathcal{A}$.

The minimizers of the right-hand side are called *equilibrium states* and they have several different characterizations.

Asymptotical Properties

We keep the notation of the previous section but we consider one-dimensional chains, $\nu = 1$. Let ω be translation-invariant state on $\mathcal{A}$ and we fix a positive number $\varepsilon < 1$. We have in our mind that ε is small and say that a sequence of projection $Q_n \in \mathcal{A}_{[1,n]}$ is of high probability if $\omega(Q_n) \ge 1 - \varepsilon$. The size of Q_n, the cardinality of a maximal pairwise orthogonal family of projections contained in Q_n, is given by $\mathrm{tr}_n Q_n$. (The subscript n in tr_n indicates that the algebraic trace functional on $\mathcal{A}_n$ is meant here.) The theorem below says that the entropy density of ω governs asymptotically the rank of the high-probability projections.

Theorem 7 *Assume that ω is an ergodic translation-invariant state of $\mathcal{A}$. Then the limit relation*

$$\lim_{n\to\infty} \frac{1}{n} \inf\{\log \mathrm{tr}_n Q_n\} = s(\omega)$$

holds, when the infimum is over all projections $Q_n \in \mathcal{A}_{[1,n]}$ such that $\omega_n(Q_n) \ge 1 - \varepsilon$.

This result is strongly related to data compression. When ω is interpreted as a stationary quantum source (with possible memory), then efficient and reliable data compression needs a subspace of small dimension and the range of Q_n can play this role. The entropy density is the maximal rate of reliable compression.

It is interesting that one can impose further requirements on the high-probability projections and the statement of the theorem remains true.

1. The partial trace of Q_{n+1} over $\mathcal{A}_{n+1}$ is Q_n;
2. $e^{n(s-\varepsilon)} \le \mathrm{tr}\, Q_n \le e^{n(s+\varepsilon)}$ if n is large enough; and
3. if $q \le Q_n$ is a minimal projection (in $\mathcal{A}_{[1,n]}$), then $\omega(q) \le e^{-n(s-\varepsilon)}$ if n is large enough.

In (2) and (3) s stands for $s(\omega)$. Let D_n be the density matrix of the restriction of ω to $\mathcal{A}_{[1,n]}$. It follows that for an eigenvalue λ of $Q_n D_n Q_n$ the inequality

$$s - \varepsilon \le -\frac{\log \lambda}{n}$$

holds.

From the point of view of data compression, it is important if the sequence $Q_n \in \mathcal{A}_{[1,n]}$ works universally for many states. Indeed, in this case the compression algorithm can be universal for several quantum sources.

Theorem 8 *Let $R > 0$. There is a projection $Q_n \in \mathcal{A}_{[1,n]}$ such that*

$$\limsup_n \frac{1}{n} \log \mathrm{tr} Q_n \le R \qquad [24]$$

and for any ergodic state ω on $\mathcal{A}$ such that $s(\omega) < R$ the relation

$$\lim_n \omega(Q_n) = 1 \qquad [25]$$

holds.

In the simplest quantum hypothesis testing problem, one has to decide between two states of a system. The state ρ_0 is the null hypothesis and ρ_1 is the alternative hypothesis. The problem is to decide which hypothesis is true. The decision is performed by a two-valued measurement $\{T, I - T\}$, where $0 \le T \le I$ is an observable. T corresponds to the acceptance of ρ_0 and $I - T$ corresponds to the acceptance of ρ_1. T is called a test. When the measurement value is 0, the hypothesis ρ_0 is accepted, otherwise the alternative hypothesis ρ_1 is accepted. The quantity $\alpha[T] = \mathrm{tr}\rho_0(I - T)$ is interpreted as the probability that the null hypothesis is true but the alternative hypothesis is accepted. This is the error of the first kind. Similarly, $\beta[T] = \mathrm{tr}\rho_1 T$ is the probability that the alternative hypothesis is true but the null hypothesis is accepted. It is called the error of the second kind.

Now we fix a formalism for an asymptotic theory of the hypothesis testing. Suppose that a sequence $(\mathcal{H}_n)$ of Hilbert spaces is given, $(\rho_0^{(n)})$ and $(\rho_1^{(n)})$ are density matrices on $\mathcal{H}_n$. The typical example we have in mind is $\rho_0^{(n)} = \rho_0 \otimes \rho_0 \otimes \cdots \otimes \rho_0$ and $\rho_1^{(n)} = \rho_1 \otimes \rho_1 \otimes \cdots \otimes \rho_1$. A positive contraction $T_n \in \mathcal{B}(\mathcal{H}_n)$ is considered as a test on a composite system. Now the errors of the first and second kind depend on n: $\alpha_n[T_n] = \mathrm{tr}\rho_0^{(n)}(I - T_n)$ and $\beta_n[T_n] = \mathrm{tr}\rho_1^{(n)} T_n$.

Set

$$\beta^*(n, \varepsilon) = \inf\{\mathrm{tr}\rho_1^{(n)} A_n\} \qquad [26]$$

where the infimum is over all $A_n \in \mathcal{B}(\mathcal{H}_n)$ such that $0 \le A_n \le I$ and $\mathrm{tr}\rho_0^{(n)}(I - A_n) \le \varepsilon$. In other words, this is the infimum of the error of the second kind when the error of the first kind is at most ε. The importance of this quantity is in the customary approach to hypothesis testing.

The following result is the quantum Stein lemma.

Theorem 9 *In the above setting, the relation*

$$\lim_{n\to\infty} \frac{1}{n} \log \beta^*(n, \varepsilon) = -S(\rho_0 \| \rho_1)$$

holds for every $0 < \varepsilon < 1$.

Bibliographic Notes

A book about entropy in several contexts is Greven *et al.* (2003). The role of von Neumann in the mathematization of the quantum entropy is described in Petz *et al.* (2001); this work contains also his gedanken experiment.

The first proof of the strong subadditivity appeared in Lieb and Ruskai (1973) and a didactical elementary approach is in Nielsen and Petz (2005). The structure of the case of equality was obtained in Hayden *et al.* (2004). The quantum relative entropy was introduced by Umegaki in 1962 and their properties and its axiomatization are contained in the monograph by Ohya and Petz (1993). The monotonicity of the relative entropy is called Uhlmann's theorem, see Uhlmann (1977) and Ohya and Petz (1993).

The rigorous and comprehensive treatment of quantum lattice systems was one of the early successes of the algebraic approach to quantum statistical thermodynamics. The subject is well summarized in Bratteli and Robinson (1981). The book by Sewell (1986) contains more physics and has less in mathematical technicalities. For many interesting entropy results concerning mean field systems, see, for example, Ragio and Werner (1991).

The high probability subspace theorem is due to Ohya and Petz – Petz (1992) and Ohya and Petz (1993) for product states – and was extended to

some algebraic and Gibbs states by Hiai and Petz. The application to data compression was first observed by Schumacher (1995). The chained property of the high-probability subspaces was studied in Bjelaković *et al.* (2003) and the universality is from Kaltchenko and Yang (2003).

A weak form of the quantum Stein lemma was proved in Hiai and Petz (1991) and the stated form is due to Nagaoka and Ogawa (2000). An extension to the case where $\rho_0^{(n)}$ is not a product was given in Bjelaković and Siegmund-Schultze (2004).

Other surveys about quantum entropy are Petz (1992) and Schumacher and Westmoreland (2002).

See also: Asymptotic Structure and Conformal Infinity; Capacities Enhanced by Entanglement; Channels in Quantum Information Theory; Entropy and Qualitative Transversality; Positive Maps on C^*-Algebras; von Neumann Algebras: Introduction, Modular Theory and Classification Theory; von Neumann Algebras: Subfactor Theory.

Further Reading

Bjelaković I, Krüger T, Siegmund-Schultze R, and Szkoła A (2003) Chained typical subspaces – a quantum version of Breiman's therem. *Preprint*.

Bjelaković I and Siegmund-Schultze R (2004) An ergodic theorem for the quantum relative entropy. *Communications in Mathematical Physics* 247: 697–712.

Bratteli O and Robinson DW (1981) *Operator Algebras and Quantum Statistical Mechanics. 2. Equilibrium States. Models in Quantum Statistical Mechanics*, Texts and Monographs in Physics, (2nd edn., 1997). Berlin: Springer.

Greven A, Keller G, and Warnecke G (2003) *Entropy*. Princeton and Oxford: Princeton University Press.

Hayden P, Jozsa R, Petz D, and Winter A (2004) Structure of states which satisfy strong subadditivity of quantum entropy with equality. *Communications in Mathematical Physics* 246: 359–374.

Hiai F and Petz D (1991) The proper formula for relative entropy and its asymptotics in quantum probability. *Communications in Mathematical Physics* 143: 99–114.

Kaltchenko A and Yang E-H (2003) Universal compression of ergodic quantum sources. *Quantum Information and Computation* 3: 359–375.

Lieb EH and Ruskai MB (1973) Proof of the strong subadditivity of quantum mechanical entropy. *Journal of Mathematical Physics* 14: 1938–1941.

Nielsen MA and Petz D (2005) A simple proof of the strong subadditivity inequality. *Quantum Information and Computation* 5: 507–513.

Ogawa T and Nagaoka H (2000) Strong converse and Stein's lemma in quantum hypothesis testing. *IEEE Transactions on Information Theory* 46: 2428–2433.

Ohya M and Petz D (1993) *Quantum Entropy and Its Use*, Texts and Monographs in Physics, (2nd edn., 2004). Berlin: Springer.

Petz D (1992) Entropy in quantum probability. In: Accardi L (ed.) *Quantum Probability and Related Topics VII*, pp. 275–297. Singapore: World Scientific.

Petz D (2001) Entropy, von Neumann and the von Neumann entropy. In: Rédei M and Stöltzner M (eds.) *John von Neumann and the Foundations of Quantum Physics*. Dordrecht: Kluwer.

Raggio GA and Werner RF (1991) The Gibbs variational principle for inhomogeneous mean field systems. *Helvetica Physica Acta* 64: 633–667.

Schumacher B (1995) Quantum coding. *Physical Review A* 51: 2738–2747.

Schumacher B and Westmoreland MD (2002) Relative entropy in quantum information theory. In: *Quantum Computation and Information*, (Washington, DC, 2000), Contemp. Math. vol. 305, pp. 265–289. Providence, RI: American Mathematical Society.

Sewell GL (1986) *Quantum Theory of Collective Phenomena*, New York: Clarendon.

Uhlmann A (1977) Relative entropy and the Wigner–Yanase–Dyson–Lieb concavity in an interpolation theory. *Communications in Mathematical Physics* 54: 21–32.

von Neumann J (1932) *Mathematische Grundlagen der Quantenmechanik*. Berlin: Springer. (In English: von Neumann J, *Mathematical Foundations of Quantum Mechanics*. Princeton: Princeton University Press.)

Quantum Ergodicity and Mixing of Eigenfunctions

S Zelditch, Johns Hopkins University, Baltimore, MD, USA

Quantum ergodicity and mixing belong to the field of quantum chaos, which studies quantizations of "chaotic" classical Hamiltonian systems. The basic question is: how does the chaos of the classical dynamics impact on the eigenvalues/eigenfunctions of the quantum Hamiltonian $\hat{H}$ and on long-time dynamics generated by $\hat{H}$?

These problems lie at the foundations of the semiclassical limit, that is, the limit as the Planck constant $\hbar \to 0$ or the energy $E \to \infty$. More generally, one could ask what impact any dynamical feature of a classical mechanical system (e.g., complete integrability, KAM, and ergodicity) has on the eigenfunctions and eigenvalues of the quantization.

Over the last 30 years or so, these questions have been studied rather systematically by both mathematicians and physicists. There is an extensive literature comparing classical and quantum dynamics of model systems, such as comparing the geodesic flow and wave group on a compact (or finite-volume) hyperbolic surface, or comparing classical and quantum billiards on the Sinai billiard or the Bunimovich stadium, or comparing the

discrete dynamical system generated by a hyperbolic torus automorphism and its quantization by the metaplectic representation. As these models indicate, the basic problems and phenomena are richly embodied in simple, low-dimensional examples in much the same way that two-dimensional toy statistical mechanical models already illustrate complex problems on phase transitions. The principles established for simple models should apply to far more complex systems such as atoms and molecules in strong magnetic fields.

The conjectural picture which has emerged from many computer experiments and heuristic arguments on these simple model systems is roughly that there exists a length scale in which quantum chaotic systems exhibit universal behavior. At this length scale, the eigenvalues resemble eigenvalues of random matrices of large size and the eigenfunctions resemble random waves. A small sample of the original physics articles suggesting this picture is Berry (1977), Bohigas *et al.* (1984), Feingold and Peres (1986), and Heller (1984).

This article reviews some of the rigorous mathematical results in quantum chaos, particularly those on eigenfunctions of quantizations of classically ergodic or mixing systems. They support the conjectural picture of random waves up to two moments, that is, on the level of means and variances. A few results also exist on higher moments in very special cases. But from the mathematical point of view, the conjectural links to random matrices or random waves remain very much open at this time. A key difficulty is that the length scale on which universal behavior should occur is far below the resolving power of any known mathematical techniques, even in the simplest model problems. The main evidence for the random matrix and random wave connections comes from numerous computer experiments of model cases in the physics literature. We will not review numerical results here, but to get a well-rounded view of the field, it is important to understand the computer experiments (see, e.g., Bäcker *et al.* (1998a, b) and Barnett (2005)).

The model quantum systems that have been most intensively studied in mathematical quantum chaos are Laplacians or Schrödinger operators on compact (or finite-volume) Riemannian manifolds, with or without boundary, and quantizations of symplectic maps on compact Kähler manifolds. Similar techniques and results apply in both settings, so for the sake of coherence we concentrate on the Laplacian on a compact Riemannian manifold with "chaotic" geodesic flow and only briefly allude to the setting of "quantum maps." Additionally, two main kinds of methods are in use: (1) methods of semiclassical (or microlocal) analysis, which apply to general Laplacians (and more general Schrödinger operators), and (2) methods of number theory and automorphic forms, which apply to arithmetic models such as arithmetic hyperbolic manifolds or quantum cat maps. Arithmetic models are far more "explicitly solvable" than general chaotic systems, and the results obtained for them are far sharper than the results of semiclassical analysis. This article is primarily devoted to the general results on Laplacians obtained by semiclassical analysis; *see* Arithmetic Quantum Chaos for results by J Marklov. For background on semiclassical analysis, see Heller (1984).

Wave Group and Geodesic Flow

The model quantum Hamiltonians we will discuss are Laplacians Δ on compact Riemannian manifolds (M, g) (with or without boundary). The classical phase space in this setting is the cotangent bundle T^*M of M, equipped with its canonical symplectic form $\sum_i \mathrm{d}x_i \wedge \mathrm{d}\xi_i$. The metric defines the Hamiltonian

$$H(x,\xi) = |\xi|_g = \sqrt{\sum_{ij=1}^{n} g^{ij}(x)\xi_i\xi_j}$$

on T^*M, where

$$g_{ij} = g\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right)$$

$[g^{ij}]$ is the inverse matrix to $[g_{ij}]$. We denote the volume density of (M, g) by dVol and the corresponding inner product on $L^2(M)$ by $\langle f, g\rangle$. The unit (co-) ball bundle is denoted $B^*M = \{(x,\xi) : |\xi| \leq 1\}$.

The Hamiltonian flow Φ^t of H is the geodesic flow. By definition, $\Phi^t(x,\xi) = (x_t, \xi_t)$, where (x_t, ξ_t) is the terminal tangent vector at time t of the unit speed geodesic starting at x in the direction ξ. Here and below, we often identify T^*M with the tangent bundle TM using the metric to simplify the geometric description. The geodesic flow preserves the energy surfaces $\{H = E\}$ which are the co-sphere bundles $S^*_E M$. Due to the homogeneity of H, the flow on any energy surface $\{H = E\}$ is equivalent to that on the co-sphere bundle $S^*M = \{H = 1\}$. (This homogeneity could be broken by adding a potential $V \in C^\infty(M)$ to form a semiclassical Schrödinger operator $-\hbar^2\Delta + V$, whose underlying Hamiltonian flow is generated by $|\xi|_g^2 + V(x)$.) *See* h-Pseudodifferential Operators and Applications.

The quantization of the Hamiltonian H is the square root $\sqrt{\Delta}$ of the positive Laplacian

$$\Delta = -\frac{1}{\sqrt{g}}\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}g^{ij}g\frac{\partial}{\partial x_j}$$

of (M, g). Here, $g = \det[g_{ij}]$. We choose to work with $\sqrt{\Delta}$ rather than Δ since the former generates the wave

$$U_t = \mathrm{e}^{\mathrm{i}t\sqrt{\Delta}}$$

which is the quantization of the geodesic flow Φ^t.

By the last statement we mean that U_t is related to Φ^t in several essentially equivalent ways:

1. singularities of waves, that is, solutions $U_t\psi$ of the wave equation, propagate along geodesics;
2. U_t is a Fourier integral operator (= quantum map) associated to the canonical relation defined by the graph of Φ^t in $T^*M \times T^*M$; and
3. Egorov's theorem holds.

We only define the latter since it plays an important role in studying eigenfunctions. As with any quantum theory, there is an algebra of observables on the Hilbert space $L^2(M, \mathrm{dvol}_g)$ which quantizes T^*M. Here, dvol_g is the volume form of the metric. The algebra is that $\Psi^*(M)$ of pseudodifferential operators ψDO's of all orders, though we often restrict to the subalgebra Ψ^0 of ψ DO's of order zero. We denote by $\Psi^m(M)$ the subspace of pseudodifferential operators of order m. The algebra is defined by constructing a quantization Op from an algebra of symbols $a \in S^m(T^*M)$ of order m (polyhomogeneous functions on $T^*M\backslash 0$) to Ψ^m. The map Op is not unique. In the reverse direction is the symbol map $\sigma_A: \Psi^m \to S^m(T^*M)$ which takes an operator $\mathrm{Op}(a)$ to the homogeneous term a_m of order m in a.

Egorov's theorem for the wave group concerns the conjugations

$$\alpha_t(A) := U_t A U_t^*, \quad A \in \Psi^m(M) \qquad [1]$$

Such a conjugation defines the quantum evolution of observables in the Heisenberg picture, and, since the early days of quantum mechanics, it was known to correspond to the classical evolution

$$V_t(a) := a \circ \Phi^t \qquad [2]$$

of observables $a \in C^\infty(S^*M)$. Egorov's theorem is the rigorous version of this correspondence: it states that α_t defines an order-preserving automorphism of $\Psi^*(M)$, that is, $\alpha_t(A) \in \Psi^m(M)$ if $A \in \Psi^m(M)$, and that

$$\sigma_{U_t A U_t^*}(x,\xi) = \sigma_A(\Phi^t(x,\xi)) := V_t(\sigma_A), \quad (x,\xi) \in T^*M\backslash 0 \qquad [3]$$

This formula is almost universally taken to be the definition of quantization of a flow or map in the physics literature.

The key difficulty in quantum chaos is that it involves a comparison between long-time dynamical properties of Φ^t and U_t through the symbol map and similar classical limits. The classical dynamics defines the "principal symbol" behavior of U_t and the "error" $U_t A U_t^* - Op(\sigma_A \circ \Phi^t)$ typically grows exponentially in time. This is just the first example of a ubiquitous "exponential barrier" in the subject.

Eigenvalues and Eigenfunctions of Δ

The eigenvalue problem on a compact Riemannian manifold

$$\Delta\varphi_j = \lambda_j^2\varphi_j, \quad \langle\varphi_j, \varphi_k\rangle = \delta_{jk}$$

is dual under the Fourier transform to the wave equation. Here, $\{\varphi_j\}$ is a choice of orthonormal basis of eigenfunctions, which is not unique if the eigenvalues have multiplicities >1. The individual eigenfunctions are difficult to study directly, and so one generally forms the spectral projections kernel,

$$E(\lambda, x, y) = \sum_{j:\lambda_j \le \lambda} \varphi_j(x)\varphi_j(y) \qquad [4]$$

Semiclassical asymptotics is the study of the $\lambda \to \infty$ limit of the spectral data $\{\varphi_j, \lambda_j\}$ or of $E(\lambda, x, y)$. The (Schwartz) kernel of the wave group can be represented in terms of the spectral data by

$$U_t(x,y) = \sum_j \mathrm{e}^{\mathrm{i}t\lambda_j}\varphi_j(x)\varphi_j(y)$$

or equivalently as the Fourier transform $\int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}t\lambda}\,\mathrm{d}E(\lambda, x, y)$ of the spectral projections. Hence, spectral asymptotics is often studied through the large-time behavior of the wave group.

The link between spectral theory and geometry, and the source of Egorov's theorem for the wave group, is the construction of a parametrix (or WKB formula) for the wave kernel. For small times t, the simplest is the Hadamard parametrix,

$$U_t(x,y) \sim \int_0^\infty \mathrm{e}^{\mathrm{i}\theta(r^2(x,y)-t^2)}\sum_{k=0}^\infty U_k(x,y)\theta^{((d-3)/2)-k}\,\mathrm{d}\theta \quad (t < \mathrm{inj}(M,g)) \qquad [5]$$

where $r(x,y)$ is the distance between points, $U_0(x,y) = \Theta^{-1/2}(x,y)$ is the volume 1/2-density, $\mathrm{inj}(M,g)$ is the injectivity radius, and the higher Hadamard coefficients are obtained by solving transport equations along geodesics. The parametrix is asymptotic to the wave kernel in the sense of

smoothness, that is, the difference of the two sides of [5] is smooth. The relation [5] may be iterated using $U_{tm} = U_t^m$ to obtain a parametrix for long times. This is obviously complicated and not necessarily the best long-time parametrix construction, but it illustrates again the difficulty of a long-time analysis.

Weyl Law and Local Weyl Law

A fundamental and classical result in spectral asymptotics is Weyl's law on counting eigenvalues:

$$\begin{aligned} N(\lambda) &= \#\{j : \lambda_j \le \lambda\} \\ &= \frac{|B_n|}{(2\pi)^n}\mathrm{Vol}(M,g)\lambda^n + O(\lambda^{n-1}) \end{aligned} \qquad [6]$$

Here, $|B_n|$ is the Euclidean volume of the unit ball and $\mathrm{Vol}(M,g)$ is the volume of M with respect to the metric g. An equivalent formula which emphasizes the correspondence between classical and quantum mechanics is

$$\mathrm{tr}E_\lambda = \frac{\mathrm{Vol}(|\xi|_g \le \lambda)}{(2\pi)^n} \qquad [7]$$

where Vol is the symplectic volume measure relative to the natural symplectic form $\sum_{j=1}^n \mathrm{d}x_j \wedge \mathrm{d}\xi_j$ on T^*M. Thus, the dimension of the space where $H = \sqrt{\Delta}$ is $\le \lambda$ is asymptotically the volume where its symbol $|\xi|_g \le \lambda$.

The remainder term in Weyl's law is sharp on the standard sphere, where all geodesics are periodic, but is not sharp on (M,g) for which the set of periodic geodesics has measure zero (Duistermaat–Guillemin, Ivrii) (*see* Semiclassical Spectra and Closed Orbits). When the set of periodic geodesics has measure zero (as is the case for ergodic systems), one has

$$\begin{aligned} N(\lambda) &= \#\{j : \lambda_j \le \lambda\} \\ &= \frac{|B_n|}{(2\pi)^n}\mathrm{Vol}(M,g)\lambda^n + o(\lambda^{n-1}) \end{aligned} \qquad [8]$$

The remainder is then of smaller order than the derivative of the principal term, and one then has asymptotics in shorter intervals:

$$\begin{aligned} N([\lambda,\lambda+1]) &= \#\{j : \lambda_j \in [\lambda,\lambda+1]\} \\ &= n\frac{|B_n|}{(2\pi)^n}\mathrm{Vol}(M,g)\lambda^{n-1} + o(\lambda^{n-1}) \end{aligned} \qquad [9]$$

Physicists tend to write $\lambda \sim h^{-1}$ and to average over intervals of this width. Then mean spacing between the eigenvalues in this interval is $\sim C_n\mathrm{Vol}(M,g)^{-1}\times \lambda^{-(n-1)}$, where C_n is a constant depending on the dimension.

An important generalization is the "local Weyl law" concerning the traces $\mathrm{tr}AE(\lambda)$, where $A \in \Psi^m(M)$. It asserts that

$$\begin{aligned} &\sum_{\lambda_j \le \lambda}\langle A\varphi_j, \varphi_j\rangle \\ &\quad = \frac{1}{(2\pi)^n}\int_{B^*M}\sigma_A \mathrm{d}x\,\mathrm{d}\xi\ \lambda^n + O(\lambda^{n-1}) \end{aligned} \qquad [10]$$

There is also a pointwise local Weyl law:

$$\sum_{\lambda_j \le \lambda}|\varphi_j(x)|^2 = \frac{1}{(2\pi)^n}|B^n|\lambda^n + R(\lambda,x) \qquad [11]$$

where $R(\lambda,x) = O(\lambda^{n-1})$ uniformly in x. Again, when the periodic geodesics form a set of measure zero in S^*M, one could average over the shorter interval $[\lambda,\lambda+1]$. Combining the Weyl and local Weyl law, we find the surface average of σ_A is a limit of traces:

$$\begin{aligned} \omega(A) &:= \frac{1}{\mu(S^*M)}\int_{S^*M}\sigma_A\,\mathrm{d}\mu \\ &= \lim_{\lambda\to\infty}\frac{1}{N(\lambda)}\sum_{\lambda_j\le\lambda}\langle A\varphi_j,\varphi_j\rangle \end{aligned} \qquad [12]$$

Here, μ is the "Liouville measure" on S^*M, that is, the surface measure $\mathrm{d}\mu = \mathrm{d}x\,\mathrm{d}\xi/\mathrm{d}H$ induced by the Hamiltonian $H = |\xi|_g$ and by the symplectic volume measure $\mathrm{d}x\,\mathrm{d}\xi$ on T^*M.

Problems on Asymptotics Eigenfunctions

Eigenfunctions arise in quantum mechanics as stationary states, that is, states ψ for which the probability measure $|\psi(t,x)|^2\mathrm{dvol}$ is constant in time where $\psi(t,x) = U_t\psi(x)$ is the evolving state. This follows from the fact that

$$U_t\varphi_k = \mathrm{e}^{\mathrm{i}t\lambda_k}\varphi_k \qquad [13]$$

and that $|\mathrm{e}^{\mathrm{i}t\lambda_k}| = 1$. They are the basic modes of the quantum system. One would like to know the behavior as $\lambda_j \to \infty$ (or $\hbar \to 0$ in the semiclassical setting) of invariants such as:

1. matrix elements $\langle A\varphi_j,\varphi_j\rangle$ of observables in this state;
2. transition elements $\langle A\varphi_i,\varphi_j\rangle$ between states;
3. size properties as measured by L^p norms $\|\varphi_j\|_{L^p}$;
4. value distribution as measured by the distribution function $\mathrm{Vol}\{x \in M : |\varphi_j(x)|^2 > t\}$; and
5. shape properties, for example, distribution of zeros and critical points of φ_j.

Let us introduce some problems which have motivated much of the work in this area.

Problem 1 Let $\mathcal{Q}$ denote the set of "quantum limits," that is, weak* limit points of the sequence $\{\Phi_k\}$ of distributions on the classical phase space S^*M, defined by

$$\int_X a \, d\Phi_k := \langle Op(a)\varphi_k, \varphi_k \rangle$$

where $a \in C^\infty(S^*M)$.

The set $\mathcal{Q}$ is independent of the definition of Op. It follows almost immediately from Egorov's theorem that $\mathcal{Q} \subset \mathcal{M}_I$, where $\mathcal{M}_I$ is the convex set of invariant probability measures for the geodesic flow. Furthermore, they are time-reversal invariant, that is, invariant under $(x, \xi) \to (x, -\xi)$ since the eigenfunctions are real valued.

To see this, it is helpful to introduce the linear functionals on Ψ^0:

$$\rho_k(A) = \langle Op(a)\varphi_k, \varphi_k \rangle \qquad [14]$$

We observe that $\rho_k(I) = 1$, $\rho_k(A) \geq 0$ if $A \geq 0$, and that

$$\rho_k\left(U_t A U_t^*\right) = \rho_k(A) \qquad [15]$$

Indeed, if $A \geq 0$ then $A = B^*B$ for some $B \in \Psi^0$ and we can move B^* to the right-hand side. Similarly, [15] is proved by moving U_t to the right-hand side and using [13]. These properties mean that ρ_j is an "invariant state" on the algebra Ψ^0. More precisely, one should take the closure of Ψ^0 in the operator norm. An invariant state is the analog in quantum statistical mechanics of an invariant probability measure.

The next important fact about the states ρ_k is that any weak limit of the sequence $\{\rho_k\}$ on Ψ^0 is an invariant probability measure on $C(S^*M)$, that is, a positive linear functional on $C(S^*M)$ rather than just a state on Ψ^0. This follows from the fact that $\langle K\varphi_j, \varphi_j \rangle \to 0$ for any compact operator K, and so any limit of $\langle A\varphi_k, \varphi_k \rangle$ is equally a limit of $\langle (A + K)\varphi_k, \varphi_k \rangle$. Hence, any limit is bounded by $\inf_K \|A + K\|$ (the infimum taken over compact operators), and for any $A \in \Psi^0$, $\|\sigma_A\|_{L^\infty} = \inf_K \|A + K\|$. Hence, any weak limit is bounded by a constant times $\|\sigma_A\|_{L^\infty}$ and is therefore continuous on $C(S^*M)$. It is a positive functional since each ρ_j, and hence any limit, is a probability measure. By Egorov's theorem and the invariance of the ρ_k, any limit of $\rho_k(A)$ is a limit of $\rho_k(Op(\sigma_A \circ \Phi^t))$ and hence the limit measure is invariant.

Problem 1 is thus to identify which invariant measures in $\mathcal{M}_I$ show up as weak limits of the functionals ρ_k or equivalently the distributions $d\Phi_k$. The weak limits reflect the concentration and oscillation properties of eigenfunctions. Here are some possibilities:

1. Normalized Liouville measure. In fact, the functional ω of [12] is also a state on Ψ^0 for the reason explained above. A subsequence $\{\varphi_{k_j}\}$ of eigenfunctions is considered diffuse if $\rho_{k_j} \to \omega$.
2. A periodic orbit measure μ_γ defined by

$$\mu_\gamma(A) = \frac{1}{L_\gamma} \int_\gamma \sigma_A \, ds$$

   where L_γ is the length of γ. A sequence of eigenfunctions for which $\rho_{k_j} \to \mu_\gamma$ obviously concentrates (or strongly "scars") on the closed geodesic.
3. A finite sum of periodic orbit measures.
4. A delta-function along an invariant Lagrangian manifold $\Lambda \subset S^*M$. The associated eigenfunctions are viewed as "localizing" along Λ.
5. A more general invariant measure which is singular with respect to $d\mu$.

All of these possibilities can and do happen in different examples. If $d\Phi_{k_j} \to \omega$, then in particular we have

$$\frac{1}{\mathrm{Vol}(M)} \int_E |\varphi_{k_j}(x)|^2 d\mathrm{Vol} \to \frac{\mathrm{Vol}(E)}{\mathrm{Vol}(M)}$$

for any measurable set E whose boundary has measure zero. Interpreting $|\varphi_{k_j}(x)|^2 d\mathrm{Vol}$ as the probability density of finding a particle of energy λ_k^2 at x, this result means that the sequence of probabilities tends to uniform measure.

However, $d\Phi_{k_j} \to \omega$ is much stronger since it says that the eigenfunctions become diffuse on the energy surface S^*M and not just on the configuration space M. As an example, consider the flat torus $\mathbb{R}^n/\mathbb{Z}^n$. An orthonormal basis of eigenfunctions is furnished by the standard exponentials $e^{2\pi i\langle k, x\rangle}$ with $k \in \mathbb{Z}^n$. Obviously, $|e^{2\pi i\langle k, x\rangle}|^2 = 1$, so the eigenfunctions are already diffuse in configuration space. On the other hand, they are far from diffuse in phase space, and localize on invariant Lagrangian tori in S^*M. Indeed, by definition of pseudodifferential operator, $Ae^{2\pi i\langle k, x\rangle} = a(x, k)\, e^{2\pi i\langle k, x\rangle}$, where $a(x, k)$ is the complete symbol. Thus,

$$\begin{aligned}\langle Ae^{2\pi i\langle k, x\rangle}, e^{2\pi i\langle k, x\rangle}\rangle &= \int_{\mathbb{R}^n/\mathbb{Z}^n} a(x, k)\, dx \\ &\sim \int_{\mathbb{R}^n/\mathbb{Z}^n} \sigma_A\left(x, \frac{k}{|k|}\right) dx\end{aligned}$$

A subsequence $e^{2\pi i\langle k_j, x\rangle}$ of eigenfunctions has a weak limit if and only if $k_j/|k_j|$ tends to a limit vector ξ_0 in the unit sphere in $\mathbb{R}^n$. In this case, the associated

weak* limit is $\int_{\mathbb{R}^n/\mathbb{Z}^n} \sigma_A(x,\xi_0)\mathrm{d}x$, that is, the delta-function on the invariant torus $T_{\xi_0} \subset S^*M$ defined by the constant momentum condition $\xi = \xi_0$. The eigenfunctions are said to localize on this invariant torus for Φ^t.

The flat torus is a model of a completely integrable system on both the classical and quantum levels. Another example is that of the standard round sphere S^n. In this case, the author and D Jakobson showed that absolutely any invariant measure $\nu \in \mathcal{M}_I$ can arise as a weak limit of a sequence of eigenfunctions. This reflects the huge degeneracy (multiplicities) of the eigenvalues.

On the other hand, if the geodesic flow is ergodic, one would expect the eigenfunctions to be diffuse in phase space. In the next section, we will discuss the rigorous results on this problem.

Off-diagonal matrix elements

$$\rho_{jk}(A) = \langle A\varphi_i, \varphi_j\rangle \qquad [16]$$

are also important as transition amplitudes between states. They no longer define states since $\rho_{jk}(I) = 0$, are positive, or invariant. Indeed, $\rho_{jk}(U_t A U_t^*) = \mathrm{e}^{\mathrm{i}t(\lambda_j-\lambda_k)}\rho_{jk}(A)$, so they are eigenvectors of the automorphism α_t of [1]. A sequence of such matrix elements cannot have a weak limit unless the spectral gap $\lambda_j - \lambda_k$ tends to a limit $\tau \in \mathbb{R}$. In this case, by the same discussion as above, any weak limit of the functionals ρ_{jk} will be an eigenmeasure of the geodesic flow which transforms by $\mathrm{e}^{\mathrm{i}\tau t}$ under the action of Φ^t. Examples of such eigenmeasures are orbital Fourier coefficients

$$\frac{1}{L_\gamma}\int_0^{L_\gamma} \mathrm{e}^{-\mathrm{i}\tau t}\sigma_A(\Phi^t(x,\xi))\,\mathrm{d}t$$

along a periodic orbit. Here, $\tau \in (2\pi/L_\gamma)\mathbb{Z}$. We denote by $\mathcal{Q}_\tau$ such eigenmeasures of the geodesic flow. Problem 1 has the following extension to off-diagonal elements:

Problem 2 Determine the set $\mathcal{Q}_\tau$ of "quantum limits," that is, weak* limit points of the sequence $\{\Phi_{kj}\}$ of distributions on the classical phase space S^*M, defined by

$$\int_X a\mathrm{d}\Phi_{kj} := \langle \mathrm{O}p(a)\varphi_k, \varphi_j\rangle$$

where $\lambda_j - \lambda_k = \tau + o(1)$ and where $a \in C^\infty(S^*M)$, or equivalently of the functionals ρ_{jk}.

As will be discussed in the section "Quantum weak mixing," the asymptotics of off-diagonal elements depends on the weak mixing properties of the geodesic flow and not just its ergodicity.

Matrix elements of eigenfunctions are quadratic forms. More "nonlinear" problems involve the L^p-norms or the distribution functions of eigenfunctions. Estimates of the L^∞-norms can be obtained from the local Weyl law [10]. Since the jump in the left-hand side at λ is $\sum_{j:\lambda_j=\lambda}|\varphi_j(x)|^2$ and the jump in the right-hand side is the jump of $R(\lambda,x)$, this implies

$$\sum_{j:\lambda_j=\lambda}|\varphi_j(x)|^2 = O(\lambda^{n-1}) \Longrightarrow \|\varphi_j\|_{L^\infty} = O(\lambda^{\frac{n-1}{2}}) \qquad [17]$$

For general L^p-norms, the following bounds were proved by C Sogge for any compact Riemannian manifold:

$$\frac{\|\varphi_j\|_p}{\|\varphi\|_2} = O(\lambda^{\delta(p)}), \quad 2 \le p \le \infty \qquad [18]$$

where

$$\delta(p) = \begin{cases} n\left(\frac{1}{2}-\frac{1}{p}\right)-\frac{1}{2}, & \frac{2(n+1)}{n-1} \le p \le \infty \\ \frac{n-1}{2}\left(\frac{1}{2}-\frac{1}{p}\right), & 2 \le p \le \frac{2(n+1)}{n-1} \end{cases} \qquad [19]$$

These estimates are sharp on the unit sphere $S^n \subset \mathbb{R}^{n+1}$. The extremal eigenfunctions are the zonal spherical harmonics, which are the L^2-normalized spectral projection kernels $\Pi_N(x,x_0)/\|\Pi_N(\,\cdot\,,x_0)\|$ centered at any x_0. However, they are not sharp for generic (M,g), and it is natural to ask how "chaotic dynamics" might influence L^p-norms.

Problem 3 Improve the estimates $\|\varphi_j\|_p/\|\varphi\|_2 = O(\lambda^{\delta(p)})$ for (M,g) with ergodic or mixing geodesic flow.

C Sogge and the author have proved that if a sequence of eigenfunctions attains the bounds in [17], then there must exist a point x_0 so that a positive measure of geodesics starting at x_0 in $S^*_{x_0}M$ returns to x_0 at a fixed time T. In the real analytic case, all return so x_0 is a perfect recurrent point. In dimension 2, such a perfect recurrent point cannot occur if the geodesic flow is ergodic; hence $\|\varphi_j\|_{L^\infty} = o(\lambda^{(n-1)/2})$ on any real analytic surface with ergodic geodesic flow. This shows that none of the L^p-estimates above the critical index are sharp for real analytic surfaces with ergodic geodesic flow, and the problem is the extent to which they can be improved.

The random wave model (see the section "Random waves and orthonormal bases") predicts that eigenfunctions of Riemannian manifolds with chaotic geodesic flow should have the bounds $\|\varphi_\lambda\|_{L^p} = O(1)$ for $p < \infty$ and that $\|\varphi_\lambda\|_{L^\infty} < \sqrt{\log\lambda}$. But there are

no rigorous estimates at this time close to such predictions. The best general estimate to date on negatively curved compact manifolds (which are models of chaotic geodesic flow) is just the logarithmic improvement

$$||\varphi_j||_{L^\infty} = O\left(\frac{\lambda^{n-1}}{\log \lambda}\right)$$

on the standard remainder term in the local Weyl law. This was known for compact hyperbolic manifolds from the Selberg trace formula, and similar estimates hold manifolds without conjugate points (P Bérard). The exponential growth of the geodesic flow again causes a barrier in improving the estimate beyond the logarithm. In the analogous setting of quantum "cat maps," which are models of chaotic classical dynamics, there exist arbitrarily large eigenvalues with multiplicities of the order $O(\lambda^{n-1}/\log \lambda)$; the L^∞-norm of the L^2-normalized projection kernel onto an eigenspace of this multiplicity is of order of the square root of the multiplicity (Faure *et al.* 2003). This raises doubt that the logarithmic estimate can be improved by general dynamical arguments. Further discussion of L^∞-norms, as well as zeros, will be given at the end of the next section for ergodic systems.

Quantum Ergodicity

In this section, we discuss results on the problems stated above when the geodesic flow of (M, g) is assumed to be ergodic. Let us recall that this means that Liouville measure is an ergodic measure for Φ^t. This is a spectral property of the operator V_t of [2] on $L^2(S^*M, \mathrm{d}\mu)$, namely that V_t has 1 as an eigenvalue of multiplicity 1. That is, the only invariant L^2-functions (with respect to Liouville measure) are the constant functions. This implies that the only invariant sets have Liouville measure 0 or 1 and (Birkhoff's ergodic theorem) that time averages of functions are constant almost everywhere (equal to the space average).

In this case, there is a general result which originated in the work of Schnirelman and was developed into the following theorem by Zelditch, Colin de Verdière, and Sunada (manifolds without boundary), and Gérard–Leichtnam and Zelditch–Zworski (manifolds with boundary). The following discussion is based on the articles (Zelditch 1996b, c, Zelditch and Zworski 1996), which contain further references to the literature.

Theorem 1 *Let (M, g) be a compact Riemannian manifold (possibly with boundary), and let $\{\lambda_j, \varphi_j\}$ be the spectral data of its Laplacian Δ. Then the geodesic flow G^t is ergodic on $(S^*M, \mathrm{d}\mu)$ if and only if, for every $A \in \Psi^o(M)$, we have*:

(i) $\lim_{\lambda\to\infty} \frac{1}{N(\lambda)} \sum_{\lambda_j \le \lambda} |(A\varphi_j, \varphi_j) - \omega(A)|^2 = 0.$

(ii) $(\forall \epsilon)(\exists \delta) \limsup_{\lambda\to\infty} \frac{1}{N(\lambda)} \sum_{j\neq k\,:\,\lambda_j,\lambda_k \le \lambda, |\lambda_j - \lambda_k| < \delta} |(A\varphi_j, \varphi_k)|^2 < \epsilon.$

This implies that there exists a subsequence $\{\varphi_{j_k}\}$ of eigenfunctions whose indices j_k have counting density 1 for which $\langle A\varphi_{j_k}, \varphi_{j_k}\rangle \to \omega(A)$. We will call the eigenfunctions in such a sequence "ergodic eigenfunctions." One can sharpen the results by averaging over eigenvalues in the shorter interval $[\lambda, \lambda + 1]$ rather than in $[0, \lambda]$.

There is also an ergodicity result for boundary values of eigenfunctions on domains with boundary and with Dirichlet, Neumann, or Robin boundary conditions (Gérard–Leichtnam, Hassell–Zelditch, Burq). This corresponds to the fact that the billiard map on $B^*\partial M$ is ergodic.

The first statement (i) is essentially a convexity result. It remains true if one replaces the square by any convex function φ on the spectrum of A,

$$\frac{1}{N(E)} \sum_{\lambda_j \le E} \varphi(\langle A\varphi_k, \varphi_k\rangle - \omega(A)) \to 0 \qquad [20]$$

Before sketching a proof, we point out a somewhat heuristic "picture proof" of the theorem. Namely, ergodicity of the geodesic flow is equivalent to the statement that Liouville measure is an extreme point of the compact convex set $\mathcal{M}_I$. In fact, it further implies that ω is an extreme point of the compact convex set $\mathcal{E}_{\mathbb{R}}$ of invariant states for α_t of eqn [1]; see Ruelle (1969) for background. But the local Weyl law says that ω is also the limit of the convex combination

$$\frac{1}{N(E)} \sum_{\lambda_j \le E} \rho_j$$

An extreme point cannot be written as a convex combination of other states unless all the states in the combination are equal to it. In our case, ω is only a limit of convex combinations so it need not (and does not) equal each term. However, almost all terms in the sequence must tend to ω, and that is equivalent to [1].

Sketch of Proof of Theorem 1(i) As mentioned above, this is a convexity result and with no additional effort we can consider more general sums of the form. We then have

$$\sum_{\lambda_j \le E} \varphi(\langle A\varphi_k, \varphi_k\rangle - \omega(A)) = \sum_{\lambda_j \le E} \varphi(\langle\langle A\rangle_T - \omega(A)\varphi_k, \varphi_k\rangle) \qquad [21]$$

where

$$\langle A\rangle_T = \frac{1}{2T}\int_{-T}^{T} U_t A U_t^*\, \mathrm{d}t$$

We then apply the Peierls–Bogoliubov inequality

$$\sum_{j=1}^{n} \varphi((B\varphi_j,\varphi_j)) \le \operatorname{tr}\varphi(B)$$

with $B = \Pi_E[\langle A\rangle_T - \omega(A)]\Pi_E$ to get

$$\sum_{\lambda_j\le E} \varphi(\langle\langle A\rangle_T - \omega(A)\varphi_k,\varphi_k\rangle) \le \operatorname{tr}\varphi(\Pi_E[\langle A\rangle_T - \omega(A)]\Pi_E) \qquad [22]$$

Here, Π_E is the spectral projection for $\hat{H}$ corresponding to the interval $[0,E]$. From the Berezin inequality we then have (if $\varphi(0)=0$):

$$\begin{aligned}&\frac{1}{N(E)}\operatorname{tr}\varphi(\Pi_E[\langle A\rangle_T - \omega(A)]\Pi_E)\\ &\quad\le \frac{1}{N(E)}\operatorname{tr}\Pi_E\varphi([\langle A\rangle_T - \omega(A)])\Pi_E\\ &\quad\to \omega_E(\varphi(\langle A\rangle_T - \omega(A))),\ \text{as } E\to\infty\end{aligned}$$

As long as φ is smooth, $\varphi(\langle A\rangle_T - \omega(A))$ is a pseudodifferential operator of order zero with principal symbol $\varphi(\langle\sigma_A\rangle_T - \omega(A))$. By the assumption that $\omega_E\to\omega$ we get

$$\begin{aligned}&\lim_{E\to\infty}\frac{1}{N(E)}\sum_{\lambda_j\le E}\varphi(\langle A\varphi_k,\varphi_k\rangle - \omega(A))\\ &\quad\le \int_{\{H=1\}}\varphi(\langle\sigma_A\rangle_T - \omega(A))\,\mathrm{d}\mu\end{aligned}$$

where

$$\langle\sigma_A\rangle_T = \frac{1}{2T}\int_{-T}^{T}\sigma_A\circ\Phi^t\,\mathrm{d}t$$

As $T\to\infty$ the right-hand side approaches $\varphi(0)=0$ by the dominated convergence theorem and by Birkhoff's ergodic theorem. Since the left-hand side is independent of T, this implies that

$$\lim_{E\to\infty}\frac{1}{N(E)}\sum_{\lambda_j\le E}\varphi(\langle A\varphi_k,\varphi_k\rangle - \omega(A)) = 0$$

for any smooth convex φ on $\mathrm{Spec}(A)$ with $\varphi(0)=0$. □

As mentioned above, the statement of Theorem 1(i) is equivalent to saying that there is a subsequence $\{\varphi_{j_k}\}$ of counting density 1 for which $\rho_{j_k}\to\omega$. The above proof does not and cannot settle the question whether there exist exceptional sparse subsequences of eigenfunctions of density zero tending to other invariant measures. To see this, we observe that the proof is so general that it applies to seemingly very different situations. In place of the distributions $\{\Phi_j\}$ we may consider the set μ_γ of periodic orbit measures for a hyperbolic flow on a compact manifold X. That is,

$$\mu_\gamma(f) = \frac{1}{T_\gamma}\int_\gamma f \quad \text{for } f\in C(X)$$

where γ is a closed orbit and T_γ is its period. According to the Bowen–Margulis equidistribution theorem for closed orbits of hyperbolic flows, we have

$$\frac{1}{\Pi(T)}\sum_{\gamma:T_\gamma\le T}\frac{1}{|\det(I-P_\gamma)|}\mu_\gamma\to\mu$$

where as above μ is the Liouville measure, where P_γ is the linear Poincaré map and where $\Pi(T)$ is the normalizing factor which makes the left side a probability measure, that is, defined by the integral of 1 against the sum. An exact repetition of the previous argument shows that up to a sparse subsequence of γ's, $\mu_\gamma\to\mu$ individually. Yet clearly, the whole sequence does not tend to $\mathrm{d}\mu$: for instance, one could choose the sequence of iterates γ^k of a fixed closed orbit.

Quantum Ergodicity in Terms of Operator Time and Space Averages

The first part of the result above may be reformulated as a relation between operator time and space averages.

Definition Let $A\in\Psi^0$ be an observable and define its time average to be:

$$\langle A\rangle := \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}U_t^* A U_t\,\mathrm{d}t$$

and its space average to be scalar operator

$$\omega(A)\cdot I$$

Here, the limit is taken in the weak operator topology (i.e., one matrix element at a time). To see what is involved, we consider matrix elements with respect to the eigenfunctions. We have

$$\left(\frac{1}{2T}\int_{-T}^{T}U_t^* A U_t\,\mathrm{d}t\varphi_i,\varphi_j\right) = \frac{\sin T(\lambda_i-\lambda_j)}{T(\lambda_i-\lambda_j)}(A\varphi_i,\varphi_j)$$

from which it is clear that the matrix element tends to zero as $T\to\infty$ unless $\lambda_i=\lambda_j$. However, there is no uniformity in the rate at which it goes to zero since the spacing $\lambda_i-\lambda_j$ could be uncontrollably small.

In these terms, Theorem 1(i) states that

$$\langle A\rangle = \omega(A)I + K, \quad \text{where } \lim_{\lambda\to\infty} \omega_\lambda(K^*K) \to 0 \quad [23]$$

where $\omega_\lambda(A) = \operatorname{tr} E(\lambda)A$. Thus, the time average equals the space average plus a term K which is semiclassically small in the sense that its Hilbert–Schmidt norm square $\|E_\lambda K\|^2_{\mathrm{HS}}$ in the span of the eigenfunctions of eigenvalue $\leq\lambda$ is $o(N(\lambda))$.

This is not exactly equivalent to Theorem 1(i) since it is independent of the choice of orthonormal basis, while the previous result depends on the choice of basis. However, when all eigenvalues have multiplicity 1, then the two are equivalent. To see the equivalence, note that $\langle A\rangle$ commutes with $\sqrt{\Delta}$ and hence is diagonal in the basis $\{\varphi_j\}$ of joint eigenfunctions of $\langle A\rangle$ and of U_t. Hence, K is the diagonal matrix with entries $\langle A\varphi_k,\varphi_k\rangle - \omega(A)$. The condition is therefore equivalent to

$$\lim_{E\to\infty} \frac{1}{N(E)} \sum_{\lambda_j\leq E} |\langle A\varphi_k, \varphi_k\rangle - \omega(A)|^2 = 0$$

Since all the terms are positive, no cancellation is possible and this condition is equivalent to the existence of a subset $\mathcal{S}\subset\mathbb{N}$ of density 1 such that $\mathcal{Q}_{\mathcal{S}} := \{\mathrm{d}\Phi_k : k\in\mathcal{S}\}$ has only ω as a weak* limit point. As above, one says that the sequence of eigenfunctions is ergodic.

One could take this restatement of Theorem 1(i) as a semiclassical definition of quantum ergodicity. Two natural questions arise. First:

Problem 4 Suppose the geodesic flow Φ^t of (M,g) is ergodic on S^*M. Is the operator K in

$$\langle A\rangle = \omega(A) + K$$

a compact operator? In this case, $\sqrt{\Delta}$ is said to be quantum uniquely ergodic (QUE). If ergodicity is not sufficient for the QUE property, what extra conditions need to be added?

Compactness would imply that $\langle K\varphi_k,\varphi_k\rangle \to 0$, hence $\langle A\varphi_k,\varphi_k\rangle \to \omega(A)$ along the entire sequence. Quite a lot of attention has been focused on this problem in the last decade. It is probable that ergodicity is not by itself sufficient for the QUE property of general Riemann manifold. For instance, it is believed that there exist modes of asymptotic bouncing ball type which concentrate on the invariant Lagrangian cylinder (with boundary) formed by bouncing ball orbits of the Bunimovich stadium (see e.g., Heller (1984) for more on such "scarring"). Further, Faure *et al.* (2003) have shown that QUE does not hold for the hyperbolic system defined by a quantum cat map on the torus. Since the methods applicable to eigenfunctions of quantum maps and of Laplacians have much in common, this negative result shows that there cannot exist a universal structural proof of QUE.

The principal positive result available at this time is the recent proof by Lindenstrauss of the QUE property for the orthonormal basis of Laplace–Hecke eigenfunctions on arithmetic hyperbolic surfaces. It is generally believed that the spectrum of the Laplace eigenvalues is of multiplicity 1 for such surfaces, so this should imply QUE completely for these surfaces. Earlier partial results on Hecke eigenfunctions are due to Rudnick–Sarnak, Wolpert, and others. For references and further discussion onf Hecke eigenfunctions, see Rudnick and Sarnak (1994) (*see* Arithmetic Quantum Chaos).

So far we have not mentioned Theorem 1(ii). In the next section, we will describe a similar but more general result for mixing systems and the relevance of (ii) will become clear. An interesting open problem is the extent to which (ii) is actually necessary for the equivalence to classical ergodicity.

Problem 5 Converse QE: What can be said of the classical limit of a quantum ergodic system, that is, a system for which $\langle A\rangle = \omega(A) + K$, where K is compact? Is it necessarily ergodic?

Very little is known on this converse problem at present. It is known that if there exists an open set in S^*M filled by periodic orbits, then the Laplacian cannot be quantum ergodic (see Marklof and O'Keefe (2005) for recent results and references). But no proof exists at this time that KAM systems, which have Cantor-like positive measure invariant sets, are not quantum ergodic. It is known that there exists a positive proportion of approximate eigenfunctions (quasimodes) which localize on the invariant tori, but it has not been proved that a positive proportion of actual eigenfunctions has this localization property.

Further Problems and Results on Ergodic Eigenfunctions

Ergodicity is also known to have an impact on the distribution of zeros. The complex zeros in Kähler phase spaces of ergodic eigenfunctions of quantum ergodic maps become uniformly distributed with respect to the Kähler volume form (Nonnenmacher–Voros, Shiffman–Zelditch). An interesting problem is whether the real analog is true:

Problem 6 Ergodicity and equidistribution of nodal sets. Let $\mathcal{N}_{\varphi_j}\subset M$ denote the nodal set (zero set) of φ_j, and equip it with its hypersurface volume form $\mathrm{d}\mathcal{H}^{n-1}$ induced by g. Let (M,g) have ergodic geodesic flow, and suppose that $\{\varphi_j\}$ is an ergodic

sequence of eigenfunctions. Are the following asymptotics valid?

$$\int_{\mathcal{N}_{\varphi_j}} f \, \mathrm{d}\mathcal{H}^{n-1} \sim \lambda_j \frac{1}{\mathrm{Vol}(M,g)} \int_M f \, \mathrm{dVol}$$

This is predicted by the random wave model of the section "Random waves and orthonormal bases." An equidistribution law for the complex zeros is known which gives some evidence for the validity of this limit formula. Let (M,g) be a compact real analytic Riemannian manifold and let $\varphi_j^{\mathbb{C}}$ be the holomorphic extension of the real analytic eigenfunction φ_j to the complexification $M_{\mathbb{C}}$ of M (its Grauert tube). Then, if the geodesic flow is ergodic and if φ_j is an ergodic sequence of eigenfunctions, the normalized current of integration $(1/\lambda_j) Z_{\varphi_j^{\mathbb{C}}}$ over the complex zero set of $\varphi_j^{\mathbb{C}}$ tends weakly to $(\mathrm{i}/\pi)\bar{\partial}\partial|\xi_g|$. This current is singular along the zero section.

Finally, we mention some results on L^∞-norms of eigenfunctions on arithmetic hyperbolic manifolds of dimensions 2 and 3. It was proved by Iwaniec–Sarnak that the joint eigenfunctions of Δ and the Hecke operators on arithmetic hyperbolic surfaces have the upper bound $\|\varphi_j\|_\infty = O_\epsilon(\lambda_j^{5/48+\epsilon})$ for all j and $\epsilon > 0$, and the lower bound $\|\varphi_j\|_\infty \geq c\sqrt{\log\log \lambda_j}$ for some constant $c > 0$ and infinitely many j. Rudnick and Sarnak (1994) proved that there exists an arithmetic hyperbolic manifold and a subsequence φ_{j_k} of eigenfunctions with $\|\varphi_{j_k}\|_{L^\infty} \gg \lambda_{j_k}^{1/4}$, contradicting the random wave model predictions.

Quantum Weak Mixing

There are parallel results on quantizations of weak-mixing geodesic flows which are the subject of this section. First we recall the classical definition: the geodesic flow of (M,g) is weak mixing if the operator V_t has purely continuous spectrum on the orthogonal complement of the constant functions in $L^2(S^*M, \mathrm{d}\mu)$. Hence, like ergodicity, it is a spectral property of the geodesic flow.

We have:

Theorem 2 (Zelditch 1996c). *The geodesic flow Φ^t of (M,g) is weak mixing if and only if the conditions (i) and (ii) of Theorem 1 hold and additionally, for any $A \in \Psi^0(M)$,*

$$(\forall \epsilon)(\exists \delta) \limsup_{\lambda\to\infty} \frac{1}{N(\lambda)} \sum_{\substack{j\neq k:\, \lambda_j,\lambda_k \leq \lambda \\ |\lambda_j - \lambda_k - \tau| < \delta}} |(A\varphi_j, \varphi_k)|^2 < \epsilon$$
$$(\forall \tau \in \mathbb{R})$$

The restriction $j \neq k$ is of course redundant unless $\tau = 0$, in which case the statement coincides with quantum ergodicity. This result follows from the general asymptotic formula, valid for any compact Riemannian manifold (M,g), that

$$\frac{1}{N(\lambda)} \sum_{i\neq j, \lambda_i,\lambda_j \leq \lambda} |\langle A\varphi_i, \varphi_j\rangle|^2 \left|\frac{\sin T(\lambda_i - \lambda_j - \tau)}{T(\lambda_i - \lambda_j - \tau)}\right|^2$$
$$\sim \left\| \frac{1}{2T} \int_{-T}^{T} \mathrm{e}^{\mathrm{i}t\tau} V_t(\sigma_A) \right\|_2^2 - \left|\frac{\sin T\tau}{T\tau}\right|^2 \omega(A)^2 \qquad [24]$$

In the case of weak-mixing geodesic flows, the right-hand side tends to 0 as $T \to \infty$. As with diagonal sums, the sharper result is true where one averages over the short intervals $[\lambda, \lambda + 1]$.

Spectral Measures and Matrix Elements

Theorem 2 is based on expressing the spectral measures of the geodesic flow in terms of matrix elements. The main limit formula is

$$\int_{\tau-\varepsilon}^{\tau+\varepsilon} \mathrm{d}\mu_{\sigma_A} := \lim_{\lambda\to\infty} \frac{1}{N(\lambda)} \sum_{\substack{i,j:\, \lambda_j \leq \lambda, \\ |\lambda_i - \lambda_j - \tau| < \varepsilon}} |\langle A\varphi_i, \varphi_j\rangle|^2 \qquad [25]$$

where $\mathrm{d}\mu_{\sigma_A}$ is the spectral measure for the geodesic flow corresponding to the principal symbol of A, $\sigma_A \in C^\infty(S^*M, \mathrm{d}\mu)$. Recall that the spectral measure of V_t corresponding to $f \in L^2$ is the measure $\mathrm{d}\mu_f$ defined by

$$\langle V_t f, f\rangle_{L^2(S^*M)} = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}t\tau} \, \mathrm{d}\mu_f(\tau)$$

The limit formula [25] is equivalent to the dual formula (under the Fourier transform):

$$\lim_{\lambda\to\infty} \frac{1}{N(\lambda)} \sum_{i,j:\lambda_j \leq \lambda} \mathrm{e}^{\mathrm{i}t(\lambda_i - \lambda_j)} |\langle A\varphi_i, \varphi_j\rangle|^2$$
$$= \langle V_t \sigma_A, \sigma_A\rangle_{L^2(S^*M)} \qquad [26]$$

The proof of [26] is to consider, for $A \in \Psi^\circ$, the operator $A_t^* A \in \Psi^\circ$ with $A_t = U_t^* A U_t$. By the local Weyl law,

$$\lim_{\lambda\to\infty} \frac{1}{N(\lambda)} \mathrm{tr}\, E(\lambda) A_t^* A = \langle V_t \sigma_A, \sigma_A\rangle_{L^2(S^*M)}$$

The right-hand-side of [25] defines a measure $\mathrm{d}m_A$ on $\mathbb{R}$ and [26] says

$$\int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}t\tau} \, \mathrm{d}m_A(\tau) = \langle V_t \sigma_A, \sigma_A\rangle_{L^2(S^*M)} = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}t\tau} \, \mathrm{d}\mu_{\sigma_A}(\tau)$$

Since weak-mixing systems are ergodic, it is not necessary to average in both indices along an ergodic subsequence:

$$\lim_{\lambda_j \to \infty} \langle A_t^* A\varphi_j, \varphi_j \rangle = \sum_j e^{it(\lambda_i - \lambda_j)} |\langle A\varphi_i, \varphi_j \rangle|^2 = \langle V_t \sigma_A, \sigma_A \rangle_{L^2(S^*M)} \quad [27]$$

Dually, one has

$$\lim_{\lambda_j \to \infty} \sum_{i\,:\,|\lambda_i - \lambda_j - \tau| < \varepsilon} |\langle A\varphi_i, \varphi_j \rangle|^2 = \int_{\tau-\varepsilon}^{\tau+\varepsilon} d\mu_{\sigma_A} \quad [28]$$

For QUE systems, these limit formulas are valid for the full sequence of eigenfunctions.

Rate of Quantum Ergodicity and Mixing

A quantitative refinement of quantum ergodicity is to ask at what rate the sums in Theorem 1(i) tend to zero, that is, to establish a rate of quantum ergodicity. More generally, we consider "variances" of matrix elements. For diagonal matrix elements, we define

$$V_A(\lambda) := \frac{1}{N(\lambda)} \sum_{j:\lambda_j \le \lambda} |\langle A\varphi_j, \varphi_j \rangle - \omega(A)|^2 \quad [29]$$

In the off-diagonal case, one may view $|\langle A\varphi_i, \varphi_j \rangle|^2$ as analogous to $|\langle A\varphi_j, \varphi_j \rangle - \omega(A)|^2$. However, the sums in [25] are double sums while those of [29] are single. One may also average over the shorter intervals $[\lambda, \lambda + 1]$.

Quantum Chaos Conjectures

First, consider off-diagonal matrix elements. One conjecture is that it is not necessary to sum in j in [28]: each individual term has the asymptotics consistent with [28]. This is implicitly conjectured by Feingold–Peres (1986) (see [11]) in the form

$$|\langle A\varphi_i, \varphi_j \rangle|^2 \simeq \frac{C_A\left(\frac{E_i - E_j}{\hbar}\right)}{2\pi\rho(E)} \quad [30]$$

where

$$C_A(\tau) = \int_{-\infty}^{\infty} e^{-i\tau t} \langle V_t \sigma_A, \sigma_A \rangle \, dt$$

In our notation, $\lambda_j = \hbar^{-1} E_j$ and $\rho(E)\,dE \sim dN(\lambda)$. There are $\sim C\lambda^{n-1}$ eigenvalues λ_i in the interval $[\lambda_j - \tau - \epsilon, \lambda_j - \tau + \epsilon]$, so [30] states that individual terms have the asymptotics of [28].

On the basis of the analogy between $|\langle A\varphi_i, \varphi_j \rangle|^2$ and $|\langle A\varphi_j, \varphi_j \rangle - \omega(A)|^2$, it is conjectured in Feingold and Peres (1986) that

$$V_A(\lambda) \sim \frac{C_{A-\omega(A)I}(0)}{\lambda^{n-1}\,\mathrm{vol}(\Omega)}$$

The idea is that $\varphi_\pm = (1/\sqrt{2})(\varphi_i \pm \varphi_j)$ have the same matrix element asymptotics as eigenfunctions when $\lambda_i - \lambda_j$ is sufficiently small. But then $2\langle A\varphi_+, \varphi_- \rangle = \langle A\varphi_i, \varphi_i \rangle - \langle A\varphi_j, \varphi_j \rangle$ when $A^* = A$. Since we are taking a difference, we may replace each matrix element $\langle A\varphi_i, \varphi_i \rangle$ by $\langle A\varphi_i, \varphi_i \rangle - \omega(A)$ (and also for φ_j). The conjecture then assumes that $\langle A\varphi_i, \varphi_i \rangle - \omega(A)$ has the same order of magnitude as $\langle A\varphi_i, \varphi_i \rangle - \langle A\varphi_j, \varphi_j \rangle$. Dynamical grounds for this conjecture are given in Eckhardt *et al.* (1995). The order of magnitude is predicted by some natural random wave models, as discussed in the next section.

Rigorous results

At this time, the strongest variance result is an asymptotic formula for the diagonal variance proved by Luo and Sarnak (2004) for special Hecke eigenfunctions on the quotient $\boldsymbol{H}^2/\mathrm{SL}(2,\mathbb{Z})$ of the upper half plane by the modular group. Their result pertains to holomorphic Hecke eigenforms, but the analogous statement for smooth Maass–Hecke eigenfunctions is expected to hold by similar methods, so we state the result as a theorem/conjecture. Note that $\boldsymbol{H}^2/\mathrm{SL}(2,\mathbb{Z})$ is a noncompact finite-area surface whose Laplacian Δ has both a discrete and a continuous spectrum. The discrete Hecke eigenfunctions are joint eigenfunctions of Δ and the Hecke operators T_p.

Theorem/Conjecture 1 (Luo and Sarnak 2004). *Let $\{\varphi_k\}$ denote the orthonormal basis of Hecke eigenfunctions for $\boldsymbol{H}^2/\mathrm{SL}(2,\mathbb{Z})$. Then there exists a quadratic form $B(f)$ on $C_0^\infty(\boldsymbol{H}^2/\mathrm{SL}(2,\mathbb{Z}))$ such that*

$$\frac{1}{N(\lambda)} \sum_{\lambda_j \le \lambda} \left| \int_X f|\varphi_j|^2 \mathrm{dvol} - \frac{1}{\mathrm{Vol}(X)} \int_X f \mathrm{dVol} \right|^2 = \frac{B(f,f)}{\lambda} + o\left(\frac{1}{\lambda}\right)$$

When the multiplier $f = \varphi_\lambda$ is itself an eigenfunction, Luo–Sarnak have shown that

$$B(\varphi_\lambda, \varphi_\lambda) = C_{\varphi_\lambda}(0) L(\tfrac{1}{2}, \varphi_\lambda)$$

where $L(\frac{1}{2}, \varphi_\lambda)$ is a certain L-function. Thus, the conjectured classical variance is multiplied by an arithmetic factor depending on the multiplier. A crucial fact in the proof is that the quadratic form B is diagonalized by the φ_λ.

The only rigorous result to date which is valid on general Riemannian manifolds with hyperbolic geodesic flow is the logarithmic decay:

Theorem 3 (*Zelditch*). *For any* (M, g) *with hyperbolic geodesic flow*,

$$\frac{1}{N(\lambda)}\sum_{\lambda_j\le\lambda}|(A\varphi_j,\varphi_j)-\omega(A)|^{2p}=\frac{1}{(\log\lambda)^p}$$

The logarithm reflects the exponential blow-up in time of remainder estimates for traces involving the wave group associated to hyperbolic flows. It would be surprising if the logarithmic decay is sharp for Laplacians. However, a recent result of R Schubert shows that the estimate is sharp in the case of two-dimensional hyperbolic quantum cat maps. Hence, the estimate cannot be improved by semiclassical arguments that hold in both settings.

Random Waves and Orthonormal Bases

We have mentioned that the random wave model provides a kind of guideline for what to conjecture about eigenfunctions of quantum chaotic system. In this final section, we briefly discuss random wave models and what they predict.

By a random wave model, one means a probability measure on a space of functions. To deal with orthonormal bases rather than individual functions, one sets a probability measure on a space of orthonormal bases, that is, on a unitary group. We denote expected values relative to a given probability measure by $\boldsymbol{E}$. We now consider some specific Gaussian models and what they predict about variances.

As a model for quantum chaotic eigenfunctions in plane domains, Berry (1977) suggested using the Euclidean random wave model at fixed energy. A rigorous version of such a model is as follows: let $\mathcal{E}_\lambda$ denote the space of (tempered) eigenfunctions of eigenvalue λ^2 of the Euclidean Laplacian Δ on $\mathbb{R}^n$. It is spanned by exponentials $e^{i\langle k,x\rangle}$ with $k\in\mathbb{R}^n, |k|=\lambda$. The infinite-dimensional space $\mathcal{E}_\lambda$ is a unitary representation of the Euclidean motion group and carries an invariant inner product. The inner product defines an associated Gaussian measure whose covariance kernel $C_\lambda(x,y)=\boldsymbol{E}f(x)\bar{f}(y)$ is the derivative at λ of the spectral function

$$E(\lambda,x,y)=(2\pi)^{-n}\int_{|\xi|\le\lambda}e^{i\langle x-y,\xi\rangle}\,d\xi,\quad \xi\in\mathbb{R}^n \qquad [31]$$

Thus,

$$\begin{aligned}C_\lambda(x,y)&=\frac{d}{d\lambda}E(\lambda,x,y)\\&=(2\pi)^{-n}\int_{|\xi|=\lambda}e^{i\langle x-y,\xi\rangle}dS\\&=(2\pi)^{-n}\lambda^{n-1}\int_{|\xi|=1}e^{i\lambda\langle x-y,\xi\rangle}\,dS\end{aligned}\qquad [32]$$

where dS is the usual surface measure. With this definition, $C_\lambda(x,x)\sim\lambda^{n-1}$. In order to make $E(f(x)^2)=1$ consistent with normalized eigenfunctions, we divide by λ^{n-1} to define

$$\hat{C}_\lambda(x,y)=(2\pi)^{-n}\int_{|\xi|=1}e^{i\lambda\langle x-y,\xi\rangle}dS$$

One could express the integral as a Bessel function to rewrite this as

$$\Gamma\left(\frac{n-1}{2}\right)|\lambda|x-y||^{-(n-2)/2}J_{(n-2)/2}(\lambda|x-y|)$$

Wick's formula in this ensemble gives

$$\boldsymbol{E}\varphi(x)^2\varphi(y)^2=\frac{1}{\mathrm{Vol}(\Omega)^2}[1+2C_\lambda(x,y)^2]$$

Thus, in dimension n we have

$$\begin{aligned}&\boldsymbol{E}\left[\int\!\!\int V(x)V(y)\varphi(x)^2\varphi(y)^2dxdy-\bar{V}^2\right]\\&=\frac{2}{\mathrm{Vol}(\Omega)^2}\int_\Omega\int_\Omega\hat{C}_\lambda(x,y)^2V(x)V(y)dx\,dy\\&\sim\frac{1}{\lambda^{n-1}\mathrm{Vol}(\Omega)^2}\int_\Omega\int_\Omega\frac{V(x)V(y)}{|x-y|^{n-1}}\cos(|x-y|\lambda)^2dx\,dy\end{aligned}$$

In the last line, we used the stationary-phase asymptotics

$$\begin{aligned}&(2\pi)^{-n}\lambda^{n-1}\int_{|\xi|=1}e^{i\lambda\langle x-y,\xi\rangle}dS\\&\quad\sim C_n(\lambda|x-y|)^{-(n-1)/2}\cos(|x-y|\lambda)\end{aligned}\qquad [33]$$

Thus, the variances have order $\lambda^{-(n-1)}$ in dimension n, consistent with the conjectures in Feingold and Peres (1986) and Eckhardt *et al.* (1995).

This model is often used to obtain predictions on eigenfunctions of chaotic systems. By construction, it is tied to Euclidean geometry and only pertains directly to individual eigenfunctions of a fixed eigenvalue. It is based on the infinite-dimensional multiplicity of eigenfunctions of fixed eigenvalue of the Euclidean Laplacian on $\mathbb{R}^n$. There also exist random wave models on a curved Riemannian manifold (M,g), which model individual eigenfunctions and also random orthonormal bases

(Zelditch 1996a). Thus, one can compare the behavior of sums over eigenvalues of the orthonormal basis of eigenfunctions of Δ with that of a random orthonormal basis. Instead of taking Gaussian random combinations of Euclidean plane waves of a fixed eigenvalue, one takes Gaussian random combinations $\sum_{j:\lambda_j\in[\lambda,\lambda+1]} c_j\varphi_j$ of the eigenfunctions of (M,g) with eigenvalues in a short interval in the sense above. Equivalently, one takes random combinations with $\sum_j |c_j|^2 = 1$. These random waves are globally adapted to (M,g). The statistical results depend on the measure of the set of periodic geodesics of (M,g); thus, as discussed in Kaplan and Heller (1998), different random wave models make different predictions about off-diagonal variances.

Fix a compact Riemannian manifold (M,g) and partition the spectrum of $\sqrt{\Delta}$ into the intervals $I_k = [k, k+1]$. Let $\Pi_k = E(k+1) - E(k)$ be the kernel of the spectral projections for $\sqrt{\Delta}$ corresponding to the interval I_k. Its kernel $\Pi_k(x,y)$ is the covariance kernel of Gaussian random combinations $\sum_{j:\lambda_j\in I_k} c_j\varphi_j$ and is analogous to $C_\lambda(x,y)$ in the Euclidean case; it is of course not the derivative $\mathrm{d}E(\lambda,x,y)$ but the difference of the spectral projector over I_k. We denote by $N(k)$ the number of eigenvalues in I_k and put $\mathcal{H}_k = \mathrm{ran}\Pi_k$ (the range of Π_k). We define a "random" orthonormal basis of $\mathcal{H}_k$ by changing the basis of eigenfunctions $\{\varphi_j\}$ of Δ in $\mathcal{H}_k$ by a random element of the unitary group $U(\mathcal{H}_k)$ of the finite-dimensional Hilbert space $\mathcal{H}_k$. We then define a random orthonormal basis of $L^2(M)$ by taking the product over all the spectral intervals in our partition. More precisely, we define the infinite-dimensional unitary group

$$U(\infty) = \prod_{k=1}^{\infty} U(\mathcal{H}_k)$$

of sequences $(U_1, U_2, \dots)$, with $U_k \in U(\mathcal{H}_k)$. We equip $U(\infty)$ with the product

$$\mathrm{d}\nu_\infty = \prod_{k=1}^{\infty} \mathrm{d}\nu_k$$

of the unit mass Haar measures $\mathrm{d}\nu_k$ on $U(\mathcal{H}_k)$: we then define a random orthonormal basis of $L^2(M)$ to be obtained by applying a random element $U \in U(\infty)$ to the orthonormal basis $\Phi = \{\varphi_j\}$ of eigenfunctions of $\sqrt{\Delta}$.

Assuming the set of periodic geodesics of (M,g) has measure zero, the Weyl remainder results [8] and strong Szegö limit asymptotics of Guillemin–Okikiolu and Laptev–Robert–Safarov give two term asymptotics for the traces $\Pi_k A \Pi_k, (\Pi_k A \Pi_k)^2$ for any pseudodifferential operator A. Combining the strong Szegö asymptotics with the arguments of Zelditch (1996a), random orthonormal bases can be proved to satisfy the following variance asymptotics:

1. $E(\sum_{j:\lambda_j\in I_k} |(AU\varphi_j, U\varphi_j) - \omega(A)|^2 \sim (\omega(A^*A) - \omega(A)^2)$;

2. $E(\sum_{i\neq j:\lambda_j,\lambda_i\in I_k} \left|\frac{\sin T(\lambda_i-\lambda_j-\tau)}{T(\lambda_i-\lambda_j-\tau)}\right|^2 |(AU\varphi_j, U\varphi_i)|^2 \sim \left\{2\left|\frac{\sin\tau T}{\tau T}\right|^2 + \frac{1}{N(k)}\sum_{i\neq j}\left|\frac{\sin T(\lambda_i-\lambda_j-\tau)}{T(\lambda_i-\lambda_j-\tau)}\right|^2\right\} \times(\omega(A^*A) - \omega(A)^2)$

See also: Arithmetic Quantum Chaos; Determinantal Random Fields; Eigenfunctions of Quantum Completely Integrable Systems; Fractal Dimensions in Dynamics; *h*-Pseudodifferential Operators and Applications; Number Theory in Physics; Regularization for Dynamical Zeta Functions; Semiclassical Spectra and Closed Orbits.

Further Reading

Bäcker A, Schubert R, and Stifter P (1998a) Rate of quantum ergodicity in Euclidean billiards. *Physical Review E (3)* 57(5): part A 5425–5447.

Bäcker A, Schubert R, and Stifter P (1998b) Rate of quantum ergodicity in Euclidean billiards – erratum. *Physical Review E, (3)* 58(4): 5192 (math-ph/0512030).

Barnett A (2005) Asymptotic rate of quantum ergodicity in chaotic Euclidean billiards.

Berry MV (1977) Regular and irregular semiclassical wavefunctions. *Journal of Physics A* 10(12): 2083–2091.

Bohigas O, Giannoni MJ, and Schmit C (1984) Characterization of chaotic quantum spectra and universality of level fluctuation laws. *Physical Review Letters* 52(1): 1–4.

Eckhardt B, Fishman S, Keating J, Agam O, Main J *et al.* (1995) Approach to ergodocity in quantum wave functions. *Physical Review E* 52: 5893–5903.

Faure F, Nonnenmacher S, and De Bièvre S (2003) Scarred eigenstates for quantum cat maps of minimal periods. *Communications in Mathematical Physics* 239(3): 449–492.

Feingold M and Peres A (1986) Distribution of matrix elements of chaotic systems. *Physical Review A (3)* 34(1): 591–595.

Heller EJ (1984) Bound-state eigenfunctions of classically chaotic Hamiltonian systems: Scars of periodic orbits. *Physical Review Letters* 53(16): 1515–1518.

Kaplan L and Heller EJ (1998) Weak quantum ergodicity. *Physica D* 121(1–2): 1–18 (Appendix by S. Zelditch.).

Luo W and Sarnak P (2004) Quantum variance for Hecke eigenforms. *Annales Scientifiques de l'Ecole Normale Superieure* 37: 769–799.

Marklof J and O'Keefe S (2005) Weyl's law and quantum ergodicity for maps with divided phase space (with an appendix 'Converse quantum ergodicity' by Steve Zelditch). *Nonlinearity* 18(1): 277–304.

Rudnick Z and Sarnak P (1994) The behaviour of eigenstates of arithmetic hyperbolic manifolds. *Communications in Mathematical Physics* 161: 195–213.

Ruelle D (1969) *Statistical Mechanics: Rigorous Results*. New York: W. A. Benjamin.

Sarnak P (1995) Arithmetic quantum chaos. In: *The Schur Lectures (Tel Aviv, 1992)*, *Israel Mathematical Conference Proc*, vol. 8 (1995), pp. 183–236.
Zelditch S (1996a) A random matrix model for quantum mixing. *International Mathematics Research Notices* 3: 115–137.
Zelditch S (1996b) Quantum ergodicity of C^* dynamical systems. *Communications in Mathematical Physics* 177: 507–528.
Zelditch S (1996c) Quantum Mixing. *Journal of Functional Analysis* 140: 68–86.
Zelditch S and Zworski M (1996) Ergodicity of eigenfunctions for ergodic billiards. *Communications in Mathematical Physics* 175: 673–682.

Quantum Error Correction and Fault Tolerance

D Gottesman, Perimeter Institute, Waterloo, ON, Canada

Quantum Error Correction

Building a quantum computer or a quantum communications device in the real world means having to deal with errors. Any qubit stored unprotected or one transmitted through a communications channel will inevitably come out at least slightly changed. The theory of quantum error-correcting codes (QECCs) has been developed to counteract noise introduced in this way. By adding extra qubits and carefully encoding the quantum state we wish to protect, a quantum system can be insulated to a great extent against errors.

To build a quantum computer, we face an even more daunting task: if our quantum gates are imperfect, everything we do will add to the error. The theory of fault-tolerant quantum computation tells us how to perform operations on states encoded in a QECC without compromising the code's ability to protect against errors.

In general, a QECC is a subspace of a Hilbert space designed so that any of a set of possible errors can be corrected by an appropriate quantum operation. Specifically:

Definition 1 Let $\mathcal{H}_n$ be a 2^n-dimensional Hilbert space (n qubits), and let C be a K-dimensional subspace of $\mathcal{H}_n$. Then C is an $((n, K))$ (binary) QECC correcting the set of errors $\mathcal{E} = \{E_a\}$ iff $\exists \mathcal{R}$ s.t. $\mathcal{R}$ is a quantum operation and $(\mathcal{R} \circ E_a)(|\psi\rangle) = |\psi\rangle$ for all $E_a \in \mathcal{E}, |\psi\rangle \in C$.

$\mathcal{R}$ is called the "recovery" or "decoding" operation and serves to actually perform the correction of the state. The decoder is sometimes also taken to map $\mathcal{H}_n$ into an unencoded Hilbert space $\mathcal{H}_{\log K}$ isomorphic to C. This should be distinguished from the "encoding" operation which maps $\mathcal{H}_{\log K}$ into $\mathcal{H}_n$, determining the imbedding of C. The computational complexity of the encoder is frequently a great deal lower than that of the decoder. In particular, the task of determining what error has occurred can be computationally difficult (NP-hard, in fact), and designing codes with efficient decoding algorithms is an important task in quantum error correction, as in classical error correction.

This article will cover only binary quantum codes, built with qubits as registers, but all of the techniques discussed here can be generalized to higher-dimensional registers, or "qudits."

To determine whether a given subspace is able to correct a given set of errors, we can apply the quantum error-correction conditions (Bennett *et al.* 1996, Knill and Laflamme 1997):

Theorem 1 *A QECC C corrects the set of errors $\mathcal{E}$ iff*

$$\langle \psi_i | E_a^\dagger E_b | \psi_j \rangle = C_{ab} \delta_{ij} \qquad [1]$$

where $E_a, E_b \in \mathcal{E}$ and $\{|\psi_i\rangle\}$ form an orthonormal basis for C.

The salient point in these error-correction conditions is that the matrix element C_{ab} does not depend on the encoded basis states i and j, which, roughly speaking, indicates that neither the environment nor the decoding operation learns any information about the encoded state. We can imagine the various possible errors taking the subspace C into other subspaces of $\mathcal{H}_n$, and we want those subspaces to be isomorphic to C, and to be distinguishable from each other by an appropriate measurement. For instance, if $C_{ab} = \delta_{ab}$, then the various erroneous subspaces are orthogonal to each other.

Because of the linearity of quantum mechanics, we can always take the set of errors $\mathcal{E}$ to be a linear space: if a QECC corrects E_a and E_b, it will also correct $\alpha E_a + \beta E_b$ using the same recovery operation. In addition, if we write any superoperator $\mathcal{S}$ in terms of its operator-sum representation $\mathcal{S}(\rho) \mapsto \sum A_k \rho A_k^\dagger$, a QECC that corrects the set of errors $\{A_k\}$ automatically corrects $\mathcal{S}$ as well. Thus, it is sufficient in general to check that the error-correction conditions hold for a basis of errors.

Frequently, we are interested in codes that correct any error affecting t or fewer physical qubits. In that case, let us consider tensor products of the Pauli matrices

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
$$Y = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad [2]$$

Define the Pauli group $\mathcal{P}_n$ as the group consisting of tensor products of I, X, Y, and Z on n qubits, with an overall phase of ± 1 or $\pm \mathrm{i}$. The weight $\mathrm{wt}(P)$ of a Pauli operator $P \in \mathcal{P}_n$ is the number of qubits on which it acts as X, Y, or Z (i.e., not as the identity). Then the Pauli operators of weight t or less form a basis for the set of all errors acting on t or fewer qubits, so a QECC which corrects these Pauli operators corrects all errors acting on up to t qubits. If we have a channel which causes errors independently with probability $O(\epsilon)$ on each qubit in the QECC, then the code will allow us to decode a correct state except with probability $O(\epsilon^{t+1})$, which is the probability of having more than t errors. We get a similar result in the case where the noise is a general quantum operation on each qubit which differs from the identity by something of size $O(\epsilon)$.

Definition 2 The distance d of an $((n, K))$ QECC is the smallest weight of a nontrivial Pauli operator $E \in \mathcal{P}_n$ s.t. the equation

$$\langle \psi_i | E | \psi_j \rangle = C(E)\delta_{ij} \qquad [3]$$

fails.

We use the notation $((n, K, d))$ to refer to an $((n, K))$ QECC with distance d. Note that for $P, Q \in \mathcal{P}_n, \mathrm{wt}(PQ) \leq \mathrm{wt}(P) + \mathrm{wt}(Q)$. Then by comparing the definition of distance with the quantum error-correction conditions, we immediately see that a QECC corrects t general errors iff its distance $d > 2t$. If we are instead interested in "erasure" errors, when the location of the error is known but not its precise nature, a distance d code corrects $d - 1$ erasure errors. If we only wish to detect errors, a distance d code can detect errors on up to $d - 1$ qubits.

One of the central problems in the theory of quantum error correction is to find codes which maximize the ratios $(\log K)/n$ and d/n, so they can encode as many qubits as possible and correct as many errors as possible. Conversely, we are also interested in the problem of setting upper bounds on achievable values of $(\log K)/n$ and d/n. The quantum Singleton bound (or Knill–Laflamme (1997) bound) states that any $((n, K, d))$ QECC must satisfy

$$n - \log K \geq 2d - 2 \qquad [4]$$

We can set a lower bound on the existence of QECCs using the quantum Gilbert–Varshamov bound, which states that, for large n, an $((n, 2^k, d))$ QECC exists provided that

$$k/n \leq 1 - (d/n)\log 3 - h(d/n) \qquad [5]$$

where $h(x) = -x \log x - (1 - x)\log(1 - x)$ is the binary Hamming entropy. Note that the Gilbert–Varshamov bound simply states that codes at least this good exist; it does not suggest that better codes cannot exist.

Stabilizer Codes

In order to better manipulate and discover QECCs, it is helpful to have a more detailed mathematical structure to work with. The most widely used structure gives a class of codes known as "stabilizer codes" (Calderbank *et al.* 1998, Gottesman 1996). They are less general than arbitrary quantum codes, but have a number of useful properties that make them easier to work with than the general QECC.

Definition 3 Let $S \subset \mathcal{P}_n$ be an abelian subgroup of the Pauli group that does not contain -1 or $\pm \mathrm{i}$, and let $C(S) = \{|\psi\rangle \text{ s.t. } P|\psi\rangle = |\psi\rangle \forall P \in S\}$. Then $C(S)$ is a stabilizer code and S is its stabilizer.

Because of the simple structure of the Pauli group, any abelian subgroup has order 2^{n-k} for some k and can easily be specified by giving a set of $n - k$ commuting generators.

The code words of the QECC are by definition in the $+1$-eigenspace of all elements of the stabilizer, but an error E acting on a code word will move the state into the -1-eigenspace of any stabilizer element M which anticommutes with E:

$$M(E|\psi\rangle) = -EM|\psi\rangle = -E|\psi\rangle \qquad [6]$$

Thus, measuring the eigenvalues of the generators of S tells us information about the error that has occurred. The set of such eigenvalues can be represented as an $(n - k)$-dimensional binary vector known as the "error syndrome." Note that the error syndrome does not tell us anything about the encoded state, only about the error that has occurred.

Theorem 2 *Let S be a stabilizer with $n - k$ generators, and let $S^{\perp} = \{E \in \mathcal{P}_n \text{ s.t. } [E, M] = 0 \ \forall M \in S\}$. Then S encodes k qubits and has distance d, where d is the smallest weight of an operator in $S^{\perp} \setminus S$.*

We use the notation $[[n, k, d]]$ to a refer to such a stabilizer code. Note that the square brackets specify that the code is a stabilizer code, and that the middle term k refers to the number of encoded qubits, and not the dimension 2^k of the encoded subspace, as for the general QECC (whose dimension might not be a power of 2).

$S^{\perp}$ is the set of Pauli operators that commute with all elements of the stabilizer. They would therefore appear to be those errors which cannot be detected by the code. However, the theorem specifies the distance of the code by considering $S^{\perp} \backslash S$. A Pauli operator $P \in S$ cannot be detected by the code, but there is in fact no need to detect it, since all code words remain fixed under P, making it equivalent to the identity operation. A distance d stabilizer code which has nontrivial $P \in S$ with $\mathrm{wt}(P) < d$ is called degenerate, whereas one which does not is nondegenerate. The phenomenon of degeneracy has no analog for classical error-correcting codes, and makes the study of quantum codes substantially more difficult than the study of classical error correction. For instance, a standard bound on classical error correction is the Hamming bound (or sphere-packing bound), but the analogous quantum Hamming bound

$$k/n \leq 1 - (t/n)\log 3 - h(t/n) \qquad [7]$$

for $[[n, k, 2t+1]]$ codes (when n is large) is only known to apply to nondegenerate quantum codes (though in fact we do not know of any degenerate QECCs that violate the quantum Hamming bound).

An example of a stabilizer code is the 5-qubit code, a [[5,1,3]] code whose stabilizer can be generated by

$$\begin{array}{l} X \otimes Z \otimes Z \otimes X \otimes I \\ I \otimes X \otimes Z \otimes Z \otimes X \\ X \otimes I \otimes X \otimes Z \otimes Z \\ Z \otimes X \otimes I \otimes X \otimes Z \end{array}$$

The 5-qubit code is a nondegenerate code, and is the smallest possible QECC which corrects 1 error (as one can see from the quantum Singleton bound).

It is frequently useful to consider other representations of stabilizer codes. For instance, $P \in \mathcal{P}_n$ can be represented by a pair of n-bit binary vectors $(p_X \,|\, p_Z)$, where p_X is 1 for any location where P has an X or Y tensor factor and is 0 elsewhere, and p_Z is 1 for any location where P has a Y or Z tensor factor. Two Pauli operators $P = (p_X|p_Z)$ and $Q = (q_X|q_Z)$ commute iff $p_X \cdot q_Z + p_Z \cdot q_X = 0$. Then the stabilizer for a code becomes a pair of $(n-k) \times n$ binary matrices, and most interesting properties can be determined by an appropriate linear algebra exercise. Another useful representation is to map the single-qubit Pauli operators I, X, Y, Z to the finite field GF(4), which sets up a connection between stabilizer codes and a subset of classical codes on four-dimensional registers.

CSS Codes

CSS codes are a very useful class of stabilizer codes invented by Calderbank and Shor (1996), and by Steane (1996). The construction takes two binary classical linear codes and produces a quantum code, and can therefore take advantage of much existing knowledge from classical coding theory. In addition, CSS codes have some very useful properties which make them excellent choices for fault-tolerant quantum computation.

A classical $[n, k, d]$ linear code (n physical bits, k logical bits, classical distance d) can be defined in terms of an $(n-k) \times n$ binary "parity check" matrix H – every classical code word v must satisfy $Hv = 0$. Each row of the parity check matrix can be converted into a Pauli operator by replacing each 0 with an I operator and each 1 with a Z operator. Then the stabilizer code generated by these operators is precisely a quantum version of the classical error-correcting code given by H. If the classical distance $d = 2t+1$, the quantum code can correct t bit flip (X) errors, just as could the classical code.

If we want to make a QECC that can also correct phase (Z) errors, we should choose two classical codes C_1 and C_2, with parity check matrices H_1 and H_2. Let C_1 be an $[n, k_1, d_1]$ code and let C_2 be an $[n, k_2, d_2]$ code. We convert H_1 into stabilizer generators as above, replacing each 0 with I and each 1 with Z. For H_2, we perform the same procedure, but each 1 is instead replaced by X. The code will be able to correct bit flip (X) errors as if it had a distance d_1 and to correct phase (Z) errors as if it had a distance d_2. Since these two operations are completely separate, it can also correct Y errors as both a bit flip and a phase error. Thus, the distance of the quantum code is at least $\min(d_1, d_2)$, but might be higher because of the possibility of degeneracy.

However, in order to have a stabilizer code at all, the generators produced by the above procedure must commute. Define the dual $C^{\perp}$ of a classical code C as the set of vectors w s.t. $w \cdot v = 0$ for all $v \in C$. Then the Z generators from H_1 will all commute with the X generators from H_2 iff $C_2^{\perp} \subseteq C_1$ (or equivalently, $C_1^{\perp} \subseteq C_2$). When this is true, C_1 and C_2 define an $[[n, k_1 + k_2 - n, d]]$ stabilizer code, where $d \geq \min(d_1, d_2)$.

The smallest distance-3 CSS code is the 7-qubit code, a [[7, 1, 3]] QECC created from the classical Hamming code (consisting of all sums of classical strings 1111000, 1100110, 1010101, and 1111111). The encoded $|\bar{0}\rangle$ for this code consists of the superposition of all even-weight classical code words and the encoded $|\bar{1}\rangle$ is the superposition of all odd-weight classical code words. The 7-qubit code is much studied because its properties make it particularly well suited to fault-tolerant quantum computation.

Fault Tolerance

Given a QECC, we can attempt to supplement it with protocols for performing fault-tolerant operations. The basic design principle of a fault-tolerant protocol is that an error in a single location – either a faulty gate or noise on a quiescent qubit – should not be able to alter more than a single qubit in each block of the QECC. If this condition is satisfied, t separate single-qubit or single-gate failures are required for a distance $2t + 1$ code to fail.

Particular caution is necessary, as computational gates can cause errors to propagate from their original location onto qubits that were previously correct. In general, a gate coupling pairs of qubits allows errors to spread in both directions across the coupling.

The solution is to use transversal gates whenever possible (Shor 1996). A transversal operation is one in which the ith qubit in each block of a QECC interacts only with the ith qubit of other blocks of the code or of special ancilla states. An operation consisting only of single-qubit gates is automatically transversal. A transversal operation has the virtue that an error occurring on the third qubit in a block, say, can only ever propagate to the third qubit of other blocks of the code, no matter what other sequence of gates we perform before a complete error-correction procedure.

In the case of certain codes, such as the 7-qubit code, a number of different gates can be performed transversally. Unfortunately, it does not appear to be possible to perform universal quantum computations using just transversal gates. We therefore have to resort to more complicated techniques. First we create special encoded ancilla states in a non-fault-tolerant way, but perform some sort of check on them (in addition to error correction) to make sure they are not too far off from the goal. Then we interact the ancilla with the encoded data qubits using gates from our stock of transversal gates and perform a fault-tolerant measurement. Then we complete the operation with a further transversal gate which depends on the outcome of the measurement.

Fault-Tolerant Gates

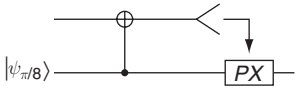
We will focus on stabilizer codes. Universal fault tolerance is known to be possible for any stabilizer code, but in most cases the more complicated type of construction is needed for all but a few gates. The Pauli group $\mathcal{P}_k$, however, can be performed transversally on any stabilizer code. Indeed, the set $S^{\perp} \backslash S$ of undetectable errors is a boon in this case, as it allows us to perform these gates. In particular, each coset $S^{\perp}/S$ corresponds to a different logical Pauli operator (with S itself corresponding to the identity). On a stabilizer code, therefore, logical Pauli operations can be performed via a transversal Pauli operation on the physical qubits.

Stabilizer codes have a special relationship to a finite subgroup $\mathcal{C}_n$ of the unitary group $U(2^n)$ frequently called the "Clifford group." The Clifford group on n qubits is defined as the set of unitary operations which conjugate the Pauli group $\mathcal{P}_n$ into itself; $\mathcal{C}_n$ can be generated by the Hadamard transform, the controlled-NOT (CNOT), and the single-qubit $\pi/4$ phase rotation diag(1, i). The set of stabilizer codes is exactly the set of codes which can be created by a Clifford group encoder circuit using $|0\rangle$ ancilla states.

Some stabilizer codes have interesting symmetries under the action of certain Clifford group elements, and these symmetries result in transversal gate operations. A particularly useful fact is that a transversal CNOT gate (i.e., CNOT acting between the ith qubit of one block of the QECC and the ith qubit of a second block for all i) acts as a logical CNOT gate on the encoded qubits for any CSS code. Furthermore, for the 7-qubit code, transversal Hadamard performs a logical Hadamard, and the transversal $\pi/4$ rotation performs a logical $-\pi/4$ rotation. Thus, for the 7-qubit code, the full logical Clifford group is accessible via transversal operations.

Unfortunately, the Clifford group by itself does not have much computational power: it can be efficiently simulated on a classical computer. We need to add some additional gate outside the Clifford group to allow universal quantum computation; a single gate will suffice, such as the single-qubit $\pi/8$ phase rotation diag(1, $\exp(i\pi/4)$). Note that this gives us a finite generating set of gates. However, by taking appropriate products, we get an infinite set of gates, one that is dense in the unitary group $U(2^n)$, allowing universal quantum computation.

The following circuit performs a $\pi/8$ rotation, given an ancilla state $|\psi_{\pi/8}\rangle = |0\rangle + \exp(i\pi/4)|1\rangle$:



Here P is the $\pi/4$ phase rotation diag(1, i), and X is the bit flip. The product is in the Clifford group, and is only performed if the measurement outcome is 1. Therefore, given the ability to perform fault-tolerant Clifford group operations, fault-tolerant measurements, and to prepare the encoded $|\psi_{\pi/8}\rangle$ state, we have universal fault-tolerant quantum computation. A slight generalization of the fault-tolerant measurement procedure below can be used to fault-tolerantly verify the $|\psi_{\pi/8}\rangle$ state, which is a $+1$ eigenstate of PX. Using this or another verification procedure, we can check a non-fault-tolerant construction.

Fault-Tolerant Measurement and Error Correction

Since all our gates are unreliable, including those used to correct errors, we will need some sort of fault-tolerant quantum error-correction procedure. A number of different techniques have been developed. All of them share some basic features: they involve creation and verification of specialized ancilla states, and use transversal gates which interact the data block with the ancilla state.

The simplest method, due to Shor, is very general but also requires the most overhead and is frequently the most susceptible to noise. Note that the following procedure can be used to measure (non-fault-tolerantly) the eigenvalue of any (possibly multiqubit) Pauli operator M: produce an ancilla qubit in the state $|+\rangle = |0\rangle + |1\rangle$. Perform a controlled-$M$ operation from the ancilla to the state being measured. In the case where M is a multiqubit Pauli operator, this can be broken down into a sequence of controlled-X, controlled-Y, and controlled-Z operations. Then measure the ancilla in the basis of $|+\rangle$ and $|-\rangle = |0\rangle - |1\rangle$. If the state is a $+1$ eigenvector of M, the ancilla will be $|+\rangle$, and if the state is a -1 eigenvector, the ancilla will be $|-\rangle$.

The advantage of this procedure is that it measures just M and nothing more. The disadvantage is that it is not transversal, and thus not fault tolerant. Instead of the unencoded $|+\rangle$ state, we must use a more complex ancilla state $|00\ldots0\rangle + |11\ldots1\rangle$ known as a "cat" state. The cat state contains as many qubits as the operator M to be measured, and we perform the controlled-X, -Y, or -Z operations transversally from the appropriate qubits of the cat state to the appropriate qubits in the data block. Since, assuming the cat state is correct, all of its qubits are either $|0\rangle$ or $|1\rangle$, the procedure either leaves the data state alone or performs M on it uniformly. A $+1$ eigenstate in the data therefore leaves us with $|00\ldots0\rangle + |11\ldots1\rangle$ in the ancilla and a -1 eigenstate leaves us with $|00\ldots0\rangle - |11\ldots1\rangle$. In either case, the final state still tells us nothing about the data beyond the eigenvalue of M. If we perform a Hadamard transform and then measure each qubit in the ancilla, we get either a random even-weight string (for eigenvalue $+1$) or an odd-weight string (for eigenvalue -1).

The procedure is transversal, so an error on a single qubit in the initial cat state or in a single gate during the interaction will only produce one error in the data. However, the initial construction of the cat state is not fault tolerant, so a single-gate error then could eventually produce two errors in the data block. Therefore, we must be careful and use some sort of technique to verify the cat state, for instance, by checking if random pairs of qubits are the same. Also, note that a single phase error in the cat state will cause the final measurement outcome to be wrong (even and odd switch places), so we should repeat the measurement procedure multiple times for greater reliability.

We can then make a full fault-tolerant error-correction procedure by performing the above measurement technique for each generator of the stabilizer. Each measurement gives us one bit of the error syndrome, which we then decipher classically to determine the actual error.

More sophisticated techniques for fault-tolerant error correction involve less interaction with the data but at the cost of more complicated ancilla states. A procedure due to Steane uses (for CSS codes) one ancilla in a logical $|\bar{0}\rangle$ state of the same code and one ancilla in a logical $|\bar{0}\rangle + |\bar{1}\rangle$ state. A procedure due to Knill (for any stabilizer code) teleports the data qubit through an ancilla consisting of two blocks of the QECC containing an encoded Bell state $|\overline{00}\rangle + |\overline{11}\rangle$. Because the ancillas in Steane and Knill error correction are more complicated than the cat state, it is especially important to verify the ancillas before using them.

The Threshold for Fault Tolerance

In an unencoded protocol, even one error can destroy the computation, but a fully fault-tolerant protocol will give the right answer unless multiple

errors occur before they can be corrected. On the other hand, the fault-tolerant protocol is larger, requiring more qubits and more time to do each operation, and therefore providing more opportunities for errors. If errors occur on the physical qubits independently at random with probability p per gate or time step, the fault-tolerant protocol has probability of logical error for a single logical gate or time step at most Cp^2, where C is a constant that depends on the design of the fault-tolerant circuitry (assume the QECC has distance 3, as for the 7-qubit code). When $p < p_t = 1/C$, the fault tolerance helps, decreasing the logical error rate. p_t is the "threshold" for fault-tolerant quantum computation. If the error rate is higher than the threshold, the extra overhead means that errors will occur faster than they can be reliably corrected, and we are better off with an unencoded system.

To further lower the logical error rate, we turn to a family of codes known as "concatenated codes" (Aharonov and Ben-Or, Kitaev 1997, Knill *et al.* 1998). Given a code word of a particular $[[n, 1]]$ QECC, we can take each physical qubit and again encode it using the same code, producing an $[[n^2, 1]]$ QECC. We could repeat this procedure to get an n^3-qubit code, and so forth. The fault-tolerant procedures concatenate as well, and after L levels of concatenation, the effective logical error rate is $p_t(p/p_t)^{2^L}$ (for a base code correcting 1 error). Therefore, if p is below the threshold p_t, we can achieve an arbitrarily good error rate ϵ per logical gate or time step using only poly$(\log \epsilon)$ resources, which is excellent theoretical scaling.

Unfortunately, the practical requirements for this result are not nearly so good. The best rigorous proofs of the threshold to date show that the threshold is at least 2×10^{-5} (meaning one error per 50,000 operations). Optimized simulations of fault-tolerant protocols suggest that the true threshold may be as high as 5%, but to tolerate this much error, existing protocols require enormous overhead, perhaps increasing the number of gates and qubits by a factor of a million or more for typical computations. For lower physical error rates, overhead requirements are more modest, particularly if we only attempt to optimize for calculations of a given size, but are still larger than one would like.

Furthermore, these calculations make a number of assumptions about the physical properties of the computer. The errors are assumed to be independent and uncorrelated between qubits except when a gate connects them. It is assumed that measurements and classical computations can be performed quickly and reliably, and that quantum gates can be performed between arbitrary pairs of qubits in the computer, irrespective of their physical proximity. Of these, only the assumption of independent errors is at all necessary, and that can be considerably relaxed to allow short-range correlations and certain kinds of non-Markovian environments. However, the effects of relaxing these assumptions on the threshold value and overhead requirements have not been well studied.

Further Reading

Aharonov D and Ben-Or M (1999) Fault-tolerant quantum computation with constant error rate, quant-ph/9906129.

Bennett C, DiVincenzo D, Smolin J, and Wootters W (1996) Mixed state entanglement and quantum error correction. *Physical Review A* 54: 3824–3851 (quant-ph/9604024).

Calderbank AR and Shor PW (1996) Good quantum error-correcting codes exist. *Physical Review A* 54: 1098–1105 (quant-ph/9512032).

Calderbank AR, Rains EM, Shor PW, and Sloane NJA (1998) Quantum error correction via codes over GF(4). *IEEE Transactions on Information Theory* 44: 1369–1387 (quant-ph/9605005).

Gottesman D (1996) Class of quantum error-correcting codes saturating the quantum Hamming bound. *Physical Review A* 54: 1862–1868 (quant-ph/9604038).

Kitaev AY (1997) Quantum error correction with imperfect gates. In: Hirota O, Holeva AS, and Caves CM (eds.) *Quantum Communication, Computing, and Measurement (Proc. 3rd Int. Conf. of Quantum Communication and Measurement)*, pp. 181–188. New York: Plenum.

Knill E and Laflamme R (1997) A theory of quantum error-correcting codes. *Physical Review A* 55: 900–911 (quant-ph/9604034).

Knill E, Laflamme R, and Zurek WH (1998) Resilient quantum computation. *Science* 279: 342–345.

Shor PW (1996) Fault-tolerant quantum computation. In: *Proc. 35th Ann. Symp. on Fundamentals of Computer Science*, pp. 56–65. (quant-ph/9605011). Los Alamitos: IEEE Press.

Steane AM (1996) Multiple particle interference and quantum error correction. *Proceedings of the Royal Society of London A* 452: 2551–2577 (quant-ph/9601029).

Quantum Field Theory in Curved Spacetime

B S Kay, University of York, York, UK

Introduction and Preliminaries

Quantum Field Theory (QFT) in curved spacetime is a hybrid approximate theory in which quantum matter fields are assumed to propagate in a fixed classical background gravitational field. Its basic physical prediction is that strong gravitational fields can polarize the vacuum and, when time dependent, lead to pair creation just as a strong and/or time-dependent electromagnetic field can polarize the vacuum and/or give rise to pair creation of charged particles. One expects it to be a good approximation to full quantum gravity provided the typical frequencies of the gravitational background are very much less than the Planck frequency $(c^5/G\hbar)^{1/2} \sim 10^{43}\,\mathrm{s}^{-1}$) and provided, with a suitable measure for energy, the energy of created particles is very much less than the energy of the background gravitational field or of its matter sources. Undoubtedly, the most important prediction of the theory is the Hawking effect, according to which a, say spherically symmetric, classical black hole of mass M will emit thermal radiation at the Hawking temperature $T=(8\pi M)^{-1}$ (here and from now on, we use Planck units where $G, c, \hbar$ and, k (Boltzmann's constant) are all taken to be 1).

On the mathematical side, the need to formulate the laws and derive the general properties of QFT on nonflat spacetimes forces one to state and prove results in local terms and, as a byproduct, thereby leads to an improved perspective on flat-spacetime QFT too. It is also interesting to formulate QFT on idealized spacetimes with particular global geometrical features. Thus, QFT on spacetimes with bifurcate Killing horizons is intimately related to the Hawking effect; QFT on spacetimes with closed timelike curves is intimately related to the question whether the laws of physics permit the manufacture of a time machine.

As is standard in general relativity, a curved spacetime is modeled mathematically as a (paracompact, Hausdorff) manifold $\mathcal{M}$ equipped with a pseudo-Riemannian metric g of signature $(-,+++)$ (we follow the conventions of the standard text by Misner *et al.* (1973)). We shall also assume, except where otherwise stated, our spacetime to be *globally hyperbolic*, that is, that $\mathcal{M}$ admits a *global time coordinate*, by which we mean a global coordinate t such that each constant-t surface is a smooth Cauchy surface, that is, a smooth spacelike 3-surface cut exactly once by each inextendible causal curve. (Without this default assumption, extra problems arise for QFT which we shall briefly mention in connection with the "time machine" question discussed later.) In view of this definition, globally hyperbolic spacetimes are clearly time-orientable and we shall assume a choice of time-orientation has been made so we can talk about the "future" and "past" directions. Modern formulations of the subject take, as the fundamental mathematical structure modeling the quantum field, a $*$-algebra $\mathcal{A}$ (with identity I) together with a family of local sub $*$-algebras $\mathcal{A}(\mathcal{O})$ labeled by bounded open regions $\mathcal{O}$ of the spacetime $(\mathcal{M}, g)$ and satisfying the *isotony* or *net* condition that $\mathcal{O}_1 \subset \mathcal{O}_2$ implies $\mathcal{A}(\mathcal{O}_1)$ is a subalgebra of $\mathcal{A}(\mathcal{O}_2)$ as well as the condition that whenever two bounded open regions $\mathcal{O}_1$ and $\mathcal{O}_2$ are spacelike separated, then $\mathcal{A}(\mathcal{O}_1)$ and $\mathcal{A}(\mathcal{O}_2)$ commute.

Standard concepts and techniques from algebraic quantum theory are then applicable: In particular, *states* are defined to be positive (this means $\omega(A^*A) \geq 0\ \forall A \in \mathcal{A}$) normalized (this means $\omega(I)=1$) linear functionals on $\mathcal{A}$. One distinguishes between *pure* states and *mixed* states, only the latter being writable as nontrivial convex combinations of other states. To each state, ω, the *GNS construction* associates a representation, ρ_ω, of $\mathcal{A}$ on a Hilbert space $\mathcal{H}_\omega$ together with a cyclic vector $\Omega \in \mathcal{H}_\omega$ such that

$$\omega(A) = \langle \Omega | \rho_\omega(A) \Omega \rangle$$

(and the *GNS triple* $(\rho_\omega, \mathcal{H}, \Omega)$ is unique up to equivalence). There are often technical advantages in formulating things so that the $*$-algebra is a C^*-algebra. Then the GNS representation is as everywhere-defined bounded operators and is irreducible if and only if the state is pure. A useful concept, due to Haag, is the *folium* of a given state ω which may be defined to be the set of all states ω_σ which arise in the form $\mathrm{tr}(\sigma\rho_\omega(\cdot))$, where σ ranges over the density operators (trace-class operators with unit trace) on $\mathcal{H}_\omega$.

Given a state, ω, and an automorphism, α, which preserves the state (i.e., $\omega \circ \alpha = \omega$) then there will be a unitary operator, U, on H_ω which *implements* α in the sense that $\rho_\omega(\alpha(A)) = U^{-1}\rho_\omega(A)U$ and U is chosen uniquely by the condition $U\Omega = \Omega$.

On a *stationary* spacetime, that is, one which admits a one-parameter group of isometries whose integral curves are everywhere timelike, the algebra will inherit a one-parameter group (i.e., satisfying $\alpha(t_1) \circ \alpha(t_2) = \alpha(t_1 + t_2)$) of time-translation

automorphisms, $\alpha(t)$, and, given any stationary state (i.e., one which satisfies $\omega \circ \alpha(t) = \omega \ \forall t \in \boldsymbol{R}$), these will be implemented by a one-parameter group of unitaries, $U(t)$, on its GNS Hilbert space satisfying $U(t)\Omega = \Omega$. If $U(t)$ is strongly continuous so that it takes the form e^{-iHt} and if the Hamiltonian, H, is positive, then ω is said to be a "ground state." Typically one expects *ground states* to exist and often be unique.

Another important class of stationary states for the algebra of a stationary spacetime is the class of *KMS states*, ω^β, at inverse temperature β; these have the physical interpretation of thermal equilibrium states. In the GNS representation of one of these, the automorphisms are also implemented by a strongly continuous unitary group, e^{-iHt}, which preserves Ω but (in place of H positive) there is a complex conjugation, J, on H_ω such that

$$e^{-\beta H/2}\rho_\omega(A)\Omega = J\rho_\omega(A^*)\Omega \qquad [1]$$

for all $A \in \mathcal{A}$. An attractive feature of the subject is that its main qualitative features are already present for linear field theories and, unusually in comparison with other questions in QFT, these are susceptible of a straightforward explicit and rigorous mathematical formulation. In fact, as our principal example, we give, in the next section a construction for the field algebra for the quantized real linear Klein–Gordon equation

$$(\Box_g - m^2 - V)\phi = 0 \qquad [2]$$

of mass m on a globally hyperbolic spacetime $(\mathcal{M}, g)$. Here, $\Box_g$ denotes the Laplace–Beltrami operator $g^{ab}\nabla_a\partial_b$ $(= (|\det(g)|)^{-1/2}\partial_a((|\det(g)|^{1/2}g^{ab}\partial_b))$. We include a scalar external background classical field, V, in addition to the external gravitational field represented by g. In case m is zero, taking V to equal $R/6$, where R denotes the Riemann scalar, makes the equation *conformally invariant*.

The main new feature of QFT in curved spacetime (present already for linear field theories) is that, in a general (neither flat nor stationary) spacetime there will not be any single preferred state but rather a family of preferred states, members of which are best regarded as on an equal footing with one another. It is this feature which makes the above algebraic framework particularly suitable, indeed essential, to a clear formulation of the subject. Conceptually, it is this feature which takes the most getting used to. In particular, one must realize that, as we shall explain later, the interpretation of a state as having a particular "particle content" is in general problematic because it can only be relative to a particular choice of "vacuum" state and, depending on the spacetime of interest, there may be one state or several states or, frequently, no states at all which deserve the name "vacuum" and even when there are states which deserve this name, they will often only be defined in some approximate or asymptotic or transient sense or only on some subregion of the spacetime.

Concomitantly, one does not expect global observables such as the "particle number" or the quantum Hamiltonian of flat-spacetime free-field theory to generalize to a curved spacetime context, and for this reason local observables play a central role in the theory. The quantized stress–energy tensor is a particularly natural and important such local observable and the theory of this is central to the whole subject. A brief introduction to it is given in a later section.

This is followed by a further section on the Hawking and Unruh effects and then a brief section on the problems of extending the theory beyond the "default" setting, to nonglobally hyperbolic spacetimes. Finally, we briefly mention a number of other interesting and active areas of the subject as well as issuing a few *warnings* to be borne in mind when reading the literature.

Construction of $*$-Algebra(s) for a Real Linear Scalar Field on Globally Hyperbolic Spacetimes and Some General Theorems

On a globally hyperbolic spacetime, the classical equation [2] admits well-defined *advanced* and *retarded Green* functions (strictly bidistributions) Δ^A and Δ^R and the standard covariant quantum free real (or "Hermitian") scalar field commutation relations familiar from Minkowski spacetime free-field theory naturally generalize to the (heuristic) equation

$$[\hat{\phi}(x), \hat{\phi}(y)] = i\Delta(x, y)I$$

where Δ is the *Lichnérowicz commutator function* $\Delta = \Delta^A - \Delta^R$. Here, the "$\hat{}$" on the quantum field $\hat{\phi}$ serves to distinguish it from a classical solution ϕ. In mathematical work, one does not assign a meaning to the field at a point itself, but rather aims to assign meaning to *smeared fields* $\hat{\phi}(F)$ for all real-valued test functions $F \in C_0^\infty(\mathcal{M})$ which are then to be interpreted as standing for $\int_{\mathcal{M}} \hat{\phi}(x)f(x)|\det(g)|^{1/2}\, d^4x$. In fact, it is straightforward to define a *minimal field algebra* (see below) $\mathcal{A}_{\min}$ generated by such $\hat{\phi}(F)$ which satisfy the suitably smeared version

$$[\hat{\phi}(F), \hat{\phi}(G)] = i\Delta(F, G)I$$

of the above commutation relations together with Hermiticity (i.e., $\hat{\phi}(F)^* = \hat{\phi}(F)$), the property of being a weak solution of eqn [2] (i.e., $\hat{\phi}((\Box_g - m^2 - V)F) = 0\ \forall F \in C_0^\infty(\mathcal{M})$) and linearity in test functions. There is a technically different alternative formulation of this minimal algebra, which is known as the *Weyl algebra*, which is constructed to be the C^*-algebra generated by operators $W(F)$ (to be interpreted as standing for $\exp(\mathrm{i}\int_{\mathcal{M}} \hat{\phi}(x)f(x)|\det(g)|^{1/2}\,\mathrm{d}^4x)$ satisfying

$$W(F_1)W(F_2) = \exp(-\mathrm{i}\Delta(F_1,F_2)/2)W(F_1+F_2)$$

together with $W(F)^* = W(-F)$ and $W((\Box_g - m^2 - V)F) = I$. With either the minimal algebra or the Weyl algebra one can define, for each bounded open region $\mathcal{O}$, subalgebras $A(\mathcal{O})$ as generated by the $\hat{\phi}(\cdot)$ (or the $W(\cdot)$) smeared with test functions supported in $\mathcal{O}$ and verify that they satisfy the above "net" condition and commutativity at spacelike separation.

Specifying a state, ω, on $\mathcal{A}_{\min}$ is tantamount to specifying its collection of n-point distributions (i.e., smeared n-point functions) $\omega(\hat{\phi}(F_1)\ldots\hat{\phi}(F_n))$. (In the case of the Weyl algebra, one restricts attention to "regular" states for which the map $F \to \omega(W(F))$ is sufficiently often differentiable on finite-dimensional subspaces of $C_0^\infty(\mathcal{M})$ and defines the n-point distributions in terms of derivatives with respect to suitable parameters of expectation values of suitable Weyl algebra elements.) A particular role is played in the theory by the *quasifree* states for which all the *truncated* n-point distributions except for $n=2$ vanish. Thus, all the n-point distributions for odd n vanish while the four-point distribution is made out of the two-point distribution according to

$$\begin{aligned}\omega(\hat{\phi}(F_1)\hat{\phi}(F_2)\hat{\phi}(F_3)\hat{\phi}(F_4)) &= \omega(\hat{\phi}(F_1)\hat{\phi}(F_2))\omega(\hat{\phi}(F_3)\hat{\phi}(F_4)) \\ &\quad + \omega(\hat{\phi}(F_1)\hat{\phi}(F_3))\omega(\hat{\phi}(F_2)\hat{\phi}(F_4)) \\ &\quad + \omega(\hat{\phi}(F_1)\hat{\phi}(F_4))\omega(\hat{\phi}(F_2)\hat{\phi}(F_3))\end{aligned}$$

etc. The anticommutator distribution

$$G(F_1,F_2) = \omega(\hat{\phi}(F_1)\hat{\phi}(F_2)) + \omega(\hat{\phi}(F_2)\hat{\phi}(F_1)) \qquad [3]$$

of a quasifree state (or indeed of any state) will satisfy the following conditions (for all test functions F, F_1, F_2, etc.):

C1. *Symmetry*

$$G(F_1,F_2) = G(F_2,F_1)$$

C2. *Weak bisolution property*

$$\begin{aligned}G((\Box_g - m^2 - V)F_1, F_2) &= 0 \\ &= G(F_1, (\Box_g - m^2 - V)F_2)\end{aligned}$$

C3. *Positivity*

$$\begin{aligned}G(F,F) \geq 0 \text{ and } G(F_1,F_1)^{1/2}G(F_2,F_2)^{1/2} \\ \geq |\Delta(F_1,F_2)|\end{aligned}$$

and it can be shown that, to every bilinear functional G on $C_0^\infty(\mathcal{M})$ satisfying (C1)–(C3), there is a quasifree state with two-point distribution $(1/2)(G + \mathrm{i}\Delta)$. One further declares a quasifree state to be *physically admissible* only if (for pairs of points in sufficiently small convex neighborhoods)

C4. *Hadamard condition*

$$\text{“}G(x_1,x_2) = \frac{1}{2\pi^2}\left(u(x_1,x_2)\mathrm{P}\frac{1}{\sigma} + v(x_1,x_2)\log|\sigma| + w(x_1,x_2)\right)\text{”}$$

This last condition expresses the requirement that (locally) the two-point distribution actually "is" (in the usual sense in which one says that a distribution "is" a function) a smooth function for pairs of non-null-separated points. At the same time, it requires that the two-point distribution be singular at pairs of null-separated points and locally specifies the nature of the singularity for such pairs of points with a leading "principal part of $1/\sigma$" type singularity and a subleading "$\log|\sigma|$" singularity, where σ denotes the square of the geodesic distance between x_1 and x_2. u (which satisfies $u(x_1,x_2) = 1$ when $x_1 = x_2$) and v are certain smooth two-point functions determined in terms of the local geometry and the local values of V by something called the *Hadamard procedure* while the smooth two-point function w depends on the state. We shall omit the details. The important point is that this Hadamard condition on the two-point distribution is believed to be the correct generalization to a curved spacetime of the well-known universal short-distance behavior shared by the truncated two-point distributions of all physically relevant states for the special case of our theory when the spacetime is flat (and V vanishes). In the latter case, u reduces to 1, and v to a simple power series $\sum_{n=0}^{\infty} v_n\sigma^n$ with $v_0 = m^2/4$, etc.

Actually, it is known (this is the content of "Kay's conjecture" which was proved by M Radzikowski in 1992) that (C1)–(C4) together imply that the two-point distribution is nonsingular at all pairs of (not necessarily close together) spacelike separated points. More important than this result itself is a reformulation of the Hadamard condition in terms of the concepts of *microlocal analysis* which Radzikowski originally introduced as a tool towards its proof.

C4′. *Wave front set* (*or microlocal*) *spectrum condition*

$$\begin{aligned}&\mathrm{WF}(G+\mathrm{i}\Delta)\\&\quad=\{(x_1,p_1;x_2,p_2)\in T^*(\mathcal{M}\times\mathcal{M})\setminus\mathbf{0}|x_1 \text{ and } \mathrm{x}_2\\&\qquad\text{lie on a single null geodesic, } p_1 \text{ is tangent to}\\&\qquad\text{that null geodesic and future pointing, and}\\&\qquad p_2 \text{ when parallel transported along that null}\\&\qquad\text{geodesic from } x_2 \text{ to } x_1 \text{ equals } -p_1\}\end{aligned}$$

For the gist of what this means, it suffices to know that to say that an element (x,p) of the cotangent bundle of a manifold (excluding the zero section $\mathbf{0}$) is in the *wave front set*, WF, of a given distribution on that manifold may be expressed informally by saying that that distribution is singular *at* the point x *in the direction* p. (And here the notion is applied to $G+\mathrm{i}\Delta$, thought of as a distribution on the manifold $\mathcal{M}\times\mathcal{M}$.)

We remark that generically (and, e.g., always if the spatial sections are compact and $m^2+V(x)$ is everywhere positive) the Weyl algebra for eqn [2] on a given stationary spacetime will have a unique ground state and unique KMS states at each temperature and these will be quasifree and Hadamard.

Quasifree states are important also because of a theorem of R Verch (1994, in verification of another conjecture of Kay) that (in the Weyl algebra framework) on the algebra of any bounded open region, the folia of the quasifree Hadamard states coincide. With this result one can extend the notion of physical admissibility to not-necessarily-quasifree states by demanding that, to be admissible, a state belong to the resulting common folium when restricted to the algebra of each bounded open region; equivalently, that it be a *locally normal* state on the resulting natural extension of the net of local Weyl algebras to a net of local W^*-algebras.

Particle Creation and the Limitations of the Particle Concept

Global hyperbolicity also entails that the Cauchy problem is well posed for the classical field equation [2] in the sense that for every Cauchy surface, $\mathcal{C}$, and every pair (f,p) of Cauchy data in $C_0^\infty(\mathcal{C})$, there exists a unique solution ϕ in $C_0^\infty(\mathcal{M})$ such that $f=\phi|_{\mathcal{C}}$ and $p=|\det(g)|^{1/2}g^{ab}\partial_b\phi|_{\mathcal{C}}$. Moreover, ϕ has compact support on all other Cauchy surfaces. Given a global time coordinate t, increasing towards the future, foliating $\mathcal{M}$ into a family of constant-t Cauchy surfaces, $\mathcal{C}_t$, and given a choice of global timelike vector field τ^a (e.g., $\tau^a=g^{ab}\partial_b t$) enabling one to identify all the $\mathcal{C}_t$, say with $\mathcal{C}_0$, by identifying points cut by the same integral curve of τ^a, a single such classical solution ϕ may be pictured as a family $\{(f_t,p_t): t\in\boldsymbol{R}\}$ of time-evolving Cauchy data on $\mathcal{C}_0$. Moreover, since [2] implies, for each pair of classical solutions, ϕ_1,ϕ_2, the conservation (i.e., $\partial_a j^a=0$) of the current $j^a=|\det(g)|^{1/2}g^{ab}(\phi_1\partial_b\phi_2-\phi_2\partial_b\phi_1)$, the symplectic form (on $C_0^\infty(\mathcal{C})\times C_0^\infty(\mathcal{C})$)

$$\sigma((f_t^1,p_t^1);(f_t^2,p_t^2))=\int_{\mathcal{C}_0}(f_t^1p_t^2-p_t^1f_t^2)\mathrm{d}^3x$$

will be conserved in time.

Corresponding to this picture of classical dynamics, one expects there to be a description of quantum dynamics in terms of a family of sharp-time quantum fields (φ_t,π_t) on $\mathcal{C}_0$, satisfying heuristic canonical commutation relations

$$\begin{aligned}[\varphi_t(\boldsymbol{x}),\varphi_t(\boldsymbol{y})]&=0\\[\pi_t(\boldsymbol{x}),\pi_t(\boldsymbol{y})]&=0\\[\varphi_t(\boldsymbol{x}),\pi_t(\boldsymbol{y})]&=\mathrm{i}\delta^3(\boldsymbol{x},\boldsymbol{y})I\end{aligned}$$

and evolving in time according to the same dynamics as the Cauchy data of a classical solution. (Both these expectations are correct because the field equation is linear.) An elegant way to make rigorous mathematical sense of these expectations is in terms of a $*$-algebra with identity generated by Hermitian objects "$\sigma((\varphi_0,\pi_0);(f,p))$" ("symplectically smeared sharp-time fields at $t=0$") satisfying linearity in f and p together with the commutation relations

$$\begin{aligned}&[\sigma((\varphi_0,\pi_0);(f^1,p^1)),\sigma((\varphi_0,\pi_0);(f^2,p^2))]\\&\quad=\mathrm{i}\sigma((f^1,p^1);(f^2,p^2))I\end{aligned}$$

and to define (symplectically smeared) time-t sharp-time fields by demanding

$$\sigma((\varphi_t,\pi_t);(f_t,p_t))=\sigma((\varphi_0,\pi_0);(f_0,p_0))$$

where (f_t,p_t) is the classical time-evolute of (f_0,p_0). This $*$-algebra of sharp-time fields may be identified with the (minimal) field $*$-algebra of the previous section, the $\hat{\phi}(F)$ of the previous section being identified with $\sigma((\varphi_0,\pi_0);(f,p))$, where (f,p) are the Cauchy data at $t=0$ of $\Delta*F$. (This identification is of course many–one since $\hat{\phi}(F)=0$ whenever F arises as $(\Box_g-m^2-V)G$ for some test function $G\in C_0^\infty(\mathcal{M})$.)

Specializing momentarily to the case of the free scalar field $(\Box-m^2)\phi=0$ $(m\neq 0)$ in Minkowski space with a flat $t=0$ Cauchy surface, the "symplectically smeared" two-point function of the usual ground state ("Minkowski vacuum state"), ω_0, is given, in this formalism, by

$$\begin{aligned}&\omega_0(\sigma((\varphi,\pi);(f^1,p^1))\sigma((\varphi,\pi);(f^2,p^2)))\\&\quad=\tfrac{1}{2}(\langle f^1|\mu f^2\rangle+\langle p^1|\mu^{-1}p^2\rangle\\&\qquad+\mathrm{i}\sigma((f^1,p^1);(f^2,p^2)))\end{aligned}\qquad[4]$$

where the inner products are in the *one-particle Hilbert space* $\mathcal{H} = L^2_C(\boldsymbol{R}^3)$ and $\mu = (m^2 - \nabla^2)^{1/2}$. The GNS representation of this state may be concretely realized on the familiar *Fock space* $\mathcal{F}(\mathcal{H})$ over $\mathcal{H}$ by

$$\rho_0(\sigma((\varphi,\pi);(f,p))) = -\mathrm{i}(\hat{a}^\dagger(\mathrm{a}) - (\hat{a}^\dagger(\mathrm{a}))^*)$$

where a denotes the element of $\mathcal{H}$:

$$\mathrm{a} = \frac{(\mu^{1/2} f + \mathrm{i}\mu^{-1/2} p)}{\sqrt{2}}$$

(we note in passing that, if we equip $\mathcal{H}$ with the symplectic form $2\,\mathrm{Im}\langle\cdot|\cdot\rangle$, then $K:(f,p)\mapsto \mathrm{a}$ is a symplectic map) and $\hat{a}^\dagger(\mathrm{a})$ is the usual smeared *creation operator* ($=$ "$\int \hat{a}^\dagger(\boldsymbol{x})\mathrm{a}(\boldsymbol{x})\mathrm{d}^3\boldsymbol{x}$") on $\mathcal{F}(\mathcal{H})$ satisfying

$$[(\hat{a}^\dagger(\mathrm{a}^1))^*, \hat{a}^\dagger(\mathrm{a}^2)] = \langle \mathrm{a}^1|\mathrm{a}^2\rangle_{\mathcal{H}} I$$

The usual (smeared) *annihilation operator*, $\hat{a}(\mathrm{a})$, is $(\hat{a}^\dagger(C\mathrm{a}))^*$, where C is the natural complex conjugation, $\mathrm{a}\mapsto \mathrm{a}^*$ on $\mathcal{H}$. Both of these operators annihilate the *Fock vacuum vector* $\Omega^{\mathcal{F}}$. In this representation, the one-parameter group of time-translation automorphisms

$$\alpha(t):\sigma((\varphi_0,\pi_0);(f,p))\mapsto \sigma((\varphi_t,\pi_t);(f,p)) \quad [5]$$

is implemented by $\exp(-\mathrm{i}Ht)$ where H is the second quantization of μ (i.e., the operator otherwise known as $\int \mu(k)\hat{a}^\dagger(k)\hat{a}(k)\mathrm{d}^3k$) on $\mathcal{F}(\mathcal{H})$.

The most straightforward (albeit physically artificial) situation involving "particle creation" in a curved spacetime concerns a globally hyperbolic spacetime which, outside of a compact region, is isometric to Minkowski space with a compact region removed – that is, to a globally hyperbolic spacetime which is flat except inside a localized "bump" of curvature (see **Figure 1**). (One could also allow the function V in [2] to be nonzero inside the bump.) On the field algebra (defined as in the previous section) of such a spacetime, there will be an "in" vacuum state (which may be identified with the Minkowski vacuum to the past of the bump) and an "out" vacuum state (which may be identified with the Minkowski vacuum to the future of the bump) and one expects, for example, the "in vacuum" to arise as a many-particle state in the GNS representation of the "out vacuum" corresponding to the creation of particles out of the vacuum by the bump of curvature.

In the formalism of this section, if we choose our global time coordinate on such a spacetime so that, say, the $t=0$ surface is to the past of the bump and the $t=T$ surface to its future, then the single automorphism $\alpha(T)$ (defined as in [5]) encodes the overall effect of the bump of curvature on the quantum field and one can ask whether it is implemented by a unitary operator in the GNS representation of the Minkowski vacuum state [4].

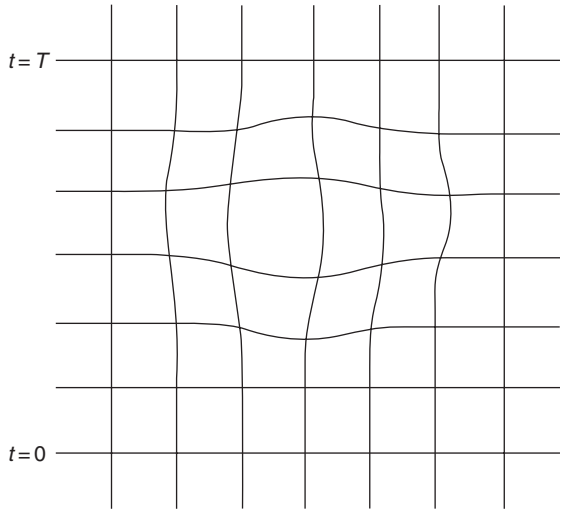


Figure 1 A spacetime which is flat outside of a compact bump of curvature.

This question may be answered by referring to the real linear map $\mathcal{T}:\mathcal{H}\to\mathcal{H}$ which sends $\mathrm{a}_T = 2^{-1/2}(\mu^{1/2} f_T + \mathrm{i}\mu^{-1/2} p_T)$ to $\mathrm{a}_0 = 2^{-1/2}(\mu^{1/2} f_0 + \mathrm{i}\mu^{-1/2} p_0)$. By the conservation in time of σ and the symplecticity, noted in passing above, of the map $K:(f,p)\mapsto \mathrm{a}$, this satisfies the defining relation

$$\mathrm{Im}\langle \mathcal{T}\mathrm{a}^1|\mathcal{T}\mathrm{a}^2\rangle = \mathrm{Im}\langle \mathrm{a}^1|\mathrm{a}^2\rangle$$

of a *classical Bogoliubov transformation*. Splitting $\mathcal{T}$ into its complex-linear and complex-antilinear parts by writing

$$\mathcal{T} = \alpha + \beta C$$

where α and β are complex-linear operators, this relation may alternatively be expressed in terms of the pair of relations

$$\alpha^*\alpha - \bar{\beta}^*\bar{\beta} = I, \qquad \bar{\alpha}^*\bar{\beta} = \beta^*\alpha$$

where $\bar{\alpha} = C\alpha C$, $\bar{\beta} = C\beta C$.

We remark that there is an easy-to-visualize equivalent way of defining α and β in terms of the analysis, to the past of the bump, into *positive- and negative-frequency parts* of complex solutions to [2] which are purely *positive frequency* to the future of the bump. In fact, if, for any element $\mathrm{a}\in\mathcal{H}$, we identify the positive-frequency solution to the Minkowski-space Klein–Gordon equation

$$\phi^{\mathrm{pos}}_{\mathrm{out}}(t,\boldsymbol{x}) = ((2\mu)^{-1/2}\exp(-\mathrm{i}\mu t)\mathrm{a})(\boldsymbol{x})$$

with a complex solution to [2] to the future of the bump, then (it may easily be seen) to the past of the bump, this same solution will be identifiable with

the (partly positive-frequency, partly negative-frequency) Minkowski-space Klein–Gordon solution

$$\phi_{\text{in}}(t,\boldsymbol{x}) = \left((2\mu)^{-1/2}\exp(-\mathrm{i}\mu t)\alpha\mathrm{a}\right)(\boldsymbol{x}) + \left((2\mu)^{-1/2}\exp(\mathrm{i}\mu t)\bar{\beta}\mathrm{a}\right)(\boldsymbol{x})$$

and this could be taken to be the defining equation for the operators α and β.

It is then known (by a 1962 theorem of Shale) that the automorphism [5] (strictly, its Weyl algebra counterpart) will be unitarily implemented if and only if β is a Hilbert–Schmidt operator on $\mathcal{H}$. Wald (1979, in case $m \geq 0$) and Dimock (1979, in case $m \neq 0$) have verified that this condition is satisfied in the case of our bump-of-curvature situation. In that case, if we denote the unitary implementor by U, we have the following results:

R1. The expectation value $\langle U\Omega|N(\mathrm{a})U\Omega\rangle_{\mathcal{F}(\mathcal{H})}$ of the number operator, $N(\mathrm{a}) = \hat{a}^\dagger(\mathrm{a})\hat{a}(\mathrm{a})$, where a is a normalized element of $\mathcal{H}$, is equal to $\langle\beta\mathrm{a}\,|\,\beta\mathrm{a}\rangle_{\mathcal{H}}$.

R2. First note that there exists an orthonormal basis of vectors, $e_i, (i = 1\ldots\infty)$, in $\mathcal{H}$ such that the (Hilbert–Schmidt) operator $\bar{\beta}^*\bar{\alpha}^{*-1}$ has the canonical form $\sum_i \lambda_i\langle Ce_i|\cdot\rangle|e_i\rangle$. We then have (up to an undetermined phase)

$$U\Omega = N\exp\left(-\frac{1}{2}\sum_i \lambda_i\hat{a}^\dagger(e_i)\hat{a}^\dagger(e_i)\right)\Omega$$

where the normalization constant N is chosen so that $\|U\Omega\| = 1$. This formula makes manifest that the particles are created in pairs.

We remark that, identifying elements, a, of $\mathcal{H}$ with positive-frequency solutions (below, we shall call them "modes") as explained above, result (R1) may alternatively be expressed by saying that the expectation value, $\omega_{\text{in}}(N(\mathrm{a}))$, in the *in-vacuum state* of the occupation number, $N(\mathrm{a})$, of a *normalized mode*, a, to the future of the bump, is given by $\langle\beta\mathrm{a}|\beta\mathrm{a}\rangle_{\mathcal{H}}$.

This formalism and the results, (R1) and (R2) above, will generalize (at least heuristically, and sometimes rigorously – see especially the rigorous scattering-theoretic work in the 1980s by Dimock and Kay and more recently by A Bachelot and others) to more realistic spacetimes which are only asymptotically flat or asymptotically stationary. In favorable cases, one will still have notions of classical solutions which are positive frequency asymptotically towards the future/past, and, in consequence, one will have well-defined asymptotic notions of "vacuum" and "particles." Also, in, for example, cosmological, models where the background spacetime is slowly varying in time, one can define approximate *adiabatic* notions of classical positive-frequency solutions, and hence also of quantum "vacuum" and "particles" at each finite value of the cosmological time. But, at times where the gravitational field is rapidly varying, one does not expect there to be any sensible notion of "particles." And, in a rapidly time-varying background gravitational field which never settles down, one does not expect there to be any sensible particle interpretation of the theory at all. To understand these statements, it suffices to consider the $(1+0)$-dimensional Klein–Gordon equation with an external potential V:

$$\left(-\frac{\mathrm{d}^2}{\mathrm{d}t^2} - m^2 - V(t)\right)\phi = 0$$

which is of course a system of one degree of freedom, mathematically equivalent to the harmonic oscillator with a time-varying angular frequency $\varpi(t) = (m^2 + V(t))^{1/2}$. One could of course express its quantum theory in terms of a time-evolving Schrödinger wave function $\Psi(\varphi, t)$ and attempt to give this a particle interpretation at each time, s, by expanding $\Psi(\varphi, s)$ in terms of the harmonic oscillator wave functions for a harmonic oscillator with some particular choice of angular frequency. But the problem is, as is easy to convince oneself, that there is no such good choice. For example, one might think that a good choice would be to take, at time s, the set of harmonic oscillator wave functions with angular frequency $\varpi(s)$. (This is sometimes known as the method of "instantaneous diagonalization of the Hamiltonian.") But suppose we were to apply this prescription to the case of a smooth $V(\cdot)$ which is constant in time until time 0 and assume the initial state is the usual vacuum state. Then at some positive time s, the number of particles predicted to be present is the same as the number of particles predicted to be present on the same prescription at all times after s for a $\hat{V}(\cdot)$ which is equal to $V(\cdot)$ up to time s and then takes the constant value $V(s)$ for all later times (see Figure 2). But $\hat{V}(\cdot)$ will generically have a sharp corner in its graph (i.e., a

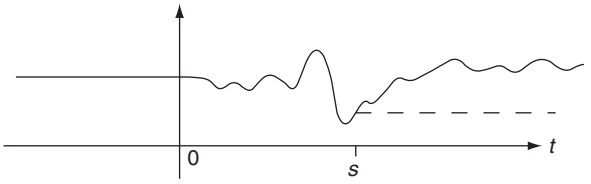


Figure 2 Plots of ϖ against t for the two potentials V (continuous line) and $\hat{V}$ (continuous line upto s and then dashed line) which play a role in our critique of "instantaneous diagonalization."

discontinuity in its time derivative) at time s, and one would expect a large part of the particle production in the latter situation to be accounted for by the presence of this sharp corner – and therefore a large part of the predicted particle production in the case of $V(\cdot)$ to be spurious.

Back in $1+3$ dimensions, even where a good notion of particles is possible, it depends on the choice of time evolution, as is dramatically illustrated by the Unruh effect discussed in the relevant section.

Theory of the Stress–Energy Tensor

To orient ideas, consider first the free (minimally coupled) scalar field, $(\Box - m^2)\phi = 0$, in Minkowski space. If one quantizes this system in the usual Minkowski-vacuum representation, then the expectation value of the *renormalized stress-energy tensor* (which in this case is the same thing as the *normal ordered stress–energy tensor*) in a vector state Ψ in the Fock space will be given by the formal *point-splitting* expression

$$\begin{aligned}&\langle\Psi|T_{ab}(x)\Psi\rangle\\ &\quad=\lim_{(x_1,x_2)\to(x,x)}\left(\partial^1_a\partial^2_b-\tfrac{1}{2}\eta_{ab}(\eta^{cd}\partial^1_c\partial^2_d+m^2)\right)\\ &\qquad\times(\langle\Psi|\rho_0(\phi(x_1)\phi(x_2))\Psi\rangle\\ &\qquad-\langle\Omega^{\mathcal{F}}|\rho_0(\phi(x_1)\phi(x_2))\Omega^{\mathcal{F}}\rangle)\end{aligned}\qquad[6]$$

where η_{ab} is the usual Minkowski metric. A sufficient condition for the limit here to be finite and well defined would, for example, be for Ψ to consist of a (normalized) finite superposition of n-particle vectors of form $\hat{a}^\dagger(\mathrm{a}_1),\ldots,\hat{a}^\dagger(\mathrm{a}_n)\Omega^{\mathcal{F}}$ where the smearing functions $\mathrm{a}_1,\ldots,\mathrm{a}_n$ are all C^∞ elements of $\mathcal{H}$ (i.e., of $L^2_C(\boldsymbol{R}^3)$. The reason this works is that the two-point function in such states shares the same short-distance singularity as the Minkowski-vacuum two-point function. For exactly the same reason, one obtains a well-defined finite limit if one defines the expectation value of the stress–energy tensor in any physically admissible quasifree state by the expression

$$\begin{aligned}&\omega(T_{ab}(x))\\ &\quad=\lim_{(x_1,x_2)\to(x,x)}\left(\partial^1_a\partial^2_b-\tfrac{1}{2}\eta_{ab}(\eta^{cd}\partial^1_c\partial^2_d+m^2)\right)\\ &\qquad\times(\omega(\phi(x_1)\phi(x_2))-\omega_0(\phi(x_1)\phi(x_2)))\end{aligned}\qquad[7]$$

This latter point-splitting formula generalizes to a definition for the *expectation value of the renormalized stress–energy tensor* for an arbitrary physically admissible quasifree state (or indeed for an arbitrary state whose two-point function has *Hadamard form* – i.e., whose anticommutator function satisfies condition (C4)) on the minimal field algebra and to other linear field theories (including the stress tensor for a conformally coupled linear scalar field) on a general globally hyperbolic spacetime (and the result obtained agrees with that obtained by other methods, including *dimensional regularization* and *zeta-function regularization*). However, the generalization to a curved spacetime involves a number of important new features which we now briefly list (see Wald (1978) for details).

First, the subtraction term which replaces $\omega_0(\phi(x_1)\phi(x_2))$ is, in general, not the expectation value of $\phi(x_1)\phi(x_2)$ in any particular state, but rather a particular *locally constructed Hadamard two-point function* whose physical interpretation is more subtle; the renormalization is thus in general not to be regarded as a normal ordering. Second, the immediate result of the resulting limiting process will not be covariantly conserved and, in order to obtain a covariantly conserved quantity, one needs to add a particular *local geometrical correction term*. The upshot of this is that the resulting expected stress–energy tensor is covariantly conserved but possesses a (state-independent) *anomalous trace*. In particular, for a massless conformally coupled linear scalar field, one has (for all physically admissible quasifree states, ω) the *trace anomaly formula*

$$\omega(T^a_a(x))=(2880\pi^2)^{-1}\left(C_{abcd}C^{abcd}+R_{ab}R^{ab}-\tfrac{1}{3}R^2\right)$$

plus an arbitrary multiple of $\Box R$. In fact, in general, the thus-defined renormalized stress–energy tensor operator (see below) is only defined up to a *finite renormalization ambiguity* which consists of the addition of arbitrary multiples of the functional derivatives with respect to g_{ab} of the quantities

$$I_n=\int_{\mathcal{M}}F_n(x)|\det(g)|^{1/2}\mathrm{d}^4x$$

where n ranges from 1 to 4 with $F_1=1$, $F_2=R$, $F_3=R^2$, and $F_4=R_{ab}R^{ab}$. In the Minkowski-space case, only the first of these ambiguities arises and it is implicitly resolved in the formulas [6], [7] inasmuch as these effectively incorporate the *renormalization condition* that $\omega_0(T_{ab})=0$. (For the same reason, the locally flat example we give below has no ambiguity.)

One expects, in both flat and curved cases, that, for test functions, $F\in C^\infty_0(\mathcal{M})$, there will exist operators $T_{ab}(F)$ which are *affiliated to* the net of

local W^*-algebras referred to earlier and that it is meaningful to write

$$\int_{\mathcal{M}} \omega(T_{ab}(x))F(x)|\det(g)|^{1/2}\mathrm{d}^4x = \omega(T_{ab}(F))$$

provided that, by ω on the right-hand side, we understand the extension of ω from the Weyl algebra to this net. ($T_{ab}(F)$ is however not expected to belong to the minimal algebra or be affiliated to the Weyl algebra.)

An interesting simple example of a renormalized stress–energy tensor calculation is the so-called *Casimir effect* calculation for a linear scalar field on a (for further simplicity, $(1+1)$-dimensional) timelike cylinder spacetime of radius R (see Figure 3). This spacetime is globally hyperbolic and stationary and, while locally flat, globally distinct from Minkowski space. As a result, while – provided the regions $\mathcal{O}$ are sufficiently small (such as the diamond region in Figure 3) – elements $\mathcal{A}(\mathcal{O})$ of the minimal net of local algebras on this spacetime will be identifiable, in an obvious way, with elements of the minimal net of local algebras on Minkowski space, the stationary ground state ω_{cylinder} will, when restricted to such thus-identified regions, be distinct from the Minkowski vacuum state ω_0. The resulting renormalized stress–energy tensor (as first pointed out in Kay (1979)), definable, once the above identification has been made, exactly as in [7]) turns out, in the massless case, to be nonzero and, interestingly, to have a (in the natural coordinates, constant) *negative energy-density* T_{00}. In fact, in this massless case,

$$\omega_{\text{cylinder}}(T_{ab}) = \frac{1}{24\pi R^2}\eta_{ab}$$

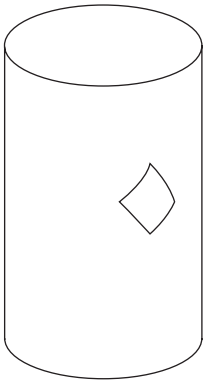


Figure 3 The timelike cylinder spacetime of radius R with a diamond region isometric to a piece of Minkowski space. See Kay (1979). Casimir effect in quantum field theory. (Original title: The Casimir effect without magic.) *Physical Review D* 20: 3052–3062. Reprinted with permission © 1979 by the American Physical Society.

Hawking and Unruh Effects

The original calculation by Hawking (1975) concerned a model spacetime for a star which collapses to a black hole. For simplicity, we shall only discuss the spherically symmetric case (see Figure 4). Adopting a similar "mode" viewpoint to that mentioned after results (R1) and (R2) discussed earlier, the result of the calculation may be stated as follows: For a real linear scalar field satisfying [2] with $m=0$ (and $V=0$) on this spacetime, the expectation value $\omega_{\text{in}}(N(\mathrm{a}_{\varpi,\ell}))$ of the occupation number of a one-particle outgoing mode $\mathrm{a}_{\varpi,\ell}$) localized (as far as a normalized mode can be) around ϖ in angular-frequency space and about retarded time v, and with angular momentum "quantum number" ℓ, in the in-vacuum state (i.e., on the minimal algebra for a real scalar field on this model spacetime) ω_{in} is, at late retarded times, given by the formula

$$\omega_{\text{in}}(N(\mathrm{a}_{\varpi,\ell})) = \frac{\Gamma(\varpi,\ell)}{\exp(8\pi M\varpi)-1}$$

where M is the mass of the black hole and the *absorption factor* (alternatively known as *gray-body factor*) $\Gamma(\varpi,\ell)$ is equal to the norm-squared of that part of the one-particle mode $\mathrm{a}_{\varpi,\ell}$ which, viewed as a complex positive-frequency classical solution propagating backwards in time from late retarded times, would be absorbed by the black hole. (Note the independence of the right-hand side of this formula from the retarded time, v.) This calculation can be understood as an application of result (R1)

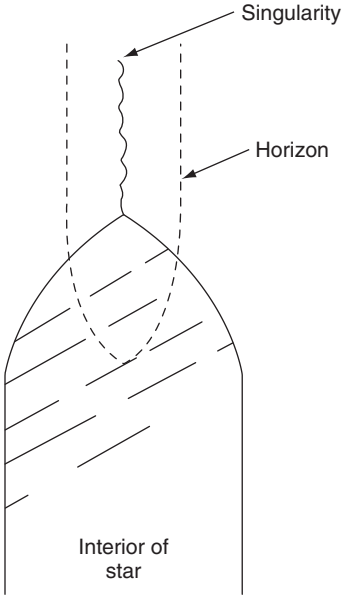


Figure 4 The spacetime of a star collapsing to a spherical black hole.

(even though the spacetime is more complicated than one with a localized "bump of curvature" and even though the relevant overall time evolution will not be unitarily implemented, the result still applies when suitably interpreted) and the heart of the calculation is an asymptotic estimate of the relevant "β" Bogoliubov coefficient which turns out to be dependent on the geometrical optics of rays which pass through the star just before the formation of the horizon. This result suggests that the in-vacuum state is indistinguishable at late retarded times from a state of blackbody radiation at the *Hawking temperature*, $T_{\text{Hawking}} = 1/8\pi M$, in Minkowski space from a blackbody (gray body) with the same absorption factor. This was confirmed by further work by many authors. Much of that work, as well as the original result of Hawking was partially heuristic but later work by Dimock and Kay (1987), by Fredenhagen and Haag (1990), and by Bachelot (1999) and others has put different aspects of it on a rigorous mathematical footing. The result generalizes to nonzero mass and higher spin fields to interacting fields as well as to other types of black hole and the formula for the Hawking temperature generalizes to

$$T_{\text{Hawking}} = \kappa/2\pi$$

where κ is the *surface gravity* of the black hole.

This result suggests that there is something fundamentally "thermal" about quantum fields on black-hole backgrounds and this is confirmed by a number of mathematical results. In particular, the theorems in the two papers Kay and Wald (1991) and Kay (1993), combined together, tell us that there is a unique state on the Weyl algebra for the *maximally extended Schwarzschild spacetime* (a.k.a. *Kruskal–Szekeres spacetime*) (see **Figure 5**) which is invariant under the *Schwarzschild isometry group* and whose two-point function has Hadamard form. Moreover, they tell us that this state, when restricted to a single wedge (i.e., the exterior Schwarzschild spacetime) is necessarily a KMS state at the Hawking temperature. This unique state is known as the *Hartle–Hawking–Israel state*. These results in fact apply more generally to a wide class of globally hyperbolic spacetimes with *bifurcate Killing horizons* including de Sitter space – where the unique state is sometimes called the *Euclidean* and sometimes the *Bunch–Davies* vacuum state – as well as to Minkowski space, in which case the unique state is the usual Minkowski vacuum state, the analog of the exterior Schwarzschild wedge is a so-called *Rindler wedge*, and the relevant isometry group is a one-parameter family of wedge-preserving Lorentz boosts. In the latter situation, the fact that the Minkowski vacuum state is a KMS state (at "temperature" $1/2\pi$)

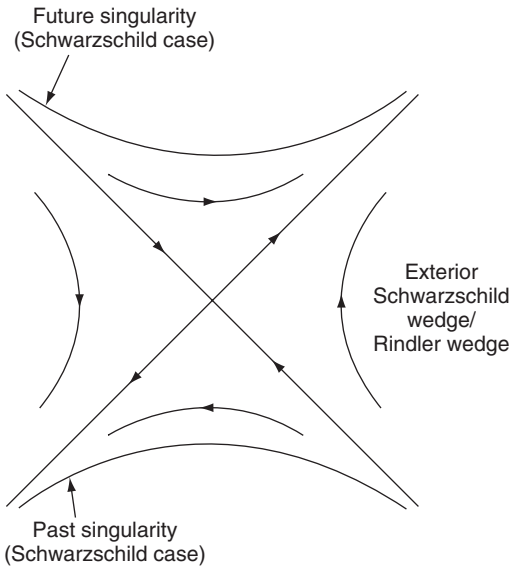


Figure 5 The geometry of maximally extended Schwarzschild (/or Minkowski) spacetime. In the Schwarzschild case, every point represents a 2-sphere (/in the Minkowski case, a 2-plane). The curves with arrows on them indicate the Schwarzschild time evolution (/one-parameter family of Lorentz boosts). These curves include the (straight lines at right angles) event horizons (/Killing horizons).

when restricted to a Rindler wedge and regarded with respect to the time evolution consisting of the wedge-preserving one-parameter family of Lorentz boosts is known as the *Unruh effect* (1975). This latter property of the Minkowski vacuum in fact generalizes to general *Wightman QFTs* and is in fact an immediate consequence of a combination of the *Reeh–Schlieder theorem* (applied to a Rindler wedge) and the *Bisognano–Wichmann theorem* (1975). The latter theorem says that the defining relation [1] of a KMS state holds if, in [1], we identify the operator J with the complex conjugation which implements wedge reflection and H with the self-adjoint generator of the unitary implementor of Lorentz boosts. We remark that the Unruh effect illustrates how the concept of "vacuum" (when meaningful at all) is dependent on the choice of time evolution under consideration. Thus, the usual Minkowski vacuum is a ground state with respect to the usual Minkowski time evolution but not (when restricted to a Rindler wedge) with respect to a one-parameter family of Lorentz boosts; with respect to these, it is, instead, a KMS state.

Nonglobally Hyperbolic Spacetimes and the "Time Machine" Question

Hawking (1992) argued that a spacetime in which a time machine gets manufactured should be modeled (see **Figure 6**) by a spacetime with an initial *globally*

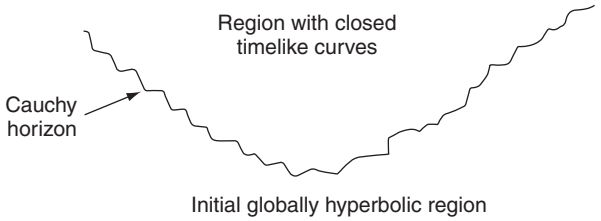


Figure 6 The schematic geometry of a spacetime in which a time machine gets manufactured.

hyperbolic region with a region containing closed timelike curves to its future and such that the future boundary of the globally hyperbolic region is a *compactly generated Cauchy horizon*. On such a spacetime, Kay *et al.* (1997) proved that it is impossible for any distributional bisolution which satisfies (even a certain weakened version of) the Hadamard condition on the initial globally hyperbolic region to continue to satisfy that condition on the full spacetime – the (weakened) Hadamard condition being necessarily violated at at least one point *on* the Cauchy horizon. This result implies that, however one extends a state, satisfying our conditions (C1)–(C4), on the minimal algebra for [2] on the initial globally hyperbolic region, the expectation value of its stress–energy tensor must necessarily become singular on the Cauchy horizon. This result, together with many heuristic results and specific examples considered by many other authors appears to support the validity of the (Hawking 1992) *chronology protection conjecture* to the effect that it is impossible in principle to manufacture a time machine. However, there are potential loopholes in the physical interpretation of this result as pointed out by Visser (1997), as well as other claims by various authors that one can nevertheless violate the chronology protection conjecture. For a recent discussion on this question, we refer to Visser (2003).

Other Related Topics and Some Warnings

There is a vast computational literature, calculating the expectation values of stress–energy tensors in states of interest for scalar and higher spin linear fields (and also some work for interacting fields) on interesting cosmological and black-hole backgrounds. QFT on de Sitter and anti-de Sitter space is a big subject area in its own right with recent renewed interest because of its relevance to *string theory* and *holography*. Also important on black-hole backgrounds is the calculation of gray-body factors, again with renewed interest because of relevance to string theory and to *brane-world* scenarios.

There are many further mathematically rigorous results on algebraic and axiomatic QFT in a curved spacetime setting, including versions of *PCT*, *spin-statistics* and *Reeh–Schlieder* theorems and also rigorous *energy inequalities* bounding the extent to which expected energy densities can be negative, etc.

There is much mathematical work controlling scattering theory on black holes, partly with a view to further elucidating the Hawking effect.

Perturbative renormalization theory of interacting quantum fields in curved spacetime is also now a highly developed subject.

Beyond QFT in a fixed curved spacetime is *semiclassical gravity* which takes into account the *back-reaction* of the expectation value of the stress–energy tensor on the classical gravitational background. There are also interesting condensed matter analogs of the Hawking effect such as *dumb holes*.

Readers exploring the wider literature, or doing further research on the subject should be aware that the word "vacuum" is sometimes used to mean "ground state" and sometimes just to mean "quasifree state." They should be cautious of attempts to define particles on Cauchy surfaces in *instantaneous diagonalization* schemes (cf. the remarks at the end of the section "Particle creation and the limitations of the particle concept"). When studying (or performing) calculations of the "expectation value of the stress–energy tensor" it is always important to ask oneself with respect to which state the expectation value is being taken. It is also important to remember to check that candidate two-point (anticommutator) functions satisfy the positivity condition (C3) discussed earlier. Typically, two-point distributions obtained via *mode sums* automatically satisfy condition (C3) (and condition (C4)), but those obtained via *image* methods do not always satisfy it. (When they do not, the presence of nonlocal spacelike singularities is often a tell-tale sign as can be inferred from Kay's conjecture/Radzikowski's theorem discussed earlier.) There are a number of apparent implicit assertions in the literature that some such two-point functions arise from "states" when of course they cannot. Some of these concern proposed analogs to the Hartle–Hawking–Israel state for the (appropriate maximal globally hyperbolic portion of the maximally extended) Kerr spacetime. That they cannot belong to states is clear from a theorem in Kay and Wald (1991) which states that there is no stationary Hadamard state on this spacetime at all. Others of them concern claimed "states" on spacetimes such as those discussed in the previous section which, if they really were states would seem to be in conflict with the chronology protection conjecture. Finally, beware states (such as the so-called α-*vacua* of de Sitter spacetime) whose two-point

distributions violate the "Hadamard" condition (C4) and which therefore do not have a well-defined finite expectation value for the renormalized stress–energy tensor.

See also: AdS/CFT Correspondence; Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Black Hole Mechanics; Bosons and Fermions in External Fields; Integrability and Quantum Field Theory; Quantum Fields with Indefinite Metric: Non-Trivial Models; Quantum Fields with Topological Defects; Quantum Geometry and Its Applications; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools; Thermal Quantum Field Theory.

Further Reading

Birrell ND and Davies PCW (1982) *Quantum Fields in Curved Space*. Cambridge: Cambridge University Press.

Brunetti R, Fredenhagen K, and Verch R (2003) The generally covariant locality principle – a new paradigm for local quantum physics. *Communications in Mathematical Physics* 237: 31–68.

DeWitt BS (1975) Quantum field theory in curved space-time. *Physics Reports* 19(6): 295–357.

Dimock J (1980) Algebras of local observables on a manifold. *Communications in Mathematical Physics* 77: 219–228.

Haag R (1996) *Local Quantum Physics*. Berlin: Springer.

Hartle JB and Hawking SW (1976) Path-integral derivation of black-hole radiance. *Physical Review D* 13: 2188–2203.

Hawking SW (1975) Particle creation by black holes. *Communications in Mathematical Physics* 43: 199–220.

Hawking SW (1992) The chronology protection conjecture. *Physical Review D* 46: 603–611.

Israel W (1976) Thermo-field dynamics of black holes. *Physics Letters A* 57: 107–110.

Kay BS (1979) Casimir effect in quantum field theory. (Original title: The Casimir effect without magic.) *Physical Review D* 20: 3052–3062.

Kay BS (1993) Sufficient conditions for quasifree states and an improved uniqueness theorem for quantum fields on space-times with horizons. *Journal of Mathematical Physics* 34: 4519–4539.

Kay BS (2000) Application of linear hyperbolic PDE to linear quantum fields in curved spacetimes: especially black holes, time machines and a new semi-local vacuum concept. *Journées Équations aux Dérivées Partielles*, Nantes, 5–9 juin 2000, GDR 1151 (CNRS): IX1–IX19. (Also available at http://www.math.sciences.univ-nantes.fr or as gr-qc/0103056.)

Kay BS, Radzikowski MJ, and Wald RM (1997) Quantum field theory on spacetimes with a compactly generated Cauchy horizon. *Communications in Mathematical Physics* 183: 533–556.

Kay BS and Wald RM (1991) Theorems on the uniqueness and thermal properties of stationary, nonsingular, quasifree states on spacetimes with a bifurcate Killing horizon. *Physics Reports* 207(2): 49–136.

Misner CW, Thorne KS, and Wheeler JA (1973) *Gravitation*. San Francisco: W.H. Freeman.

Unruh W (1976) Notes on black hole evaporation. *Physical Review D* 14: 870–892.

Visser M (2003) The quantum physics of chronology protection. In: Gibbons GW, Shellard EPS, and Rankin SJ (eds.) *The Future of Theoretical Physics and Cosmology*. Cambridge: Cambridge University Press.

Wald RM (1978) Trace anomaly of a conformally invariant quantum field in a curved spacetime. *Physical Review D* 17: 1477–1484.

Wald RM (1994) *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics*. Chicago: University of Chicago Press.

Quantum Field Theory: A Brief Introduction

L H Ryder, University of Kent, Canterbury, UK

By any account quantum field theory occupies a prominent place in the history of mathematical physics. This article is, however, not intended to serve as an overview of this subject, but has the more modest aim of identifying a few areas which seem to me interesting and significant.

Historical Remarks; Second Quantization

At the time when quantum field theory was at the forefront of theoretical physics its *raison d'être* was to complete the quantum description of the subatomic world. Quantum mechanics had been amazingly successful in solving almost the whole of atomic physics by making explicit the quantum (wave) nature of the electron, according to the formulations of Heisenberg and Schrödinger. The introduction of the quantum idea into physics, however, by Planck in 1900 closely followed by Einstein in 1905 was the proposal of a quantum (particular) aspect of the electromagnetic field – the photon. In the mid-1920s the only force in nature to be considered was the electromagnetic interaction; this was before the theories of Yukawa and Fermi, concerning the strong and weak nuclear forces. Dirac, Heisenberg, Jordan, and others then addressed themselves to finding a formulation of quantum electrodynamics (QED) comparable in mathematical sophistication to the Heisenberg–Schrödinger formulation of quantum mechanics – which Planck's and Einstein's theories were not.

The idea that was pursued, at least in the early stages, was that the Schrödinger wave function ψ, taken as a wave field, should be "quantized"; Dirac

seems to have taken this as a model for photons. Jordan further proposed that electrons should be treated as the quanta of an electron field, but recognized that their fermionic nature would modify the quantization procedure. This generic idea involved what was called "second quantization" – of a field into a particle.

One of the earliest quantization rules was Bohr's condition relating to the periodic orbits of electrons in atoms, $J = \int p\, \mathrm{d}q = nh$. At the hands of Heisenberg and Dirac this became upgraded to the commutation relation

$$[q, p] = \mathrm{i}\hbar$$

where the operators p and q are "observables." In their papers on quantum field theory, Dirac, Jordan and Wigner, and Heisenberg introduced creation and annihilation operators which had the function, as their name implied, of creating and destroying single particles – quanta of the field. These operators obeyed the commutation rules (with $[A, B] = AB - BA$)

$$[b_r, b_s^*] = \delta_{rs}, \quad [b_r, b_s] = [b_r^*, b_s^*] = 0$$

when the field quanta were bosons, and the anti-commutation rules

$$\{b_r, b_s^*\} = \delta_{rs}, \quad \{b_r, b_s\} = \{b_r^*, b_s^*\} = 0$$

(with $\{A, B\} = AB + BA$) when the field quanta were fermions (e.g., electrons). These steps constitute second quantization, but it may be noted that the creation and annihilation operators are not observables, as p and q are in the Heisenberg commutation relation. In addition, the second quantization conditions do not involve Planck's constant. "First" and "second" quantization are therefore not so similar as one might like to think.

The question of what exactly is being quantized was in fact the source of some confusion. In his paper of 1927, Dirac's attention is focussed on electromagnetic radiation, but he nevertheless discusses the difference between "a light-wave and the de Broglie or Schrödinger wave associated with the light-quanta." As Dirac points out, "their intensities are to be interpreted in different ways. The number of light quanta per unit volume associated with a monochromatic light-wave equals the energy per unit volume of the wave divided by the energy $(2\pi h)\nu$ of a single light quantum. On the other hand a monochromatic de Broglie wave of amplitude a (multiplied into the imaginary exponential factor) must be interpreted as representing a^2 light quanta per unit volume for all frequencies." There are at least two problematic issues here. First, is the Schrödinger wave function ψ to be considered as a "real" field, whose quanta result in "real" particles, or is it a probability field, whose significance lies in Born's probabilistic interpretation of quantum mechanics? Born wrote in 1926, "[Einstein said that] the waves are present only to show the corpuscular light quanta the way, and he spoke in the sense of a "ghost field". This determines the probability that a light quantum, the bearer of energy and momentum, takes a certain path; however, the field itself has no energy and no momentum." This is the first problem. The second one concerns the nature of the quantization itself. Is this a quantization of field energy, or a quantization of the field itself, as a substantial entity? If the field is real, the second of these does not imply the first.

Ambiguities surrounding the idea of second quantization survived into the 1960s. Wigner is recorded as saying, in an interview in 1963, "just as we get photons by quantising the electromagnetic fields, so we should be able to get material particles by quantising the Schrödinger field." And Rosenfeld, also in an interview in 1963, said, "in some sense or other, Jordan himself took the wave function, the probability amplitude, physically more seriously than most people [did]."

It would seem we are justified in concluding that the idea of second quantization contains flaws, but an even clearer indication of the need for rethinking is provided by the story of the Dirac equation. This is a wave equation for the electron, compatible with special relativity, and taking explicit account of its spin being $(1/2)\hbar$. The equation famously had both positive- and negative-energy solutions. This potential disaster was converted by Dirac into a triumph by reinterpreting the (absence of) negative-energy solutions as (positive-energy) antiparticles – positrons, particles with positive charge but the same mass and spin as the electron. Positrons were eventually discovered by Anderson. It was later shown that the existence of antiparticles is a general feature of quantum field theory, not just a peculiarity of spin-1/2 particles. The significance of this discovery, however, is that the twin requirements of relativity and quantum theory are not compatible with a single-particle state; rather, these requirements result in a two-particle state. Thus, in some sense the requirements of relativity and quantum mechanics already start to take us down the road to a quantum theory of fields.

Quantum field theory is then constructed on the following sort of framework: "classical" theories for fields with any spin may be written down and these are quantized by reinterpreting the field variables as operators and imposing Heisenberg-type commutation relations on the field and its corresponding

"momentum" variable. So, for example, for spinless fields we have the equal-time commutation relation

$$[\phi(\boldsymbol{x}, t), \pi(\boldsymbol{y}, t)] = \mathrm{i}\hbar\delta^{(3)}(\boldsymbol{x} - \boldsymbol{y})$$

where $\pi = \partial \mathrm{L}/\partial(\partial_0\phi)$ and L is the Lagrange density. The mass and spin of particles are defined with reference to the Poincaré group (thereby incorporating special relativity) and the quantum requirement is the familiar one that physical states are represented by vectors in Hilbert space. The rest follows: as Weinberg says, "quantum field theory is the way it is because (with certain qualifications) this is the only way to reconcile quantum mechanics with special relativity."

Renormalization

A notorious problem in quantum field theory is the occurrence of infinities. In QED, for example, the electron acquires a self-energy – and therefore a contribution to its mass – by virtue of the emission and reabsorption of virtual photons. It turns out that this self-energy is infinite – it is given by a divergent integral – even in the lowest order of perturbation theory. In the early days, this was recognized as being a serious problem, and in fact it turns out to be a generic problem in quantum field theory. It was realized by Dyson, however, that in some field theories these divergences may be dealt with by redefining a small number of parameters (e.g., in QED, the electron mass, charge, and field amplitude) so that thereafter the theory is finite to all orders of perturbation theory. Such theories are called renormalizable, and QED is a renormalizable field theory.

Some important field theories, however, are not renormalizable; an example is Fermi's theory of weak interactions. To lowest order in perturbation theory, Fermi's theory works well (e.g., in accounting for the electron spectrum in neutron beta decay), but to higher orders divergent results are obtained, which cannot be waved away by redefining a finite number of parameters; that is to say, as the order of perturbation increases, so also does the number of parameters to be redefined. Nonrenormalizable theories of this type have traditionally been regarded as highly undesirable, not to say rather nasty.

The modern view of renormalization is, however, somewhat different. The problem with nonrenormalizable theories is that, in order to calculate a physical process to all orders in perturbation theory, an infinite number of parameters must be renormalized, so the theory has no predictive power. In practice, however, we do not need to calculate to all orders in perturbation theory, since any physical process (say a scattering process or a particle decay) will only be observed at a finite energy and comparison of theory and experiment therefore only requires calculation up to a finite order of perturbation theory. So even nonrenormalizable theories are perfectly acceptable as low-energy theories. This amounts to a philosophy of effective field theories; an effective field theory is a model which holds good up to a particular energy scale, or equivalently down to a particular length scale.

An important addition to the theoretical armoury is the renormalization group. Renormalization is implemented first of all by a scheme of regularization, which enables the divergences to be exhibited explicitly. The simplest type of regularization is the introduction of a cutoff in the momentum integrals, but in modern particle physics the favored scheme is dimensional regularization. The dimensionality of the integrals in momentum space is taken to be $d = 4 - \varepsilon$ and the divergent quantities have an explicit dependence on ε (which, of course, as the "real" world is approached, approaches zero). At the same time, a mass parameter μ is introduced in order to define dimensionless quantities, for example, a dimensionless coupling constant. The renormalized quantities then depend on the "bare" (unrenormalized) quantities and on μ and ε. The arbitrariness of μ enables a differential equation, for scattering amplitudes, for example, to be written down. While at first sight this renormalization group equation might seem to have no physical importance, in fact it gives a powerful way of studying scattering behavior at large momenta.

Most interestingly, the concept of the renormalization group also arises in condensed matter physics. Here, rather than, for example, a cutoff in momentum space, the relevant parameter is a distance scale. In the Ising model in statistical mechanics, for example, in which spins are located on a lattice, the parameter is the lattice spacing. To construct a theory that describes the physics on the macroscopic scale involves integrating out the details on the microscopic scale and one way to do this is via the "block spin" transformation originally introduced by Kadanoff. In this way the renormalization group has had a large impact in condensed matter physics, for example, in the study of critical phenomena.

Particle Physics and Cosmology

Probably the most spectacular success of quantum field theory in the twentieth century has been in particle physics. The "standard model" accounts for the strong, electromagnetic, and weak interactions

between elementary particles with outstanding success. The interactions are generalizations of Maxwell's electrodynamics, which is invariant under a symmetry group U(1) of gauge transformations. An enlargement of this group to $SU(2) \otimes U(1)$ accounts for the unified electroweak interaction (the unification resulting from the fact that the two U(1)'s above are not exactly the same; there is some on-diagonal mixing), and the strong interactions between quarks, which binds them into hadrons, are invariant under an SU(3) group of gauge transformations. The gauge fields are the photon γ, the W and Z bosons (both heavy; of the order of 100 times the proton mass), and the (massless) gluons mediating the force between quarks (quantum chromodynamics, QCD). An important feature of the standard model is spontaneous symmetry breaking, which is the mechanism by which the W and Z particles acquire a mass (but the photon does not, and neither do the gluons). This goes by the name of the Higgs mechanism.

The quantization of the standard model is most successfully carried out using the path-integral formalism, rather than canonical quantization, and the proof of the renormalizability of the model (of nonabelian gauge theories with spontaneous symmetry breaking) was given by 't Hooft. Details of these topics are now available in many textbooks.

Confidence that this is a realistic model of elementary particles – that is to say, of quarks and leptons – depends, of course, on particular experiments and their interpretation and an important milestone on this journey was Feynman's quark–parton model of deep inelastic electron–proton scattering. The interpretation of the data required a picture of an electron scattering from an individual quark in the proton, and this in turn required a negligible interaction between quarks; in other words, that at small distances (inside the proton) the quarks are (almost) free – despite the fact that at large distances they most certainly are not! The proof, by Gross, Politzer, and Wilczek, that nonabelian gauge are indeed asymptotically free (asymptotic in momentum space, that is) was therefore an important event in helping to establish the credibility of the standard model.

A characteristic contribution of quantum field theory to our view of the physical world is its picture of the vacuum, as being populated with virtual particle–antiparticle pairs. A consequence of this is the phenomenon of vacuum polarization – that the presence of an electric charge in free space polarizes these virtual pairs. This in turns leads to the phenomenon of screening in QED, and antiscreening in QCD, SU(3) having a more complicated structure than U(1). It also leads to a nonzero (in fact, quadratically divergent!) value for the energy of the vacuum. This is in effect the contribution of the zero-point energies of all the oscillators in the Fourier expansion of the scalar field operator. In any other interaction than gravity, this zero-point energy may be ignored, but in gravity it may be expected to have observable consequences, and indeed it turns out that it plays the same role as a cosmological constant Λ, and therefore acts as an agent of acceleration, rather than deceleration, of the universe.

A final topic worth noting is one whose existence would have been inconceivable in the early days of this subject. The nonlinearity of the (nonabelian) gauge field equations and the existence of a nontrivial group space allows new types of topologically nontrivial solutions to these equations: solitons, bounces, instantons, sphalerons, and so on. Effects such as fractional spin and nonconservation of fermion number also appear, and, on the cosmological scale, domain walls and cosmic strings. There is something here for theoretical physicists of many differing interests.

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; BRST Quantization; Constrained Systems; Constructive Quantum Field Theory; Deformation Quantization; Electroweak Theory; Euclidean Field Theory; Exact Renormalization Group; Integrability and Quantum Field Theory; Nonperturbative and Topological Aspects of Gauge Theory; Perturbative Renormalization Theory and BRST; Quantum Chromodynamics; Quantum Electrodynamics and Its Precision Tests; Quantum Fields with Indefinite Metric: Non-Trivial Models; Quantum Fields with Topological Defects; Renormalization: General Theory; Standard Model of Particle Physics; Symmetries and Conservation Laws; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Topological Defects and Their Homotopy Classification; Topological Quantum Field Theory: Overview; Twistors.

Further Reading

Cao TY (1997) *Conceptual Developments of 20th Century Field Theories*. Cambridge: Cambridge University Press.

Davies P (ed.) (1989) *The New Physics*. Cambridge: Cambridge University Press.

Gross F (1993) *Relativistic Quantum Mechanics and Field Theory*. New York: Wiley.

Huang K (1998) *Quantum Field Theory*. New York: Wiley.

Itzykson C and Zuber J-B (1980) *Quantum Field Theory*. New York: McGraw-Hill.

Maggiore M (2005) *A Modern Introduction to Quantum Field Theory*. Oxford: Oxford University Press.

Rubakov V (2002) *Classical Theory of Gauge Fields*. Princeton: Princeton University Press.

Schweber SS (1994) *QED and the Men Who Made It*. Princeton: Princeton University Press.

Schwinger J (ed.) (1958) *Quantum Electrodynamics*. New York: Dover.

't Hooft G (1997) *In Search of the Ultimate Building Blocks*. Cambridge: Cambridge University Press.

Weinberg S (1995, 1996) *The Quantum Theory of Fields*, vol. 1 and 2. Cambridge: Cambridge University Press.
Wentzel G (1960) Quantum theory of fields (until 1947). In: Fierz M and Weisskopf VF (eds.) *Theoretical Physics in the Twentieth Century*, (Reprinted in Mehra J (1973) *The Physicist's Conception of Nature*. Dordrecht: Reidel.) New York: Interscience.
Zee A (2003) *Quantum Field Theory in a Nutshell*. Princeton: Princeton University Press.

Quantum Fields with Indefinite Metric: Non-Trivial Models

S Albeverio and H Gottschalk, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

Introduction

The nonperturbative construction of quantum field models with nontrivial scattering in arbitrary dimension d of the underlying Minkowski spacetime is much simpler in the framework of quantum field theory with indefinite metric than in the positive-metric case. In particular, there exist a number of solutions in the physical dimension $d = 4$, where up to now no positive-metric solutions are known. The reasons why this is so are reviewed in this article, and some examples obtained by analytic continuation from the solutions of Euclidean covariant stochastic partial differential equations (SPDEs) driven by non-Gaussian white noise are discussed.

The Hilbert Space Structure Condition

It has been proved by F Strocchi that a quantum gauge field in a local, covariant gauge cannot act on a Hilbert space with a positive-definite inner product. But it is possible to overcome this obstacle by passing from a Hilbert space representation of the algebra of the quantum field to Krein space representations in order to preserve locality and covariance under the Poincaré group.

A Krein space $\mathcal{K}$ is an inner-product space which also is a Hilbert space with respect to some auxiliary scalar product. The relation between the inner product $\langle .,. \rangle$ and the auxiliary scalar product $(.,.)$ is given by a self-adjoint linear operator $J : \mathcal{K} \to \mathcal{K}$ with $J^2 = \mathbf{1}_{\mathcal{K}}$ and $\langle .,. \rangle = (.,J.)$. J is called the metric operator. A quantum field acting on such a space is called a quantum field with indefinite metric. The formal definition is as follows.

Let $\mathcal{D} \subseteq \mathcal{K}$ be a dense linear space and $\Omega \in \mathcal{D}$ a distinguished vector (henceforth called the vacuum). Let $\mathcal{S} = \mathcal{S}(\mathbb{R}^d, \mathbb{C}^N)$ be the space of Schwartz test functions with values in $\mathbb{C}^N$. A quantum field ϕ by definition is a linear mapping from $\mathcal{S}$ to the linear operators on $\mathcal{D}$. One usually assumes that $\mathcal{D}$ is generated as the linear span of vectors generated by repeated application of field operators to the vacuum. The following properties should hold for the quantum field ϕ:

1. *Temperedness*: $f_n \to f$ in $\mathcal{S} \Rightarrow \langle \Psi, \phi(f_n)\Phi \rangle \to \langle \Psi, \phi(f)\Psi \rangle \ \forall \Psi, \Phi \in \mathcal{S}$.
2. *Covariance*: There exists a weakly continuous representation U of the covering of the orthochronous, proper Poincaré group $\tilde{P}^{\uparrow}_{+}$ by linear operators on $\mathcal{D}$ which is J-unitary, that is, $U^{[*]} = U^{-1}$ with $U^{[*]} = JU^*J|_{\mathcal{D}}$ and leaves Ω invariant. ϕ is said to be covariant with respect to U and a representation τ of the covering of the orthochronous, proper Lorentz group $\tilde{L}^{\uparrow}_{+}$ if $U(g)\phi(f)U(g)^{-1} = \phi(f_g)$, where $f_g(x) = \tau(\Lambda)f(\Lambda^{-1}(x-a))$, $g = \{\Lambda, a\}$, $\Lambda \in \tilde{L}^{\uparrow}_{+}$, $a \in \mathbb{R}^d$.
3. *Spectrality*: Let $U(a)$, $a \in \mathbb{R}^d$, be the representation of the translation group and let $\sigma = \cup_{\Psi,\Phi\in\mathcal{D}} \mathrm{supp}\mathcal{F}(\langle \Psi, U(.)\Phi \rangle)$ with $\mathcal{F}$ the Fourier transform (in the sense of tempered distributions). Formally, σ is the joint spectrum of the generators of spacetime translations $U(a)$. The spectral condition then demands that $\sigma \subseteq \bar{V}^{+}_{0}$, the closed forward light cone in energy–momentum space.
4. *Locality*: There is a decomposition $\mathbb{C}^N = \oplus_\kappa V_\kappa$ such that for each $f, h \in \mathcal{S}$ taking values in a V_κ and having spacelike separated supports one has either $[\phi(f), \phi(h)] = 0$ or $\{\phi(f), \phi(h)\} = 0$, where $[.,.]$ is the commutator and $\{.,.\}$ the anticommutator.
5. *Hermiticity*: There is an involution $*$ on $\mathcal{S}$ such that $\phi(f)^{[*]} = \phi(f^*)$.

The quantum-mechanical interpretation of the inner product of two vectors in $\mathcal{K}$ as a probability amplitude, however, gets lost. It has to be restored by the construction of a physical subspace of $\mathcal{K}$ where the restriction of the inner product is non-negative. This is called the Gupter–Bleuler gauge procedure. Typically, one first considers the problem of constructing quantum fields with indefinite metric, that is, the dynamical problem is addressed. This is often followed by the construction of the physical states, which involves implementation of quantum constraints.

The vacuum expectation values (VEVs), also called Wightman functions, of the quantum field theory with indefinite metric (IMQFT) are defined as

$$W_n(f_1 \otimes \cdots \otimes f_n) = \langle \Omega, \phi(f_1) \cdots \phi(f_n)\Omega \rangle \quad f_1, \ldots, f_n \in \mathcal{S} \qquad [1]$$

An axiomatic framework for (unconstrained) IMQFT has been suggested by G Morchio and F Strocchi in terms of the Wightman functions $W_n \in \mathcal{S}'$, $n \in \mathbb{N}_0$. Previous work on the topic had been done by J Yngvason. These generalized Wightman axioms of Morchio and Strocchi replace the positivity condition on the Wightman functions by a so-called Hilbert space structure condition (HSSC): for $n \in \mathbb{N}_0$ there exist p_n a Hilbert seminorm on $\mathcal{S}^{\otimes n}$ such that

$$|W_{n+m}(f \otimes h)| \leq p_n(f) p_m(h) \quad \forall n, m \in \mathbb{N}_0 \quad f \in \mathcal{S}^{\otimes n},\ h \in \mathcal{S}^{\otimes m} \qquad [2]$$

This condition makes sure that a field algebra on a Krein space with VEVs equal to the given set of Wightman functions can be constructed. The remaining axioms of the Wightman framework – temperedness, covariance, spectral condition, locality, and Hermiticity – remain the same. Clustering of Wightman functions is assumed at least for massive theories:

$$\lim_{t \to \infty} W_{n+m}(f \otimes h_{ta}) = W_n(f) W_m(h) \quad \forall n, m \in \mathbb{N}_0 \quad f \in \mathcal{S}^{\otimes n},\ h \in \mathcal{S}^{\otimes m} \qquad [3]$$

for spacelike $a \in \mathbb{R}^d$. It fails to hold in certain physical contexts where multiple vacua (also called Θ-vacua) accompanied with massless Goldstone bosons occur due to spontaneous symmetry breaking.

In the original Wightman axioms, there are essentially two nonlinear axioms: positivity and clustering. Here nonlinear means that checking that condition involves more than one VEV with a given number of field operators. The cluster condition can be linearized by an operation on the Wightman functions called "truncation." The equations

$$W_n(f_1 \otimes \cdots \otimes f_n) = \sum_{I \in \mathcal{P}^{(n)}} \prod_{\substack{\{j_1, \ldots, j_l\} \in I \\ j_1 < j_2 < \cdots < j_l}} W_n^T(f_{j_1} \otimes \cdots \otimes f_{j_l}) \qquad [4]$$

recursively define the truncated Wightman functions W_n^T for $n \in \mathbb{N}$. Here $\mathcal{P}^{(n)}$ stands for the set of all partitions of $\{1, \ldots, n\}$ into disjoint, nonempty sets. Unfortunately, the positivity condition (at least when combined with nontrivial scattering) becomes highly nonlinear for truncated Wightman functions. This can be seen as one explanation why it is so difficult to find nontrivial (i.e., corresponding to nontrivial interactions) solutions to the Wightman axioms.

But it turns out that, in contrast to positivity, the HSSC is essentially linear for truncated Wightman functions.

Theorem 1 *If there exists a Schwartz norm $||\cdot||$ on $\mathcal{S}$ such that W_n^T is continuous with respect to $||\cdot||^{\otimes n}$ for $n \in \mathbb{N}$ then the associated sequence of Wightman functions $\{W_n\}$ fulfills the HSSC* [2].

Note that $||\cdot||^{\otimes n}$ is well defined as $\mathcal{S}$ is a nuclear space. This theorem makes it much easier to construct IMQFTs. In particular, all known solutions of the linear program for truncated Wightman functions lead to an abundance of mathematical solutions to the axioms of IMQFT, as long as the singularities of truncated Wightman functions in position and energy–momentum space do not become increasingly stronger with growing n. For example, the perturbative solutions to Wightman functions of Ostendorf and Steinmann provide solutions when the perturbation series is truncated at a given order.

Relativistic Fields from Euclidean Stochastic Equations

In the classical work on constructive quantum field theory, relativistic fields in spacetime dimensions $d = 2$ and 3 have been constructed by analytic continuation from Euclidean random fields. This, in particular, has led to firm connections between quantum field theory and equilibrium statistical mechanics. Let us discuss one specific class of solutions of the axioms of IMQFT for arbitrary d which also stem from random fields related to an ensemble of statistical mechanics of classical, continuous particles. Mathematically, this is connected with using random fields with Poisson distribution. As in constructive QFT, the moments, also called Schwinger functions, of the random field can be analytically continued from Euclidean imaginary time to relativistic real time. That this is possible results from an explicit calculation. Axiomatic results cannot be used, as they depend on positivity or reflection positivity in the Euclidean spacetime, respectively.

By definition, a mixing Euclidean covariant random field φ is an almost surely linear mapping from $\mathcal{S}_{\mathbb{R}} = \mathcal{S}(\mathbb{R}^d, \mathbb{R}^N)$ to the space of real-valued

measurable functions (random variables) on some probability space that fulfills the following properties:

1. *Temperedness*: $f_n \to f$ in $\mathcal{S}_\mathbb{R} \Rightarrow \varphi(f_n) \xrightarrow{\mathcal{L}} \varphi(f)$.
2. *Covariance*: $\varphi(f) \stackrel{\mathcal{L}}{=} \varphi(f_g)\ \forall f \in S_\mathbb{R}$, $g = \{\Lambda, a\}$, $\Lambda \in \mathrm{SO}(d)$, $a \in \mathbb{R}^d$, $f_g(x) = \tau(\Lambda) f(\Lambda^{-1}(x-a))$ for some continuous representation $\tau : \mathrm{SO}(d) \to \mathrm{GL}(N)$.
3. *Mixing*: $\lim_{t\to\infty} \mathbb{E}[AB_{ta}] = \mathbb{E}[A]\mathbb{E}[B]$ for all square-integrable random variables $A = A(\varphi)$, $B = B(\varphi)$, and $B_{ta} = B(\varphi_{ta})$, $\varphi_{ta}(f) = \varphi(f_{ta})\ \forall f \in S_\mathbb{R}$, $a \in \mathbb{R}^d \setminus \{0\}$.

The mixing condition in the Euclidean spacetime plays the same role as the cluster property in the generalized Wightman axioms.

In particular, we consider random fields φ obtained as solutions of the SPDE $D\varphi = \eta$. In this equation, η is a noise field, that is, η is τ-covariant for some representation of $\mathrm{SO}(d)$, $\eta(f)$ has infinitely divisible probability law and $\eta(f)$, $\eta(h)$ are independent $\forall f, h \in \mathcal{S}_\mathbb{R}$ with $\mathrm{supp}\, f \cap \mathrm{supp}\, h = \emptyset$. D is a τ-covariant (i.e., $\tau(\Lambda) D \tau(\Lambda)^{-1} = D \quad \forall \Lambda \in \mathrm{SO}(d)$) partial differential operator with constant coefficients (also pseudodifferential operators D could be considered). From the classification of infinitely divisible probability laws, it is known that η essentially consists of Gaussian white noise and Poisson fields and derivatives thereof. Such a Gauss–Poisson noise field by the Bochner–Minlos theorem is characterized by its Fourier transform. Direct relations with QFT arise if one chooses

$$\mathbb{E}[\mathrm{e}^{\mathrm{i}\eta(f)}] = \exp\left\{ \int_{\mathbb{R}^d} \psi(f) - f \cdot \bar{\sigma}^2 p(-\Delta) f \, \mathrm{d}x \right\} \quad f \in \mathcal{S}_\mathbb{R} \qquad [5]$$

where $\psi : \mathbb{R}^N \to \mathbb{C}$ is a Lévy function,

$$\psi(t) = \mathrm{i}a \cdot t - \frac{t \cdot \sigma^2 t}{2} + z \int_{\mathbb{R}^N \setminus \{0\}} (\mathrm{e}^{\mathrm{i}t\cdot s} - 1)\, \mathrm{d}r(s) \quad t \in \mathbb{R}^N \qquad [6]$$

Here the centered dot represents a τ-invariant scalar product on $\mathbb{R}^N$, σ a positive-semidefinite τ-invariant $N \times N$ matrix, $z \geq 0$ a real number and r is a τ-invariant probability measure on $\mathbb{R}^n \setminus \{0\}$ with all moments. Further, $\bar{\sigma}^2_{\alpha,\beta} = (\partial^2 \psi(t)/\partial t_\alpha \partial t_\beta)|_{t=0}$, and $p : [0,\infty) \to [0,\infty)$ is a polynomial depending on D. If $\hat{D}^{-1}$, the Fourier-transformed inverse of D, exists, it can be represented by

$$\hat{D}^{-1}(k) = \frac{Q_E(k)}{\prod_{l=1}^P (|k|^2 + m_l^2)^{\nu_l}} \qquad [7]$$

Here $Q_E(k)$ is a complex $N \times N$ matrix with polynomial entries being τ-covariant, $\tau(\Lambda) Q_E (\Lambda^{-1}k) \tau(\Lambda)^{-1} = Q_E(k) \quad \forall \Lambda \in \mathrm{SO}(d), \quad k \in \mathbb{R}^d$. $\nu_l \in \mathbb{N}$ and $m_1 \in \mathbb{C} \setminus (-\infty, 0)$ are parameters with the interpretation of the mass spectrum $(m_1, \ldots, m_P)$ and $(\nu_1, \ldots, \nu_P)$ the dipole degrees of the related masses. We restrict ourselves to the case of positive mass spectrum where $m_l > 0$, and in this case

$$p(t) = p(t, D) = \frac{\prod_{l=1}^P (t + m_l^2)^{\nu_l}}{\prod_{l=1}^P m_l^{2\nu_l}}, \quad t > 0 \qquad [8]$$

One can show that φ obtained as the unique solution of the SPDE $D\varphi = \eta$ is a Euclidean covariant, mixing random field. The Schwinger functions (moments) of φ are given by

$$S_n(f_1 \otimes \cdots \otimes f_n) = E[\varphi(f_1) \cdots \varphi(f_n)], \quad f_1, \ldots, f_n \in \mathcal{S}_\mathbb{R} \qquad [9]$$

Now the Schwinger functions can be calculated explicitly. They are determined by the truncated Schwinger functions, cf. [4], as follows: for $n = 2$,

$$S^{\mathrm{T}}_{2,\alpha_1,\alpha_2}(x_1, x_2) = \frac{Q^{\mathrm{E}}_{2,\alpha_1,\alpha_2}(-\mathrm{i}\underline{\nabla}_2)}{\prod_{l=1}^N m_l^{2\nu_l}} \left[\prod_{l=1}^N (-\Delta + m_l^2)^{-\nu_l} \right](x_1 - x_2) \qquad [10]$$

and for $n \geq 3$

$$S^{\mathrm{T}}_{n,\alpha_1\cdots\alpha_n}(x_1, \ldots, x_n) = Q^{\mathrm{E}}_{n,\alpha_1\cdots\alpha_n}(-\mathrm{i}\underline{\nabla}_n) \times \int_{\mathbb{R}^d} \prod_{j=1}^n \left[\prod_{l=1}^N (-\Delta + m_l^2)^{-\nu_l} \right](x_j - x)\, \mathrm{d}x \qquad [11]$$

where

$$Q^{\mathrm{E}}_{n,\alpha_1\cdots\alpha_n}(-\mathrm{i}\underline{\nabla}_n) = C^{\beta_1\cdots\beta_n} \prod_{l=1}^n Q_{E,\beta_l,\alpha_l}\left(-\mathrm{i}\frac{\partial}{\partial x_l} \right) \qquad [12]$$

with

$$C_{\beta_1\cdots\beta_n} = (-\mathrm{i})^n \frac{\partial^n \psi(t)}{\partial t_{\beta_1} \cdots \partial t_{\beta_n}}\bigg|_{t=0} \qquad [13]$$

and the Einstein convention of summation and raising/lowering of indices on $\mathbb{R}^N$ with respect to the invariant inner product $\cdot$ is applied. The Schwinger functions fulfill the requirements of τ-covariance, symmetry, clustering, and Hermiticity from the Osterwalder–Schrader axioms of Euclidean QFT.

While there is no known general reason why a relativistic QFT should exist for a given set of Schwinger functions, one can take advantage of the explicit formulas [10]–[13] in order to calculate the analytic continuation from Euclidean to relativistic times explicitly.

It simplifies the considerations to exclude dipole fields, that is, one assumes that $\nu_l = 1$ for $l = 1, \ldots, n$. In physical terms, the no-dipole condition guarantees that the asymptotic fields in Minkowski spacetime fulfill the Klein–Gordon equation and thus generate particles in the usual sense if applied to the vacuum. If this condition is not imposed, asymptotic fields might only fulfill a dipole equation $(\Box + m^2)^2 \phi^{\text{in/out}} = 0$ or a related hyperbolic equation of even higher order, and the particle states generated by application of such fields to the vacuum require a gauge fixing (constraints) in order to obtain a physical interpretation. Given the no-dipole condition, one obtains by expansion into partial fractions

$$\frac{1}{\prod_{l=1}^{P}(|k|^2 + m_l^2)} = \sum_{l=1}^{N} \frac{b_l}{(|k|^2 + m_l^2)} \qquad [14]$$

with $b_l \in (0, \infty)$ uniquely determined and $b_l \neq 0$. For the truncated Schwinger functions, this implies $(n \geq 3)$ that

$$\begin{aligned} &S^{\mathrm{T}}_{n,\alpha_1\cdots\alpha_n}(x_1, \ldots, x_n) \\ &\quad = Q^{\mathrm{E}}_{n,\alpha_1\cdots\alpha_n}(-\mathrm{i}\underline{\nabla}_n) \sum_{l_1,\ldots,l_n=1}^{P} \\ &\qquad \times \prod_{r=1}^{n} b_{l_r} \int_{\mathbb{R}^d} \prod_{j=1}^{n} (-\Delta + m_{l_j}^2)^{-1}(x - x_j)\, \mathrm{d}x \end{aligned} \qquad [15]$$

At this point, a lengthy calculation yields a representation of the functions $\int_{\mathbb{R}^d} \prod_{j=1}^{n} (-\Delta + m_j^2)^{-1}(x - x_j)\, \mathrm{d}x$ as the Fourier–Laplace transform of a distribution $\hat{W}^{\mathrm{T}}_{n,m_1,\ldots,m_n}$ that fulfills the spectral condition. This is equivalent to the statement that the analytic continuation of such functions to relativistic times yields $W^{\mathrm{T}}_{n,m_1,\ldots,m_n}$, where the latter distribution is the inverse Fourier transform of $\hat{W}^{\mathrm{T}}_{n,m_1,\ldots,m_n}$. This distribution up to a constant that can be integrated into Q^{E} is given by

$$\left\{ \sum_{j=1}^{n} \prod_{l=1}^{j-1} \delta^-_{m_l}(k_l) \frac{(-1)}{k^2 - m_j^2} \prod_{l=j+1}^{n} \delta^+_{m_l}(k_l) \right\} \delta\left(\sum_{l=1}^{n} k_l \right) \qquad [16]$$

Here $\delta^{\pm}_m(k) = \theta(\pm k^0)\delta(k^2 - m^2)$, where θ is the Heaviside step function and $k^2 = k^{0^2} - |\boldsymbol{k}|^2$. On the other hand, the partial differential operator Q^{E}_n can be analytically continued in momentum space:

$$\begin{aligned} &Q^{\mathrm{M}}_n((k_1^0, \boldsymbol{k}_1), \ldots, (k_n^0, \boldsymbol{k}_n)) \\ &\quad = Q^{\mathrm{E}}_n((\mathrm{i}k_1^0, \boldsymbol{k}_1), \ldots, (\mathrm{i}k_n^0, \boldsymbol{k}_n)) \end{aligned} \qquad [17]$$

$k_1, \ldots, k_n \in \mathbb{R}^d$. With the definition

$$\hat{W}^{\mathrm{T}}_{2,\alpha_1\alpha_2}(k_1, k_2) = (2\pi)^{(d+1)} \frac{Q^{\mathrm{M}}_{2,\alpha_1\alpha_2}(k_1, k_2)}{\prod_{l=1}^{N} m_l^2} \times \sum_{l=1}^{N} b_l\, \delta^-_{m_l}(k_1)\delta(k_1 + k_2) \qquad [18]$$

and

$$\begin{aligned} &\hat{W}^{\mathrm{T}}_{n,\alpha_1\cdots\alpha_n}(k_1, \ldots, k_n) \\ &\quad = Q^{\mathrm{M}}_{n,\alpha_1\cdots\alpha_n}(k_1, \ldots, k_n) \\ &\qquad \times \sum_{l_1,\ldots,l_n=1}^{N} \prod_{j=1}^{n} b_{l_j}\, \hat{W}^{\mathrm{T}}_{n,m_{l_1},\ldots,m_{l_n}}(k_1, \ldots, k_n) \end{aligned} \qquad [19]$$

the analytic continuation of Schwinger functions can be summarized as follows:

Theorem 2 *The truncated Schwinger functions $\mathcal{S}^{\mathrm{T}}_n$ have a Fourier–Laplace representation with $\hat{W}^{\mathrm{T}}_n$ defined in eqns [18] and [19]. Equivalently, $\mathcal{S}^{\mathrm{T}}_n$ is the analytic continuation of W^{T}_n from purely real relativistic time to purely imaginary Euclidean time. The truncated Wightman functions W^{T}_n fulfill the requirements of temperedness, relativistic covariance with respect to the representation of the orthochronous, proper Lorentz group $\tilde{\tau} : \mathrm{L}^{\uparrow}_{+}(d) \to \mathrm{Gl}(\mathrm{L})$, locality, spectral property, and cluster property. Here $\tilde{\tau}$ is obtained by analytic continuation of τ to a representation of the proper complex Lorentz group over $\mathbb{C}^d$ (which contains $\mathrm{SO}(d)$ as a real submanifold) and restriction of this representation to the real orthochronous proper Lorentz group.*

Again making use of the explicit formula in Theorem 2, the condition of Theorem 1 can be verified. This proves the existence of IMQFT models associated with the class of random fields under discussion.

Theorem 3 *The Wightman functions defined in Theorem 2 fulfill the HSSC [2]. In particular, there exists a QFT with indefinite metric such that the Wightman functions are given as the VEVs of that IMQFT.*

Nontrivial Scattering

Theories as described in Theorem 2 obviously have trivial scattering behavior if the noise field η is Gaussian, that is, if, in [7], $z = 0$. In the case where there is also a Poisson component in η, that is, $z > 0$, higher-order truncated Wightman functions do not vanish and such relativistic theories have nontrivial scattering.

Before the scattering of the models can be discussed, some comments about scattering in IMQFT in general are in order. The scattering

theory in axiomatic QFT, Haag–Ruelle theory, relies on positivity. In fact, one can show that in the class of models under discussion, the LSZ asymptotic condition is violated if dipole degrees of freedom are admitted. In that case more complicated asymptotic conditions have to be used. In any case, the Haag–Ruelle theory cannot be adapted to IMQFT.

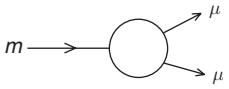
Nevertheless, asymptotic fields and states can be constructed in IMQFT if one imposes a no-dipole condition in a mathematically precise way. Then the LSZ asymptotic condition leads to the construction of mixed VEVs of asymptotic in- and out-fields with local fields. The collection of such VEVs is called the form-factor functional. After constructing this collection of mixed VEVs, one can try to check the HSSC for this functional and obtains a Krein space representation for the algebra generated by in- local and out-fields.

Following this line, asymptotic in- and out-particle states can be constructed for the given mass spectrum $(m_1, \ldots, m_P)$. If $a^{\text{in/out}\dagger}_{\alpha,l}(k)$, $l = 1, \ldots, P$, denotes the creation operator for an incoming/outgoing particle with mass m_l, spin component α, and energy–momentum k, the following scattering amplitude can be derived for r incoming particles with masses $m_{l_1}, \ldots, m_{l_r}$ and $n - r$ outgoing particles with masses $m_{l_{r+1}}, \ldots, m_{l_n}$:

$$\left\langle a^{\text{in}\dagger}_{\alpha_1,l_1}(k_1)\cdots a^{\text{in}\dagger}_{\alpha_r,l_r}(k_r)\Omega, a^{\text{out}\dagger}_{\alpha_{r+1},l_{r+1}}(k_{r+1})\cdots a^{\text{out}\dagger}_{\alpha_n,l_n}(k_n)\Omega\right\rangle^{\mathrm{T}} = -(2\pi)\mathrm{i}Q^{\mathrm{M}}_{\alpha_1,\ldots,\alpha_n}(-k_1,\ldots,-k_r,k_{r+1},\ldots,k_n) \times \prod_{j=1}^{n}\delta^{+}_{m_{l_j}}(k_j)\,\delta(K^{\text{in}} - K^{\text{out}}) \qquad [20]$$

$K^{\text{in/out}}$ stand for the total energy–momentum of in- and out-particles, that is, $K^{\text{in}} = \sum_{j=1}^{r} k_j$ and $K^{\text{out}} = \sum_{j=r+1}^{n} k_j$.

Two immediate consequences can be drawn from [20]. First, choosing a model with nonvanishing Poisson part such that $C_{\beta_1\beta_2\beta_3} \neq 0$ and a differential operator D containing in its mass spectrum the masses m and μ with $m > 2\mu$, one gets a nonvanishing scattering amplitude for the process



[21]

even though in- and out-particle states consist of particles with well-defined sharp masses. Thus, for the incoming particle, the energy uncertainty, which for a particle at rest is proportional to the mass uncertainty, vanishes but still the particle undergoes a nontrivial decay and must have a finite decay time. This appears to be a contradiction to the energy–time uncertainty relation, which therefore seems to have an unclear status in IMQFT (i.e., in QFT including gauge fields). The origin of this inequality, which of course is experimentally very well tested, apparently has to be located in the constraints, that is, in the procedure of implementing a gauge, of the theory and not in the unconstrained IMQFT.

Second, one can replace somewhat artificially the polynomials Q^{M}_n in [17] by any other symmetric and relativistically covariant polynomial. If the sequence of the "new" Q^{M}_n is of uniformly bounded degree in any of the arguments $k_1, \ldots, k_n$, the redefined Wightman functions in [17] still fulfill the requirements of Theorem 1 and thus define a new relativistic, local IMQFT. The scattering amplitudes of such a theory are again well defined and given by [20]. For example, in the case of only one scalar particle with mass m, one can show that arbitrary Lorentz-invariant scattering behavior of bosonic particles can be reproduced by such theories for energies below an arbitrary maximal energy up to arbitrary precision. This kind of interpolation theorem shows that the outcome of an arbitrary scattering experiment can be reproduced within the formalism of (unconstrained) IMQFT as long as it is in agreement with the general requirements of Poincaré invariance and statistics.

List of Symbols

$\to$	converges to
$\xrightarrow{\mathcal{L}}$	convergence in law
$\mathbb{N}$	set of natural numbers
$\mathbb{N}_0$	set of natural numbers and zero
$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
$\mathbf{1}$	identity mapping
$\vert_{\mathcal{D}}$	restricted to $\mathcal{D}$
x^0 and $\boldsymbol{x}$	time and spatial part of $x = (x^0, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^{d-1}$
∇_n	gradient operator on $\mathbb{R}^{dn}$

See also: Algebraic Approach to Quantum Field Theory; Euclidean Field Theory; Indefinite Metric; Perturbative Renormalization Theory and BRST; Quantum Field Theory in Curved Spacetime; Quantum Field Theory: A Brief Introduction; Stochastic Differential Equations.

Further Reading

Albeverio S and Gottschalk H (2001) Scattering theory for quantum fields with indefinite metric. *Communications in Mathematical Physics* 216: 491–513.

Albeverio S, Gottschalk H, and Wu J-L (1997) Models of local relativistic quantum fields with indefinite metric (in all dimensions). *Communications in Mathematical Physics* 184: 509–531.

Gottschalk H (2001) On the stability of one particle states generated by quantum fields fulfilling Yang–Feldman equations. *Reports on Mathematical Physics* 47(2): 241–246.

Grothaus M and Streit L (1999) Construction of relativistic quantum fields in the framework of white noise analysis. *Journal of Mathematical Physics* 40(11): 5387.
Morchio G and Strocchi F (1980) Infrared singularities, vacuum structure and pure phases in local quantum field theory. *Ann. Inst. H. Poincaré* 33: 251.
Steinmann O (2000) *Perturbative Quantum Electrodynamics and Axiomatic Field Theory*. Berlin: Springer.
Strocchi F (1993) *Selected Topics on the General Properties of Quantum Field Theory*. Lecture Notes in Physics, vol. 51. Singapore: World Scientific.

Quantum Fields with Topological Defects

M Blasone and G Vitiello, Università degli Studi di Salerno, Baronissi (SA), Italy
P Jizba, Czech Technical University, Prague, Czech Republic

Introduction

The ordered patterns we observe in condensed matter and in high-energy physics are created by the quantum dynamics. Macroscopic systems exhibiting some kind of ordering, such as superconductors, ferromagnets, and crystals, are described by the underlying quantum dynamics. Even the large-scale structures in the universe, as well as the ordering in the biological systems appear to be the manifestation of the microscopic dynamics governing the elementary components of these systems. Thus, we talk of macroscopic quantum systems: these are quantum systems in the sense that, although they behave classically, some of their macroscopic features nevertheless cannot be understood without recourse to quantum theory.

The question then arises how the quantum dynamics generates the observed macroscopic properties. In other words, how it happens that the macroscopic scale characterizing those systems is dynamically generated out of the microscopic scale of the quantum elementary components (Umezawa 1993, Umezawa *et al.* 1982).

Moreover, we also observe a variety of phenomena where quantum particles coexist and interact with extended macroscopic objects which show a classical behavior, for example, vortices in superconductors and superfluids, magnetic domains in ferromagnets, dislocations and other topological defects (grain boundaries, point defects, etc.) in crystals, and so on.

We are thus also faced with the question of the quantum origin of topological defects and their interaction with quanta (Umezawa 1993, Umezawa *et al.* 1982): this is a crucial issue for the understanding of symmetry-breaking phase transitions and structure formation in a wide range of systems from condensed matter to cosmology (Kibble 1976, Zurek 1997, Volovik 2003).

Here, we will review how the generation of ordered structures and extended objects is explained in quantum field theory (QFT). We follow Umezawa (1993) and Umezawa *et al.* (1982) in our presentation. We will consider systems in which spontaneous symmetry breaking (SSB) occurs and show that topological defects originate by inhomogeneous (localized) condensation of quanta. The approach followed here is alternative to the usual one (Rajaraman 1982), in which one starts from the classical soliton solutions and then "quantizes" them, as well as to the QFT method based on dual (disorder) fields (Kleinert 1989).

In the next section we introduce some general features of QFT useful for our discussion and treat some aspects of SSB and the rearrangement of symmetry. Next we discuss the boson transformation theorem and the topological singularities of the boson condensate. We then present, as an example, a model with U(1) gauge invariance in which SSB, rearrangement of symmetry, and topological defects are present (Matsumoto *et al.* 1975a, b). There we show how macroscopic fields and currents are obtained from the microscopic quantum dynamics. The Nielsen–Olesen vortex solution is explicitly obtained as an example. The final section is devoted to conclusions.

Symmetry and Order in QFT: A Dynamical Problem

QFT deals with systems with infinitely many degrees of freedom. The fields used for their description are operator fields whose mathematical significance is fully specified only when the state space where they operate is also assigned. This is the space of the states, or physical phase, of the system under given boundary conditions. A change in the boundary conditions may result in the transition of the system from one phase to another. For example, a change of temperature from above to below the critical temperature may induce the transition from the

normal to the superconducting phase in a metal. The identification of the state space where the field operators have to be realized is thus a physically nontrivial problem in QFT. In this respect, the QFT structure is drastically different from the one of quantum mechanics (QM). The reason is the following.

The von Neumann theorem (1955) in QM states that for systems with a finite number of degrees of freedom all the irreducible representations of the canonical commutation relations are unitarily equivalent. Therefore, in QM the physical system can only live in one single physical phase: unitary equivalence means indeed physical equivalence and thus there is no room (no representations) for physically different phases. Such a situation drastically changes in QFT where systems with infinitely many degrees of freedom are treated. In such a case, the von Neumann theorem does not hold and infinitely many unitarily inequivalent representations of the canonical commutation relations do in fact exist (Umezawa 1993, Umezawa *et al.* 1982). It is such richness of QFT that allows the description of different physical phases.

QFT as a Two-Level Theory

In the perturbative approach, any quantum experiment or observation can be schematized as a scattering process where one prepares a set of free (noninteracting) particles (incoming particles or in-fields) which are then made to collide at some later time in some region of space (spacetime region of interaction). The products of the collision are expected to emerge out of the interaction region as free particles (outgoing particles or out-fields). Correspondingly, one has the in-field and the out-field state space. The interaction region is where the dynamics operates: given the in-fields and the in-states, the dynamics determines the out-fields and the out-states.

The incoming particles and the outgoing ones (also called quasiparticles in solid state physics) are well distinguishable and localizable particles only far away from the interaction region, at a time much before ($t=-\infty$) and much after ($t=+\infty$) the interaction time: in- and out-fields are thus said to be asymptotic fields, and for them the interaction forces are assumed not to operate (switched off).

The only regions accessible to observations are those far away (in space and in time) from the interaction region, that is, the asymptotic regions (the in- and out-regions). It is so since, at the quantum level, observations performed in the interaction region or vacuum fluctuations occurring there may drastically interfere with the interacting objects, thus changing their nature. Besides the asymptotic fields, one then also introduces dynamical or Heisenberg fields, that is, the fields in terms of which the dynamics is given. Since the interaction region is precluded from observation, we do not observe Heisenberg fields. Observables are thus solely described in terms of asymptotic fields.

Summing up, QFT is a "two-level" theory: one level is the interaction level where the dynamics is specified by assigning the equations for the Heisenberg fields. The other level is the physical level, the one of the asymptotic fields and of the physical state space directly accessible to observations. The equations for the physical fields are equations for free fields, describing the observed incoming/outgoing particles.

To be specific, let the Heisenberg operator fields be generically denoted by $\psi_{\mathrm{H}}(x)$ and the physical operator fields by $\varphi_{\mathrm{in}}(x)$. For definiteness, we choose to work with the in-fields, although the set of out-fields would work equally well. They are both assumed to satisfy equal-time canonical (anti)-commutation relations.

For brevity, we omit considerations on the renormalization procedure, which are not essential for the conclusions we will reach. The Heisenberg field equations and the free-field equations are written as

$$\Lambda(\partial)\psi_{\mathrm{H}}(x)=\mathcal{J}[\psi_{\mathrm{H}}](x) \qquad [1]$$

$$\Lambda(\partial)\varphi_{\mathrm{in}}(x)=0 \qquad [2]$$

where $\Lambda(\partial)$ is a differential operator, $x\equiv(t,\boldsymbol{x})$ and $\mathcal{J}$ is some functional of the ψ_{H} fields, describing the interaction.

Equation [1] can be formally recast in the following integral form (Yang–Feldman equation):

$$\psi_{\mathrm{H}}(x)=\varphi_{\mathrm{in}}(x)+\Lambda^{-1}(\partial)*\mathcal{J}[\psi_{\mathrm{H}}](x) \qquad [3]$$

where $*$ denotes convolution. The symbol $\Lambda^{-1}(\partial)$ denotes formally the Green function for $\varphi_{\mathrm{in}}(x)$. The precise form of Green's function is specified by the boundary conditions. Equation [3] can be solved by iteration, thus giving an expression for the Heisenberg fields $\psi_{\mathrm{H}}(x)$ in terms of powers of the $\varphi_{\mathrm{in}}(x)$ fields; this is the Haag expansion in the LSZ formalism (or "dynamical map" in the language of Umezawa 1993 and Umezawa *et al.* 1982), which might be formally written as

$$\psi_{\mathrm{H}}(x)=F[x;\varphi_{\mathrm{in}}] \qquad [4]$$

(A (formal) closed form for the dynamical map is obtained in the closed time path (CTP) formalism (Blasone and Jizba 2002). Then the Haag expansion [4] is directly applicable to both equilibrium and nonequilibrium situations.)

We stress that the equality in the dynamical map [4] is a "weak" equality, which means that it must be understood as an equality among matrix elements computed in the Hilbert space of the physical particles.

We observe that mathematical consistency in the above procedure requires that the set of φ_{in} fields must be an irreducible set; however, it may happen that not all the elements of the set are known from the beginning. For example, there might be composite (bound states) fields or even elementary quanta whose existence is ignored in a first recognition. Then the computation of the matrix elements in physical states will lead to the detection of unexpected poles in the Green's functions, which signal the existence of the ignored quanta. One thus introduces the fields corresponding to these quanta and repeats the computation. This way of proceeding is called the self- consistent method (Umezawa 1993, Umezawa *et al.* 1982). Thus it is not necessary to have a one-to-one correspondence between the sets $\{\psi^j_{\text{H}}\}$ and $\{\varphi^i_{\text{in}}\}$, as it happens whenever the set $\{\varphi^i_{\text{in}}\}$ includes composite particles.

The Dynamical Rearrangement of Symmetry

As already mentioned, in QFT the Fock space for the physical states is not unique since one may have several physical phases, for example, for a metal the normal phase and the superconducting phase, and so on. Fock spaces describing different phases are unitarily inequivalent spaces and correspondingly we have different expectation values for certain observables and even different irreducible sets of physical quanta. Thus, finding the dynamical map involves singling out the Fock space where the dynamics has to be realized.

Let us now suppose that the Heisenberg field equations are invariant under some group G of transformations of ψ_{H}:

$$\psi_{\text{H}}(x) \to \psi'_{\text{H}}(x) = g[\psi_{\text{H}}(x)] \qquad [5]$$

with $g \in G$. The symmetry is spontaneously broken when the vacuum state in the Fock space $\mathcal{H}$ is not invariant under the group G but only under one of its subgroups (Umezawa 1993, Umezawa *et al.* 1982).

On the other hand, eqn [4] implies that when ψ_{H} is transformed as in [5], then

$$\varphi_{\text{in}}(x) \to \varphi'_{\text{in}}(x) = g'[\varphi_{\text{in}}(x)] \qquad [6]$$

with g' belonging to some group of transformations G' and such that

$$g[\psi_{\text{H}}(x)] = F[g'[\varphi_{\text{in}}(x)]] \qquad [7]$$

When symmetry is spontaneously broken it is $G' \neq G$, with G' the group contraction of G; when symmetry is not broken then $G' = G$.

Since G is the invariance group of the dynamics, eqn [4] requires that G' is the group under which free fields equations are invariant, that is, also φ'_{in} is a solution of [2]. Since eqn [4] is a weak equality, G' depends on the choice of the Fock space $\mathcal{H}$ among the physically realizable unitarily inequivalent state spaces. Thus, we see that the (same) original invariance of the dynamics may manifest itself in different symmetry groups for the φ_{in} fields according to different choices of the physical state space. Since this process is constrained by the dynamical equations [1], it is called the dynamical rearrangement of symmetry (Umezawa 1993, Umezawa *et al.* 1982).

In conclusion, different ordering patterns appear to be different manifestations of the same basic dynamical invariance. The discovery of the process of the dynamical rearrangement of symmetry leads to a unified understanding of the dynamical generation of many observable ordered patterns. This is the phenomenon of the dynamical generation of order. The contraction of the symmetry group is the mathematical structure controlling the dynamical rearrangement of the symmetry. For a qualitative presentation see Vitiello (2001).

One can now ask which ones are the carriers of the ordering information among the system elementary constituents and how the long-range correlations and the coherence observed in ordered patterns are generated and sustained. The answer is in the fact that SSB implies the appearance of bosons (Goldstone 1961, Goldstone *et al.* 1962, Nambu and Jona-Lasinio 1961), the so-called Nambu–Goldstone (NG) modes or quanta. They manifest as long-range correlations and thus they are responsible of the above-mentioned change of scale, from microscopic to macroscopic. The coherent boson condensation of NG modes turns out to be the mechanism by which order is generated, as we will see in an explicit example in a later section.

The "Boson Transformation" Method

We now discuss the quantum origin of extended objects (defects) and show how they naturally emerge as macroscopic objects (inhomogeneous condensates) from the quantum dynamics. At zero temperature, the classical soliton solutions are then recovered in the Born approximation. This approach is known as the "boson transformation" method (Umezawa 1993, Umezawa *et al.* 1982).

The Boson Transformation Theorem

Let us consider, for simplicity, the case of a dynamical model involving one scalar field ψ_{H} and one asymptotic field φ_{in} satisfying eqns [1] and [2], respectively.

As already remarked, the dynamical map is valid only in a weak sense, that is, as a relation among matrix elements. This implies that eqn [4] is not unique, since different sets of asymptotic fields and the corresponding Hilbert spaces can be used in its construction. Let us indeed consider a c–number function $f(x)$, satisfying the φ_{in} equations of motion [2]:

$$\Lambda(\partial)f(x) = 0 \qquad [8]$$

The boson transformation theorem (Umezawa 1993, Umezawa *et al.* 1982) states that the field

$$\psi^f_{\mathrm{H}}(x) = F[x; \varphi_{\mathrm{in}} + f] \qquad [9]$$

is also a solution of the Heisenberg equation [1]. The corresponding Yang–Feldman equation takes the form

$$\psi^f_{\mathrm{H}}(x) = \varphi_{\mathrm{in}}(x) + f(x) + \Lambda^{-1}(\partial) * \mathcal{J}[\psi^f_{\mathrm{H}}](x) \qquad [10]$$

The difference between the two solutions ψ_{H} and ψ^f_{H} is only in the boundary conditions. An important point is that the expansion in [9] is obtained from that in [4] by the spacetime-dependent translation

$$\varphi_{\mathrm{in}}(x) \to \varphi_{\mathrm{in}}(x) + f(x) \qquad [11]$$

The essence of the boson transformation theorem is that the dynamics embodied in eqn [1] contains an internal freedom, represented by the possible choices of the function $f(x)$, satisfying the free-field equation [8].

We also observe that the transformation [11] is a canonical transformation since it leaves invariant the canonical form of commutation relations.

Let $|0\rangle$ denote the vacuum for the free field φ_{in}. The vacuum expectation value of eqn [10] gives

$$\begin{aligned}\phi^f(x) &\equiv \langle 0|\psi^f_{\mathrm{H}}(x)|0\rangle \\ &= f(x) + \left\langle 0\middle|\left[\Lambda^{-1}(\partial) * \mathcal{J}[\psi^f_{\mathrm{H}}](x)\right]\middle|0\right\rangle \end{aligned} \qquad [12]$$

The c–number field $\phi^f(x)$ is the order parameter. We remark that it is fully determined by the quantum dynamics. In the classical or Born approximation, which consists in taking $\langle 0|\mathcal{J}[\psi^f_{\mathrm{H}}]|0\rangle = \mathcal{J}[\phi^f]$, that is, neglecting all the contractions of the physical fields, we define $\phi^f_{\mathrm{cl}}(x) \equiv \lim_{\hbar\to 0}\phi^f(x)$. In this limit, we have

$$\Lambda(\partial)\phi^f_{\mathrm{cl}}(x) = \mathcal{J}[\phi^f_{\mathrm{cl}}](x) \qquad [13]$$

that is, $\phi^f_{\mathrm{cl}}(x)$ provides the solution of the classical Euler–Lagrange equation.

Beyond the classical level, in general, the form of this equation changes. The Yang–Feldman equation [10] gives not only the equation for the order parameter, eqn [13], but also, at higher orders in $\hbar$, the dynamics of the physical quanta in the potential generated by the "macroscopic object" $\phi^f(x)$ (Umezawa 1993, Umezawa *et al.* 1982).

One can show (Umezawa 1993, Umezawa *et al.* 1982) that the class of solutions of eqn [8] which lead to topologically nontrivial (i.e., carrying a nonzero topological charge) solutions of eqn [13], are those which have some sort of singularity with respect to Fourier transform. These can be either divergent singularities or topological singularities. The first are associated to a divergence of $f(x)$ for $|\boldsymbol{x}| = \infty$, at least in some direction. Topological singularities are instead present when $f(x)$ is not single-valued, that is, it is path dependent. In both cases, the macroscopic object described by the order parameter, carries a nonzero topological charge.

Topological Singularities and Massless Bosons

An important result is that the boson transformation functions carrying topological singularities are only allowed for massless bosons (Umezawa 1993, Umezawa *et al.* 1982).

Consider a generic boson field χ_{in} satisfying the equation

$$(\partial^2 + m^2)\chi_{\mathrm{in}}(x) = 0 \qquad [14]$$

and suppose that the function $f(x)$ for the boson transformation $\chi_{\mathrm{in}}(x) \to \chi_{\mathrm{in}}(x) + f(x)$ carries a topological singularity. It is then not single-valued and thus path dependent:

$$G^+_{\mu\nu}(x) \equiv [\partial_\mu, \partial_\nu] f(x) \neq 0, \quad \text{for certain } \mu, \nu, x \qquad [15]$$

On the other hand, $\partial_\mu f(x)$, which is related with observables, is single-valued, that is, $[\partial_\rho, \partial_\nu]$ $\partial_\mu f(x) = 0$. Recall that $f(x)$ is solution of the χ_{in} equation:

$$(\partial^2 + m^2)f(x) = 0 \qquad [16]$$

From the definition of $G^+_{\mu\nu}(x)$ and the regularity of $\partial_\mu f(x)$, it follows, by computing $\partial^\mu G^+_{\mu\nu}(x)$, that

$$\partial_\mu f(x) = \frac{1}{\partial^2 + m^2}\partial^\lambda G^+_{\lambda\mu}(x) \qquad [17]$$

This equation and the antisymmetric nature of $G^+_{\mu\nu}(x)$ then lead to $\partial^2 f(x) = 0$, which in turn implies $m = 0$. Thus, we conclude that [15] is only compatible with massless equation for χ_{in}.

The topological charge is defined as

$$N_T = \int_C dl^\mu \partial_\mu f = \int_S dS_\mu \epsilon^{\mu\nu\sigma} \partial_\nu \partial_\sigma f$$
$$= \frac{1}{2} \int_S dS^{\mu\nu} G^+_{\mu\nu} \quad [18]$$

Here C is a contour enclosing the singularity and S a surface with C as boundary. N_T does not depend on the path C provided this does not cross the singularity. The dual tensor $G^{\mu\nu}(x)$ is

$$G^{\mu\nu}(x) \equiv -\tfrac{1}{2} \epsilon^{\mu\nu\lambda\rho} G^+_{\lambda\rho}(x) \quad [19]$$

and satisfies the continuity equation

$$\partial_\mu G^{\mu\nu}(x) = 0$$
$$\Leftrightarrow \quad \partial_\mu G^+_{\lambda\rho}(x) + \partial_\rho G^+_{\mu\lambda}(x) + \partial_\lambda G^+_{\rho\mu}(x) = 0 \quad [20]$$

Equation [20] completely characterizes the topological singularity (Umezawa 1993, Umezawa *et al.* 1982).

An Example: The Anderson–Higgs–Kibble Mechanism and the Vortex Solution

We consider a model of a complex scalar field $\phi(x)$ interacting with a gauge field $A_\mu(x)$ (Anderson 1958, Higgs 1960, Kibble 1967). The lagrangian density $\mathcal{L}[\phi(x), \phi^*(x), A_\mu(x)]$ is invariant under the global and the local U(1) gauge transformations (we do not assume a particular form for the Lagrangian density, so the following results are quite general):

$$\phi(x) \to e^{i\theta} \phi(x), \quad A_\mu(x) \to A_\mu(x) \quad [21]$$

$$\phi(x) \to e^{ie_0 \lambda(x)} \phi(x), \quad A_\mu(x) \to A_\mu(x) + \partial_\mu \lambda(x) \quad [22]$$

respectively, where $\lambda(x) \to 0$ for $|x_0| \to \infty$ and/or $|\boldsymbol{x}| \to \infty$ and e_0 is the coupling constant. We work in the Lorentz gauge $\partial_\mu A^\mu(x) = 0$. The generating functional, including the gauge constraint, is (Matsumoto *et al.* 1975a, b)

$$\mathcal{Z}[J, K] = \frac{1}{\mathcal{N}} \int [dA_\mu][d\phi][d\phi^*][dB]$$
$$\times \exp\left[i\, \mathcal{S}[A_\mu, B, \phi]\right] \quad [23]$$

$$\mathcal{S} = \int d^4x \Big[\mathcal{L}(x) + B(x)\partial^\mu A_\mu(x)$$
$$+ K^*(x)\phi(x) + K(x)\phi^*(x)$$
$$+ J^\mu(x) A_\mu(x) + i\epsilon |\phi(x) - v|^2\Big]$$
$$\mathcal{N} = \int [dA_\mu][d\phi][d\phi^*][dB]$$
$$\times \exp\left[i \int d^4x \left(\mathcal{L}(x) + i\epsilon |\phi(x) - v|^2\right)\right]$$

$B(x)$ is an auxiliary field which implements the gauge-fixing condition (Matsumoto *et al.* 1975a, b). Notice the ϵ-term where v is a complex number; its rôle is to specify the condition of symmetry breaking under which we want to compute the functional integral and it may be given the physical meaning of a small external field triggering the symmetry breaking (Matsumoto *et al.* 1975a, b). The limit $\epsilon \to 0$ must be made at the end of the computations. We will use the notation

$$\langle F[\phi] \rangle_{\epsilon, J, K} \equiv \frac{1}{\mathcal{N}} \int [dA_\mu][d\phi][d\phi^*][dB] F[\phi]$$
$$\times \exp\left[i\mathcal{S}[A_\mu, B, \phi]\right] \quad [24]$$

with $\langle F[\phi] \rangle_\epsilon \equiv \langle F[\phi] \rangle_{\epsilon, J=K=0}$ and $\langle F[\phi] \rangle \equiv \lim_{\epsilon \to 0} \langle F[\phi] \rangle_\epsilon$.

The fields ϕ, A_μ, and B appearing in the generating functional are c-number fields. In the following, the Heisenberg operator fields corresponding to them will be denoted by $\phi_H, A_{H\mu}$, and B_H, respectively. Thus, the spontaneous symmetry breaking condition is expressed by $\langle 0 | \phi_H(x) | 0 \rangle \equiv \tilde{v} \neq 0$, with $\tilde{v}$ constant.

Since in the functional integral formalism the functional average of a given c-number field gives the vacuum expectation value of the corresponding operator field, for example, $\langle F[\phi] \rangle \equiv \langle 0 | F[\phi_H] | 0 \rangle$, we have $\lim_{\epsilon \to 0} \langle \phi(x) \rangle_\epsilon \equiv \langle 0 | \phi_H(x) | 0 \rangle = \tilde{v}$.

Let us introduce the following decompositions:

$$\phi(x) = \frac{1}{\sqrt{2}} [\psi(x) + i\chi(x)]$$
$$K(x) = \frac{1}{\sqrt{2}} [K_1(x) + iK_2(x)]$$
$$\rho(x) \equiv \psi(x) - \langle \psi(x) \rangle_\epsilon$$

Note that $\langle \chi(x) \rangle_\epsilon = 0$ because of the invariance under $\chi \to -\chi$.

The Goldstone Theorem

Since the functional integral [23] is invariant under the global transformation [21], we have that $\partial \mathcal{Z}[J, K]/\partial\theta = 0$ and subsequent derivatives with respect to K_1 and K_2 lead to

$$\langle \psi(x) \rangle_\epsilon = \sqrt{2} \epsilon v \int d^4y \langle \chi(x) \chi(y) \rangle_\epsilon$$
$$= \sqrt{2} \epsilon v \Delta_\chi(\epsilon, 0) \quad [25]$$

In momentum space the propagator for the field χ has the general form

$$\Delta_\chi(0, p) = \lim_{\epsilon \to 0} \left[\frac{Z_\chi}{p^2 - m_\chi^2 + i\epsilon a_\chi}\right.$$
$$\left. + (\text{continuum contributions})\right] \quad [26]$$

Here Z_χ and a_χ are renormalization constants. The integration in eqn [25] picks up the pole contribution at $p^2=0$, and leads to

$$\tilde{v}=\sqrt{2}\frac{Z_\chi}{a_\chi}v \Leftrightarrow m_\chi=0, \quad \tilde{v}=0 \Leftrightarrow m_\chi\neq 0 \qquad [27]$$

The Goldstone theorem (Goldstone 1961, Goldstone *et al.* 1962) is thus proved: if the symmetry is spontaneously broken ($\tilde{v}\neq 0$), a massless mode must exist, whose field is $\chi(x)$, that is, the NG boson mode. Since it is massless, it manifests as a long-range correlation mode. (Notice that in the present case of a complex scalar field model, the NG mode is an elementary field. In other models, it may appear as a bound state, for example, the magnon in (anti)ferromagnets.) Note that

$$\frac{\partial}{\partial v}\langle\psi(x)\rangle_\epsilon=\sqrt{2}\epsilon\int \mathrm{d}^4y\langle\rho(x)\rho(y)\rangle_\epsilon \qquad [28]$$

and because $m_\rho\neq 0$, the right-hand side of this equation vanishes in the limit $\epsilon\to 0$; therefore, $\tilde{v}$ is independent of $|v|$, although the phase of $|v|$ determines the one of $\tilde{v}$ (from eqn [25]): as in ferromagnets, once an external magnetic field is switched on, the system is magnetized independently of the strength of the external field.

The Dynamical Map and the Field Equations

Observing that the change of variables [21] (and/or [22]) does not affect the generating functional, we may obtain the Ward–Takahashi identities. Also, using $B(x)\to B(x)+\lambda(x)$ in [23] gives $\langle\partial^\mu A_\mu(x)\rangle_{\epsilon,J,K}=0$. One then finds the following two-point function pole structures (Matsumoto *et al.* 1975a, b):

$$\langle B(x)\chi(y)\rangle=\lim_{\epsilon\to 0}\left\{\frac{-\mathrm{i}}{(2\pi)^4}\int \mathrm{d}^4p\,\mathrm{e}^{-\mathrm{i}p(x-y)}\frac{e_0\tilde{v}}{p^2+\mathrm{i}\epsilon a_\chi}\right\} \qquad [29]$$

$$\langle B(x)A^\mu(y)\rangle=\partial_x^\mu\frac{\mathrm{i}}{(2\pi)^4}\int \mathrm{d}^4p\,\mathrm{e}^{-\mathrm{i}p(x-y)}\frac{1}{p^2} \qquad [30]$$

$$\langle B(x)B(y)\rangle=\lim_{\epsilon\to 0}\left\{\frac{-\mathrm{i}}{(2\pi)^4}\int \mathrm{d}^4p\,\mathrm{e}^{-\mathrm{i}p(x-y)}\frac{(e_0\tilde{v})^2}{Z_\chi}\times\left[\frac{1}{p^2+\mathrm{i}\epsilon a_\chi}-\frac{1}{p^2}\right]\right\} \qquad [31]$$

The absence of branch-cut singularities in propagators [29]–[31] suggests that $B(x)$ obeys a free-field equation. In addition, eqn [31] indicates that the model contains a massless negative-norm state (ghost) besides the NG massless mode χ. Moreover, it can be shown (Matsumoto *et al.* 1975a, b) that a massive vector field U^μ_{in} also exists in the theory. Note that because of the invariance $(\chi, A_\mu, B)\to(-\chi, -A_\mu, -B)$, all the other two-point functions must vanish.

The dynamical maps expressing the Heisenberg operator fields in terms of the asymptotic operator fields are found to be (Matsumoto *et al.* 1975a, b)

$$\phi_{\mathrm{H}}(x)=:\exp\left\{\mathrm{i}\frac{Z_\chi^{1/2}}{\tilde{v}}\chi_{\mathrm{in}}(x)\right\}\left[\tilde{v}+Z_\rho^{1/2}\rho_{\mathrm{in}}(x)+\mathcal{F}[\rho_{\mathrm{in}}, U^\mu_{\mathrm{in}}, \partial(\chi_{\mathrm{in}}-b_{\mathrm{in}})]\right]: \qquad [32]$$

$$A^\mu_{\mathrm{H}}(x)=Z_3^{1/2}U^\mu_{\mathrm{in}}(x)+\frac{Z_\chi^{1/2}}{e_0\tilde{v}}\partial^\mu b_{\mathrm{in}}(x)+:\mathcal{F}^\mu[\rho_{\mathrm{in}}, U^\mu_{\mathrm{in}}, \partial(\chi_{\mathrm{in}}-b_{\mathrm{in}})]: \qquad [33]$$

$$B_{\mathrm{H}}(x)=\frac{e_0\tilde{v}}{Z_\chi^{1/2}}[b_{\mathrm{in}}(x)-\chi_{\mathrm{in}}(x)]+c \qquad [34]$$

where $:\ldots:$ denotes the normal ordering and the functionals $\mathcal{F}$ and $\mathcal{F}^\mu$ are to be determined within a particular model. In eqns [32]–[34], χ_{in} denotes the NG mode, b_{in} the ghost mode, U^μ_{in} the massive vector field, and ρ_{in} the massive matter field. In eqn [34] c is a c-number constant, whose value is irrelevant since only derivatives of B appear in the field equations (see below). Z_3 represents the wave function renormalization for U^μ_{in}. The corresponding field equations are

$$\partial^2\chi_{\mathrm{in}}(x)=0, \qquad \partial^2 b_{\mathrm{in}}(x)=0, \qquad (\partial^2+m_\rho^2)\rho_{\mathrm{in}}(x)=0 \qquad [35]$$

$$(\partial^2+m_V^2)U^\mu_{\mathrm{in}}(x)=0, \quad \partial_\mu U^\mu_{\mathrm{in}}(x)=0 \qquad [36]$$

with $m_V{}^2=(Z_3/Z_\chi)(e_0\tilde{v})^2$. The field equations for B_{H} and $A_{\mathrm{H}\mu}$ read (Matsumoto *et al.* 1975a, b)

$$\partial^2 B_{\mathrm{H}}(x)=0, \quad -\partial^2 A_{\mathrm{H}\mu}(x)=j_{\mathrm{H}\mu}(x)-\partial_\mu B_{\mathrm{H}}(x) \qquad [37]$$

with $j_{\mathrm{H}\mu}(x)=\delta\mathcal{L}(x)/\delta A^\mu_{\mathrm{H}}(x)$. One may then require that the current $j_{\mathrm{H}\mu}$ is the only source of the gauge field $A_{\mathrm{H}\mu}$ in any observable process. This amounts to impose the condition: ${}_p\langle b|\partial_\mu B_{\mathrm{H}}(x)|a\rangle_p=0$, that is,

$$(-\partial^2)_p\langle b|A^0_{\mathrm{H}\mu}(x)|a\rangle_p={}_p\langle b|\,j_{\mathrm{H}\mu}(x)|a\rangle_p \qquad [38]$$

where $|a\rangle_p$ and $|b\rangle_p$ denote two generic physical states and $A^{0\mu}_{\mathrm{H}}(x)\equiv A^\mu_{\mathrm{H}}(x)-e_0\tilde{v}:\partial^\mu b_{\mathrm{in}}(x):$. Equations [38] are the classical Maxwell equations. The condition ${}_p\langle b|\partial_\mu B_{\mathrm{H}}(x)|a\rangle_p=0$ leads to the Gupta–Bleuler–like condition

$$[\chi^{(-)}_{\mathrm{in}}(x)-b^{(-)}_{\mathrm{in}}(x)]|a\rangle_p=0 \qquad [39]$$

where $\chi^{(-)}_{\mathrm{in}}$ and $b^{(-)}_{\mathrm{in}}$ are the positive-frequency parts of the corresponding fields. Thus, we see that χ_{in} and b_{in} cannot participate in any observable reaction.

This is confirmed by the fact that they are present in the S-matrix in the combination $(\chi_{\text{in}} - b_{\text{in}})$ (Matsumoto *et al.* 1975a, b). It is to be remarked, however, that the NG boson does not disappear from the theory: we shall see below that there are situations in which the NG fields do have observable effects.

The Dynamical Rearrangement of Symmetry and the Classical Fields and Currents

From eqns [32]–[33] we see that the local gauge transformations of the Heisenberg fields

$$\begin{aligned} &\phi_{\text{H}}(x) \to \mathrm{e}^{\mathrm{i}e_0\lambda(x)}\phi_{\text{H}}(x) \\ &A^{\mu}_{\text{H}}(x) \to A^{\mu}_{\text{H}}(x) + \partial^{\mu}\lambda(x), \quad B_{\text{H}}(x) \to B_{\text{H}}(x) \end{aligned} \tag{40}$$

with $\partial^2\lambda(x) = 0$, are induced by the in-field transformations

$$\begin{aligned} &\chi_{\text{in}}(x) \to \chi_{\text{in}}(x) + \frac{e_0\tilde{v}}{Z_{\chi}^{1/2}}\lambda(x) \\ &b_{\text{in}}(x) \to b_{\text{in}}(x) + \frac{e_0\tilde{v}}{Z_{\chi}^{1/2}}\lambda(x) \\ &\rho_{\text{in}}(x) \to \rho_{\text{in}}(x), \quad U^{\mu}_{\text{in}}(x) \to U^{\mu}_{\text{in}}(x) \end{aligned} \tag{41}$$

On the other hand, the global phase transformation $\phi_{\text{H}}(x) \to \mathrm{e}^{\mathrm{i}\theta}\phi_{\text{H}}(x)$ is induced by

$$\begin{aligned} &\chi_{\text{in}}(x) \to \chi_{\text{in}}(x) + \frac{\tilde{v}}{Z_{\chi}^{1/2}}\theta f(x), \quad b_{\text{in}}(x) \to b_{\text{in}}(x) \\ &\rho_{\text{in}}(x) \to \rho_{\text{in}}(x), \quad U^{\mu}_{\text{in}}(x) \to U^{\mu}_{\text{in}}(x) \end{aligned} \tag{42}$$

with $\partial^2 f(x) = 0$ and the limit $f(x) \to 1$ to be performed at the end of computations. Note that under the above transformations, the in-field equations and the S-matrix are invariant and that B_{H} is changed by an irrelevant c-number (in the limit $f \to 1$).

Consider now the boson transformation $\chi_{\text{in}}(x) \to \chi_{\text{in}}(x) + \alpha(x)$: in local gauge theories the boson transformation must be compatible with the Heisenberg field equations but also with the physical state condition [39]. Under the boson transformation with $\alpha(x) = \tilde{v}Z_{\chi}^{-1/2}\theta f(x)$ and $\partial^2 f(x) = 0, B_{\text{H}}$ changes as

$$B_{\text{H}}(x) \to B_{\text{H}}(x) - \frac{e_0\tilde{v}^2}{Z_{\chi}}f(x) \tag{43}$$

eqn [38] is thus violated when the Gupta–Bleuler-like condition is imposed. In order to restore it, the shift in B_{H} must be compensated by means of the following transformation on U^{μ}_{in}:

$$U^{\mu}_{\text{in}}(x) \to U^{\mu}_{\text{in}}(x) + Z_3^{-1/2}a^{\mu}(x), \quad \partial_{\mu}a^{\mu}(x) = 0 \tag{44}$$

with a convenient c-number function $a^{\mu}(x)$. The dynamical maps of the various Heisenberg operators are not affected by [44] since they contain U^{μ}_{in} and B_{H} in a combination such that the changes of B_{H} and of U^{μ}_{in} compensate each other provided

$$(\partial^2 + m_V^2)a_{\mu}(x) = \frac{m_V^2}{e_0}\partial_{\mu}f(x) \tag{45}$$

Equation [45] thus obtained is the Maxwell equation for the massive potential vector a_{μ} (Matsumoto *et al.* 1975a, b). The classical ground state current j^{μ} turns out to be

$$j^{\mu}(x) \equiv \langle 0|j^{\mu}_{\text{H}}(x)|0\rangle = m_V^2\left[a^{\mu}(x) - \frac{1}{e_0}\partial^{\mu}f(x)\right] \tag{46}$$

The term $m_V^2 a^{\mu}(x)$ is the Meissner current, while $(m_V^2/e_0)\partial^{\mu}f(x)$ is the boson current. The key point here is that both the macroscopic field and current are given in terms of the boson condensation function $f(x)$.

Two remarks are in order: first, note that the terms proportional to $\partial^{\mu}f(x)$ are related to observable effects, for example, the boson current which acts as the source of the classical field. Second, note that the macroscopic ground state effects do not occur for regular $f(x)(G^{+}_{\mu\nu}(x) = 0)$. In fact, from [45] we obtain $a_{\mu}(x) = (1/e_0)\partial_{\mu}f(x)$ for regular $f(x)$ which implies zero classical current $(j_{\mu} = 0)$ and zero classical field $(F_{\mu\nu} = \partial_{\mu}a_{\nu} - \partial_{\nu}a_{\mu})$, since the Meissner and the boson current cancel each other.

In conclusion, the vacuum current appears only when $f(x)$ has topological singularities and these can be created only by condensation of massless bosons, that is, when SSB occurs. This explains why topological defects appear in the process of phase transitions, where NG modes are present and gradients in their condensate densities are nonzero (Kibble 1976, Zurek 1997).

On the other hand, the appearance of spacetime order parameter is no guarantee that persistent ground state currents (and fields) will exist: if $f(x)$ is a regular function, the spacetime dependence of $\tilde{v}$ can be gauged away by an appropriate gauge transformation.

Since, as already mentioned, the boson transformation with regular $f(x)$ does not affect observable quantities, the S-matrix is actually given by

$$S = :S\left[\rho_{\text{in}}, U^{\mu}_{\text{in}} - \frac{1}{m_V}\partial(\chi_{\text{in}} - b_{\text{in}})\right]: \tag{47}$$

This is indeed independent of the boson transformation with regular $f(x)$:

$$\begin{aligned} S \to S' = :S\Big[&\rho_{\text{in}}, U^{\mu}_{\text{in}} - \frac{1}{m_V}\partial(\chi_{\text{in}} - b_{\text{in}}) \\ &+ Z_3^{-1/2}\left(a^{\mu} - \frac{1}{e_0}\partial^{\mu}f\right)\Big]: \end{aligned} \tag{48}$$

since $a_\mu(x) = (1/e_0)\partial_\mu f(x)$ for regular $f(x)$. However, $S' \neq S$ for singular $f(x)$: S' includes the interaction of the quanta U^μ_{in} and ϕ_{in} with the classically behaving macroscopic defects (Umezawa 1993, Umezawa *et al.* 1982).

The Vortex Solution

Below we consider the example of the Nielsen–Olesen vortex string solution. We show which one is the boson function $f(x)$ controlling the nonhomogeneous NG boson condensation in terms of which the string solution is described. For brevity, we only report the results of the computations. The detailed derivation as well as the discussion of further examples can be found in (Umezawa 1993, Umezawa *et al.* 1982).

In the present U(1) problem, the electromagnetic tensor and the vacuum current are (Umezawa 1993, Umezawa *et al.* 1982, Matsumoto *et al.* 1975a, b)

$$\begin{aligned} F_{\mu\nu}(x) &= \partial_\mu a_\nu(x) - \partial_\nu a_\mu(x) \\ &= 2\pi \frac{m_V^2}{e_0} \int \mathrm{d}^4x' \Delta_c(x-x') G^+_{\mu\nu}(x') \end{aligned} \quad [49]$$

$$j_\mu(x) = -2\pi \frac{m_V^2}{e_0} \int \mathrm{d}^4x' \Delta_c(x-x') \partial^\nu_{x'} G^+_{\nu\mu}(x') \quad [50]$$

respectively, and satisfy $\partial^\mu F_{\mu\nu}(x) = -j_\nu(x)$. In these equations,

$$\Delta_c(x-x') = \frac{1}{(2\pi)^4} \int \mathrm{d}^4p\, \mathrm{e}^{-\mathrm{i}p(x-x')} \frac{1}{p^2 - m_V^2 + \mathrm{i}\epsilon} \quad [51]$$

The line singularity for the vortex (or string) solution can be parametrized by a single line parameter σ and by the time parameter τ. A static vortex solution is obtained by setting $y_0(\tau,\sigma) = \tau$ and $\boldsymbol{y}(\tau,\sigma) = \boldsymbol{y}(\sigma)$, with y denoting the line coordinate. $G^+_{\mu\nu}(x)$ is nonzero only on the line at y (we can consider more lines but let us limit to only one line, for simplicity). Thus, we have

$$\begin{aligned} &G_{0i}(x) = \int \mathrm{d}\sigma \frac{\mathrm{d}y_i(\sigma)}{\mathrm{d}\sigma} \delta^3[\boldsymbol{x} - \boldsymbol{y}(\sigma)] \;\; G_{ij}(x) = 0 \\ &G^+_{ij}(x) = -\epsilon_{ijk} G_{0k}(x), \quad G^+_{0i}(x) = 0 \end{aligned} \quad [52]$$

Equation [49] shows that these vortices are purely magnetic. We obtain

$$\begin{aligned} \partial_0 f(x) &= 0 \\ \partial_i f(x) &= \frac{1}{(2\pi)^2} \int \mathrm{d}\sigma \epsilon_{ijk} \frac{\mathrm{d}y_k(\sigma)}{\mathrm{d}\sigma} \partial^x_j \\ &\quad \times \int \mathrm{d}^3p \frac{\mathrm{e}^{\mathrm{i}\boldsymbol{p}\cdot(\boldsymbol{x} - \boldsymbol{y}(\sigma))}}{\boldsymbol{p}^2} \end{aligned} \quad [53]$$

that is, by using the identity $(2\pi)^{-2} \int \mathrm{d}^3p (\mathrm{e}^{\mathrm{i}p\cdot x}/\boldsymbol{p}^2) = 1/2|\boldsymbol{x}|$,

$$\nabla f(x) = -\frac{1}{2} \int \mathrm{d}\sigma \frac{\mathrm{d}y_k(\sigma)}{\mathrm{d}\sigma} \wedge \nabla_x \frac{1}{|\boldsymbol{x} - \boldsymbol{y}(\sigma)|} \quad [54]$$

Note that $\nabla^2 f(x) = 0$ is satisfied.

A straight infinitely long vortex is specified by $y_i(\sigma) = \sigma\delta_{i3}$ with $-\infty < \sigma < \infty$. The only nonvanishing component of $G^{\mu\nu}(x)$ are $G^{03}(x) = G^+_{12}(x) = \delta(x_1)\delta(x_2)$. Equation [54] gives (Umezawa 1993, Umezawa *et al.* 1982, Matsumoto 1975a, b)

$$\begin{aligned} \frac{\partial}{\partial x_1} f(x) &= \frac{1}{2} \int \mathrm{d}\sigma \frac{\partial}{\partial x_2} [x_1^2 + x_2^2 + (x_3 - \sigma)^2]^{-1/2} \\ &= -\frac{x_2}{x_1^2 + x_2^2} \\ \frac{\partial}{\partial x_2} f(x) &= \frac{x_1}{x_1^2 + x_2^2}, \quad \frac{\partial}{\partial x_3} f(x) = 0 \end{aligned} \quad [55]$$

and then

$$f(x) = \tan^{-1}\left(\frac{x_2}{x_1}\right) = \theta(x) \quad [56]$$

We have thus determined the boson transformation function corresponding to a particular vortex solution. The vector potential is

$$\begin{aligned} a_1(x) &= -\frac{m_V^2}{2e_0} \int \mathrm{d}^4x' \Delta_c(x-x') \frac{x'_2}{x'^2_1 + x'^2_2} \\ a_2(x) &= \frac{m_V^2}{2e_0} \int \mathrm{d}^4x' \Delta_c(x-x') \frac{x'_1}{x'^2_1 + x'^2_2} \\ a_3(x) &= a_0(x) = 0 \end{aligned} \quad [57]$$

and the only nonvanishing component of $F_{\mu\nu}$:

$$\begin{aligned} F_{12}(x) &= -2\pi \frac{m_V^2}{e_0} \int \mathrm{d}^4x' \Delta_c(x-x') \delta(x'_1) \delta(x'_2) \\ &= \frac{m_V^2}{e_0} K_0\left(m_V \sqrt{x_1^2 + x_2^2}\right) \end{aligned} \quad [58]$$

Finally, the vacuum current eqn [50] is given by

$$\begin{aligned} j_1(x) &= -\frac{m_V^3}{e_0} \frac{x_2}{\sqrt{x_1^2 + x_2^2}} K_1\left(m_V \sqrt{x_1^2 + x_2^2}\right) \\ j_2(x) &= \frac{m_V^3}{e_0} \frac{x_1}{\sqrt{x_1^2 + x_2^2}} K_1\left(m_V \sqrt{x_1^2 + x_2^2}\right) \\ j_3(x) &= j_0(x) = 0 \end{aligned} \quad [59]$$

We observe that these results are the same of the Nielsen–Olesen vortex solution. Notice that we did not specify the potential in our model but only the invariance properties. Thus, the invariance properties of the dynamics determine the characteristics of the topological solutions. The vortex solution

manifests the original U(1) symmetry through the cylindrical angle θ which is the parameter of the U(1) representation in the coordinate space.

Conclusions

We have discussed how topological defects arise as inhomogeneous condensates in QFT. Topological defects are shown to have a genuine quantum nature. The approach reviewed here goes under the name of "boson transformation method" and relies on the existence of unitarily inequivalent representations of the field algebra in QFT.

Describing quantum fields with topological defects amounts then to properly choose the physical Fock space for representing the Heisenberg field operators. Once the boundary conditions corresponding to a particular soliton sector are found, the Heisenberg field operators embodied with such conditions contain the full information about the defects, the quanta and their mutual interaction. One can thus calculate Green's functions for particles in the presence of defects. The extension to finite temperature is discussed in Blasone and Jizba (2002) and Manka and Vitiello (1990).

As an example we have discussed a model with U(1) gauge invariance and SSB and we have obtained the Nielsen–Olesen vortex solution in terms of localized condensation of Goldstone bosons. These thus appear to play a physical role, although, in the presence of gauge fields, they do not show up in the physical spectrum as excitation quanta. The function $f(x)$ controlling the condensation of the NG bosons must be singular in order to produce observable effects. Boson transformations with regular $f(x)$ only amount to gauge transformations. For the treatment of topological defects in nonabelian gauge theories, see Manka and Vitiello (1990).

Finally, when there are no NG modes, as in the case of the kink solution or the sine-Gordon solution, the boson transformation function has to carry divergence singularity at spatial infinity (Umezawa 1993, Umezawa *et al.* 1982, Blasone and Jizba 2002). The boson transformation has also been discussed in connection with the Bäklund transformation at a classical level and the confinement of the constituent quanta in the coherent condensation domain.

For further reading on quantum fields with topological defects, see Blasone *et al.* (2006).

Acknowledgments

The authors thank MIUR, INFN, INFM, and the ESF network COSLAB for partial financial support.

See also: Abelian Higgs Vortices; Algebraic Approach to Quantum Field Theory; Quantum Field Theory: A Brief Introduction; Quantum Field Theory in Curved Spacetime; Symmetries in Quantum Field Theory: Algebraic Aspects; Symmetries in Quantum Field Theory of Lower Spacetime Dimensions; Topological Defects and their Homotopy Classification.

Further Reading

Anderson PW (1958) Coherent excited states in the theory of superconductivity: gauge invariance and the Meissner effect. *Physical Review* 110: 827–835.

Blasone M and Jizba P (2002) Topological defects as inhomogeneous condensates in quantum field theory: kinks in $(1+1)$ dimensional $\lambda\psi^4$ theory. *Annals of Physics* 295: 230–260.

Blasone M, Jizba P, and Vitiello G (2006) *Spontaneous Breakdown of Symmetry and Topological Defects*, London: Imperial College Press. (in preparation).

Goldstone J (1961) Field theories with "superconductor" solutions. *Nuovo Cimento* 19: 154–164.

Goldstone J, Salam A, and Weinberg S (1962) Broken symmetries. *Physical Review* 127: 965–970.

Higgs P (1960) Spontaneous symmetry breakdown without massless bosons. *Physical Review* 145: 1156–1163.

Kibble TWB (1967) Symmetry breaking in non-abelian gauge theories. *Physical Review* 155: 1554–1561.

Kibble TWB (1976) Topology of cosmic domains and strings. *Journal of Physics A* 9: 1387–1398.

Kibble TWD (1980) Some implications of a cosmological phase transition. *Physics Reports* 67: 183–199.

Kleinert H (1989) *Gauge Fields in Condensed Matter*, vols. I & II Singapore: World Scientific.

Manka R and Vitiello G (1990) Topological solitons and temperature effects in gauge field theory. *Annals of Physics* 199: 61–83.

Matsumoto H, Papastamatiou NJ, Umezawa H, and Vitiello G (1975a) Dynamical rearrangement in Anderson–Higgs–Kibble mechanism. *Nuclear Physics B* 97: 61–89.

Matsumoto H, Papastamatiou NJ, and Umezawa H (1975b) The boson transformation and the vortex solutions. *Nuclear Physics B* 97: 90–124.

Nambu Y and Jona-Lasinio G (1961) Dynamical model of elementary particles based on an analogy with superconductivity. I. *Physical Review* 122: 345–358.

Nambu Y and Jona-Lasinio G (1961) Dynamical model of elementary particles based on an analogy with superconductivity. II. *Physical Review* 124: 246–254.

Rajaraman R (1982) *Solitons and Instantons: An Introduction to Solitons and Instantons in Quantum Field Theory*. Amsterdam: North-Holland.

Umezawa H (1993) *Advanced Field Theory: Micro, Macro and Thermal Physics*. New York: American Institute of Physics.

Umezawa H, Matsumoto H, and Tachiki M (1982) *Thermo Field Dynamics and Condensed States*. Amsterdam: North-Holland.

Vitiello G (2001) *My Double Unveiled*. Amsterdam: John Benjamins.

Volovik GE (2003) *The Universe in a Helium Droplet*. Oxford: Clarendon.

von Neumann J (1955) *Mathematical Foundation of Quantum Mechanics*. Princeton: Princeton University Press.

Zurek WH (1997) Cosmological experiments in condensed matter systems. *Physics Reports* 276: 177–221.

Quantum Geometry and Its Applications

A Ashtekar, Pennsylvania State University, University Park, PA, USA
J Lewandowski, Uniwersyte Warszawski, Warsaw, Poland

Introduction

In general relativity, the gravitational field is encoded in the Riemannian geometry of spacetime. Much of the conceptual compactness and mathematical elegance of the theory can be traced back to this central idea. The encoding is also directly responsible for the most dramatic ramifications of the theory: the big bang, black holes, and gravitational waves. However, it also leads one to the conclusion that spacetime itself must end and physics must come to a halt at the big bang and inside black holes, where the gravitational field becomes singular. But this reasoning ignores quantum physics entirely. When the curvature becomes large, of the order of $1/\ell_{\rm Pl}^2 = c^3/G\hbar$, quantum effects dominate and predictions of general relativity can no longer be trusted. In this "Planck regime," one must use an appropriate synthesis of general relativity and quantum physics, that is, a quantum gravity theory. The predictions of this theory are likely to be quite different from those of general relativity. In the real, quantum world, evolution may be completely nonsingular. Physics may not come to a halt and quantum theory could extend classical spacetime.

There are a number of different approaches to quantum gravity. One natural avenue is to retain the interplay between gravity and geometry but now use "quantum" Riemannian geometry in place of the standard, classical one. This is the key idea underlying loop quantum gravity. There are several calculations which indicate that the well-known failure of the standard perturbative approach to quantum gravity may be primarily due to its basic assumption that spacetime can be modeled as a smooth continuum at all scales. In loop quantum gravity, one adopts a nonperturbative approach. There is no smooth metric in the background. Geometry is not only dynamical but quantum mechanical from "birth." Its fundamental excitations turn out to be one dimensional and polymer-like. The smooth continuum is only a coarse-grained approximation. While a fully satisfactory quantum gravity theory still awaits us (in any approach), detailed investigations have been carried out to completion in simplified models – called mini- and midi-superspaces. They show that quantum spacetime does not end at singularities. Rather, quantum geometry serves as a "bridge" to another large classical spacetime.

This article will focus on structural issues from the perspective of mathematical physics. For complementary perspectives and further details, *see* Loop Quantum Gravity, Canonical General Relativity, Quantum Cosmology, Black Hole Mechanics, and Spin Foams in this Encyclopedia.

Basic Framework

The starting point is a Hamiltonian formulation of general relativity based on spin connections (Ashtekar 1987). Here, the phase space $\mathbf{\Gamma}$ consists of canonically conjugate pairs $(A, \mathbb{P})$, where A is a connection on a 3-manifold M and $\mathbb{P}$ a 2-form, both of which take values in the Lie algebra su(2). Since $\mathbf{\Gamma}$ can also be thought of as the phase space of the SU(2) Yang–Mills theory, in this approach there is a unified kinematic framework for general relativity that describes gravity and the gauge theories which describe the other three basic forces of nature. The connection A enables one to parallel transport chiral spinors (such as the left-handed fermions of the standard electroweak model) along curves in M. Its curvature is directly related to the electric and magnetic parts of the spacetime "Riemann tensor." The dual P of $\mathbb{P}$ plays a double role (the dual is defined via $\int_M \mathbb{P} \wedge \omega = \int_M P \lrcorner \omega$ for any 1-form ω on M). Being the momentum canonically conjugate to A, it is analogous to the Yang–Mills electric field. But (apart from a constant), it is also an orthonormal triad (with density weight 1) on M and therefore determines the positive-definite ("spatial") 3-metric, and hence the Riemannian geometry of M. This dual role of P is a reflection of the fact that now SU(2) is the (double cover of the) group of rotations of the orthonormal spatial triads on M itself rather than of rotations in an "internal" space associated with M.

To pass to quantum theory, one first constructs an algebra of "elementary" functions on $\mathbf{\Gamma}$ (analogous to the phase-space functions x and p in the case of a particle) which are to have unambiguous operator analogs. The holonomies

$$h_e(A) := \mathcal{P} \exp - \int_e A \qquad [1]$$

associated with a curve/edge e on M are (SU(2)-valued) configuration functions on $\mathbf{\Gamma}$. Similarly,

given a 2-surface S on M, and an su(2)-valued (test) function f on M,

$$P_{S,f} := \int_S \mathrm{tr}(f\,\mathbb{P}) \qquad [2]$$

is a momentum function on $\mathbf{\Gamma}$, where tr is over the su(2) indices. (For simplicity of presentation, all fields are assumed to be smooth and curves/edges e and surfaces S, finite and piecewise analytic in a specific sense. The extension to smooth curves and surfaces was carried out by Bacz and Sawin, Lewandowski and Thiemann, and Fleischhack. It is technically more involved but the final results are qualitatively the same.) The symplectic structure on $\mathbf{\Gamma}$ enables one to calculate the Poisson brackets $\{h_e, P_{S,f}\}$. The result is a linear combination of holonomies and can be written as a Lie derivative,

$$\{h_e, P_{S,f}\} = \mathcal{L}_{X_{S,f}} h_e \qquad [3]$$

where $X_{S,f}$ is a derivation on the ring generated by holonomy functions, and can therefore be regarded as a vector field on the configuration space $\mathcal{A}$ of connections. This is a familiar situation in classical mechanics of systems whose configuration space is a finite-dimensional manifold. Functions h_e and vector fields $X_{S,f}$ generate a Lie algebra. As in quantum mechanics on manifolds, the first step is to promote this algebra to a quantum algebra by demanding that the commutator be given by $i\hbar$ times the Lie bracket. The result is a $\star$-algebra $\mathfrak{a}$, analogous to the algebra generated by operators $\exp i\lambda\hat{x}$ and $\hat{p}$ in quantum mechanics. By exponentiating the momentum operators $\hat{P}_{S,f}$ one obtains $\mathfrak{W}$, the analog of the quantum-mechanical Weyl algebra generated by $\exp i\lambda\hat{x}$ and $\exp i\mu\hat{p}$.

The main task is to obtain the appropriate representation of these algebras. In that representation, quantum Riemannian geometry can be probed through the momentum operators $\hat{P}_{S,f}$, which stem from classical orthonormal triads. As in quantum mechanics on manifolds or simple field theories in flat space, it is convenient to divide the task into two parts. In the first, one focuses on the algebra $\mathfrak{C}$ generated by the configuration operators $\hat{h}_c$ and finds all its representations, and in the second one considers the momentum operators $\hat{P}_{S,f}$ to restrict the freedom.

$\mathfrak{C}$ is called the holonomy algebra. It is naturally endowed with the structure of an abelian $C^\star$ algebra (with identity), whence one can apply the powerful machinery made available by the Gel'fand theory. This theory tells us that $\mathfrak{C}$ determines a unique compact, Hausdorff space $\bar{\mathcal{A}}$ such that the $C^\star$ algebra of all continuous functions on $\mathcal{A}$ is naturally isomorphic to $\mathfrak{C}$. $\bar{\mathcal{A}}$ is called the Gel'fand spectrum of $\mathfrak{C}$. It has been shown to consist of "generalized connections" $\bar{A}$ defined as follows: $\bar{A}$ assigns to any oriented edge e in M an element $\bar{A}(e)$ of SU(2) (a "holonomy") such that $\bar{A}(e^{-1}) = [\bar{A}(e)]^{-1}$; and, if the endpoint of e_1 is the starting point of e_2, then $\bar{A}(e_1 \circ e_2) = \bar{A}(e_1)\cdot \bar{A}(e_2)$. Clearly, every smooth connection A is a generalized connection. In fact, the space $\mathcal{A}$ of smooth connections has been shown to be dense in $\bar{\mathcal{A}}$ (with respect to the natural Gel'fand topology thereon). But $\bar{\mathcal{A}}$ has many more "distributional elements." The Gel'fand theory guarantees that every representation of the $C^\star$ algebra $\mathfrak{C}$ is a direct sum of representations of the following type: the underlying Hilbert space is $\mathcal{H} = L^2(\bar{\mathcal{A}}, \mathrm{d}\mu)$ for some measure μ on $\bar{\mathcal{A}}$ and (regarded as functions on $\bar{\mathcal{A}}$) elements of $\mathfrak{C}$ act by multiplication. Since there are many inequivalent measures on $\bar{\mathcal{A}}$, there is a multitude of representations of $\mathfrak{C}$. A key question is how many of them can be extended to representations of the full algebra $\mathfrak{a}$ (or $\mathfrak{W}$) without having to introduce any "background fields" which would compromise diffeomorphism covariance. Quite surprisingly, the requirement that the representation be cyclic with respect to a state which is invariant under the action of the (appropriately defined) group Diff M of piecewise-analytic diffeomorphisms on M singles out a unique irreducible representation. This result was established for $\mathfrak{a}$ by Lewandowski, Okołów, Sahlmann and Thiemann, and for $\mathfrak{W}$ by Fleischhack. It is the quantum geometry analog to the seminal results by Segal and others that characterized the Fock vacuum in Minkowskian field theories. However, while that result assumes not only Poincaré invariance but also specific (namely free) dynamics, it is striking that the present uniqueness theorems make no such restriction on dynamics. The requirement of diffeomorphism invariance is surprisingly strong and makes the "background-independent" quantum geometry framework surprisingly tight.

This representation had been constructed by Ashtekar, Baez, and Lewandowski some ten years before its uniqueness was established. The underlying Hilbert space is given by $\mathcal{H} = L^2(\bar{\mathcal{A}}, \mathrm{d}\mu_o)$ where μ_o is a diffeomorphism-invariant, faithful, regular Borel measure on $\bar{\mathcal{A}}$, constructed from the normalized Haar measure on SU(2). Typical quantum states can be visualized as follows. Fix: (1) a graph α on M (by a graph on M we mean a set of a finite number of embedded, oriented intervals called edges; if two edges intersect, they do so only at one or both ends, called vertices), and (2) a smooth function ψ on $[\mathrm{SU}(2)]^n$. Then, the function

$$\Psi_\gamma(\bar{A}) := \psi(\bar{A}(e_1), \ldots, \bar{A}(e_n)) \qquad [4]$$

on $\bar{\mathcal{A}}$ is an element of $\mathcal{H}$. Such states are said to be "cylindrical" with respect to the graph α and their space is denoted by Cyl_α. These are "typical states" in the sense that $\mathrm{Cyl} := \cup_\alpha \mathrm{Cyl}_\alpha$ is dense in $\mathcal{H}$. Finally, as ensured by the Gel'fand theory, the holonomy (or configuration) operators $\hat{h}_e$ act just by multiplication. The momentum operators $\hat{P}_{S,f}$ act as Lie derivatives: $\hat{P}_{S,f}\Psi = -\mathrm{i}\hbar\mathcal{L}_{X_{S,f}}\Psi$.

Remark Given any graph α in M, and a labeling of each of its edges by a nontrivial irreducible representation of SU(2) (i.e., by a nonzero half integer j), one can construct a finite-dimensional Hilbert space $\mathcal{H}_{\alpha,j}$, which can be thought of as the state space of a spin system "living on" the graph α. The full Hilbert space admits a simple decomposition: $\mathcal{H} = \oplus_{\alpha,j}\,\mathcal{H}_{\alpha,j}$. This is called the spin-network decomposition. The geometric operators discussed in the next section leave each $\mathcal{H}_{\alpha,j}$ invariant. Therefore, the availability of this decomposition greatly simplifies the task of analyzing their properties.

Geometric Operators

In the classical theory, $E := 8\pi G\gamma P$ has the interpretation of an orthonormal triad field (or a "moving frame") on M (with density weight 1). Here, γ is a dimensionless, strictly positive number, called the Barbero–Immirzi parameter, which arises as follows. Because of emphasis on connections, in the classical theory the first-order Palatini action is a more natural starting point than the second-order Einstein–Hilbert action. Now, there is a freedom to add a term to the Palatini action which vanishes when Bianchi identities are satisfied and therefore does not change the equations of motion. γ arises as the coefficient of this term. In some respects γ is analogous to the θ parameter of Yang–Mills theory. Indeed, while theories corresponding to any permissible values of γ are related by a canonical transformation classically, quantum mechanically this transformation is not unitarily implementable. Therefore, although there is a unique representation of the algebra $\mathfrak{a}$ (or $\mathfrak{W}$), there is a one-parameter family of inequivalent representations of the algebra of geometric operators generated by suitable functions of orthonormal triads E, each labeled by the value of γ. This is a genuine quantization ambiguity. As with the θ ambiguity in QCD, the actual value of γ in nature has to be determined experimentally. The current strategy in quantum geometry is to fix its value through a thought experiment involving black hole thermodynamics (see below).

The basic object in quantum Riemannian geometry is the triad flux operator $\hat{E}_{S,f} := 8\pi G\gamma\,\hat{P}_{S,f}$. It is self-adjoint and all its eigenvalues are discrete. To define other geometric operators such as the area operator $\hat{A}_S$ associated with a surface S or a volume operator $\hat{V}_R$ associated with a region R, one first expresses the corresponding phase-space functions in terms of the "elementary" functions E_{S_i,f_i} using suitable surfaces S_i and test functions f_i and then promotes E_{S_i,f_i} to operators. Even though the classical expressions are typically nonpolynomial functions of E_{S_i,f_i}, the final operators are all well defined, self-adjoint and with purely discrete eigenvalues. Therefore, in the sense of the word used in elementary quantum mechanics (e.g., of the hydrogen atom), one says that geometry is quantized. Because the theory has no background metric or indeed any other background field, all geometric operators transform covariantly under the action of the Diff M. This diffeomorphism covariance makes the final expressions of operators rather simple. In the case of the area operator, for example, the action of $\hat{A}_S$ on a state Ψ_α [4] depends entirely on the points of intersection of the surface S and the graph α and involves only right- and left-invariant vector fields on copies of SU(2) associated with edges of α which intersect S. In the case of the volume operator $\hat{V}_R$, the action depends on the vertices of α contained in R and, at each vertex, involves the right- and left-invariant vector fields on copies of SU(2) associated with edges that meet at each vertex.

To display the explicit expressions of these operators, let us first define on Cyl_α three basic operators $\hat{J}_j^{(v,e)}$, with $j \in \{1,2,3\}$, associated with the pair consisting of an edge e of α and a vertex v of e:

$$\hat{J}_j^{(v,e)}\Psi_\alpha(\bar{A}) = \begin{cases} \mathrm{i}\dfrac{\mathrm{d}}{\mathrm{d}t}|_{t=0}\psi_\alpha(\ldots, U_e(\bar{A})\exp(t\tau_j),\ldots) \\ \quad \text{if } e \text{ begins at } v \\ \mathrm{i}\dfrac{\mathrm{d}}{\mathrm{d}t}|_{t=0}\psi_\alpha(\ldots, \exp(-t\tau_j)U_e(\bar{A}),\ldots) \\ \quad \text{if } e \text{ ends at } v \end{cases} \qquad [5]$$

where τ_j denotes a basis in su(2) and "..." stands for the rest of the arguments of Ψ_α which remain unaffected. The quantum area operator A_s is assigned to a finite two-dimensional submanifold S in M. Given a cylindrical state we can always represent it in the form [4] using a graph α adapted to S, such that every edge e either intersects S at exactly one endpoint, or is contained in the closure $\bar{S}$, or does not intersect $\bar{S}$. For each vertex v in S of the graph α, the family of edges intersecting v can be divided into three classes: edges $\{e_1,\ldots,e_u\}$ lying on one side (say "above") S, edges $\{e_{u+1},\ldots,e_{u+d}\}$ lying

on the other side (say "below"), and edges contained in S. To each v we assign a generalized Laplace operator

$$\Delta_{S,v} = -\eta^{ij}\left(\sum_{I=1}^{u}\hat{J}_i^{(v,e_I)} - \sum_{I=u+1}^{u+d}\hat{J}_i^{(v,e_I)}\right) \times \left(\sum_{K=1}^{u}\hat{J}_j^{(v,e_K)} - \sum_{K=u+1}^{u+d}\hat{J}_j^{(v,e_K)}\right) \quad [6]$$

where η_{ij} stands for $-1/2$ the Killing form on su(2). Now, the action of the quantum area operator $\hat{A}_S$ on Ψ_α is defined as follows:

$$\hat{A}_S\Psi_\alpha = 4\pi\gamma\ell_{\mathrm{Pl}}^2\sum_{v\in S}\sqrt{-\Delta_{S,v}}\Psi_\alpha \quad [7]$$

The quantum area operator has played the most important role in applications. Its complete spectrum is known in a closed form. Consider arbitrary sets $j_I^{(u)}, j_I^{(d)}$, and $j_I^{(u+d)}$ of half-integers, subject to the condition

$$j_I^{(u+d)} \in \{|j_I^{(u)} - j_I^{(d)}|, |j_I^{(u)} - j_I^{(d)}| + 1, \ldots, j_I^{(u)} + j_I^{(d)}\} \quad [8]$$

where I runs over any finite number of integers. The general eigenvalues of the area operator are given by:

$$a_S = 4\pi\gamma\ell_{\mathrm{Pl}}^2\sum_I\left(2j_I^{(u)}(j_I^{(u)}+1) + 2j_I^{(d)}(j_I^{(d)}+1) - j_I^{(u+d)}(j_I^{(u+d)}+1)\right)^{1/2} \quad [9]$$

On the physically interesting sector of SU(2)-gauge-invariant subspace $\mathcal{H}_{\mathrm{inv}}$ of $\mathcal{H}$, the lowest eigenvalue of $\hat{A}_S$ – "the area gap" – depends on some global properties of S. Specifically, it "knows" whether the surface is open, or a 2-sphere, or, if M is a 3-torus, a (nontrivial) 2-torus in M. Finally, on $\mathcal{H}_{\mathrm{inv}}$, one is often interested only in the subspace of states Ψ_α, where α has no edges which lie within a given surface S. Then, the expression of eigenvalues simplifies considerably:

$$a_S = 8\pi\gamma\ell_{\mathrm{Pl}}^2\sum_I\sqrt{j_I(j_I+1)} \quad [10]$$

To display the action of the quantum volume operator $\hat{V}_R$, for each vertex v of a given graph α, let us first define an operator $\hat{q}_v$ on Cyl_α.

$$\hat{q}_v = (8\pi\gamma\ell_{\mathrm{Pl}}^2)^3\frac{1}{48} \times \sum_{e,e',e''}\epsilon(e,e',e'')c^{ijk}\hat{J}_i^{(v,e)}\hat{J}_j^{(v,e')}\hat{J}_k^{(v,e'')} \quad [11]$$

where e, e', and e'' run over the set of edges intersecting v, $\epsilon(e,e',e'')$ takes values ±1 or 0 depending on the orientation of the half-lines tangent to the edges at v, $[\tau_i,\tau_j] = c^k{}_{ij}\tau_k$ and the indices are raised by the tensor η_{ij}. The action of the quantum volume operator on a cylindrical state [4] is then given by

$$\hat{V}_R\Psi_\alpha = \kappa_o\sum_{v\in R}\sqrt{|\hat{q}_v|}.\Psi_\alpha \quad [12]$$

Here, κ_o is an overall, independent of a graph, constant resulting from an averaging.

The volume operator plays an unexpectedly important role in the definition of both the gravitational and matter contributions to the scalar constraint operator which dictates dynamics. Finally, a notable property of the volume operator is the following. Let $R(p,\epsilon)$ be a family of neighborhoods of a point $p \in M$. Then, as indicated above, $\hat{V}_{R(p,\epsilon)}\Psi_\alpha = 0$ if α has no vertex in the neighborhood. However, if α has a vertex at p

$$\lim_{\epsilon\to0}\hat{V}_{R(x,\epsilon)}\Psi_\alpha$$

exists but is not necessarily zero. This is a reflection of the "distributional" nature of quantum geometry.

Remark States $\Psi_\alpha \in \mathrm{Cyl}$ have support only on the graph α. In particular, they are simply annihilated by geometric operators such as $\hat{A}_S$ and $\hat{V}_R$ if the support of the surface S and the region R does not intersect the support of α. In this sense the fundamental excitations of geometry are one dimensional and geometry is polymer-like. States Ψ_α, where α is just a "small graph," are highly quantum mechanical – like states in QED representing just a few photons. Just as coherent states in QED require an infinite superposition of such highly quantum states, to obtain a semiclassical state approximating a given classical geometry, one has to superpose a very large number of such elementary states. More precisely, in the Gel'fand triplet $\mathrm{Cyl} \subset \mathcal{H} \subset \mathrm{Cyl}^\star$, semiclassical states belong to the dual $\mathrm{Cyl}^\star$ of Cyl.

Applications

Since quantum Riemannian geometry underlies loop quantum gravity and spin-foam models, all results obtained in these frameworks can be regarded as its applications. Among these, there are two which have led to resolutions of long-standing issues. The first concerns black hole entropy, and the second, quantum nature of the big bang.

Black Holes

Seminal advances in fundamentals of black hole physics in the mid-1970s suggested that the entropy of large black holes is given by $S_{\mathrm{BH}} = (a_{\mathrm{hor}}/4\ell_{\mathrm{Pl}}^2)$,

where $a_{\rm hor}$ is the horizon area. This immediately raised a challenge to potential quantum gravity theories: give a statistical mechanical derivation of this relation. For familiar thermodynamic systems, a statistical mechanical derivation begins with an identification the microscopic degrees of freedom. For a classical gas, these are carried by molecules; for the black body radiation, by photons; and for a ferromagnet, by Heisenberg spins. What about black holes? The microscopic building blocks cannot be gravitons because the discussion involves stationary black holes. Furthermore, the number of microscopic states is absolutely huge: some $\exp 10^{77}$ for a solar mass black hole, a number that completely dwarfs the number of states of systems one normally encounters in statistical mechanics. Where does this huge number come from? In loop quantum gravity, this is the number of states of the "quantum horizon geometry."

The idea behind the calculation can be heuristically explained using the "It from Bit" argument, put forward by Wheeler in the 1990s. Divide the black hole horizon into elementary cells, each with one Planck unit of area, $\ell^2_{\rm Pl}$, and assign to each cell two microstates. Then the total number of states $\mathcal{N}$ is given by $\mathcal{N}=2^n$, where $n=(a_{\rm hor}/\ell^2_{\rm Pl})$ is the number of elementary cells, whence entropy is given by $S=\ln\mathcal{N}\sim a_{\rm hor}$. Thus, apart from a numerical coefficient, the entropy (It) is accounted for by assigning two states (Bit) to each elementary cell. This qualitative picture is simple and attractive. However, the detailed derivation in quantum geometry has several new features.

First, Wheeler's argument would apply to any 2-surface, while in quantum geometry the surface must represent a horizon in equilibrium. This requirement is encoded in a certain boundary condition that the canonically conjugate pair $(A,\mathbb{P})$ must satisfy at the surface and plays a crucial role in the quantum theory. Second, the area of each elementary cell is not a fixed multiple of $\ell^2_{\rm Pl}$ but is given by [10], where I labels the elementary cells and j_I can be any half-integer (such that the sum is within a small neighborhood of the classical area of the black hole under consideration). Finally, the number of quantum states associated with an elementary cell labeled by j_I is not 2 but $(2j_I+1)$.

The detailed theory of the quantum horizon geometry and the standard statistical mechanical reasoning is then used to calculate the entropy and the temperature. For large black holes, the leading contribution to entropy is proportional to the horizon area, in agreement with quantum field theory in curved spacetimes. (The subleading term $-(1/2)\ln(a_{\rm hor}/\ell^2_{\rm Pl})$ is a quantum gravity correction to Hawking's semiclassical result. This correction, with the $-1/2$ factor, is robust in the sense that it also arises in other approaches.) However, as one would expect, the proportionality factor depends on the Barbero–Immirzi parameter γ and so far loop quantum gravity does not have an independent way to determine its value. The current strategy is to determine γ by requiring that, for the Schwarzschild black hole, the leading term agrees exactly with Hawking's semiclassical answer. This requirement implies that γ is the root of algebraic equation and its value is given by $\gamma\approx 0.2735$. Now, quantum geometry theory is completely fixed. One can calculate entropy of other black holes, with angular momentum and distortion. A nontrivial check on the strategy is that for all these cases, the coefficient in the leading-order term again agrees with Hawking's semiclassical result.

The detailed analysis involves a number of structures of interest to mathematical physics. First, the intrinsic horizon geometry is described by a U(1) Chern–Simons theory on a punctured 2-sphere (the horizon), the level k of the theory being given by $k=a_{\rm hor}/4\pi\gamma\ell^2_{\rm Pl}$. The punctures are simply the intersections of the excitations of the polymer geometry in the bulk with the horizon 2-surface. Second, because of the horizon boundary conditions, in the classical theory the gauge group SU(2) is reduced to U(1) at the horizon. At each puncture, it is further reduced to the discrete subgroup $\mathbb{Z}_k$ of U(1), sometimes referred to as a "quantum U(1) group." Third, the "surface phase space" associated with the horizon is represented by a noncommutative torus. Finally, the surface Chern–Simons theory is entirely unrelated to the bulk quantum geometry theory but the quantum horizon boundary condition requires that the spectrum of a certain operator in the Chern–Simons theory must be identical to that of another operator in the bulk theory. The surprising fact is that there is an exact agreement. Without this seamless matching, a coherent description of the quantum horizon geometry would not have been possible.

The main weakness of this approach to black hole entropy stems from the Barbero–Immirzi ambiguity. The argument would be much more compelling if the value of γ were determined by independent considerations, without reference to black hole entropy. (By contrast, for extremal black holes, string theory provides the correct coefficient without any adjustable parameter. The AdS/CFT duality hypothesis (as well as other semiquantitative) arguments have been used to encompass certain black holes which are away from extremality. But in these cases, it is not known if the numerical coefficient is

1/4 as in Hawking's analysis.) It's primary strengths are twofold. First, the calculation encompasses all realistic black holes – not just extremal or near-extremal – including the astrophysical ones, which may be highly distorted. Hairy black holes of mathematical physics and cosmological horizons are also encompassed. Second, in contrast to other approaches, one works directly with the physical, curved geometry around black holes rather than with a flat-space system which has the same number of states as the black hole of interest.

The Big Bang

Most of the work in physical cosmology is carried out using spatially homogeneous and isotropic models and perturbations thereon. Therefore, to explore the quantum nature of the big bang, it is natural to begin by assuming these symmetries. Then the spacetime metric is determined simply by the scale factor $a(t)$ and matter fields $\phi(t)$ which depend only on time. Thus, because of symmetries, one is left with only a finite number of degrees of freedom. Therefore, field-theoretic difficulties are bypassed and passage to quantum theory is simplified. This strategy was introduced already in the late 1960s and early 1970s by DeWitt and Misner. Quantum Einstein's equations now reduce to a single differential equation of the type

$$\frac{\partial^2}{\partial a^2}(f(a)\Psi(a,\phi)) = \text{const.}\,\hat{H}_\phi\,\Psi(a,\phi) \qquad [13]$$

on the wave function $\Psi(a,\phi)$, where $\hat{H}_\phi$ is the matter Hamiltonian and $f(a)$ reflects the freedom in factor ordering. Since the scale factor a vanishes at the big bang, one has to analyze the equation and its solutions near $a=0$. Unfortunately, because of the standard form of the matter Hamiltonian, coefficients in the equation diverge at $a=0$ and the evolution cannot be continued across the singularity unless one introduces unphysical matter or a new principle. A well-known example of new input is the Hartle–Hawking boundary condition which posits that the universe starts out without any boundary and a metric with positive-definite signature and later makes a transition to a Lorentzian metric.

Bojowald and others have shown that the situation is quite different in loop quantum cosmology because quantum geometry effects make a qualitative difference near the big bang. As in older quantum cosmologies, one carries out a symmetry reduction at the classical level. The final result differs from older theories only in minor ways. In the homogeneous, isotropic case, the freedom in the choice of the connection is encoded in a single function $c(t)$ and, in that of the momentum/triad, in another function $p(t)$. The scale factor is given by $a^2=|p|$. (The variable p itself can assume both signs; positive if the triad is left handed and negative if it is right handed. p vanishes at degenerate triads which are permissible in this approach.) The system again has only a finite number of degrees of freedom. However, quantum theory turns out to be inequivalent to that used in older quantum cosmologies.

This surprising result comes about as follows. Recall that in quantum geometry, one has well-defined holonomy operators $\hat{h}$ but there is no operator corresponding to the connection itself. In quantum mechanics, the analog would be for operators $\hat{U}(\lambda)$ corresponding to the classical functions $\exp \mathrm{i}\lambda x$ to exist but not be weakly continuous in λ; the operator $\hat{x}$ would then not exist. Once the requirement of weak continuity is dropped, von Neumann's uniqueness theorem no longer holds and the Weyl algebra can have inequivalent irreducible representations. The one used in loop quantum cosmology is the direct analog of full quantum geometry. While the space $\mathcal{A}$ of smooth connections reduces just to the real line $\mathbb{R}$, the space $\bar{\mathcal{A}}$ of generalized connections reduces to the Bohr compactification $\bar{\mathbb{R}}_{\text{Bohr}}$ of the real line. (This space was introduced by the mathematician Harold Bohr (Nils' brother) in his theory of almost-periodic functions. It arises in the present application because holonomies turn out to be almost periodic functions of c.) The Hilbert space of states is thus $\mathcal{H}=L^2(\bar{\mathbb{R}}_{\text{Bohr}}, \mathrm{d}\mu_o)$ where μ_o is the Haar measure on (the abelian group) $\bar{\mathbb{R}}_{\text{Bohr}}$. As in full quantum geometry, the holonomies act by multiplication and the triad/momentum operator $\hat{p}$ via Lie derivatives.

To facilitate comparison with older quantum cosmologies, it is convenient to use a representation in which $\hat{p}$ is diagonal. Then, quantum states are functions $\Psi(p,\phi)$. But the Wheeler–DeWitt equation is now replaced by a difference equation:

$$C^+(p)\,\Psi(p+4p_o,\phi) + C^o(p)\,\Psi(p,\phi) + C^-(p)\,\Psi(p-4p_o)(\phi) = \text{const.}\,\hat{H}_\phi\Psi(p,\phi) \qquad [14]$$

where p_o is determined by the lowest eigenvalue of the area operator ("area gap") and the coefficients $C^{\pm}(p)$ and $C^o(p)$ are functions of p. In a backward "evolution," given Ψ at $p+4$ and p, such a "recursion relation" determines Ψ at $p-4$, provided C^- does not vanish at $p-4$. The coefficients are well behaved and nowhere vanishing, whence the evolution does not stop at any finite p, either in the past or in the future. Thus, near $p=0$ this equation is drastically different from the Wheeler–DeWitt equation [13]. However, for large p – that is, when the universe is large – it is well

approximated by [13] and smooth solutions of [13] are approximate solutions of the fundamental discrete equation [14] in a precise sense.

To complete quantization, one has to introduce a suitable Hilbert space structure on the space of solutions to [14], identify physically interesting operators and analyze their properties. For simple matter fields, this program has been completed. With this machinery at hand, one begins with semiclassical states which are peaked at configurations approximating the classical universe at late times (e.g., now) and evolves backwards. Numerical simulations show that the state remains peaked at the classical solution till very early times when the matter density becomes of the order of Planck density. This provides, in particular, a justification, from first principles, for the assumption that spacetime can be taken to be classical even at the onset of the inflationary era, just a few Planck times after the (classical) big bang. While one would expect a result along these lines to hold on physical grounds, technically it is nontrivial to obtain semiclassicality over such huge domains. However, in the Planck regime near the big bang, there are major deviations from the classical behavior. Effectively, gravity becomes repulsive, the collapse is halted and then the universe re-expands. Thus, rather than modifying spacetime structure just in a tiny region near the singularity, quantum geometry effects open a bridge to another large classical universe. These are dramatic modifications of the classical theory.

For over three decades, hopes have been expressed that quantum gravity would provide new insights into the true nature of the big bang. Thanks to quantum geometry effects, these hopes have been realized and many of the long-standing questions have been answered. While the final picture has some similarities with other approaches, (e.g., "cyclic universes," or pre-big-bang cosmology), only in loop quantum cosmology is there a fully deterministic evolution across what was the classical big-bang. However, so far, detailed results have been obtained only in simple models. The major open issue is the inclusion of perturbations and subsequent comparison with observations.

See also: Algebraic Approach to Quantum Field Theory; Black Hole Mechanics; Canonical General Relativity; Knot Invariants and Quantum Gravity; Loop Quantum Gravity; Quantum Cosmology; Quantum Dynamics in Loop Quantum Gravity; Quantum Fields Theory in Curved Spacetime; Spacetime Topology, Causal Structure and Singularities; Spin Foams; Wheeler–De Witt Theory.

Further Reading

Ashtekar A (1987) New Hamiltonian formulation of general relativity. *Physical Review D* 36: 1587–1602.

Ashtekar A and Krishnan B (2004) Isolated and dynamical horizons and their applications. *Living Reviews in Relativity* 10: 1–78 (gr-qc/0407042).

Ashtekar A and Lewandowski L (2004) Background independent quantum gravity: a status report. *Classical Quantum Gravity* 21: R53–R152.

Bojowald M and Morales-Tecotl HA (2004) Cosmological applications of loop quantum gravity. *Lecture Notes in Physics* 646: 421–462 (also available at gr-qc/0306008).

Gambini R and Pullin J (1996) *Loops, Knots, Gauge Theories and Quantum Gravity*. Cambridge: Cambridge University Press.

Perez A (2003) Spin foam models for quantum gravity. *Classical Quantum Gravity* 20: R43–R104.

Rovelli C (2004) *Quantum Gravity*. Cambridge: Cambridge University Press.

Thiemann T (2005) *Introduction to Modern Canonical Quantum General Relativity*, Cambridge: Cambridge University Press. (draft available as gr-qc/0110034).

Quantum Group Differentials, Bundles and Gauge Theory

T Brzeziński, University of Wales Swansea, Swansea, UK

Introduction

Mathematics of classical gauge theories is contained in the theory of principal and associated vector bundles. Principal bundles describe pure gauge fields and their transformations, while the associated bundles contain matter fields. A structure group of a bundle has a meaning of a gauge group, while the base manifold is a spacetime for the theory. In this article, we review the theory of bundles in which a structure group is a quantum group and base space or spacetime might be noncommutative. To fully deal with geometric aspects, we first review differential geometry of quantum groups. Then we describe the theory of quantum principal bundles, connections on such bundles, gauge transformations, associated vector bundles and their sections. We indicate that, for a certain class of quantum principal bundles, sections of an associated bundle become vector bundles of noncommutative geometry *à la* Connes, that is,

finite projective modules. The theory is illustrated by two explicit examples that can be viewed as deformations of the classical magnetic monopole and the instanton.

Differential Structures on Algebras

Algebraic Conventions

Throughout this article, A (P etc.,) will be an associative unital complex algebra. To gain some geometric intuition the reader can think of A as an algebra of continuous complex functions on a compact (Hausdorff) space X, $C(X)$, with product given by pointwise multiplication $fg(x) = f(x)g(x)$, and with the unit provided by a constant function $x \mapsto 1$. The algebra $C(X)$ is commutative, but, in what follows, we do not assume that A is a commutative algebra. By an A-bimodule we mean a vector space with mutually commuting left and right actions of A. All modules are unital (i.e., the unit element of A acts trivially). On elements, the multiplication in an algebra or an action of A on a module is denoted by juxtaposition.

Differential Calculus on an Algebra

A first-order differential calculus on A is a pair $(\Omega^1(A), \mathrm{d})$, where $\Omega^1(A)$ is an A-bimodule and $\mathrm{d}: A \to \Omega^1(A)$ is a linear map such that:

1. for all $a, b \in A$, $\mathrm{d}(ab) = (\mathrm{d}a)b + a\mathrm{d}b$ (the Leibniz rule); and
2. every $\omega \in \Omega^1(A)$ can be written as $\omega = \sum_i a_i \mathrm{d}b_i$ for some $a_i, b_i \in A$.

Elements of $\Omega^1(A)$ are called differential 1-forms and the map d is called an exterior derivative. As a motivating example, take $A = C(X)$ and $\Omega^1(A)$ the space of 1-forms on X (sections of the cotangent bundle T^*X), and d the usual exterior differential. Higher-differential forms corresponding to $(\Omega^1(A), \mathrm{d})$ are defined as elements of a differential graded algebra $\Omega(A)$. This is an algebra which can be decomposed into the direct sum of A-bimodules $\Omega^n(A)$, that is, $\Omega(A) = A \oplus \Omega^1(A) \oplus \Omega^2(A) \oplus \cdots$. In addition to $\mathrm{d}: A \to \Omega^1(A)$, there are maps $\mathrm{d}_n: \Omega^n(A) \to \Omega^{n+1}(A)$ such that, for all $\omega_n \in \Omega^n(A)$, $\omega_k \in \Omega^k(A)$,

1. $\mathrm{d}_1 \circ \mathrm{d} = 0$ and $\mathrm{d}_{n+1} \circ \mathrm{d}_n = 0, n = 1, 2, \ldots$;
2. $\omega_n \omega_k \in \Omega^{n+k}(A)$; and
3. $\mathrm{d}_{n+k}(\omega_n \omega_k) = (\mathrm{d}_n \omega_n)\omega_k + (-1)^n \omega_n (\mathrm{d}_k \omega_k)$.

Elements of $\Omega^n(A)$ are known as "differential n-forms." $\Omega^n(A)$ contains all linear combinations of expressions $a_0\, \mathrm{d}a_1\, \mathrm{d}a_2 \cdots \mathrm{d}a_n$ with $a_0, \ldots, a_n \in A$. One says that $\Omega(A)$ satisfies the "density condition" if any element of $\Omega^n(A)$ is of the above form, for any n. To simplify notation, one writes d for d_n.

As an example of $\Omega(A)$, take $A = C(X)$ and then the exterior algebra $\Omega(X)$ for $\Omega(A)$. The exterior algebra satisfies density condition as any n-form can be written as $f(x) \wedge \mathrm{d}g(x) \wedge \mathrm{d}h(x) \wedge \cdots$. The wedge product is anticommutative, but for a noncommutative algebra A, the anticommutativity of the product in $\Omega(A)$ cannot be generally required.

The Universal Differential Calculus

Any algebra A comes equipped with a universal differential calculus denoted by $(\Omega^1 A, \mathrm{d})$. $\Omega^1 A$ is defined as the kernel of the multiplication map, that is, $\Omega^1 A := \{\sum_i a_i \otimes b_i \in A \otimes A \mid \sum_i a_i b_i = 0\} \subseteq A \otimes A$. The derivative is defined by $\mathrm{d}(a) = 1 \otimes a - a \otimes 1$. The n-forms are defined as $\Omega^n A = \Omega^1 A \otimes_A \Omega^1 A \otimes_A \cdots \otimes_A \Omega^1 A$ (n-copies of $\Omega^1 A$). $\Omega^n A$ can be identified with a subspace of $A \otimes A \otimes \cdots \otimes A$ ($n+1$-copies of A) consisting of all such elements that vanish upon multiplication of any two consecutive factors. With this identification, higher derivatives read

$$\mathrm{d}\Big(\sum_i a_0^i \otimes a_1^i \otimes \cdots \otimes a_n^i\Big) = \sum_{k=0}^{n+1} \sum_i (-1)^k a_0^i \otimes a_1^i \otimes \cdots \otimes a_{k-1}^i \otimes 1 \otimes a_k^i \cdots \otimes a_n^i$$

The universal differential calculus satisfies the density condition.

This calculus captures very little (if any) of the geometry of the underlying algebra A, but it has the universality property, that is, any differential calculus on A can be obtained as a quotient of ΩA. In other words, any differential calculus $\Omega(A)$ is fully determined by a system of A-sub-bimodules $N_n \in A^{\otimes n+1}$ (or homogeneous ideals in the algebra ΩA), so that $\Omega^n(A) = \Omega^n A / N_n$. The differentials d in $\Omega(A)$ are derived from universal differentials via the canonical projections $\pi_n: \Omega^n A \to \Omega^n(A)$.

Typical examples of algebras in quantum geometry are given by generators and relations, that is, $A = \mathbb{C}\langle x_1, \ldots, x_n \rangle / \langle R_i(x_1, \ldots, x_n) \rangle$, where $\mathbb{C}\langle x_1, \ldots, x_n \rangle$ is a free algebra on generators x_k and $R_i(x_1, \ldots, x_n)$ are polynomials, so that $R_i(x_1, \ldots, x_n) = 0$ in A. Correspondingly, the modules $\Omega^n(A)$ are given by generators and relations. If $\Omega(A)$ satisfies the density condition, that the whole of $\Omega(A)$ must be generated by some 1-forms. The sub-bimodules N_n contain relations satisfied by these generators.

$*$-Calculi

If A is a $*$-algebra, then a calculus is called a "$*$-calculus" provided $\Omega(A)$ is a graded $*$-algebra, and $\mathrm{d}(\rho^*) = (\mathrm{d}\rho)^*$, for all $\rho \in \Omega(A)$.

Differential Structures on Quantum Groups (Hopf Algebras)

Hopf Algebra Preliminaries

From now on, A is a Hopf algebra (quantum group), with a coproduct $\Delta : A \to A \otimes A$, counit $\varepsilon : A \to \mathbb{C}$ and antipode S. We use Sweedler's notation $\Delta(a) = \sum a_{(1)} \otimes a_{(2)}$. We also write $A^+ = \ker \varepsilon$ (the augmentation ideal).

For any algebra P, the convolution product of linear maps $f, g : A \to P$ is a linear map $f * g : A \to P$, defined by $f * g(a) = \sum f(a_{(1)}) * g(a_{(2)})$. A map $f : A \to P$ is said to be convolution invertible, provided there exists $f^{-1} : A \to P$ such that $f * f^{-1} = f * f^{-1} = 1\varepsilon$.

An A-coaction on a comodule V, $\varrho : V \to V \otimes A$, is denoted by $\varrho(v) = \sum v_{(0)} \otimes v_{(1)}$. The right adjoint coaction in A is a map

$$\mathrm{Ad} : A \to A \otimes A,$$
$$\mathrm{Ad}(a) = \sum a_{(2)} \otimes (Sa_{(1)})a_{(3)}$$

A subspace B of A is said to be "Ad-invariant" provided $\mathrm{Ad}(B) \subseteq B \otimes A$. For example, A^+ is such a space.

Covariant Differential Calculi

For Hopf algebras one can study calculi that are covariant with respect to Δ. For $A = \mathbb{C}[G]$ (an algebra of functions on a Lie group), this corresponds to the covariance of a differential structure on G with respect to regular representations.

A first-order differential calculus $\Omega^1(A)$ on a quantum group A is said to be left-covariant, if there exists a linear map $\Delta_{\mathrm{L}} : \Omega^1(A) \to A \otimes \Omega^1(A)$ (called a left coaction) such that, for all $a, b \in A$,

$$\Delta_{\mathrm{L}}(a\mathrm{d}b) = \sum a_{(1)}b_{(1)} \otimes a_{(2)}\mathrm{d}b_{(2)}$$

$\Omega^1(A)$ is called a right-covariant differential calculus if there exists a linear map $\Delta_{\mathrm{R}} : \Omega^1(A) \to \Omega^1(A) \otimes A$ (called a right coaction) such that, for all $a, b \in A$,

$$\Delta_{\mathrm{R}}(a\mathrm{d}b) = \sum a_{(1)}\mathrm{d}b_{(1)} \otimes a_{(2)}b_{(2)}$$

If $\Omega^1(A)$ is both left- and right-covariant, it is called a "bicovariant differential calculus." A bicovariant $\Omega^1(A)$ has a structure of a Hopf A-bimodule, that is, it is an A-bimodule and an A-bicomodule such that the coactions are compatible with actions.

The universal calculus on A is bicovariant with coactions

$$\Delta_{\mathrm{R}}^U\Big(\sum_i a^i \otimes b^i\Big) = \sum_i a^i_{(1)} \otimes b^i_{(1)} \otimes a^i_{(2)}b^i_{(2)},$$
$$\Delta_{\mathrm{L}}^U\Big(\sum_i a^i \otimes b^i\Big) = \sum_i a^i_{(1)}b^i_{(1)} \otimes a^i_{(2)} \otimes b^i_{(2)}$$

Since $\Omega^1(A) = \Omega^1 A/N$ for an A-sub-bimodule $N \in \Omega^1 A$, the calculus $\Omega^1(A)$ is left (resp. right) covariant if and only if $\Delta_{\mathrm{L}}^U(N) \subseteq A \otimes N$ (resp. $\Delta_{\mathrm{R}}^U(N) \subseteq N \otimes A$).

The Woronowicz Theorems

A form ω in a left-covariant differential calculus $\Omega^1(A)$ is said to be left-invariant provided $\Delta_{\mathrm{L}}(\omega) = 1 \otimes \omega$. $\Omega^1(A)$ is a free A-module with basis given by left-invariant forms, that is, one can choose a set of left-invariant forms ω^i such that any 1-form ρ can be uniquely written as a finite sum $\rho = \sum_i a_i\omega^i, a_i \in A$.

The first Woronowicz theorem states that there is a one-to-one correspondence between left-covariant calculi on A and right A-ideals $Q \subseteq A^+$. The correspondence is provided by the map

$$\kappa : A \otimes Q \to N, \qquad a \otimes q \mapsto \sum aSq_{(1)} \otimes q_{(2)}$$

where N is such that $\Omega^1(A) = (\Omega^1 A)/N$. The inverse of κ reads $\kappa^{-1}(\sum_i a^i \otimes b^i) = \sum_i a^i b^i_{(1)} \otimes b^i_{(2)}$. The map κ induces the map $\bar{\kappa} : A^+/Q \to \Omega^1(A)$, via $\omega([a]) = [\kappa(1 \otimes a)]$ where $[-]$ denotes cosets in A^+/Q and in $\Omega^1(A) = (\Omega^1 A)/N$. This establishes a one-to-one correspondence between the space $\Lambda^1 = A^+/Q$ and the space of left-invariant 1-forms in $\Omega^1(A)$. The dual space to Λ^1, that is, the space of linear functionals $\Lambda^1 \to \mathbb{C}$, is often termed a "quantum Lie algebra" or a "quantum tangent space" corresponding to a left-covariant calculus $\Omega^1(A)$. The dimension of Λ^1 is known as a dimension of $\Omega^1(A)$.

The definitions and analysis of right-covariant differential calculi are done in a symmetric manner. For a bicovariant calculus, a form ω that is both left- and right-invariant, is termed a "bi-invariant" form.

The second Woronowicz theorem states a one-to-one correspondence between bicovariant differential calculi and Ad-invariant A-ideals $Q \subseteq A^+$ (cf. the subsection "Hopf algebra preliminaries"). The correspondence is provided by the map κ above. For the universal calculus, Q is trivial, and hence $\Lambda^1 = A^+ = \ker(\varepsilon)$.

Higher-order Bicovariant Calculi

Given a first-order bicovariant calculus $\Omega^1(A)$, one constructs a braiding operator, known as the

"Woronowicz braiding" $\tau:\Omega^1(A)\otimes_A\Omega^1(A)\to\Omega^1(A)\otimes_A\Omega^1(A)$ by setting $\tau(a\omega\otimes_A\eta)=a\eta\otimes_A\omega$ for all $a\in A$, and any left-invariant ω and right-invariant η, and then extending it A-linearly to the whole of $\Omega^1(A)\otimes_A\Omega^1(A)$. This operator satisfies the braid relation $(\mathrm{id}\otimes_A\tau)\circ(\tau\otimes_A\mathrm{id})\circ(\mathrm{id}\otimes_A\tau)=(\tau\otimes_A\mathrm{id})\circ(\mathrm{id}\otimes_A\tau)\circ(\tau\otimes_A\mathrm{id})$, and is invertible provided the antipode S is invertible. The Woronowicz braiding is used to define symmetric forms as those invariant under τ. One then defines exterior 2-forms as elements of $\Omega^1(A)\otimes_A\Omega^1(A)/\ker(\mathrm{id}-\tau)$, and introduces the wedge product. The wedge product is not in general anticommutative, but one does have $\omega\wedge\eta=-\eta\wedge\omega$ for bi-invariant ω,η. This construction is extended to higher forms and leads to the definition of the exterior algebra $\Omega(A)$. To define exterior n-forms, one maps any permutation on n-elements to the corresponding element of the braid group generated by τ and then takes the quotient of the nth tensor power of $\Omega^1(A)$ by all elements corresponding to even permutations. The differential $\mathrm{d}:A\to\Omega^1(A)$ is extended to an exterior differential in the whole of $\Omega(A)$ in the following way. First, $\Omega^1(A)$ is extended by a one-dimensional A-bimodule generated by a form θ that is required to be bi-invariant. The resulting extended bimodule (which, in general, is not a first-order differential calculus, as θ is not necessarily of the form $\sum_i a_i\,\mathrm{d}b_i$, for some $a_i,b_i\in A$) is then determined from the relation $\mathrm{d}a=\theta a-a\theta$ for all $a\in A$. Higher exterior derivative is then defined by $\mathrm{d}\rho=\theta\wedge\rho-(-1)^n\rho\wedge\theta$, for any $\rho\in\Omega^n(A)$.

The algebra $\Omega(A)$ is a $\mathbb{Z}_2$-graded differential Hopf algebra, that is, it has a coproduct such that

$$\Delta(\omega\wedge\eta)=\sum(-1)^{|\omega_{(2)}||\eta_{(1)}|}\omega_{(1)}\wedge\eta_{(1)}\otimes\omega_{(2)}\wedge\eta_{(2)}$$

where $|\omega_{(2)}|$ etc., denotes the degree of a homogeneous component in the decomposition of $\Delta(\omega)$. Furthermore,

$$\Delta(\mathrm{d}\omega)=\sum\left(\mathrm{d}\omega_{(1)}\otimes\omega_{(2)}+(-1)^{|\omega_{(1)}|}\omega_{(1)}\otimes\mathrm{d}\omega_{(2)}\right)$$

On the 1-forms this coproduct is simply the sum $\Delta_{\mathrm{L}}+\Delta_{\mathrm{R}}$.

Classification

There is no unique covariant differential calculus on A, so classification of covariant differential calculi is an important problem. For example, it is known that the quantum group $\mathrm{SU}_q(2)$ admits a left-covariant three-dimensional calculus, but there is no three-dimensional bicovariant calculus. On the other hand, there are two four-dimensional bicovariant calculi on $\mathrm{SU}_q(2)$. Differential calculi are classified for standard quantum groups such as $\mathrm{SL}_q(N)$ or $\mathrm{Sp}_q(N)$.

General classification results are based on the equivalence between the category of Hopf bimodules of a finite-dimensional Hopf algebra A and that of Yetter–Drinfeld or crossed modules of A. These are the modules of the Drinfeld double of A. As a result, in the case of a finite-dimensional factorizable coquasitriangular Hopf algebra A with a dual Hopf algebra H, the bicovariant $\Omega^1(A)$ are in one-to-one correspondence with two-sided ideals in H^+. If, in addition, A is semisimple, then (coirreducible) calculi are in one-to-one correspondence with nontrivial irreducible representations of H. This can be extended to infinite-dimensional algebras, provided one works over a field of formal power series in the deformation parameter.

Quantum Group Principal Bundles

Quantum Principal Bundles

In classical geometry, a (topological) principal bundle is a locally compact Hausdorff space with a (continuous) free and proper action of a locally compact group (e.g., a Lie group). In terms of algebras of functions this gives rise to the following structure. A is a Hopf algebra (the model is functions on a group G), P is a right A-comodule algebra with a coaction $\Delta_P:P\to P\otimes A$ (the model is functions on a total space X). Let $B=\{b\in P\,|\,\Delta_P(b)=b\otimes 1\}$ be the coinvariant subalgebra (the model is functions on a base manifold $M=X/G$). Fix a bicovariant calculus $\Omega^1(A)$, with the corresponding Q and $\Lambda^1=A^+/Q$ as in the subsection "The Woronowicz theorems." Take a differential calculus $\Omega^1(P)=\Omega^1P/N_P$ such that:

1. $\Delta_{\Omega^1P}(N_P)\subseteq N_P\otimes A$, where for all $\sum_i p^i\otimes q^i\in\Omega^1P$,

$$\Delta_{\Omega^1P}\left(\sum_i p^i\otimes q^i\right)=\sum_i p^i_{(0)}\otimes q^i_{(0)}\otimes p^i_{(1)}q^i_{(1)}\in\Omega^1P\otimes A$$

2. $\tilde{\chi}(N_P)\subseteq N_P\otimes Q$, where

$$\tilde{\chi}:\Omega^1P\to P\otimes A^+,\qquad\sum_i p^i\otimes q^i\mapsto\sum_i p^i\Delta_P(q^i)=\sum_i p^iq^i_{(0)}\otimes q^i_{(1)}$$

3. $N_B=N_P\cap\Omega^1B$ gives rise to a differential structure $\Omega^1(B)=\Omega^1B/N_B$ on B. Condition (1) ensures that Δ_{Ω^1P} descends to a coaction $\Delta_{\Omega^1(P)}:\Omega^1(P)\to\Omega^1(P)\otimes A$, while (2) allows for defining a map

$$\mathrm{ver}:\Omega^1(P)\to P\otimes\Lambda^1,\qquad\mathrm{ver}([\omega])=[\tilde{\chi}(\omega)]$$

Since B is a subalgebra of P, the P-bimodule

$$P\Omega^1(B)P := \left\{\sum_i p^i(\mathrm{d}b^i)q^i | p^i, q^i \in P, b^i \in B\right\}$$

is a sub-bimodule of $\Omega^1 P$, known as horizontal forms. P is called a "quantum principal bundle" over B with quantum structure group A and calculi $\Omega^1(A)$ and $\Omega^1(P)$ provided the following sequence;

$$0 \longrightarrow P\Omega^1(B)P \longrightarrow \Omega^1(P) \xrightarrow{\text{ver}} P \otimes \Lambda^1 \longrightarrow 0$$

is exact. This definition reflects the geometric content of principal bundles, but is not restricted to any specific differential calculus. The surjectivity of ver corresponds to the freeness of the (co)action, while the condition $\ker(\text{ver}) = P\Omega^1(B)P$ corresponds to identification of vertical vector fields as those that are annihilated by horizontal forms.

The Universal Calculus Case

In the universal calculus case, both N_P and Q in the previous subsection are trivial, and $\text{ver} = \tilde{\chi}$. Universal horizontal forms $P(\Omega^1 B)P$ coincide with the kernel of the canonical projection $P \otimes_B P \to P \otimes P$. The exactness of the sequence in the last subsection is equivalent to the requirement that the map

$$\begin{aligned} &\text{can} : P \otimes_B P \to P \otimes A \\ &p \otimes_B q \mapsto p\Delta_P(q) = \sum pq_{(0)} \otimes q_{(1)} \end{aligned}$$

be bijective. In algebra, such an inclusion of algebras $B \subseteq P$ is known as a Hopf–Galois extension. Thus, a geometric notion of a quantum principal bundle with the universal calculus is the same as the algebraic notion of a Hopf–Galois extension.

If (2) in the previous subsection is replaced by stronger conditions $\tilde{\chi}(N_P) = N_P \otimes Q$ and $(N_P \cap \ker\tilde{\chi}) \subseteq P(\Omega^1 B)P$, then exactness of the sequence in the previous subsection is equivalent to the bijectivity of "can." Thus, although defined in a purely algebraic way, the notion of a Hopf–Galois extension carries deep geometric meaning. It therefore makes sense to consider primarily Hopf–Galois extensions and then specify differential structure in such a way that this stronger version of (2) is satisfied. Henceforth, unless specified otherwise, a quantum principal bundle is taken with the universal differential calculus.

Quantum Homogeneous Bundles

Suppose that P is a Hopf algebra, and that there is a Hopf algebra surjection $\pi : P \to A$. This induces a coaction of P on A via $\Delta_P = (\mathrm{id} \otimes \pi) \circ \Delta$, where now Δ is a coproduct in P. P is a quantum principal A-bundle over the coinvariants B, provided $\ker\pi \subseteq B^+P$, where $B^+ = B \cap P^+$. B is a left quantum homogeneous space in the sense that $\Delta(B) \subseteq P \otimes B$, and P is known as a quantum homogeneous bundle. An example of this is the standard quantum 2-sphere – a quantum homogeneous space of $\mathrm{SU}_q(2)$ (see the subsection "The Dirac q-monopole"). This construction reflects the classical construction of a principal bundle over a homogeneous space, since every homogeneous space of a group G can be identified with a quotient G/H, where $H \subseteq G$ is a subgroup. Not every quantum homogeneous space can be obtained in this way (e.g., nonstandard quantum 2-spheres), as quantum groups P do not have sufficiently many quantum subgroups A (in a sense of Hopf algebra projections $\pi : P \to A$). To study gauge theory on general quantum homogeneous spaces, more general notion of a bundle needs to be developed (see the subsection "Generalizations of quantum principal bundles").

A general differential calculus on a quantum homogeneous bundle is specified by choosing a left-covariant calculus on P with an ideal $Q_P \in P^+$ such that $(\mathrm{id} \otimes \pi) \circ \mathrm{Ad}(Q_P) \subseteq Q_P \otimes A$. A bicovariant calculus on A is then given by $Q_A = \pi(Q_P)$.

Quantum Trivial Bundles

A quantum principal bundle (with the universal differential calculus) is said to be "trivial" or "cleft" provided there exists a linear map $\Phi : A \to P$ such that

1. $\Phi(1) = 1$ (unitality);
2. $\Delta_P \circ \Phi = (\Phi \otimes \mathrm{id}) \circ \Delta$ (*colinearity* or *covariance*); and
3. Φ is convolution invertible (cf. the subsection "Hopf algebra preliminaries").

Φ is called a trivialization. In this case, P is isomorphic to $B \otimes A$ as a left B-module and right A-comodule via the map $B \otimes A \to P, b \otimes a \mapsto b\Phi(a)$. In particular, an A-covariant (i.e., colinear) algebra map $j : A \to P$ is a trivialization (the convolution inverse of j is $j \circ S$).

Based on trivial bundles, locally trivial bundles can be constructed by choosing a compatible covering of B (in terms of ideals).

At this point, the reader should be warned that the notion of a trivial quantum principal bundle includes bundles which are not trivial classically (i.e., do not correspond to functions on the Cartesian product of spaces). As an example, consider the Möbius strip viewed as a $\mathbb{Z}_2$-principal bundle over the circle S^1. Obviously, this is not a trivial bundle (the Möbius strip is not isomorphic to

$S^1 \times \mathbb{Z}_2$). It can be shown, however, that the quantum principal bundle corresponding to the Möbius strip has a trivialization Φ in the above sense.

Generalizations of Quantum Principal Bundles

In the case of majority of quantum homogeneous spaces, the map π in the subsection "Quantum homogeneous bundles" is a coalgebra and right P-module map, but not an algebra map. Thus, the induced coaction is not an algebra map either. To cover examples like these, one needs to introduce a generalization of quantum principal bundles. Consider an algebra P that is also a right comodule of a coalgebra C with coaction Δ_P. Define

$$B := \Big\{ b \in P | \forall p \in P, \Delta_P(bp) = b\Delta_P(p) = \sum bp_{(0)} \otimes p_{(1)} \Big\}$$

B is a subalgebra of P. P is a principal coalgebra-bundle over B or $B \subseteq P$ is a coalgebra-Galois extension provided the map

$$\begin{aligned} \mathrm{can} : P \otimes_B P &\to P \otimes C \\ p \otimes_B q &\mapsto p\Delta_P(q) = \sum pq_{(0)} \otimes q_{(1)} \end{aligned}$$

is bijective. This purely algebraic requirement induces a rich symmetry structure on P, given in terms of entwining, which allows one for developing various differential geometric notions such as those discussed in the next section. The lack of space does not permit us to describe this theory here.

Connections, Gauge Transformations, Matter Fields

Connections and Connection Forms

A "connection" in a quantum principal bundle with calculi $\Omega^1(P), \Omega^1(A)$ is a left P-linear map $\Pi : \Omega^1(P) \to \Omega^1(P)$ such that:

1. $\Pi \circ \Pi = \Pi$ (Π is a projection);
2. $\ker \Pi = P\Omega^1(B)P$; and
3. $\Delta_{\Omega^1(P)} \circ \Pi = (\Pi \otimes \mathrm{id}) \circ \Delta_{\Omega^1(P)}$ (colinearity or covariance).

The exact sequence in the subsection "Quantum principal bundles" implies that Π is a left P-linear projection if and only if there exists a left P-linear map $\sigma : P \otimes \Lambda^1 \to \Omega^1(P)$ such that $\mathrm{ver} \circ \sigma = \mathrm{id}$. Since σ is left P-linear, it is fully specified by its action on Λ^1. This leads to the equivalent definition of a connection as a connection form or a gauge field, that is, a map $\omega : \Lambda^1 \to \Omega^1(P)$ such that:

1. for all $\lambda \in \Lambda^1$, $\mathrm{ver}(\omega(\lambda)) = 1 \otimes \lambda$; and
2. $\Delta_{\Omega^1(P)} \circ \omega = (\omega \otimes \mathrm{id}) \circ \mathrm{Ad}^1_\Lambda$ (Ad-covariance), where Ad_{Λ^1} is a projection of the adjoint coaction to Λ^1, that is, $\mathrm{Ad}_{\Lambda^1}([a]) = [\mathrm{Ad}(a)]$ (well defined, because Q is Ad-invariant for a bicovariant calculus, see the subsection "The Woronowicz theorems").

The correspondence between connections and connection 1-forms is given by the formula

$$\prod(p\mathrm{d}q) = \sum pq_{(0)}\omega([q_{(1)}])$$

In the universal differential calculus case, $\Lambda^1 = A^+$, hence ω can be viewed as a map $\omega : A \to \Omega^1 P$, such that $\omega(1) = 0$. The map $\mathcal{F}_\omega : A \to \Omega^2 P$, given by $\mathcal{F}_\omega = \mathrm{d}\omega + \omega * \omega$ is called a "curvature" of ω. The curvature satisfies the Bianchi identity, $\mathrm{d}\mathcal{F}_\omega = \mathcal{F}_\omega * \omega - \omega * \mathcal{F}_\omega$.

In the case of a trivial bundle with trivialization Φ and universal calculus, any linear map $\beta : A \to \Omega^1 B$ such that $\beta(1) = 0$ defines a connection 1-form

$$\omega = \Phi^{-1} * \mathrm{d}\Phi + \Phi^{-1} * \beta * \Phi$$

The corresponding curvature is $\mathcal{F}_\omega = \Phi^{-1} * \mathcal{F}_\beta * \Phi$, where $\mathcal{F}_\beta = \mathrm{d}\beta + \beta * \beta$.

In the case of a quantum homogeneous bundle with calculus determined by $Q_P \in P^+$ and $Q_A = \pi(Q_P)$ (cf. the subsection "Quantum homogeneous bundles"), a canonical connection form can be assigned to any algebra map $i : A \to P$ such that

1. $\pi \circ i = \mathrm{id}$ (i-splits π);
2. $\varepsilon_P \circ i = \varepsilon_A$ (co-unitality);
3. $(\mathrm{id} \otimes \pi) \circ \mathrm{Ad}_P \circ i = (i \otimes \mathrm{id}) \circ \mathrm{Ad}_A$ (Ad-covariance); and
4. $i(Q_A) \subseteq Q_P$ (differentiability).

Explicitly, $\omega([a]) = \sum (Si(a)_{(1)})\mathrm{d}i(a)_{(2)}$.

Covariant Derivative: Strong Connections

A covariant derivative associated to a connection Π is a map $\mathrm{D} : P \to P\Omega^1(B)P, p \mapsto \mathrm{d}p - \Pi(\mathrm{d}p)$. A covariant derivative maps elements of P into horizontal forms, since $\ker \Pi = P\Omega^1(B)P$, and satisfies the Leibniz rule $\mathrm{D}(bp) = (\mathrm{d}b)p + b\mathrm{D}p$, for all $b \in B$, $p \in P$.

A connection is "strong" provided $\mathrm{D}(p) \in \Omega^1(B)P$. A covariant derivative of a strong connection is a connection on module P in the sense of Connes. Furthermore, in the universal calculus case, and when A has invertible antipode, the existence of strong connections leads to rich gauge theory of associated bundles (cf. the subsection "Associated bundles: matter fields"). A connection in a trivial bundle

described in the subsection "Connections and connection forms" is strong (and every strong connection in a trivial bundle is of this form). Assuming invertibility of the antipode in A, a canonical connection in a quantum homogeneous bundle described in that subsection is strong provided Ad-covariance (3) is replaced by conditions $(\mathrm{id}\otimes\pi)\circ\Delta\circ i=(i\otimes\mathrm{id})\circ\Delta_A$ (right covariance) and $(\pi\otimes\mathrm{id})\circ\Delta\circ i=(\mathrm{id}\otimes i)\circ\Delta_A$ (left covariance), where Δ is a coproduct in P, and Δ_A is a coproduct in A.

In the universal calculus case, the map D can be extended to a map $\mathrm{D}:\Omega^1P\to\Omega^2P$ via the formula $\mathrm{D}(\rho)=\mathrm{d}\rho+\sum\rho_{(0)}\omega(\rho_{(1)})$. Then $\mathrm{D}\circ\mathrm{D}(p)=\sum p_{(0)}\mathcal{F}_\omega(p_{(1)})$, where $\mathcal{F}_\omega$ is the curvature of ω (cf. the subsection "Connections and connection forms"). This explains the relationship between a curvature understood as the square of a covariant derivative and $\mathcal{F}_\omega$.

Bundle Automorphisms and Gauge Transformations

A quantum bundle automorphism is a left B-linear right A-covariant (i.e., colinear) automorphism $F:P\to P$ such that $F(1)=1$. Bundle automorphisms form a group with operation $FG=G\circ F$. This group is isomorphic to the group $\mathcal{G}(P)$ of gauge transformations, that is, maps $f:A\to P$ that satisfy the following conditions:

1. $f(1_A)=1_P$ (unitality);
2. $\Delta_P\circ f=(f\otimes\mathrm{id})\circ\mathrm{Ad}$ (Ad-covariance); and
3. f is convolution invertible (cf. the subsection "Hopf algebra preliminaries").

The product in $\mathcal{G}(P)$ is the convolution product (cf. the subsection "Hopf algebra preliminaries"). The group of gauge transformations acts on the space of (strong) connection forms ω via the formula

$$f\triangleright\omega=f*\omega*f^{-1}+f*\mathrm{d}f^{-1},\quad\forall f\in\mathcal{G}(P)$$

This resembles the gauge transformation law of a gauge field in the standard gauge theory. The curvature transforms covariantly as $\mathcal{F}_{f\triangleright\omega}=f*\mathcal{F}_\omega*f^{-1}$.

In the case of a trivial principal bundle, gauge transformations correspond to a change of the trivialization and can be identified with convolution-invertible maps $\gamma:A\to B$ such that $\gamma(1)=1$. A map $\beta:A\to\Omega^1B$ that induces a connection as in the subsection "Connections and connection forms" is transformed to $\gamma*\beta*\gamma^{-1}+\gamma*\mathrm{d}\gamma^{-1}$, and the curvature $\mathcal{F}_\beta\mapsto\gamma*\mathcal{F}_\beta*\gamma^{-1}$.

Associated Bundles: Matter Fields

Given a right A-comodule (corepresentation) $\varrho:V\to V\otimes A$ one defines a quantum vector bundle associated to P as

$$E=\Big\{\sum_i v^i\otimes p^i\in V\otimes P\Big|\sum_i v^i_{(0)}\otimes p^i_{(0)}\otimes v^i_{(1)}p^i_{(1)}=\sum_i v^i\otimes p^i\otimes 1\Big\}\subseteq V\otimes P$$

E is a right B-module with product $(\sum_i v^i\otimes p^i)b=\sum_i v^i\otimes p^ib$. A right B-linear map $s:E\to B$ is called a section of E. The space of sections $\Gamma(E)$ is a left B-module via $(bs)(p)=bs(p)$.

The theory of associated bundles is particularly rich when A has a bijective antipode and P has a strong connection form ω. In this case, $\Gamma(E)$ is isomorphic to the left B-module Γ_ϱ of maps $\phi:V\to P$ such that $\Delta_P\circ\phi=(\phi\otimes\mathrm{id})\circ\varrho$. If V is finite dimensional, then Γ_ϱ is a finite projective B-module, that is, it is a module of sections of a noncommutative vector bundle in the sense of Connes. The strong connection induces a map $\nabla:\Gamma_\varrho\to\Omega^1B\otimes_B\Gamma_\varrho$, given by $\nabla(\phi)(v)=\mathrm{d}\phi(v)+\sum\phi(v_{(0)})\omega(v_{(1)})$. ∇ is a connection in the sense of Connes (in a projective left B-module), that is, for all $b\in B,\phi\in\Gamma_\varrho,\nabla(b\phi)=\mathrm{d}b\otimes_B\phi+b\nabla(\phi)$.

In the case of a trivial bundle, Γ_ϱ can be identified with the space of linear maps $V\to B$. Thus, sections of an associated bundle correspond to pullbacks of matter fields, as in the classical local gauge theory matter fields are defined as functions on a spacetime with values in a representation (vector) space of the gauge group.

The Dirac q-Monopole

This is an example of a strong connection in a quantum homogeneous bundle (cf. the subsection "Quantum homogeneous bundles"). $P=\mathrm{SU}_q(2)$ is a matrix Hopf $*$-algebra with matrix of generators

$$\begin{pmatrix}a & -qc^*\\ c & a^*\end{pmatrix}$$

and relations

$$ac=qca,\qquad ac^*=qc^*a,\qquad cc^*=c^*c$$
$$a^*a+c^*c=1,\qquad aa^*+q^2cc^*=1$$

where q is a real parameter. $A=\mathbb{C}[U(1)]$ is a Hopf $*$-algebra generated by unitary and group-like u (i.e., $uu^*=u^*u=1,\Delta(u)=u\otimes u$). The $*$-projection $\pi:P\to A$ is defined by $\pi(a)=u$. The coinvariant subalgebra B is generated by $x=cc^*$, $z=ac^*$, $z^*=ca^*$. The elements x and z satisfy relations

$$x^*=x,\qquad zx=q^2xz,$$
$$zz^*=q^2x(1-q^2x),\qquad z^*z=x(1-x)$$

Thus, B is the algebra of functions on the standard quantum 2-sphere. A strong connection is obtained from a bicovariant $*$-map $i: A \to P$ given by $i(u^n) = a^n$ (cf. the subsections "Quantum homogeneous bundles," "Connections and connection forms," and "Covariant derivative: strong connections"). Explicitly, the connection form reads

$$\omega(u^n) = \sum_{k=0}^{n} \binom{n}{k}_{q^{-2}} c^{*k} a^{*n-k} \mathrm{d}(a^{n-k} c^k)$$
$$\omega(u^{*n}) = \sum_{k=0}^{n} q^{2k} \binom{n}{k}_{q^{-2}} a^{n-k} c^k \mathrm{d}(c^{*k} a^{*n-k})$$

where the deformed binomial coefficients are defined for any number x by

$$\binom{n}{k}_x = \frac{(x^n - 1)(x^{n-1} - 1) \cdots (x^{k+1} - 1)}{(x^{n-k} - 1)(x^{n-k-1} - 1) \cdots (x - 1)}$$

There is a family $V_n, n \in \mathbb{Z}$ of one-dimensional corepresentations of $\mathbb{C}[\mathrm{U}(1)]$ with $V_n = \mathbb{C}$ and $\varrho^n(1) = 1 \otimes u^n, n \geq 0$ and $\varrho^n(1) = 1 \otimes u^{*n}, n < 0$. This leads to the family of finite projective modules $\Gamma_n = \Gamma_{\varrho^n}$ as described in the subsection "Associated bundles: matter fields." The Hermitian projectors $\boldsymbol{e}(n)$ of these modules come out as, for $n > 0$,

$$e(n)_{ij} = \sqrt{\binom{n}{i}_{q^{-2}} \binom{n}{j}_{q^{-2}}}\, a^{n-i} c^i c^{*j} a^{*n-j},$$
$$i, j = 0, 1, \ldots, n$$
$$e(-n)_{ij} = q^{i+j} \sqrt{\binom{n}{i}_{q^{-2}} \binom{n}{j}_{q^{-2}}}\, c^{*i} a^{*n-i} a^{n-j} c^j,$$
$$i, j = 0, 1, \ldots, n$$

The $\boldsymbol{e}(n)$ describe q-monopoles of magnetic charge $-n$. For example, the charge-1 projector explicitly reads

$$\begin{pmatrix} 1 - x & z^* \\ z & q^2 x \end{pmatrix}$$

and reduces to the usual charge-1 Dirac monopole projector when $q = 1$. The covariant derivatives ∇ are Levi-Civita or Grassmann connections in modules Γ_n corresponding to projectors $\boldsymbol{e}(n)$.

The q-Instanton

This is an example of a coalgebra bundle and the associated vector bundle, which is a deformation of an instanton (with instanton number 1). $P = \mathbb{C}[S_q^7]$ is the $*$-algebra of polynomial functions on the quantum 7-sphere. As a $*$-algebra it is defined by generators z_1, z_2, z_3, z_4 and relations

$$z_i z_j = q z_j z_i \quad (\text{for } i < j)$$
$$z_j^* z_i = q z_i z_j^* \quad (\text{for } i \neq j)$$
$$z_k^* z_k = z_k z_k^* + (1 - q^2) \sum_{j<k} z_j z_j^*,$$
$$\sum_{k=1} z_k z_k^* = 1$$

where $q \in \mathbb{R}$. The coaction of the $*$-Hopf algebra $A = \mathrm{SU}_q(2)$ (cf. the previous subsection) on P is constructed as follows. Start with the quantum group $\mathrm{U}_q(4)$, generated by a matrix $\boldsymbol{t} = (t_{ij})_{i,j=1}^4$ and view $\mathbb{C}[S_q^7]$ as a right quantum homogeneous space of $\mathrm{U}_q(4)$ generated by the bottom row in $\boldsymbol{t}$. Thus, there is a right coaction of $\mathrm{U}_q(4)$ on $\mathbb{C}[S_q^7]$ obtained by the restriction of the coproduct in $\mathrm{U}_q(4)$. Next, project $\mathrm{U}_q(4)$ to $\mathrm{SU}_q(2)$ by a suitable coideal and a right ideal in $\mathrm{U}_q(4)$. The corresponding canonical surjection $r: \mathrm{U}_q(4) \to \mathrm{SU}_q(2)$ is a coalgebra map, characterized as a right $\mathrm{U}_q(4)$-module map by $r(t_{11}t_{22} - qt_{12}t_{21}) = 1$ and

$$r(\boldsymbol{t}) = \begin{pmatrix} \boldsymbol{u} & 0 \\ 0 & \boldsymbol{u} \end{pmatrix}, \qquad \bar{\boldsymbol{u}} = \begin{pmatrix} u_{22} & -u_{21} \\ -u_{12} & u_{11} \end{pmatrix}$$

where $\boldsymbol{u} = (u_{ij})_{i,j=1}^2$ is the matrix of generators of $\mathrm{SU}_q(2)$ (cf. the previous subsection). When applied to the coaction of $\mathrm{U}_q(4)$ on $\mathbb{C}[S_q^7]$, r induces the required coaction $\Delta_P: \mathbb{C}[S_q^7] \to \mathbb{C}[S_q^7] \otimes \mathrm{SU}_q(2)$. Explicitly, the coaction comes out on generators as $\Delta_P(z_j) = \sum_i z_i \otimes r(t_{ij})$. The coaction Δ_P is not an algebra map. The coinvariant subalgebra B is a $*$-algebra generated by

$$a = z_1 z_4^* - z_2 z_3^*$$
$$b = z_1 z_3 + q^{-1} z_2 z_4$$
$$R = z_1 z_1^* + z_2 z_2^*$$

The elements a, a^*, b, b^*, R satisfy the following relations:

$$Ra = q^{-2} aR, \qquad Rb = q^2 bR$$
$$ab = q^3 ba, \qquad ab^* = q^{-1} b^* a$$
$$aa^* + q^2 bb^* = R(1 - q^2 R)$$
$$aa^* = q^2 a^* a + (1 - q^2) R^2$$
$$b^* b = q^4 bb^* + (1 - q^2) R$$

Hence B can be understood as a deformation of the algebra of functions on the 4-sphere and is denoted by $\mathbb{C}[\Sigma_q^4]$. One can show that the map "can" in the subsection "Generalizations of quantum principal bundles" is bijective, hence there is an $\mathrm{SU}_q(2)$-coalgebra principal bundle with the total space the quantum 7-sphere $\mathbb{C}[S_q^7]$ and the base space the

quantum 4-sphere $\mathbb{C}[\Sigma_q^4]$. By abstract arguments that involve cosemisimplicity of $SU_q(2)$, one can prove that there exists a strong connection in this bundle; this is the q-deformed instanton field. At the time of writing this article, however, the explicit form of this connection is not known.

On the other hand, following the classical construction of an instanton, one can take the fundamental two-dimensional corepresentation $V = \mathbb{C}^2$ of $SU_q(2)$ and explicitly construct q-instanton projection with instanton number 1. Writing e_1, e_2 for the basis of V, the coaction $\varrho : V \to V \otimes SU_q(2)$ is given by

$$(e_j) = \sum_i e_i \otimes u_{ij}$$

The associated bundle (cf. the subsection "Associated bundles: matter fields") is a finite projective left module over $\mathbb{C}[\Sigma_q^4]$. The corresponding q-instanton projector comes out as

$$\begin{pmatrix} q^2R & 0 & qa & q^2b \\ 0 & q^2R & qb^* & -q^3a^* \\ qa^* & qb & 1-R & 0 \\ q^2b^* & -q^3a & 0 & 1-q^4R \end{pmatrix}$$

See also: Bicrossproduct Hopf Algebras and Noncommutative Spacetime; Hopf Algebras and q-Deformation Quantum Groups; Noncommutative Tori, Yang–Mills, and String Theory.

Further Reading

Bonechi F, Ciccoli N, and Tarlini M (2003) Noncommutative instantons and the 4-sphere from quantum groups. *Communications in Mathematical Physics* 226: 419–432.

Bonechi F, Ciccoli N, Dąbrowski L, and Tarlini M (2004) Bijectivity of the canonical map for the noncommutative instanton bundle. *Journal of Geometry and Physics* 51: 71–81.

Brzeziński T and Majid S (1993) Quantum group gauge theory on quantum spaces. *Communications in Mathematical Physics* 157: 591–638.

Brzeziński T and Majid S (1995) Quantum group gauge theory on quantum spaces – Erratum. *Communications in Mathematical Physics* 167: 235.

Brzeziński T and Majid S (2000) Quantum geometry of algebra factorisations and coalgebra bundles. *Communications in Mathematical Physics* 213: 491–521.

Calow D and Matthes R (2002) Connections on locally trivial quantum principal fibre bundles. *Journal of Geometry and Physics* 41: 114–165.

Dąbrowski L, Grosse H, and Hajac PM (2001) Strong connections and Chern–Connes pairing in the Hopf–Galois theory. *Communications in Mathematical Physics* 220: 301–331.

Hajac PM and Majid S (1999) Projective module description of the q-monopole. *Communications in Mathematical Physics* 206: 247–264.

Heckenberger I and Schmüdgen K (1998) Classification of bicovariant differential calculi on the quantum groups $SL_q(n+1)$ and $Sp_q(2n)$. *Journal für die Reine und Angewandte Mathematik* 502: 141–162.

Klimyk A and Schmüdgen K (1997) *Quantum Groups and Their Representations*. Berlin: Springer.

Majid S (1998) Classification of bicovariant differential calculi. *Journal of Geometry and Physics* 25: 119–140.

Majid S (1999) Quantum and braided Riemannian geometry. *Journal of Geometry and Physics* 30: 113–146.

Majid S (2000) *Foundations of Quantum Group Theory*, 1st pbk. edn. Cambridge: Cambridge University Press.

Wess J and Zumino B (1990) Covariant differential calculus on the quantum hyperplane, *Nuclear Physics B. Proceedings Supplement* 18B: 302–312.

Woronowicz SL (1989) Differential calculus on compact matrix pseudogroups (quantum groups). *Communications in Mathematical Physics* 122: 125–170.

Quantum Hall Effect

K Hannabuss, University of Oxford, Oxford, UK

Introduction

When a current flows in a thin sample with a transverse magnetic field B, the Lorentz force deflects the trajectories of the charge carriers, producing an excess charge on one side and a charge deficiency on the other, and creating a potential difference across the conductor perpendicular to both the direct current and the magnetic field. This is known as the Hall effect, in honour of E H Hall, who, inspired by a remark of Maxwell, first demonstrated it in thin samples of gold foil in October 1879 (Hall's subsequent measurements of the potential difference showed that the carriers could be positively or negatively charged for different materials). A schematic diagram of Hall's experiment and the lateral separation of charges is shown in Figure 1.

Equilibrium is reached when the magnetic force balances that from the potential difference E due to the displaced charge. When the charge carriers are electrons, with the electron density n, and the electron current J, this gives $neE = JB$. Comparison with Ohm's law, $J = \sigma E$, gives conductance (the reciprocal of resistance) to be $\sigma = ne/B$. More generally, considering the currents and fields as

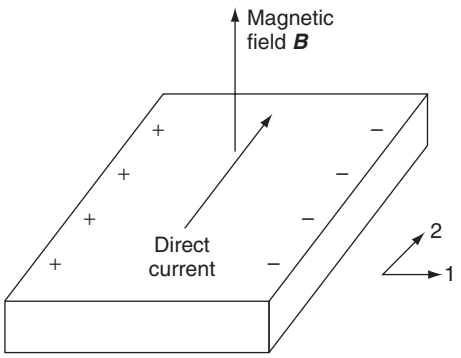


Figure 1 Schematic diagram of charge separation in Hall's experiment.

vectors, σ is represented by a matrix. Rescaling by the sample thickness δ, the diagonal components of $\delta\sigma$ give the direct conductivity $\sigma_{\|}$ and its off-diagonal elements give the Hall conductivity: $\sigma_H = \delta\sigma_{21}$. (For systems symmetric under 90° rotations, $\sigma_{11} = \sigma_{22}$ and $\sigma_{12} = -\sigma_{21}$.) In quantum theory, one usually works in terms of the filling fraction $\nu = n\delta h/eB$ and then $\sigma_H = \nu e^2/h$.

In 1980 von Klitzing, Dorda, and Pepper discovered that at very low temperatures in very high magnetic fields, the Hall conductivity σ_H is quantized as integral multiples of $e^2/\hbar$, a fact known as the integer quantum hall effect (IQHE). The integer multiples were accurate to 1 part in 10^8, and the effect was exceptionally robust against changes in the geometry of the samples and in the experimental parameters. Indeed, the unprecedented accuracy of the effect led to its adoption as the international standard for resistance in 1990.

More precisely, the Hall conductivity was no longer proportional to the filling fraction ν, but the graph of σ_H against ν displayed a sequence of jumps, as shown in Figure 2. In this figure, the conductivity has plateau at the integer multiples of $e^2/\hbar$, and jumps between them within fairly small ranges of the filling fraction. Moreover, the direct conductivity vanishes where the Hall conductivity takes its constant integral values.

These results raise numerous questions.

1. Why does the conductivity take such precise integer values, and why are they so stable under changes of the geometry and physical parameters?
2. Why does the direct conductivity vanish, except in regions where the Hall conductivity jumps between integer values, and how are such jumps possible?

Moreover, any theory must also explain why these features are not present under the more normal

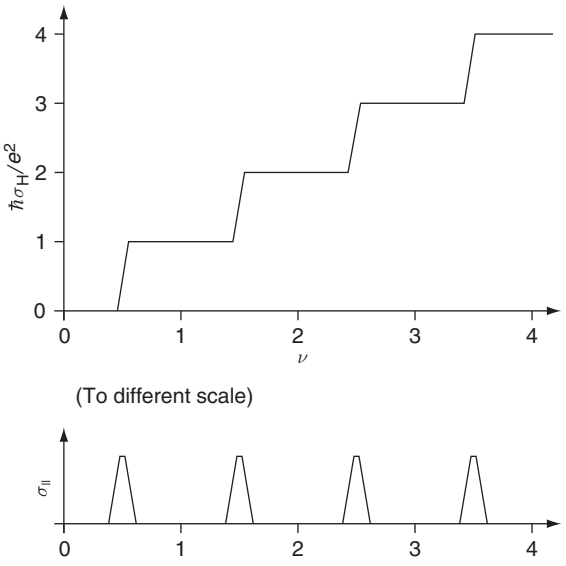


Figure 2 Schematic diagram of the Hall and direct conductivities plotted against the filling factor ν.

conditions of the classical Hall effect. The following features seem to play a role, and in the case of the first three, even in the classical effect.

1. As Hall discovered, the samples must be very thin to exhibit even the classical effect. (Nowadays they are often a surface layer between two semiconductors.)
2. The samples are macroscopic and much larger than the quantum wavelengths appearing in the problem.
3. The electric field is small enough that nonlinear effects are negligible.
4. The quantum effect appears only at a very low temperature.

The first of these suggests that we should idealize to the case where the motion of the charge carriers is restricted to a two-dimensional region, and the second that we may work in the thermodynamic limit where the conducting surface is the whole of $\boldsymbol{R}^2$. The third and fourth ensure both that the linear Ohm's law should be adequate, and also that it should be enough to consider the limiting cases of very weak electric fields and zero temperature. Multiple limits of this sort raise delicate mathematical issues. Indeed, many plausible models of the effect turn out, on careful analysis, to predict vanishing Hall conductivity.

A theoretical explanation of the quantization of the conductivity was soon suggested by Laughlin. Exploiting the apparent independence of sample geometry, he considered a cylindrical conductor where quantization followed on consideration of

the flux tubes threading it. Laughlin's choice of a particular configuration precluded investigation of the influence of changing geometry. This was soon provided by Thouless, Kohmoto, Nightingale, and de Nijs, who argued (from a lattice version of the problem) that the conductivity could be identified with the Chern character of a line bundle over a Brillouin zone (a quotient of momentum space by the action of the reciprocal crystal lattice), so that it had to be integral and the stability of the effect was a consequence of the topological nature of σ. Unfortunately, whilst suggestive, this explanation worked only under the physically implausible constraint that the magnetic flux through a crystal cell was rational, offered no explanation of the link between the Hall and direct conductivities, and, working with a periodic Hamiltonian, made no allowance for the impurities and disorder usually important in solid-state problems of this sort.

Notwithstanding these deficiencies, this model contained important insights, which inspired Bellissard to model the effect using Connes' newly developed noncommutative geometry (Bellissard 1986, Connes 1986). (Kunz produced a Hilbert space theory at about the same time, but that has been rather less influential.) Connes' work turned out to contain all the relevant concepts and tools needed to provide a good understanding of the effect, based on interpreting the conductivity as a noncommutative Chern character for a noncommutative version of the Brillouin zone. In fact, the techniques of noncommutative geometry seemed to fit the quantum Hall effect so well that this has become one of the standard examples of the theory.

Even whilst the theorists were struggling to explain the experiments, observations by Tsui, Störmer, and Gossard showed that, with suitable care, fractional Hall conductivities could also be observed, although these were far less stable than those given by integers. One, therefore, distinguishes between IQHE and the fractional quantum Hall effect (FQHE), and this survey concentrates largely on the former. One simplifying feature of the IQHE is that it seems to be comprehensible at the level of individual noninteracting electrons, whereas the FQHE certainly involves some kind of interaction and many-body theory.

This article presents an outline of the connection between noncommutative differential geometry and the IQHE, and concludes by discussing some of the approaches to the FQHE, and some other applications of noncommutative geometry and mathematical directions suggested by the theory. The sections alternate between the physical model and the mathematical abstraction from it.

There are good surveys of the area (Bellissard *et al.* 1994, McCann 1998) explaining how the mathematical model arises out of the physics, the mathematical models themselves. As well as being the standard reference for noncommutative geometry, Connes (1994) discusses the Hall effect. These resources contain good bibliographies, which may be consulted for further references.

Electron Motion in a Magnetic Field

The following discussion restricts attention to motion in two dimensions, with electrons as the charge carriers, and no interactions between them. (The first condition is essential; the second could be relaxed a little to allow sufficiently long-lived quasi-particles.) A single free electron with mass m and charge e moving in the x_1–x_2 plane with a constant transverse magnetic field $\boldsymbol{B}$ in the positive x_3-direction, can be described by the Landau Hamiltonian

$$H_{\mathrm{L}} = |\boldsymbol{P} - e\boldsymbol{A}|^2/2m \qquad [1]$$

where $\boldsymbol{A} = \frac{1}{2}\boldsymbol{B} \times \boldsymbol{X}$ is a magnetic vector potential that gives rise to $\boldsymbol{B}$. This problem is exactly solvable by, for example, introducing $\boldsymbol{K}_{\pm} = (K^1_{\pm}, K^2_{\pm}) = \boldsymbol{P} \mp e\boldsymbol{A}$. The components of $\boldsymbol{K}_+$ and $\boldsymbol{K}_-$ commute with each other, but $[K^1_{\pm}, K^2_{\pm}] = \pm \mathrm{i}\hbar eB$. Comparison with the harmonic oscillator shows that the energy spectrum of $H_{\mathrm{L}} = [(K^1_+)^2 + (K^2_+)^2]/2m$ is $\{(n + \frac{1}{2})\hbar eB/m\colon n \in \boldsymbol{Z}\}$. Since H_{L} commutes with the components of $\boldsymbol{K}_-$, each of these Landau energy levels is infinitely degenerate, and the filling fraction ν measures what proportion of states in the Landau levels are filled. The frequency $\omega_{\mathrm{c}} = eB/m$ is the cyclotron frequency for classical circular orbits in the magnetic field.

The degeneracy of the Landau Hamiltonian can also be understood in terms of the magnetic translations obtained by exponentiating the connection defined by the magnetic potential $\boldsymbol{A}$: $\nabla_j = \partial_j + \mathrm{i}eA_j/\hbar = \mathrm{i}K^j_-/\hbar$. More precisely, we set

$$U(\boldsymbol{a}) = \exp(-\mathrm{i}\boldsymbol{a} \cdot \nabla) = \exp(-\mathrm{i}\boldsymbol{a} \cdot \boldsymbol{K}_-/\hbar) \qquad [2]$$

which clearly commutes with H_{L}, expressing the translational symmetry of this model. The curvature $[\nabla_1, \nabla_2] = B$ of the connection manifests itself in the identities

$$U(\boldsymbol{a})U(\boldsymbol{b}) = \mathrm{e}^{(1/2)\mathrm{i}\phi}U(\boldsymbol{a} + \boldsymbol{b}) = \mathrm{e}^{\mathrm{i}\phi}U(\boldsymbol{b})U(\boldsymbol{a}) \qquad [3]$$

where $\phi = e\boldsymbol{B} \cdot (\boldsymbol{a} \times \boldsymbol{b})/\hbar$ measures the magnetic flux through the parallelogram spanned by $\boldsymbol{a}$ and $\boldsymbol{b}$. These show that U is a projective representation of $\boldsymbol{R}^2$ with projective multiplier $\lambda(\boldsymbol{a}, \boldsymbol{b}) = \exp(\frac{1}{2}\mathrm{i}\phi)$.

The significance of this is that, unless ϕ is an integer multiple of 2π, $U(\boldsymbol{a})$ and $U(\boldsymbol{b})$ generate a noncommutative algebra. This replaces the commutative algebra of functions on two-dimensional momentum space and leads naturally to a noncommutative geometry.

The unembellished Landau Hamiltonian cannot describe the Hall effect without adding an electric potential $e\boldsymbol{E}\cdot\boldsymbol{X}$ to drive the current in the sample. (Alternatively, and useful for the later discussion, one could use the radiation gauge in which, instead of introducing a scalar potential, a time-dependent term is added to $\boldsymbol{A}$ so that $\boldsymbol{E}=-\partial\boldsymbol{A}/\partial t$.)

The quantum Hall effect also depends crucially on the effects of impurities in the conducting material. These can be modeled by adding a random potential V_ω with ω in a compact probability space Ω to obtain $H_\omega = H_{\mathrm{L}} + e\boldsymbol{E}\cdot\boldsymbol{X} + V_\omega(X)$. A continuous function f on Ω can be interpreted as a random variable, and its expectation $\tau_\Omega(f)$ gives a trace on the C*-algebra $C(\Omega)$ (i.e., a positive linear functional such that $\tau_\Omega(AB)=\tau_\Omega(BA)$).

Although the magnetic translations commute with H_{L}, they do not generally commute with the potentials so they act on Ω, but, on the other hand, the physics of a disordered system and its translates should be the same, so we assume that the probability measure and hence also τ_Ω are invariant under magnetic translations. (As noted earlier, we work in the thermodynamic limit, where the Hall sample expands to fill $\boldsymbol{R}^2$, so we do not need to worry about translations moving the sample itself.) Then Ω with the magnetic translation action can be interpreted as the noncommutative Brillouin zone. (A space Ω can be reconstructed from the magnetic translations of the resolvents of the Hamiltonians (Bellisard *et al.* 1994).)

The current $\boldsymbol{J}$ may be defined as the functional derivative of the Hamiltonian with respect to the vector potential $\boldsymbol{A}$ or, in components, $J_k=\delta_k H = \delta H/\delta A_k$. For the Landau Hamiltonian, this gives

$$\mathrm{i}\hbar\delta_k H_{\mathrm{L}} = \mathrm{i}e\hbar(P_k - eA_k)/m = e[X_k, H_{\mathrm{L}}] \qquad [4]$$

a relation which persists for $H=H_{\mathrm{L}}+V(X)$ whenever the potential V is independent of $\boldsymbol{A}$, so that $\delta_k H = -\mathrm{i}e[X_k,H]/\hbar = e\,\mathrm{d}X_k/\mathrm{d}t$, the charge times velocity, as one might expect. The operator functional calculus delivers a similar formula for derivations of the spectral projections of H. We have $\delta_k = e\partial_k/\hbar$, where, in view of the commutation relations, $\partial_k = -\mathrm{i}[X_k,\cdot\,]$ can be regarded as a momentum-space derivative, confirming that we are dealing with the differential geometry of momentum space.

We now wish to calculate the expected current $\langle J_k\rangle$, in a thermal state with chemical potential μ at inverse temperature $\beta=1/kT$ (where k is Boltzmann's constant and the temperature is T(kelvin). Using the Fermi–Dirac distribution, the grand canonical expectation value is

$$\langle J_k\rangle = \mathrm{tr}\left[\left(1+\mathrm{e}^{-\beta(H-\mu)}\right)^{-1} J_k\right] \qquad [5]$$

Since the quantum Hall effect occurs at low temperatures (large β) and for weak fields, we formally proceed to those limits. Then $(1+\mathrm{e}^{-\beta(H-\mu)})^{-1}$ tends to the projection P_{F} onto the states with energy less than the Fermi energy E_{F} in the absence of the electric field. The limiting expected current is, therefore, $\mathrm{tr}(P_{\mathrm{F}}J_k)=\mathrm{tr}(P_{\mathrm{F}}\delta_k H)$, where H is now the Hamiltonian including the electric field (without which there would be no current).

A detailed calculation of the Hall conductivity using the Kubo–Greenwood formula shows that the conductivity matrix is actually

$$\sigma_{kj} = \mathrm{i}(e^2/\hbar)\mathrm{tr}(P_{\mathrm{F}}[\partial_j P_{\mathrm{F}}, \partial_k P_{\mathrm{F}}]) \qquad [6]$$

In particular, this immediately implies that the direct conductivity terms σ_{jj} vanish, as observation suggested. The derivation of [6] requires great care, and references may be found in the surveys, but a formal argument in the next section may lend this expression some plausibility.

The Noncommutative Geometry

The principal ingredient for noncommutative geometry is an algebra, and thus we shall now consider a class of algebras broad enough to include the physical example.

The action of the magnetic translations on Ω defines automorphisms of the C*-algebra $C(\Omega)$, which permit the construction of a twisted crossed-product algebra, in which these automorphisms are represented by conjugation. Because much of the theory has been formulated with lattice approximations using $\boldsymbol{Z}^2$ rather than $\boldsymbol{R}^2$, it is useful to work more generally with a separable locally compact abelian group G with continuous multiplier λ, and a homomorphism α to automorphisms of a C*-algebra $\mathcal{A}_1$ with trace τ_1, which will in practice be the commutative algebra $C(\Omega)$ with τ_Ω. The twisted crossed product $\mathcal{A}=C(\mathcal{A}_1, G, \lambda)$ can be constructed as the norm completion of the continuous compactly

supported functions from G to $\mathcal{A}_1$ with the product, adjoint and norm

$$(f^*g)(x) = \int_G \lambda(y, x-y) f(y)(\alpha_y g)(x-y)\,\mathrm{d}y \quad [7]$$

$$f^*(x) = \lambda(x,-x)^{-1} f(-x)^* \quad [8]$$

$$\|f\| = \max\left\{\int_G \|f(x)\|_{\mathcal{A}_1}\,\mathrm{d}x, \int_G \|f^*(x)\|_{\mathcal{A}_1}\,\mathrm{d}x\right\} \quad [9]$$

integration being with respect to the Haar measure. The crossed-product algebra is noncommutative, both because of the action of G and due to the multiplier λ. It has a trace $\tau[f] = \tau_1[f(0)]$ and, when $G = \boldsymbol{R}^2$, has derivations given by $\partial_k f = -\mathrm{i}x_k f(\boldsymbol{x})$.

As an example, consider the case of periodic potentials invariant under translation by vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Then the group $G \cong \boldsymbol{Z}^2$ generated by $\boldsymbol{a}$ and $\boldsymbol{b}$ acts trivially on Ω and the crossed-product algebra is just a product of $\mathcal{A}_1$ and the twisted group algebra of complex-valued functions $C(\mathbf{C}, G, \lambda)$, generated by $U(\boldsymbol{a})$ and $U(\boldsymbol{b})$. We already noted that the algebra is commutative only when the flux $\phi \in 2\pi\boldsymbol{Z}$, in which case it is just the convolution algebra of $\boldsymbol{Z}^2$, which by Fourier transforming (effectively setting $U(\boldsymbol{a}) = \mathrm{e}^{\mathrm{i}\alpha}$ and $U(\boldsymbol{b}) = \mathrm{e}^{\mathrm{i}\beta}$) is the algebra $C(\boldsymbol{T}^2)$, with torus coordinates α and β. For fluxes which are rational multiples of 2π we obtain a matrix algebra, whilst irrational fluxes give an infinite-dimensional irrational rotation algebra or noncommutative torus, a standard example in noncommutative geometry.

Any $*$-representation ρ of $\mathcal{A}_1$ on a Hilbert space $\mathcal{H}_\rho$ can be induced to a $*$-representation π_ρ of the twisted crossed product on $\mathcal{H} = L^2(G, \mathcal{H}_\rho)$ by setting

$$(\pi_\rho(f)\psi)(x) = \int_G \lambda(x, y-x)^{-1}\rho(\alpha_{-x}f)(y-x)\psi(y)\,\mathrm{d}y \quad [10]$$

for $f \in \mathcal{A}$ and $\psi \in \mathcal{H}$. When $\mathcal{A}_1 = C(\Omega)$, we may take ρ to be a one-dimensional irreducible $*$-representations given by evaluating the function at a point $\omega \in \Omega$.

When $G = \boldsymbol{R}^2$, it is easy to construct a Fredholm module from π_ρ. The space $\mathcal{H}_2 = \mathcal{H} \otimes \mathbf{C}^2$ has actions π_ρ of $\mathcal{A}$ on the first factor and of the Pauli spin matrices $\sigma_1, \sigma_2, \sigma_3$, on the second. It may be regarded as a graded module with grading operator σ_3, and

$$F = (x_1^2 + x_2^2)^{-1/2}(x_1\sigma_1 + x_2\sigma_2) \quad [11]$$

provides a Connes–Fredholm involution which anticommutes with σ_3. Detailed technical results of Connes show how to use the supertrace on $\mathcal{H}_2$ and the Dixmier trace to interpret the physically important quantities in this setting.

We now turn to the formal derivation of the key alternative expression for the conductivity. In the abstract algebraic setting, when $p \in \mathcal{A}$ is a projection in the domain of a derivation δ the derivative of $(1-p)p = 0$ gives

$$0 = \delta((1-p)p) = (1-p)\delta(p) - \delta(p)p \quad [12]$$

and then an easy calculation leads to

$$[p, [\delta p, p]] = 2p(\delta p)p - (\delta p)p^2 - p^2(\delta p) = -\delta p \quad [13]$$

In the identity for elements a, b, c, and $h \in \mathcal{A}$

$$\begin{aligned}\tau([a,[b,c]]h) - \tau(c[[h,a],b]) \\ = \tau([a,[b,c]h]) + \tau([b,c[h,a]]) = 0\end{aligned} \quad [14]$$

we set $a = c = p$ and $b = \delta p$ to obtain

$$\tau([p,[\delta p,p]]h) = \tau(p[[h,p],\delta p]) \quad [15]$$

Combining this with [12] when $\tau \circ \delta = 0$, one obtains

$$\begin{aligned}\tau(p\delta h) &= \tau(\delta(ph)) - \tau(\delta(p)h) \\ &= \tau([p,[\delta p,p]]h) = \tau(p[[h,p],\delta p])\end{aligned} \quad [16]$$

The Hall Conductivity and Anderson Localisation

Substituting $p = P_\mathrm{F}$ and $h = H$ in formula [16] would give the current $\mathrm{tr}(P_\mathrm{F}[[H, P_\mathrm{F}], \delta P_\mathrm{F}])$. Since δ_k is proportional to the commutator with X_k, it is true that $\mathrm{tr} \circ \delta_k = 0$, but, unfortunately, P_F need not lie in the domain of δ_k, and H is unbounded, further compounding the difficulties. These are serious problems, although the situation is not quite as bad as it seems. Without the electrostatic term $e\boldsymbol{E}\cdot\boldsymbol{X}$ in H, P_F would have been a spectral projection with which H would commute, so that

$$[H, P_\mathrm{F}] = e[\boldsymbol{E}\cdot\boldsymbol{X}, P_\mathrm{F}] = eE_j[X_j, P_\mathrm{F}] = \mathrm{i}eE_j\partial_j P_\mathrm{F} \quad [17]$$

and H disappears from the formula, to be replaced by $\partial_j P_\mathrm{F}$. This would give the expected current $\mathrm{i}(e^2/\hbar)\mathrm{tr}(P_\mathrm{F}[\partial_j P_\mathrm{F}, \partial_k P_\mathrm{F}])E_j$, and the conductivity matrix

$$\sigma_{kj} = \mathrm{i}(e^2/\hbar)\mathrm{tr}(P_\mathrm{F}[\partial_j P_\mathrm{F}, \partial_k P_\mathrm{F}]) \quad [18]$$

given earlier (there is no need to scale by the thickness in two dimensions).

However it is derived, this expression for the conductivity only makes sense under suitable conditions, otherwise $\mathrm{tr}(P_F[\partial_j P_F, \partial_k P_F])$ might either be undefined (because P_F is not differentiable) or might not be trace class. There is a simple condition sufficient to handle both these difficulties, which also leads to an interesting physical insight. From the obvious inequality

$$0 \leq \mathrm{tr}\left[P_F(\partial_1 P_F \pm i\partial_2 P_F)^*(\partial_1 P_F \pm i\partial_2 P_F)\right] \quad [19]$$

$$= \mathrm{tr}\left[P_F\left((\partial_1 P_F)^2 + (\partial_2 P_F)^2\right)\right] \pm i\,\mathrm{tr}(P_F[\partial_1 P_F, \partial_2 P_F]) \quad [20]$$

and the fact that $1 \geq P_F$, we deduce that

$$\mathrm{tr}\left[\left((\partial_1 P_F)^2 + (\partial_2 P_F)^2\right)\right] \geq \mathrm{tr}\left[P_F\left((\partial_1 P_F)^2 + (\partial_2 P_F)^2\right)\right] \geq |\mathrm{tr}(P_F[\partial_1 P_F, \partial_2 P_F])| \quad [21]$$

Thus, if $\mathrm{tr}[((\partial_1 P_F)^2 + (\partial_2 P_F)^2)]$ exists and is finite, then our expression for the conductivity is well defined. Mathematically, this is a Sobolev type condition. To see the physical significance, we recall that $\partial_k P_F = -i[X_k, P_F]$, so that the condition is equivalent to the finiteness of $\mathrm{tr}[(X_1^2 + X_2^2)P_F{}^2] - \mathrm{tr}[(X_1 P_F)^2 + (X_2 P_F)^2]$.

This condition imposes a requirement for some localization in the system (when P_F is a rank-1 projection, it reduces to the requirement that the variance $\langle X_1^2 + X_2^2\rangle - \langle X_1\rangle^2 - \langle X_2\rangle^2$ be finite). This links with a much older observation of Anderson that the interference caused by impurities in a crystal, which cancel at long range, should, at smaller distances, cause localized clumping. The mathematical development of this idea by Pastur provides an appropriate tool for handling the conditions for the valdiity of the conductivity formula. The impurities generating Anderson localization are provided in this model by the random potential in the Hamiltonian. It also leads us to restrict attention to the dense subalgebra $\mathcal{A}_0$ of $f \in \mathcal{A}$, where $\tau[(\partial_1 f)^*(\partial_1 f) + (\partial_2 f)^*(\partial_2 f)] < \infty$.

The Integral Quantum Hall Effect

Having identified the features of physical interest, we can return to the abstract algebraic description with conductivity $i(e^2/\hbar)\tau(p[\partial_j p, \partial_k p])$. The key observation is that this can be interpreted as the Connes pairing between a cyclic cocycle c_τ on $\mathcal{A}_0$ and the projection p whose stable equivalence class represents an element of the C*-algebraic K-theory, $K_0(\mathcal{A})$. Such pairings give noncommutative Chern characters. The cyclic cocycle is a trilinear form defined on elements $a_0, a_1, a_2 \in \mathcal{A}_0$ by

$$c_\tau(a_0, a_1, a_2) = \tau[a_0(\delta_1 a_1 \delta_2 a_2 - \delta_2 a_1 \delta_1 a_2)] \quad [22]$$

This is easily shown to be cyclic, $c_\tau(a_0, a_1, a_2) = c_\tau(a_1, a_2, a_0)$, and to satisfy the cyclic 2-cocycle condition

$$c_\tau(a_0 a_1, a_2, a_3) - c_\tau(a_0, a_1 a_2, a_3) + c_\tau(a_0, a_1, a_2 a_3) - c_\tau(a_3 a_0, a_1, a_2) = 0 \quad [23]$$

The Hall conductivity $\sigma_{21} = ic_\tau(p, p, p)e^2/\hbar$ can now be interpreted as the noncommut ative Chern character defined by the projection p.

This interpretation of the Hall conductivity clears the way to prove that it is integral, and there are several different routes to this.

One approach is to identify the conductivity with some kind of index which is clearly integral. Bellissard worked with the Fredholm module where, by results of Connes, the Chern character is interpreted as the index of the Fredholm operator $\pi_\rho(p)F\pi_\rho(p)$. Avron, Seiler and Simon have interpreted the conductivity as a relative index $\dim[\ker(P_F - Q_F - 1)] - \dim[\ker(Q_F - P_F - 1)]$ of the projections P_F and its conjugate $Q_F = uP_F u^*$ by an off-diagonal element u of F. This is particularly interesting as the conjugation by u can be interpreted as a nonsingular gauge transformation of exactly the kind introduced by Laughlin in his original explanation of the quantum Hall effect in terms of singular flux tubes piercing a cylindrical conductor.

Xia suggested another approach rewriting $\mathcal{A}$ as a repeated crossed product with $\boldsymbol{R}$, which allows us to calculate $K_0(\mathcal{A})$, using either Connes' Thom isomorphism theorem or the Takai duality theorem for stable algebras to get

$$K_0(\mathcal{A}) = K_0\left[C(\mathcal{A}_1, G, \lambda)\right] \cong K_0(\mathcal{A}_1) \quad [24]$$

which, when $\mathcal{A}_1 = C(\Omega)$, is just $K^0(\Omega)$, leading to identification as a topological index. For the simplest case of $\Omega = \boldsymbol{T}^2$, this gives $K^0(\Omega) \cong \boldsymbol{Z}^2$. The image of τ, and so also c_τ, actually sits in just one component, leading to quantization of the Hall conductivity.

The two questions posed in the introduction can now be answered as follows: The Hall conductivity can be identified with a topological index which can take only integer values, and therefore does not respond to continuous changes in any of the physical parameters until the change brings the system into a region where one of the background assumptions fails, such as a breakdown in the localization condition. The same conditions also ensure that the direct current vanishes. Roughly speaking, the

plateaus occur when the Fermi energy is in a gap in the extended (nonlocalized) spectrum.

This brief overview has omitted many of the interesting features of the detailed theory, which can be found in the surveys, such as the fact that low-lying energy levels do not contribute to the conductivity, and Shubin's theorem identifying $\tau(p)$ as the integrated density of states. Harper's equation describing a discrete lattice analog of the IQHE has been a test-bed for many of the ideas, and various results were first proved in that setting. The FQHE was discovered during an unsuccessful search for a Wigner crystal phase transition, but analysis of discrete models provides strong evidence that Hall conductors have very complicated phase diagrams.

The Fractional Quantum Hall Effect

As mentioned in the introduction, by the time IQHE had been understood theoretically, it had been found that, with appropriate care, fractional conductivities could also be observed, although they were much less precise and stable than the integer values, and the plateaus less pronounced. Although there have been many phenomenological explanations, there is as yet no mathematical understanding from quantum field theory as compelling as that for the integer effect. We shall briefly summarize some of the main lines of attack.

The first explanation, again due to Laughlin, has also provided the basis for many subsequent treatments of the problem. The wave functions of the oscillator-like Landau Hamiltonian can conveniently be represented in the Bargmann–Segal Fock space of holomorphic functions f on $\boldsymbol{R}^2 \sim \boldsymbol{C}$ which are square-integrable with respect to a Gaussian measure. Incorporating the measure into the functions, these have the form $f(z)\exp(-|z|^2/2)$. Many particle wave functions are similarly realized in terms of holomorphic functions on $\boldsymbol{C}^N$, and must be antisymmetric under odd permutations of the particles to describe fermions. This quickly leads one to consider functions of the form

$$\prod_{r<s}(z_r - z_s)^k \exp\left(-\sum_j |z_j|^2/2\right) \qquad [25]$$

for odd integers $k > 0$, and their multiples by even holomorphic functions. The lowest energy where such a wave function occurs is when $k = 1$, and larger values of k have the effect of dividing the Hall conductivity by k, which produces fractional conductivities.

Halperin suggested quite early that counterflowing currents in the interior of a sample would tend to cancel, so that most of the current would be carried near the edge of the sample. There are several mathematical derivations of this, by, for example, Macris, Martin, and Pulé, and by Fröhlich, Graf, and Walcher. The K-theory of the boundary and bulk of a sample can be linked by exact sequences such as those of the commutative theory (Kellendonk *et al.* 2000), and even in the IQHE boundary and bulk conductivities can be used (Schulz-Baldes *et al.* 2002).

It has been fairly clear that whilst the IQHE can already be understood in terms of the motion of a single electron, the fractional effect is a many-body cooperative effect. One attempt to simplify the description is to work with an incompressible quantum fluid, and for edge currents one should study the boundary theory of such a fluid, in which the dominant contribution to the action is a Chern–Simons term, with conductivity as a coefficient. For an annular sample, this leads, in a suitable limit, to a chiral Luttinger model on the boundary circles, which can then be tackled mathematically using the representation theory of loop groups. This leads to some elegant mathematics, including extensions to multiple coupled bands, with conductivities described by Cartan matrices, as explained in the International Congress of Mathematicians (ICM) survey (Fröhlich 1995), and in the review by Fröhlich and Studer (1993).

The theory of composite fermions provides another physical approach in which field-theoretic effects result in the electrons sharing their charges in such a way as to produce fractional charges, and there is experimental evidence of such fractional charges in studies of tunneling from one edge to another. Then the FQHE is easily understood by simply replacing the electron charge e by e/k in the appropriate formulas.

Susskind has suggested combining noncommutative geometry with the theory of incompressible quantum fluids, an idea taken up by Polychronakos (2001). There are intriguing mathematical parallels with work by Berest and Wilson on ideals in the Weyl algebra and the Calogero–Moser model.

Further Developments

Bellissard and others have extended the use of noncommutative geometrical methods into other parts of solid-state theory, where they clarify a number of the physical ideas. This is particularly useful in the case of quasicrystals, which are not easily handled by the conventional methods (Bellissard *et al.* 2000). Some ideas in string theory resemble higher-dimensional analogs, and higher-dimensional versions of the quantum Hall effect have also been studied by Hu and Zhang.

Finally, we conclude with some mathematical extensions of the theory. We have seen that, for periodic systems, the noncommutative Brillouin

zone can be a noncommutative torus, and it is possible to consider noncommutative versions of Riemann surfaces of higher genera. Carey *et al.* (1998) studied the effect in a noncommutative hyperbolic geometry with a discrete group action, generalizing the action of a Fuchsian group on the unit disc. This provides a tractable example in which one has an edge (albeit rather different from the normal physical situations) and also examples of a Hall effect in higher-genus noncommutative Riemann surfaces closely related to those of Klimek and Lesznewski. Natsumé and Nest have subsequently shown that these are deformation quantizations of the commutative Riemann surface theory in the sense of Rieffel. Coverings of noncommutative Riemann surfaces, which might provide an analoge of composite fermions, have been investigated by Marcolli and Mathai (1999, 2001).

See also: C^*-Algebras and Their Classification; Chern–Simons Models: Rigorous Results; Fractional Quantum Hall Effect; Hopf Algebras and *q*-Deformation Quantum Groups; Localization for Quasiperiodic Potentials; Noncommutative Geometry and the Standard Model; Noncommutative Tori, Yang–Mills, and String Theory; Schrödinger Operators.

Further Reading

Bellissard J (1986) *K*-theory of C^*-algebras, in solid state physics. In: Dorlas T, Hugenholtz NM, and Winnink M (eds.) *Statistical Mechanics and Field Theory: Mathematical Aspects*, Springer Lecture Notes in Physics vol. 257, pp. 99–156. Berlin: Springer.

Bellissard J, Herrmann DJL, and Zarrouati M (2000) Hulls of a periodic solids and gap labelling theorem. In: Baake M and Moody RV (eds.) *Directions in Mathematical Quasi-Crystals*, CRM Monograph Series vol. 13, pp. 207–258. Providence, RI: American Mathematical Society.

Bellissard J, van Elst A, and Schulz-Baldes H (1994) The noncommutative geometry of the quantum Hall effect. *Journal of Mathematical Physics* 35: 5373–5457.

Carey AL, Hannabuss KC, Mathai V, and McCann P (1998) Quantum Hall effect on the hyperbolic plane. *Communications in Mathematical Physics* 190: 629–673.

Connes A (1986) Non-commutative differential geometry. *Publications of the Institut des Hautes Etudes Scientifiques* 62: 257–360.

Connes A (1994) *Non Commutative Geometry*. San Diego: Academic Press.

Fröhlich J (1995) *The Fractional Quantum Hall Effect, Chern–Simons Theory and Integral Lattices*. Proceedings of the International Congress of Mathematicians, 1994, Zürich, 75–105. Basel: Birkhäuser Verlag.

Fröhlich J and Studer UM (1993) Gauge invariance and current algebra in nonrelativistic many-body theory. *Reviews of Modern Physics* 65: 733–802.

Kellendonk J, Richter T, and Schulz-Baldes M (2002) Edge current channels and Chern numbers in the integral quantum Hall effect. *Reviews in Mathematical Physics* 14: 87–119.

Marcolli M and Mathai V (1999) Twisted index theory on good orbifolds. I. Non-commutative Bloch theory. *Communications in Contemporary Mathematics* 1: 553–587.

Marcolli M and Mathai V (2001) Twisted index theory on good orbifolds. II. Fractional quantum numbers. *Communications in Mathematical Physics* 217: 55–87.

McCann PJ (1998) Geometry and the integer quantum Hall effect. In: Carey AL and Murray MK (eds.) *Geometric Analysis and Lie Theory in Mathematics and Physics*, pp. 122–208. Australian Mathematical Society Lecture Series. Cambridge: Cambridge University Press.

Polychronakos A (2001) Quantum Hall states as matrix Chern–Simons theory. *The Journal of High Energy Physics* 4(paper 11): 20.

Schulz-Baldes M, Kellendonk J, and Richter T (2000) Simultaneous quantization of edge and bulk Hall conductivity. *Journal of Physics A* 33: L27–L32.

Quantum Mechanical Scattering Theory

D R Yafaev, Université de Rennes, Rennes, France

Introduction

Scattering theory is concerned with the study of the large-time behavior of solutions of the time-dependent Schrödinger equation [1] for a system with a Hamiltonian H:

$$i\partial u/\partial t = Hu, \quad u(0) = f \qquad [1]$$

Being a part of the perturbation theory, scattering theory describes the asymptotics of $u(t)$ as $t \to +\infty$ or $t \to -\infty$ in terms of solutions of the Schrödinger equation for a "free" system with a Hamiltonian H_0. Of course, eqn [1] has a unique solution $u(t) = \exp(-iHt)f$, while the solution of the same equation with the operator H_0 and the initial data $u_0(0) = f_0$ is given by the formula $u_0(t) = \exp(-iH_0t)f_0$. From the viewpoint of scattering theory, the function $u(t)$ has free asymptotics as $t \to \pm\infty$ if for appropriate initial data $f_0^{\pm}$ eqn [2] holds:

$$\lim_{t\to\pm\infty} \|u(t) - u_0^{\pm}(t)\| = 0 \qquad [2]$$

Here and throughout this article a relation containing the signs "$\pm$" is understood as two independent equalities. We emphasize that initial data $f_0^{\pm}$ are different for $t \to +\infty$ and $t \to -\infty$ and

$u_0^{\pm}(t) = \exp(-iH_0t)f_0^{\pm}$. Equation [2] leads to a connection between the corresponding initial data $f_0^{\pm}$ and f given by

$$f = \lim_{t\to\pm\infty} \exp(iHt)\exp(-iH_0t)f_0^{\pm} \qquad [3]$$

If f is an eigenvector of H, that is, $Hf = \lambda f$, then obviously $u(t) = e^{-i\lambda t}f$. On the contrary, if f belongs to the (absolutely) continuous subspace of H, then necessarily $u(t)$ has the free asymptotics as $t \to \pm\infty$. This result is known as asymptotic completeness.

The Schrödinger operator $H = -\Delta + V(x)$ in the space $\mathcal{H} = L_2(\mathbb{R}^d)$ with a real potential V decaying at infinity is a typical Hamiltonian of scattering theory. The operator H describes a particle in an external potential V or two interacting particles. Asymptotically (as $t \to +\infty$ or $t \to -\infty$), particles may either form a bound state or be free (a scattering state). Of course, a bound (scattering) state at $-\infty$ remains the same at $+\infty$. To be more precise, suppose that

$$|V(x)| \le C(1 + |x|)^{-\rho} \qquad [4]$$

where $\rho > 1$. Then relation [2] can be justified with the kinetic energy operator $H_0 = -\Delta$ playing the role of the unperturbed operator.

As discussed in Landau and Lifshitz (1965) (see also Amrein *et al.* (1977), Pearson (1988), and Yafaev (2000)), in scattering experiments one sends a beam of particles of energy $\lambda > 0$ in a direction ω. Such a beam is described by the plane wave

$$\psi_0(x;\omega,\lambda) = \exp(ik\langle\omega, x\rangle), \quad \lambda = k^2 > 0$$

(which satisfies of course the free equation $-\Delta\psi_0 = \lambda\psi_0$). The scattered particles are described for large distances by the outgoing spherical wave

$$a(\hat{x},\omega;\lambda)|x|^{-(d-1)/2}\exp(ik|x|)$$

Here $\hat{x} = x|x|^{-1}$ is the direction of observation and the coefficient $a(\hat{x},\omega;\lambda)$ is known as the scattering amplitude. This means that quantum particles subject to a potential $V(x)$ are described by the solution ψ of eqn [5] with asymptotics [6] at infinity:

$$-\Delta\psi + V(x)\psi = \lambda\psi \qquad [5]$$

$$\begin{aligned}\psi(x;\omega,\lambda) = {} & \exp(ik\langle\omega, x\rangle) \\ & + a(\hat{x},\omega;\lambda)|x|^{-(d-1)/2}\exp(ik|x|) \\ & + o\left(|x|^{-(d-1)/2}\right)\end{aligned} \qquad [6]$$

The existence of such solutions requires of course a proof. The differential scattering crosssection defined by eqn [7] gives us the part of particles scattered in a solid angle $d\hat{x}$:

$$d\sigma(\hat{x},\omega;\lambda) = |a(\hat{x},\omega;\lambda)|^2 d\hat{x} \qquad [7]$$

As discussed below, the temporal asymptotics of solutions of the time-dependent Schrödinger equation [1] are closely related to the asymptotics at large distances of solutions of the stationary Schrödinger equation [5].

Time-Dependent Scattering Theory and Møller Operators

If $V(x) \to 0$ as $|x| \to \infty$, then the essential spectrum of the Schrödinger operator $H = -\Delta + V(x)$ covers the whole positive half-line, whereas the negative spectrum of H consists of eigenvalues accumulating, perhaps, at the point zero only.

Scattering theory requires a more advanced classification of the spectrum based on measure theory. Consider a self-adjoint operator H defined on domain $\mathcal{D}(H)$ in a Hilbert space $\mathcal{H}$. Let E be its spectral family. Then the space $\mathcal{H}$ can be decomposed into the orthogonal sum of invariant subspaces $\mathcal{H}^{(p)}, \mathcal{H}^{(sc)}$ and $\mathcal{H}^{(ac)}$. The subspace $\mathcal{H}^{(p)}$ is spanned by eigenvectors of H and the subspaces $\mathcal{H}^{(sc)}, \mathcal{H}^{(ac)}$ are distinguished by the condition that the measure $(E(X)f, f)$ (here $X \subset \mathbb{R}$ is a Borel set) is singularly or absolutely continuous with respect to the Lebesgue measure for all $f \in \mathcal{H}^{(sc)}$ or $f \in \mathcal{H}^{(ac)}$. Typically (in applications to quantum-mechanical problems) the singularly continuous part is absent, that is, $\mathcal{H}^{(sc)} = \{0\}$. We denote by $H^{(ac)}$ the restriction of H on its absolutely continuous subspace $\mathcal{H}^{(ac)}$ and by $P^{(ac)}$ the orthogonal projection on this subspace. The same objects for the operator H_0 will be endowed with the index "0."

Equation [3] motivates the following fundamental definition. The wave, or Møller, operator $W_{\pm} = W_{\pm}(H, H_0)$ for a pair of self-adjoint operators H_0 and H is defined by eqn [8] provided that the corresponding strong limit exists:

$$W_{\pm} = \operatorname*{s-lim}_{t\to\pm\infty} \exp(iHt)\exp(-iH_0t)P_0^{(ac)} \qquad [8]$$

The wave operator is isometric on $\mathcal{H}_0^{(ac)}$ and enjoys the intertwining property

$$HW_{\pm} = W_{\pm}H_0 \qquad [9]$$

Therefore, its range $\operatorname{Ran} W_{\pm}$ is contained in the absolutely continuous subspace $\mathcal{H}^{(ac)}$ of the operator H.

The operator $W_{\pm}(H, H_0)$ is said to be complete if eqn [10] holds:

$$\operatorname{Ran} W_{\pm}(H, H_0) = \mathcal{H}^{(ac)} \qquad [10]$$

It is easy to see that the completeness of $W_\pm(H,H_0)$ is equivalent to the existence of the "inverse" wave operator $W_\pm(H_0,H)$. Thus, if the wave operator $W_\pm(H,H_0)$ exists and is complete, then the operators $H_0^{(\mathrm{ac})}$ and $H^{(\mathrm{ac})}$ are unitarily equivalent. We emphasize that scattering theory studies not arbitrary unitary equivalence but only the "canonical" one realized by the wave operators.

Along with the wave operators an important role in scattering theory is played by the scattering operator defined by eqn [11] where W_+^* is the operator adjoint to W_+:

$$S = S(H,H_0) = W_+^*(H,H_0)W_-(H,H_0) \qquad [11]$$

The operator S commutes with H_0 and hence reduces to multiplication by the operator function $S(\lambda)=S(\lambda;H,H_0)$ in a representation of $\mathcal{H}_0^{(\mathrm{ac})}$ which is diagonal for $H_0^{(\mathrm{ac})}$. The operator $S(\lambda)$ is known as the scattering matrix. The scattering operator [11] is unitary on the subspace $\mathcal{H}_0^{(\mathrm{ac})}$ provided the wave operators $W_\pm(H,H_0)$ exist and are complete. The scattering operator $S(H,H_0)$ connects the asymptotics of the solutions of eqn [1] as $t\to-\infty$ and as $t\to+\infty$ in terms of the free problem, that is $S(H,H_0):f_0^-\mapsto f_0^+$, where $f_0^\pm$ are the same as in eqn [2]. The scattering operator and the scattering matrix are usually of great interest in mathematical physics problems, because they connect the "initial" and the "final" characteristics of the process directly, bypassing its consideration for finite times.

The definition of the wave operators can be extended to self-adjoint operators acting in different spaces. Let H_0 and H be self-adjoint operators in Hilbert spaces $\mathcal{H}_0$ and $\mathcal{H}$, respectively, and let "identification" $J:\mathcal{H}_0\to\mathcal{H}$ be a bounded operator. Then the wave operator $W_\pm=W_\pm(H,H_0;J)$ for the triple H_0,H, and J is defined by eqn [12] provided again that the strong limit there exists:

$$W_\pm = \operatorname*{s-lim}_{t\to\pm\infty}\exp(\mathrm{i}Ht)J\exp(-\mathrm{i}H_0t)P_0^{(\mathrm{ac})} \qquad [12]$$

Intertwining property [9] is preserved for wave operator [12]. This operator is isometric on $\mathcal{H}_0^{(\mathrm{ac})}$ if and only if

$$\lim_{t\to\pm\infty}\|J\exp(-\mathrm{i}H_0t)f_0\| = \|f_0\|$$

for all $f_0\in\mathcal{H}_0^{(\mathrm{ac})}$. Since

$$\operatorname*{s-lim}_{|t|\to\infty}K\exp(-\mathrm{i}H_0t)P_0^{(\mathrm{ac})} = 0$$

for a compact operator K, wave operators [12] corresponding to identifications J_1 and J_2 coincide if J_2-J_1 is compact or, at least, the operators $(J_2-J_1)E_0(X)$ are compact for all bounded intervals X. Consideration of wave operators [12] with $J\neq I$ may of course be of interest also in the case $\mathcal{H}_0=\mathcal{H}$.

It suffices to verify the existence of limits [8] or [12] on some set dense in the absolutely continuous subspace $\mathcal{H}_0^{(\mathrm{ac})}$ of the operator H_0. The following simple but convenient condition for the existence of wave operators is usually called Cook's criterion. Suppose that $H_0=H_0^{(\mathrm{ac})}$ and that the operator J maps domain $\mathcal{D}(H_0)$ of the operator H_0 into $\mathcal{D}(H)$. Let

$$\int_0^{\pm\infty}\|(HJ-JH_0)\exp(-\mathrm{i}H_0t)f\|\mathrm{d}t<\infty$$

for all f from some set $D_0\subset\mathcal{D}(H_0)$ dense in $\mathcal{H}_0$. Then the wave operator $W_\pm(H,H_0;J)$ exists.

This result is often useful in applications since the operator $\exp(-\mathrm{i}H_0t)$ is known explicitly. For example, it works with $J=I$ for the pair

$$H_0=-\Delta, \qquad H=H_0+V(x) \qquad [13]$$

if $V(x)$ satisfies estimate [4] with $\rho>1$. On the other hand, different proofs of the existence of the wave operators $W_\pm(H_0,H;J^*)$ require new mathematical tools. There are two essentially different approaches in scattering theory: the trace-class and smooth methods.

Time-Independent Scattering Theory

The approach in scattering theory relying on definition [8] is called time dependent. An alternative possibility is to change the definition of wave operators replacing the unitary groups by the corresponding resolvents $R_0(z)=(H_0-z)^{-1}$ and $R(z)=(H-z)^{-1}$. They are related by a simple identity

$$\begin{aligned}R(z) &= R_0(z)-R_0(z)VR(z)\\ &= R_0(z)-R(z)VR_0(z)\end{aligned} \qquad [14]$$

where $V=H-H_0$ and $\operatorname{Im}z\neq0$. In the stationary approach in place of limits [8] one has to study the boundary values (in a suitable topology) of the resolvents as the spectral parameter z tends to the real axis. An important advantage of the stationary approach is that it gives convenient formulas for the wave operators and the scattering matrix.

Let us discuss here the stationary formulation of the scattering problem for operators [13] in the Hilbert space $\mathcal{H}=L_2(\mathbb{R}^d)$ in terms of solutions of the Schrödinger equation [5]. If $V(x)$ satisfies estimate [4] with $\rho>(d+1)/2$, then for all $\lambda>0$ and all unit vectors $\omega\in\mathbb{S}^{d-1}$, eqn [5] has the solution $\psi(x;\omega,\lambda)$ with asymptotics [6] as $|x|\to\infty$. Moreover, the scattering amplitude $a(\hat{x},\omega;\lambda)$ belongs to the space

$L_2(\mathbb{S}^{d-1})$ in the variable $\hat{x}$ uniformly in $\omega \in \mathbb{S}^{d-1}$, and it can be expressed via $\psi(x;\omega,\lambda)$ by the formula

$$a(\theta,\omega;\lambda) = -\gamma_d(\lambda)\int_{\mathbb{R}^d} e^{-ik\langle\theta,x\rangle} V(x)\psi(x;\omega,\lambda)\,dx$$

where

$$\gamma_d(\lambda) = e^{-\pi i(d-3)/4}2^{-1}(2\pi)^{-(d-1)/2}\lambda^{(d-3)/4}$$

Let us define two sets of scattering solutions, or eigenfunctions of the continuous spectrum, by the formulas

$$\psi_-(x;\omega,\lambda) = \overline{\psi(x;\omega,\lambda)} \text{ and } \psi_+(x;\omega,\lambda) = \overline{\psi(x;-\omega,\lambda)}$$

In terms of boundary values of the resolvent, the functions $\psi_\pm(\omega,\lambda)$ can be constructed by the formula

$$\psi_\pm(\omega,\lambda) = \psi_0(\omega,\lambda) - R(\lambda \mp i0)V\psi_0(\omega,\lambda) \qquad [15]$$

Obviously, functions [15] satisfy eqn [5]. Using resolvent identity [14], it is easy to derive the Lippmann–Schwinger equation

$$\psi_\pm(\omega,\lambda) = \psi_0(\omega,\lambda) - R_0(\lambda \mp i0)V\psi_\pm(\omega,\lambda)$$

for $\psi_\pm(\omega,\lambda)$. Asymptotics [6] can be deduced from the formula

$$(R_0(\lambda \pm i0)f)(x) = c_\pm(\lambda)(\Gamma_0(\lambda)f)(\pm\hat{x})|x|^{-(d-1)/2} \times \exp(\pm ik|x|) + O(|x|^{-(d+1)/2})$$

where $f \in C_0^\infty(\mathbb{R}^d)$, $c_\pm(\lambda) = \pi^{1/2}\lambda^{-1/4}e^{\mp i\pi(d-3)/4}$ and the operator $\Gamma_0(\lambda)$ defined by eqn [16] is (up to the numerical factor) the restriction of the Fourier transform $\hat{f} = \mathcal{F}_0 f$ onto the sphere of radius $\lambda^{1/2}$:

$$(\Gamma_0(\lambda)f)(\omega) = 2^{-1/2}\lambda^{(d-2)/4}\hat{f}(\lambda^{1/2}\omega), \quad \omega \in \mathbb{S}^{d-1} \qquad [16]$$

The wave operators $W_\pm(H,H_0)$ can be constructed in terms of the solutions $\psi_\pm$. Set $\xi = \lambda^{1/2}\omega$ (ξ is the momentum variable), write $\psi_\pm(x,\xi)$ instead of $\psi_\pm(x;\omega,\lambda)$, and consider two transformations

$$(\mathcal{F}_\pm f)(\xi) = (2\pi)^{-d/2}\int_{\mathbb{R}^d} \overline{\psi_\pm(x,\xi)}\, f(x)\,dx \qquad [17]$$

(defined initially, e.g., on the Schwartz class $\mathcal{S}(\mathbb{R}^d)$) of the space $L_2(\mathbb{R}^d)$ into itself. The operators $\mathcal{F}_\pm$ can be regarded as generalized Fourier transforms, and both of them coincide with the usual Fourier transform $\mathcal{F}_0$ if $V = 0$. It follows from eqns [5], [17] that under the action of $\mathcal{F}_\pm$ the operator H goes over into multiplication by $|\xi|^2$, that is,

$$(\mathcal{F}_\pm H f)(\xi) = |\xi|^2(\mathcal{F}_\pm f)(\xi)$$

Moreover, with the help of eqn [15], it can be shown that $\mathcal{F}_\pm$ is an isometry on $\mathcal{H}^{(ac)}$, it is zero on $\mathcal{H} \ominus \mathcal{H}^{(ac)}$, and its range Ran $\mathcal{F}_\pm = L_2(\mathbb{R}^d)$. This is equivalent to eqns [18]:

$$\mathcal{F}_\pm^*\mathcal{F}_\pm = P^{(ac)}, \qquad \mathcal{F}_\pm\mathcal{F}_\pm^* = I \qquad [18]$$

Hence any function $f \in \mathcal{H}^{(ac)}$ admits the expansion in the generalized Fourier integral

$$f(x) = (2\pi)^{-d/2}\int_{\mathbb{R}^d} \psi_\pm(x,\xi)(\mathcal{F}_\pm f)(\xi)\,d\xi$$

It can also be deduced from eqn [6] that the vector

$$(\mathcal{F}_\pm^* - \mathcal{F}_0^*)\exp(-i|\xi|^2 t)\hat{f}$$

tends to zero as $t \to \pm\infty$ for an arbitrary $\hat{f} \in L_2(\mathbb{R}^d)$. This implies the existence of the wave operators $W_\pm = W_\pm(H,H_0)$ for pair [13] and gives the representation

$$W_\pm = \mathcal{F}_\pm^*\mathcal{F}_0 \qquad [19]$$

Completeness of $W_\pm$ follows from eqn [19] and the first equation in [18]. The second equality in [18] is equivalent to the isometricity of $W_\pm$. Formula [19] is an example of a stationary representation for the wave operator. It formally implies that

$$W_\pm\psi_0(\omega,\lambda) = \psi_\pm(\omega,\lambda)$$

which means that each wave operator establishes a one-to-one correspondence between eigenfunctions of the continuous spectrum of the operators H_0 and H.

The main ideas of the stationary approach go back to Friedrichs (1965), and Povzner. The inverse problem of reconstruction of a potential V given the scattering amplitude a (see eqn [6]) is treated in Faddeev (1976).

The Trace-Class Method

Recall that the class $\mathfrak{S}_p, p \geq 1$, consists of compact operators T such that the norm

$$\|T\|_p = \left(\sum_n \lambda_n^p(|T|)\right)^{1/p}, \quad |T| = (T^*T)^{1/2}$$

is finite. Eigenvalues $\lambda_n(|T|) =: s_n(T)$ of a non-negative operator $|T|$ are called singular numbers of T. In particular, $\mathfrak{S}_1$ is the trace class and $\mathfrak{S}_2$ is the Hilbert–Schmidt class.

The trace-class method (see Reed and Simon (1976) or Yafaev (1992) for a detailed presentation) makes no assumptions about the "unperturbed" operator H_0. Its basic result is the following theorem of Kato and Rosenblum. If $V = H - H_0$ belongs to the trace class

$\mathfrak{S}_1$, then the wave operators $W_\pm(H, H_0)$ exist and are complete. In particular, the operators $H_0^{(\mathrm{ac})}$ and $H^{(\mathrm{ac})}$ are unitarily equivalent. This can be considered as a far advanced extension of the H Weyl theorem, which states the stability of the essential spectrum under compact perturbations.

The condition $V \in \mathfrak{S}_1$ in the Kato–Rosenblum theorem cannot be relaxed in the framework of operator ideals $\mathfrak{S}_p$. This follows from the Weyl–von Neumann–Kuroda theorem. Let H_0 be an arbitrary self-adjoint operator. For any $p > 1$ and any $\varepsilon > 0$ there exists a self-adjoint operator V such that $V \in \mathfrak{S}_p$, $\|V\|_p < \varepsilon$ and the operator $H = H_0 + V$ has purely point spectrum. Of course, such an operator H has no absolutely continuous part. At the same time, the operator H_0 may be absolutely continuous. In this case, the wave operators $W_\pm(H, H_0)$ do not exist.

Although sharp in the abstract framework, the Kato–Rosenblum theorem cannot directly be applied to the theory of differential operators where a perturbation is usually an operator of multiplication and hence is not even compact. We mention its two generalizations applicable to this theory. The first, the Birman–Kato–Kreĭn theorem, claims that the wave operators $W_\pm(H, H_0)$ exist and are complete provided

$$R^n(z) - R_0^n(z) \in \mathfrak{S}_1$$

for some $n = 1, 2, \ldots$ and all z with $\operatorname{Im} z \neq 0$. The second, the Birman theorem, asserts that the same is true if $\mathcal{D}(H) = \mathcal{D}(H_0)$ or $\mathcal{D}(|H|^{1/2}) = \mathcal{D}(|H_0|^{1/2})$ and

$$E(X)(H - H_0)E_0(X) \in \mathfrak{S}_1$$

for all bounded intervals X.

The wave operators enjoy the following property known as the Birman invariance principle. Suppose that $\varphi(H) - \varphi(H_0) \in \mathfrak{S}_1$ for a real function φ such that its derivative φ' is absolutely continuous and $\varphi'(\lambda) > 0$. Then the wave operators $W_\pm(H, H_0)$ exist and eqn [20] holds:

$$W_\pm(H, H_0) = W_\pm(\varphi(H), \varphi(H_0)) \qquad [20]$$

A direct generalization of the Kato–Rosenblum theorem to the operators acting in different spaces is due to Pearson. Suppose that H_0 and H are self-adjoint operators in spaces $\mathcal{H}_0$ and $\mathcal{H}$, respectively, $J : \mathcal{H}_0 \to \mathcal{H}$ is a bounded operator and $V = HJ - JH_0 \in \mathfrak{S}_1$. Then the wave operators $W_\pm(H, H_0; J)$ and $W_\pm(H_0, H; J^*)$ exist.

Although rather sophisticated, the proof relies only on the following elementary lemma of Rosenblum. For a self-adjoint operator H, consider the set $\mathfrak{R} \subset \mathcal{H}^{(\mathrm{ac})}$ of elements f such that

$$r_H^2(f) := \operatorname{ess\,sup} \mathrm{d}(E(\lambda)f, f)/\mathrm{d}\lambda < \infty$$

If $K : \mathcal{H} \to \mathcal{G}$ ($\mathcal{G}$ is some Hilbert space) is a Hilbert–Schmidt operator, then for all $f \in \mathfrak{R}$

$$\int_{-\infty}^{\infty} \|K \exp(-\mathrm{i}Ht)f\|^2 \, \mathrm{d}t \leq 2\pi r_H^2(f)\|K\|_2^2 \qquad [21]$$

Moreover, the set $\mathfrak{R}$ is dense in $\mathcal{H}^{(\mathrm{ac})}$.

The Pearson theorem allows to simplify considerably the original proofs of different generalizations of the Kato–Rosenblum theorem.

A typical application of the trace-class theory is the following result. Suppose that

$$\mathcal{H} = L_2(\mathbb{R}^d),\ H_0 = -\Delta + V_0(x),\ H = H_0 + V(x) \qquad [22]$$

where the functions V_0 and V are real, $V_0 \in L_\infty(\mathbb{R}^d)$ and V satisfies estimate [4] for some $\rho > d$. Then the wave operators $W_\pm(H, H_0)$ exist and are complete.

The Smooth Method

The smooth method (see Kuroda (1978), Reed and Simon (1979), or Yafaev (1992), for a detailed presentation) relies on a certain regularity of the perturbation in the spectral representation of the operator H_0. There are different ways to understand regularity. For example, in the Friedrichs–Faddeev model H_0 acts as multiplication by independent variable in the space $\mathcal{H} = L_2(\Lambda; \mathfrak{N})$, where Λ is an interval and $\mathfrak{N}$ is an auxiliary Hilbert space. The perturbation V is an integral operator with sufficiently smooth kernel.

Another possibility is to use the concept of H-smoothness introduced by Kato. An H-bounded operator K is called H-smooth if, for all $f \in \mathcal{D}(H)$,

$$\int_{-\infty}^{\infty} \|K \exp(-\mathrm{i}Ht)f\|^2 \, \mathrm{d}t \leq C\|f\|^2 \qquad [23]$$

(cf. eqns [21] and [23]). Here and below, C are different positive numbers whose precise values are inessential. It is important that this definition admits equivalent reformulations in terms of the resolvent or of the spectral family. Thus, K is H-smooth if and only if

$$\sup_{\lambda \in \mathbb{R}, \varepsilon > 0} \|K(R(\lambda + \mathrm{i}\varepsilon) - R(\lambda - \mathrm{i}\varepsilon))K^*\| < \infty$$

or if and only if

$$\sup |X|^{-1} \|KE(X)\|^2 < \infty$$

for all intervals $X \subset \mathbb{R}$.

In applications the assumption of H-smoothness of an operator K imposes too stringent conditions on the operator H. In particular, the operator H is necessarily absolutely continuous if kernel of K is trivial. This assumption excludes eigenvalues and other singular points in the spectrum of H, for

example, the bottom of the continuous spectrum for the Schrödinger operator with decaying potential or edges of bands if the spectrum has the band structure. The notion of local H-smoothness suggested by Lavine is considerably more flexible. By definition, K is called H-smooth on a Borel set $X \subset \mathbb{R}$ if the operator $KE(X)$ is H-smooth. Note that, under the assumption

$$\sup_{\lambda \in X, \varepsilon > 0} \|K(R(\lambda + \mathrm{i}\varepsilon) - R(\lambda - \mathrm{i}\varepsilon))K^*\| < \infty \qquad [24]$$

the operator K is H-smooth on the closure of X.

The following Kato–Lavine theorem is simple but very useful. Suppose that

$$HJ - JH_0 = K^*K_0$$

where the operators K_0 and K are H_0-smooth and H-smooth, respectively, on an arbitrary compact subinterval of some interval Λ. Then the wave operators

$$W_\pm(H, H_0; JE_0(\Lambda)) \text{ and } W_\pm(H_0, H; J^*E(\Lambda))$$

exist (and are adjoint to each other).

This result cannot usually be applied directly since the verification of H_0- and especially of H-smoothness may be a difficult problem. Let us briefly explain how it can be done on the example of pair [10], where the potential $V(x)$ satisfies estimate [4] for some $\rho > 1$. Let us start with the operator $H_0 = -\Delta$. Denote by $L_2^{(l)} = L_2^{(l)}(\mathbb{R}^d)$ the Hilbert space with the norm $\|f\|_l = \|\langle x \rangle^l f\|$, where $\langle x \rangle = (1 + |x|^2)^{1/2}$. Let the operator $\Gamma_0(\lambda)$ be defined by eqn [16], and let $X \subset (0, \infty)$ be some compact interval. Set $\mathfrak{N} = L_2(\mathbb{S}^{d-1})$. If $f \in L_2^{(l)}$ with $l > 1/2$, then, by the Sobolev trace theorem,

$$\begin{aligned} \|\Gamma_0(\lambda)f\|_{\mathfrak{N}} &\le C\|f\|_l \\ \|\Gamma_0(\lambda)f - \Gamma_0(\lambda')f\|_{\mathfrak{N}} &\le C|\lambda - \lambda'|^\alpha \|f\|_l \end{aligned} \qquad [25]$$

for an arbitrary $\alpha \le l - 1/2, \alpha < 1$ and all $\lambda, \lambda' \in X$. These estimates imply that the function

$$(E_0(\lambda)f, f) = \int_{|\xi|^2 < \lambda} |\hat{f}(\xi)|^2 \, \mathrm{d}\xi \qquad [26]$$

is differentiable and the derivative

$$\mathrm{d}(E_0(\lambda)f, f)/\mathrm{d}\lambda = \|\Gamma_0(\lambda)f\|_{\mathfrak{N}}^2, \quad f \in L_2^{(l)}, \ l > 1/2$$

is Hölder-continuous in $\lambda > 0$ (uniformly in f, $\|f\|_l \le 1$). Therefore, applying the Privalov theorem to the Cauchy integral

$$(R_0(z)f, f) = \int_0^\infty (\lambda - z)^{-1} \, \mathrm{d}(E_0(\lambda)f, f)$$

we obtain that the analytic operator function

$$\mathcal{R}_0(z) = \langle x \rangle^{-l} R_0(z) \langle x \rangle^{-l}, \quad l > 1/2$$

considered in the space $\mathcal{H}$, is continuous in norm in the closed complex plane $\mathbb{C}$ cut along $(0, \infty)$ with possible exception of the point $z = 0$. This implies H_0-smoothness of the operator $\langle x \rangle^{-l}, l > 1/2$, on all compact intervals $X \subset (0, \infty)$.

To obtain a similar result for the operator H, we proceed from the resolvent identity [14]. Let $\mathcal{R}(z) = \langle x \rangle^{-l} R(z) \langle x \rangle^{-l}$, and let B be the operator of multiplication by the bounded function $(1 + |x|)^\rho V(x)$. If

$$f + \mathcal{R}_0(z)Bf = 0$$

then $\psi = R_0(z)\langle x \rangle^{-l} Bf$ satisfies the Schrödinger equation $H\psi = z\psi$. Since H is self-adjoint, this implies that $\psi = 0$ and hence $f = 0$. Using eqn [14], we obtain that

$$\mathcal{R}(z) = (I + \mathcal{R}_0(z)B)^{-1}\mathcal{R}_0(z), \quad \mathrm{Im}\, z \ne 0 \qquad [27]$$

because the inverse operator here exists by the Fredholm alternative.

The operator function $(I + \mathcal{R}_0(z)B)^{-1}$ is analytic in the complex plane cut along $(0, \infty)$ with possible exception of poles (coinciding with eigenvalues of H) on the negative half-axis. Moreover, $(I + \mathcal{R}_0(z)B)^{-1}$ is continuous up to the cut except the set $\mathcal{N} \subset (0, \infty)$ of λ where at least one of the homogeneous equations

$$f + \mathcal{R}_0(\lambda \pm \mathrm{i}0)Bf = 0 \qquad [28]$$

has a nontrivial solution. It follows from eqn [27] that the same is true for the operator function $\mathcal{R}(z)$. It can be shown that the set $\mathcal{N}$ is closed and has the Lebesgue measure zero. Let $\Lambda = (0, \infty) \setminus \mathcal{N}$; then $\Lambda = \cup_n \Lambda_n$ where Λ_n are disjoint open intervals. By condition [24], the operator $\langle x \rangle^{-l}, l > 1/2$, is H-smooth on any strictly interior subinterval of every Λ_n. Applying the Kato–Lavine theorem, we see that the wave operators $W_\pm(H, H_0; E_0(\Lambda_n))$ and $W_\pm(H_0, H; E(\Lambda_n))$ exist for all n. Since $E_0(\Lambda) = I$ and $E(\Lambda) = P^{(\mathrm{ac})}$, this implies the existence of $W_\pm(H, H_0)$ and $W_\pm(H_0, H)$. Thus, the wave operators $W_\pm(H, H_0)$ for pair [13] exist and are complete if estimate [4] holds for some $\rho > 1$.

Compared to the trace-class method, conditions on the perturbation $V(x)$ are less restrictive, while the class of admissible "free" problems is essentially more narrow (in eqn [22] $V_0(x)$ is an arbitrary bounded function). It is not known whether the wave operators $W_\pm(H, H_0)$ exist for all pairs [22] such that $V_0 \in L_\infty$ and V satisfies [4] for some $\rho > 1$.

It is important that the smooth method allows one to prove the absence of the singular continuous spectrum. Note first that the continuity of $\mathcal{R}(z)$ implies that the operator H is absolutely continuous on the subspace $E(\Lambda)\mathcal{H}$. Therefore, the singular

positive spectrum of H is necessarily contained in $\mathcal{N}$. To prove that its continuous part is empty, it suffices to check that the set $\mathcal{N}$ consists of eigenvalues of the operator H. In terms of $u = \langle x\rangle^{-l}Bf, l = \rho/2$, eqn [28] can be rewritten as

$$u + VR_0(\lambda \pm \mathrm{i}0)u = 0 \qquad [29]$$

Multiplying this equation by $R_0(\lambda \pm \mathrm{i}0)u$ and taking the imaginary part of the scalar product, we see that

$$\pi\, \mathrm{d}(E_0(\lambda)u, u)/\mathrm{d}\lambda = \mathrm{Im}(R_0(\lambda \pm \mathrm{i}0)u, u) = 0$$

According to eqn [26], this implies that

$$\hat{u}(\xi) = 0 \quad \text{for} \quad |\xi| = \lambda^{1/2} \qquad [30]$$

It follows from eqn [29] that

$$\psi = R_0(\lambda \pm \mathrm{i}0)u \qquad [31]$$

that is, $\hat{\psi}(\xi) = (|\xi|^2 - \lambda \mp \mathrm{i}0)^{-1}\hat{u}(\xi)$, is a formal (because of the singularity of the denominator) solution of Schrödinger equation [5]. Therefore, one needs only to verify that $\psi \in L_2(\mathbb{R}^d)$. Since $u \in L_2^{(l)}$, where $l = \rho/2$, this is a direct consequence of [25] and [30] if $\rho > 2$. In the general case, one uses that under assumption [30] the function $(|\xi|^2 - \lambda)^{-1}\hat{u}(\xi)$ belongs to the space $L_2^{(p)}$ for any $p < l - 1$. By virtue of condition [4] where $\rho > 1$, eqn [29] now shows that actually $u \in L_2^{(p)}$ for any $p < l + \rho - 1$. Repeating these arguments, we obtain, after n steps, that $u \in L_2^{(p)}$ for any $p < l + n(\rho - 1)$. For n large enough, this implies that $u \in L_2^{(p)}$ for $p > 1$, and consequently function [31] belongs to $L_2(\mathbb{R}^d)$.

Similar arguments show that eigenvalues of H have finite multiplicity and do not have positive accumulation points. For the proof of boundedness of the set of eigenvalues, one uses additionally the estimate

$$\|\mathcal{R}_0(\lambda \pm \mathrm{i}0)\| = O(\lambda^{-1/2}), \quad \lambda \to \infty \qquad [32]$$

Actually, according to Kato theorem the Schrödinger operator H does not have positive eigenvalues.

There exists also a purely time-dependent approach, the Enss method (see Perry (1983)), which relies on an advanced study of the free evolution operator $\exp(-\mathrm{i}H_0t)$.

The Scattering Matrix

The operator $H_0 = -\Delta$ can of course be diagonalized by the classical Fourier transform. To put it slightly differently, set

$$(F_0f)(\lambda) = \Gamma_0(\lambda)f$$

where the operator $\Gamma_0(\lambda)$ is defined by eqn [16]. Then

$$F_0 : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}_+; \mathfrak{N}), \quad \mathfrak{N} = L_2(\mathbb{S}^{d-1})$$

is a unitary operator and $(F_0H_0f)(\lambda) = \lambda(F_0f)(\lambda)$.

Under assumption [4] where $\rho > 1$, the scattering operator S for pair [13] is defined by eqn [11]. It is unitary on the space $\mathcal{H} = L_2(\mathbb{R}^d)$ and commutes with the operator H_0. It follows that $(F_0Sf)(\lambda) = S(\lambda)(F_0f)(\lambda), \lambda > 0$, where the unitary operator $S(\lambda) : \mathfrak{N} \to \mathfrak{N}$ is known as the scattering matrix. The scattering matrix $S(\lambda)$ for the pair H_0, H can be computed in terms of the scattering amplitude. Namely, $S(\lambda)$ acts in the space $L_2(\mathbb{S}^{d-1})$, and $S(\lambda) - I$ is the integral operator whose kernel is the scattering amplitude. More precisely,

$$(S(\lambda)f)(\theta) = f(\theta) + 2\mathrm{i}\lambda^{1/2}\overline{\gamma_d(\lambda)} \int_{\mathbb{S}^{d-1}} a(\theta, \omega; \lambda)f(\omega)\, \mathrm{d}\omega$$

In operator notation, this representation can be rewritten as

$$S(\lambda) = I - 2\pi\, \mathrm{i}\Gamma_0(\lambda)(V - VR(\lambda + \mathrm{i}0)V)\Gamma_0^*(\lambda) \qquad [33]$$

The right-hand side here is correctly defined as a bounded operator in the space $\mathfrak{N}$ and is continuous in $\lambda > 0$. Moreover, the operator $S(\lambda) - I$ is compact since $\Gamma_0(\lambda)\langle x\rangle^{-l} : \mathcal{H} \to \mathfrak{N}$ is compact for $l > 1/2$ by virtue of the Sobolev trace theorem.

It follows that the spectrum of the operator $S(\lambda)$ consists of eigenvalues of finite multiplicity, except possibly the point 1, lying on the unit circle and accumulating at the point 1 only. In the general case, eigenvalues of $S(\lambda)$ play the role of scattering phases or shifts considered often for radial potentials $V(x) = V(|x|)$.

The scattering amplitude is singular on the diagonal $\theta = \omega$ only. Moreover, this singularity is weaker for potentials with faster decay at infinity (for ρ bigger). If $\rho > (d + 1)/2$, then the operator $S(\lambda) - I$ belongs to the Hilbert–Schmidt class. In this case the total scattering cross section

$$\sigma(\omega; \lambda) = \int_{\mathbb{S}^{d-1}} |a(\theta, \omega; \lambda)|^2\, \mathrm{d}\theta$$

is finite for all energies $\lambda > 0$ and all incident directions $\omega \in \mathbb{S}^{d-1}$. If $\rho > d$, then the operator $S(\lambda) - I$ belongs to the trace class. In this case, the scattering amplitude $a(\theta, \omega; \lambda)$ is a continuous function of $\theta, \omega \in \mathbb{S}^{d-1}$ (and $\lambda > 0$). The unitarity of the operator $S(\lambda)$ implies the optical theorem

$$\sigma(\omega; \lambda) = \lambda^{-1/2}\mathrm{Im}\big(\gamma_d^{-1}(\lambda)a(\omega, \omega; \lambda)\big)$$

Using resolvent identity [14], one deduces from eqn [33] the Born expansion

$$S(\lambda) = I - 2\pi \mathrm{i} \sum_{n=0}^{\infty} (-1)^n \Gamma_0(\lambda) V (R_0(\lambda + \mathrm{i}0) V)^n \Gamma_0^*(\lambda)$$

This series is norm-convergent for small potentials V and according to estimate [32] for high energies λ.

Long-Range Interactions

Potentials decaying at infinity as the Coulomb potential

$$V(x) = \gamma |x|^{-1}, \quad d \geq 3$$

or slower are called long-range. More precisely, it is required that

$$|\partial^\alpha V(x)| \leq C(1 + |x|)^{-\rho - |\alpha|}, \quad \rho \in (0, 1] \qquad [34]$$

for all derivatives of V up to some order. In the long-range case, the wave operators $W_\pm(H, H_0)$ do not exist, and the asymptotic dynamics should be properly modified. It can be done in a time-dependent way either in the coordinate or momentum representations. For example, in the coordinate representation, the free evolution $\exp(-\mathrm{i}H_0 t)$ should be replaced in definition [8] of wave operators by unitary operators $U_0(t)$ defined by

$$(U_0(t) f)(x) = \exp(\mathrm{i}\,\Xi(x, t))(2\mathrm{i}t)^{-d/2} \hat{f}(x/(2t))$$

where $\hat{f}$ is the Fourier transform of f. For short-range potentials we can set $\Xi(x, t) = (4t)^{-1}|x|^2$. In the long-range case the phase function $\Xi(x, t)$ should be chosen as a (perhaps, approximate) solution of the eikonal equation

$$\partial \Xi / \partial t + |\nabla \Xi|^2 + V = 0$$

In particular, we can set

$$\Xi(x, t) = (4t)^{-1}|x|^2 - t \int_0^1 V(sx)\, \mathrm{d}s$$

if $\rho > 1/2$ in [34]. For the Coulomb potential,

$$\Xi(x, t) = (4t)^{-1}|x|^2 - \gamma t |x|^{-1} \ln |t|$$

(the singularity at $x = 0$ is inessential here). Thus, both in short- and long-range cases solutions of the time-dependent Schrödinger equation "live" in a region of the configuration space where $|x|$ is of order $|t|$. Long-range potentials change only asymptotic phases of these solutions.

Another possibility is a time-independent modification in the phase space. Let us consider wave operators $W_\pm(H, H_0; J)$, where J is a pseudodifferential operator,

$$(Jf)(x) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\langle x, \xi \rangle}\, \mathrm{e}^{\mathrm{i}\Phi(x, \xi)} \zeta(x, \xi) \hat{f}(\xi)\, \mathrm{d}\xi$$

with oscillating symbol $\exp(\mathrm{i}\Phi(x, \xi))\zeta(x, \xi)$. Due to the conservation of energy, we may suppose that $\zeta(x, \xi)$ contains a factor $\psi(|\xi|^2)$ with $\psi \in C_0^\infty(0, \infty)$. Set

$$\varphi(x, \xi) = \langle x, \xi \rangle + \Phi(x, \xi)$$

The perturbation $HJ - JH_0$ is also a pseudodifferential operator, and its symbol is short-range (it is $O(|x|^{-1-\varepsilon}), \varepsilon > 0$, as $|x| \to \infty$) if $\exp(\mathrm{i}\varphi(x, \xi))$ is an approximate eigenfunction of the operator H corresponding to the "eigenvalue" $|\xi|^2$. This leads to the eikonal equation

$$|\nabla_x \varphi(x, \xi)|^2 + V(x) = |\xi|^2$$

The notorious difficulty (for $d \geq 2$) of this method is that the eikonal equation does not have (even approximate) solutions such that $|\nabla_x \Phi(x, \xi)| \to 0$ as $|x| \to \infty$ and the arising error term is short-range. However, it is easy to construct functions $\varphi = \varphi_\pm$ satisfying these conditions if a conical neighborhood of the direction $\mp\xi$ is removed from $\mathbb{R}^d$. For example,

$$\Phi_\pm(x, \xi) = \pm 2^{-1} \int_0^\infty (V(x \pm \tau\xi) - V(\pm\tau\xi))\, \mathrm{d}\tau$$

if $\rho > 1/2$ in eqn [34]. Then the cutoff function $\zeta(x, \xi) = \zeta_\pm(x, \xi)$ should be homogeneous of order zero in the variable x and it should be equal to zero in a neighborhood of the direction $\mp\xi$. We emphasize that now we have a couple of different identifications $J = J_\pm$.

The long-range problem is essentially more difficult than the short-range one. The limiting absorption principle remains true in this case, but its proof cannot be performed within perturbation theory. The simplest proof relies on the Mourre estimate (see Cycon *et al.* (1987)) for the commutator $\mathrm{i}[H, A]$ of H with the generator of dilations

$$A = -\mathrm{i} \sum_{j=1}^{d} (x_j \partial_j + \partial_j x_j)$$

The Mourre estimate affirms that, for all $\lambda > 0$,

$$\mathrm{i}E(\Lambda_\lambda)[H, A]E(\Lambda_\lambda) \geq c(\lambda)E(\Lambda_\lambda), \quad c(\lambda) > 0 \qquad [35]$$

if $\Lambda_\lambda = (\lambda - \varepsilon, \lambda + \varepsilon)$ and ε is small enough. For the free operator H_0, this estimate takes the form $\mathrm{i}[H_0, A] = 4H_0$ and can be regarded as a commutation relation. Estimate [35] means that the observable

$$(A\mathrm{e}^{-\mathrm{i}Ht} f,\, \mathrm{e}^{-\mathrm{i}Ht} f)$$

is a strictly increasing function of t for all $f \in \mathcal{H}^{(ac)}$. The H-smoothness of the operator $\langle x\rangle^{-l}, l > 1/2$, is deduced from this fact by some arguments of abstract nature (they do not really use concrete forms of the operators H and A).

However, the limiting absorption principle is not sufficient for construction of scattering theory in the long-range case, and it should be supplemented by an additional estimate. To formulate it, denote by

$$(\nabla^{\perp}u)(x) = (\nabla u)(x) - |x|^{-2}\langle(\nabla u)(x), x\rangle x$$

the orthonal projection of a vector $(\nabla u)(x)$ on the plane orthogonal to x. Then the operator $K = \langle x\rangle^{-1/2}\nabla^{\perp}$ is H-smooth on any compact $X \subset (0,\infty)$. This result is formulated as an estimate (either on the resolvent or on the unitary group of H), which we refer to as the radiation estimate. This estimate is not very astonishing from the viewpoint of analogy with the classical mechanics. Indeed, in the case of free motion, the vector $x(t)$ of the position of a particle is directed asymptotically as its momentum ξ. Regarded as a pseudodifferential operator, $\nabla^{\perp}$ has symbol $\xi - |x|^{-2}\langle\xi, x\rangle x$, which equals zero if $x = \gamma\xi$ for some $\gamma \in \mathbb{R}$. Thus, $\nabla^{\perp}$ removes the part of the phase space where a classical particle propagates. The proof of the radiation estimate is based on the inequality

$$K^*K \le C_0[H, \partial_r] + C_1\langle x\rangle^{-1-\rho}, \quad \partial_r = \partial/\partial|x|$$

which can be obtained by a direct calculation. Since the integral

$$\mathrm{i}\int_0^t ([H, \partial_r]\mathrm{e}^{-\mathrm{i}Hs}f, \mathrm{e}^{-\mathrm{i}Hs}f)\,\mathrm{d}s = (\partial_r \mathrm{e}^{-\mathrm{i}Ht}f, \mathrm{e}^{-\mathrm{i}Ht}f) - (\partial_r f, f)$$

is bounded by $C(X)\|f\|^2$ for $f \in E(X)f$ and the operator $\langle x\rangle^{-(1+\rho)/2}$ is H-smooth on X, this implies H-smoothness of the operator $KE(X)$.

Calculating the perturbation $HJ_{\pm} - J_{\pm}H_0$, we see that it is a sum of two pseudodifferential operators. The first of them is short-range and thus can be taken into account by the limiting absorption principle. The symbol of the second one contains first derivatives (in the variable x) of the cutoff function $\zeta_{\pm}(x,\xi)$ and hence decreases at infinity as $|x|^{-1}$ only. This operator factorizes into a product of H_0- and H-smooth operators according to the radiation estimate. Thus, all wave operators $W_{\pm}(H, H_0; J_{\pm})$ and $W_{\pm}(H_0, H; J_{\pm}^*)$ exist. These operators are isometric since the operators $J_{\pm}$ are in some sense close to unitary operators. The isometricity of $W_{\pm}(H_0, H; J_{\pm}^*)$ is equivalent to the completeness of $W_{\pm}(H, H_0; J_{\pm})$.

Although the modified wave operators enjoy basically the same properties as in the short-range case, properties of the scattering matrices in the short- and long-range cases are drastically different. Here we note only that for long-range potentials, due to a wild diagonal singularity of kernel of the scattering matrix, its spectrum covers the whole unit circle.

Different aspects of long-range scattering are discussed in Dereziński and Gérard (1997), Pearson (1988), Saitō (1979), and Yafaev (2000).

See also: *N*-Particle Quantum Scattering; Quantum Dynamical Semigroups; Random Matrix Theory in Physics; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools; Schrödinger Operators; Spectral Theory for Linear Operators.

Further Reading

Amrein WO, Jauch JM, and Sinha KB (1977) *Scattering Theory in Quantum Mechanics*. New York: Benjamin.

Cycon H, Froese R, Kirsh W, and Simon B (1987) *Schrödinger Operators*, Texts and Monographs in Physics. Berlin: Springer.

Dereziński J and Gérard C (1997) *Scattering Theory of Classical and Quantum N Particle Systems*. Berlin: Springer.

Faddeev LD (1976) Inverse problem of quantum scattering theory II. *J. Soviet. Math.* 5: 334–396.

Friedrichs K (1965) *Perturbation of Spectra in Hilbert Space*. Providence, RI: American Mathematical Society.

Kuroda ST (1978) *An Introduction to Scattering Theory, Lecture Notes Series No. 51*. Aarhus University.

Landau LD and Lifshitz EM (1965) *Quantum Mechanics*. Pergamon.

Pearson D (1988) *Quantum Scattering and Spectral Theory*. London: Academic Press.

Perry PA (1983) *Scattering Theory by the Enss Method*. vol. 1. Harwood: Math. Reports.

Reed M and Simon B (1979) *Methods of Modern Mathematical Physics III*. New York: Academic Press.

Saitō Y (1979) *Spectral Representation for Schrödinger Operators with Long-Range Potentials*. Springer Lecture Notes in Math, vol. 727.

Yafaev DR (1992) *Mathematical Scattering Theory*. Providence, RI: American Mathematical Society.

Yafaev DR (2000) *Scattering Theory: Some Old and New Problems*, Springer Lecture Notes in Math., vol. 1735.

Quantum Mechanics: Foundations

R Penrose, University of Oxford, Oxford, UK

The Framework of Quantum Mechanics

In 1900, Max Planck initiated the quantum revolution by presenting the hypothesis that radiation is emitted or absorbed only in "quanta," each of energy $h\nu$, for frequency ν (where h was a new fundamental constant of Nature). By this device, he explained the precise shape of the puzzling blackbody spectrum. Then, in 1905, Albert Einstein introduced the concept of the photon, according to which light, of frequency ν would, in appropriate circumstances, behave as though it were constituted as individual particles, each of energy $h\nu$, rather than as continuous waves, and he was able to explain the conundrum posed by the photoelectric effect by this means. Later, in 1923, Prince Louis de Broglie proposed that, conversely, all particles behave like waves, the energy being Planck's $\hbar\nu$ and the momentum being $h\lambda^{-1}$, where λ is the wavelength, which was later strikingly confirmed in a famous experiment of Davisson and Germer in 1927. Some years earlier, in 1913, Niels Bohr had used another aspect of this curious quantum "discreteness," explaining the stable electron orbits in hydrogen by the assumption that (orbital) angular momentum must be quantized in units of $\hbar(=h/2\pi)$.

All this provided a very remarkable collection of facts and concepts, albeit somewhat disjointed, explaining a variety of previously baffling physical phenomena, where a certain discreteness seemed to be entering Nature at a fundamental level, where previously there had been continuity, and where there was an overriding theme of a confusion as to whether – or in what circumstances – waves or particles provide better pictures of reality. Moreover, no clear and consistent picture of an actual "quantum-level reality" as yet seemed to arise out of all this. Then, in 1925, Heisenberg introduced his "matrix mechanics," subsequently developed into a more complete theory by Born, Heisenberg and Jordan, and then more fully by Dirac. Some six months after Heisenberg, in 1926, Schrödinger introduced his very different-looking "wave mechanics," which he subsequently showed was equivalent to Heisenberg's scheme. These became encompassed into a comprehensive framework through the transformation theory of Dirac, which he put together in his famous book *The Principles of Quantum Mechanics*, first published in 1930. Later, von Neumann set the framework on a more rigorous basis in his 1932 book, *Mathematische Grundlageen der Quantenmechanik* (later translated as *Mathematical Foundations of Quantum Mechanics*, 1955).

This formalism, now well known to physicists, is based on the presence of a quantum state $|\psi\rangle$ (Dirac's "ket" notation being adopted here). In Schrödinger's description, $|\psi\rangle$ is to evolve by unitary evolution, according to the Schrödinger equation

$$\mathrm{i}\hbar\frac{\partial|\psi\rangle}{\partial t} = \mathrm{H}|\psi\rangle$$

where H is the quantum Hamiltonian. The totality of allowable states $|\psi\rangle$ constitutes a Hilbert space $\mathbf{H}$ and the Schrödinger equation provides a continuous one-parameter family of unitary transformations of $\mathbf{H}$. The letter U is used here for the "quantum-level" evolution whereby the state $|\psi\rangle$ evolves in time according to this unitary Schrödinger evolution. However, we must be careful not to demand an interpretation of this evolution similar to that which we adopt for a classical theory, such as is provided by Maxwell's equations for the electromagnetic field. In Maxwell's theory, the evolution that his equations provide is accepted as very closely mirroring the actual way in which a physically real electromagnetic field evolves with time. In quantum mechanics, however, it is a highly contentious matter how we should regard the "reality" of the unitarily evolving state $|\psi\rangle$.

One of the key difficulties resides in the fact that the world that we actually observe about us rather blatantly does not accord with such a unitarily evolving $|\psi\rangle$. Indeed, the standard way that the quantum formalism is to be interpreted is very far from the mere following of such a picture. So long as no "measurement" is deemed to have been taking place, this U-evolution procedure would be adopted, but upon measurement, the state is taken to behave in a very different way, namely to "jump" instantaneously to some eigenstate $|\phi\rangle$ of the quantum operator Q which is taken to represent the measurement, with probability given by the Born rule

$$|\langle\phi|\psi\rangle|^2$$

if we assume that both $|\psi\rangle$ and $|\phi\rangle$ are normalized ($\langle\psi|\psi\rangle = 1 = \langle\phi|\phi\rangle$); otherwise we can express this probability simply as

$$\frac{\langle\phi|\psi\rangle\langle\psi|\phi\rangle}{\langle\psi|\psi\rangle\langle\phi|\phi\rangle}$$

(The operator Q is normally taken to be self-adjoint, so that $Q = Q^*$ and its eigenvalues are real, but more

generally complex eigenvalues are accommodated if we allow Q to be normal, that is, $QQ^* = Q^*Q$. In each case we require the eigenvectors of Q to span the Hilbert space **H**.) This "evolution procedure" of the quantum state is very different from U, owing both to its discontinuity and its indeterminacy. The letter R will be used for this, standing for the "reduction" of the quantum state (sometimes referred to as the "collapse of the wave function"). This strange hybrid, whereby U and R are alternated, with U holding between measurements and R holding at measurements, is the standard procedure that is pragmatically adopted in conventional quantum mechanics, and which works so marvelously well, with no known discrepancy between the theory and observation. (In his classic account, von Neumann (1932, 1955), "R" is referred to as his "process I" and "U" as his "process II.") However, there appears to be no consensus whatever about the relation between this mathematical procedure and what is "really" going on in the physical world. This is the kind of issue that will be of concern to us here.

Quantum Reality

The discussion here will be given only in the Schrödinger picture, for the reason that the issues appear to be clearer with this description. In the Heisenberg picture, the state $|\psi\rangle$ does not evolve in time, and all dynamics is taken up in the time evolution of the dynamical variables. But this evolution does not refer to the evolution of specific systems, the "state" of any particular system being defined to remain constant in time. Since the Schrödinger and Heisenberg pictures are deemed to be equivalent (at least for the "normal" systems that are under consideration here), we do not lose anything substantial by sticking to Schrödinger's description, whereas there does seem to be a significant gain in understanding of what the formalism is actually telling us.

There are, however, many different attitudes that are expressed as to the "reality" of $|\psi\rangle$. (There is an unfortunate possibility of confusion here in the two uses of the word "real" that come into the discussion here. In the quantum formalism, the state is mathematically a "complex" rather than a "real" entity, whereas our present concern is not directly to do with this, but with the "ontology" of the quantum description.) According to what is commonly regarded as the standard – "Copenhagen" – interpretation of quantum mechanics (due primarily to Bohr, Heisenberg, and Pauli), the quantum state $|\psi\rangle$ is not taken as a description of a quantum-level reality at all, but merely as a description of the observer's knowledge of the of the quantum system under consideration. According to this view, the "jumping" that the quantum state undergoes is regarded as unsurprising, since it does not represent a sudden change in the reality of the situation, but merely in the observer's knowledge, as new information becomes available, when the result of some measurement becomes known to the observer. According to this view, there is no objective quantum reality described by $|\psi\rangle$. Whether or not there might be some objective quantum-level reality with some other mathematical description seems to be left open by this viewpoint, but the impression given is that there might well not be any such quantum-level reality at all, in the sense that it becomes meaningless to ask for a description of "actual reality" at quantum-relevant scales.

Of course some connection with the real world is necessary, in order that the quantum formalism can relate to the results of experiment. In the Copenhagen viewpoint, the experimenter's measuring apparatus is taken to be a classical-level entity, which can be ascribed a real ontological status. When the Geiger counter "clicks" or when the pointer "points" to some position on a dial, or when the track in the cloud chamber "becomes visible" – these are taken to be real events. The intervening description in terms of a quantum state vector $|\psi\rangle$ is not ascribed a reality. The role of $|\psi\rangle$ is merely to provide a calculational procedure whereby the different outcomes of an experiment can be assigned probabilities. Reality comes about only when the result of the measurement is manifested, not before.

A difficulty with this viewpoint is that it is hard to draw a clear line between those entities which are considered to have an actual reality, such as the experimental apparatus or a human observer, and the elemental constituents of those entities, which are such things as electrons or protons or neutrons or quarks, which are to be treated quantum mechanically and therefore, on the "Copenhagen" view, their mathematical descriptions are denied such an honored ontological status. Moreover, there is no limit to the number of particles that can partake in a quantum state. According to current quantum mechanics, the most accurate mathematical procedure for describing a system with a large number of particles would indeed be to use a unitarily evolving quantum state. What reasons can be presented for or against the viewpoint that this gives us a reasonable description of an actual reality? Can our perceived reality arise as some kind of statistical limit when very large numbers of constituents are involved?

Before entering into the more subtle and contentious issues of the nature of "quantum reality," it

is appropriate that one of the very basic mathematical aspects of the quantum formalism be addressed first. It is an accepted aspect of the quantum formalism that a state-vector such as $|\psi\rangle$ should not, in any case, be thought of as providing a unique mathematical description of a "physical reality" for the simple reason that $|\psi\rangle$ and $z|\psi\rangle$, where z is any nonzero complex number, describe precisely the same physical situation. It is a common, but not really necessary, practice to demand that $|\psi\rangle$ be normalized to unity: $\langle\psi|\psi\rangle = 1$, in which case the freedom in $|\psi\rangle$ is reduced to the multiplication by a phase factor $|\psi\rangle \mapsto e^{i\theta}|\psi\rangle$. Either way, the physically distinguishable states constitute a projective Hilbert space $\mathbb{P}\mathbf{H}$, where each point of $\mathbb{P}\mathbf{H}$ corresponds to a one-dimensional linear subspace of the Hilbert space $\mathbf{H}$. The issue, therefore, is whether quantum reality can be described in terms of the points of a projective Hilbert space $\mathbb{P}\mathbf{H}$.

Reality in Spin-1/2 Systems

As a general comment, it seems that for systems with a small number of degrees of freedom – that is, for a Hilbert space $\mathbf{H}^n$ of small finite dimension n – it seems more reasonable to assign a reality to the elements of $\mathbb{P}\mathbf{H}^n$ than is the case when n is large. Let us begin with a particularly simple case, where $n = 2$, and $\mathbf{H}^2$ describes the two-dimensional space of spin states of a massive particle of spin 1/2, such as an electron, proton, or quark, or suitable atom. Here we can take as an orthonormal pair of basis states $|\uparrow\rangle$ and $|\downarrow\rangle$, representing right-handed spin about the "up" and "down" directions, respectively. Clearly there is nothing special about these particular directions, so any other state of spin, of direction $|\nearrow\rangle$ say, is just as "real" as the original two. Indeed, we always find

$$|\nearrow\rangle = w|\uparrow\rangle + z|\downarrow\rangle$$

for some pair of complex numbers z and w (not both zero). The different possible ratios $z:w$ give us a complex plane (of zw^{-1}) compactified by a point at infinity (where $w = 0$) – a "Riemann sphere" – which is a realization of the complex projective 1-space $\mathbb{P}\mathbf{H}^2$.

There does indeed seem to be something "real" about the spin state of such a spin-1/2 particle or atom. We might imagine preparing the spin of a suitable spin-1/2 atom using a Stern–Gerlach apparatus (*see* Introductory Article: Quantum Mechanics) oriented in some chosen direction. The atom seems to "know" the direction of its spin, because if we measure it again in the same direction it has to be prepared to give us the answer "YES," to the second measurement, with certainty, and that direction for its spin state is the only one that can guarantee this answer. (We are, of course, considering only "ideal" measurements, for the purpose of argument.) Moreover, we could imagine that between the two measurements, some appropriate magnetic field had been introduced so as to rotate the spin direction in some very specific way, so that the spin state is now some other direction such as $|\searrow\rangle$. By rotating our second Stern–Gerlach apparatus to agree with this new direction, we must again get certainty for the YES answer, the guaranteeing of this by the rotated state seeming now to give a "reality" to this new state $|\searrow\rangle$. The quantum formalism does not allow us to ascertain an unknown direction of spin. But it does allow for us to "confirm" (or "refute") a proposed direction for the spin state, in the sense that if the proposed direction is incorrect, then there is a nonzero probability of refutation. Only the correct direction can be guaranteed to give the YES answer.

EPR–Bohm and Bell's Theorem

For a pair of particles or atoms of spin 1/2, the issue of the "reality" of spin states becomes less clear. Consider, for example, the EPR–Bohm example (where "EPR" stands for Einstein–Podolski–Rosen) whereby an initial state of spin 0 decays into two spin-1/2 atoms, traveling in opposite directions (east E, and west W). If a suitable Stern–Gerlach apparatus is set up to measure the spin of the atom at E, finding an answer $|\searrow\rangle$, say, then this immediately ensures that the state at W is the oppositely pointing $|\nwarrow\rangle$, which can subsequently be "confirmed" by measurement at W. This, then, seems to provide a "reality" for the spin state $|\nwarrow\rangle$ at W as soon as the E measurement has been performed, but not before. Now, let us suppose that some orientation different from $\nwarrow$ had actually been set up for the measurement at W, namely that which would have given YES for the direction $\leftarrow$. This measurement can certainly give the answer YES upon encountering $|\nwarrow\rangle$ (with a certain nonzero probability, namely $(1 + \cos\theta)/2$, where θ is the angle between $\nwarrow$ and $\leftarrow$). So far, this provides us with no problem with the "reality" of the spin state of the atom at W, since it would have been $|\nwarrow\rangle$ before the measurement at W and would have "collapsed" (by the R-process) to $|\leftarrow\rangle$ after the measurement. But now suppose that the measurement at W had actually been performed momentarily before the measurement at E, rather than just after it. Then there is no reason that the W-measurement would encounter $|\nwarrow\rangle$, rather than some other direction, but the result $|\leftarrow\rangle$ of the measurement at W now seems to force the state at E to be $|\rightarrow\rangle$. Indeed, the two measurements, at E and

at W, might have been spacelike separated, and because of the requirements of special relativity there would be no meaning to say which of the two measurements – at E or at W – had "actually" occurred first. One seems to obtain a different picture of "reality" depending on this ordering.

In fact, the calculations of probabilities come out the same whichever picture is used, so if one asks only for a calculational procedure for the probabilities, rather than an actual picture of quantum reality, these considerations are not problematic. But they do provide profound difficulties for any view of quantum reality that is entirely local. The difficulty is made particularly clear in a theorem due to John Bell (1964, 1966a, b) which showed that on the basis of the assumptions of local realism, there are particular relations between the conditional probabilities, which must hold in any situation of this kind; moreover, these inequalities can be violated in various situations in standard quantum mechanics. (See, most specifically, Clauser *et al.* (1969).) Several experiments that were subsequently performed (notably Aspect *et al.* (1982)) confirmed the expectations of quantum mechanics, thereby presenting profound difficulties for any local realistic model of the world. There are also situations of this kind which involve only yes/no questions, so that actual probabilities do not need to be considered, see Kochen and Specker (1967), Peres (1991), Hardy (1993), Conway and Kochen (2002). Basically: if one insists on realism, then one must give up locality. Moreover, nonlocal realistic models, consistent with the requirements of special relativity, are not easy to construct (*see* Quantum Mechanics: Generalizations), and have so far proved elusive.

Other Aspects of Quantum Nonlocality

Problems of this kind occur even at the more elementary level of single particles, if one tries to consider that an ordinary particle wave function (position-space description of $|\psi\rangle$) might be just some kind of "local disturbance," like an ordinary classical wave. Consider the wave function spreading out from a localized source, to be detected at a perpendicular screen some distance away. The detection of the particle at any one place on the screen immediately forbids the detection of that particle at any other place on the screen, and if we are to think of this information as being transmitted as a classical signal to all other places on the screen, then we are confronted with problems of superluminary communication. Again, any "realistic" picture of this process would require nonlocal ingredients, which are difficult to square with the requirements of special relativity. (It is possible that these difficulties might be resolved within some kind of nonlocal geometry, such as that supplied by twistor theory (*see* Twistors; Twistor Theory: Some Applications); see, particularly, Penrose (2005).)

These types of issues are made even more dramatic and problematic in the procedure of "quantum teleportation," whereby the information in a quantum state (e.g., the unknown actual direction ↗ in some quantum state $|\nearrow\rangle$) can be transported from one experimenter A to another one B, by merely the sending of a small finite number of classical bits of information from A to B, where before this classical information is transmitted, A and B must each be in possession of one member of an EPR pair. More explicitly, we may suppose A (Alice) is presented with a spin-1/2 state $|\nearrow\rangle$, but is not told the direction ↗. She has in her possession another spin-1/2 state which is an EPR–Bohm partner of a spin-1/2 state in the possession of B (Bob). She combines this $|\nearrow\rangle$ with her EPR atom and then performs a measurement which distinguishes the four orthogonal "Bell states"

$$\begin{aligned} 0:&\ |\uparrow\rangle|\downarrow\rangle - |\downarrow\rangle|\uparrow\rangle \\ 1:&\ |\uparrow\rangle|\uparrow\rangle - |\downarrow\rangle|\downarrow\rangle \\ 2:&\ |\uparrow\rangle|\uparrow\rangle + |\downarrow\rangle|\downarrow\rangle \\ 3:&\ |\uparrow\rangle|\downarrow\rangle + |\downarrow\rangle|\uparrow\rangle \end{aligned}$$

where the first state in each product refers to her unknown state and the second refers to her EPR atom. The result of this measurement is conveyed to Bob by an ordinary classical signal, coded by the indicated numbers 0, 1, 2, 3. On receiving Alice's message, Bob takes the other member of the EPR pair and performs the following rotation on it:

0: leave alone
1: 180° about x-axis
2: 180° about y-axis
3: 180° about z-axis

This achieves the successful "teleporting" of $|\nearrow\rangle$ from A to B, despite the fact that only 2 bits of classical information have been signaled. It is the acausal EPR–Bohm connection that provides the transmission of "quantum information" in a classically acausal way. Again, we see the essentially nonlocal (or acausal) nature of any attempted "realistic" picture of quantum phenomena. It may be regarded as inappropriate to use the term "information" for something that is propagated acausally and cannot be directly used for signaling. It has been suggested, accordingly, that a term such as "quanglement" might be more appropriate to use for this concept; see Penrose (2002, 2004).

The preceding arguments illustrate how quantum systems involving even just a few particles can exhibit features quite unlike the ordinary behavior of classical particles. This was pointed out by Schrödinger (1935), and he referred to this key property of composite quantum systems as "entanglement." An entangled quantum state (vector) is an element of a product Hilbert space $\mathbf{H}^m \otimes \mathbf{H}^n$ which cannot be written as a tensor product of elements $|\psi\rangle|\phi\rangle$, with $|\psi\rangle \in \mathbf{H}^m$ and $|\phi\rangle \in \mathbf{H}^n$, where $\mathbf{H}^m$ refers to one part of the system and $\mathbf{H}^n$ refers to another part, usually taken to be physically widely separated from the first. EPR systems are a clear example, and we begin to see very nonclassical, effectively nonlocal behavior with entangled systems generally. A puzzling aspect of this is that the vast majority of states are indeed entangled, and the more parts that a system has, the more entangled it becomes (where the generalization of this notion to more than two parts is evident). One might have expected that "big" quantum systems with large numbers of parts ought to behave more and more like classical systems when they get larger and more complicated. However, we see that this is very far from being the case. There is no good reason why a large quantum system, left on its own to evolve simply according to $\boldsymbol{U}$ should actually resemble a classical system, except in very special circumstances. Something of the nature of the $\boldsymbol{R}$ process seems to be needed in order that classical behaviour can "emerge."

Schrödinger's Cat

To clarify the nature of the problem we must consider a key feature of the $\boldsymbol{U}$ formalism, namely "linearity," which is supposed to hold no matter how large or complicated is the quantum system under consideration. Recall the quantum superposition principle, which allows us to construct arbitrary combinations of states

$$|\psi\rangle = w|\chi\rangle + z|\phi\rangle$$

from two given states $|\chi\rangle$ and $|\phi\rangle$. Quantum linearity tells us that if

$$|\chi\rangle \rightsquigarrow |\chi'\rangle \qquad \text{and} \qquad |\phi\rangle \rightsquigarrow |\phi'\rangle$$

where the symbol "$\rightsquigarrow$" expresses how a state will have evolved after a specified time period T, then

$$|\psi\rangle = w|\chi\rangle + z|\phi\rangle \rightsquigarrow |\psi'\rangle = w|\chi'\rangle + z|\phi'\rangle$$

Let us now consider how this might be applied in a particular, rather outlandish situation. Let us suppose that the $|\chi\rangle$-evolution consists of a photon going in one direction, encountering a detector, which is connected to some murderous device which kills a cat. The $|\phi\rangle$-evolution, on the other hand, consists of the photon going in some other direction, missing the detector so that the murderous device is not activated, and the cat is left alive. These two alternatives would each be perfectly plausible evolutions which might take place in the physical world. Now, by use of a beam splitter (effectively a "half-silvered mirror") we can easily arrange for the initial state of the photon to be the superposition $w|\chi\rangle + z|\phi\rangle$ of the two. Then by quantum linearity we find, as the final result, the superposed state $w|\chi'\rangle + z|\phi'\rangle$, in which the cat is in a superposition of life and death (a "Schrödinger's cat").

We note that the two individual final states $|\chi'\rangle$ and $|\phi'\rangle$ would each involve not just the cat but also its environment, fully entangled with the cat's state, and perhaps also some human observer looking at the cat. In the latter case, $|\chi'\rangle$ would involve the observer in a state of unhappily perceiving a dead cat, and $|\phi'\rangle$ happily perceiving a live one. Two of the "conventional standpoints" with regard to the measurement problem are of relevance here. According to the standpoint of environmental decoherence, the details of the environmental degrees of freedom are completely inaccessible, and it is deemed to be appropriate to construct a density matrix to describe the situation, which is a partial trace $\boldsymbol{D}$ of the quantity $|\psi\rangle\langle\psi|$, constructed by tracing out over all the environmental degrees of freedom:

$$\boldsymbol{D} = \text{trace over environment}\{|\psi\rangle\langle\psi|\}$$

The density matrix tends to be regarded as a more appropriate quantity than the ket $|\psi\rangle$ to represent the physical situation, although this represents something of an "ontology shift" from the point of view that was being held previously. Under appropriate assumptions, $\boldsymbol{D}$ may now be shown to attain a form that is close to being diagonal in a basis with respect to which the cat is either dead or alive, and then, by a second "ontology shift" $\boldsymbol{D}$ is re-read as describing a probability mixture of these two states.

According to the second "conventional standpoint" under consideration here, it is not logical to take this detour through a density-matrix description, and instead one should maintain a consistent ontology by following the evolution of the state $|\psi\rangle$ itself throughout. The "real" resulting physical state is then taken to be actually $|\psi'\rangle$, which involves the superposition of a dead and live cat. Of course this "reality" does not agree with the reality that we actually perceive, so the position is taken that a conscious mind would not actually be able to function in such a superposed condition, and would have to settle into a state of perception of either a dead cat or a live one, these two alternatives occurring with probabilities as given by the Born rule stated above. It may be argued that this conclusion depends

upon some appropriate theory of how conscious minds actually perceive things, and this appears to be lacking.

A good many physicists might argue that none of these attempts at resolution of the measurement problem is satisfactory, including "Copenhagen," although the latter at least has the advantage of offering a pragmatic, if not fully logical, stance. Such physicists might take the position that it is necessary to move away from the precise version of quantum theory that we have at present, and turn to one of its modifications. Some major candidates for modification are discussed in Quantum Mechanics: Generalizations. Most of these actually make predictions that, at some stage, would differ from those of standard quantum mechanics. So it becomes an experimental matter to ascertain the plausibility of these schemes. In addition, there are reinterpretations which do not change quantum theory's predictions, such as the de Broglie–Bohm model. In this, there are two levels of "reality," a firmer one with a particle or position-space ontology, and a secondary one containing waves which guide the behavior at the firmer level. It is clear, however, that these issues will remain the subject of debate for many years to come.

See also: Functional Integration in Quantum Physics; Normal Forms and Semiclassical Approximation; Quantum Mechanics: Generalizations; Twistor Theory: Some Applications [In Integrable Systems, Complex Geometry and String Theory]; Twistors.

Further Reading

Aspect A, Grangier P, and Roger G (1982) Experimental realization of Einstein–Podolsky–Rosen–Bohm. *Gedankenexperiment*: a new violation of Bell's inequalities. *Physical Review Letters* 48: 91–94.

Bell JS (1964) On the Einstein Podolsky Rosen paradox. *Physics* 1: 195–200. Reprinted in *Quantum Theory and Measurement*, eds., Wheeler JA and Zurek WH (Princeton Univ. Press, Princeton, 1983).

Bell JS (1966a) On the problem of hidden variables in quantum theory. *Reviews in Modern Physics* 38: 447–452.

Bell JS (1966b) *Speakable and Unspeakable in Quantum Mechanics*. Cambridge: Cambridge University Press. Reprint 1987.

Clauser JF, Horne MA, Shimony A, and Holt RA (1969) Proposed experiment to test local hidden-variable theories. *Physical Review Letters* 23: 880–884.

Conway J and Kochen S (2002) The geometry of the quantum paradoxes. In: Bertlmann RA and Zeilinger A (eds.) *Quantum [Un]speakables: From Bell to Quantum Information*, Ch. 18 (ISBN 3-540-42756-2). Berlin: Springer.

Hardy L (1993) Nonlocality for two particles without inequalities for almost all entangled states. *Physical Review Letters* 71(11): 1665.

Kochen S and Specker EP (1967) *Journal of Mathematics and Mechanics* 17: 59.

Penrose R (2002) John Bell, state reduction, and quanglement. In: Reinhold A, Bertlmann, and Zeilinger A (eds.) *Quantum [Un]speakables: From Bell to Quantum Information*, pp. 319–331. Berlin: Springer.

Penrose R (2004) *The Road to Reality: A Complete Guide to the Laws of the Universe*. London: Jonathan Cape.

Penrose R (2005) The twistor approach to space-time structures. In: Ashtekar A (ed.) *In 100 Years of Relativity; Space-Time Structure: Einstein and Beyond*. Singapore: World Scientific.

Peres A (1991) Two simple proofs of the Kochen–Specker theorem. *Journal of Physics A: Mathematical and General* 24: L175–L178.

Schrödinger E (1935) Probability relations between separated systems. *Proceedings of the Cambridge Philosophical Society* 31: 555–563.

von Neumann J (1932) *Mathematische Grundlageen der Quantenmechanik*. Berlin: Springer.

von Neumann J (1955) *Mathematical Foundations of Quantum Mechanics*. Princeton: Princeton University Press.

Quantum Mechanics: Generalizations

P Pearle, Hamilton College, Clinton, NY, USA
A Valentini, Perimeter Institute for Theoretical Physics, Waterloo, ON, Canada

Introduction

According to the so-called "Copenhagen Interpretation," standard quantum theory is limited to describing experimental situations. It is at once remarkably successful in its predictions, and remarkably ill-defined in its conceptual structure: what is an experiment? what physical objects do or do not require quantization? how are the states realized in nature to be characterized? how and when is the wave-function "collapse postulate" to be invoked? Because of its success, one may suspect that quantum theory can be promoted from a theory of measurement to a theory of reality. But, that requires there to be an unambiguous specification (S) of the possible real states of nature and their probabilities of being realized.

There are several approaches that attempt to achieve S. The more conservative approaches (e.g., consistent histories, environmental decoherence, many worlds) do not produce any predictions that differ from the standard ones because they do not tamper with the usual basic mathematical

formalism. Rather, they utilize structures compatible with standard quantum theory to elucidate S. These approaches, which will not be discussed in this article, have arguably been less successful so far at achieving S than approaches that introduce significant alterations to quantum theory.

This article will largely deal with the two most well-developed realistic models that reproduce quantum theory in some limit and yield potentially new and testable physics outside that limit. First, the pilot-wave model, which will be discussed in the broader context of "hidden-variables theories." Second, the continuous spontaneous localization (CSL) model, which describes wave-function collapse as a physical process. Other related models will also be discussed briefly.

Due to bibliographic space limitations, this article contains a number of uncited references, of the form "[author] in [year]." Those in the next section can be found in Valentini (2002b, 2004a,b) or at www.arxiv.org. Those in the subsequent sections can be found in Adler (2004), Bassi and Ghirardi (2003), Pearle (1999) (or in subsequent papers by these authors, or directly, at www.arxiv.org), and in Wallstrom (1994).

Hidden Variables and Quantum Nonequilibrium

A deterministic hidden-variables theory defines a mapping $\omega = \omega(M, \lambda)$ from initial hidden parameters λ (defined, e.g., at the time of preparation of a quantum state) to final outcomes ω of quantum measurements. The mapping depends on macroscopic experimental settings M, and fixes the outcome for each run of the experiment. Bell's theorem of 1964 shows that, for entangled quantum states of widely separated systems, the mapping must be nonlocal: some outcomes for (at least) one system must depend on the setting for another distant system.

In a viable theory, the statistics of quantum measurement outcomes – over an ensemble of experimental trials with fixed settings M – will agree with quantum theory for some special distribution $\rho_{\rm QT}(\lambda)$ of hidden variables. For example, expectation values will coincide with the predictions of the Born rule

$$\langle\omega\rangle_{\rm QT} \equiv \int {\rm d}\lambda\, \rho_{\rm QT}(\lambda)\omega(M,\lambda) = {\rm tr}(\hat{\rho}\hat{\Omega})$$

for an appropriate density operator $\hat{\rho}$ and Hermitian observable $\hat{\Omega}$. (As is customary in this context, $\int {\rm d}\lambda$ is to be understood as a generalized sum.) However, given the mapping $\omega = \omega(M, \lambda)$ for individual trials, one may, in principle, consider nonstandard distributions $\rho(\lambda) \neq \rho_{\rm QT}(\lambda)$ that yield statistics outside the domain of ordinary quantum theory (Valentini 1991, 2002a). We may say that such distributions correspond to a state of quantum nonequilibrium.

Quantum nonequilibrium is characterized by the breakdown of a number of basic quantum constraints. In particular, nonlocal signals appear at the statistical level. We shall first illustrate this for the hidden-variables model of de Broglie and Bohm. Then we shall generalize the discussion to all (deterministic) hidden-variables theories.

At present there is no experimental evidence for quantum nonequilibrium in nature. However, from a hidden-variables perspective, it is natural to explore the theoretical properties of nonequilibrium distributions, and to search experimentally for the statistical anomalies associated with them.

From this point of view, quantum theory is a special case of a wider physics, much as thermal physics is a special case of a wider (nonequilibrium) physics. (The special distribution $\rho_{\rm QT}(\lambda)$ is analogous to, say, Maxwell's distribution of molecular speeds.) Quantum physics may be compared with the physics of global thermal equilibrium, which is characterized by constraints – such as the impossibility of converting heat into work (in the absence of temperature differences) – that are not fundamental but contingent on the state. Similarly, quantum constraints such as statistical locality (the impossibility of converting entanglement into a practical signal) are seen as contingencies of $\rho_{\rm QT}(\lambda)$.

Pilot-Wave Theory

The de Broglie–Bohm "pilot-wave theory" – as it was originally called by de Broglie, who first presented it at the Fifth Solvay Congress in 1927 – is the classic example of a deterministic hidden-variables theory of broad scope (Bohm 1952, Bell 1987, Holland 1993). We shall use it to illustrate the above ideas. Later, the discussion will be generalized to arbitrary theories.

In pilot-wave dynamics, an individual closed system with (configuration-space) wave function $\Psi(X, t)$ satisfying the Schrödinger equation

$$i\hbar\frac{\partial\Psi}{\partial t} = \hat{H}\Psi \qquad [1]$$

has an actual configuration $X(t)$ with velocity

$$\dot{X}(t) = \frac{J(X,t)}{|\Psi(X,t)|^2} \qquad [2]$$

where $J = J[\Psi] = J(X,t)$ satisfies the continuity equation

$$\frac{\partial|\Psi|^2}{\partial t} + \nabla \cdot J = 0 \qquad [3]$$

(which follows from [1]). In quantum theory, J is the "probability current." In pilot-wave theory, Ψ is an objective physical field (on configuration space) guiding the motion of an individual system.

Here, the objective state (or ontology) for a closed system is given by Ψ and X. A probability distribution for X – discussed below – completes an unambiguous specification S (as mentioned in the introduction).

Pilot-wave dynamics may be applied to any quantum system with a locally conserved current in configuration space. Thus, X may represent a many-body system, or the configuration of a continuous field, or perhaps some other entity.

For example, at low energies, for a system of N particles with positions $\boldsymbol{x}_i(t)$ and masses m_i $(i = 1, 2, \ldots, N)$, with an external potential V, [1] (with $X \equiv (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)$) reads

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} = \sum_{i=1}^{N} -\frac{\hbar^2}{2m_i}\nabla_i^2\Psi + V\Psi \qquad [4]$$

while [2] has components

$$\frac{\mathrm{d}\boldsymbol{x}_i}{\mathrm{d}t} = \frac{\hbar}{m_i}\mathrm{Im}\left(\frac{\nabla_i\Psi}{\Psi}\right) = \frac{\nabla_i S}{m_i} \qquad [5]$$

(where $\Psi = |\Psi|\mathrm{e}^{(\mathrm{i}/\hbar)S}$).

In general, [1] and [2] determine $X(t)$ for an individual system, given the initial conditions $X(0), \Psi(X,0)$ at $t = 0$. For an arbitrary initial distribution $P(X,0)$, over an ensemble with the same wave function $\Psi(X,0)$, the evolution $P(X,t)$ of the distribution is given by the continuity equation

$$\frac{\partial P}{\partial t} + \nabla \cdot (P\dot{X}) = 0 \qquad [6]$$

The outcome of an experiment is determined by $X(0), \Psi(X,0)$, which may be identified with λ. For an ensemble with the same $\Psi(X,0)$, we have $\lambda = X(0)$.

Quantum equilibrium From [3] and [6], if we assume $P(X,0) = |\Psi(X,0)|^2$ at $t = 0$, we obtain $P(X,t) = |\Psi(X,t)|^2$ – the Born-rule distribution of configurations – at all times t.

Quantum measurements are, like any other process, described and explained in terms of evolving configurations. For measurement devices whose pointer readings reduce to configurations, the distribution of outcomes of quantum measurements will match the statistical predictions of quantum theory (Bohm 1952, Bell 1987, Dürr *et al.* 2003). Thus, quantum theory emerges phenomenologically for a "quantum equilibrium" ensemble with distribution $P(X,t) = |\Psi(X,t)|^2$ (or $\rho(\lambda) = \rho_{\mathrm{QT}}(\lambda)$).

Quantum nonequilibrium In principle, as we saw for general hidden-variables theories, we may consider a nonequilibrium distribution $P(X,0) \neq |\Psi(X,0)|^2$ of initial configurations while retaining the same deterministic dynamics [1], [2] for individual systems (Valentini 1991). The time evolution of $P(X,t)$ will be determined by [6].

As we shall see, in appropriate circumstances (with a sufficiently complicated velocity field $\dot{X}$), [6] generates relaxation $P \to |\Psi|^2$ on a coarse-grained level, much as the analogous classical evolution on phase space generates thermal relaxation. But for as long as the ensemble is in nonequilibrium, the statistics of outcomes of quantum measurements will disagree with quantum theory.

Quantum nonequilibrium may have existed in the very early universe, with relaxation to equilibrium occurring soon after the big bang. Thus, a hidden-variables analog of the classical thermodynamic "heat death of the universe" may have actually taken place (Valentini 1991). Even so, relic cosmological particles that decoupled sufficiently early could still be in nonequilibrium today, as suggested by Valentini in 1996 and 2001. It has also been speculated that nonequilibrium could be generated in systems entangled with degrees of freedom behind a black-hole event horizon (Valentini 2004a).

Experimental searches for nonequilibrium have been proposed. Nonequilibrium could be detected by the statistical analysis of random samples of particles taken from a parent population of (for example) relics from the early universe. Once the parent distribution is known, the rest of the population could be used as a resource, to perform tasks that are currently impossible (Valentini 2002b).

H-Theorem: Relaxation to Equilibrium

Before discussing the potential uses of nonequilibrium, we should first explain why all systems probed so far have been found in the equilibrium state $P = |\Psi|^2$. This distribution may be accounted for along the lines of classical statistical mechanics, noting that all currently accessible systems have had a long and violent astrophysical history.

Dividing configuration space into small cells, and introducing coarse-grained quantities $\bar{P}, \overline{|\Psi|^2}$, a general argument for relaxation $\bar{P} \to \overline{|\Psi|^2}$ is based on an

analog of the classical coarse-graining H-theorem. The coarse-grained H-function

$$\bar{H} = \int \mathrm{d}X\ \bar{P} \ln(\bar{P}/\overline{|\Psi|^2}) \qquad [7]$$

(minus the relative entropy of $\bar{P}$ with respect to $|\Psi|^2$) obeys the H-theorem (Valentini 1991)

$$\bar{H}(t) \leq \bar{H}(0)$$

(assuming no initial fine-grained microstructure in P and $|\Psi|^2$). Here, $\bar{H} \geq 0$ for all $\bar{P}, \overline{|\Psi|^2}$ and $\bar{H} = 0$ if and only if $\bar{P} = \overline{|\Psi|^2}$ everywhere.

The H-theorem expresses the fact that P and $|\Psi|^2$ behave like two "fluids" that are "stirred" by the same velocity field $\dot{X}$, so that P and $|\Psi|^2$ tend to become indistinguishable on a coarse-grained level. Like its classical analog, the theorem provides a general understanding of how equilibrium is approached, while not proving that equilibrium is actually reached. (And of course, for some simple systems – such as a particle in the ground state of a box, for which the velocity field $\nabla S/m$ vanishes – there is no relaxation at all.) A strict decrease of $\bar{H}(t)$ immediately after $t = 0$ is guaranteed if $\dot{X}_0 \cdot \nabla(P_0/|\Psi_0|^2)$ has nonzero spatial variance over a coarse-graining cell, as shown by Valentini in 1992 and 2001.

A relaxation timescale τ may be defined by $1/\tau^2 \equiv -(\mathrm{d}^2\bar{H}/\mathrm{d}t^2)_0/\bar{H}_0$. For a single particle with quantum energy spread ΔE, a crude estimate given by Valentini in 2001 yields $\tau \sim (1/\varepsilon)\hbar^2/m^{1/2}(\Delta E)^{3/2}$, where ε is the coarse-graining length. For wave functions that are superpositions of many energy eigenfunctions, the velocity field (generally) varies rapidly, and detailed numerical simulations (in two dimensions) show that relaxation occurs with an approximately exponential decay $\bar{H}(t) \approx \bar{H}_0 \mathrm{e}^{-t/t_c}$, with a time constant t_c of order τ (Valentini and Westman 2005).

Equilibrium is then to be expected for particles emerging from the violence of the big bang. The possibility is still open that relics from very early times may not have reached equilibrium before decoupling.

Nonlocal Signaling

We now show how nonequilibrium, if it were ever discovered, could be used for nonlocal signaling.

Pilot-wave dynamics is nonlocal. For a pair of particles A, B with entangled wave function $\Psi(\boldsymbol{x}_A, \boldsymbol{x}_B, t)$, the velocity $\dot{\boldsymbol{x}}_A(t) = \nabla_A S(\boldsymbol{x}_A, \boldsymbol{x}_B, t)/m_A$ of A depends instantaneously on $\boldsymbol{x}_B$, and local operations at B – such as switching on a potential – instantaneously affect the motion of A. For an ensemble $P(\boldsymbol{x}_A, \boldsymbol{x}_B, t) = |\Psi(\boldsymbol{x}_A, \boldsymbol{x}_B, t)|^2$, local operations at B have no statistical effect at A: the individual nonlocal effects vanish upon averaging over an equilibrium ensemble.

Nonlocality is (generally) hidden by statistical noise only in quantum equilibrium. If instead $P(\boldsymbol{x}_A, \boldsymbol{x}_B, 0) \neq |\Psi(\boldsymbol{x}_A, \boldsymbol{x}_B, 0)|^2$, a local change in the Hamiltonian at B generally induces an instantaneous change in the marginal $p_A(\boldsymbol{x}_A, t) \equiv \int \mathrm{d}^3\boldsymbol{x}_B P(\boldsymbol{x}_A, \boldsymbol{x}_B, t)$ at A. For example, in one dimension a sudden change $\hat{H}_B \to \hat{H}'_B$ in the Hamiltonian at B induces a change $\Delta p_A \equiv p_A(x_A, t) - p_A(x_A, 0)$ (for small t) (Valentini 1991),

$$\Delta p_A = -\frac{t^2}{4m}\frac{\partial}{\partial x_A}\left(a(x_A)\int \mathrm{d}x_B\ b(x_B) \times \frac{P(x_A, x_B, 0) - |\Psi(x_A, x_B, 0)|^2}{|\Psi(x_A, x_B, 0)|^2}\right) \qquad [8]$$

(Here $m_A = m_B = m$, $a(x_A)$ depends on $\Psi(x_A, x_B, 0)$, while $b(x_B)$ also depends on $\hat{H}'_B$ and vanishes if $\hat{H}'_B = \hat{H}_B$.) The signal is generally nonzero if $P_0 \neq |\Psi_0|^2$.

Nonlocal signals do not lead to causal paradoxes if, at the hidden-variable level, there is a preferred foliation of spacetime with a time parameter that defines a fundamental causal sequence. Such signals, if they were observed, would define an absolute simultaneity as discussed by Valentini in 1992 and 2005. Note that in pilot-wave field theory, Lorentz invariance emerges as a phenomenological symmetry of the equilibrium state, conditional on the structure of the field-theoretical Hamiltonian (as discussed by Bohm and Hiley in 1984, Bohm, Hiley and Kaloyerou in 1987, and Valentini in 1992 and 1996).

Subquantum Measurement

In principle, nonequilibrium particles could also be used to perform "subquantum measurements" on ordinary, equilibrium systems. We illustrate this with an exactly solvable one-dimensional model (Valentini 2002b).

Consider an apparatus "pointer" coordinate y, with known wave function $g_0(y)$ and known (ensemble) distribution $\pi_0(y) \neq |g_0(y)|^2$, where $\pi_0(y)$ has been deduced by statistical analysis of random samples from a parent population with known wave function $g_0(y)$. (We assume that relaxation may be neglected: for example, if g_0 is a box ground state, $\dot{y} = 0$ and $\pi_0(y)$ is static.) Consider also a "system" coordinate x with known wave function $\psi_0(x)$ and known distribution $\rho_0(x) = |\psi_0(x)|^2$. If $\pi_0(y)$ is arbitrarily narrow, x_0 can be measured without

disturbing $\psi_0(x)$, to arbitrary accuracy (violating the uncertainty principle).

To do this, at $t=0$ we switch on an interaction Hamiltonian $\hat{H}=a\hat{x}\hat{p}_y$, where a is a constant and p_y is canonically conjugate to y. For relatively large a, we may neglect the Hamiltonians of x and y. For $\Psi=\Psi(x,y,t)$, we then have $\partial\Psi/\partial t=-ax\partial\Psi/\partial y$. For $|\Psi|^2$ we have the continuity equation $\partial|\Psi|^2/\partial t=-ax\partial|\Psi|^2/\partial y$, which implies the hidden-variable velocity fields $\dot{x}=0, \dot{y}=ax$ and trajectories $x(t)=x_0$, $y(t)=y_0+ax_0t$.

The initial product $\Psi_0(x,y)=\psi_0(x)g_0(y)$ evolves into $\Psi(x,y,t)=\psi_0(x)g_0(y-axt)$. For $at\to 0$ (with a large but fixed), $\Psi(x,y,t)\to\psi_0(x)g_0(y)$ and $\psi_0(x)$ is undisturbed: for small at, a standard quantum pointer with the coordinate y would yield negligible information about x_0. Yet, for arbitrarily small at, the hidden-variable pointer coordinate $y(t)=y_0+ax_0t$ does contain complete information about x_0 (and $x(t)=x_0$). This "subquantum" information will be visible to us if $\pi_0(y)$ is sufficiently narrow.

For, over an ensemble of similar experiments, with initial joint distribution $P_0(x,y)=|\psi_0(x)|^2\pi_0(y)$ (equilibrium for x and nonequilibrium for y), the continuity equation $\partial P/\partial t=-ax\partial P/\partial y$ implies that $P(x,y,t)=|\psi_0(x)|^2\pi_0(y-axt)$. If $\pi_0(y)$ is localized around $y=0$ ($\pi_0(y)=0$ for $|y|>w/2$), then a standard (faithful) measurement of y with result y_{meas} will imply that x lies in the interval $(y_{\text{meas}}/at-w/2at, y_{\text{meas}}/at+w/2at)$ (so that $P(x,y,t)\neq 0$). Taking the simultaneous limits $at\to 0$, $w\to 0$, with $w/at\to 0$, the midpoint $y_{\text{meas}}/at\to x_0$ (since $y_{\text{meas}}=y_0+ax_0t$ and $|y_0|\leq w/2$), while the error $w/2at\to 0$.

If w is arbitrarily small, a sequence of such measurements will determine the hidden trajectory $x(t)$ without disturbing $\psi(x,t)$, to arbitrary accuracy.

Subquantum Information and Computation

From a hidden-variables perspective, immense physical resources are hidden from us by equilibrium statistical noise. Quantum nonequilibrium would probably be as useful technologically as thermal or chemical nonequilibrium.

Distinguishing nonorthogonal states In quantum theory, nonorthogonal states $|\psi_1\rangle, |\psi_2\rangle$ ($\langle\psi_1|\psi_2\rangle\neq 0$) cannot be distinguished without disturbing them. This theorem breaks down in quantum nonequilibrium (Valentini 2002b). For example, if $|\psi_1\rangle, |\psi_2\rangle$ are distinct states of a single spinless particle, then the associated de Broglie–Bohm velocity fields will in general be different, even if $\langle\psi_1|\psi_2\rangle\neq 0$, and so will the hidden-variable trajectories. Subquantum measurement of the trajectories could then distinguish the states $|\psi_1\rangle, |\psi_2\rangle$.

Breaking quantum cryptography The security of standard protocols for quantum key distribution depends on the validity of the laws of quantum theory. These protocols would become insecure given the availability of nonequilibrium systems (Valentini 2002b).

The protocols known as BB84 and B92 depend on the impossibility of distinguishing nonorthogonal quantum states without disturbing them. An eavesdropper in possession of nonequilibrium particles could distinguish the nonorthogonal states being transmitted between two parties, and so read the supposedly secret key. Further, if subquantum measurements allow an eavesdropper to predict quantum measurement outcomes at each "wing" of a (bipartite) entangled state, then the EPR (Einstein–Podolsky–Rosen) protocol also becomes insecure.

Subquantum computation It has been suggested that nonequilibrium physics would be computationally more powerful than quantum theory, because of the ability to distinguish nonorthogonal states (Valentini 2002b). However, this ability depends on the (less-than-quantum) dispersion w of the nonequilibrium ensemble. A well-defined model of computational complexity requires that the resources be quantified in some way. Here, a key question is how the required w scales with the size of the computational task. So far, no rigorous results are known.

Extension to All Deterministic Hidden-Variables Theories

Let us now discuss arbitrary (deterministic) theories.

Nonlocal signaling Consider a pair of two-state quantum systems A and B, which are widely separated and in the singlet state. Quantum measurements of observables $\hat{\sigma}_A\equiv\boldsymbol{m}_A\cdot\hat{\boldsymbol{\sigma}}_A, \hat{\sigma}_B\equiv\boldsymbol{m}_B\cdot\hat{\boldsymbol{\sigma}}_B$ (where $\boldsymbol{m}_A, \boldsymbol{m}_B$ are unit vectors in Bloch space and $\hat{\boldsymbol{\sigma}}_A, \hat{\boldsymbol{\sigma}}_B$ are Pauli spin operators) yield outcomes $\sigma_A, \sigma_B=\pm 1$, in the ratio $1:1$ at each wing, with a correlation $\langle\hat{\sigma}_A\hat{\sigma}_B\rangle=-\boldsymbol{m}_A\cdot\boldsymbol{m}_B$. Bell's theorem shows that for a hidden-variables theory to reproduce this correlation – upon averaging over an equilibrium ensemble with distribution $\rho_{\text{QT}}(\lambda)$ – it must take the nonlocal form

$$\sigma_A=\sigma_A(\boldsymbol{m}_A,\boldsymbol{m}_B,\lambda),\qquad \sigma_B=\sigma_B(\boldsymbol{m}_A,\boldsymbol{m}_B,\lambda) \quad [9]$$

More precisely, to obtain $\langle\sigma_A\sigma_B\rangle_{\text{QT}}=-\boldsymbol{m}_A\cdot\boldsymbol{m}_B$ (where $\langle\sigma_A\sigma_B\rangle_{\text{QT}}\equiv\int d\lambda\rho_{\text{QT}}(\lambda)\sigma_A\sigma_B$), at least one of

σ_A, σ_B must depend on the measurement setting at the distant wing. Without loss of generality, we assume that σ_A depends on $\boldsymbol{m}_B$.

For an arbitrary nonequilibrium ensemble with distribution $\rho(\lambda) \neq \rho_{\rm QT}(\lambda)$, in general $\langle \sigma_A \sigma_B \rangle \equiv \int {\rm d}\lambda\, \rho(\lambda) \sigma_A \sigma_B$ differs from $-\boldsymbol{m}_A \cdot \boldsymbol{m}_B$, and the outcomes $\sigma_A, \sigma_B = \pm 1$ occur in a ratio different from 1 : 1. Further, a change of setting $\boldsymbol{m}_B \rightarrow \boldsymbol{m}'_B$ at B will generally induce a change in the outcome statistics at A, yielding a nonlocal signal at the statistical level. To see this, note that, in a nonlocal theory, the "transition sets"

$$\begin{aligned} T_A(-,+) \equiv \{\lambda | \sigma_A(\boldsymbol{m}_A, \boldsymbol{m}_B, \lambda) = -1, \\ \sigma_A(\boldsymbol{m}_A, \boldsymbol{m}'_B, \lambda) = +1\} \\ T_A(+,-) \equiv \{\lambda | \sigma_A(\boldsymbol{m}_A, \boldsymbol{m}_B, \lambda) = +1, \\ \sigma_A(\boldsymbol{m}_A, \boldsymbol{m}'_B, \lambda) = -1\} \end{aligned}$$

cannot be empty for arbitrary settings. Yet, in quantum equilibrium, the outcomes $\sigma_A = \pm 1$ occur in the ratio 1 : 1 for all settings, so the transition sets must have equal equilibrium measure, $\mu_{\rm QT}[T_A(-,+)] = \mu_{\rm QT}[T_A(+,-)]$ ($\mathrm{d}\mu_{\rm QT} \equiv \rho_{\rm QT}(\lambda)\mathrm{d}\lambda$). That is, the fraction of the equilibrium ensemble making the transition $\sigma_A = -1 \rightarrow \sigma_A = +1$ under $\boldsymbol{m}_B \rightarrow \boldsymbol{m}'_B$ must equal the fraction making the reverse transition $\sigma_A = +1 \rightarrow \sigma_A = -1$. (This "detailed balancing" is analogous to the principle of detailed balance in statistical mechanics.) Since $T_A(-,+), T_A(+,-)$ are fixed by the deterministic mapping, they are independent of the ensemble distribution $\rho(\lambda)$. Thus, for $\rho(\lambda) \neq \rho_{\rm QT}(\lambda)$, in general $\mu[T_A(-,+)] \neq \mu[T_A(+,-)]$ ($\mathrm{d}\mu \equiv \rho(\lambda) d\lambda$): the fraction of the nonequilibrium ensemble making the transition $\sigma_A = -1 \rightarrow \sigma_A = +1$ will not in general balance the fraction making the reverse transition. The outcome ratio at A will then change under $\boldsymbol{m}_B \rightarrow \boldsymbol{m}'_B$ and there will be an instantaneous signal at the statistical level from B to A (Valentini 2002a).

Thus, in any deterministic hidden-variables theory, nonequilibrium distributions $\rho(\lambda) \neq \rho_{\rm QT}(\lambda)$ generally allow entanglement to be used for nonlocal signalling (just as, in ordinary statistical physics, differences of temperature make it possible to convert heat into work).

Experimental signature of nonequilibrium Quantum expectations are additive, $\langle c_1 \hat{\Omega}_1 + c_2 \hat{\Omega}_2 \rangle = c_1 \langle \hat{\Omega}_1 \rangle + c_2 \langle \hat{\Omega}_2 \rangle$, even for noncommuting observables ($[\hat{\Omega}_1, \hat{\Omega}_2] \neq 0$, with c_1, c_2 real). As emphasized by Bell in 1966, this seemingly trivial consequence of the (linearity of the) Born rule $\langle \hat{\Omega} \rangle = {\rm tr}(\hat{\rho}\hat{\Omega})$ is remarkable because it relates statistics from distinct, "incompatible" experiments. In nonequilibrium, such additivity generically breaks down (Valentini 2004b).

Further, for a two-state system with observables $\boldsymbol{m} \cdot \hat{\boldsymbol{\sigma}}$, the "dot-product" structure of the quantum expectation $\langle \boldsymbol{m} \cdot \hat{\boldsymbol{\sigma}} \rangle = {\rm tr}(\hat{\rho} \boldsymbol{m} \cdot \hat{\boldsymbol{\sigma}}) = \boldsymbol{m} \cdot \boldsymbol{P}$ (for some Bloch vector $\boldsymbol{P}$) is equivalent to expectation additivity (Valentini 2004b). Nonadditive expectations then provide a convenient signature of nonequilibrium for any two-state system. For example, the sinusoidal modulation of the quantum transmission probability for a single photon through a polarizer

$$p^{+}_{\rm QT}(\Theta) = \tfrac{1}{2}(1 + \langle \boldsymbol{m} \cdot \hat{\boldsymbol{\sigma}} \rangle) = \tfrac{1}{2}(1 + P \cos 2\Theta) \qquad [10]$$

(where an angle θ on the Bloch sphere corresponds to a physical angle $\Theta = \theta/2$) will generically break down in nonequilibrium. Deviations from [10] would provide an unambiguous violation of quantum theory (Valentini 2004b).

Such deviations were searched for by Papaliolios in 1967, using laboratory photons and successive polarization measurements over very short times, to test a hidden-variables theory (distinct from pilot-wave theory) due to Bohm and Bub (1966), in which quantum measurements generate nonequilibrium for short times. Experimentally, successive measurements over timescales $\sim 10^{-13}$ s agreed with the (quantum) sinusoidal modulation $\cos^2 \Theta$ to $\lesssim 1\%$. Similar tests might be performed with photons of a more exotic origin.

Continuous Spontaneous Localization Model (CSL)

The basic postulate of CSL is that the state vector $|\psi, t\rangle$ represents reality. Since, for example, in describing a measurement, the usual Schrödinger evolution readily takes a real state into a nonreal state, that is, into a superposition of real states (such as apparatus states describing different experimental outcomes), CSL requires a modification of Schrödinger's evolution. To the Hamiltonian is added a term which depends upon a classical randomly fluctuating field $w(\boldsymbol{x}, t)$ and a mass-density operator $\hat{A}(\boldsymbol{x}, t)$. This term acts to collapse a superposition of states, which differ in their spatial distribution of mass density, to one of these states. The rate of collapse is very slow for a superposition involving a few particles, but very fast for a superposition of macroscopically different states. Thus, very rapidly, what you see (in nature) is what you get (from the theory). Each state vector evolving under each $w(\boldsymbol{x}, t)$ corresponds to a realizable state, and a rule is given for how to associate a probability with each. In this way, an

unambiguous specification S, as mentioned in the introduction, is achieved.

Requirements for Stochastic Collapse Dynamics

Consider a normalized state vector $|\psi,t\rangle = \sum_n \alpha_n(t)|a_n\rangle (\langle a_n|a_{n'}\rangle = \delta_{nn'})$ which undergoes a stochastic dynamical collapse process. This means that, starting from the initial superposition at $t=0$, for each run of the process, the squared amplitudes $x_n(t) \equiv |\alpha_n(t)|^2$ fluctuate until all but one vanish, that is, $x_m(\infty)=1, (x_{\neq m}(\infty)=0)$ with probability $x_m(0)$.

This may be achieved simply, assuming negligible effect of the usual Schrödinger evolution, if the stochastic process enjoys the following properties (Pearle 1979):

$$\sum_n x_n(t) = 1 \qquad [11a]$$

$$\overline{x_n(t)} = x_n(0) \qquad [11b]$$

$$\overline{x_n(\infty)x_m(\infty)} = 0 \quad \text{for } m \neq n \qquad [11c]$$

where the overbar indicates the ensemble average at the indicated time. The only way that a sum of products of non-negative terms can vanish is for at least one term in each product to vanish. Thus, according to [11c], for each run, at least one of each pair $\{x_n(\infty), x_m(\infty)\}(n \neq m)$ must vanish. This means that at most one $x_n(\infty)$ might not vanish and, by [11a], applied at $t=\infty$, one $x_n(\infty)$ must not vanish and, in fact, must equal 1: hence, each run produces collapse. Now, let the probability of the outcome $\{x_n(\infty)=1, x_{\neq n}(\infty)=0\}$ be denoted P_n. Since $\overline{x_n(\infty)} = 1\cdot P_n + \sum_{m\neq n} 0 \cdot P_m = P_n$ then, according to the Martingale property [11b], applied at $t=\infty, P_n = x_n(0)$: hence, the ensemble of runs produces the probability postulated by the usual "collapse rule" of standard quantum theory.

A (nonquantum) stochastic process which obeys these equations is the gambler's ruin game. Suppose one gambler initially possesses the fraction $x_1(0)$ of their joint wealth, and the other has the fraction $x_2(0)$. They toss a coin: heads, a dollar goes from gambler 1 to gambler 2, tails the dollar goes the other way. [11a] is satisfied since the sum of money in the game remains constant, [11b] holds because it is a fair game, and [11c] holds because each game eventually ends. Thus, gambler i wins all the money with probability $x_i(0)$.

CSL in Essence

Consider the (nonunitary) Schrödinger picture evolution equation

$$|\psi,t\rangle_w = \mathcal{T}\exp\left(-\int_0^l \mathrm{d}t'\{i\hat{H} + (4\lambda)^{-1}[w(t') - 2\lambda\hat{A}]^2\}\right)|\psi,0\rangle \qquad [12]$$

where $\hat{H}$ is the usual Hamiltonian, $w(t')$ is an arbitrary function of white noise class, $\hat{A}$ is a Hermitian operator $(\hat{A}|a_n\rangle = a_n|a_n\rangle)$, λ is a collapse rate parameter, $\mathcal{T}$ is the time-ordering operator and $\hbar = 1$. Associated with this, the probability rule

$$P_t(w)Dw \equiv {}_w\langle\psi,t|\psi,t\rangle_w \prod_{j=0}^{t/\mathrm{d}t} \mathrm{d}w(t_j)/(2\pi\lambda/\mathrm{d}t)^{1/2} \qquad [13]$$

is defined, which gives the probability that nature chooses a noise which lies in the range $\{w(t'), w(t') + \mathrm{d}w(t')\}$ for $0 \leq t' \leq t$ (for calculational purposes, time is discretized, with $t_0 = 0$).

Equations [12] and [13] contain the essential features of CSL, and are all that is needed to discuss the simplest collapse behavior. Set $\hat{H}=0$, so there is no competition between collapse and the usual Schrödinger evolution, and let the initial state vector be $|\psi,0\rangle = \sum_n \alpha_n|a_n\rangle$. Equations [12] and [13] become

$$|\psi,t\rangle_w = \sum_n \alpha_n|a_n\rangle \exp\left(-(4\lambda)^{-1}\int_0^l \mathrm{d}t'[w(t') - 2\lambda a_n]^2\right) \qquad [14a]$$

$$P_t(w) = \sum_n |\alpha_n|^2 \exp\left(-(2\lambda)^{-1}\int_0^l \mathrm{d}t'[w(t') - 2\lambda a_n]^2\right) \qquad [14b]$$

When the unnormalized state vector in [14a] is divided by $P_t^{1/2}(w)$ and so normalized, the squared amplitudes are

$$x_n(t) = |\alpha_n|^2 \exp\left(-(2\lambda)^{-1} \times \int_0^t \mathrm{d}t'[w(t') - 2\lambda a_n]^2\right)/P_t(w)$$

which are readily shown to satisfy [11a], [11b], and [11c] in the form $\overline{x_n^{1/2}(\infty)x_m^{1/2}(\infty)} = 0 (m \neq n)$ (which does not change the argument in the last subsection, but makes for an easier calculation). Thus, [14a] and [14b] describe collapse dynamics.

To describe collapse to a joint eigenstate of a set of mutually commuting operators $\hat{A}^r$, replace $(4\lambda)^{-1}[w(t') - 2\lambda\hat{A}]^2$ in the exponent of [12] by $\sum_r (4\lambda)^{-1}[w^r(t') - 2\lambda\hat{A}^r]^2$. The interaction picture state vector in this case is [12] multiplied by $\exp(\mathrm{i}\hat{H}t)$:

$$|\psi, t\rangle_w = \mathcal{T}\exp\left(-(4\lambda)^{-1}\int_0^l \mathrm{d}t' \times \sum_r [w^r(t') - 2\lambda\hat{A}^r(t')]^2\right)|\psi, 0\rangle \quad [15]$$

where $\hat{A}^r(t') \equiv \exp(\mathrm{i}\hat{H}t')\hat{A}^r\exp(-\mathrm{i}\hat{H}t')$. The density matrix follows from [15], and [13]:

$$\begin{aligned}\hat{\rho}(t) &\equiv \int P_t(w)Dw|\psi, t\rangle_{ww}\langle\psi, t|/P_t(w) \\ &= \mathcal{T}\exp\left(-\lambda/2\int_0^t \mathrm{d}t' \times \sum_r [\hat{A}^r_{\mathrm{L}}(t') - \hat{A}^r_{\mathrm{R}}(t')]^2\right)\hat{\rho}(0) \quad [16]\end{aligned}$$

where $\hat{A}^r_{\mathrm{L}}(t')(\hat{A}^r_{\mathrm{R}}(t'))$ appears to the left (right) of $\hat{\rho}(0)$, and is time-ordered (time reverse-ordered). In the example described by [14], the density matrix [16] is

$$\hat{\rho}(t) = \sum_{n,m} \mathrm{e}^{-(\lambda t/2)(a_n - a_m)^2}\alpha_n\alpha_m^*|a_n\rangle\langle a_m|$$

which encapsulates the ensemble's collapse behavior.

CSL

The CSL proposal (Pearle 1989) is that collapse is engendered by distinctions between states at each point of space, so the index r of $\hat{A}^r$ in [15] becomes $\boldsymbol{x}$,

$$|\psi, t\rangle_w = \mathcal{T}\exp\left(-(4\lambda)^{-1}\int_0^t\int \mathrm{d}t'\,\mathrm{d}\boldsymbol{x}' \times [w(\boldsymbol{x}', t') - 2\lambda\hat{A}(\boldsymbol{x}', t')]^2\right)|\psi, 0\rangle \quad [17]$$

and the distinction looked at is mass density. However, one cannot make the choice $\hat{A}(\boldsymbol{x}, 0) = \hat{M}(\boldsymbol{x})$, where $\hat{M}(\boldsymbol{x}) = \sum_i m_i\hat{\xi}_i^\dagger(\boldsymbol{x})\hat{\xi}_i(\boldsymbol{x})$ is the mass-density operator (m_i is the mass of the ith type of particle, so $m_{\mathrm{e}}, m_{\mathrm{p}}, m_{\mathrm{n}}, \ldots$ are the masses, respectively, of electrons, protons, neutrons…, and $\hat{\xi}_i^\dagger(\boldsymbol{x})$ is the creation operator for such a particle at location $\boldsymbol{x}$), because this entails an infinite rate of energy increase of particles ([23] with $a = 0$). Instead, adapting a "Gaussian smearing" idea from the Ghirardi *et al.* (1986) spontaneous localization (SL) model (see the subsection "Spontaneous localization model"), choose $\hat{A}^{\boldsymbol{x}}$ as, essentially, proportional to the mass in a sphere of radius a about $\boldsymbol{x}$:

$$\hat{A}(\boldsymbol{x}, t) \equiv \mathrm{e}^{\mathrm{i}\hat{H}t}\frac{1}{(\pi a^2)^{3/4}} \times \int \mathrm{d}z\frac{\hat{M}(z)}{m_{\mathrm{p}}}\mathrm{e}^{-(2a^2)^{-1}(x-z)^2}\mathrm{e}^{-\mathrm{i}\hat{H}t} \quad [18]$$

The parameter value choices of SL, $\lambda \approx 10^{-16}\,\mathrm{s}^{-1}$ (according to [17] and [18], the collapse rate for protons) and $a \approx 10^{-5}$ cm are, so far, consistent with experiment (see the next subsection), and will be adopted here.

The density matrix associated with [17] is, as in [16],

$$\hat{\rho}(t) = \mathcal{T}\exp\left(-(\lambda/2)\int_0^t \mathrm{d}t'\,\mathrm{d}\boldsymbol{x}'[\hat{A}_{\mathrm{L}}(\boldsymbol{x}', t') - \hat{A}_{\mathrm{R}}(\boldsymbol{x}', t')]^2\right)\hat{\rho}(0) \quad [19]$$

which satisfies the differential equation

$$\frac{\mathrm{d}\hat{\rho}(t)}{\mathrm{d}t} = -\frac{\lambda}{2}\int \mathrm{d}\boldsymbol{x}'[\hat{A}(\boldsymbol{x}', t), [\hat{A}(\boldsymbol{x}', t), \hat{\rho}(t)]] \quad [20]$$

of Lindblad–Kossakowski form.

Consequences of CSL

Since the state vector dynamics of CSL is different from that of standard quantum theory, there are phenomena for which the two make different predictions, allowing for experimental tests. Consider an N-particle system with position operators $\hat{X}_i(\hat{X}_i|\boldsymbol{x}\rangle = x_i|\boldsymbol{x}\rangle)$. Substitution of $\hat{A}(\boldsymbol{x}')$ from [18] in the Schrödinger picture version of [20], integration over $\boldsymbol{x}'$, and utilization of

$$f(z)\hat{M}(z)|\boldsymbol{x}\rangle = \sum_{i=1}^N m_i f(\hat{\boldsymbol{X}}_i)\delta(z - \hat{\boldsymbol{X}}_i)|\boldsymbol{x}\rangle$$

results in

$$\begin{aligned}\frac{\mathrm{d}\hat{\rho}(t)}{\mathrm{d}t} = -\mathrm{i}[\hat{\rho}(t), \hat{H}] - \frac{\lambda}{2}\sum_{i=1}^N\sum_{j=1}^N\frac{m_i}{m_{\mathrm{p}}}\frac{m_j}{m_{\mathrm{p}}} \\ \times\left[\mathrm{e}^{-(4a^2)^{-1}(\hat{\boldsymbol{X}}_{\mathrm{L}i} - \hat{\boldsymbol{X}}_{\mathrm{L}j})^2} + \mathrm{e}^{-(4a^2)^{-1}(\hat{\boldsymbol{X}}_{\mathrm{R}i} - \hat{\boldsymbol{X}}_{\mathrm{R}j})^2}\right. \\ \left. - 2\mathrm{e}^{-(4a^2)^{-1}(\hat{\boldsymbol{X}}_{\mathrm{L}i} - \hat{\boldsymbol{X}}_{\mathrm{R}j})^2}\right]\hat{\rho}(t) \quad [21]\end{aligned}$$

which is a useful form for calculations first suggested by Pearle and Squires in 1994.

Interference Consider the collapse rate of an initial state $|\phi\rangle = \alpha_1|1\rangle + \alpha_2|2\rangle$, where $|1\rangle, |2\rangle$ describe a

clump of matter, of size $\ll a$, at different locations with separation $\gg a$. Electrons may be neglected because of their small collapse rate compared to the much more massive nucleons, and the nucleon mass difference may be neglected. In using [21] to calculate $d\langle 1|\hat{\rho}(t)|2\rangle/dt$, since $\exp[-(4a^2)^{-1}(\hat{X}_i - \hat{X}_j)^2] \approx 1$ when acting on state $|1\rangle$ or $|2\rangle$, and ≈ 0 when $\hat{X}_i$ acts on $|1\rangle$ and $\hat{X}_j$ acts on $|2\rangle$, [21] yields, for N nucleons, the collapse rate λN^2:

$$\frac{d\langle 1|\hat{\rho}(t)|2\rangle}{dt} = -i\langle 1|[\hat{\rho}(t), \hat{H}]|2\rangle - \lambda N^2 \langle 1|\hat{\rho}(t)|2\rangle \quad [22]$$

If the clump undergoes a two-slit interference experiment, where the size and separation conditions above are satisfied for a time ΔT, and if the result agrees with the standard quantum theory prediction to 1%, it also agrees with CSL provided $\lambda^{-1} > 100N^2\Delta T$. So far, interference experiments with N as large as $\approx 10^3$ have been performed, by Nairz, Arndt, and Zeilinger in 2000. The SL value of $\lambda^{-1} \approx 10^{16}$ would be testable, that is, the quantum-predicted interference pattern would be "washed out" to 1% accuracy, if the clump were an $\approx 10^{-6}$ cm radius sphere of mercury, which contains $N \approx 10^8$ nucleons, interfered for $\Delta T = 0.01$ s. Currently envisioned but not yet performed experiments (e.g., by Marshall, Simon, Penrose, and Bouwmester in 2003) have been analyzed (e.g., by Bassi, Ippoliti, and Adler in 2004 and by Adler in 2005), which involve a superposition of a larger clump of matter in slightly displaced positions, entangled with a photon whose interference pattern is measured: these proposed experiments are still too crude to detect the SL value of λ, or the gravitationally based collapse rate proposed by Penrose in 1996 (see the next section and papers by Christian in 1999 and 2005).

Bound state excitation Collapse narrows wave packets, thereby imparting energy to particles. If $\hat{H} = \sum_{i=1}^{N} \hat{P}_i^2/2m_i + \hat{V}(x_1, \ldots, x_N)$, it is straightforward to calculate from [21] that

$$\frac{d}{dt}\langle \hat{H}\rangle \equiv \frac{d}{dt}\mathrm{tr}[\hat{H}\hat{\rho}(t)] = \sum_{i=1}^{N} \frac{3\lambda\hbar^2}{4m_i a^2} \quad [23]$$

For a nucleon, the mean rate of energy increase is quite small, $\approx 3 \times 10^{-25}$ eV s^{-1}. However, deviations from the mean can be significantly greater.

Equation [21] predicts excitation of atoms and nuclei. Let $|E_0\rangle$ be an initial bound energy eigenstate. Expanding [21] in a power series in (bound state size/a)^2, the excitation rate of state $|E\rangle$ is

$$\Gamma \equiv \frac{d\langle E|\hat{\rho}(t)|E\rangle}{dt}\Big|_{t=0} = \frac{\lambda}{2a^2}\left\langle E\Big|\sum_{i=1}^{N}\frac{m_i}{m_p}\hat{X}_i\Big|E_0\right\rangle\left\langle E_0\Big|\sum_{i=1}^{N}\frac{m_i}{m_p}\hat{X}_i\Big|E\right\rangle + O(\text{size}/a)^4 \quad [24]$$

Since $|E_0\rangle, |E\rangle$ are eigenstates of the center-of-mass operator $\sum_{i=1}^{N} m_i\hat{X}_i/\sum_{i=1}^{N} m_i$ with eigenvalue 0, the dipole contribution explicitly given in [24] vanishes identically. This leaves the quadrupole contribution as the leading term, which is too small to be measured at present.

However, the choice of $\hat{A}(x)$ as mass-density operator was made only after experimental indication. Let g_i replace m_i/m_p in [21] and [24], so that λg_i^2 is the collapse rate for the ith particle. Then, experiments looking for the radiation expected from "spontaneously" excited atoms and nuclei, in large amounts of matter for a long time, as shown by Collett, Pearle, Avignone, and Nussinov in 1995, Pearle, Ring, Collar, and Avignone in 1999, and Jones, Pearle, and Ring in 2004, have placed the following limits:

$$\left|\frac{g_e}{g_p} - \frac{m_e}{m_p}\right| < \frac{12m_e}{m_p}, \quad \left|\frac{g_n}{g_p} - \frac{m_n}{m_p}\right| < \frac{3(m_n - m_p)}{m_p}$$

Random walk According to [17] and [13], the center-of-mass wave packet, of a piece of matter of size $\approx a$ or smaller, containing N nucleons, achieves equilibrium size s in a characteristic time τ_s, and undergoes a random walk through a root-mean-square distance ΔQ:

$$s \approx \left[\frac{a^2\hbar}{\lambda m_p N^3}\right]^{1/4}, \qquad \tau_s \approx \frac{Nm_p s^2}{\hbar}$$
$$\Delta Q \approx \frac{\hbar\lambda^{1/2}t^{3/2}}{m_p a} \quad [25]$$

The results in [25] were obtained by Collett and Pearle in 2003. These quantitative results can be qualitatively understood as follows.

In time Δt, the usual Schrödinger equation expands a wave packet of size s to $\approx s + (\hbar/Nm_p s)\Delta t$. CSL collapse, by itself, narrows the wave packet to $\approx s[1 - \lambda N^2(s/a)^2\Delta t]$. The condition of no change in s is the result quoted above. τ_s is the time it takes the Schrödinger evolution to expand a wave packet near size s to size s: $(\hbar/Nm_p s)\tau_s \approx s$.

The $t^{3/2}$ dependence of ΔQ arises because this is a random walk without damping (unlike Brownian motion, where $\Delta Q \sim t^{1/2}$). The mean energy increase $\approx \lambda N \hbar^2 m_p^{-1} a^{-2} t$ of [23] implies the root-mean-square velocity increase $\approx [\lambda \hbar^2 m_p^{-2} a^{-2} t]^{1/2}$, whose product with t is ΔQ.

For example, a sphere of density $1\,\mathrm{cm}^{-3}$ and radius 10^{-5} cm has $s \approx 4 \times 10^{-7}$ cm, $\tau_s \approx 0.6$ s and $\Delta Q \approx 5[t \text{ in days}]^{3/2}$ cm. At the low pressure of 5×10^{-17} torr at 4.2 K reported by Gabrielse's group in 1990, the mean collision time with gas molecules is $\approx$80 min, over which $\Delta Q \approx 0.7$ mm. Thus, observation of this effect should be feasible.

Further Remarks

It is possible to define energy for the $w(\boldsymbol{x}, t)$ field so that total energy is conserved: as the particles gain energy, the w-field loses energy, as shown by Pearle in 2005.

Attempts to construct a special-relativistic CSL-type model have not yet succeeded, although Pearle in 1990, 1992, and 1999, Ghirardi, Grassi, and Pearle in 1990, and Nicrosini and Rimini in 2003 have made valiant attempts. The problem is that the white noise field $w(\boldsymbol{x}, t)$ contains all wavelengths and frequencies, exciting the vacuum in lowest order in λ to produce particles at the unacceptable rate of infinite energy/per second per cubic centimeter. Collapse models which utilize a colored noise field w have a similar problem in higher orders. In 2005, Pearle suggested a quasi-relativistic model which reduces to CSL in the low-speed limit.

CSL is a phenomenological model which describes dynamical collapse so as to achieve S. Besides needing decisive experimental verification, it needs identification of the $w(\boldsymbol{x}, t)$ field with a physical entity.

Other collapse models which have been investigated are briefly described below.

Spontaneous Localization Model

The SL model of Ghirardi *et al.* (1986), although superseded by CSL, is historically important and conceptually valuable. Let $\hat{H} = 0$ for simplicity, and consider a single particle whose wave function at time t is $\psi(\boldsymbol{x}, t)$. Over the next interval $\mathrm{d}t$, with probability $1 - \lambda \mathrm{d}t$, it does not change. With probability $\lambda \mathrm{d}t$ it does change, by being "spontaneously localized" or "hit." A hit means that the new (unnormalized) wave function suddenly becomes

$$\psi(\boldsymbol{x}, t + \mathrm{d}t) = \psi(\boldsymbol{x}, t)(\pi a^2)^{-3/4}\, \mathrm{e}^{-(2a^2)^{-1}(\boldsymbol{x}-\boldsymbol{z})^2}$$

with probability

$$\lambda \mathrm{d}t\, \mathrm{d}\boldsymbol{z} \int \mathrm{d}\boldsymbol{x} |\psi(\boldsymbol{x}, t + \mathrm{d}t)|^2$$

Thus $\boldsymbol{z}$, the "center" of the hit, is most likely to be located where the wave function is large. For a single particle in the superposition described in the subsection "Interference," a single hit is overwhelmingly likely to reduce the wave function to one or the other location, with total probability $|\alpha_i|^2$, at the rate λ.

For an N-particle clump, it is considered that each particle has the same independent probability, $\lambda \mathrm{d}t$, of being hit. But, for the example in the subsection "Interference," a single hit on any particle in one location of the clump has the effect of multiplying the wave function part describing the clump in the other location by the tail of the Gaussian, thereby collapsing the wave function at the rate λN.

By use of the Gaussian hit rather than a delta-function hit, SL solves the problem of giving too much energy to particles as mentioned in the subsection "CSL." By the hypothesis of independent particle hits, SL also solves the problem of achieving a slow collapse rate for a superposition of small objects and a fast collapse rate for a superposition of large objects. However, the hits on individual particles destroys the (anti-) symmetry of wave functions. The CSL collapse toward mass density eigenstates removes that problem. Also, while SL modifies the Schrödinger evolution of a wave function, it involves discontinuous dynamics and so is not described by a modified Schrödinger equation as is CSL.

Other Models

For a single (low-energy) particle, the polar decomposition $\Psi = R\mathrm{e}^{(\mathrm{i}/\hbar)S}$ of the Schrödinger equation implies two real equations,

$$\frac{\partial R^2}{\partial t} + \boldsymbol{\nabla} \cdot \left(R^2 \frac{\boldsymbol{\nabla} S}{m} \right) = 0 \qquad [26]$$

(the continuity equation for $R^2 = |\Psi|^2$) and

$$\frac{\partial S}{\partial t} + \frac{(\boldsymbol{\nabla} S)^2}{2m} + V + Q = 0 \qquad [27]$$

where $Q \equiv -(\hbar^2/2m)\nabla^2 R/R$ is the "quantum potential." (These equations have an obvious generalisation to higher-dimensional configuration space.) In 1926, Madelung proposed that one should start from [26] and [27] – regarded as hydrodynamical equations for a classical charged fluid with mass density mR^2 and fluid velocity $\boldsymbol{\nabla} S/m$ – and construct $\Psi = R\mathrm{e}^{(\mathrm{i}/\hbar)S}$ from the solutions.

This "hydrodynamical" interpretation suffers from many difficulties, especially for many-body systems. In any case, a criticism by Wallstrom (1994) seems decisive: [26] and [27] (and their higher-dimensional analogs) are not, in fact, equivalent to the Schrödinger equation. For, as usually understood, the quantum wave function Ψ is a single-valued and continuous complex field, which typically possesses nodes ($\Psi = 0$), in the neighborhood of which the phase S is multivalued, with values differing by integral multiples of $2\pi\hbar$. If one allows S in [26], [27] to be multivalued, there is no reason why the allowed values should differ by integral multiples of $2\pi\hbar$, and in general Ψ will not be single-valued. On the other hand, if one restricts S in [26], [27] to be single-valued, one will exclude wave functions – such as those of nonzero angular momentum – with a multivalued phase. (This problem does not exist in pilot-wave theory as we have presented it here, where Ψ is regarded as a basic entity.)

Stochastic mechanics, introduced by Fényes in 1952 and Nelson (1966), has particle trajectories $\boldsymbol{x}(t)$ obeying a "forward" stochastic differential equation $\mathrm{d}\boldsymbol{x}(t) = \boldsymbol{b}(\boldsymbol{x}(t), t)\mathrm{d}t + \mathrm{d}\boldsymbol{w}(t)$, where $\boldsymbol{b}$ is a drift (equal to the mean forward velocity) and $\boldsymbol{w}$ a Wiener process, and also a similar "backward" equation. Defining the "current velocity" $\boldsymbol{v} = (1/2)(\boldsymbol{b} + \boldsymbol{b}_*)$, where $\boldsymbol{b}_*$ is the mean backward velocity, and using an appropriate time-symmetric definition of mean acceleration, one may impose a stochastic version of Newton's second law. If one assumes, in addition, that $\boldsymbol{v}$ is a gradient ($\boldsymbol{v} = \nabla S/m$ for some S), then one obtains [26], [27] with $R \equiv \sqrt{\rho}$, where ρ is the particle density. Defining $\Psi \equiv \sqrt{\rho}\mathrm{e}^{(\mathrm{i}/\hbar)S}$, it appears that one recovers the Schrödinger equation for the derived quantity Ψ. However, again, there is no reason why S should have the specific multivalued structure required for the phase of a single-valued complex field. It then seems that, despite appearances, quantum theory cannot in fact be recovered from stochastic mechanics (Wallstrom 1994). The same problem occurs in models that use stochastic mechanics as an intermediate step (e.g., Markopoulou and Smolin in 2004): the Schrödinger equation is obtained only for exceptional, nodeless wave functions.

Bohm and Bub (1966) first proposed dynamical wave-function collapse through deterministic evolution. Their collapse outcome is determined by the value of a Wiener–Siegel hidden variable (a variable distributed uniformly over the unit hypersphere in a Hilbert space identical to that of the state vector). In 1976, Pearle proposed dynamical wave-function collapse equations where the collapse outcome is determined by a random variable, and suggested (Pearle 1979) that the modified Schrödinger equation be formulated as an Itô stochastic differential equation, a suggestion which has been widely followed. (The equation for the state vector given here, which is physically more transparent, has its time derivative equivalent to a Stratonovich stochastic differential equation, which is readily converted to the Itô form.) The importance of requiring that the density matrix describing collapse be of the Lindblad–Kossakowski form was emphasized by Gisin in 1984 and Diosi in 1988. The stochastic differential Schrödinger equation that achieves this was found independently by Diosi in 1988 and by Belavkin, Gisin, and Pearle in separate papers in 1989 (see Ghirardi *et al.* 1990).

A gravitationally motivated stochastic collapse dynamics was proposed by Diosi in 1989 (and somewhat corrected by Ghirardi *et al.* in 1990). Penrose emphasized in 1996 that a quantum state, such as that describing a mass in a superposition of two places, puts the associated spacetime geometry also in a superposition, and has argued that this should lead to wave-function collapse. He suggests that the collapse time should be $\sim\hbar/\Delta E$, where ΔE is the gravitational potential energy change obtained by actually displacing two such masses: for example, the collapse time $\approx\hbar/(Gm^2/R)$, where the mass is m, its size is R, and the displacement is $\approx R$ or larger. No specific dynamics is offered, just the vision that this will be a property of a correct future quantum theory of gravity.

Collapse to energy eigenstates was first proposed by Bedford and Wang in 1975 and 1977 and, in the context of stochastic collapse (e.g., [11] with $\hat{A} = \hat{H}$), by Milburn in 1991 and Hughston in 1996, but it has been argued by Finkelstein in 1993 and Pearle in 2004 that such energy-driven collapse cannot give a satisfactory picture of the macroscopic world. Percival in 1995 and in a 1998 book, and Fivel in 1997 have discussed energy-driven collapse for microscopic situations.

Adler (2004) has presented a classical theory (a hidden-variables theory) from which it is argued that quantum theory "emerges" at the ensemble level. The classical variables are $N \times N$ matrix field amplitudes at points of space. They obey appropriate classical Hamiltonian dynamical equations which he calls "trace dynamics," since the expressions for Hamiltonian, Lagrangian, Poisson bracket, etc., have the form of the trace of products of matrices and their sums with constant coefficients. Using classical statistical mechanics, canonical ensemble averages of (suitably projected) products of fields are analyzed and it is argued that they obey all the properties associated with Wightman functions, from which quantum field theory, and its nonrelativistic-limit quantum mechanics, may be derived. As well as obtaining the algebra of quantum theory in this way,

it is argued that statistical fluctuations around the canonical ensemble can give rise to the behavior of wave-function collapse, of the kind discussed here, both energy-driven and CSL-type mass-density-driven collapse so that, with the latter, comes the Born probability interpretation of the algebra. The Hamiltonian needed for this theory to work is not provided but, as the argument progresses, its necessary features are delimited.

See also: Quantum Mechanics: Foundations.

Further Reading

Adler SL (2004) *Quantum Theory as an Emergent Phenomenon*. Cambridge: Cambridge University Press.

Bassi A and Ghirardi GC (2003) Dynamical reduction models. *Physics Reports* 379: 257–426 (ArXiv: quant-ph/0302164).

Bell JS (1987) *Speakable and Unspeakable in Quantum Mechanics*. Cambridge: Cambridge University Press.

Bohm D (1952) A suggested interpretation of the quantum theory in terms of 'hidden' variables. I and II. *Physical Review* 85: 166–179 and 180–193.

Bohm D and Bub J (1966) A proposed solution of the measurement problem in quantum mechanics by a hidden variable theory. *Reviews of Modern Physics* 38: 453–469.

Dürr D, Goldstein S, and Zanghì N (2003) Quantum equilibrium and the role of operators as observables in quantum theory, ArXiv: quant-ph/0308038.

Ghirardi G, Rimini A, and Weber T (1986) Unified dynamics for microscopic and macroscopic systems. *Physical Review D* 34: 470–491.

Ghirardi G, Pearle P, and Rimini A (1990) Markov processes in Hilbert space and continuous spontaneous localization of systems of identical particles. *Physical Review A* 42: 78–89.

Holland PR (1993) *The Quantum Theory of Motion: An Account of the de Broglie–Bohm Causal Interpretation of Quantum Mechanics*. Cambridge: Cambridge University Press.

Nelson E (1966) Derivation of the Schrödinger equation from Newtonian mechanics. *Physical Review* 150: 1079–1085.

Pearle P (1979) Toward explaining why events occur. *International Journal of Theoretical Physics* 18: 489–518.

Pearle P (1989) Combining stochastic dynamical state-vector reduction with spontaneous localization. *Physical Review A* 39: 2277–2289.

Pearle P (1999) Collapse models. In: Petruccione F and Breuer HP (eds.) *Open Systems and Measurement in Relativistic Quantum Theory*, pp. 195–234. Heidelberg: Springer. (ArXiv: quant-ph/9901077).

Valentini A (1991) Signal-locality, uncertainty, and the subquantum *H*-theorem. I and II. *Physics Letters A* 156: 5–11 and 158: 1–8.

Valentini A (2002a) Signal-locality in hidden-variables theories. *Physics Letters A* 297: 273–278.

Valentini A (2002b) Subquantum information and computation. *Pramana – Journal of Physics* 59: 269–277 (ArXiv: quant-ph/0203049).

Valentini A (2004a) Black holes, information loss, and hidden variables, ArXiv: hep-th/0407032.

Valentini A (2004b) Universal signature of non-quantum systems. *Physics Letters A* 332: 187–193 (ArXiv: quant-ph/0309107).

Valentini A and Westman H (2005) Dynamical origin of quantum probabilities. *Proceedings of the Royal Society of London Series A* 461: 253–272.

Wallstrom TC (1994) Inequivalence between the Schrödinger equation and the Madelung hydrodynamic equations. *Physical Review A* 49: 1613–1617.

Quantum Mechanics: Weak Measurements

L Diósi, Research Institute for Particle and Nuclear Physics, Budapest, Hungary

Introduction

In quantum theory, the mean value of a certain observable $\hat{A}$ in a (pure) quantum state $|i\rangle$ is defined by the quadratic form:

$$\langle \hat{A} \rangle_i =: \langle i|\hat{A}|i\rangle \qquad [1]$$

Here $\hat{A}$ is Hermitian operator on the Hilbert space $\mathcal{H}$ of states. We use Dirac formalism. The above mean is interpreted statistically. No other forms had been known to possess a statistical interpretation in standard quantum theory. One can, nonetheless, try to extend the notion of mean for normalized bilinear expressions (Aharonov *et al.* 1988):

$$A_{\mathrm{w}} =: \frac{\langle f|\hat{A}|i\rangle}{\langle f|i\rangle} \qquad [2]$$

However unusual is this structure, standard quantum theory provides a plausible statistical interpretation for it, too. The two pure states $|i\rangle, |f\rangle$ play the roles of the prepared initial and the postselected final states, respectively. The statistical interpretation relies upon the concept of weak measurement. In a single weak measurement, the notorious decoherence is chosen asymptotically small. In physical terms, the coupling between the measured state and the meter is assumed asymptotically weak. The novel mean value [2] is called the (complex) weak value.

The concept of quantum weak measurement (Aharonov *et al.* 1988) provides particular

conclusions on postselected ensembles. Weak measurements have been instrumental in the interpretation of time-continuous quantum measurements on single states as well. Yet, weak measurement itself can properly be illuminated in the context of classical statistics. Classical weak measurement as well as postselection and time-continuous measurement are straightforward concepts leading to conclusions that are natural in classical statistics. In quantum context, the case is radically different and certain paradoxical conclusions follow from weak measurements. Therefore, we first introduce the classical notion of weak measurement on postselected ensembles and, alternatively, in time-continuous measurement on a single state. Certain idioms from statistical physics will be borrowed and certain not genuinely quantum notions from quantum theory will be anticipated. The quantum counterpart of weak measurement, postselection, and continuous measurement will be presented afterwards. The apparent redundancy of the parallel presentations is of reason: the reader can separate what is common in classical and quantum weak measurements from what is genuinely quantum.

Classical Weak Measurement

Given a normalized probability density $\rho(X)$ over the phase space $\{X\}$, which we call the state, the mean value of a real function $A(X)$ is defined as

$$\langle A\rangle_\rho =: \int \mathrm{d}X\, A\rho \qquad [3]$$

Let the outcome of an (unbiased) measurement of A be denoted by a. Its stochastic expectation value $E[a]$ coincides with the mean [3]:

$$E[a] = \langle A\rangle_\rho \qquad [4]$$

Performing a large number N of independent measurements of A on the elements of the ensemble of identically prepared states, the arithmetic mean $\bar{a}$ of the outcomes yields a reliable estimate of $E[a]$ and, this way, of the theoretical mean $\langle A\rangle_\rho$.

Suppose, for concreteness, the measurement outcome a is subject to a Gaussian stochastic error of standard dispersion $\sigma > 0$. The probability distribution of a and the update of the state corresponding to the Bayesian inference are described as

$$p(a) = \langle G_\sigma(a - A)\rangle_\rho \qquad [5]$$

$$\rho \to \frac{1}{p(a)} G_\sigma(a - A)\rho \qquad [6]$$

respectively. Here G_σ is the central Gaussian distribution of variance σ. Note that, as expected, eqn [5] implies eqn [4]. Nonzero σ means that the measurement is nonideal, yet the expectation value $E[a]$ remains calculable reliably if the statistics N is suitably large.

Suppose the spread of A in state ρ is finite:

$$\Delta_\rho^2 A =: \langle A^2\rangle_\rho - \langle A\rangle_\rho^2 < \infty \qquad [7]$$

Weak measurement will be defined in the asymptotic limit (eqns [8] and [9]) where both the stochastic error of the measurement and the measurement statistics go to infinity. It is crucial that their rate is kept constant:

$$\sigma, N \to \infty \qquad [8]$$

$$\Delta^2 =: \frac{\sigma^2}{N} = \text{const.} \qquad [9]$$

Obviously for asymptotically large σ, the precision of individual measurements becomes extremely weak. This incapacity is fully compensated by the asymptotically large statistics N. In the weak measurement limit (eqns [8] and [9]), the probability distribution p_w of the arithmetic mean $\bar{a}$ of the N independent outcomes converges to a Gaussian distribution:

$$p_\mathrm{w}(\bar{a}) \to G_\Delta\left(\bar{a} - \langle A\rangle_\rho\right) \qquad [10]$$

The Gaussian is centered at the mean $\langle A\rangle_\rho$, and the variance of the Gaussian is given by the constant rate [9]. Consequently, the mean [3] is reliably calculable on a statistics N growing like $\sim\sigma^2$.

With an eye on quantum theory, we consider two situations – postselection and time-continuous measurement – of weak measurement in classical statistics.

Postselection

For the preselected state ρ, we introduce postselection via the real function $\Pi(X)$, where $0 \le \Pi \le 1$. The postselected mean value of a certain real function $A(X)$ is defined by

$${}_\Pi\langle A\rangle_\rho =: \frac{\langle \Pi A\rangle_\rho}{\langle \Pi\rangle_\rho} \qquad [11]$$

where $\langle \Pi\rangle_\rho$ is the rate of postselection. Postselection means that after having obtained the outcome a regarding the measurement of A, we measure the function Π, too, in ideal measurement with random outcome π upon which we base the following random decision. With probability π, we include the current a into the statistics and we discard it

with probability $1-\pi$. Then the coincidence of $E[a]$ and ${}_{\Pi}\langle A\rangle_{\rho}$, as in eqn [4], remains valid:

$$E[a] = {}_{\Pi}\langle A\rangle_{\rho} \qquad [12]$$

Therefore, a large ensemble of postselected states allows one to estimate the postselected mean ${}_{\Pi}\langle A\rangle_{\rho}$.

Classical postselection allows introducing the effective postselected state:

$$\rho_{\Pi} =: \frac{\Pi\rho}{\langle\Pi\rangle_{\rho}} \qquad [13]$$

Then the postselected mean [11] of A in state ρ can, by eqn [14], be expressed as the common mean of A in the effective postselected state ρ_{Π}:

$${}_{\Pi}\langle A\rangle_{\rho} = \langle A\rangle_{\rho_{\Pi}} \qquad [14]$$

As we shall see later, quantum postselection is more subtle and cannot be reduced to common statistics, that is, to that without postselection. The quantum counterpart of postselected mean does not exist unless we combine postselection and weak measurement.

Time-Continuous Measurement

For time-continuous measurement, one abandons the ensemble of identical states. One supposes that a single time-dependent state ρ_t is undergoing an infinite sequence of measurements (eqns [5] and [6]) of A employed at times $t=\delta t, t=2\delta t, t=3\delta t, \ldots$. The rate $\nu =: 1/\delta t$ goes to infinity together with the mean squared error σ^2. Their rate is kept constant:

$$\sigma, \nu \to \infty \qquad [15]$$

$$g^2 =: \frac{\sigma^2}{\nu} = \text{const.} \qquad [16]$$

In the weak measurement limit (eqns [15] and [16]), the infinite frequent weak measurements of A constitute the model of time-continuous measurement. Even the weak measurements will significantly influence the original state ρ_0, due to the accumulated effect of the infinitely many Bayesian updates [6]. The resulting theory of time-continuous measurement is described by coupled Gaussian processes [17] and [18] for the primitive function α_t of the time-dependent measurement outcome and, respectively, for the time-dependent Bayesian conditional state ρ_t:

$$\mathrm{d}\alpha_t = \langle A\rangle_{\rho_t}\,\mathrm{d}t + g\,\mathrm{d}W_t \qquad [17]$$

$$\mathrm{d}\rho_t = g^{-1}\left(A - \langle A\rangle_{\rho_t}\right)\rho_t\,\mathrm{d}W_t \qquad [18]$$

Here $\mathrm{d}W_t$ is the Itô differential of the Wiener process.

Equations [17] and [18] are the special case of the Kushner–Stratonovich equations of time-continuous Bayesian inference conditioned on the continuous measurement of A yielding the time-dependent outcome value a_t. Formal time derivatives of both sides of eqn [17] yield the heuristic equation

$$a_t = \langle A\rangle_{\rho_t} + g\xi_t \qquad [19]$$

Accordingly, the current measurement outcome is always equal to the current mean plus a term proportional to standard white noise ξ_t. This plausible feature of the model survives in the quantum context as well. As for the other equation [18], it describes the gradual concentration of the distribution ρ_t in such a way that the variance $\Delta_{\rho_t}A$ tends to zero while $\langle A\rangle_{\rho_t}$ tends to a random asymptotic value. The details of the convergence depend on the character of the continuously measured function $A(X)$. Consider a stepwise $A(X)$:

$$A(X) = \sum_{\lambda} a^{\lambda} P^{\lambda}(X) \qquad [20]$$

The real values a^{λ} are step heights all differing from each other. The indicator functions P^{λ} take values 0 or 1 and form a complete set of pairwise disjoint functions on the phase space:

$$\sum_{\lambda} P^{\lambda} \equiv 1 \qquad [21]$$

$$P^{\lambda}P^{\mu} = \delta_{\lambda\mu}P^{\lambda} \qquad [22]$$

In a single ideal measurement of A, the outcome a is one of the a^{λ}'s singled out at random. The probability distribution of the measurement outcome and the corresponding Bayesian update of the state are given by

$$p^{\lambda} = \langle P^{\lambda}\rangle_{\rho_0} \qquad [23]$$

$$\rho_0 \to \frac{1}{p^{\lambda}}P^{\lambda}\rho_0 =: \rho^{\lambda} \qquad [24]$$

respectively. Equations [17] and [18] of time-continuous measurement are a connatural time-continuous resolution of the "sudden" ideal measurement (eqns [23] and [24]) in a sense that they reproduce it in the limit $t \to \infty$. The states ρ^{λ} are trivial stationary states of the eqn [18]. It can be shown that they are indeed approached with probability p^{λ} for $t \to \infty$.

Quantum Weak Measurement

In quantum theory, states in a given complex Hilbert space $\mathcal{H}$ are represented by non-negative density operators $\hat{\rho}$, normalized by $\mathrm{tr}\,\hat{\rho}=1$. Like the

classical states ρ, the quantum state $\hat{\rho}$ is interpreted statistically, referring to an ensemble of states with the same $\hat{\rho}$. Given a Hermitian operator $\hat{A}$, called observable, its theoretical mean value in state $\hat{\rho}$ is defined by

$$\langle \hat{A} \rangle_{\hat{\rho}} = \mathrm{tr}(\hat{A}\hat{\rho}) \qquad [25]$$

Let the outcome of an (unbiased) quantum measurement of $\hat{A}$ be denoted by a. Its stochastic expectation value $E[a]$ coincides with the mean [25]:

$$E[a] = \langle \hat{A} \rangle_{\hat{\rho}} \qquad [26]$$

Performing a large number N of independent measurements of $\hat{A}$ on the elements of the ensemble of identically prepared states, the arithmetic mean $\bar{a}$ of the outcomes yields a reliable estimate of $E[a]$ and, this way, of the theoretical mean $\langle \hat{A} \rangle_{\hat{\rho}}$. If the measurement outcome a contains a Gaussian stochastic error of standard dispersion σ, then the probability distribution of a and the update, called collapse in quantum theory, of the state are described by eqns [27] and [28], respectively. (We adopt the notational convenience of physics literature to omit the unit operator $\hat{I}$ from trivial expressions like $a\hat{I}$.)

$$p(a) = \left\langle G_\sigma(a - \hat{A}) \right\rangle_{\hat{\rho}} \qquad [27]$$

$$\hat{\rho} \to \frac{1}{p(a)} G_\sigma^{1/2}(a - \hat{A})\hat{\rho} G_\sigma^{1/2}(a - \hat{A}) \qquad [28]$$

Nonzero σ means that the measurement is nonideal, but the expectation value $E[a]$ remains calculable reliably if N is suitably large.

Weak quantum measurement, like its classical counterpart, requires finite spread of the observable $\hat{A}$ on state $\hat{\rho}$:

$$\Delta_{\hat{\rho}}^2 \hat{A} =: \langle \hat{A}^2 \rangle_{\hat{\rho}} - \langle \hat{A} \rangle_{\hat{\rho}}^2 < \infty \qquad [29]$$

Weak quantum measurement, too, will be defined in the asymptotic limit [8] introduced for classical weak measurement. Single quantum measurements can no more distinguish between the eigenvalues of $\hat{A}$. Yet, the expectation value $E[a]$ of the outcome a remains calculable on a statistics N growing like $\sim\sigma^2$.

Both in quantum theory and classical statistics, the emergence of nonideal measurements from ideal ones is guaranteed by general theorems. For completeness of this article, we prove the emergence of the nonideal quantum measurement (eqns [27] and [28]) from the standard von Neumann theory of ideal quantum measurements (von Neumann 1955). The source of the statistical error of dispersion σ is associated with the state $\hat{\rho}_M$ in the complex Hilbert space $\mathcal{L}^2$ of a hypothetic meter. Suppose $R \in (-\infty, \infty)$ is the position of the "pointer." Let its initial state $\hat{\rho}_M$ be a pure central Gaussian state of width σ; then the density operator $\hat{\rho}_M$ in Dirac position basis takes the form

$$\hat{\rho}_M = \int \mathrm{d}R \int \mathrm{d}R' G_\sigma^{1/2}(R) G_\sigma^{1/2}(R') |R\rangle\langle R'| \qquad [30]$$

We are looking for a certain dynamical interaction to transmit the "value" of the observable $\hat{A}$ onto the pointer position $\hat{R}$. To model the interaction, we define the unitary transformation [31] to act on the tensor space $\mathcal{H} \otimes \mathcal{L}^2$:

$$\hat{U} = \exp(\mathrm{i}\hat{A} \otimes \hat{K}) \qquad [31]$$

Here $\hat{K}$ is the canonical momentum operator conjugated to $\hat{R}$:

$$\exp(\mathrm{i}a\hat{K})|R\rangle = |R + a\rangle \qquad [32]$$

The unitary operator $\hat{U}$ transforms the initial uncorrelated quantum state into the desired correlated composite state:

$$\hat{\Sigma} =: \hat{U}\hat{\rho} \otimes \hat{\rho}_M \hat{U}^\dagger \qquad [33]$$

Equations [30]–[33] yield the expression [34] for the state $\hat{\Sigma}$:

$$\hat{\Sigma} = \int \mathrm{d}R \int \mathrm{d}R' G_\sigma^{1/2}(R - \hat{A})\hat{\rho} G_\sigma^{1/2} \times (R' - \hat{A}) \otimes |R\rangle\langle R'| \qquad [34]$$

Let us write the pointer's coordinate operator $\hat{R}$ into the standard form [35] in Dirac position basis:

$$\hat{R} = \int \mathrm{d}a |a\rangle\langle a| \qquad [35]$$

The notation anticipates that, when pointer $\hat{R}$ is measured ideally, the outcome a plays the role of the nonideally measured value of the observable $\hat{A}$. Indeed, let us consider the ideal von Neumann measurement of the pointer position on the correlated composite state $\hat{\Sigma}$. The probability of the outcome a and the collapse of the composite state are given by the following standard equations:

$$p(a) = \mathrm{tr}\left[(\hat{I} \otimes |a\rangle\langle a|)\hat{\Sigma}\right] \qquad [36]$$

$$\hat{\Sigma} \to \frac{1}{p(a)}\left[(\hat{I} \otimes |a\rangle\langle a|)\hat{\Sigma}(\hat{I} \otimes |a\rangle\langle a|)\right] \qquad [37]$$

respectively. We insert eqn [34] into eqns [36] and [37]. Furthermore, we take the trace over $\mathcal{L}^2$ of both sides of eqn [37]. In such a way, as expected, eqns [36] and [37] of ideal measurement of $\hat{R}$ yield the

earlier postulated eqns [27] and [28] of nonideal measurement of $\hat{A}$.

Quantum Postselection

A quantum postselection is defined by a Hermitian operator satisfying $\hat{0} \leq \hat{\Pi} \leq \hat{I}$. The corresponding postselected mean value of a certain observable $\hat{A}$ is defined by

$$ {}_{\hat{\Pi}}\langle \hat{A} \rangle_{\hat{\rho}} =: \mathrm{Re} \frac{\langle \hat{\Pi} \hat{A} \rangle_{\hat{\rho}}}{\langle \hat{\Pi} \rangle_{\hat{\rho}}} \qquad [38] $$

The denominator $\langle \hat{\Pi} \rangle_{\hat{\rho}}$ is the rate of quantum postselection. Quantum postselection means that after the measurement of $\hat{A}$, we measure the observable $\hat{\Pi}$ in ideal quantum measurement and we make a statistical decision on the basis of the outcome π. With probability π, we include the case in question into the statistics while we discard it with probability $1 - \pi$. By analogy with the classical case [12], one may ask whether the stochastic expectation value $E[a]$ of the postselected measurement outcome does coincide with

$$ E[a] \stackrel{?}{=} {}_{\hat{\Pi}}\langle \hat{A} \rangle_{\hat{\rho}} \qquad [39] $$

Contrary to the classical case, the quantum equation [39] does not hold. The quantum counterparts of classical equations [12]–[14] do not exist at all. Nonetheless, the quantum postselected mean ${}_{\hat{\Pi}}\langle \hat{A} \rangle_{\hat{\rho}}$ possesses statistical interpretation although restricted to the context of weak quantum measurements. In the weak measurement limit (eqns [8] and [9]), a postselected analog of classical equation [10] holds for the arithmetic mean $\bar{a}$ of postselected weak quantum measurements:

$$ p_{\mathrm{w}}(\bar{a}) \to G_{\Delta}\left(\bar{a} - {}_{\hat{\Pi}}\langle \hat{A} \rangle_{\hat{\rho}} \right) \qquad [40] $$

The Gaussian is centered at the postselected mean ${}_{\hat{\Pi}}\langle \hat{A} \rangle_{\hat{\rho}}$, and the variance of the Gaussian is given by the constant rate [9]. Consequently, the mean [38] becomes calculable on a statistics N growing like $\sim \sigma^2$.

Since the statistical interpretation of the postselected quantum mean [38] is only possible for weak measurements, therefore ${}_{\hat{\Pi}}\langle \hat{A} \rangle_{\hat{\rho}}$ is called the (real) weak value of $\hat{A}$. Consider the special case when both the state $\hat{\rho} = |i\rangle\langle i|$ and the postselected operator $\hat{\Pi} = |f\rangle\langle f|$ are pure states. Then the weak value ${}_{\hat{\Pi}}\langle \hat{A} \rangle_{\hat{\rho}}$ takes, in usual notations, a particular form [41] yielding the real part of the complex weak value A_{w} [1]:

$$ {}_{f}\langle \hat{A} \rangle_{i} =: \mathrm{Re} \frac{\langle f | \hat{A} | i \rangle}{\langle f | i \rangle} \qquad [41] $$

The interpretation of postselection itself reduces to a simple procedure. One performs the von Neumann ideal measurement of the Hermitian projector $|f\rangle\langle f|$, then includes the case if the outcome is 1 and discards it if the outcome is 0. The rate of postselection is $|\langle f | i \rangle|^2$. We note that a certain statistical interpretation of $\mathrm{Im}\, A_{\mathrm{w}}$, too, exists although it relies upon the details of the "meter."

We outline a heuristic proof of the central equation [40]. One considers the nonideal measurement (eqns [27] and [28]) of $\hat{A}$ followed by the ideal measurement of $\hat{\Pi}$. Then the joint distribution of the corresponding outcomes is given by eqn [42]. The probability distribution of the postselected outcomes a is defined by eqn [43], and takes the concrete form [44]. The constant $\mathcal{N}$ assures normalization:

$$ p(\pi, a) = \mathrm{tr}\left(\delta(\pi - \hat{\Pi}) G_{\sigma}^{1/2}(a - \hat{A}) \hat{\rho} G_{\sigma}^{1/2}(a - \hat{A}) \right) \qquad [42] $$

$$ p(a) =: \frac{1}{\mathcal{N}} \int \pi p(\pi, a)\, \mathrm{d}\pi \qquad [43] $$

$$ p(a) =: \frac{1}{\mathcal{N}} \left\langle G_{\sigma}^{1/2}(a - \hat{A}) \hat{\Pi} G_{\sigma}^{1/2}(a - \hat{A}) \right\rangle_{\hat{\rho}} \qquad [44] $$

Suppose, for simplicity, that $\hat{A}$ is bounded. When $\sigma \to \infty$, eqn [44] yields the first two moments of the outcome a:

$$ E[a] \to {}_{\hat{\Pi}}\langle \hat{A} \rangle_{\hat{\rho}} \qquad [45] $$

$$ E[a^2] \sim \sigma^2 \qquad [46] $$

Hence, by virtue of the central limit theorem, the probability distribution [40] follows for the average $\bar{a}$ of postselected outcomes in the weak measurement limit (eqns [8] and [9]).

Quantum Weak-Value Anomaly

Unlike in classical postselection, effective postselected quantum states cannot be introduced. We can ask whether eqn [47] defines a correct postselected quantum state:

$$ \hat{\rho}_{\hat{\Pi}}^{?} =: \mathrm{Herm} \frac{\hat{\Pi} \hat{\rho}}{\langle \hat{\Pi} \rangle_{\hat{\rho}}} \qquad [47] $$

This pseudo-state satisfies the quantum counterpart of the classical equation [14]:

$$ {}_{\hat{\Pi}}\langle \hat{A} \rangle_{\hat{\rho}} = \mathrm{tr}\left(\hat{A} \hat{\rho}_{\hat{\Pi}}^{?} \right) \qquad [48] $$

In general, however, the operator $\hat{\rho}_{\hat{\Pi}}^{?}$ is not a density operator since it may be indefinite. Therefore, eqn [47] does not define a quantum state. Equation [48] does not guarantee that the quantum weak value

${}_{\hat{\Pi}}\langle\hat{A}\rangle_{\hat{\rho}}$ lies within the range of the eigenvalues of the observable $\hat{A}$.

Let us see a simple example for such anomalous weak values in the two-dimensional Hilbert space. Consider the pure initial state given by eqn [49] and the postselected pure state by eqn [50], where $\phi \in [0, \pi]$ is a certain angular parameter.

$$|i\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} e^{i\phi/2} \\ e^{-i\phi/2} \end{bmatrix} \quad [49]$$

$$|f\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} e^{-i\phi/2} \\ e^{i\phi/2} \end{bmatrix} \quad [50]$$

The probability of successful postselection is $\cos^2\phi$. If $\phi \neq \pi/2$, then the postselected pseudo-state follows from eqn [47]:

$$\hat{\rho}^{?}_{\hat{\Pi}} = \frac{1}{2} \begin{bmatrix} 1 & \cos^{-1}\phi \\ \cos^{-1}\phi & 1 \end{bmatrix} \quad [51]$$

This matrix is indefinite unless $\phi = 0$, its two eigenvalues are $1 \pm \cos^{-1}\phi$. The smaller the postselection rate $\cos^2\phi$, the larger is the violation of the positivity of the pseudo-density operator. Let the weakly measured observable take the form

$$\hat{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad [52]$$

Its eigenvalues are ± 1. We express its weak value from eqns [41], [49], and [50] or, equivalently, from eqns [48] and [51]:

$${}_f\langle\hat{A}\rangle_i = \frac{1}{\cos\phi} \quad [53]$$

This weak value of $\hat{A}$ lies outside the range of the eigenvalues of $\hat{A}$. The anomaly can be arbitrarily large if the rate $\cos^2\phi$ of postselection decreases.

Striking consequences follow from this anomaly if we turn to the statistical interpretation. For concreteness, suppose $\phi = 2\pi/3$ so that ${}_f\langle\hat{A}\rangle_i = 2$. On average, 75% of the statistics N will be lost in postselection. We learnt from eqn [40] that the arithmetic mean $\bar{a}$ of the postselected outcomes of independent weak measurements converges stochastically to the weak value upto the Gaussian fluctuation Δ, as expressed symbolically by

$$\bar{a} = 2 \pm \Delta \quad [54]$$

Let us approximate the asymptotically large error σ of our weak measurements by $\sigma = 10$ which is already well beyond the scale of the eigenvalues ± 1 of the observable $\hat{A}$. The Gaussian error Δ derives from eqn [9] after replacing N by the size of the postselected statistics which is approximately $N/4$:

$$\Delta^2 = 400/N \quad [55]$$

Accordingly, if $N = 3600$ independent quantum measurements of precision $\sigma = 10$ are performed regarding the observable $\hat{A}$, then the arithmetic mean $\bar{a}$ of the ~ 900 postselected outcomes a will be 2 ± 0.33. This exceeds significantly the largest eigenvalue of the measured observable $\hat{A}$. Quantum postselection appears to bias the otherwise unbiased nonideal weak measurements.

Quantum Time-Continuous Measurement

The mathematical construction of time-continuous quantum measurement is similar to the classical one. We consider the weak measurement limit (eqns [15] and [16]) of an infinite sequence of nonideal quantum measurements of the observable $\hat{A}$ at $t = \delta t, 2\delta t, \ldots$, on the time-dependent state $\hat{\rho}_t$. The resulting theory of time-continuous quantum measurement is incorporated in the coupled stochastic equations [56] and [57] for the primitive function α_t of the time-dependent outcome and the conditional time-dependent state $\hat{\rho}_t$, respectively (Diósi 1988):

$$d\alpha_t = \langle\hat{A}\rangle_{\hat{\rho}_t} dt + g dW_t \quad [56]$$

$$d\hat{\rho}_t = -\tfrac{1}{8} g^{-2}[\hat{A}, [\hat{A}, \hat{\rho}_t]]\, dt + g^{-1}\,\mathrm{Herm}\left(\hat{A} - \langle\hat{A}\rangle_{\hat{\rho}_t}\right)\hat{\rho}_t\, dW_t \quad [57]$$

Equation [56] and its classical counterpart [17] are perfectly similar. There is a remarkable difference between eqn [57] and its classical counterpart [18]. In the latter, the stochastic average of the state is constant: $E[d\rho_t] = 0$, expressing the fact that classical measurements do not alter the original ensemble if we "ignore" the outcomes of the measurements. On the contrary, quantum measurements introduce irreversible changes to the original ensemble, a phenomenon called decoherence in the physics literature. Equation [57] implies the closed linear first-order differential equation [58] for the stochastic average of the quantum state $\hat{\rho}_t$ under time-continuous measurement of the observable $\hat{A}$:

$$\frac{dE[\hat{\rho}_t]}{dt} = -\tfrac{1}{8} g^{-2}[\hat{A}, [\hat{A}, E[\hat{\rho}_t]]] \quad [58]$$

This is the basic irreversible equation to model the gradual loss of quantum coherence (decoherence) under time-continuous measurement. In fact, the very equation models decoherence under the influence of a large class of interactions, for example, with thermal reservoirs or complex environments. In

two-dimensional Hilbert space, for instance, we can consider the initial pure state $\langle i| =: [\cos\phi, \sin\phi]$ and the time-continuous measurement of the diagonal observable [59] on it. The solution of eqn [58] is given by eqn [60]:

$$\hat{A} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \qquad [59]$$

$$E[\hat{\rho}_t] = \begin{bmatrix} \cos^2\phi & e^{-t/4g^2}\cos\phi\sin\phi \\ e^{-t/4g^2}\cos\phi\sin\phi & \sin^2\phi \end{bmatrix} \qquad [60]$$

The off-diagonal elements of this density matrix go to zero, that is, the coherent superposition represented by the initial pure state becomes an incoherent mixture represented by the diagonal density matrix $\hat{\rho}_\infty$.

Apart from the phenomenon of decoherence, the stochastic equations show remarkable similarity with the classical equations of time-continuous measurement. The heuristic form of eqn [56] is eqn [61] of invariable interpretation with respect to the classical equation [19]:

$$a_t = \langle \hat{A} \rangle_{\hat{\rho}_t} + g\xi_t \qquad [61]$$

Equation [57] describes what is called the time-continuous collapse of the quantum state under time-continuous quantum measurement of $\hat{A}$. For concreteness, we assume discrete spectrum for $\hat{A}$ and consider the spectral expansion

$$\hat{A} = \sum_\lambda a^\lambda \hat{P}^\lambda \qquad [62]$$

The real values a^λ are nondegenerate eigenvalues. The Hermitian projectors $\hat{P}^\lambda$ form a complete orthogonal set:

$$\sum_\lambda \hat{P}^\lambda \equiv \hat{I} \qquad [63]$$

$$\hat{P}^\lambda \hat{P}^\mu = \delta_{\lambda\mu}\hat{P}^\lambda \qquad [64]$$

In a single ideal measurement of $\hat{A}$, the outcome a is one of the a^λ's singled out at random. The probability distribution of the measurement outcome and the corresponding collapse of the state are given by

$$p^\lambda = \langle \hat{P}^\lambda \rangle_{\hat{\rho}_0} \qquad [65]$$

$$\hat{\rho}_0 \to \frac{1}{p^\lambda}\hat{P}^\lambda\hat{\rho}_0\hat{P}^\lambda =: \hat{\rho}^\lambda \qquad [66]$$

respectively. Equations [56] and [57] of continuous measurements are an obvious time-continuous resolution of the "sudden" ideal quantum measurement (eqns [65] and [66]) in a sense that they reproduce it in the limit $t \to \infty$. The states $\hat{\rho}^\lambda$ are stationary states of eqn [57]. It can be shown that they are indeed approached with probability p^λ for $t \to \infty$ (Gisin 1984).

Related Contexts

In addition to the two particular examples as in postselection and in time-continuous measurement, respectively, presented above, the weak measurement limit itself has further variants. A most natural example is the usual thermodynamic limit in standard statistical physics. Then weak measurements concern a certain additive microscopic observable (e.g., the spin) of each constituent and the weak value represents the corresponding additive macroscopic parameter (e.g., the magnetization) in the infinite volume limit. This example indicates that weak values have natural interpretation despite the apparent artificial conditions of their definition. It is important that the weak value, with or without postselection, plays the physical role similar to that of the common mean $\langle \hat{A} \rangle_{\hat{\rho}}$. If, between their pre- and postselection, the states $\hat{\rho}$ become weakly coupled with the state of another quantum system via the observable $\hat{A}$, their average influence will be as if $\hat{A}$ took the weak value ${}_{\hat{\Pi}}\langle \hat{A} \rangle_{\hat{\rho}}$. Weak measurements also open a specific loophole to circumvent quantum limitations related to the irreversible disturbances that quantum measurements cause to the measured state. Noncommuting observables become simultaneously measurable in the weak limit: simultaneous weak values of noncommuting observables will exist.

Literally, weak measurement had been coined in 1988 for quantum measurements with (pre- and) postselection, and became the tool of a certain time-symmetric statistical interpretation of quantum states. Foundational applications target the paradoxical problem of pre- and retrodiction in quantum theory. In a broad sense, however, the very principle of weak measurement encapsulates the trade between asymptotically weak precision and asymptotically large statistics. Its relevance in different fields has not yet been fully explored and a growing number of foundational, theoretical, and experimental applications are being considered in the literature – predominantly in the context of quantum physics. Since specialized monographs or textbooks on quantum weak measurement are not yet available, the reader is mostly referred to research articles, like the recent one by Aharonov and Botero (2005), covering many topics of postselected quantum weak values.

Nomenclature

a	measurement outcome
$\bar{a}$	arithmetic mean of measurement outcomes
$\hat{A}$	Hermitian operator, quantum observable
$A(X)$	real phase-space function
$E[\dots]$	stochastic expectation value
$\langle f\|\hat{A}\|i\rangle$	matrix element
$\langle f\|i\rangle$	inner product
$\mathcal{H}$	Hilbert space
$\mathcal{L}^2$	space of Lebesgue square-integrable complex functions
p	probability distribution
tr	trace
$\hat{U}$	unitary operator
W_t	Wiener process
ξ_t	white noise process
${}_{\Pi}\langle\dots\rangle_{\rho}$	postselected mean value
$\hat{\rho}$	density operator
$\rho(X)$	phase-space distribution
$\otimes$	direct product
$\dagger$	operator adjoint
$\|\dots\rangle$	state vector
$\langle\dots\|$	adjoint state vector
$\langle\dots\rangle_{\rho}$	mean value

Further Reading

Aharonov Y, Albert DZ, and Vaidman L (1988) How the result of measurement of a component of the spin of a spin-1/2 particle can turn out to be 100. *Physical Review Letters* 60: 1351–1354.

Aharonov Y and Botero A (2005) Quantum averages of weak values. *Physical Review A* (in print).

Aharonov Y, Popescu S, Rohrlich D, and Vaidman L (1993) Measurements, errors, and negative kinetic energy. *Physical Review A* 48: 4084–4090.

Aharonov Y and Vaidman L (1990) Properties of a quantum system during the time interval between two measurements. *Physical Review A* 41: 11–20.

Belavkin VP (1989) A new equation for a continuous nondemolition measurement. *Physics Letters A* 140: 355–358.

Diósi L (1988) Continuous quantum measurement and Itô-formalism. *Physics Letters A* 129: 419–432.

Gisin N (1984) Quantum measurements and stochastic processes. *Physical Review Letters* 52: 1657–1660.

Giulini D, Joos E, Kiefer C, Kupsch J, Stamatescu IO *et al.* (1996) *Decoherence and the Appearance of a Classical World in Quantum Theory*. Berlin: Springer.

Gough J and Sobolev A (2004) Stochastic Schrödinger equations as limit of discrete filtering. *Open Systems and Information Dynamics* 11: 1–21.

Kraus K (1983) *States, Effects, and Operations: Fundamental Notions of Quantum Theory*. Berlin: Springer.

von Neumann J (1955) *Mathematical Foundations of Quantum Mechanics*. Princeton: Princeton University Press.

Nielsen MA and Chuang IL (2000) *Quantum Computation and Quantum Information*. Cambridge: Cambridge University Press.

Stratonovich R (1968) *Conditional Markov Processes and Their Application to the Theory of Optimal Control*. New York: Elsevier.

Wiseman HM (2002) Weak values, quantum trajectories, and the cavity-QED experiment on wave-particle correlation. *Physical Review A* 65: 32111–32116.

Quantum *n*-Body Problem

R G Littlejohn, University of California at Berkeley, Berkeley, CA, USA

Introduction

This article concerns the nonrelativistic quantum mechanics of isolated systems of n particles interacting by means of a scalar potential, what we shall call the "quantum n-body problem." Such systems are described by the kinetic-plus-potential Hamiltonian,

$$H = T + V = \sum_{\alpha=1}^{n} \frac{|\boldsymbol{P}_\alpha|^2}{2m_\alpha} + V(\boldsymbol{R}_1, \dots, \boldsymbol{R}_n) \qquad [1]$$

where $\boldsymbol{R}_\alpha, \boldsymbol{P}_\alpha, \alpha = 1, \dots, n$ are the positions and momenta of the n particles in three-dimensional space, m_α are the masses, and V is the potential energy. This Hamiltonian also occurs in the "classical n-body problem," in which V is usually assumed to consist of the sum of the pairwise gravitational interactions of the particles. In this article, we shall only assume that V (hence H) is invariant under translations, proper rotations, parity, and permutations of identical particles. The Hamiltonian H is also invariant under time reversal. This Hamiltonian describes the dynamics of isolated atoms, molecules, and nuclei, with varying degrees of approximation, including the case of molecules in the Born–Oppenheimer approximation, in which V is the Born–Oppenheimer potential. We shall ignore the spin of the particles, and treat the wave function Ψ as a scalar. We assume that Ψ is an eigenfunction of H, $H\Psi = E\Psi$. In practice, the value of n typically ranges from 2 to several hundred. Often the cases $n = 3$ and $n = 4$ are of special interest. In this article, we shall assume that $n \geq 3$, since $n = 2$ is the trivial case of central-force motion. The quantum n-body problem is not to be confused with the "quantum

many-body problem," which usually refers to the quantum mechanics of large numbers of identical particles, such as the electrons in a solid.

Of particular interest is the "reduction" of the Hamiltonian [1], that is, the elimination of those degrees of freedom that can be eliminated due to the continuous symmetries of translations and rotations. A basic problem is to write down the reduced Hamiltonian and to make its analytical and geometrical properties clear. In the following we shall present this reduction in two stages, dealing first with the translations and second with the proper rotations. In each stage, we shall describe the reduction first in coordinate language and then in geometrical language. The discrete symmetries of parity, time reversal, and permutation of identical particles are handled by standard methods of group representation theory, and will not be discussed here.

There has been considerable interest in mathematical circles in recent years in the reduction of dynamical systems with symmetry, and the quantum n-body problem is one of the most important such systems from a physical standpoint. As such, the basic theory of the quantum n-body problem has received considerable attention in the physical literature going back to the birth of quantum mechanics, and continues to be of great practical importance. This article and the bibliography attempt to bridge these two centers of interest.

Reduction by Translations: Coordinate Description

We begin with a coordinate description of the reduction of the system [1] by translations. The coordinates $(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n)$ are coordinates on the configuration space of the system, called the "original configuration space" or OCS. The OCS is $\mathbb{R}^{3n}$. The original system has $3n$ degrees of freedom. The translation group acts on configuration space by $\boldsymbol{R}_\alpha \mapsto \boldsymbol{R}_\alpha + \xi$, for $\alpha = 1, \ldots, n$, where ξ is a displacement vector. It acts on wave functions by $\Psi(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n) \mapsto \Psi(\boldsymbol{R}_1 - \xi, \ldots, \boldsymbol{R}_n - \xi)$.

To reduce the system by translations, we perform a linear coordinate transformation on the OCS, taking us from the original vectors $(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n)$ to a new set of n vectors $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1}, \boldsymbol{R}_{\mathrm{CM}})$, where $\boldsymbol{R}_{\mathrm{CM}}$ is the center-of-mass position,

$$\boldsymbol{R}_{\mathrm{CM}} = \frac{1}{M}\sum_{\alpha=1}^{n} m_\alpha \boldsymbol{R}_\alpha \qquad [2]$$

where $M = \sum_\alpha m_\alpha$ is the total mass of the system, and the other $n-1$ vectors of the new coordinate system, $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1})$, are required to be translationally invariant, that is, independent linear functions of the relative particle positions $\boldsymbol{R}_\alpha - \boldsymbol{R}_\beta$. We denote the momenta conjugate to $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1}, \boldsymbol{R}_{\mathrm{CM}})$ by $(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{n-1}, \boldsymbol{P}_{\mathrm{CM}})$, of which $\boldsymbol{P}_{\mathrm{CM}}$ turns out to be the total momentum of the system,

$$\boldsymbol{P}_{\mathrm{CM}} = \sum_{\alpha=1}^{n} \boldsymbol{P}_\alpha \qquad [3]$$

Under such a coordinate transformation, the potential energy becomes simply a function of the $n-1$ relative vectors, $V(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1})$, whereas the kinetic energy becomes

$$T = \frac{|\boldsymbol{P}_{\mathrm{CM}}|^2}{2M} + \frac{1}{2}\sum_{\alpha,\beta=1}^{n-1} K^{\alpha\beta}\boldsymbol{p}_\alpha \cdot \boldsymbol{p}_\beta \qquad [4]$$

where $K^{\alpha\beta}$ is a symmetric tensor (the "inverse mass tensor").

The vectors $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1})$ specify the positions of n particles relative to their center of mass. As described so far, these vectors need only be independent, translationally invariant linear combinations of the particle postitions. However, it is convenient to choose them so that the inverse mass tensor becomes proportional to the identity, $K^{\alpha\beta} = (1/M)\delta_{\alpha\beta}$. An elegant way of doing this is the method of Jacobi vectors, which involves splitting the original set of particles into two nonempty subsets, which are then split into smaller subsets, etc., until only subsets of a single particle remain. The process can be represented by a tree growing downward, with the original n particles as the root, and the ends of the branches at the bottom each containing one particle. Then the vectors $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1})$ (the Jacobi vectors) are chosen to be proportional to the differences between the centers of mass of the two subsets at each splitting. With the right constants of proportionality, the kinetic energy becomes

$$T = \frac{1}{2M}|\boldsymbol{P}_{\mathrm{CM}}|^2 + \frac{1}{2M}\sum_{\alpha=0}^{n-1}|\boldsymbol{p}_\alpha|^2 \qquad [5]$$

Henceforth, we shall assume that the vectors $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1})$ are Jacobi vectors with conjugate momenta $(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{n-1})$.

The choice of Jacobi vectors is not unique. In the first place, there is a discrete set of possible ways of splitting the original set of n particles into subsets (of forming trees), each of which leads to the same form [5] of the kinetic energy. More generally, the kinetic energy [5] is invariant under transformations

$$\boldsymbol{r}'_\alpha = \sum_{\beta=1}^{n-1} Q_{\alpha\beta}\boldsymbol{r}_\beta \qquad [6]$$

where $Q_{\alpha\beta}$ is an orthogonal matrix, $Q \in O(n-1)$. Such transformations are called "kinematic rotations." The discrete choices of trees in forming the Jacobi vectors are equivalent to a discrete set of kinematic rotations $Q_{\alpha\beta}$ that map one standard choice of Jacobi vectors into the others.

Since the momentum $\boldsymbol{P}_{\mathrm{CM}}$ of the center of mass commutes with H, the eigenfunctions Ψ of H can be chosen to have the form

$$\Psi(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_n) = \exp(\mathrm{i}\boldsymbol{R}_{\mathrm{CM}} \cdot \boldsymbol{P}_{\mathrm{CM}}/\hbar)\psi(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1}) \qquad [7]$$

This causes ψ to be an eigenfunction of the "translation-reduced Hamiltonian," $H_{\mathrm{tr}}\psi = E_{\mathrm{tr}}\psi$, where

$$H_{\mathrm{tr}} = \frac{1}{2M}\sum_{\alpha=0}^{n-1} |\boldsymbol{p}_\alpha|^2 + V(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1}) \qquad [8]$$

The kinetic energy of the center of mass, $|\boldsymbol{P}_{\mathrm{CM}}|^2/2M$, has been discarded from both H_{tr} and E_{tr}, which represent physically the energy of the system about its center of mass.

Reduction by Translations: Geometrical Description

The kinetic energy T in eqn [1] specifies a metric $\mathrm{d}s^2 = \sum_\alpha m_\alpha |\mathrm{d}\boldsymbol{R}_\alpha|^2$ on the OCS ($=\mathbb{R}^{3n}$). The translation group ($=\mathbb{R}^3$) acts freely on the OCS, with an action that is generated by $\boldsymbol{P}_{\mathrm{CM}}$. This action defines an orthogonal decomposition of the OCS, $\mathbb{R}^{3n} = \mathbb{R}^3 \oplus \mathbb{R}^{3n-3}$, where $\mathbb{R}^3$ is the orbit of the origin (the other orbits of the translation group action are parallel spaces), and $\mathbb{R}^{3n-3}$ is the orthogonal subspace (henceforth the "translation-reduced configuration space" or TRCS for short). The TRCS is physically the space of configurations relative to the center of mass. The vectors $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{n-1})$ are coordinates on the TRCS. The TRCS possesses a metric which is the projection of the metric on the OCS onto the TRCS by means of the translation group action. The metric can be projected because translations preserve the original metric (they are isometries). Jacobi vectors are Euclidean coordinates on the TRCS with respect to this metric.

The tree method of constructing Jacobi vectors can be understood in terms of certain group actions which take place as each subset of particles is split into two further subsets. The group action in question leaves the center of mass of the original subset invariant, while moving the two new subsets apart along a line. This motion in the configuration space is orthogonal to all the other group actions that are created in the process of splitting subsets of particles, including the original action of the translation group. Thus, each splitting of a subset of particles generates a three-dimensional subspace of the OCS, on which one of the $\boldsymbol{r}_\alpha$ are coordinates. The conjugate momentum $\boldsymbol{p}_\alpha$ is the generator of the group action moving the two new subsets apart. The final result is that the OCS is decomposed into n orthogonal, three-dimensional subspaces, one of which contains the action of the original translation group, and the others of which represent the decomposition of the TRCS into $n-1$, three-dimensional orthogonal subspaces.

The TRCS can also be seen as a global section of a flat, trivial, principal fiber bundle created by the action of the translation group on the OCS. Alternatively, the TRCS can be seen as the quotient space, $\mathbb{R}^{3n}/\mathbb{R}^3$. The construction is fairly simple because the translation group is Abelian.

The wave function ψ can be seen as a member of the Hilbert space of wave functions on the TRCS, upon which the reduced Hamiltonian H_{tr} of eqn [8] acts. Alternatively, it can be seen as the function obtained by restricting Ψ on the OCS to the TRCS, where Ψ has a dependence along the orbits of the translation group given by $\exp(\mathrm{i}\boldsymbol{R}_{\mathrm{CM}} \cdot \boldsymbol{P}_{\mathrm{CM}}/\hbar)$, that is, by an irreducible representation (irrep) of the translation group.

Reduction by Rotations: Coordinate Description

The Hamiltonian H_{tr} acts on wave functions ψ defined on the TRCS and has $3n-3$ degrees of freedom. Consider a coordinate transformation to eliminate further degrees of freedom due to the rotational invariance. This coordinate transformation takes us from the Jacobi vectors $\{\boldsymbol{r}_\alpha, \alpha = 1, \ldots, n-1\}$ to orientational and shape coordinates. Shape coordinates are a set of $3n-6$ coordinates $\{q^\mu, \mu = 1, \ldots, 3n-6\}$ that specify the shape of the n-particle system, that is, they are $3n-6$ independent functions of the interparticle distances (hence rotationally invariant). We will call the space upon which the q^μ are coordinates "shape space." For example, in the case of the three-body problem, shape space is the space of all triangles.

As for orientational coordinates, to define them it is necessary first to define a "body frame." We assume we are already given one frame, the "space frame," a fixed inertial frame. The body frame is a 3-frame attached in a conventional way to each shape of the system of particles, which rotates with the particles. The orientational coordinates, to be

denoted by $\{\theta^i, i=1,2,3\}$, are three coordinates (e.g., Euler angles) specifying the SO(3) rotation that maps the space frame into the body frame. We shall write the new coordinates collectively as $\{\theta^i, q^\mu\}$.

There is a great deal of arbitrariness in the choice of a body frame, since for a given shape a body frame can be attached in many ways, the different choices being related by proper rotations. The only requirement is that the body frame should change smoothly as the shape changes. Popular choices for the body frame are the principal axis and Eckart frames.

When the potential energy is transformed to the new coordinates, it becomes a function only of the $\{q^\mu\}$, that is, of the shape. The potential can be written as $V=V(q)$. V is a scalar field on shape space.

The transformation of the kinetic energy is more complicated. When the (Euclidean) metric tensor on the TRCS is transformed to orientational and shape coordinates there results a $(3n-3)\times(3n-3)$ component matrix which may be partitioned into blocks according to the coordinates $\{\theta^i, q^\mu\}$, that is, according to $3n-3=3+(3n-6)$. This matrix cannot be made diagonal or even block diagonal by any choice of orientational or shape coordinates, or by any choice of body frame.

The components of the metric tensor in the new coordinates are conveniently expressed in terms of three fields on shape space. The first is the moment-of-inertia tensor E, which describes the 3×3 upper block of the metric tensor. Its components are given by

$$E_{ij}=M\sum_{\alpha=1}^{n-1}\left(|\boldsymbol{r}_\alpha|^2\,\delta_{ij}-r_{\alpha i}r_{\alpha j}\right) \qquad [9]$$

The vectors and tensors in this equation can be referred either to the space frame or the body frame, but the body frame is more convenient because then the components of the vectors r_α are functions only of the shape coordinates q^μ. Thus, the body frame components E_{ij} of the moment-of-inertia tensor define a field on shape space.

The second field is the "gauge potential" $\boldsymbol{A}_\mu$, an object with $3(3n-6)$ components $A^i_\mu, i=1,2,3$, $\mu=1,\ldots,3n-6$, which describes the off-diagonal blocks of the metric tensor. It is defined by

$$\boldsymbol{A}_\mu=\mathsf{E}^{-1}\left(M\sum_{\alpha=1}^{n-1}\boldsymbol{r}_\alpha\times\frac{\partial\boldsymbol{r}_\alpha}{\partial q^\mu}\right) \qquad [10]$$

in which all vectors are understood to be referred to the body frame (so the partial derivatives make sense). The gauge potential $\boldsymbol{A}_\mu$ is responsible for the "falling cat" phenomenon, in which a flexible body of zero angular momentum nevertheless manages to rotate.

The third field is the $(3n-6)\times(3n-6)$ lower block of the metric tensor on the TRCS, an object with two shape indices. It is given by

$$g_{\mu\nu}=M\sum_{\alpha=1}^{n-1}\left(\frac{\partial\boldsymbol{r}_\alpha}{\partial q^\mu}\cdot\frac{\partial\boldsymbol{r}_\alpha}{\partial q^\nu}\right)-\boldsymbol{A}_\mu\cdot\mathsf{E}\cdot\boldsymbol{A}_\nu \qquad [11]$$

where again the vectors are referred to the body frame. The notation suggests (correctly) that $g_{\mu\nu}$ is the metric tensor on shape space.

On transforming the wave function from the Jacobi vectors to coordinates (θ^i, q^μ), it is convenient to introduce a Jacobian factor, $\psi(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_{n-1})=D^{1/4}\phi(\theta^i,q^\mu)$, where $D=(\det\mathsf{E})(\det g_{\mu\nu})$. This causes the new wave function ϕ to have the normalization

$$\int \mathrm{d}\mathsf{R}\left(\prod_{\mu=1}^{3n-6}\mathrm{d}q^\mu\right)|\phi|^2 \qquad [12]$$

where $\mathrm{d}\mathsf{R}$ is the Haar measure on the group SO(3). The factor D depends only on the q^μ, not the θ^i. Then the Schrödinger equation can be written as $H_{\mathrm{tr}}\phi=E_{\mathrm{tr}}\phi$, where H_{tr} is a differential operator involving $\partial/\partial\theta^i$ and $\partial/\partial q^\mu$.

The orientational derivatives $\partial/\partial\theta^i$ in H_{tr} are conveniently expressed in terms of the angular momentum operator $\boldsymbol{L}$. When acting on the original wave function Ψ on the OCS, the angular momentum is

$$\boldsymbol{L}=\sum_{\alpha=1}^{n}\boldsymbol{R}_\alpha\times\boldsymbol{P}_\alpha \qquad [13]$$

When this is transformed to the coordinates $(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_{n-1},\boldsymbol{R}_{\mathrm{CM}})$, it becomes $\boldsymbol{L}=\boldsymbol{L}_{\mathrm{CM}}+\boldsymbol{L}_{\mathrm{tr}}$, where $\boldsymbol{L}_{\mathrm{CM}}=\boldsymbol{R}_{\mathrm{CM}}\times\boldsymbol{P}_{\mathrm{CM}}$, and

$$\boldsymbol{L}_{\mathrm{tr}}=\sum_{\alpha=1}^{n-1}\boldsymbol{r}_\alpha\times\boldsymbol{p}_\alpha \qquad [14]$$

Physically, $\boldsymbol{L}_{\mathrm{tr}}$ is the angular momentum of the system about the center of mass.

We shall henceforth drop the "tr" on $H_{\mathrm{tr}}, E_{\mathrm{tr}}$, and $\boldsymbol{L}_{\mathrm{tr}}$, thereby restricting attention to the energy and angular momentum about the center of mass.

The angular momentum $\boldsymbol{L}$, when acting on wave functions $\psi(\boldsymbol{r}_1,\ldots,\boldsymbol{r}_{n-1})$ on the TRCS, is a vector of differential operators involving $\partial/\partial\boldsymbol{r}_\alpha$. When these are transformed to orientational and shape coordinates, the components of $\boldsymbol{L}$ become differential operators involving only orientational derivatives, $\partial/\partial\theta^i$. There are no shape derivatives, $\partial/\partial q^\mu$, since $\boldsymbol{L}$ generates rotations, that is, changes in orientation, not shape. Thus, one can solve for the operators $\partial/\partial\theta^i$ in terms of the components of $\boldsymbol{L}$. This is true

both for the space and the body components of $\boldsymbol{L}$, although the differential operators are not the same in the two cases. The space components of $\boldsymbol{L}$ satisfy the usual angular momentum commutation relations, $[L_i, L_j] = i\hbar\epsilon_{ijk} L_k$, while the body components of satisfy $[L_i, L_j] = -i\hbar\epsilon_{ijk} L_k$ (with a minus sign relative to the space commutation relations).

Thus, the Hamiltonian can be expressed in terms of $\boldsymbol{L}$ and the shape momentum operators, $p_\mu = -i\hbar\partial/\partial q^\mu$. The result is

$$H = \tfrac{1}{2}\boldsymbol{L}\cdot\mathsf{I}^{-1}\cdot\boldsymbol{L} + \tfrac{1}{2}(p_\mu - \boldsymbol{L}\cdot\boldsymbol{A}_\mu)g^{\mu\nu}(p_\nu - \boldsymbol{L}\cdot\boldsymbol{A}_\nu) + V_2(q) + V(q) \qquad [15]$$

where all vectors are referred to the body frame, where $g^{\mu\nu}$ is the contravariant metric tensor on shape space, and where V_2 is given by

$$V_2 = \frac{\hbar^2}{2} D^{-1/4} \frac{\partial}{\partial q^\mu}\left(g^{\mu\nu}\frac{\partial D^{1/4}}{\partial q^\nu}\right) \qquad [16]$$

V_2 looks like a potential (it is a function of only q), hence the notation, but physically it belongs to the kinetic energy. It is sometimes called an "extrapotential." It arises from nonclassical commutators in the transformation of the kinetic energy (hence the $\hbar^2$ dependence). The first term of eqn [15] is the kinetic energy of rotation, also called the "vertical" kinetic energy, the next two terms are the remainder of the kinetic energy, somewhat imprecisely thought of as the kinetic energy of vibrations or changes in shape, also called the "horizontal kinetic energy," and the final term is the (true) potential, discussed above.

Since the Hamiltonian commutes with the angular momentum, $[H, \boldsymbol{L}] = 0$, ϕ can be chosen to be simultaneous eigenfunctions of L^2 and L_z (the latter being the space component), as well as of energy. Let ϕ_{lm} be these eigenfunctions, where l and m are the quantum numbers of L^2 and L_z, respectively. Then by the transformation properties of ϕ under rotations, we can write

$$\phi_{lm}(\theta^i, q^\mu) = \sum_{k=-l}^{+l} \chi_{lk}(q^\mu)\, D^l_{km}(\theta^i) \qquad [17]$$

where D is a standard rotation matrix and χ_{lk} are functions only of q^μ. In these equations we use the phase and other standard conventions of the theory of rotations. The wave function χ is a function only of q^μ and can loosely be thought of as the wave function on shape space. It is not a scalar like Ψ, ψ, or ϕ, but rather has $2l+1$ components indexed by k.

The Schrödinger equation for χ can be written as $H\chi = E\chi$, where H has the same form as in eqn [15], except that now the components of the angular momentum L_i are interpreted, no longer as differential operators in θ^i, but as $(2l+1)\times(2l+1)$ matrices that act on the "spinor" χ. These matrices are the transposes of the usual angular momentum matrices in angular momentum theory, that is, $(L_i)_{kk'} = \langle k'|L_i|k\rangle$.

This is the final form of the Schrödinger equation after all reductions by all continuous symmetries have been carried out. The fully reduced system has $3n-5$ degrees of freedom ($3n-6$ for the shape coordinates, and one for the "spinor" index k).

Reduction by Rotations: Geometrical Description

The proper rotation group SO(3) acts on the OCS by $\boldsymbol{R}_\alpha \mapsto \mathsf{R}\boldsymbol{R}_\alpha$, and on the TRCS by $\boldsymbol{r}_\alpha \mapsto \mathsf{R}\boldsymbol{r}_\alpha$, where $\mathsf{R} \in \mathrm{SO}(3)$. Rotations acting on the OCS do not commute with translations, but the action preserves the translation fibers, and thus can be projected onto the TRCS.

The action of SO(3) on the TRCS is effective but not free, that is, most orbits are diffeomorphic to SO(3), but a subset of measure zero (the "singular" orbits) are diffeomorphic to S^2 or a single point. Configurations of the n-particle system in which the particles do not lie on a line ("noncollinear shapes") have SO(3) orbits, those in which the particles do lie on a line but are not coincident have S^2 orbits, and the n-body collision (a single shape) has an orbit that is a single point. Thus, the action of SO(3) on the TRCS foliates the TRCS into a $(3n-6)$-parameter family of copies of SO(3), plus the singular orbits. If we exclude the singular orbits, then the TRCS has the structure of an SO(3) principal fiber bundle. In general, the bundle is not trivial. Shape space may be defined as the quotient space under the SO(3) action. Omitting the singular shapes, shape space is the base space of the bundle. The coordinates q^μ introduced above are coordinates on shape space. The singular shapes and orbits are physically accessible, and there are important questions regarding the behavior of the system in their neighborhood.

The definition of a body frame is equivalent to the choice of a section of the fiber bundle, generally only locally defined over some region of shape space. A configuration (a point in the TRCS) on the section defines an orientation of the n-particle system for the given shape, which serves as a reference orientation to which others can be referred. We think of the reference orientation as one in which the space and body frames coincide; in other orientations of the same shape, the body frame has been rotated with the body to a new orientation. The choice of the section (body frame) allows us to

impose coordinates on each (nonsingular) rotation fiber, that is, we label points on the fiber by the rotation that takes us from the section to the actual configuration in question. This is why a choice of body frame is necessary before defining orientational coordinates. Sections are only defined locally. Popular choices of body frame, such as the principal axis frame, imply multivalued sections, unless branch cuts are introduced. Orientational coordinates are simply coordinates on the group manifold SO(3), transferred to the nonsingular rotation fibers, with the group identity element mapped onto the point where the fiber intersects the section.

The metric tensor determines much of the geometry of the reduction by rotations. Since the metric on the TRCS is SO(3)-invariant, horizontal subspaces in the SO(3) fiber bundle (the TRCS minus the singular orbits) can be defined as the spaces orthogonal to the fibers (hence orthogonal to the vertical subspaces). This is a standard construction in Kaluza–Klein theories, which reappears here. Thus, the bundle has a connection, induced by the metric.

The moment-of-inertia tensor is the metric tensor restricted to a fiber, evaluated in a basis of left- (body frame) or right-invariant (space frame) vector fields on SO(3), which are transported to the fibers to create a basis of vertical vector fields.

The coordinate description of the connection is the gauge potential $\boldsymbol{A}_\mu$, in which the μ index refers to shape coordinates q^μ, and the components of the 3-vector $\boldsymbol{A}$ refer to the standard set of left- or right-invariant vector fields on SO(3). The coordinate representative of the curvature 2-form is conveniently denoted by $\boldsymbol{B}_{\mu\nu}$, defined by

$$\boldsymbol{B}_{\mu\nu} = \frac{\partial \boldsymbol{A}_\nu}{\partial q^\mu} - \frac{\partial \boldsymbol{A}_\mu}{\partial q^\nu} - \boldsymbol{A}_\mu \times \boldsymbol{A}_\nu \qquad [18]$$

where it is understood that body frame components are used. Direct calculation shows that it is nonzero, hence the fiber bundle is not flat, for any value of $n \geq 3$. The curvature form $\boldsymbol{B}_{\mu\nu}$ appears in the classical equation of motion and in the quantum commutation relations.

The field $\boldsymbol{B}_{\mu\nu}$ satisfies differential equations on shape space that have the form of Yang–Mills field equations. It is interesting that the sources of this field are singularities of the monopole type, located on the singular shapes. In the case $n = 3$, the source is a single monopole located at the three-body collision, which is similar to a Dirac monopole in electromagnetic theory.

The $(3n-6)$-dimensional horizontal subspaces of the TRCS are annihilated by three differential forms, whose values on a velocity vector of the system are the components of the classical angular momentum $\boldsymbol{L}$ (body or space components, depending on the basis of forms). Thus, horizontal motions are those for which $\boldsymbol{L} = 0$, and horizontal lifts of curves in shape space are motions of the system with vanishing angular momentum. Since angular momentum is conserved, such motions are generated by the classical equations of motion and are physically allowed. For loops in shape space, the holonomy generated by the horizontal lift is physically the rotation that a flexible body experiences when it is carried under conditions of vanishing angular momentum from an initial shape, through intermediate shapes and back to the initial shape. An example is the rotation generated by the "falling cat."

Since the metric on the TRCS is SO(3)-invariant, it may be projected onto shape space, which therefore is a Riemannian manifold in its own right. The projected metric is $ds^2 = g_{\mu\nu}\, dq^\mu\, dq^\nu$. This metric is not flat (the Riemann curvature tensor is nonzero for all values $n \geq 3$). Geodesics in shape space have horizontal lifts that are free particle motions ($V = 0$) of zero angular momentum. Conversely, such motions project onto geodesics on shape space.

A popular choice of body frame in molecular physics is the Eckart frame, which has advantages for the description of small vibrations and other purposes. The section defining the Eckart frame is a flat vector subspace of the TRCS of dimension $3n-6$ that is orthogonal (horizontal) to a particular fiber (over an equilibrium shape) at a particular orientation.

The geometrical meaning of eqn [17] is that rotations act on a set of wave functions ϕ that span an irrep of SO(3) by multiplication by the representative element of the group. In standard physics notation, l indexes the irrep, and m indexes the basis vectors spanning the irrep. Thus, the values of these wave functions at any point on the fiber are known once their values are given at a reference point. A convenient choice for the reference point is the point on the section, and the wave functions χ_{lk} are simply the values of the ϕ_{lm} on this reference point (with a change of notation, $m \to k$). Thus, the wave functions χ_{lk} are properly not "wave functions on shape space," but rather wave functions on the section.

Shape space in the case $n = 3$ is homeomorphic to the region $x_3 \geq 0$ of $\mathbb{R}^3$, and in the case $n = 4$ to $\mathbb{R}^6$. A convenient tool for understanding the structure of shape space is by its foliation under the action of the kinematic rotations, eqn [5]. The kinematic rotations commute with ordinary rotations, and hence have an action on shape space. This action preserves the eigenvalues of the moment-of-inertia tensor.

Concluding Remarks

The quantum n-body problem provides an interesting example in which nonabelian gauge theories find application in nonrelativistic quantum mechanics. The fields E, $\boldsymbol{A}_\mu$, and $g_{\mu\nu}$, and fields derived from them such as the curvature tensor $\boldsymbol{B}_{\mu\nu}$ and the Riemann curvature tensor derived from $g_{\mu\nu}$, satisfy a complex set of differential equations on shape space that can be derived by considering the vanishing of the Riemann tensor on the TRCS. The resulting field equations are useful in perturbation theory, for example, in the study of small vibrations of a molecule. This means of constructing field equations on the base space of a bundle is standard in Kaluza–Klein theories, which are an important line of thinking in modern attempts to understand gauge field theories in particle physics.

The rotations generated by flexible bodies of vanishing angular momentum (the "falling cat") are an example of a "geometric phase," that is, a nonabelian generalization of "Berry's phase." It is interesting how the associated gauge potential $\boldsymbol{A}_\mu$ in this problem plays a role in the dynamics of the n-particle system.

The Hamiltonian [15] is the starting point for numerous practical calculations, for example, the numerical evaluation of energy levels, cross-sections and reaction rates in molecular physics. One can compute, for example, chemical reaction rates for molecular processes in atmospheric or astrophysical contexts, where experiments would be difficult or expensive. The numerical analysis of the Hamiltonian [15] usually requires the introduction of a basis set and the processing of large matrices. Current techniques for basis set selection are not very satisfactory, and this is an area where research into wavelets and numerical analysis could have an impact.

See also: Bosons and Fermions in External Fields; Gravitational *N*-Body Problem (Classical); Integrable Systems: Overview.

Further Reading

Abraham R and Marsden JE (1978) *Foundations of Mechanics*, 2nd edn. Reading, MA: Benjamin/Cummings.

Aquilanti V and Cavalli S (1986) Coordinates for molecular dynamics: orthogonal local frames. *Journal of Chemical Physics* 85: 1355–1361.

Berry MV (1984) Quantal phase factors accompanying adiabatic changes. *Proceedings of the Royal Society of London. Series A* 392: 45–57.

Coquereaux R (1988) *Riemannian Geometry, Fiber bundles, Kaluza–Klein Theories and All That*. World Scientific Lecture Notes in Physics, vol. 16. Singapore: World Scientific.

Ezra GS (1982) *Symmetry Properties of Molecules*. New York: Springer.

Iwai T (1987) A gauge theory for the quantum planar three-body problem. *Journal of Mathematical Physics* 28: 964–974.

Kupperman A (1993) A new look at symmetrized hyperspherical coordinates. *Advances in Molecular Vibrations and Collision Dynamics* vol. 2B: 117–186.

Littlejohn RG and Reinsch M (1997) Gauge fields in the separation of rotations and internal motions in the n-body problem. *Reviews of Modern Physics* 69: 213–275.

Mead CA (1992) The geometric phase in molecular systems. *Reviews of Modern Physics* 64: 51–85.

Mitchell KA and Littlejohn RG (2000) Kinematic orbits and the structure of the internal space for systems of five or more bodies. *Journal of Physics A* 33: 1395–1416.

Naber GL (1997) *Topology, Geometry and Gauge Fields*. New York: Springer.

Wilson EB Jr., Decius JC, and Cross PC (1955) *Molecular Vibrations: the Theory of Infrared and Raman Vibrational Spectra*. New York: McGraw-Hill.

Quantum Phase Transitions

S Sachdev, Yale University, New Haven, CT, USA

Introduction

The study of second-order phase transitions at nonzero temperatures has a long and distinguished history in statistical mechanics. Many key physical phenomena, such as the loss of ferromagnetism in iron at the Curie temperature or the critical endpoint of CO_2, are now understood in precise quantitative detail. This understanding began in the work of Onsager, and is based upon what may now be called the Landau–Ginzburg–Wilson theory. The content of this sophisticated theory may be summarized in a few basic principles: (1) The collective thermal fluctuations near second-order transitions can be accurately described by simple classical models, that is, quantum-mechanical effects can be entirely neglected. (2) The classical models identify an "order parameter," a collective variable which has to be treated on par with other thermodynamic variables, and whose correlations exhibit distinct behavior in the phases on either side of the transition. (3) The thermal fluctuations of the order parameter near the transition are controlled by a continuum field theory whose structure is usually completly dictated by simple symmetry considerations.

This article will not consider such nonzero temperature phase transitions, but will instead describe second-order phase transitions at the absolute zero of temperature. Such transitions are driven by quantum fluctuations mandated by the Heisenberg uncertainty principle: one can imagine moving across the quantum critical point by effectively "tuning the value of Planck's constant, $\hbar$." Clearly, quantum mechanics plays a central role at such transitions, unlike the situation at nonzero temperatures. The reader may object that absolute zero is an idealization not realized by any experimental system; hence, the study of quantum phase transitions is a subject only of academic interest. As we will illustrate below, knowledge of the zero-temperature quantum critical points of a system is often the key to understanding its finite-temperature properties, and in some cases the influence of a zero-temperature critical point can be detected at temperatures as high as ambient room temperature.

We will begin in the following section by introducing some simple lattice models which exhibit quantum phase transitions. Next the theory of the critical point in these models is based upon a natural extension of the Landau–Ginzburg–Wilson (LGW) method, and this will be presented. This section will also describe the consequences of a zero-temperature critical point on the nonzero temperature properties. Finally, we will consider more complex models in which quantum interference effects play a more subtle role, and which cannot be described in the LGW framework: such quantum critical points are likely to play a central role in understanding many of the correlated electron systems of current interest.

Simple Models

Quantum Ising Chain

This is a simple model of N qubits, labeled by the index $j = 1, \ldots, N$. On each "site" j there are two qubit quantum states $|\uparrow\rangle_j$ and $|\downarrow\rangle_j$ (in practice, these could be two magnetic states of an ion at site j in a crystal). The Hilbert space therefore consists of 2^N states, each consisting of a tensor product of the states on each site. We introduce the Pauli spin operators, $\hat{\sigma}_j^\alpha$, on each site j, with $\alpha = x, y, z$:

$$\hat{\sigma}^x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \hat{\sigma}^y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$
$$\hat{\sigma}^z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad [1]$$

These operators clearly act on the two states of the qubit on site j, and the Pauli operators on different sites commute.

The quantum Ising chain is defined by the simple Hamiltonian

$$H_{\mathrm{I}} = -J \sum_{j=1}^{N-1} \hat{\sigma}_j^z \hat{\sigma}_{j+1}^z - gJ \sum_{j=1}^{N} \hat{\sigma}_j^x \qquad [2]$$

where $J > 0$ sets the energy scale, and $g \geq 0$ is a dimensionless coupling constant. In the thermodynamic limit ($N \to \infty$), the ground state of H_{I} exhibits a second-order quantum phase transition as g is tuned across a critical value $g = g_c$ (for the specific case of H_{I} it is known that $g_c = 1$), as we will now illustrate.

First, consider the ground state of H_{I} for $g \ll 1$. At $g = 0$, there are two degenerate "ferromagnetically ordered" ground states

$$|\Uparrow\rangle = \prod_{j=1}^{N} |\uparrow\rangle_j, \qquad |\Downarrow\rangle = \prod_{j=1}^{N} |\downarrow\rangle_j \qquad [3]$$

Each of these states breaks a discrete "Ising" symmetry of the Hamiltonian rotations of all spins by 180° about the x-axis. These states are more succinctly characterized by defining the ferromagnetic moment, N_0, by

$$N_0 = \langle \Uparrow | \hat{\sigma}_j^z | \Uparrow \rangle = -\langle \Downarrow | \hat{\sigma}_j^z | \Downarrow \rangle \qquad [4]$$

At $g = 0$ we clearly have $N_0 = 1$. A key point is that in the thermodynamic limit, this simple picture of the ground state survives for a finite range of small g (indeed, for all $g < g_c$), but with $0 < N_0 < 1$. The quantum tunneling between the two ferromagnetic ground states is exponentially small in N (and so can be neglected in the thermodynamic limit), and so the ground state remains 2-fold degenerate and the discrete Ising symmetry remains broken. The change in the wave functions of these states from eqn [3] can be easily determined by perturbation theory in g: these small g quantum fluctuations reduce the value of N_0 from unity but do not cause the ferromagnetism to disappear.

Now consider the ground state of H_{I} for $g \gg 1$. At $g = \infty$ there is a single nondegenerate ground state which fully preserves all symmetries of H_{I}:

$$|\Rightarrow\rangle = 2^{-N/2} \prod_{j=1}^{N} \left(|\uparrow\rangle_j + |\downarrow\rangle_j \right) \qquad [5]$$

It is easy to verify that this state has no ferromagnetic moment $N_0 = \langle \Rightarrow | \hat{\sigma}_j^z | \Rightarrow \rangle = 0$. Further, perturbation theory in $1/g$ shows that these features of the ground state are preserved for a finite range of large g values

(indeed, for all $g > g_c$). One can visualize this ground state as one in which strong quantum fluctuations have destroyed the ferromagnetism, with the local magnetic moments quantum tunneling between "up" and "down" on a timescale of order $\hbar/J$.

Given the very distinct signatures of the small g and large g ground states, it is clear that the ground state cannot evolve smoothly as a function of g. These must be at least one point of nonanalyticity as a function of g: for H_I it is known that there is only a single nonanalytic point, and this is at the location of a second-order quantum phase transition at $g = g_c = 1$.

The character of the excitations above the ground state also undergoes a qualitative change across the quantum critical point. In both the $g < g_c$ and $g > g_c$ phases, these excitations can be described in the Landau quasiparticle scheme, that is, as superpositions of nearly independent particle-like excitations; a single well-isolated quasiparticle has an infinite lifetime at low excitation energies. However, the physical nature of the quasiparticles is very different in the two phases. In the ferromagnetic phase, with $g < g_c$, the quasiparticles are domain walls between regions of opposite magnetization:

$$|j, j+1\rangle = \prod_{k=1}^{j} |\uparrow\rangle_k \prod_{\ell=j+1}^{N} |\downarrow\rangle_\ell \qquad [6]$$

This is the exact wave function of a stationary quasiparticle excitation between sites j and $j+1$ at $g = 0$; for small nonzero g the quasiparticle acquires a "cloud" of further spin-flips and also becomes mobile. However its qualitative interpretation as a domain wall between the two degenerate ground states remains valid for all $g < g_c$. In contrast, for $g > g_c$, there is no ferromagnetism, and the nondegenerate paramagnetic state has a distinct quasiparticle excitation:

$$|j\rangle = 2^{-N/2}\left(|\uparrow\rangle_j - |\downarrow\rangle_j\right) \prod_{k \neq j} \left(|\uparrow\rangle_k + |\downarrow\rangle_k\right) \qquad [7]$$

This is a stationary "flipped spin" quasiparticle at site j, with its wave function exact at $g = \infty$. Again, this quasiparticle is mobile and applicable for all $g > g_c$, but there is no smooth connection between eqns [7] and [6].

Coupled Dimer Antiferromagnet

This model also involves qubits, but they are now placed on the sites, j, of a two-dimensional square lattice. Models in this class describe the magnetic excitations of many experimentally important spin gap compounds.

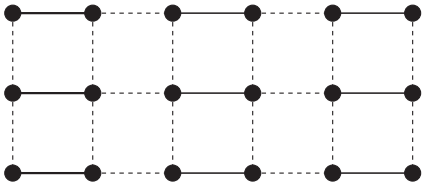


Figure 1 The coupled dimer antiferromagnet. Qubits (i.e., $S = 1/2$ spins) are placed on the sites, the $\mathcal{A}$ links are shown as full lines, and the $\mathcal{B}$ links as dashed lines.

The Hamiltonian of the dimer antiferromagnet is illustrated in Figure 1 and is given by

$$H_d = J \sum_{\langle jk \rangle \in A} \left(\hat{\sigma}_j^x \hat{\sigma}_k^x + \hat{\sigma}_j^y \hat{\sigma}_k^y + \hat{\sigma}_j^z \hat{\sigma}_k^z\right) + \frac{J}{g} \sum_{\langle jk \rangle \in B} \left(\hat{\sigma}_j^x \hat{\sigma}_k^x + \hat{\sigma}_j^y \hat{\sigma}_k^y + \hat{\sigma}_j^z \hat{\sigma}_k^z\right) \qquad [8]$$

where $J > 0$ is the exchange constant, $g \geq 1$ is the dimensionless coupling, and the set of nearest-neighbor links $\mathcal{A}$ and $\mathcal{B}$ are defined in Figure 1. An important property of H_d is that it is now invariant under the full O(3) group of spin rotations under which the $\hat{\sigma}^\alpha$ transform as ordinary vectors (in contrast to the Z_2 symmetry group of H_I). In analogy with H_I, we will find that H_d undergoes a quantum phase transition from a paramagnetic phase which preserves all symmetries of the Hamiltonian at large g, to an antiferromagnetic phase which breaks the O(3) symmetry at small g. This transition occurs at a critical value $g = g_c$, and the best current numerical estimate is $1/g_c = 0.52337(3)$.

As in the previous section, we can establish the existence of such a quantum phase transition by contrasting the disparate physical properties at large g with those at $g \approx 1$. At $g = \infty$ the exact ground state of H_d is

$$|\text{spin gap}\rangle = \prod_{\langle jk \rangle \in A} \frac{1}{\sqrt{2}} \left(|\uparrow\rangle_j |\downarrow\rangle_k - |\downarrow\rangle_j |\uparrow\rangle_k\right) \qquad [9]$$

and is illustrated in Figure 2. This state is nondegenerate and invariant under spin rotations, and so is a paramagnet: the qubits are paired into spin singlet valence bonds across all the $\mathcal{A}$ links.

The excitations above the ground state are created by breaking a valence bond, so that the pair of spins form a spin triplet with total spin $S = 1$ – this is illustrated in Figure 3. It costs a large energy to create this excitation, and at finite g the

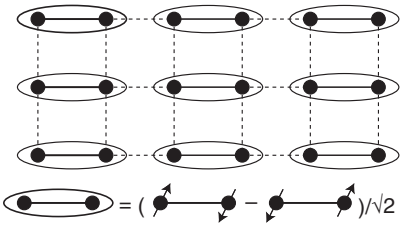


Figure 2 The paramagnetic state of H_d for $g > g_c$. The state illustrated is the exact ground state for $g = \infty$, and it is adiabatically connected to the ground state for all $g > g_c$.

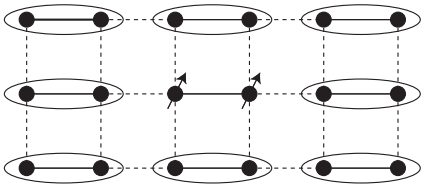


Figure 3 The triplon excitation of the $g > g_c$ paramagnet. The stationary triplon is an eigenstate only for $g = \infty$ but it becomes mobile for finite g.

triplet can hop from link to link, creating a gapped "triplon" quasiparticle excitation. This is similar to the large g paramagnet for H_I, with the important difference that each quasiparticle is now 3-fold degenerate.

At $g = 1$, the ground state of H_d is not known exactly. However, at this point H_d becomes equivalent to the nearest-neighbor square lattice antiferromagnet, and this is known to have antiferromagnetic order in the ground state, as illustrated in **Figure 4**. This state is similar to the ferromagnetic ground state of H_I, with the difference that the magnetic moment now acquires a staggered pattern on the two sublattices, rather than the uniform moment of the ferromagnet. Thus, in this ground state

$$\langle \mathrm{AF}|\hat{\sigma}_j^{\alpha}|\mathrm{AF}\rangle = N_0 \eta_j n_{\alpha} \qquad [10]$$

where $0 < N_0 < 1$ is the antiferromagnetic moment, $\eta_j = \pm 1$ identifies the two sublattices in **Figure 4**, and n_{α} is an arbitrary unit vector specifying the

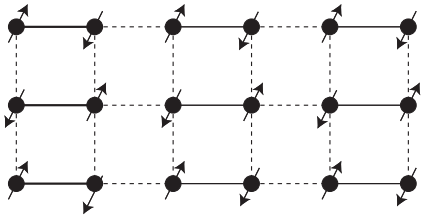


Figure 4 Schematic of the ground state with antiferromagnetic order with $g < g_c$.

orientation of the spontaneous magnetic moment which breaks the O(3) spin rotation invariance of H_d. The excitations above this antiferromagnet are also distinct from those of the paramagnet: they are a doublet of spin waves consisting of a spatial variation in the local orientation, n_{α}, of the antiferromagnetic order: the energy of this excitation vanishes in the limit of long wavelengths, in contrast to the finite energy gap of the triplon excitation of the paramagnet.

As with H_I, we can conclude from the distinct characters of the ground states and excitations for $g \gg 1$ and $g \approx 1$ that there must be a quantum critical point at some intermediate $g = g_c$.

Quantum Criticality

The simple considerations of the previous section have given a rather complete description (based on the quasiparticle picture) of the physics for $g \ll g_c$ and $g \gg g_c$. We turn, finally, to the region $g \approx g_c$. For the specific models discussed in the previous section, a useful description is obtained by a method that is a generalization of the LGW method developed earlier for thermal phase transitions. However, some aspects of the critical behavior (e.g., the general forms of eqns [13]–[15]) will apply also to the quantum critical point of the section "Beyond LGW theory."

Following the canonical LGW strategy, we need to identify a collective order parameter which distinguishes the two phases. This is clearly given by the ferromagnetic moment in eqn [4] for the quantum Ising chain, and the antiferromagnetic moment in eqn [10] for the coupled dimer antiferromagnet. We coarse-grain these moments over some finite averaging region, and at long wavelengths this yields a real order parameter field ϕ_a, with the index $a = 1, \ldots, n$. For the Ising case we have $n = 1$ and ϕ_a is a measure of the local average of N_0 as defined in eqn [4]. For the antiferromagnet, a extends over the three values x, y, z (so $n = 3$), and three components of ϕ_a specify the magnitude and orientation of the local antiferromagnetic order in eqn [10]; note the average orientation of a specific spin at site j is η_j times the local value of ϕ_a.

The second step in the LGW approach is to write down a general field theory for the order parameter, consistent with all symmetries of the underlying model. As we are dealing with a quantum transition, the field theory has to extend over spacetime, with the temporal fluctuations representing the sum over histories in the Feynman path-integral approach. With this reasoning, the proposed partition function

for the vicinity of the critical point takes the following form:

$$\mathcal{Z}_\phi = \int \mathcal{D}\phi_a(x,\tau) \times \exp\left[-\int \mathrm{d}^d x\, \mathrm{d}\tau \left(\frac{1}{2}\left((\partial_\tau \phi_a)^2 + c^2(\nabla_x \phi_a)^2 + s\phi_a^2\right) + \frac{u}{4!}\left(\phi_a^2\right)^2\right)\right] \quad [11]$$

Here τ is imaginary time; there is an implied summation over the n values of the index a, c is a velocity, and s and $u > 0$ are coupling constants. This is a field theory in $d+1$ spacetime dimensions, in which the Ising chain corresponds to $d=1$ and the dimer antiferromagnet to $d=2$. The quantum phase transition is accessed by tuning the "mass" s: there is a quantum critical point at $s=s_c$ and the $s<s_c (s>s_c)$ regions correspond to the $g<g_c (g>g_c)$ regions of the lattice models. The $s<s_c$ phase has $\langle\phi_a\rangle \neq 0$ and this corresponds to the spontaneous breaking of spin rotation symmetry noted in eqns [4] and [10] for the lattice models. The $s>s_c$ phase is the paramagnet with $\langle\phi_a\rangle = 0$. The excitations in this phase can be understood as small harmonic oscillations of ϕ_a about the point (in field space) $\phi_a = 0$. A glance at eqn [11] shows that there are n such oscillators for each wave vector. These oscillators clearly constitute the $g>g_c$ quasiparticles found earlier in eqn [7] for the Ising chain (with $n=1$) and the triplon quasiparticle (with $n=3$) illustrated in **Figure 3** for the dimer antiferromagnet.

We have now seen that there is a perfect correspondence between the phases of the quantum field theory $\mathcal{Z}_\phi$ and those of the lattice models H_I and H_d. The power of the representation in eqn [11] is that it also allows us to get a simple description of the quantum critical point. In particular, readers may already have noticed that if we interpret the temporal direction τ in eqn [11] as another spatial direction, then $\mathcal{Z}_\phi$ is simply the classical partition function for a thermal phase transition in a ferromagnet in $d+1$ dimensions: this is the canonical model for which the LGW theory was originally developed. We can now take over standard results for this classical critical point, and obtain some useful predictions for the quantum critical point of $\mathcal{Z}_\phi$. It is useful to express these in terms of the dynamic susceptibility defined by

$$\chi(k,\omega) = \frac{\mathrm{i}}{\hbar}\int \mathrm{d}^d x \times \int_0^\infty \mathrm{d}t \left\langle \left[\hat{\phi}(x,t), \hat{\phi}(0,0)\right]\right\rangle_T \mathrm{e}^{-\mathrm{i}kx + \mathrm{i}\omega t} \quad [12]$$

Here $\hat{\phi}$ is the Heisenberg field operator corresponding to the path integral in eqn [11], the square brackets represent a commutator, and the angular brackets an average over the partition function at a temperature T. The structure of χ can be deduced from the knowledge that the quantum correlators of $\mathcal{Z}_\phi$ are related by analytic continuation in time to the corresponding correlators of the classical statistical mechanics problem in $d+1$ dimensions. The latter are known to diverge at the critical point as $\sim 1/p^{2-\eta}$ where p is the $(d+1)$-dimensional momentum, η is defined to be the anomalous dimension of the order parameter ($\eta = 1/4$ for the quantum Ising chain). Knowing this, we can deduce the form of the quantum correlator in eqn [12] at the zero-temperature quantum critical point

$$\chi(k,\omega) \sim \frac{1}{(c^2k^2 - \omega^2)^{1-\eta/2}}, \quad T=0, \quad g=g_c \quad [13]$$

The most important property of eqn [13] is the absence of a quasiparticle pole in the spectral density. Instead, $\mathrm{Im}(\chi(k,\omega))$ is nonzero for all $\omega > ck$, reflecting the presence of a continuum of critical excitations. Thus the stable quasiparticles found at low enough energies for all $g \neq g_c$ are absent at the quantum critical point.

We now briefly discuss the nature of the phase diagram for $T>0$ with g near g_c. In general, the interplay between quantum and thermal fluctuations near a quantum critical point can be quite complicated, and we cannot discuss it in any detail here. However, the physics of the quantum Ising chain is relatively simple, and also captures many key features found in more complex situations, and is summarized in **Figure 5**. For all $g \neq g_c$ there is a range of low temperatures ($T \lesssim |g - g_c|$) where the long time dynamics can be described using a dilute gas of thermally excited quasiparticles. Further, the

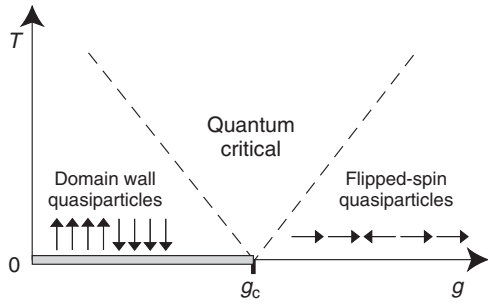


Figure 5 Nonzero temperature phase diagram of H_I. The ferromagnetic order is present only at $T=0$ on the shaded line with $g<g_c$. The dashed lines at finite T are crossovers out of the low-T quasiparticle regimes where a quasiclassical description applies. The state sketched on the paramagnetic side used the notation $|{\rightarrow}\rangle_j = 2^{-1/2}(|{\uparrow}\rangle_j + |{\downarrow}\rangle_j)$ and $|{\leftarrow}\rangle_j = 2^{-1/2}(|{\uparrow}\rangle_j - |{\downarrow}\rangle_j)$.

dynamics of these quasiparticles is quasiclassical, although we reiterate that the nature of the quasiparticles is entirely distinct on opposite sides of the quantum critical point. Most interesting, however, is the novel quantum critical region, $T \gtrsim |g - g_c|$, where neither quasiparticle picture nor a quasiclassical description are appropriate. Instead, we have to understand the influence of temperature on the critical continuum associated with eqn [13]. This is aided by scaling arguments which show that the only important frequency scale which characterizes the spectrum is $k_B T/\hbar$, and the crossovers near this scale are universal, that is, independent of specific microscopic details of the lattice Hamiltonian. Consequently, the zero-momentum dynamic susceptibility in the quantum critical region takes the following form at small frequencies:

$$\chi(k = 0, \omega) \sim \frac{1}{T^{2-\eta}} \frac{1}{(1 - i\omega/\Gamma_R)} \qquad [14]$$

This has the structure of the response of an overdamped oscillator, and the damping frequency, Γ_R, is given by the universal expression

$$\Gamma_R = \left(2 \tan \frac{\pi}{16}\right) \frac{k_B T}{\hbar} \qquad [15]$$

The numerical proportionality constant in eqn. [15] is specific to the quantum Ising chain; other models also obey eqn [15] but with a different numerical value for this constant.

Beyond LGW Theory

The quantum transitions discussed so far have turned to have a critical theory identical to that found for classical thermal transitions in $d + 1$ dimensions. Over the last decade it has become clear that there are numerous models, of key physical importance, for which such a simple classical correspondence does not exist. In these models, quantum Berry phases are crucial in establishing the nature of the phases, and of the critical boundaries between them. In less technical terms, a signature of this subtlety is an important simplifying feature which was crucial in the analyses of the section "Simple models": both models had a straightforward $g \to \infty$ limit in which we were able to write down a simple, nondegenerate, ground-state wave function of the "disordered" paramagnet. In many other models, identification of the disordered phase is not as straightforward: specifying absence of a particular magnetic order is not enough to identify a quantum state, as we still need to write down a suitable wave function. Often, subtle quantum interference effects induce new types of order in the disordered state, and such effects are entirely absent in the LGW theory.

An important example of a system displaying such phenomena is the $S = 1/2$ square lattice antiferromagnet with additional frustrating interactions. The quantum degrees of freedom are identical to those of the coupled dimer antiferromagnet, but the Hamiltonian preserves the full point-group symmetry of the square lattice:

$$H_s = \sum_{j<k} J_{jk} \left(\hat{\sigma}_j^x \hat{\sigma}_k^x + \hat{\sigma}_j^y \hat{\sigma}_k^y + \hat{\sigma}_j^z \hat{\sigma}_k^z\right) + \cdots \qquad [16]$$

Here the $J_{jk} > 0$ are short-range exchange interactions which preserve the square lattice symmetry, and the ellipses represent possible further multiple spin terms. Now imagine tuning all the non-nearest-neighbor terms as a function of some generic coupling constant g. For small g, when H_s is nearly the square lattice antiferromagnet, the ground state has antiferromagnetic order as in **Figure 4** and eqn [10]. What is now the disordered ground state for large g? One natural candidate is the spin-singlet paramagnet in **Figure 2**. However, because all nearest neighbor bonds of the square lattice are now equivalent, the state in **Figure 2** is degenerate with three other states obtained by successive 90° rotations about a lattice site. In other words, the state in **Figure 2**, when transferred to the square lattice, breaks the symmetry of lattice rotations by 90°. Consequently it has a new type of order, often called valence-bond-solid (VBS) order. It is now believed that a large class of models like H_s do indeed exhibit a second-order quantum phase transition between the antiferromagnetic state and a VBS state – see **Figure 6**. Both the existence of VBS order in the paramagnet, and of a second-order quantum transition, are features that are not predicted by LGW theory: these can only be

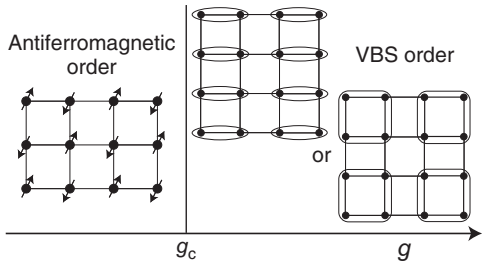


Figure 6 Phase diagram of H_s. Two possible VBS states are shown: one which is the analog of **Figure 2**, and the other in which spins form singlets in a plaquette pattern. Both VBS states have a 4-fold degeneracy due to breaking of square lattice symmetry. So the novel critical point at $g = g_c$ (described by $\mathcal{Z}_2$) has the antiferromagnetic and VBS orders vanishing as it is approached from either side: this coincident vanishing of orders is generically forbidden in LGW theories.

understood by a careful study of quantum interference effects associated with Berry phases of spin fluctuations about the antiferromagnetic state. We will not enter into details of this analysis here, but will conclude our discussion by writing down the theory so obtained for the quantum critical point in Figure 6:

$$\mathcal{Z}_z = \int \mathcal{D}z_\alpha(x,\tau)\mathcal{D}A_\mu(x,\tau) \times \exp\left(-\int \mathrm{d}^2x\,\mathrm{d}\tau\left[|(\partial_\mu - \mathrm{i}A_\mu)z_\alpha|^2 + s|z_\alpha|^2 + \frac{u}{2}(|z_\alpha|^2)^2 + \frac{1}{2e^2}(\epsilon_{\mu\nu\lambda}\partial_\nu A_\lambda)^2\right]\right) \quad [17]$$

Here μ, ν, λ are spacetime indices which extend over the two spatial directions and τ, α is a spinor index which extends over $\uparrow$, $\downarrow$, and z_α is complex spinor field. In comparing $\mathcal{Z}_z$ to $\mathcal{Z}_\phi$, note that the vector order parameter ϕ_a has been replaced by a spinor z_α, and these are related by $\phi_a = z^*_\alpha \sigma^a_{\alpha\beta} z_\beta$, where σ^a are the Pauli matrices. So the order parameter has fractionalized into the z_α. A second novel property of $\mathcal{Z}_z$ is the presence of a U(1) gauge field A_μ: this gauge force emerges near the critical point, even though the underlying model in eqn [16] only has simple two spin interactions. Studies of fractionalized critical theories like $\mathcal{Z}_c$ in other models with spin and/or charge excitations is an exciting avenue for further theoretical research.

See also: Bose–Einstein Condensates; Boundary Conformal Field Theory; Fractional Quantum Hall Effect; Ginzburg–Landau Equation; High *Tc* Superconductor Theory; Quantum Central-Limit Theorems; Quantum Spin Systems; Quantum Statistical Mechanics: Overview.

Further Reading

Matsumoto M, Yasuda C, Todo S, and Takayama H (2002) *Physical Review* B 65: 014407.
Sachdev S (1999) *Quantum Phase Transitions*. Cambridge: Cambridge University Press.
Senthil T, Balents L, Sachdev S, Vishwanath A, and Fisher MPA, http://arxiv.org/abs/cond-mat/0312617.

Quantum Spin Systems

B Nachtergaele, University of California at Davis, Davis, CA, USA

Introduction

The theory of quantum spin systems is concerned with the properties of quantum systems with an infinite number of degrees of freedom that each have a finite-dimensional state space. Occasionally, one is specifically interested in finite systems. Among the most common examples, one has an n-dimensional Hilbert space associated with each site of a d-dimensional lattice.

A model is normally defined by describing a Hamiltonian or a family of Hamiltonians, which are self-adjoint operators on the Hilbert space, and one studies their spectrum, the eigenstates, the equilibrium states, the system dynamics, and non-equilibrium stationary states, etc.

More particularly, the term "quantum spin system" often refers to such models where each degree of freedom is thought of as a spin variable, that is, there are three basic observables representing the components of the spin, S^1, S^2, and S^3, and these components transform according to a unitary representation of SU(2). The most commonly encountered situation is where the system consists of N spins, each associated with a fixed irreducible representation of SU(2). One speaks of a spin-J model if this representation is the $(2J+1)$-dimensional one. The possible values of J are $1/2, 1, 3/2, \ldots$

The spins are usually thought of as each being associated with a site in a lattice, or more generally, a vertex in a graph. In a condensed-matter-physics model, each spin may be associated with an ion in a crystalline lattice. Quantum spin systems are also used in quantum information theory and quantum computation, and show up as abstract mathematical objects in representation theory and quantum probability.

In this article we give a brief introduction to the subject, starting with a very short review of its history. The mathematical framework is sketched and the most important definitions are given. Three sections, "Symmetries and symmetry breaking," "Phase transitions," and "Dynamics," together cover the most important aspects of quantum spin systems actively pursued today.

A Very Brief History

The introduction of quantum spin systems was the result of the marriage of two developments during the 1920s. The first was the realization that angular momentum (hence, also the magnetic moment) is quantized (Pauli 1920, Stern and Gerlach 1922) and that particles such as the electron have an intrinsic

angular momentum called spin (Compton 1921, Goudsmit and Uhlenbeck 1925).

The second development was the attempt in statistical mechanics to explain ferromagnetism and the phase transition associated with it on the basis of a microscopic theory (Lenz and Ising 1925). The fundamental interaction between spins, the so-called exchange operator which is a subtle consequence of the Pauli exclusion principle, was introduced independently by Dirac and Heisenberg in 1926. With this discovery, it was realized that magnetism is a quantum effect and that a fundamental theory of magnetism requires the study of quantum-mechanical models. This realization and a large amount of subsequent work notwithstanding, some of the most fundamental questions, such as a derivation of ferromagnetism from first principles, remain open.

The first and most important quantum spin model is the Heisenberg model, so named after Heisenberg. It has been studied intensely ever since the early 1930s and its study has led to an impressive variety of new ideas in both mathematics and physics. Here, we limit ourselves to listing only some landmark developments.

Spin waves were discovered independently by Bloch and Slater in 1930 and they continue to play an essential role in our understanding of the excitation spectrum of quantum spin Hamiltonians. In two papers published in 1956, Dyson advanced the theory of spin waves by showing how interactions between spin waves can be taken into account.

In 1931, Bethe introduced the famous Bethe ansatz to show how the exact eigenvectors of the spin-1/2 Heisenberg model on the one-dimensional lattice can be found. This exact solution, directly and indirectly, led to many important developments in statistical mechanics, combinatorics, representation theory, quantum field theory and more. Hulthén used the Bethe ansatz to compute the ground-state energy of the antiferromagnetic spin-1/2 Heisenberg chain in 1938.

In their famous 1961 paper, Lieb, Schultz, and Mattis showed that some quantum spin models in one dimension can be solved exactly by mapping them into a problem of free fermions. This paper is still one of the most cited in the field.

Robinson, in 1967, laid the foundation for the mathematical framework, which we describe in the next section. Using this framework, Araki established the absence of phase transitions at positive temperatures in a large class of one-dimensional quantum spin models in 1969.

During the more recent decades, the mathematical and computational techniques used to study quantum spin models have fanned out in many directions.

When it was realized in the 1980s that the magnetic properties of complex materials play an important role in high-T_c superductivity, a variety of quantum spin models studied in the literature proliferated. This motivated a large number of theoretical and experimental studies of materials with exotic properties that are often based on quantum effects that do not have a classical analog. An example of unexpected behavior is the prediction by Haldane of the spin liquid ground state of the spin-1 Heisenberg antiferromagnetic chain in 1983. In the quest for a mathematical proof of this prediction (a quest still ongoing today), Affleck, Kennedy, Lieb, and Tasaki introduced the AKLT model in 1987. They were able to prove that the ground state of this model has all the characteristic properties predicted by Haldane for the Heisenberg chain: a unique ground state with exponential decay of correlations and a spectral gap above the ground state.

There are also particle models that are defined on a lattice, or more generally, a graph. Unlike spins, particles can hop from one site to another. These models are closely related to quantum spin systems and, in some cases, are mathematically equivalent. The best-known example of a model of lattice fermions is the Hubbard model. Such systems are not discussed further in this article.

Mathematical Framework

Quantum spin systems present an area of mathematical physics where the demands of mathematical rigor can be fully met and, in many cases, this can be done without sacrificing the ability to include all physically relevant models and phenomena. This does not mean, however, that there are few open problems remaining. But it does mean that, in general, these open problems are precisely formulated mathematical questions.

In this section we review the standard mathematical framework for quantum spin systems, in which the topics discussed in the subsequent section can be given a precise mathematical formulation. It is possible, however, to skip this section and read the rest with only a physical or intuitive understanding of the notions of observable, Hamiltonian, dynamics, symmetry, ground state, etc.

The most common mathematical setup is as follows. Let $d \geq 1$, and let $\mathcal{L}$ denote the family of finite subsets of the d-dimensional integer lattice $\mathbb{Z}^d$. For simplicity we will assume that the Hilbert space of the "spin" associated with each $x \in \mathbb{Z}^d$ has the same dimension $n \geq 2$: $\mathcal{H}_{\{x\}} \cong \mathbb{C}^n$. The Hilbert space associated with the finite volume $\Lambda \in \mathcal{L}$ is then $\mathcal{H}_\Lambda = \bigotimes_{x\in\Lambda} \mathcal{H}_x$. The algebra of observables for the spin of site x consists of the $n \times n$ complex matrices: $\mathcal{A}_{\{x\}} \cong M_n(\mathbb{C})$. For any

$\Lambda \in \mathcal{L}$, the algebra of observables for the system in Λ is given by $\mathcal{A}_\Lambda = \bigotimes_{x\in\Lambda} \mathcal{A}_{\{x\}}$. The primary observables for a quantum spin model are the spin-S matrices S^1, S^2, and S^3, where S is the half-integer such that $n = 2S + 1$. They are defined as Hermitian matrices satisfying the SU(2) commutation relations. Instead of S^1 and S^2, one often works with the spin-raising and -lowering operators, S^+ and S^-, defined by the relations $S^1 = (S^+ + S^-)/2$, and $S^2 = (S^+ - S^-)/(2\mathrm{i})$. In terms of these, the SU(2) commutation relations are

$$[S^+, S^-] = 2S^3, \quad [S^3, S^\pm] = \pm S^\pm \qquad [1]$$

where we have used the standard notation for the commutator for two elements A and B in an algebra: $[A, B] = AB - BA$. In the standard basis S^3, S^+, and S^- are given by the following matrices:

$$S^3 = \begin{pmatrix} S & & & \\ & S-1 & & \\ & & \ddots & \\ & & & -S \end{pmatrix}$$

$S^- = (S^+)^*$, and

$$S^+ = \begin{pmatrix} 0 & c_S & & & \\ & 0 & c_{S-1} & & \\ & & \ddots & \ddots & \\ & & & 0 & c_{-S+1} \\ & & & & 0 \end{pmatrix}$$

where, for $m = -S, -S+1, \ldots, S$,

$$c_m = \sqrt{S(S+1) - m(m-1)}$$

In the case $n = 2$, one often works with the Pauli matrices, σ^1, σ^2, σ^3, simply related to the spin matrices by $\sigma^j = 2S^j$, $j = 1, 2, 3$.

Most physical observables are expressed as finite sums and products of the spin matrices S_x^j, $j = 1, 2, 3$, associated with the site $x \in \Lambda$:

$$S_x^j = \bigotimes_{y\in\Lambda} A_y$$

with $A_x = S^j$, and $A_y = \mathbb{1}$ if $y \neq x$.

The $\mathcal{A}_\Lambda$ are finite-dimensional C^*-algebras for the usual operations of sum, product, and Hermitian conjugation of matrices and with identity $\mathbb{1}_\Lambda$.

If $\Lambda_0 \subset \Lambda_1$, there is a natural embedding of $\mathcal{A}_{\Lambda_0}$ into $\mathcal{A}_{\Lambda_1}$, given by

$$\mathcal{A}_{\Lambda_0} \cong \mathcal{A}_{\Lambda_0} \otimes \mathbb{1}_{\Lambda_1\setminus\Lambda_0} \subset \mathcal{A}_{\Lambda_1}$$

The algebra of local observables is then defined by

$$\mathcal{A}_{\mathrm{loc}} = \bigcup_{\Lambda\in\mathcal{L}} \mathcal{A}_\Lambda$$

Its completion is the C^*-algebra of quasilocal observables, which we will simply denote by $\mathcal{A}$.

The dynamics and symmetries of a quantum spin model are described by (groups of) automorphisms of the C^*-algebra $\mathcal{A}$, that is, bijective linear transformations α on $\mathcal{A}$ that preserve the product and * operations. Translation invariance, for example, is expressed by the translation automorphisms τ_x, $x \in \mathbb{Z}^d$, which map any subalgebra $\mathcal{A}_\Lambda$ to $\mathcal{A}_{\Lambda+x}$, in the natural way. They form a representation of the additive group $\mathbb{Z}^d$ on $\mathcal{A}$.

A translation-invariant interaction, or potential, defining a quantum spin model, is a map $\phi : \mathcal{L} \to \mathcal{A}$ with the following properties: for all $X \in \mathcal{L}$, we have $\phi(X) \in \mathcal{A}_X$, $\phi(X) = \phi(X)^*$, and for $x \in \mathbb{Z}^d$, $\phi(X + x) = \tau_x(\phi(X))$. An interaction is called finite range if there exists $R > 0$ such that $\phi(X) = 0$ whenever $\mathrm{diam}(X) > R$. The Hamiltonian in Λ is the self-adjoint element of $\mathcal{A}_\Lambda$ defined by

$$H_\Lambda = \sum_{X\subset\Lambda} \phi(X)$$

For the standard Heisenberg model the interaction is given by

$$\phi(\{x, y\}) = -J\boldsymbol{S}_x \cdot \boldsymbol{S}_y, \text{ if } |x - y| = 1 \qquad [2]$$

and $\phi(X) = 0$ in all other cases. Here, $\boldsymbol{S}_x \cdot \boldsymbol{S}_y$ is the conventional notation for $S_x^1 S_y^1 + S_x^2 S_y^2 + S_x^3 S_y^3$. The magnitude of the coupling constant J sets a natural unit of energy and is irrelevant from the mathematical point of view. Its sign, however, determines whether the model is ferromagnetic ($J > 0$), or antiferromagnetic ($J < 0$). For the classical Heisenberg model, where the role of $\boldsymbol{S}_x$ is played by a unit vector in $\mathbb{R}^3$, and which can be regarded, after rescaling by a factor S^{-2}, as the limit $S \to \infty$ of the quantum Heisenberg model, there is a simple transformation relating the ferro- and antiferromagnetic models (just map $\boldsymbol{S}_x$ to $-\boldsymbol{S}_x$ for all x in the even sublattice of $\mathbb{Z}^d$). It is easy to see that there does not exist an automorphism of $\mathcal{A}$ mapping $\boldsymbol{S}_x$ to $-\boldsymbol{S}_x$, since that would be inconsistent with the commutation relations [1]. Not only is there no exact mapping between the ferro- and the antiferromagnetic models, their ground states and equilibrium states have radically different properties. See below for the definitions and further discussion.

The dynamics (or time evolution), of the system in finite volume Λ is the one-parameter group of automorphisms of $\mathcal{A}_\Lambda$ given by

$$\alpha_t^{(\Lambda)}(A) = \mathrm{e}^{\mathrm{i}tH_\Lambda} A \mathrm{e}^{-\mathrm{i}tH_\Lambda}, \quad t \in \mathbb{R}$$

For each $t \in \mathbb{R}$, $\alpha_t^{(\Lambda)}$ is an automorphism of $\mathcal{A}$ and the family $\{\alpha_t^{(\Lambda)} \mid t \in \mathbb{R}\}$ forms a representation of the additive group $\mathbb{R}$.

Each $\alpha_t^{(\Lambda)}$ can trivially be extended to an automorphism on $\mathcal{A}$, by tensoring with the identity map. Under quite general conditions, $\alpha_t^{(\Lambda)}$ converges strongly as $\Lambda \to \mathbb{Z}^d$ in a suitable sense, that is, for every $A \in \mathcal{A}$, the limit

$$\lim_{\Lambda \uparrow \mathbb{Z}^d} \alpha_t^{(\Lambda)}(A) = \alpha_t(A)$$

exists in the norm in $\mathcal{A}$, and it can be shown that it defines a strongly continuous one-parameter group of automorphisms of $\mathcal{A}$. $\Lambda \uparrow \mathbb{Z}^d$ stands for any sequence of $\Lambda \in \mathcal{L}$ such that Λ eventually contains any given element of $\mathcal{L}$. A sufficient condition on the potential ϕ is that there exists $\lambda > 0$ such that $\|\Phi\|_\lambda$ is finite, with

$$\|\Phi\|_\lambda = \sum_{X \ni 0} \mathrm{e}^{\lambda |X|} \|\phi(X)\| \qquad [3]$$

Here $|\cdot|$ denotes the number of elements in X. One can show that, under the same conditions, δ defined on $\mathcal{A}_{\mathrm{loc}}$ by

$$\delta(A) = \lim_{\Lambda \uparrow \mathbb{Z}^d} [H_\Lambda, A]$$

is a norm-closable (unbounded) derivation on $\mathcal{A}$ and that its closure is, up to a factor i, the generator of $\{\alpha_t \mid t \in \mathbb{R}\}$, that is, formally

$$\alpha_t = \mathrm{e}^{\mathrm{i}t\delta}$$

For the class of ϕ with finite $\|\Phi\|_\lambda$ for some $\lambda > 0$, $\mathcal{A}_{\mathrm{loc}}$ is a core of analytic vectors for δ. This means that, for each $A \in \mathcal{A}_{\mathrm{loc}}$, the function $t \mapsto \alpha_t(A)$ can be extended to a function $\alpha_z(A)$ analytic in a strip $|\mathrm{Im}\, z| < a$ for some $a > 0$.

A state of the quantum spin system is a linear functional on $\mathcal{A}$ such that $\omega(A^*A) \geq 0$, for all $A \in \mathcal{A}$ (positivity), and $\omega(\mathbb{1}) = 1$ (normalization). The restriction of ω to $\mathcal{A}_\Lambda$, for each $\Lambda \in \mathcal{L}$, is uniquely determined by a density matrix, that is, $\rho_\Lambda \in \mathcal{A}_\Lambda$, such that

$$\omega(A) = \mathrm{tr}\, \rho_\Lambda A, \quad \text{for all } A \in \mathcal{A}_\Lambda$$

where tr denotes the usual trace of matrices. ρ_Λ is non-negative definite and of unit trace. If the density matrix is a one-dimensional projection, the state is called a vector state, and can be identified with a vector $\psi \in \mathcal{H}_\Lambda$, such that $\mathbb{C}\psi = \mathrm{ran}\, \rho_\Lambda$.

A ground state of the quantum spin system is a state ω satisfying the local stability inequalties:

$$\omega(A^* \delta(A)) \geq 0, \quad \text{for all } A \in A_{\mathrm{loc}} \qquad [4]$$

The states describing thermal equilibrium are characterized by the Kubo–Martin–Schwinger (KMS) condition: for any $\beta \geq 0$ (related to absolute temperature by $\beta = 1/(k_{\mathrm{B}}T)$, where k_{B} is the Boltzmann constant), ω is called β-KMS if

$$\omega(A\alpha_{\mathrm{i}\beta}(B)) = \omega(BA), \quad \text{for all } A, B \in \mathcal{A}_{\mathrm{loc}} \qquad [5]$$

The most common way to construct ground states and equilibrium states, namely solutions of [4] and [5], respectively, is by taking thermodynamic limits of finite-volume states with suitable boundary conditions. A ground state of the finite-volume Hamiltonian H_Λ is a convex combination of vector states that are eigenstates of H_Λ belonging to its smallest eigenvalue. The finite-volume equilibrium state at inverse temperature β has density matrix ρ_β defined by

$$\rho_\beta = Z(\Lambda, \beta)^{-1} \mathrm{e}^{-\beta H_\Lambda}$$

where $Z(\Lambda, \beta) = \mathrm{tr}\, \mathrm{e}^{-\beta H_\Lambda}$ is called the partition function. By considering limit points as $\Lambda \to \mathbb{Z}^d$, one can show that a quantum spin model always has at least one ground state and at least one equilibrium state for all β.

In this section, the basic concepts have so far been discussed in the most standard setup. Clearly, many generalizations are possible: one can consider non-translation-invariant models; models with random potentials; the state spaces at each site may have different dimensions; instead of $\mathbb{Z}^d$ one can consider other lattices or define models on arbitrary graphs; one can allow interactions of infinite range that satisfy weaker conditions than those imposed by the finiteness of the norm [3], or restrict to subspaces of the Hilbert space by imposing symmetries or suitable hardcore conditions; and one can study models with infinite-dimensional spins. Examples of all these types of generalizations have been considered in the literature and have interesting applications.

Symmetries and Symmetry Breaking

Many interesting properties of quantum spin systems are related to symmetries and symmetry breaking. Symmetries of a quantum spin model are realized as representations of groups, Lie algebras, or quantum (group) algebras on the Hilbert space and/or the observable algebra. The symmetry property of the model is expressed by the fact that the Hamiltonian (or the dynamics) commutes with this representation. We briefly discuss the most common symmetries.

Translation invariance. The translation automorphisms τ_x have already been defined on the

observable algebra of infinite quantum spin systems on $\mathbb{Z}^d$. One can also define translation automorphisms for finite systems with periodic boundary conditions, which are defined on the torus $\mathbb{Z}^d/T\mathbb{Z}^d$, where $T=(T_1,\ldots,T_d)$ is a positive integer vector representing the periods.

Other graph automorphisms. In general, if G is a group of automorphisms of the graph Γ, and $\mathcal{H}_\Gamma=\bigotimes_{x\in\Gamma}\mathbb{C}^n$ is the Hilbert space of a system of identical spins defined on Γ, then, for each $g\in G$, one can define a unitary U_g on $\mathcal{H}_\Gamma$ by linear extension of $U_g\bigotimes\varphi_x=\bigotimes\varphi_{g^{-1}(x)}$, where $\varphi_x\in\mathbb{C}^n$, for all $x\in\Gamma$. These unitaries form a representation of G. With the unitaries one can immediately define automorphisms of the algebra of observables: for $A\in\mathcal{A}_\Lambda$, and $U\in\mathcal{A}_\Lambda$ unitary, $\tau(A)=U^*AU$ defines an automorphism, and if U_g is a group representation, the corresponding τ_g will be, too. Common examples of graph automorphisms are the lattice symmetries of rotation and reflection. Translation symmetry and other graph automorphisms are often referred to collectively as spatial symmetries.

Local symmetries (also called gauge symmetries). Let G be a group and u_g, $g\in G$, a unitary representation of G on $\mathbb{C}^n$. Then, $U_g=\bigotimes_{x\in\Lambda}u_g$ is a representation on $\mathcal{H}_\Lambda$. The Heisenberg model [2], for example, commutes with such a representation of SU(2). It is often convenient, and generally equivalent, to work with a representation of the Lie algebra. In that case the SU(2) invariance of the Heisenberg model is expressed by the fact that H_Λ commutes with the following three operators:

$$S^i=\sum_{x\in\Lambda}S^i_x,\quad i=1,2,3$$

Note: sometimes the Hamiltonian is only symmetric under certain combinations of spatial and local symmetries. CP symmetry is an example.

For an automorphism τ, we say that a state ω is τ-invariant if $\omega\circ\tau=\tau$. If ω is τ_g-invariant for all $g\in G$, we say that ω is G-invariant.

It is easy to see that if a quantum spin model has a symmetry G, then the set of all ground states or all β-KMS states will be G-invariant, meaning that if ω is in the set, then so is $\omega\circ\tau_g$, for all $g\in G$. By a suitable averaging procedure, it is usually easy to establish that the sets of ground states or equilibrium states contain at least one G-invariant element.

An interesting situation occurs if the model is G-invariant, but there are ground states or KMS states that are not. This means that, for some $g\in G$, and some ω in the set (of ground states or KMS states), $\omega\circ\tau\neq\omega$. When this happens, one says that there is spontaneous symmetry breaking, a phenomenon that also plays an important role in quantum field theory.

The famous Hohenberg–Mermin–Wagner theorem, applied to quantum spin models, states that, as long as the interactions do not have very long range and the dimension of the lattice is 2 or less, continuous symmetries cannot be spontaneously broken in a β-KMS state for any finite β.

Quantum group symmetries. We restrict ourselves to one important example: the $\mathrm{SU}_q(2)$ invariance of the spin-1/2 XXZ Heisenberg chain with $q\in[0,1]$, and with special boundary terms. The Hamiltonian of the $\mathrm{SU}_q(2)$-invariant XXZ chain of length L is given by

$$H_L=\sum_{x=1}^{L-1}-\frac{1}{\Delta}\left(S^1_xS^1_{x+1}+S^2_xS^2_{x+1}\right)-\left(S^3_xS^3_{x+1}-1/4\right)+\frac{1}{2}\sqrt{1-\Delta^{-2}}\left(S^3_{x+1}-S^3_x\right)$$

where $q\in(0,1]$ is related to the parameter $\Delta\geq 1$ by the relation $\Delta=(q+q^{-1})/2$. When $q=0$, H_L is equivalent to the Ising chain. Thus, the XXZ model interpolates between the Ising model (the primordial classical spin system) and the isotropic Heisenberg model (the most widely studied quantum spin model). In the limit of infinite spin ($S\to\infty$), the model converges to the classical Heisenberg model (XXZ or isotropic). An interesting feature of the XXZ model are its non-translation-invariant ground states, called kink states.

In this family of models, one can see how aspects of discreteness (quantized spins) and continuous symmetry (SU(2), or quantum symmetry $\mathrm{SU}_q(2)$) are present at the same time in the quantum Heisenberg models, and the two classical limits ($q\to 0$ and $S\to\infty$) can be used as a starting point to study its properties.

Quantum group symmetry is not a special case of invariance under the action of a group. There is no group, but there is an algebra represented on the Hilbert space of each spin, for which there is a good definition of tensor product of representations, and "many" irreducible representations. In this example, the representation of $\mathrm{SU}_q(2)$ on $\mathcal{H}_{[1,L]}$ commuting with H_L is generated by

$$S^3=\sum_{x=1}^{L}\mathbb{1}_1\otimes\cdots\otimes S^3_x\otimes\mathbb{1}_{x+1}\otimes\cdots\mathbb{1}_L$$

$$S^+=\sum_{x=1}^{L}t_1\otimes\cdots\otimes t_{x-1}\otimes S^+_x\otimes\mathbb{1}_{x+1}\otimes\cdots\mathbb{1}_L$$

$$S^-=\sum_{x=1}^{L}\mathbb{1}_1\otimes\cdots\otimes S^-_x\otimes t^{-1}_{x+1}\otimes\cdots t^{-1}_L$$

where

$$t = \begin{pmatrix} q^{-1} & 0 \\ 0 & q \end{pmatrix}$$

Quantum group symmetries were discovered in exactly solvable models, starting with the spin-1/2 XXZ chain. One can exploit their representation theory to study the spectrum of the Hamiltonian in very much the same way as ordinary symmetries. The main restriction to its applicability is that the tensor product structure of the representations is inherently one-dimensional, that is, relying on an ordering from left to right. For the infinite XXZ chain the left-to-right and right-to-left orderings can be combined to generate an infinite-dimensional algebra, the quantum affine algebra $U_q(\widehat{\mathrm{sl}_2})$.

Phase Transitions

Quantum spin models of condensed matter physics often have interesting ground states. Not only are the ground states often a good approximation of the low-temperature behavior of the real systems that are modeled by it, and studying them is therefore useful, it is in many cases also a challenging mathematical problem. This is in contrast with classical lattice models for which the ground states are usually simple and easy to find. In more than one way, ground states of quantum spin systems display behavior similar to equilibrium states of classical spin systems at positive temperature.

The spin-1/2 Heisenberg antiferromagnet on $\Lambda \subset \mathbb{Z}^d$, with Hamiltonian

$$H_\Lambda = \sum_{x,y;\in\Lambda|x-y|=1} S_x \cdot S_y \qquad [6]$$

is a case in point. Even in the one-dimensional case ($d = 1$), and even though the model in that case is exactly solvable by the Bethe ansatz, its ground state is highly nontrivial. Analysis of the Bethe ansatz solution (which is not fully rigorous) shows that spin–spin correlation function decays to zero at infinity, but slower than exponentially (roughly as inverse distance squared). For $d = 2$, it is believed, but not mathematically proved, that the ground state has Néel order, that is, long-range antiferromagnetic order, accompanied by a spontaneous breaking of the SU(2) symmetry. Using reflection positivity, Dyson, Lieb, and Simon were able to prove the Néel order at sufficiently low temperature (large β), for $d \geq 3$ and all $S \geq 1/2$. This was later extended to the ground state for $d = 2$ and $S \geq 1$, and $d \geq 3$ and $S \geq 1/2$, that is, all the cases where Néel order is expected except $d = 2$, $S = 1/2$.

In contrast, no proof of long-range order in the Heisenberg ferromagnet at low temperature exists. This is rather remarkable since proving long-range order in the ground states of the ferromagnet is a trivial problem.

Of particular interest are the so-called quantum phase transitions. These are phase transitions that occur as a parameter in the Hamiltonian is varied and which are driven by the competing effects of energy and quantum fluctuations, rather than the balance between energy and entropy which drives usual equilibrium phase transitions. Since entropy does not play a role, quantum phase transitions can be oberved at zero temperature, that is, in the ground states.

An important example of a quantum phase transition occurs in the two-or higher-dimensional XY model with a magnetic field in the Z-direction. It was proved by Kennedy, Lieb, and Shastry that, at zero field, this model has off-diagonal long-range order (ODLRO), and can be interpreted as a hard-core Bose gas at half-filling. It is also clear that if the magnetic field exceeds a critical value, h_c, the model has a simple ferromagnetically ordered ground state. There are indications that there is ODLRO for all $|h| < h_c$. However, so far there is no proof that ODLRO exists for any $h \neq 0$.

What makes the ground-state problem of quantum spin systems interesting and difficult at the same time is that ground states, in general, do not minimize the expectation value of the interaction terms in the Hamiltonian individually although, loosely speaking, the expectation value of their sum (the Hamiltonian) is minimized. However, there are interesting exceptions to this rule. Two examples are the AKLT model and the ferromagnetic XXZ model.

The wide-ranging behavior of quantum spin models has required an equally wide range of mathematical approaches to study them. There is one group of methods, however, that can make a claim of substantial generality: those that start from a representation of the partition function based on the Feynman–Kac formula. Such representations turn a d-dimensional quantum spin model into a $(d + 1)$-dimensional classical problem, albeit one with some special features. This technique was pioneered by Ginibre in 1968 and was quickly adopted by a number of authors to solve a variety of problems. Techniques borrowed from classical statistical mechanics have been adapted with great success to study ground states, the low-temperature phase diagram, or the high-temperature regime of quantum spin models that can be regarded as perturbations of a classical system. More recently, it was used to develop a quantum version of Pirogov–Sinai theory which is applicable to a large class of problems, including some with low-temperature phases not related by symmetry.

Dynamics

Another feature of quantum spin systems that makes them mathematically richer than their classical couterpart is the existence of a Hamiltonian dynamics. Quite generally, the dynamics is well defined in the thermodynamic limit as a strongly continuous one-parameter group of automorphisms of the C^*-algebra of quasilocal observables. Strictly speaking, a quantum spin model is actually defined by its dynamics α_t, or by its generator δ, and not by the potential ϕ. Indeed, ϕ is not uniquely determined by α_t. In particular, it is possible to incorporate various types of boundary conditions into the definition of ϕ. This approach has proved very useful in obtaining important structural results, such as the proof by Araki of the uniqueness the KMS state at any finite β in one dimension. Another example is a characterization of equilibrium states by the energy–entropy balance inequalities, which is both physically appealing and mathematically useful: ω is a β-KMS state for a quantum spin model in the setting of the section on the mathematical framework in this article (and in fact also for more general quantum systems), if and only if the inequality

$$\beta\omega(X^*\delta(X)) \geq \omega(X^*X)\log\frac{\omega(X^*X)}{\omega(XX^*)}$$

is satisfied for all $X \in \mathcal{A}_{\text{loc}}$. This characterization and several related results were proved in a series of works by various authors (mainly Roepstorff, Araki, Fannes, Verbeure, and Sewell).

Detailed properties of the dynamics for specific models are generally lacking. One could point to the "immediate nonlocality" of the dynamics as the main difficulty. By this, we mean that, except in trivial cases, most local observables $A \in \mathcal{A}_{\text{loc}}$, become nonlocal after an arbitrarily short time, that is, $\alpha_t(A) \notin \mathcal{A}_{\text{loc}}$, for any $t \neq 0$. This nonlocality is not totally uncontrolled however. A result by Lieb and Robinson establishes that, for models with interactions that are sufficiently short range (e.g., finite range), the nonlocality propagates at a bounded speed. More precisely, under quite general conditions, there exist constants $c, v > 0$ such that, for any two local observables $A, B \in \mathcal{A}_{\{0\}}$,

$$\|[\alpha_t(A), \tau_x(B)]\| \leq 2\|A\|\|B\|e^{-c(|x|-v|t|)}$$

Attempts to understand the dynamics have generally been aimed at one of the two issues: return to equilibrium from a perturbed state, and convergence to a nonequilibrium steady state in the presence of currents. Some interesting results have been obtained although much remains to be done.

Acknowledgment

This work was supported in part by the National Science Foundation under Grant # DMS-0303316.

See also: Bethe Ansatz; Channels in Quantum Information Theory; Eight Vertex and Hard Hexagon Models; Exact Renormalization Group; Falicov–Kimball Model; Finitely Correlated States; High *Tc* Superconductor Theory; Hubbard Model; Pirogov–Sinai Theory; Quantum Central-Limit Theorems; Quantum Phase Transitions; Quantum Statistical Mechanics: Overview; Reflection Positivity and Phase Transitions; Symmetry and Symmetry Breaking in Dynamical Systems; Symmetry Breaking in Field Theory.

Further Reading

Affleck I, Kennedy T, Lieb EH, and Tasaki H (1988) Valence bond ground states in isotropic quantum antiferromagnets. *Communications in Mathematical Physics* 115: 477–528.

Aizenman M and Nachtergaele B (1994) Geometric aspects of quantum spin states. *Communications in Mathematical Physics* 164: 17–63.

Araki H (1969) Gibbs states of a one dimensional quantum lattice. *Communications in Mathematical Physics* 14: 120–157.

Borgs C, Kotecký R, and Ueltschi D (1996) Low-temperature phase diagrams for quantum perturbations of classical spin systems. *Communications in Mathematical Physics* 181: 409–446.

Bratteli O and Robinson DW (1981, 1997) *Operator Algebras and Quantum Statistical Mechanics 2. Equilibrium States. Models in Quantum Statistical Mechanics.* Berlin: Springer.

Datta N, Fernández R, and Fröhlich J (1996) Low-temperature phase diagrams of quantum lattice systems. I. Stability for quantum perturbations of classical systems with finitely-many ground states. *Journal of Statistical Physics* 84: 455–534.

Dyson F, Lieb EH, and Simon B (1978) Phase transitions in quantum spin systems with isotropic and non-isotropic interactions. *Journal of Statistical Physics* 18: 335–383.

Fannes M, Nachtergaele B, and Werner RF (1992) Finitely correlated states on quantum spin chains. *Communications in Mathematical Physics* 144: 443–490.

Kennedy T (1985) Long-range order in the anisotropic quantum ferromagnetic Heisenberg model. *Communications in Mathematical Physics* 100: 447–462.

Kennedy T and Nachtergaele B (1996) *The Heisenberg model – a bibliography.* http://math.arizona.edu/~tgk/qs.html.

Kennedy T and Tasaki H (1992) Hidden symmetry breaking and the Haldane phase in $S=1$ quantum spin chains. *Communications in Mathematical Physics* 147: 431–484.

Lieb E, Schultz T, and Mattis D (1964) Two soluble models of an antiferromagnetic chain. *Annals of Physics (NY)* 16: 407–466.

Matsui T (1990) Uniqueness of the translationally invariant ground state in quantum spin systems. *Communications in Mathematical Physics* 126: 453–467.

Mattis DC (1981, 1988) *The Theory of Magnetism. I.* Berlin: Springer.

Simon B (1993) *The Statistical Mechanics of Lattice Gases. Volume I.* Princeton: Princeton University Press.

Quantum Statistical Mechanics: Overview

L Triolo, Università di Roma "Tor Vergata", Rome, Italy

Introduction

Quantum theory actually started at the beginning of the twentieth century as a many-body theory, attempting to solve problems to which classical physics gave unsatisfactory answers.

This article aims to follow the developments of quantum statistical mechanics, hereafter called QSM, staying close to the underlying physics and sketching its methods and perspectives. The next section outlines the historical path, and the first achievements by Planck (1900) and Debye (1913); the subsequent free quantum gas theory will be recalled in the first original insights due to Fermi, Dirac (1926) and Bose, Einstein (1924–25), when many open problems began to find a coherent treatment.

In this framework, an interesting new idea appeared: the elementary units of the systems could be "particles", in the usual or in a broad meaning, a notion which includes photons, phonons, and quasiparticles of current use in condensed matter physics. The description of a classical harmonic system through independent normal modes is an example of a very fruitful use of collective variables.

The subsequent section will deal with more recent achievements, related to the properties of quantum N-body systems, which are fundamental for the derivation of their macroscopic behavior. In particular, the works by Dyson–Lenard and Lieb–Lebowitz on the stability of matter have to be recalled: a system made of electrons and ions has a thermodynamic behavior, thanks to the quantum nature of its constituents, where the Pauli exclusion principle plays an essential role.

We will then present relations that arise in quantum field theory, that is, from the second quantization methods; related technical and conceptual problems will also be presented briefly.

This is necessary for taking into account the recent works and perspectives, which will be considered in the last section. Here the new inputs and challenges from outstanding achievements in physics laboratories will be taken into account, referring to some exactly solvable models which help in understanding and in fixing the boundaries of approximate methods.

The Crisis of Classical Physics: The Quantum Free Gas

Let us briefly recall some of what Lord Kelvin called the "nineteenth century clouds" over the physics of that time (1884), and the subsequent new ideas, (Gallavotti 1999).

It is well known that the classical Dulong–Petit law of specific heat of solid crystals may be derived from the model of point particles interacting through harmonic forces; the equipartition of the mean energy among the degrees of freedom implies, for N particles, the linear dependence of the internal energy U_N on absolute temperature T, hence a constant heat capacity C_N (k_B is the Boltzmann's constant)

$$U_N = 6 \cdot \frac{1}{2} N k_B T, \qquad C_N := \frac{\partial U_N}{\partial T} = 3 N k_B \qquad [1]$$

Experimentally this is relatively well satisfied at high temperatures but it is violated for low T: one observes that U_N vanishes faster than linearly as T goes to zero, so that C_N vanishes. Moreover, the contributions to the heat capacity from the internal degrees of freedom of the molecular gases or from the free electrons in conducting solids are negligible, at room temperature: these degrees of freedom, in spite of the equipartition principle, seem frozen.

The analysis of the blackbody radiation problem from the classical point of view, that is, using equipartition among the normal modes of the electromagnetic field in the "black" cavity at temperature T, gives the following dependence, Rayleigh–Jeans law (1900), of the spectral energy density $u(\nu, T)$, on frequency ν and temperature T (c is the speed of light in vacuum):

$$u(\nu, T) = \frac{8\pi\nu^2}{c^3} k_B T \qquad [2]$$

The experimental curves for any positive T show a maximum for a frequency $\nu_{max}(T)$ which increases linearly with T according to Wien's displacement law (1893). The spectral energy density decreases fast enough to zero as $\nu \to \infty$ in such a way that the overall (integrated) energy is (finite and) proportional to T^4, according to Stefan's law (1879); the agreement with the classical form holds for low frequencies. The analytic form of the classical $u(\nu, T)$ in [2] does not present maxima and the overall radiated energy is clearly divergent (this bad behavior for large ν, present in many formulas for other models, sometimes in the corresponding "short-distance" form, is called an "ultraviolet catastrophe").

The effort by M Planck (1900) to understand the right dependence of u from ν and T was based on a thermodynamic argument about the possible energy–entropy relation, and on an assumption similar to the discretization rules on which the "old quantum theory" for the atomic structure is based. The electromagnetic field is represented, via Fourier analysis, as a set of infinitely many independent harmonic oscillators, two for every wave vector $\boldsymbol{k}$, to take into account the polarization. The frequency depends linearly on the wave number $k = |\boldsymbol{k}|$ (linear dispersion law), and the spacing becomes negligible for macroscopic dimensions of the cavity. The key idea for computing the partition function is the discretization of the phase space of each oscillator (of frequency $\nu = \omega/2\pi$). Putting there the adimensionalized Lebesgue measure $\mathrm{d}p\,\mathrm{d}q/h$, where h is a constant with physical dimensions of an action, we consider the regions R_E bounded by the constant-energy ellipses and their areas $|R_E|$, and find

$$|R_E| = \int_{R_E} \frac{\mathrm{d}p\mathrm{d}q}{h} = \frac{2\pi E}{h\omega} = \frac{E}{h\nu}$$

If these adimensional areas have integer values, that is, $E = nh\nu, n = 0, 1, 2, \ldots$, the annular region ("cell" C_n) between R_{E_n} and $R_{E_{n+1}}$ has unit area and so we approximate the partition function with the series ($\beta = 1/(k_\mathrm{B}T)$, the ubiquitous parameter in statistical mechanics, often called "inverse temperature")

$$Z_{\mathrm{discr}} = \sum_n \exp(-\beta nh\nu) = \frac{1}{1 - \exp(-\beta h\nu)}$$

In this way, the probabilistic weight given to this cell is

$$\begin{aligned} p(C_n) &= \frac{\exp(-\beta nh\nu)}{\sum_j \exp(-\beta jh\nu)} \\ &= \exp(-\beta nh\nu)(1 - \exp(-\beta h\nu)) \end{aligned} \qquad [3]$$

A well-defined value for the constant h (i.e., $h = 6.626\ldots \times 10^{-27}$ erg s, the Planck constant), combined with the usual computation for the density of states, gives a formula which quantitatively agrees with experimental data (see **Figure 1**)

$$u(\nu, T) = \frac{8\pi\nu^2}{c^3} \frac{h\nu}{\exp(h\nu/k_\mathrm{B}T) - 1} \qquad [4]$$

Moreover, for a certain range of parameters, that is, such that $\beta h\nu \ll 1$, there is agreement with the classical law.

The "quantum of light," introduced by Einstein in 1905 in his work on photoelectric effect, was later (1926) called photon by G N Lewis. The picture for representing the radiating system was one of a gas of noninteracting photons, carrying energy and momentum, and being continuously created and absorbed.

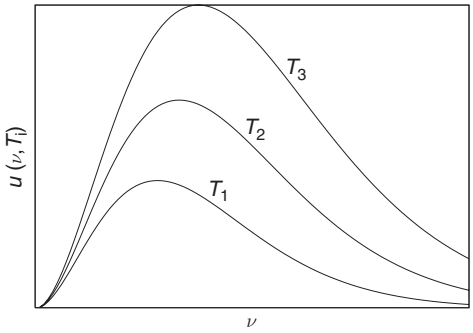


Figure 1 Dependence of the electromagnetic energy density on ν, for $T_1 < T_2 < T_3$.

A slightly different approach was used about the same time, for the problem of specific heat of crystalline solids.

The simpler model considers N points on the nodes of the lattice $\mathbb{Z}^3$, in a cubic box of side L, and interacting through harmonic forces; similarly to the radiation problem, the system is represented by a collection of independent harmonic oscillators (normal modes), which are "quantized" as before: the corresponding quanta were called phonons (by Fraenkel, in 1932) for the role of the acoustic band of frequencies. In this simplified approach (by Debye, in 1913) the different phonons are determined by a finite set of wave vectors

$$\boldsymbol{k} = \frac{2\pi}{L}\boldsymbol{n}, \quad n_i \text{ integer}, \ i = 1, 2, 3; \ |\boldsymbol{k}| \le k_\mathrm{M}$$

where the maximal modulus k_M is such that the total number of different $\boldsymbol{k}$'s is $3N$ (degrees of freedom).

Moreover, the frequency–wave number relation is simplified too, extrapolating the low-frequency (acoustic) linear relation $\nu = |\boldsymbol{k}|v_0$ (v_0 is the sound speed). In this way, the density of states which is quadratic in the frequency, has a cutoff to zero at the maximal frequency, ν_D, corresponding to $|k_\mathrm{M}|$, with an associate temperature $\Theta_\mathrm{D} = h\nu_\mathrm{D}/k_\mathrm{B}$ (Debye's temperature). The expected energy U_N in the canonical ensemble, after the computation of the canonical partition function, is given in term of the Debye function $D(\cdot)$:

$$\begin{aligned} U_N &= 3Nk_\mathrm{B}TD\left(\frac{\Theta}{T}\right) \\ D(y) &:= \frac{3}{y^3}\int_0^y \frac{x^3\mathrm{d}x}{\exp x - 1} \end{aligned} \qquad [5]$$

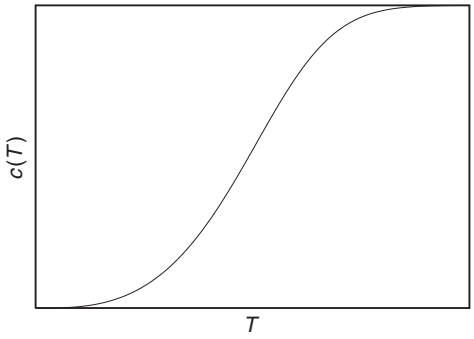


Figure 2 The specific heat of crystal solids according to Debye.

The agreement with experimental data, for the specific heat of different materials (i.e., different Θ_D), at low and high temperatures, is rather good. At low temperatures, one recovers the empirical T^3 behavior (see Figure 2). More careful measurements at low T put into evidence, for metallic solids, the role of the conduction electrons: their contribution to the heat capacity turns out to be linear in T, with a coefficient such that at room temperature it is much smaller than the lattice contribution, so that a satisfactory agreement with the classical law is found.

Soon after the beginning of quantum mechanics in its modern form (1925–26), physicists considered many-particle systems, dealing initially with the simplest situations, with a relatively easy formal apparatus, yet sufficient enough to understand in the main lines the "anomalous," that is, nonclassical, behavior.

For a system of N free particles in a cubic box of side L, quantum theory brings the labeling of the one-particle states with the wave vectors $\boldsymbol{k}$, recalling the de Broglie relation for the momentum $\boldsymbol{p}=\hbar\boldsymbol{k}$, with a possible additional spin (intrinsic angular moment) label σ

$$\boldsymbol{k}=\frac{2\pi}{L}\boldsymbol{n},\quad \boldsymbol{n}\in\mathbb{Z}^3\backslash\{0\}$$

and the statistics of the particles: because of indistinguishability, the wave function of several identical particles has to be symmetric (B–E, Bose–Einstein statistics) or antisymmetric (F–D, Fermi–Dirac statistics) in the exchange of the particles. This has the deep implication that no more than one fermion shares the same quantum state.

We may here recall the spin–statistics connection, which, in the framework of a local relativistic theory, states that integer spin particles are bosons, while particles with half-odd-integer spin are fermions.

As the state is completely defined by the knowledge of occupation numbers $n_{\boldsymbol{k},\sigma}$, we have the simple and relevant statement on the ground states for the N spinless bosons and N spin-1/2 fermion systems are described by the statement:

$$\begin{aligned}&\text{B–E system}: && n_{\boldsymbol{k}}=N\delta_{\boldsymbol{k},\boldsymbol{k}_0}\\&\text{F–D system}: && n_{\boldsymbol{k},\sigma}=\mathbf{1}(|\boldsymbol{k}|\le k_F)\forall\sigma\end{aligned}\qquad[6]$$

The constant k_F (Fermi wave number), or the equivalent $p_F=\hbar k_F$ and $\varepsilon_F=p_F{}^2/2m$ (Fermi momentum and energy, respectively) denotes the higher occupied level. In the continuum approximation, this implies the following relation between Fermi energy ε_F and density $\rho=N/L^3$:

$$\varepsilon_F=\frac{\hbar^2}{2m}(3\pi^2\rho)^{2/3}\qquad[7]$$

Going to the positive-temperature case, the grand canonical partition function is computed by considering that occupation numbers are non-negative integers for the B–E case and just 0 or 1 for the F–D case. This implies the simple formulas, with obvious meaning of symbols and leaving more details to the vast literature (see Figure 3):

$$\begin{aligned}&<n_{\boldsymbol{k},\sigma}>_{\beta,\mu}\\&=\frac{1}{\exp(\beta(\varepsilon_{\boldsymbol{k}}-\mu))\pm1},\quad +\text{for F–D},\ -\text{for B–E}\end{aligned}\qquad[8]$$

It is useful to introduce the Fermi temperature $T_F=\varepsilon_F/k_B$; using some realistic data, that is, for common metals like copper, T_F ranges roughly between 10^4–10^5 K, that is, well above the "normal," room temperatures: the quantum nature (i.e., quantum degeneracy) of the conduction electrons, modeled as free electrons, is macroscopically visible in normal conditions.

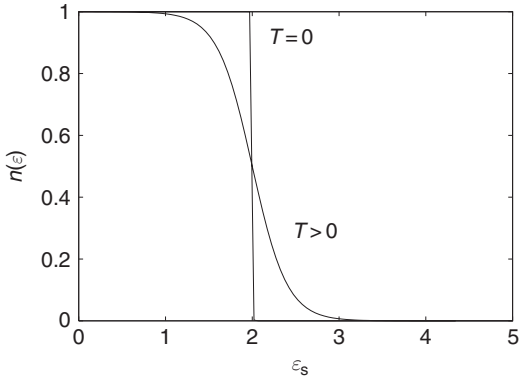


Figure 3 Expected fermionic occupation number, for $T=0$, $T>0$, and $\mu=2$.

The presence of an external field, like the periodic one given by the ionic lattice of a crystal, changes the situation in a relevant way, as the one-particle spectrum generally gets a band structure, and the allowed momenta are described in the reciprocal lattice: the Fermi sphere becomes a surface, and its structure is central for further developments.

For massive bosons, the strange superfluid features of liquid ^{4}He at low temperature, that is, below the critical value 2.17 K, led F London, just after Kapitza's discovery in 1937, to speculate that these were related to a macroscopic occupation of the ground state (B–E condensation). A more realistic model has to take into account interaction between bosons (see last section) as the microscopic interactions in superfluid liquid ^{4}He are not negligible.

Quantum *N*-Body Properties: Second Quantization

The main step in analyzing a quantum N-body system is its energy spectrum, and in particular its ground state, as it may represent a good approximation of the low-temperature states: its structure, the relations with possible symmetries of the Hamiltonian, its degeneracy, the dependence of its energy on the number of particles, are further relevant questions. The last one is related to the possibility of defining a thermodynamics for the system (Ruelle 1969). As a physically very interesting example, consider a system of electrically charged particles, N electrons with negative unit charge, and K atoms with positive charge z, say, interacting through electrostatic forces; the classical Coulomb potential as a function of distance behaves badly, as it diverges at zero and decreases slowly at infinity. The first question is about the stability: thanks to the exclusion principle, for the ground-state energy $E^0_{N,K}$ an extensive estimate from below is valid:

$$E^0_{N,K} \geq -c_0(N + Kz)$$

so that a finite-volume grand partition function exists, while for the thermodynamic limit, which involves large distances, we need more, that is, charge neutrality, which allows for screening, and a fast-decreasing effective interaction.

Let us see an example (quantum spin, Heisenberg model) belonging to the class of lattice models, where the identical microscopic elements are distinguishable by their fixed positions, that is, the nodes of a lattice like $\mathbb{Z}^d$. To any site $x \in \mathbb{Z}^d$ is associated a copy $\mathcal{H}_x$ of a $(2s+1)$-dimensional Hilbert space $\mathcal{H}$, where an irreducible unitary representation of SU(2) is given, so that the nonzero values for s are $1/2, 1, 3/2, \ldots$. For any x, the generators $S_\alpha(x), (\alpha = 1, 2, 3)$ satisfy the well-known commutation relations of the angular momentum; moreover, $\sum_\alpha {S_\alpha}^2(x) = s(s+1)\mathbf{1}$, and operators related to different sites commute. The ferromagnetic, isotropic, next-neighbors, magnetic field Hamiltonian for the finite system is

$$H_\Lambda = -J \sum_{<x,y>} \boldsymbol{S}(x) \cdot \boldsymbol{S}(y) - h \sum_x S_3(x) \qquad [9]$$

where J is the positive strength of the next-neighbors coupling ($<x, y>$ means that x and y are next neighbors); h is the intensity of the magnetic field oriented along the third axis. This model is considerably studied even now with several variants regarding possible anisotropies of the interaction, the possibly infinite range of the interaction, and the sign of J, for other (e.g., antiferromagnetic) couplings. Among the relevant results, the Mermin–Wagner theorem, at variance with the analogous classical spin model, states the absence of spontaneous magnetization in this zero-field model for $d = 2$ for any positive temperature; this can also be formulated as absence of symmetry breaking for this model (Fröhlich and Pfister in 1981 shed more light on this point).

As mentioned earlier, a useful mathematical tool for dealing with quantum systems of many particles or quasiparticles, is the occupation-number representation for the state of the system. The vector space for a system with an indefinite number of particles is the Fock space: it is the direct sum of all spaces with any number of particles, starting with the zero-particle, vacuum state. The operators which connect these subspaces are the creation and annihilation operators, very similar to the raising and lowering operators introduced by Dirac for the spectral analysis of the harmonic-oscillator Hamiltonian and the angular momentum, in the context of one-particle quantum theory.

It is perhaps worth sketching the action of these operators on the Fock space.

We consider spinless bosons first, as spin might easily be taken into account, if necessary. We suppose that a one-particle Hamiltonian has eigenfunctions labeled by a set of quantum numbers $\boldsymbol{k}$, say, as the wave vector for the purely kinetic one-particle Hamiltonian. Let $|n_{k_1}, n_{k_2}, \ldots, n_{k_p}>$ denote a vector state with $\sum_{i=1,\ldots,p} n_{k_i}$ particles, where n_{k_i} denotes the number of particles with wave vector $\boldsymbol{k}_i, i = 1, \ldots, p$; $|0>$ denotes the no-particle, vacuum state. We define the creation operators $a^*_{\boldsymbol{k}}$ as follows:

$$a^*_{\boldsymbol{k}} |\ldots n_{\boldsymbol{k}}, \ldots\rangle = \sqrt{n_{\boldsymbol{k}} + 1}\, |\ldots, n_{\boldsymbol{k}} + 1, \ldots\rangle \qquad [10]$$

Its adjoint a_k is called the annihilation operator, for its action on the vectors

$$a_k|\ldots n_k,\ldots\rangle = \sqrt{n_k}|\ldots,n_k-1,\ldots\rangle \qquad [11]$$

The operator a_k^* creates a new particle with that momentum: for any $\boldsymbol{k}$

$$a_k^* a_k|\ldots n_k,\ldots\rangle = n_k|\ldots n_k,\ldots\rangle$$

$a_k^* a_k := \hat{n}_k$ (the occupation-number operator)

The vacuum state belongs to $\mathrm{Ker}\, a_k$ for any $\boldsymbol{k}$, and the whole space is generated by application of creation operators on the vacuum state.

The following basic commutation relations, for any $\boldsymbol{k}, \boldsymbol{k}'$, are valid:

$$[a_k, a_{k'}^*] = \delta(\boldsymbol{k},\boldsymbol{k}'), \quad [a_k, a_{k'}] = [a_k^*, a_{k'}^*] = 0 \qquad [12]$$

For fermions, multiple occupancy is forbidden, so that the analogous annihilation (α_k) and creation (α_k^*) operators satisfy anticommutation relations:

$$[\alpha_k, \alpha_{k'}^*]_+ = \delta(\boldsymbol{k},\boldsymbol{k}'), \quad [\alpha_k, \alpha_{k'}]_+ = [\alpha_k^*, \alpha_{k'}^*]_+ = 0 \qquad [13]$$

The presence of spin is dealt by an additional spin label σ to these symbols, and a $\delta(\sigma,\sigma')$, where necessary.

The Hamiltonian for a system of particles, say spinless bosons, in a box Λ, made of its kinetic part together with a two-body ($\lambda v(x-y)$) interaction, is written in terms of the "field operators"; if $\{\phi_k(x)\}$ are the one-particle eigenfunctions of the single-particle purely kinetic Hamiltonian for the spinless case, and their complex conjugates are $\{\phi_k^*(x)\}$, we define the fields

$$\Phi(x) = \sum_k \phi_k(x) a_k, \quad \Phi^*(x) = \sum_k \phi_k^*(x) a_k^* \qquad [14]$$

So that the full Hamiltonian is given by

$$H_\Lambda = \int_\Lambda \mathrm{d}x\, \Phi^*(x)\left(\frac{-\hbar^2}{2m}\right)\Delta\Phi(x) + \lambda \int_\Lambda \mathrm{d}x \int_\Lambda \mathrm{d}y\, v(x-y)\Phi^*(x)\Phi(x)\Phi^*(y)\Phi(y) \qquad [15]$$

We mention that a theoretical breakthrough in the analysis of superfluidity was made by Bogoliubov (1946), who, starting from the Hamiltonian in [15], introduced the following Hamiltonian in the momentum representation:

$$H_\Lambda = \sum_k \varepsilon_k a_k^* a_k + \frac{1}{2}\sum_{k,k',q} \hat{v}_q a_{k-q}^* a_{k'+q}^* a_k a_{k'} \qquad [16]$$

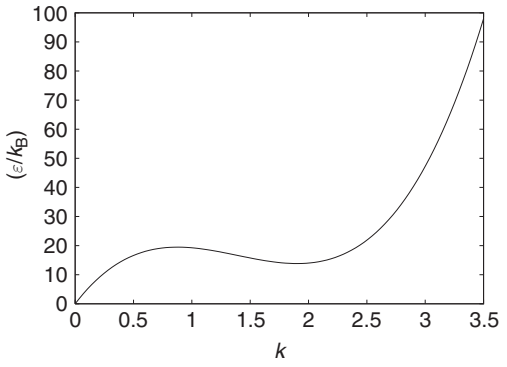


Figure 4 Excitation spectrum for superfluids.

where ε_k is the one-particle kinetic energy and $\hat{v}_q$ is the Fourier transform of the two-body potential. To study the excitation spectrum above the ground state, he introduced an approximation about the persistency of a macroscopic occupation of the ground state and a diagonalization procedure leading to new quasiparticles with a characteristic energy spectrum, linearly increasing near $|\boldsymbol{k}| = 0$, then presenting a positive minimum before the subsequent increase (see Figure 4).

Some Mathematical Tools for Macroscopic Quantum Systems

The formal apparatus of second quantization, born in the context of the quantum field theory, brought to statistical mechanics new ideas and techniques and related difficulties. For instance, the renormalization group was conceived in the 1970s to deal both with critical phenomena (i.e., power singularities of thermodynamic quantities around the critical point) and with divergences in quantum field theory. This subject is currently being developed and applied in models of quantum statistical mechanics (QSM) (Benfatto and Gallavotti, 1995).

Another issue, which has again strong relations with quantum field theory, is the algebraic formulation of QSM. This point of view, which is well suited for the analysis of infinitely extended quantum systems, uses a unified, synthetic, and rigorous language. The procedure for passing from a finite quantum system to its infinitely extended version deserves some attention.

It is well known that, for finite quantum systems, say N particles in a box Λ, an observable is represented by a self-adjoint operator A on a Hilbert space $\mathcal{H}_\Lambda$, and the normalized elements $\{|\psi>\}$ of this space are the pure states ρ_ψ which define the expectations

$$\rho_\psi(A) := <\psi|A\psi>$$

The mixed states (mixtures) are defined by convex combinations of pure states, the coefficients having an obvious statistical meaning.

Among the observables, the Hamiltonian plays a special role, as it generates the dynamics of the system, which evolves the pure states through the unitary group (Schrödinger picture)

$$\left|\psi(t) \geq \exp\left(-\frac{itH_\Lambda}{\hbar}\right)\right|\psi>$$

To the notion of equilibrium probability measure on the phase space of a classical system, corresponds the mixed state $\rho_{H_\Lambda,\beta}$ such that

$$\rho_{H_\Lambda,\beta}(A) := Z_{\Lambda,\beta}{}^{-1}\,\mathrm{tr}(\exp(-\beta H_\Lambda)A) \qquad [17]$$

The normalization factor $Z_{\Lambda,\beta} = \mathrm{tr}(\exp(-H_{\Lambda\beta}))$ is the canonical partition function.

Consider now the algebra $\mathcal{A}(\Lambda)$ of local observables; sending Λ to infinity, by induction, it is possible to define the algebra $\mathcal{A}$ of quasilocal observables. The main point is a set of algebraic relations like the canonical commutation relations (CCRs) and the canonical anticommutation relations (CARs) for the creation/annihilation operators: the observables of $\mathcal{A}$, through the GNS (Gel'fand, Naimark, and Segal) construction may be represented as operators on the appropriate Hilbert spaces, depending on the chosen state; the representations, at variance with the finite case, might be inequivalent. It is possible to define the equilibrium state for the infinite system and how to insert in a natural way the possible group invariance of the system ($\mathbb{R}^d$ or $\mathbb{Z}^d$, typically), ending with characterization of the pure phases of the system as the ergodic components in the decomposition of an equilibrium state. These states have the property that coarse-grained observables have sharp values (Ruelle 1969, Sewell 2002): if $\mathrm{Av}_l(A)$ is the space average on scale l, that is, over boxes of side l, for an ergodic state ρ,

$$\lim_{l\to\infty} \rho([\mathrm{Av}_l(A) - \rho(A)]^2) = 0$$

Another issue which is worth mentioning is the characterization of equilibrium states through the KMS (Kubo–Martin–Schwinger) condition. The strong formal similarity between the finite-volume quantum evolution operator $\alpha_t := \exp(-itH_\Lambda/\hbar)$ and the statistical equilibrium density operator $\exp(-\beta H_\Lambda)$, leads to the identity, valid for any couple of bounded observables A and B, using the short symbol $<\cdot>_{\beta,\Lambda}$ for the expectations with respect to the statistical operator:

$$<A_t B>_{\beta,\Lambda} = <BA_{t+i\hbar\beta}>_{\beta,\Lambda} \qquad [18]$$

This relation is suitably extended for infinite size, and therefore defines a KMS state; it implies some physically relevant properties like stability with respect to local disturbances and dissipativity (Sewell 2002).

A final issue in this section concerns another formalism stemming from the Feynman path-integral formulation of quantum mechanics: here a functional integral represents the statistical equilibrium density operator $W_\beta = \exp(-\beta H)$. For a d-dimensional system of N particles in a potential field ($X \in \mathbb{R}^{dN}$) $V = V(X) = \sum_{i<j}\phi(x_i - x_j)$ and Hamiltonian $H = -(1/2)\,\Delta + V$ the Feynman–Kac formula which, for a test function χ, may be written as follows:

$$(W_\beta\chi)(X) = \int P^\beta_{X,Y}(\mathrm{d}\omega)\exp\left(-\int_0^\beta \mathrm{d}s V(\omega(s))\chi(Y)\mathrm{d}Y\right)$$

where $P^\beta_{X,Y}(\mathrm{d}\omega)$ is the Wiener measure on the space of paths $\{\omega(s), s \in [0,\beta]\}$. For details on the construction and several other related features on the treatment of the different statistics, see Glimm and Jaffe (1981).

New Problems and Challenges

In this final section, we recall some phenomena which have been observed recently in physics laboratories, and which presumably deserve considerable efforts to overcome the heuristic level of explanation. About this last point, it is worth quoting a method that has been used to get results even without clear justifications of the underlying hypotheses, that is, the mean-field procedure. It started with the Curie–Weiss theory of magnetism and is based on the following drastic simplification: the microscopic element of the system feels an average interaction field due to other elements, indipendently of the positions of the latter. This method might provide relatively good results if the range of the interaction is very large, and in fact, a clear version with due limiting procedure was introduced by Kac, and applied by Lebowitz and Penrose in the 1960s for a microscopic derivation of van der Waals equation, and soon extended by Lieb to quantum systems.

We will briefly outline some aspects of three recent achievements of condensed matter physics for which modeling is still on the way of further progress: the B–E condensation, the high-T_c superconductivity, and the fractional quantum Hall effect. The first consists in trapping an ultracold (at less than 50 μK) dilute bosonic gas, for example, 10^4–10^7 atoms of ^{87}Rb, finding experimental evidence for Bose condensation. To understand the

properties of this system, an important tool is the Gross–Pitaevskii energy functional for the condensate wave function Φ,

$$E[\Phi] = \int \mathrm{d}x \left[\frac{\hbar^2}{2m} |\nabla \Phi|^2 + V_{\mathrm{ext}}(x) |\Phi|^2 + \frac{g}{2} |\Phi|^4 \right]$$

where the quartic term represents the reduced (mean-field) interaction among particles.

The second issue, that is, the high-temperature superconductivity, certainly deserves much attention. It has been observed recently in some ceramic materials well above 100 K, and a clear model which takes into account the formation of pairs and the peculiar isotropy–anisotropy aspects of the normal conductivity and superconductivity is still lacking (Mattis 2003).

Finally, let us consider the fractional quantum Hall effect; recall that the integer version, that is, a discretization of the Hall resistivity R_{H} by multiples of $h/(e^2)$, finds an explanation in terms of band spectra, formation of magnetic Landau levels, and localization from surface impurities, that is, without taking into account direct interactions among electrons.

The fractional discretization of R_{H} (Störmer 1999) has a theoretical interpretation, in terms of subtle collective behavior of the two-dimensional semiconductor electron system: the quasiparticles which represent the excitations may behave as composite fermions or bosons, or exhibit a fractional statistics (*see* Fractional Quantum Hall Effect).

This brief excursion through these new fascinating phenomena shows the rich interplay between theory and experiments: these phenomena are a source of new ideas and suggest new models for further progress.

See also: Bose–Einstein Condensates; Dynamical Systems and Thermodynamics; Exact Renormalization Group; Falicov–Kimball Model; Fermionic Systems; Finitely Correlated States; Fractional Quantum Hall Effect; High T_c Superconductor Theory; Hubbard Model; Quantum Phase Transitions; Quantum Spin Systems; Stability of Matter.

Further Reading

Benfatto G and Gallavotti G (1995) *Renormalization Group*. Princeton: Princeton University Press.

Gallavotti G (1999) *Statistical Mechanics: A Short Treatise*. Berlin: Springer.

Glimm J and Jaffe A (1981) *Quantum Physics. A Functional Integral Point of View*. New York: Springer.

Landau LD and Lifschitz EN (2000) Statistical Physics, *Course of Theoretical Physics*, 3rd edn., vol 5. Parts I and II. Oxford: Butterworth-Heinemann.

Mattis DC (2003) *Statistical Mechanics Made Simple*. NJ: World Scientific.

Ruelle D (1969) *Statistical Mechanics. Rigorous Results*. New York: Benjamin.

Sewell GL (2002) *Quantum Mechanics and Its Emergent Macrophysics*. Princeton: Princeton University Press.

Sinai YaG (1982) *Theory of Phase Transitions: Rigorous Results*. Budapest: Akadémiai Kiadó.

Störmer HL (1999) Nobel lecture: the quantum Hall effect. *Reviews of Modern Physics* 71: 875–889.

Taylor PL and Heinonen O (2002) *A Quantum Approach to Condensed Matter Physics*. Cambridge: Cambridge University Press.

Quasiperiodic Systems

P Kramer, Universität Tübingen, Tübingen, Germany

Introduction: From Periodic to Quasiperiodic Systems

Periodic systems occur in many branches of physics. Their mathematical analysis was stimulated in particular by the analysis of the periodic translational symmetry of crystals. The systematic study of the compatibility between translational and crystallographic point or reflection symmetry leads to the concept of space group symmetry. Mathematical crystallography in three dimensions (3D) culminated in 1892 in the complete classification of the 230 space groups due to Fedorov and Schoenflies (see Schwarzenberger (1980, pp. 132–135). One characteristic property of periodic systems is that their Fourier transform has a pure point spectrum. Since the Fourier spectrum is experimentally accessible through diffraction experiments, it provides a main tool for the structure determination of crystals.

With quantum mechanics in the twentieth century, it became possible to describe crystal structures quantitatively as ordered systems of atomic nuclei and electrons with electromagnetic interactions. The representation theory of crystallographic space groups now opened the way to verify the

space group symmetry of atomic systems for example from the band structure of crystals. It was then believed that in physics atomic long-range order is linked to periodicity and hence to the paradigm of the 230 space groups in 3D.

Mathematical analysis beyond this paradigm started independently in various directions. Bohr (1925) studied quasiperiodic functions and their Fourier transform. He interpreted them as restrictions of periodic functions in nD to their values on a linear subspace of orientation irrational with respect to a lattice. Mathematical crystallography in general dimension $n > 3$, including point group symmetry, was started around 1949 in work by Hermann and by Zassenhaus (see Schwarzenberger (1980)), and completed in 1978 for $n = 4$ in Brown *et al.* (1978). A different route was taken by Penrose (1974). He constructed an aperiodic tiling (covering without gaps or overlaps) of the plane. Its tiles in two rhombus shapes provide global 5-fold point symmetry and make the tiling incompatible with any periodic lattice in 2D. The connection between Penrose's aperiodic tiling and irrational subspaces in periodic structures was made by de Bruijn (1981). He interpreted the Penrose rhombus tiling as the intersection of geometric objects from cells of a hypercubic lattice in 5D with a 2D subspace, irrational and invariant under 5-fold noncrystallographic point symmetry. Kramer and Neri (1984) embedded the icosahedral group as a point group into the hypercubic lattice in 6D and constructed a 3D irrational subspace invariant under the noncrystallographic icosahedral point group. From intersections of boundaries of the hypercubic lattice cells with this subspace, they constructed a 3D tiling of global icosahedral point symmetry with two rhombohedral tiles.

Shechtman *et al.* (1984) discovered in the system *AlMn* diffraction patterns of icosahedral point symmetry. Since icosahedral symmetry is incompatible with a lattice in 3D, they concluded that there exists atomic long-range order without a lattice. The new paradigm of quasiperiodic long-range order in quasicrystals was established and since then stimulated a broad range of theoretical and experimental research.

The interplay between the notions – (1) of crystallographic symmetry in nD, $n > 3$, (2) of subspaces invariant under a point group but irrational and hence incompatible with a lattice, and (3) of discrete geometric periodic objects in nD providing quasiperiodic tilings on these subspaces – forms the mathematical basis for a new quasiperiodic long-range order found in quasicrystals. The present-day theory of quasicrystals offers the most elaborate study of quasiperiodic systems. Therefore, we shall focus in what follows on the concepts developed in this theory.

In the following section, we briefly review basic concepts of periodic systems and lattices in nD, their classification in terms of point symmetry and space groups, and their cell structure. In a section on quasiperiodic point sets and functions, a quasiperiodic system is taken as a geometric object on an irrational mD subspace in an n-dimensional space and lattice. Noncrystallographic point symmetry is shown to select the irrational subspace. Next, scaling symmetry in quasiperiodic systems is demonstrated. Then, examples of quasiperiodic systems with point and scaling symmetry are given. The penultimate section discusses quasiperiodic tilings and their windows. Finally, the notion of a fundamental domain for quasiperiodic functions compatible with a tiling is illustrated.

Concepts from Periodic Systems

A distribution $f^p(x)$ of geometric objects on Euclidean space E^n (a real linear space equipped with standard Euclidean scalar product $\langle , \rangle$ and metric) with coordinates $x \in E^n$ is called "periodic" if it is invariant under translations b^i in n linearly independent directions,

$$(\boldsymbol{p}) : f^p : f^p(x + b^i) = f^p(x), \quad i = 1, \ldots, n \qquad [1]$$

The set of all translations on E^n forms the discrete additive abelian translation group

$$T = \left\{ b \in E^n : b = \sum_i^n m_i b^i, (m_1, \ldots, m_n) \in \boldsymbol{Z}^n \right\} \qquad [2]$$

Any orbit (set of all images of an initial point) under the action $T \times E^n \rightarrow E^n$ yields a lattice Λ on E^n. Since T acts fixpoint-free, there is a one-to-one correspondence $\Lambda \leftrightarrow T$. A fundamental domain on E^n is defined as a subset of points $x \in E^n$ which contains a representative point from any orbit under T. Such a fundamental domain can be chosen, for example, as the unit cell of the lattice Λ or as the Voronoi cell (eqn [5]). By eqn [1], the functional values on E^n of a periodic function $f^p(x)$ are completely determined from its values on a fundamental domain of E^n.

Given the lattice basis $(b^1, \ldots, b^n)$ of eqn [2] in E^n, the vector components of the basis form the $n \times n$ basis matrix B of Λ. The most general change of the basis preserving the lattice is given by acting with any element h of the general linear group

$\mathrm{Gl}(n, \mathbf{Z})$, with integral matrix entries and determinant ± 1, on the lattice basis,

$$\mathrm{Gl}(n, \mathbf{Z}) \ni h : B \to B' = Bh \qquad [3]$$

The crystallographic classification of inequivalent lattices in E^n starts from $\mathrm{Gl}(n, \mathbf{Z})$. In addition to translations, it employs crystallographic point symmetry operations, (Brown *et al.* 1978, p. 9). A crystallographic point group operation of a lattice Λ is a Euclidean isometry g which belongs to a group $G \ni g$ with representations $D : G \to O(n, R)$ and $\mathcal{D} : G \to \mathrm{Gl}(n, \mathbf{Z})$ such that

$$G = \{g : D(g)B = B\mathcal{D}(g)\} \qquad [4]$$

The maximal crystallographic point group for given lattice Λ is the holohedry of Λ. The group generated by T, G is a space group which classifies the lattice. For finer details in the classification of space groups, we refer to Brown *et al.* (1978). For crystallography in E^3, this classification yields 230 space groups. Crystallography in E^n is described in Schwarzenberger (1980) and in Brown *et al.* (1978) where it is elaborated for E^4.

From a lattice $\Lambda \in E^n$ and from the Euclidean metric, one constructs a cell structure as follows: the Voronoi cell $V(b)$, centered at a lattice point $b \in \Lambda$, known in physics as the Wigner–Seitz cell, is the set of points

$$V(b) = \{x \in E^n : |x - b| \leq |x - b'|,\ b' \in \Lambda\} \qquad [5]$$

Any Voronoi cell has a hierarchy of boundaries X_p of dimension $p, 0 \leq p \leq n$ which we denote as p-boundaries.

The set of Voronoi cells at all lattice points form the Λ-periodic Voronoi complex of $\Lambda \in E^n$. The Voronoi cells and complexes associated with a lattice admit a notion of geometric duality. We denote dual objects by a star, *. They are built from convex hulls of sets of lattice points (Kramer and Schlottmann 1989) as follows. A Voronoi p-boundary X_p is shared by several Voronoi cells $V(b)$ and determines a set of lattice points

$$S(X_p) : \{b \in \Lambda : X_p \in V(b)\} \qquad [6]$$

The boundary dual to X_p is defined as the convex hull $X^*_{(n-p)} := \mathrm{conv}\{b : b \in S(X_p)\}$. $X^*_{(n-p)}$ can be shown to be an $(n-p)$-boundary of a dual Delone cell. A Delone cell D is defined as the convex hull of all lattice points whose Voronoi cells share a single vertex, called a hole of the lattice. Since these vertices fall into classes of orbits under translations, they determine translationally inequivalent classes of Delone cells $D^\alpha, D^\beta, \ldots$.

Fourier analysis applied to a periodic function $f^p(x)$ on E^n reduces to an n-fold Fourier series. The Fourier spectrum is a pure point spectrum and the Fourier coefficients can be referred to the points of a reciprocal lattice Λ° (eqn [7]) in Fourier space $E^{\circ n}$. We denote objects belonging to this Fourier space by the index $^\circ$. The basis matrix B° of the reciprocal lattice $\Lambda^\circ \in E^{\circ n}$ is obtained from B as the inverse transpose,

$$\langle b^{\circ i}, b^j \rangle = \delta^{ij} \leftrightarrow B^\circ = (B^{-1})^{\mathrm{T}} \qquad [7]$$

The values of the Fourier coefficients of $f^p(x)$ reduce to integrals over the fundamental domain of the lattice Λ. From eqns [4] and [7] it follows that the orthogonal representation of a point group G in coordinate and in Fourier space coincides. The Fourier spectrum and its point symmetry in crystals are observed in diffraction experiments.

Quasiperiodic Point Sets and Functions

Quasiperiodic functions are characterized from their Fourier spectrum (Bohr 1925) by

(**qp**$^\circ$) The Fourier point spectrum of a quasiperiodic function forms a $\mathbf{Z}$-module M° of rank $n, n > m$ on Fourier space $E^{\circ m}$.

A $\mathbf{Z}$-module of rank $n, n > m$ on $E^{\circ m}$ is defined as a set

$$M^\circ = \left\{ b^\circ : b^\circ = \sum_j^n m_j b^{\circ i},\ (m_1, \ldots, m_n) \in \mathbf{Z}^n \right\} \qquad [8]$$

with the $\mathbf{Z}$-module basis $(b^{\circ 1}, \ldots, b^{\circ n})$ linearly independent with respect to integral linear combinations. The step from a lattice Λ° to a module M° is nontrivial since the set of all module points becomes dense on $E^{\circ m}$. The Fourier coefficients of a quasiperiodic function are assigned to the discrete set of module points (eqn [8]).

Bohr in his analysis of quasiperiodic functions (Bohr 1925, II, pp. 111–125) shows that a general $\mathbf{Z}$-module M° of rank n can be taken as the projection to a subspace $E^{\circ m}$ of dimension m of a (nonunique) lattice $\Lambda^\circ \in E^{\circ n}, n > m$. It is convenient to consider in Fourier space $E^{\circ n}$ an orthogonal splitting which we denote as

$$E^{\circ n} = E^{\circ m}_{\parallel} + E^{\circ (n-m)}_{\perp}, E^{\circ m}_{\parallel} \perp E^{\circ (n-m)}_{\perp} \qquad [9]$$

A characterization of a quasiperiodic function $f^{qp}(x)$ on coordinate space is obtained as follows. From Λ° one can construct with the help of eqn [7] the lattice $\Lambda := (\Lambda^\circ)^\circ$ reciprocal to Λ° on a coordinate space E^n and associate to it via the Fourier series a quasiperiodic function on a coordinate subspace $E^m_{\parallel}$ of $E^n = E^m_{\parallel} + E^{(n-m)}_{\perp}$, equipped with a $\mathbf{Z}$-module M (eqn [11]). As a result one finds a

characterization of a quasiperiodic function in coordinate space:

(**qp**) A quasiperiodic function $f^{qp}(x_\parallel), x_\parallel \in E_\parallel^m$ can always be interpreted as the restriction to a subspace $E_\parallel^m$ of a Λ-periodic function $f^p(x)$ on E^n,

$$E^n = E_\parallel^m + E_\perp^{(n-m)}, x = x_\parallel + x_\perp$$
$$f^p(x_\parallel + c_\perp) =: f^{qp}(x_\parallel) \quad [10]$$

In the interpretation (**qp**) (eqn [10]), the $\mathbf{Z}$-modules in Fourier space (eqn [8]) and in coordinate space $E_\parallel^m$ become projections of reciprocal lattices,

$$M^\circ = \pi_\parallel(\Lambda^\circ), M = \pi_\parallel(\Lambda) \quad [11]$$

The linear independence of the module basis enforces a splitting (eqn [9]), irrational with respect to the lattice $\Lambda^\circ \in E^{\circ n}$.

As in the classification of crystal lattices, point symmetry plays a crucial role in the classification of $\mathbf{Z}$-modules for quasiperiodic systems like quasicrystals. Noncrystallographic point groups G (with a representation incompatible with any lattice) give rise to quasiperiodic systems as follows:

(**qp**) Given a point group G with orthogonal representations $D_\parallel : G \to O(m, R), D_\perp : G \to O(n-m, R)$ such that $D_\parallel$ is incompatible with any lattice in $E_\parallel^m$, we now require in E^n instead of eqn [10] a lattice Λ with basis B and a representation $\mathcal{D} : G \to \mathrm{Gl}(n, \mathbf{Z})$ such that

$$\begin{bmatrix} D_\parallel(G) & 0 \\ 0 & D_\perp(G) \end{bmatrix} B = B\,\mathcal{D}(G) \quad [12]$$

Equation [12] requires that the matrix B provides an irrational reduction of the representation $\mathcal{D}(G)$ into the two representations $D_\parallel(G), D_\perp(G)$. Periodic functions restricted as in the second line of eqn [10] are quasiperiodic.

For any finite group G, a representation $\mathcal{D}(G)$ allowing for lattice embedding can always be constructed by the technique of induced representations. Its reduction into representations $D_\parallel(G)$, $D_\perp(G)$ contained in this induced representation is obtained by standard techniques. If $D_\parallel(G)$ is noncrystallographic and inequivalent to $D_\perp(G)$, the subspace decomposition (eqn [12]) is unique.

Quasiperiodic functions compatible with tilings and their windows can be constructed from the dual cell structure (eqns [5] and [6]) of the embedding lattice (Kramer and Schlottmann 1989). Examples are given in the sections "Point symmetry in quasiperiodic systems" and "Quasiperiodic tilings and their windows".

Scaling and Quasiperiodicity

Quasiperiodic systems lack periodicity but can have scaling symmetries originating from a non-Euclidean extension of eqn [12].

Example 1: Scaling in the Square Lattice Z^2

We begin with the Fibonacci scaling on the square lattice $\mathbf{Z}^2$ of E^2. The symmetric matrix

$$h = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \in \mathrm{Gl}(2, \mathbf{Z}) \quad [13]$$

has eigenvalues

$$\lambda_1 = -\tau^{-1} = -\tau + 1, \qquad \lambda_2 = \tau := (1+\sqrt{5})/2 \quad [14]$$

Evaluation of the orthogonal eigenvectors allows us to define a lattice basis $B = (b^1, b^2)$ and rewrite the eigenvalue equation similar to eqn [12] as

$$\begin{bmatrix} -\tau^{-1} & 0 \\ 0 & \tau \end{bmatrix} B = B \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$
$$B = \begin{bmatrix} -\sqrt{\frac{-\tau+3}{5}} & \sqrt{\frac{\tau+2}{5}} \\ \sqrt{\frac{\tau+2}{5}} & \sqrt{\frac{-\tau+3}{5}} \end{bmatrix} \quad [15]$$

This relation shows that h with respect to the basis B acts as a non-Euclidean point symmetry of the square lattice and generates an infinite discrete group. Equation [15] provides an orthogonal splitting $E^2 = E_\parallel + E_\perp$. The element h acts on the two subspaces as a discrete linear scaling by $-\tau^{-1}, \tau$, respectively. It maps points of $\mathbf{Z}^2$ in E^2, hence also their projections to $E_\parallel$, into one another.

Figure 1 shows the lattice basis from eqn [15]. We choose as fundamental domain of $\mathbf{Z}^2$ two squares A, B whose boundaries are parallel or perpendicular to $E_\parallel$. A horizontal line $E_\parallel$ intersects these two squares at vertical distances varying with respect to their horizontal boundaries. The quasiperiodic restriction $f^{qp}(x_\parallel) = f^p(x_\parallel + c_\perp)$ of a $\mathbf{Z}^2$-periodic function $f^p(x)$ to a line $x = x_\parallel + c_\perp$ picks up varying functional values on these sections. Clearly, one needs all the values of f^p on its fundamental domain in E^2 to obtain all the values taken by f^{qp}.

Scaling symmetry appears in conjunction with noncrystallographic point symmetry (cf. the following section). Combined with quasiperiodic tilings, it gives rise to a hierarchy of self-similar tilings whose tiles scale with τ.

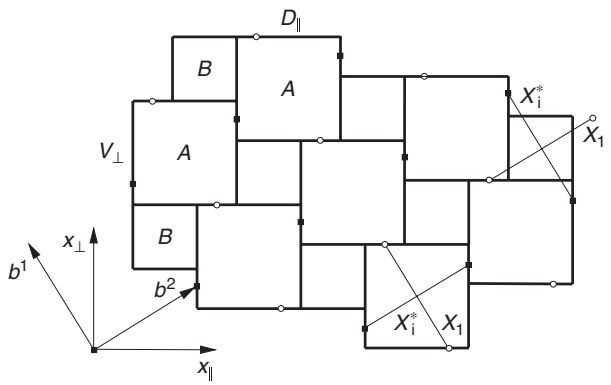


Figure 1 The square lattice with Fibonacci scaling. Lattice points are black squares, holes white circles. The vectors (b^1, b^2) indicate the lattice basis. The directions $x_\parallel, x_\perp$ of scalings by $-\tau^{-1}, \tau$ run horizontally and vertically, respectively. Perpendicular and parallel projections $V_\perp$ of Voronoi and $D_\parallel$ of Delone cells are attached to the lattice and hole points, respectively. Two different pairs of dual 1-boundaries X_1, X_1^* of Voronoi and Delone squares are marked on the right. The product polytopes $X^*_{1,\parallel} \times X_{1,\perp}$ of their projections form two squares A, B and yield a periodic tiling of E^2. A single pair A, B forms a fundamental domain of the lattice. The characteristic functions on A, B are windows for the tiles. A general quasiperiodic function $f^{qp}(x_\parallel)$ is the restriction of a periodic function $f^p(x)$, defined on A, B, to its values on a horizontal line $x = x_\parallel + c_\perp$. If the periodic function $f^p(x)$ on A, B takes only values independent of $x_\perp$, its quasiperiodic restriction $f^{qp}(x_\parallel) := f^p(x_\parallel + c_\perp)$ to this line repeats its values on the long and short tiles $A_\parallel, B_\parallel$, respectively, of the standard Fibonacci tiling. Then $A_\parallel, B_\parallel$ form a fundamental domain for quasiperiodic functions compatible with the tiling.

Point Symmetry in Quasiperiodic Systems

Quasiperiodic systems with noncrystallographic point symmetry provide the structure theory and physics of quasicrystals. We illustrate the general scheme (**qp**) of eqn [12] by examples of 5-fold and icosahedral point symmetry. For generalizations, see Janssen (1986).

Example 2: 5-Fold Point Symmetry from the Root Lattice A_4

The A_4 root lattice basis in E^4 may be derived (Baake *et al.* 1990) from five orthonormal unit vectors $(e^1, e^2, e^3, e^4, e^5)$ in E^5 as

$$\begin{aligned} B &= (b^1, b^2, b^3, b^4) \\ &:= (e^1 - e^2, e^2 - e^3, e^3 - e^4, e^4 - e^5) \end{aligned} \qquad [16]$$

As the generator of the cyclic group C_5 of 5-fold rotations, we take the cyclic permutation (12345) in cycle notation acting on the vectors $(e^1, e^2, e^3, e^4, e^5)$.

A possible choice of the basis for eqn [12] is the irrational matrix

$$\begin{aligned} B &= (b^1, b^2, b^3, b^4) \\ &= \begin{bmatrix} 1 & c & c' & c' & c \\ 0 & s & s' & -s' & -s \\ 1 & c' & c & c & c' \\ 0 & s' & -s & s & -s' \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & -1 \end{bmatrix} \\ c &= \cos\left(\frac{2\pi}{5}\right) = \frac{\tau - 1}{2}, \qquad s = \sin\left(\frac{2\pi}{5}\right) = \frac{\sqrt{\tau + 2}}{2} \\ c' &= \cos\left(\frac{4\pi}{5}\right) = -\frac{\tau}{2}, \qquad s' = \sin\left(\frac{4\pi}{5}\right) = -\frac{\sqrt{3 - \tau}}{2} \end{aligned} \qquad [17]$$

Equation [12] for the representation of the generator (12345) of the cyclic group C_5 becomes

$$\begin{aligned} &\begin{bmatrix} c & -s & 0 & 0 \\ s & c & 0 & 0 \\ 0 & 0 & c' & -s' \\ 0 & 0 & s' & c' \end{bmatrix} (b^1, b^2, b^3, b^4) \\ &\quad = (b^1, b^2, b^3, b^4) \begin{bmatrix} 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \end{aligned} \qquad [18]$$

The left of eqn [18] generates two 2D inequivalent representations of 5-fold planar rotations which are incompatible with any 2D lattice.

The lattice A_4 in addition has a scaling symmetry with a factor τ. The scaling transformation may be expressed in terms of the basis (eqn [16]) and an element $h \in Gl(4, \mathbf{Z})$ as

$$\begin{aligned} &\begin{bmatrix} -\tau & 0 & 0 & 0 \\ 0 & -\tau & 0 & 0 \\ 0 & 0 & \tau^{-1} & 0 \\ 0 & 0 & 0 & \tau^{-1} \end{bmatrix} (b^1, b^2, b^3, b^4) \\ &\quad = (b^1, b^2, b^3, b^4) \begin{bmatrix} 0 & -1 & 0 & 1 \\ 0 & -1 & -1 & 1 \\ 1 & -1 & -1 & 0 \\ 1 & 0 & -1 & 0 \end{bmatrix} \end{aligned} \qquad [19]$$

It is easily verified that the operations of scaling and of 5-fold rotation (eqns [19] and [18]) commute with one another.

Example 3: Icosahedral Point Symmetry from Lattices $\Lambda = \boldsymbol{Z}^6, D_6$

The icosahedral group $G = H_3$ has two inequivalent 3D noncrystallographic representations. H_3 allows for an induced embedding representation $\mathcal{D}: H_3 \rightarrow \mathrm{Gl}(6, \boldsymbol{Z})$, (Kramer and Neri 1984, Kramer *et al.* 1992, Kramer and Papadopolos 1997) into a hypercubic lattice $\Lambda = \boldsymbol{Z}^6$. This representation reduces into two 3D orthogonal inequivalent irreducible noncrystallographic representations $D_\parallel: H_3 \rightarrow O(3,R), D_\perp: H_3 \rightarrow O(3,R)$. The irrational basis matrix of eqn [12] for $\Lambda = \boldsymbol{Z}^6$ becomes (Kramer *et al.* 1992, p. 185, eqn (7))

$$B = (b^1, b^2, b^3, b^4, b^5, b^6) = \sqrt{\frac{1}{2(\tau+2)}} \begin{bmatrix} 0 & 1 & \overline{1} & \overline{\tau} & 0 & \tau \\ 1 & \tau & \tau & 0 & \overline{1} & 0 \\ \tau & 0 & 0 & 1 & \tau & 1 \\ 0 & \tau & \overline{\tau} & 1 & 0 & \overline{1} \\ \tau & \overline{1} & \overline{1} & 0 & \overline{\tau} & 0 \\ \overline{1} & 0 & 0 & \tau & \overline{1} & \tau \end{bmatrix} \qquad [20]$$

with $\overline{\tau} = -\tau, \overline{1} = -1$. The six basis vectors with components in the upper three rows span the so-called primitive icosahedral $\boldsymbol{Z}$-module associated with $D_\parallel$ in $E^3_\parallel$ in the sense of eqn [11]. In this space they point along the directions of six 5-fold axes of the icosahedron.

A second lattice in E^6 which admits icosahedral point symmetry is the root lattice D_6. The basis of this root lattice, often denoted as the P-lattice, is obtained from eqn [20] by a centering matrix given in Kramer *et al.* (1992, p. 185, eqn (8)). The corresponding $\boldsymbol{Z}$-module is inequivalent to the module projected from eqn [20]. The third lattice of icosahedral point symmetry in E^6 is $\Lambda = I := P^\circ$ reciprocal to the root lattice D_6. All three icosahedral modules admit (powers of) τ-scaling.

Quasiperiodic Tilings and Their Windows

Quasiperiodic sets of points arise from the general scheme (**qp**) (eqn [12]) by choosing particular periodic functions in the embedding space E^n, called the "windows," whose intersections with $E_\parallel$ are the quasiperiodic sets of points.

The window for the construction of a discrete quasiperiodic point set based on eqn [12] is given by the characteristic function $\chi(x_\perp)$ on the projection $V_\perp(x_\perp) := \pi_\perp(V(b))$ of the Voronoi cell (eqn [5]), attached to any lattice point $b \in \Lambda$.

Example 4: The Quasiperiodic Fibonacci Point Set

If in the Fibonacci system (Figure 1), one attaches to any point b of the square lattice as a window the characteristic function χ of the perpendicular projection $V_\perp(b)$ of the unit square attached to b, the function $f^{qp}(x_\parallel)$ becomes the standard quasiperiodic Fibonacci sequence of points.

The dual cell geometry of Voronoi and Delone cells and their dual boundaries (eqns [5] and [6]) allows us to construct dual canonical quasiperiodic tilings $(\mathcal{T}, \Lambda), (\mathcal{T}^*, \Lambda)$ (Kramer and Schlottmann 1989). To this end one constructs from local projections of pairs of dual boundaries $X_{m,\parallel}, X^*_{(n-m),\perp}$ or $X^*_{m,\parallel}, X_{(n-m),\perp}$ the direct product polytopes $X_{m,\parallel} \times X^*_{(n-m),\perp}$ or $X^*_\parallel \times X_\perp$ called "klotz polytopes." The characteristic functions on these polytopes form the windows for the tiles $X_{m,\parallel}, X^*_{(n-m),\parallel}$, respectively.

Example 5: The Quasiperiodic Fibonacci Tiling

The Voronoi cells V of the square lattice are squares centered at lattice points, the Delone cells D are squares centered at the vertices of Voronoi squares. The product polytopes $X^*_{1,\perp} \times X_{1,\parallel}$ from projections of dual 1-boundaries X^*_1, X_1 of Delone and Voronoi squares (cf. Figure 1) become the two types of square windows A, B. If a parallel line section $x = x_\parallel + c_\perp$ crosses one of these squares, the tile $A_\parallel$ or $B_\parallel$ is formed. The standard Fibonacci tiling results.

Example 6: Canonical Tilings from the Root Lattices A_4, D_6

The two rhombus tiles of the planar quasiperiodic Penrose pattern (Penrose 1974) $(\mathcal{T}, A_4)$ are the projections of 2-boundaries of the Voronoi complex of the root lattice $A_4 \in E^4$ (Baake *et al.* 1990). The triangle tiles of the dual tiling $(\mathcal{T}^*, A_4)$ are shown in Figure 2.

They are projections of 2-boundaries from the Delone complex of the same lattice. A full analysis of dual Voronoi and Delone boundaries of the root lattice D_6 is given in Kramer *et al.* (1992). It leads to icosahedral tilings $(\mathcal{T}, D_6)$ and $(\mathcal{T}^*, D_6)$ of E^3, (Kramer *et al.* 1992, Kramer and Papadopolos 1997, Kramer and Schlottmann 1989) and to models of icosahedral quasicrystals.

Fundamental Domains for Quasiperiodic Tilings

Canonical tilings allow us to construct quasiperiodic functions equipped with a quasiperiodic counterpart of fundamental domains or cells in crystals: assume that the tiles of a tiling $(\mathcal{T}, \Lambda)$ all are translates in E^m of a finite minimal set of prototiles $(X^1, \ldots, X^r)$. Consider the class of quasiperiodic functions which

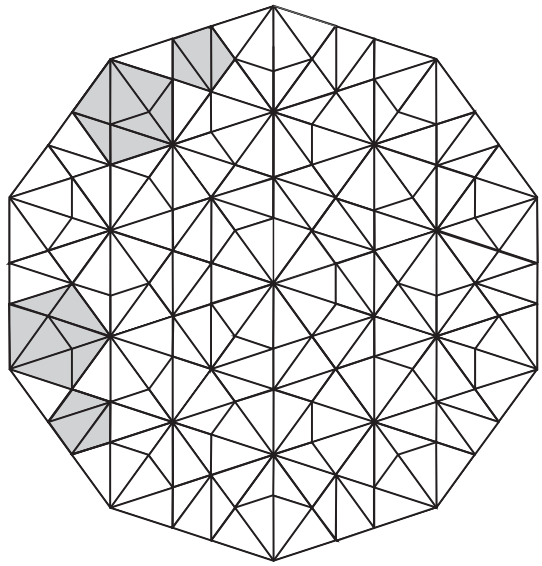


Figure 2 A patch of the planar quasiperiodic triangle tiling $(\mathcal{T}^*, A_4)$ obtained from the root lattice $A_4 \in E^4$. The tiles are two triangles, projections of 2-boundaries from the Delone cells of A_4. The vertices are projections of lattice points. The 20 shaded triangles form a set of prototiles such that any other tile is a translate of one of them. The shaded set forms a fundamental domain for the tiling.

take identical values on any translate of a prototile. These values are prescribed on the finite set of prototiles in E^m which define a fundamental domain for this class of quasiperiodic functions. Only this class of quasiperiodic functions is compatible with the tiling. It can be characterized in the scheme (**qp**) (eqn [12]) by Λ-periodic functions on E^n whose values on the tile windows of the previous section are independent of the perpendicular coordinate. A fundamental domain for the triangle tiling $(\mathcal{T}^*, A_4)$ is given by the shaded parts in Figure 2. The fundamental domain property appears in relation with the theory of covering of quasiperiodic sets (see Kramer and Papadopolos (2000)).

Example 7: Fundamental Domain for the Fibonacci Tiling

Attach to the squares A, B in Figure 1 a periodic function $f^p(x)$ with functional values independent of the perpendicular coordinate $x_\perp$ within the two squares. Consider the functional values $f^{qp}(x_\parallel) = f^p(x_\parallel + c_\perp)$ picked up on a parallel line. Clearly, these values become independent of the perpendicular coordinate of any intersection with a square A, B. The general prescription of values on a fundamental domain of $\Lambda \in E^2$ needed for a quasiperiodic function reduces to a prescription of its functional values in $E_\parallel$ on the fundamental domain formed by the two prototiles $A_\parallel, B_\parallel$.

Conclusion

For quasiperiodic systems, the general construction was introduced in the section "Quasiperiodic point sets and functions", and illustrations were given in four subsequent sections. Further reading resources are provided by the references given at the end. Here, we mention some of the many possible generalizations.

Bohr (1925) considers quasiperiodic as special cases of almost periodic systems. The module of an almost periodic function has a countable basis.

Moody (1997) discusses the notion of Meyer sets. These describe discrete sets on locally compact abelian groups and as particular cases encompass quasiperiodic systems.

Lagarias (2000) studies aperiodic sets characterized by the following properties, shared with periodic and quasiperiodic sets:

(**ap1**): inequivalent patches of points are volume bounded,
(**ap2**): pure point Fourier spectrum,
(**ap3**): linear repetitivity of patches, and
(**ap4**): self-similarity.

See also: Compact Groups and Their Representations; Finite Group Symmetry Breaking; Lie Groups: General Theory; Localization for Quasiperiodic Potentials; Symmetries and Conservation Laws; Symmetry and Symmetry Breaking in Dynamical Systems.

Further Reading

Baake M, Kramer P, Schlottmann M, and Zeidler D (1990) Planar patterns with fivefold symmetry as sections of periodic structures in 4-space. *International Journal of Modern Physics B* 4: 2217–2268.

Bohr H (1925) Zur Theorie der fastperiodischen Funktionen. I. *Acta Mathematicae* 45: 29–127.

Bohr H (1925) Zur Theorie der fastperiodischen Funktionen. II. *Acta Mathematicae* 46: 101–214.

Brown R, Bülow R, Neubüser J, Wondratschek H, and Zassenhaus H (1978) *Crystallographic Groups of Four-Dimensional Space*. New York: Wiley.

de Bruijn NG (1981) Algebraic theory of Penrose's non-periodic tilings of the plane. I. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen* 84: 39–52.

de Bruijn NG (1981) Algebraic theory of Penrose's non-periodic tilings of the plane. II. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen* 84: 53–66.

Janssen T (1986) Crystallography of quasicrystals. *Acta Crystallographica A* 42: 261–71.

Kramer P and Neri R (1984) On periodic and non-periodic space fillings of E^n obtained by projection. *Acta Crystallographica A* 40: 580–587.

Kramer P and Papadopolos Z (1997) Symmetry concepts for quasicrystals and non-commutative crystallography. In: Moody RV (ed.) *The Mathematics of Long-Range Aperiodic Order*, pp. 307–330. Dordrecht: Kluwer.

Kramer P and Papadopolos Z (eds.) (2000) *Coverings of Discrete Quasiperiodic Sets, Theory and Applications to Quasicrystals*. New York: Springer.

Kramer P, Papadopolos Z, and Zeidler D (1992) The root lattice D_6 and icosahedral quasicrystals. In: Frank A, Seligman TH, and Wolf KB (eds.) *Group Theory in Physics*, American Institute of Physics Conference Proceedings, vol. 266, pp. 179–200. New York.

Kramer P and Schlottmann M (1989) Dualization of Voronoi domains and klotz construction: a general method for the generation of proper space filling. *Journal of Physics A: Mathematical and General* 22: L1097–L1102.

Lagarias J (2000) The impact of aperiodic order on mathematics. *Material Science Engineering A* 294–296: 186–191.

Moody RV (1997) Meyer sets and their duals. In: Moody RV (ed.) *The Mathematics of Long-Range Aperiodic Order*, pp. 403–441. Dordrecht: Kluwer.

Penrose R (1974) The role of aesthetics in pure and applied mathematical research. *Bulletin of the Institute of Mathematics and its Applications* 10: 266–71.

Schwarzenberger RLE (1980) *N-Dimensional Crystallography*. San Francisco: Pitman.

Shechtman D, Blech I, Gratias D, and Cahn JW (1984) Metallic phase with long-range orientational order and no translational symmetry. *Physical Review Letters* 53: 1951–1953.

Quillen Determinant

S Scott, King's College London, London, UK

Determinants in Finite Dimensions

The determinant of a linear transformation $A : V \longrightarrow W$ acting between finite-dimensional complex vector spaces is an element $\det A$ of a complex line L_A. The abstract element $\det A$ is called the *Quillen determinant* of A, and the complex line L_A is called the determinant line of A. A choice of (linear) isomorphism

$$\phi : L_A \longrightarrow \mathbb{C} \qquad [1]$$

associates to $\det A$ the complex number

$$\det{}_\phi A := \phi(\det A) \in \mathbb{C} \qquad [2]$$

which can equivalently be written as the ratio

$$\det{}_\phi A = \frac{\det A}{\phi^{-1}(1)} \qquad [3]$$

taken in the one-dimensional complex vector space L_A relative to the canonical generator $\phi^{-1}(1)$. It is not necessarily the case that $\det A$ determines a generator for L_A; specifically, if $\dim V = m$ and $\dim W = n$, then $\det A = 0$ if $m \neq n$ (by "fiat"), while if $m = n$, then $\det A = 0$ precisely when A is not invertible. For the moment, set $m = n$.

For $k \in \{0, 1, \ldots, n\}$ the kth exterior power operator is defined by

$$\wedge^k A : \wedge^k V \longrightarrow \wedge^k W$$

$$\wedge^k A(v_1 \wedge v_2 \wedge \cdots \wedge v_k) := Av_1 \wedge Av_2 \wedge \cdots \wedge Av_k \qquad [4]$$

where $v_1, \ldots, v_k \in V$ and $\wedge^0 V := \mathbb{C}$ and $\wedge^0 A := 1$. When $k = n$, $\operatorname{Det} V := \wedge^n V$ and $\operatorname{Det} W := \wedge^n W$ are complex lines and the determinant line of A is

$$L_A := \operatorname{Det} V^* \otimes \operatorname{Det} W \qquad [5]$$

while for any basis $\{e_1, \ldots, e_n\}$ for V, with dual basis $\{e_1^*, \ldots, e_n^*\}$ for V^*,

$$\det A := e_1^* \wedge \cdots \wedge e_n^* \otimes (\wedge^n A)(e_1 \wedge \cdots \wedge e_n) \in L_A \qquad [6]$$

There is a canonical isomorphism for $A \in \operatorname{Hom}(V, W)$, $B \in \operatorname{Hom}(U, V)$

$$L_{AB} \cong L_A \otimes L_B \qquad [7]$$

coming from the isomorphism

$$\operatorname{Det} V^* \otimes \operatorname{Det} V \longrightarrow \mathbb{C} \qquad [8]$$

defined by the canonical pairing $\operatorname{Det} V^* \times \operatorname{Det} V \to \mathbb{C}$, and this preserves the determinant elements

$$\det(AB) \longleftrightarrow \det A \otimes \det B \qquad [9]$$

The Classical Determinant

When $V = W$ these constructions take on a more familiar form. Then ϕ can be chosen to be the canonical isomorphism [8] and evaluation on $\det A \in L_A$ outputs the classical determinant

$$\det{}_{\mathbb{C}} A = \sum_\sigma (-1)^\sigma a_{1,\sigma(1)} \cdots a_{n,\sigma(n)} \qquad [10]$$

where the sum is over permutations of $\{1, \ldots, n\}$ and $(a_{i,j})$ is the matrix of A with respect to any basis of V – changing the basis may change the summands on the right-hand side of [10], but not their sum. It is fundamental that when $V = W$ the classical determinant is an intrinsic invariant of the operator A, independent of the choice of basis for V; when $V \neq W$ that is no longer so since there is then no *canonical* bilinear pairing $\operatorname{Det} V^* \times \operatorname{Det} W \longrightarrow \mathbb{C}$; the choice of a non-degenerate pairing is equivalent to a choice of ϕ in [1].

The identification of [10] from [6] and [8] amounts to the identity in $\operatorname{Det} V$

$$(\wedge^n A)(e_1 \wedge \cdots \wedge e_n) = \det{}_{\mathbb{C}} A \,.\, e_1 \wedge \cdots \wedge e_n \qquad [11]$$

Since $\wedge^n(AB) = \wedge^n A \circ \wedge^n B$, [11] in turn implies the characterizing multiplicativity property of the classical determinant

$$\det_{\mathbb{C}}(AB) = \det_{\mathbb{C}} A \,.\, \det_{\mathbb{C}} B \qquad [12]$$

for $A, B \in \mathrm{End}(V)$, specializing the general fact in [7]. Similarly, the group $\mathrm{Gl}(V, \mathbb{C})$ of invertible elements of $\mathrm{End}(V)$ is identified with those A with $\det_{\mathbb{C}} A \neq 0$.

The classical determinant can also be thought of in the following ways. First, the direct sum of the operators defined in [4] yields the total exterior power operator $\wedge A : \wedge V \longrightarrow \wedge V$ on the exterior algebra $\wedge V = \oplus_{k=0}^{n} \wedge^k V$ and this has trace

$$\mathrm{tr}\,(\wedge A) = \det_{\mathbb{C}}(I + A) \qquad [13]$$

where I is the identity. Alternatively, one can do something a little more sophisticated and use the holomorphic functional calculus to define the logarithm $\log_\theta B$ of $B \in \mathrm{End}(V)$ by

$$\log_\theta B = \frac{\mathrm{i}}{2\pi} \int_{\Gamma_\theta} \log_\theta \lambda \,(B - \lambda I)^{-1}\, \mathrm{d}\lambda \qquad [14]$$

Here $\log_\theta \lambda$ is the branch of the complex logarithm defined by $\theta - 2\pi < \arg(\lambda) \leq \theta$ and Γ_θ is a positively oriented contour enclosing $\mathrm{spec}(B)$ but not any point of the spectral cut $R_\theta = \{re^{\mathrm{i}\theta} \mid r \geq 0\}$. Then, if B is invertible,

$$\mathrm{tr}\,(\log_\theta B) = \log_\theta \det_{\mathbb{C}} B \qquad [15]$$

The Fredholm Determinant

The advantage of the constructions [13] and [15] is that they extend to a restricted class of bounded linear operators on infinite-dimensional Hilbert spaces. This is consequent on the fact that both of the formulas [13] and [15] are computed as operator traces.

(Recall that a trace on a Banach algebra $\mathcal{B}$ is a linear functional $\tau : \mathcal{B} \to \mathbb{C}$ which has the property $\tau([a, b]) = 0$ for all a, b in $\mathcal{B}$, where $[a, b] := ab - ba$ is defined by the product structure on $\mathcal{B}$. Since one can define the logarithm $\log_\theta b$ of an element b of $\mathcal{B}$ with spectral cut R_θ by the formula [14], one in this case obtains a determinant $\det_{\tau,\theta}(b)$ on such elements by setting

$$\log_\theta \det_{\tau,\theta}(b) = \tau(\log_\theta b) \qquad [16]$$

If $a, b, ab \in \mathcal{B}$ have common spectral cuts θ, the trace property of τ translates into the multiplicativity property $\det_{\tau,\theta}(ab) = \det_{\tau,\theta}(a)\det_{\tau,\theta}(b)$ via a version of the Campbell–Hausdorff formula.)

The operator trace arises as follows. Let H be a complex separable Hilbert space with inner product $<,>$, let $\mathcal{C}(H)$ be the algebra of compact operators on H, and let

$$L_1 = \left\{ A \in \mathcal{C}(H) \mid \|A\|_1^2 := \sum_{i=1}^{\infty} \mu_i(A^*A) < \infty \right\} \qquad [17]$$

be the ideal of trace-class operators, where the sum is over the real discrete eigenvalues $\mu_i(A^*A) \searrow +0$ of the compact self-adjoint operator A^*A. For any orthonormal basis $\{\eta_j\}$ of H the map

$$\mathrm{tr} : L_1 \longrightarrow \mathbb{C}, \qquad A \mapsto \mathrm{tr}\,(A) := \sum_j <\eta_j, A\eta_j>$$

is a trace functional on $L_1(H)$, independent of the choice of basis. Lidskii's theorem states that

$$\mathrm{tr}\,(A) = \sum_{\lambda \in \mathrm{spec}(A)} \lambda \qquad [18]$$

with the sum over the eigenvalues of A counted up to algebraic multiplicity; for general trace-class operators this equality is highly nontrivial.

If A is trace class, then for each non-negative integer k so is each of the exterior power operators $\wedge^k A : \wedge^k H \longrightarrow \wedge^k H$, defined as in [4]. Following [13], a determinant can therefore be defined on the semigroup $I + L_1 := \{I + A \mid A \in L_1\}$ of determinant-class operators by the absolutely convergent sum

$$\det_{\mathrm{F}}(I + A) := \mathrm{tr}\,(\wedge A) = 1 + \sum_{k=1}^{\infty} \mathrm{tr}\,(\wedge^k A) \qquad [19]$$

On the other hand, since tr is tracial and $\log_\pi(I + A)$ defined by [14] is trace class, then according to [16], there is a determinant given on invertible determinant-class operators by

$$\log_\pi \det_{\mathrm{F}}(I + A) = \mathrm{tr}\,(\log_\pi(I + A)) \qquad [20]$$

which, as the left-hand side already suggests, coincides with the Fredholm determinant.

The Fredholm determinant retains the characterizing properties of the classical determinant in finite dimensions, that $\det_{\mathrm{F}} : I + L_1 \longrightarrow \mathbb{C}$ is multiplicative,

$$\det_{\mathrm{F}}((I + A)(I + B)) = \det_{\mathrm{F}}(I + A)\det_{\mathrm{F}}(I + B), \quad A, B \in L_1 \qquad [21]$$

and $\det_{\mathrm{F}}(I + A) \neq 0$ if and only if $I + A$ is invertible. It is, moreover, essentially unique; any other multiplicative functional on $I + L_1$ is equal to some power of the Fredholm determinant, or, equivalently, any trace on L_1 is a constant multiple of the operator trace. The trace property, the operator trace, and the multiplicativity of the Fredholm determinant do not, however, persist to any functional extension of the operator trace (resp. Fredholm determinant) on the

space of pseudodifferential operators of any real order acting on function spaces (fields over space-time). In quantum physics, this is a primary cause of anomalies. More precisely, determinants of differential operators arise in quantum field theories (QFTs) and string theory through the formal evaluation of their defining Feynman path integrals and the calculation of certain stable quantum numbers, which are in some sense "topological."

From the latter perspective, it is instructive to be aware also of the following, third, construction of the Fredholm determinant, which equates the existence of a nontrivial determinant to the existence of nontrivial topology of the general linear group. First, in a surprising contrast to $\mathrm{Gl}(n,\mathbb{C})$, the general linear group $\mathrm{Gl}(H)$ of an infinite-dimensional Hilbert space H with the norm topology is contractible, and hence topologically trivial. By transgression properties in cohomology, this implies any vector bundle with structure group $\mathrm{Gl}(H)$ is isomorphic to the trivial bundle. In order to recapture some topology (and hence, in applications, some physics), it is necessary to reduce to certain infinite-dimensional subgroups of $\mathrm{Gl}(H)$. The most obvious one is the group $\mathrm{Gl}(\infty)$ of of invertible operators differing from the identity by an operator of finite rank. As the inductive limit of the $\mathrm{Gl}(n,\mathbb{C})$, the cohomology and homotopy groups of $\mathrm{Gl}(\infty)$ are a stable version of those of $\mathrm{Gl}(n,\mathbb{C})$. Precisely, $\mathrm{Gl}(\infty)$ is torsion free and its cohomology ring is an exterior algebra with odd degree generators, while Bott (1959) periodicity identifies $\pi_k(\mathrm{Gl}(\infty))$ to be isomorphic to $\mathbb{Z}$ if k is odd and trivial if k is even. Topologically, it is preferable to consider the closure of $\mathrm{Gl}(\infty)$ in $\mathrm{Gl}(H)$, which yields the group $\mathrm{Gl}_{\mathrm{cpt}}(H)$ of operators differing from the identity by a compact operator, but this is now a little "too large" for analysis and differential geometry. Given our earlier comments, there is an intermediate natural choice of the Banach Lie group $\mathrm{Gl}_1(H)$ of operators differing from the identity by a trace-class operator (in fact, there is a tower of such Schatten class groups). Moreover, the inclusions $\mathrm{Gl}(\infty)\subset\mathrm{Gl}_1(H)\subset\mathrm{Gl}_{\mathrm{cpt}}(H)$ are homotopy equivalences, and so the cohomology of $\mathrm{Gl}_1(H)$ is just the exterior algebra mentioned above

$$H^*(\mathrm{Gl}_1(H))=\wedge(\omega_1,\omega_3,\omega_5,\ldots),\quad \deg\omega_j=2j-1 \qquad [22]$$

The advantage of considering $\mathrm{Gl}_1(H)$ is that precise analytical representatives for the classes ω_j can be written down:

$$\omega_j=\left(\frac{\mathrm{i}}{2\pi}\right)^j\frac{(j-1)!}{(2j-1)!}\,\Theta^{2j-1}$$

where

$$\Theta=\mathrm{tr}\,(Z^{-1}\mathrm{d}Z) \qquad [23]$$

is the 1-form on $\mathrm{Gl}_1(H)$.

This equation makes sense because the derivative $\mathrm{d}Z$ is trace class, and hence so is $Z^{-1}\mathrm{d}Z$. Now, locally $\Theta=\mathrm{d}\log\det_{\mathrm{F}}(Z)$, so that the 1-form ω_1 pulled back by a path $\sigma:S^1\to\mathrm{Gl}_1(H)$ is precisely the winding number of the curve traced out in $\mathbb{C}^*$ by the function $\det_{\mathrm{F}}(\sigma)$. In fact, this is just a special case of the Bott periodicity theorem, which tells us that the stable homotopy group $\pi_{2j-1}(\mathrm{Gl}_1(H))$ is isomorphic to $\mathbb{Z}$ and an isomorphism is defined by assigning to a map $f:S^{2j-1}\to\mathrm{Gl}_1(H)$ the integer $\int_{S^{2j-1}}f^*\omega_j\in\mathbb{Z}$ (it is not obvious *a priori* that it is an integer).

Notice that it was not necessary to have mentioned the Fredholm determinant of Z at this point. Indeed, the third definition of the Fredholm determinant is to see it as the integral of the 1-form Θ, define

$$\log_\pi\det_{\mathrm{F}}(\mathrm{I}+\mathrm{A}):=\int_\gamma\Theta \qquad [24]$$

where $\gamma:[0,1]\to\mathrm{Gl}_1(H)$ is any path with $\gamma(0)=I$ and $\gamma(1)=I+A$; this uses the connectedness of $\mathrm{Gl}_1(H)$ and independence of the choice of γ, as guaranteed by Bott periodicity.

Interestingly, this is closely tied in with the Atiyah–Singer index theorem for elliptic pseudodifferential operators (which in full generality uses the Bott periodicity theorem). Here, there is the following simple but quintessential version of that theorem which links it to the winding number of the determinant of the symbol of a differential operator

$$D=\sum_{\alpha|\le m}a_\alpha(x)D_x^\alpha \qquad [25]$$

on Euclidean space $\mathbb{R}^n$ with $\alpha=(\alpha_1,\ldots,\alpha_n)$ a multi-index of non-negative integers, $|\alpha|=\alpha_1+\cdots+\alpha_n$, and $D_x=\mathrm{i}\partial/\partial x_i$. Here D acts on $\mathrm{C}^\infty(\mathbb{R}^n,V)$ with V a finite-dimensional complex vector space and the coefficients of D are matrices varying smoothly with x which are required to decay suitably fast, $\left|D_x^\beta a_\alpha(x)\right|=\mathrm{O}(|x|^{-|\beta|})$ as $|x|\to\infty$. If the symbol σ_D of D, defined by

$$\sigma_D(x,\xi)=\sum_{\alpha|\le m}a_\alpha(x)\xi^\alpha \qquad [26]$$

with $\xi=(\xi_1,\ldots,\xi_n)\in\mathbb{R}^n$, satisfies the ellipticity condition of being invertible on the $2n-1$ sphere S^{2n-1} in (x,ξ) space, then D is a Fredholm operator. The index theorem then states

$$\mathrm{index}\,(D)=\int_{S^{2j-1}}\sigma_D^*(\omega_n)$$

the higher-dimensional analog of the winding number of the determinant.

Fredholm Operators and Determinant Line Bundles

The operators whose determinants are considered in this article are all Fredholm operators. Recall that a linear operator $A: H_1 \to H_2$ between Hilbert spaces is Fredholm if it is invertible modulo compact operators; that is, there is a "parametrix" $Q: H_2 \to H_1$ such that $QA - I$ and $AQ - I$ are compact operators on H_1 and H_2, respectively. Equivalently, the range $A(H_1)$ of A is closed in H_2, and the kernel $\mathrm{Ker}(A) = \{\eta \in H_1 \mid A\eta = 0\}$ and cokernel $\mathrm{Coker}(A) = H_2/A(H_1)$ of A are finite dimensional. (This is equally true for Banach and Frechet spaces, we restrict our attention to Hilbert spaces for brevity.) The space Fred of all such Fredholm operators with the norm topology has the homotopy type of the classifying space $\mathbb{Z} \times B\mathrm{Gl}(\infty)$. The first factor parametrizes the connected components of Fred, two Fredholm operators are in the same component if and only if they have the same index

$$\mathrm{index}\,(A) = \dim \mathrm{Ker}(A) - \dim \mathrm{Coker}(A)$$

Mostly we restrict our attention to the connected component Fred_0 of operators of index zero. The cohomology of $\mathsf{Fred}_0 \sim B\mathrm{Gl}(\infty)$ is a polynomial ring

$$H^*(\mathsf{Fred}_0) = \mathbb{R}[\mathrm{ch}_1, \mathrm{ch}_2, \mathrm{ch}_3, \ldots]$$

whose generators may be formally realized as the even degree components of the Chern character of an infinite-dimensional bundle over Fred_0. In fact, the generators ω_{2j-1} of $H^*(\mathrm{Gl}_1(H))$ are related to the ch_j through transgression, see Chern and Simons (1974). We shall be interested here in the first generator ch_1, a transgression of the Fredholm determinant "winding number 1-form" ω_1, which coincides with the real Chern class of a canonical complex line bundle $\mathrm{DET}_0 \to \mathsf{Fred}_0$. The fiber of DET_0 at $A \in \mathsf{Fred}_0$ is the determinant line $\mathrm{Det}(A)$ of the Fredholm operator A, which is defined as follows (Segal 2004).

Just as for finite-rank operators (see the subsection "Determinants in finite dimensions"), the determinant of a Fredholm operator $A: H^1 \to H^2$ exists abstractly not as a number but as an element $\det A$ of a complex line $\mathrm{Det}(A)$. For simplicity, we suppose that $\mathrm{index}\,(A) = 0$. Elements of the determinant line $\mathrm{Det}(A)$ are equivalence classes $[E, \lambda]$ of pairs (E, λ), where $E: H^1 \to H^2$ such that $A - E$ is trace class and relative to the equivalence relation $(Eq, \lambda) \sim (E, \det_{\mathrm{F}}(q)\lambda)$ for $q: H^1 \to H^1$ of determinant class and where $\det_{\mathrm{F}}(q)$ is the Fredholm determinant of q. Complex multiplication on $\mathrm{Det}(A)$ is defined by $\mu[A, \lambda] = [A, \mu\lambda]$. The abstract, or Quillen, determinant of A is the preferred element $\det A := [A, 1]$ in $\mathrm{Det}(A)$.

Here are some essential properties of the determinant line. First, $\det A$ is nonzero if and only if A is invertible. Next, quotients of abstract determinants in $\mathrm{Det}(A)$ are given by Fredholm determinants; for if $A_1: H^1 \to H^2$, $A_2: H^1 \to H^2$ are Fredholm operators such that $A_i - A$ are trace class, then if A_2 is invertible we see that $A_2^{-1}A_1$ is determinant class and hence from the definition that

$$\frac{\det(A_1)}{\det(A_2)} = \det_{\mathrm{F}}(A_2^{-1}A_1) \qquad [27]$$

where the quotient on the left-hand side is taken in $\mathrm{Det}(A)$. The principal functorial property of the determinant line is that given a commutative diagram with exact rows and Fredholm columns

$$\begin{array}{ccccccccc} 0 & \longrightarrow & H_1 & \longrightarrow & H_1' & \longrightarrow & H_1'' & \longrightarrow & 0 \\ & & \downarrow A & & \downarrow A' & & \downarrow A'' & & \\ 0 & \longrightarrow & H_2 & \longrightarrow & H_2' & \longrightarrow & H_2'' & \longrightarrow & 0 \end{array} \qquad [28]$$

then there is canonical isomorphism of complex lines

$$\mathrm{Det}(A') \cong \mathrm{Det}(A) \otimes \mathrm{Det}(A'') \qquad [29]$$

preserving the Quillen determinants $\det(A') \leftrightarrow \det(A) \otimes \det(A'')$. A consequence of this property is that given Fredholm operators $A: H_2 \to H_3$ and $B: H_1 \to H_2$, then

$$\mathrm{Det}(AB) \cong \mathrm{Det}(A) \otimes \mathrm{Det}(B)$$

with $\det(AB) \leftrightarrow \det(A) \otimes \det(B)$, generalizing the elementary property [9].

The principal context of interest for studying determinant lines is the case where one has a family $\mathcal{A} = \{A_x \mid x \in B\}$ of Fredholm operators parametrized by a manifold B, satisfying suitable continuity properties, and one aims to make sense of the determinant as a function $\mathcal{A} \to \mathbb{C}$. It is then of no difficulty to show that the corresponding family of determinant lines $\mathrm{DET}(\mathcal{A}) = \cup \mathrm{Det}(A_x)$ defines a complex line bundle over B endowed with a canonical section $\det: B \longrightarrow \mathrm{DET}(\mathcal{A})$

assigning to $x \in B$ the Quillen determinant $\det(A_x) \in \mathrm{Det}(A_x)$ (Quillen 1985, Segal 2004). To identify the Quillen determinant section with a function on $\mathcal{A}$, we need to identify a trivialization of the line bundle $\mathrm{DET}(\mathcal{A})$, giving a global basis for the fibers. This is the same thing as giving a non(or never)vanishing section $\psi : B \to \mathrm{DET}(\mathcal{A})$, with respect to which we have the regularized determinant function (cf. [3]):

$$x \mapsto \det{}_{\psi}(A_x) := \frac{\det(A_x)}{\psi(x)} \qquad [30]$$

If $\mathcal{A}$ is trivializable, so a nonzero section exists, there will be many such sections and some extra data is needed to fix a natural choice of ψ.

Each of the properties mentioned above for determinant lines carries forward to determinant line bundles in a natural way. In particular, one easily deduces from [28], or from the exact sequence

$$0 \longrightarrow \mathrm{Ker}A_x \longrightarrow H_{1,x} \xrightarrow{A_x} H_{2,x} \longrightarrow \mathrm{Coker}A_x \longrightarrow 0$$

that if the kernels $\mathrm{Ker}A_x$ have constant dimension as x varies, then there is a canonical isomorphism

$$\mathrm{Det}(\mathcal{A}) \cong \wedge^{\max}\mathrm{Ker}(\mathcal{A})^* \otimes \wedge^{\max}\mathrm{Coker}(\mathcal{A}) \qquad [31]$$

where $\mathrm{Ker}(\mathcal{A})$ is the finite-rank complex vector bundle over B with fiber $\mathrm{Ker}A_x$, and $\mathrm{Coker}(\mathcal{A})$ similarly. The interesting feature here is that it shows the determinant bundle to be the top exterior power of the index bundle $\mathrm{Ind}(\mathcal{A}) = [\mathrm{Ker}(\mathcal{A})] - [\mathrm{Coker}(\mathcal{A})] \in K(B)$ in the even K-theory of B, and in this sense determinant theory may be seen as a particular aspect of index theory – understood is the very broadest sense; in fact, the computation of determinants is usually a considerably more complex and difficult task than computing an index.

Determinant Bundles for Differential Operators over Manifolds

The Quillen determinant has been of particular interest in the case of families of Dirac operators. Such a family is associated to a C^∞ fibration $\pi : M \longrightarrow B$ of closed boundaryless finite-dimensional Riemannian manifolds of even dimension. If there is a graded Hermitian vector bundle $\mathcal{E} = \mathcal{E}^+ \oplus \mathcal{E}^- \to M$ of Clifford modules, then from the Riemannian structure one can construct a Levi-Civita connection on the vertical tangent bundle $T(M/B)$ which can be lifted to a Clifford connection on $\mathcal{E}$; for example, the spinor connection if we have a family of spin manifolds. This data yields a smooth family of first-order elliptic differential operators $\boldsymbol{D} = \{D_x : C^\infty(M_x; \mathcal{E}_x^+) \longrightarrow C^\infty(M_x; \mathcal{E}_x^-) \mid x \in B\}$ of chiral Dirac-type, with D_x a Dirac-type operator acting over the manifold $M_x = \pi^{-1}(x)$ parametrized by the fibration, along with a determinant line bundle $\mathrm{DET}(\boldsymbol{D}) \to B$ endowed with a canonical section $x \mapsto \det(D_x)$. There are various contexts in mathematics and physics in which one would like to assign to the determinant section a naturally associated smooth function (a regularized determinant) $\det{}_{\mathrm{reg}} : B \longrightarrow \mathbb{C}$, which can, for example, then be integrated. As discussed in the previous section, this depends on identifying a trivializing (nonzero) section of $\mathrm{DET}(\boldsymbol{D})$. For such a section to exist, the first Chern class $c_1(\mathrm{DET}(\boldsymbol{D}) \in H^2(B)$ must vanish, and this in turn can be computed as a term in the Atiyah–Singer (1984) index theorem for families. Indeed, this is clear from the formal identification [31] which here takes on a precise meaning.

The following simple example, which is the basic topological anomaly computation in string theory, may help to explain the type of computation. Let M_x be a copy of Σ a compact Riemann surface, so that M is a family of surfaces parametrized by B. Let $T = \cup T_x$ be the vertical complex tangent line bundle on M, where T_x is the complex tangent line bundle to M_x. Each fiber has an associated $\overline{\partial}$-operator $\overline{\partial}_x$ which we couple to the Hermitian bundle $\mathcal{E}_x := T_x^{\otimes m}$ for m a non-negative integer. In this way, we get a family $\boldsymbol{D}_\Sigma$ of $\overline{\partial}$-operators coupled to $\mathcal{E} = T^{\otimes m}$ whose index bundle is the element $\mathrm{Ind}(\boldsymbol{D}_\Sigma) = f_!(T^{\otimes m}) \in K(B)$. The Atiyah–Singer index theorem for families in this situation coincides with the Grothendieck–Riemann–Roch theorem and this says that

$$\mathrm{ch}(f_!(T^{\otimes m})) = f_*(\mathrm{ch}(T^{\otimes m})\mathrm{Todd}(T))$$

where ch is the Chern character class and $\mathrm{Todd}(T)$ is the Todd class defined for a vector bundle F whose first few terms are

$$\mathrm{Todd}(F) = 1 + \tfrac{1}{2}c_1(F) + \tfrac{1}{12}c_1(F)^2 + \cdots$$

and where $f_* : H^i(M) \to H^{i-1}(B)$ is integration over the fibers. That is, with $\xi = c_1(T)$,

$$\begin{aligned}\mathrm{ch}(f_!(T^{\otimes m})) &= f_*\big((1 + m\xi + \tfrac{1}{2}m^2\xi^2 + \cdots) \\ &\quad \times (1 + \tfrac{1}{2}\xi + \tfrac{1}{12}\xi^2 + \cdots)\big) \\ &= f_*\big(1 + (m + \tfrac{1}{2})\xi + \tfrac{1}{12}(m^2 + m + \tfrac{1}{6}) \\ &\quad \times \xi^2 + \cdots\big)\end{aligned}$$

So we have

$$c_1(f_!(T^{\otimes m})) = \tfrac{1}{12}(6m^2 + 6m + 1)f_*(\xi^2) \in H^2(B) \qquad [32]$$

But for any element of K-theory, $c_1(E) = c_1(\mathrm{DET}(E))$, and so the left-hand side of [32] is the first Chern class of the determinant line bundle $\mathrm{DET}(\boldsymbol{D}_\Sigma)$. If we take, in particular, $B = \mathrm{Conf}(\Sigma)$, the space of conformal classes of metrics on Σ (or compact subsets of this space), and couple the family $\boldsymbol{D}_\Sigma$ to a background trivial real bundle of rank $d/2$, or its negative in K-theory, then taking $m = 1$ [32] is easily seen to be modified to

$$c_1(\boldsymbol{D}_{\Sigma,-d/2}) = \frac{(d-26)}{24} f_*(\xi^2)$$

It follows for this topological anomaly to vanish one must have background spacetime of dimension $d = 26$. The idea here is that $\mathrm{Conf}(\Sigma)$ is a configuration space for bosonic strings in $\mathbb{R}^d$ with the requirement that the determinant section of the determinant line bundle be conformally invariant, corresponding to the classical invariance of the string Lagrangian defining the string path integral from which the determinant arises. That is, in order to evaluate the path integral on the reduced configuration space, one requires a trivialization of the determinant line bundle which defines a conformally invariant regularized determinant function. The above calculation says that there is a topological obstruction to this occurring when the background space dimension differs from 26.

This is the most basic example of determinant anomaly computations, which have acquired considerably more sophisticated constructions in modern versions of string theory and QFT. One immediate deficiency in the approach explained so far is that not all anomalies are topological and so even though the first Chern class of the determinant line bundle may vanish, there may still be local and global obstructions to the existence of a determinant function with the correct symmetry properties. To be more precise, one needs to say not just that a trivialization of the determinant line bundle formally exists, but to actually be able to construct a specific preferred trivialization. For this more refined objective, one needs to know more about the differential geometry of the determinant line. One approach is to fix a canonical choice of connection and, if the determinant bundle is topologically trivial, to construct a determinant section (up to phase) using the parallel transport of the connection.

The principal contribution to such a theory was made in a remarkable four page paper by Quillen (1985) in which using zeta-function regularization he presented a construction of a metric and connection on the determinant line bundle for a family of $\overline{\partial}$-operators over a Riemann surface coupled to a holomorphic vector bundle. (This is the first paper one should read on determinant line bundles; Quillen's motivation, in fact, did not come from physics but from a problem in number theory.)

To outline this construction, which was extended to general families of Dirac-type operators in Bismut and Freed (1986), first we recall that if Δ is an invertible Laplacian-type second-order elliptic differential operator acting on the space of sections of a vector bundle over a compact manifold of dimension n, then it has a spectrum consisting of real discrete eigenvalues $\{\lambda\}$ forming an unbounded subset of the positive real line. The zeta function of Δ is defined in the complex half-plane $\mathrm{Re}(s) > n/2$ by

$$\zeta(\Delta, s) = \mathrm{tr}\,(\Delta^{-s}) = \sum_\lambda \lambda^{-s}, \quad \mathrm{Re}(s) > \frac{n}{2}$$

and extends to a meromorphic function of s on the whole complex plane. It turns out that the extension has no pole at $s = 0$ and this means that we may define the zeta-function regularized determinant of Δ by

$$\det{}_\zeta(\Delta) := \exp\left(-\frac{\mathrm{d}}{\mathrm{d}s}_{|s=0} \zeta(\Delta, s)\right)$$

since $(\mathrm{d}/\mathrm{d}s)|_{s=0}\lambda^s = \log\lambda$ this formally represents a regularized product of the eigenvalues of Δ. A metric is now defined on the determinant line bundle $\mathrm{DET}(\boldsymbol{D})$ by defining the norm square of the element $\det(D_x) \in \mathrm{Det}(D_x)$ by

$$\|\det(D_x)\|^2 := \det{}_\zeta(D_x^* D_x)$$

over the subset B_0 of $x \in B$ where D_x is invertible. Elsewhere in B, one includes a factor defined by the induced L^2 metric in the kernel and cokernel. See Quillen (1985) and Bismut and Freed (1986) for full details.

A connection is defined by similarly constructing a regularized version of the connection we would define if we were working with finite-rank bundles. First, one includes in the data associated to the fibration $\pi : M \longrightarrow B$ defining the family of operators $\boldsymbol{D}$ a splitting of the tangent bundle $TM = T(M/B) \oplus \pi^*(TB)$. This assumption and the Riemannian geometry of the fibration yield a connection $\nabla^{(\pi)}$ defined along the fibers of the fibration. The connection form over B_0 is then defined by

$$\omega(x) = \mathrm{tr}_\zeta(D_x^{-1} \nabla^{(\pi)} D_x)$$

where the zeta-regularized trace tr_ζ is defined on a vertical bundle endomorphism-valued 1 form $x \mapsto A_x$ on M by

$$\mathrm{tr}_\zeta(A_x) := \mathrm{fp}_{s=0}\mathrm{tr}\left(A_x\left(D_x^* D_x\right)^{-s}\right)|^{\mathrm{mer}}$$

where the superscript indicates we are considering the meromorphically extended form, and $\mathrm{fp}_{s=0}(G(s))$ means the finite part of a meromorphic function G on $\mathbb{C}$; that is, the constant term in the Laurent expansion of $G(s)$ near $s=0$.

A theorem of Bismut and Freed, generalizing Quillen's original computation, computes the curvature $\Omega^{(\mathrm{DET}(\boldsymbol{D}))}$ of this connection to be the 2-form component in the local Atiyah–Singer families index density. This is a refined version of the topological version of that theorem which we utilized earlier; it expresses the characteristic classes on B in terms of specific canonical differential forms constructed by integrating, along the fibers of the fibration, canonically defined vertical characteristic forms. More precisely, they prove the formula (Bismut and Freed 1986 and Berline *et al.* 1992)

$$\Omega^{(\mathrm{DET}(\boldsymbol{D}))} = (2\pi\mathrm{i})^{-n/2}\left(\int_{M/B}\widehat{A}(M/B)\mathrm{ch}(\mathcal{E})\right)_{[2]} \qquad [33]$$

where $(\sigma)_{[2]} \in \Omega^2(B)$ means the 2-form component of a differential form σ on B. Here $\widehat{A}(M/B) = \det{}^{1/2}((R^{M/B}/2)/\sinh(R^{M/B}/2))$ is the vertical $\widehat{A}$-genus differential form, while $\mathrm{ch}(\mathcal{E})$ is the vertical Chern character form associated to the curvature form of the bundle $\mathcal{E}$.

This theory seems a long way from the classical theory of stable characteristic classes and the Fredholm determinant discussed in earlier sections. There are, however, interesting parallels which may guide the search for an understanding of the geometry of families of elliptic operators, of which determinants form a component. The prototypical situation where determinants arise in the quantization of gauge theory is the following. Consider the infinite-dimensional affine space $\mathcal{A}$ of connections on a complex vector bundle E with structure group G sitting over S^n the n-sphere. The Lie group G is assumed to be compact. For each connection $A \in \mathcal{A}$, we consider a Dirac operator $D_A : C^\infty(S^n, S^+ \otimes E) \longrightarrow C^\infty(S^n, S^- \otimes E)$, where E is a Hermitian vector bundle coupled to the spinor bundles $S^\pm$. The group $\mathcal{G}$ of based gauge transformations acts on $\mathcal{A}$ and symmetry properties of conservation laws lead one to be interested in constructing a determinant function on the quotient space $\mathcal{A}/\mathcal{G}$. More precisely, $g \in \mathcal{G}$ transforms D_A to $D_{g.A}$ and by equivariance the Quillen determinant section pushes down to a section of a reduced determinant line bundle over $\mathcal{A}/\mathcal{G}$. As seen earlier, the topological obstruction to realizing this determinant section as a function on $\mathcal{A}/\mathcal{G}$ can be computed from the Atiyah–Singer index theorem for families applied to the corresponding index bundle $\mathrm{Ind}(\boldsymbol{D}_{\mathcal{A}/\mathcal{G}})$ in the $K(\mathcal{A}/\mathcal{G})$ by picking out the degree-2 component in $H^2(\mathcal{A}/\mathcal{G})$ of the Chern character $\mathrm{ch}(\mathrm{Ind}(\boldsymbol{D}_{\mathcal{A}/\mathcal{G}}))$. On the other hand, it turns out that this characteristic class is the transgression of the element of $H^1(\mathcal{G}, \mathbb{Z})$ defined by the zeta-determinant trace

$$\begin{aligned}\Theta_\zeta &:= \mathrm{tr}_\zeta\left(\left(D_A^* D_{g.A}\right)^{-1} d_{\mathcal{G}}\left(D_A^* D_{g.A}\right)\right)\\ &:= \mathrm{fp}_{s=0}\mathrm{tr}\left(D_A^* D_{g.A}\right)^{-1} d_{\mathcal{G}}\left(D_A^* D_{g.A}\right)\left(D_A^* D_{g.A}\right)^{-s})|^{\mathrm{mer}}\end{aligned}$$

which counts the winding number of the zeta determinant $\mathcal{G} \longrightarrow \mathbb{C}^*$ defined by $\det_\zeta(D_A^* D_{g.A})$. This provides an interesting parallel of the classical theory described in the section "The Fredholm determinant." For more details of this and more advanced ideas take a look at Singer (1985). (A similar parallel holds between the topological derivation of the conformal anomaly outlined at the beginning of this section and what it called the Polyakov multiplicative anomaly formula for the zeta determinant of the Laplacian with respect to conformal changes in the metric on the surface.)

Aspects of more recent work in this direction have been the extension of the theory to manifolds with boundary, and how it encodes into the structures of topological and conformal field theories, see Segal (2004) and Mickelsson and Scott (2001), and more generally into M-theory (Freed and Moore 2004).

See also: Anomalies; Feynman Path Integrals; Index Theorems; Regularization for Dynamical ζ-Functions.

Further Reading

Alvarez O and Singer IM (1984) Gravitational anomalies and the families index theorem. *Communications in Mathematical Physics* 96: 409–417.

Atiyah MF and Singer IM (1984) Dirac operators coupled to vector potentials. *Proceedings of the National Academy of Sciences of the USA* 81: 2597–2600.

Bott R (1959) The stable homotopy of the classical groups. *Annals of Mathematics* 70: 313–337.

Berline N, Getzler E, and Vergne M (1992) Heat Kernels and Dirac Operators, *Grundlehren der Mathematischen Wissenschaften*, vol. 298. Berlin: Springer.

Bismut J-M and Freed DS (1986) The analysis of elliptic families I. *Communications in Mathematical Physics* 106: 159–176.

Chern SS and Simons J (1974) Characteristic forms and geometric invariants. *Annals of Mathematics* 99: 48–69.

Freed DS and Moore G (2004) Setting the quantum integrand of M-theory, arXiv:hep-th/0409135.

Mickelsson J and Scott S (2001) Functorial QFT, gauge anomalies and the Dirac determinant bundle. *Communications in Mathematical Physics* 219: 567–605.

Quillen D (1985) Determinants of Cauchy–Riemann opeators over a Riemann surface. *Functional Analysis and its Applications* 19: 31–34.

Segal G (2004) The definition of conformal field theory. In: Tillmann U (ed.) *Topology, Geometry and Quantum Field Theory*, pp. 421–577. Cambridge University Press. *Topological field theory*, "Stanford lecture notes," http://www.cgtp.duke.edu.

Singer IM (1985) Families of Dirac operators with applications to physics. *Asterique, hors serie*, 323–340.

Quivers *see* Finite-Dimensional Algebras and Quivers

R

Random Algebraic Geometry, Attractors and Flux Vacua

M R Douglas, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

Introduction

A classic question in probability theory, studied by M Kac, S O Rice, and many others, is to find the expected number and distribution of zeros or critical points of a random polynomial. The same question can be asked for random holomorphic functions or sections of bundles, and are the subject of "random algebraic geometry."

While this theory has many physical applications, in this article we focus on a variation on a standard question in the theory of disordered systems. This is to find the expected distribution of minima of a potential function randomly chosen from an ensemble, which might be chosen to model a crystal with impurities, a spin glass, or another disordered system. Now whereas standard potentials are real-valued functions, analogous functions in supersymmetric theories, such as the superpotential and the central charge, are holomorphic sections of a line bundle. Thus, one is interested in finding the distribution of critical points of a randomly chosen holomorphic section.

Two related and much-studied problems of this type are (1) the problem of finding attractor points in the sense of Ferrara, Kallosh, and Strominger, and (2) the problem of finding flux vacua as posed by Giddings, Kachru, and Polchinski. These problems involve a good deal of fascinating mathematics and are good illustrations of the general theory.

A note on general references for further reading on the subject of this article is in order. For background on random algebraic geometry and some of its other applications, as well as references in the text not listed here, consult Edelman and Kostlan (1995) and Zelditch (2001). The attractor problem is discussed in Ferrara *et al.* (1995) and Moore (2004), while IIb flux vacua were introduced in Giddings *et al.* (2002). Background on Calabi–Yau manifolds can be found in Cox and Katz (1999) and Gross *et al.* (2003).

Elementary Random Algebraic Geometry

Let us introduce this subject with the problem of finding the expected distribution of zeros of a random polynomial,

$$f(z) = c_0 + c_1 z + \cdots + c_N z^N$$

We define a random polynomial to be a probability measure on a space of polynomials. A natural choice might be independent Gaussian measures on the coefficients,

$$\mathrm{d}\mu[f] = \mathrm{d}\mu[c_0, \ldots, c_N] = \prod_{i=0}^{N} \mathrm{d}^2 c_i \frac{\sigma_i}{2\pi} \, \mathrm{e}^{-|c_i|^2/2\sigma_i^2} \quad [1]$$

We still need to choose the variances. At first the most natural choice would seem to be equal variance for each coefficient, say $\sigma_i = 1/2$. We can characterize this ensemble by its two-point function,

$$\begin{aligned} G(z_1, \bar{z}_2) &\equiv \mathbb{E}[f(z_1) f^*(\bar{z}_2)] \\ &= \int \mathrm{d}\mu[f] f(z_1) f^*(\bar{z}_2) \\ &= \sum_{n=0}^{N} (z_1 \bar{z}_2)^n \\ &= \frac{1 - z_1^{N+1} \bar{z}_2^{N+1}}{1 - z_1 \bar{z}_2} \end{aligned}$$

We now define $\mathrm{d}\mu_0(z)$ to be a measure with unit weight at each solution of $f(z) = 0$, such that its integral over a region in $\mathbb{C}$ counts the expected number of zeros in that region. It can be written in terms of the standard Dirac delta function, by multiplication by a Jacobian factor,

$$\mathrm{d}\mu_0(z) = \mathbb{E}[\delta^{(2)}(f(z)) \partial f(z) \bar{\partial} f^*(\bar{z})] \quad [2]$$

To compute this expectation value, we introduce a constrained two-point function,

$$G_{f(z)=0}(z_1, \bar{z}_2) = \frac{\mathbb{E}[\delta^{(2)}(f(z)) f(z_1) f^*(\bar{z}_2)]}{\mathbb{E}[\delta^{(2)}(f(z))]}$$

It could be explicitly computed by using the constraint $f(z) = 0$ to solve for a coefficient c_i in

the Gaussian integral, that is, projecting on the linear subspace $0 = \sum c_i z^i$. The result, in terms of $G(z_1, \bar{z}_2)$, is

$$\mathbb{E}[\delta^{(2)}(f(z))] = \frac{1}{\pi G(z, \bar{z})}$$

$$G_{f(z)=0}(z_1, \bar{z}_2) = G(z_1, \bar{z}_2) - \frac{G(z_1, \bar{z}) G(z, \bar{z}_2)}{G(z, \bar{z})}$$

as can be verified by considering

$$\mathbb{E}[\delta^{(2)}(f(z)) f(z) f^*(\bar{z}_2)] \propto G_{f(z)=0}(z, \bar{z}_2)$$
$$= G(z, \bar{z}_2) - \frac{G(z, \bar{z}) G(z, \bar{z}_2)}{G(z, \bar{z})} = 0$$

Using this, eqn [2] can be evaluated by taking derivatives:

$$\mathrm{d}\mu_0(z) = \frac{1}{G(z, \bar{z})} \lim_{z_1, z_2 \to z} D_1 \bar{D}_2 G_z(z_1, \bar{z}_2)$$
$$= \frac{1}{\pi} \partial\bar{\partial} \log G(z, \bar{z})$$

For the constant variance ensemble eqn [2],

$$\mathrm{d}\mu_0(z) = \frac{\mathrm{d}^2 z}{\pi} \left(\frac{1}{(1 - z\bar{z})^2} - \frac{(N+1)^2 (z\bar{z})^N}{(1 - (z\bar{z})^{N+1})^2} \right) \qquad [3]$$

We see that as $N \to \infty$, the zeros concentrate on the unit circle $|z| = 1$ (Hammersley 1954).

A similar formula can be derived for the distribution of roots of a real polynomial on the real axis, using $\mathrm{d}\mu(t) = \mathbb{E}[\delta(f(t)) |\mathrm{d}f/\mathrm{d}t|]$. One obtains (Kac 1943):

$$\mathrm{d}\mu_0^r(t) = \frac{\mathrm{d}t}{\pi} \sqrt{\frac{1}{(1 - t^2)^2} - \frac{(N+1)^2 t^{2N}}{(1 - t^{2N+2})^2}}$$

Integrating, one finds the expected number of real zeros of a degree N random real polynomial is $E_N \sim (2/\pi) \log N$, and as $N \to \infty$ the zeros are concentrated at $t = \pm 1$.

While concentration of measure is a fairly generic property for random polynomials, it is by no means universal. Let us consider another Gaussian ensemble, with variance $\sigma_n = N!/n!\,(N-n)!$. This choice leads to a particularly simple two-point function,

$$G(z, \bar{z}) = (1 + z\bar{z})^N \qquad [4]$$

and the distribution of zeros

$$\mathrm{d}\mu_0 = \frac{1}{\pi} \partial\bar{\partial} \log G = \frac{N\, \mathrm{d}^2 z}{\pi (1 + z\bar{z})^2} \qquad [5]$$

Rather than concentrate the zeros, in this ensemble zeros are uniformly distributed according to the volume of the Fubini–Study (SU(2)-invariant) Kähler metric

$$\omega = \partial\bar{\partial} K, \quad K = \log(1 + z\bar{z})$$

on complex projective space $\mathbb{CP}^1$.

We can better understand the different behaviors in our two examples by focusing on a Hermitian inner product (f, g) on function space, associated to the measure eqn [1] by the formal expression

$$\mathrm{d}\mu[f] = [Df]\ \mathrm{e}^{-(f,\, f)}$$

In making this precise, let us generalize a bit further and allow f to be a holomorphic section of a line bundle $\mathcal{L}$, say $\mathcal{O}(N)$ over $\mathbb{CP}^1$ in our examples. We then choose an orthonormal basis of sections $(s_i, s_j) = \delta_{ij}$, and write

$$f \equiv \sum_i c_i s_i \qquad [6]$$

and

$$\mathrm{d}\mu[f] = \frac{1}{(2\pi)^N} \prod_{i=1}^N\ \mathrm{d}^2 c_i\ \mathrm{e}^{-|c_i|^2/2}$$

We can then compute the two-point function

$$G(z_1, \bar{z}_2) \equiv \mathbb{E}[s(z_1) s^*(\bar{z}_2)] = \sum_{i=1}^N s_i(z_1) s_i^*(\bar{z}_2) \qquad [7]$$

and proceed as before.

In these terms, the simplest way to describe the measure for our first example is that it follows from the inner product on the unit circle,

$$(f, g) = \oint_{|z|=1} \frac{\mathrm{d}z}{2\pi z} f^*(z) g(z)$$

Thus, we might suspect that this has something to do with the concentration of eqn [3] on the unit circle. Indeed, this idea is made precise and generalized in Shiffman and Zelditch (2003).

Our second example belongs to a class of problems in which $\mathcal{M}$ is compact and $\mathcal{L}$ positive. In this case, the space $H^0(\mathcal{M}, \mathcal{L})$ of holomorphic sections is finite dimensional, so we can take the basis to consist of all sections. Then, if $\mathcal{M}$ is in addition Kähler, we can derive all the other data from a choice of Hermitian metric $h(f, g)$ on $\mathcal{L}$. In particular, this determines a Kähler form ω as the curvature of the metric compatible connection, and thus a volume form $\mathrm{Vol}_\omega = \omega^n/n!$. We then define the inner product to be

$$(f,\, g) = \int_{\mathcal{M}} \mathrm{Vol}_\omega\, h(f,\, g)$$

Thus, the measure equation [1] and the final distribution equation [2] are entirely determined by h. In

these terms, the underlying reason for the simplicity of eqn [5] is that we started with the SU(2)-invariant metric h, so the final distribution must be invariant as well. More generally, eqn [7] is a Szegö kernel. Taking $\mathcal{L} = \mathcal{L}_1^{\otimes N}$ for N large, this has a known asymptotic expansion, enabling a rather complete treatment (Zelditch 2001).

Our two examples also make the larger point that a wide variety of distributions are possible. Thus, to get convincing results, we must put in some information about the ensemble of random polynomials or sections which appear in the problem at hand.

The basic computation we just discussed can be vastly generalized to multiple variables, multipoint correlation functions, many different ensembles, and different counting problems. We will discuss the distribution of critical points of holomorphic sections below.

The Attractor Problem

We now turn to our physical problems. Both are posed in the context of compactification of the type IIb superstring theory on a Calabi–Yau 3-fold M. This leads to a four-dimensional effective field theory with $N = 2$ supersymmetry, determined by the geometry of M.

Let us begin by stating the attractor problem mathematically, and afterwards give its physical background. We begin by reviewing a bit of the theory of Calabi–Yau manifolds. By Yau's proof of the Calabi conjecture, the moduli space of Ricci-flat metrics on M is determined by a choice of complex structure on M, denote this J, and a choice of Kähler class. Using deformation theory, it can be shown that the moduli space of complex structures, denote this $\mathcal{M}_c(M)$, is locally a complex manifold of dimension $h^{2,1}(M)$. A point J in $\mathcal{M}_c(M)$ picks out a holomorphic 3-form $\Omega_J \in H^{3,0}(M, \mathbb{C})$, unique up to an overall choice of normalization. The converse is also true; this can be made precise by defining the period map $\mathcal{M}_c(M) \to \mathbb{P}(H^3(M, \mathbb{Z}) \otimes \mathbb{C})$ to be the class of Ω in $H^3(M, \mathbb{Z}) \otimes \mathbb{C}$ up to projective equivalence. One can prove that the period map is injective (the Torelli theorem), locally in general and globally in certain cases such as the quintic in $\mathbb{C}\mathcal{P}^4$.

Now, the data for the attractor problem is a charge, a class $\gamma \in H^3(M, \mathbb{Z})$. An attractor point for γ is then a complex structure J on M such that

$$\gamma \in H_J^{3,0}(M, \mathbb{C}) \oplus H_J^{0,3}(M, \mathbb{C}) \qquad [8]$$

This amounts to $h^{2,1}$ complex conditions on the $h^{2,1}$ complex structure moduli, so picks out isolated points in $\mathcal{M}_c(M)$, the attractor points.

There are many mathematical and physical questions one can ask about attractor points, and it would be very interesting to have a general method to find them. As emphasized by G Moore, this is one of the simplest problems arising from string theory in which integrality (here due to charge quantization) plays a central role, and thus it provides a natural point of contact between string theory and number theory. For example, one might suspect that attractor Calabi–Yau's are arithmetic, that is, are projective varieties whose defining equations live in an algebraic number field. This can be shown to always be true for $K3 \times T^2$, and there are conjectures about when this is true more generally (Moore 2004).

A simpler problem is to characterize the distribution of attractor points in $\mathcal{M}_c(M)$. As these are infinite in number, one must introduce some control parameter. While the first idea which might come to mind is to bound the magnitude of γ, since the intersection form on $H^3(M, \mathbb{Z})$ is antisymmetric, there is no natural way to do this. A better way to get a finite set is to bound the period of γ, and consider the attractor points satisfying

$$Z_{\max}^2 \geq |Z(\gamma; z)|^2 \equiv \frac{|\int_M \gamma \wedge \Omega|^2}{\int_M \Omega \wedge \bar{\Omega}} \qquad [9]$$

As an example of the type of result we will discuss below, one can show that for large $Z_{\max}$, the density of such attractor points asymptotically approaches the Weil–Peterson volume form on $\mathcal{M}_c$.

We now briefly review the origins of this problem, in the physics of 1/2 BPS (Bogomoln'yi–Prasad–Sommerfield) black holes in $N = 2$ supergravity. We begin by introducing local complex coordinates z^i on $\mathcal{M}_c(M)$. Physically, these can be thought of as massless complex scalar fields. These sit in vector multiplets of $N = 2$ supersymmetry, so there must be $h^{2,1}(M)$ vector potentials to serve as their bosonic partners under supersymmetry. These appear because the massless modes of the type IIb string include various higher rank-p form gauge potentials, in particular a self-dual 4-form which we denote C. Self-duality means that $dC = *\,dC$ up to nonlinear terms, where $*$ is the Hodge star operator in ten dimensions. Now, Kaluza–Klein reduction of this 4-form potential produces $b^3(M)$ 1-form vector potentials A_I in four dimensions. Given an explicit basis of 3-forms ω_I for $H^3(M, \mathbb{R}) \cap H^3(M, \mathbb{Z})$, this follows from the decomposition

$$C = \sum_{I=1}^{b_3} A_I \wedge \omega_I + \text{massive modes}$$

However, because of the self-duality relation, only half of these vector potentials are independent; the other half are determined in terms of them by four-dimensional electric–magnetic duality. Explicitly, given the intersection form η_{ij} on $H^3 \otimes H^3$, we have

$$\mathrm{d}A_i = \eta_{ij} *_4 \mathrm{d}A_j \qquad [10]$$

where $*_4$ denotes the Hodge star in $d = 4$. Thus we have $h^{2,1} + 1$ independent vector potentials. One of these sits in the $N = 2$ supergravity multiplet, and the rest are the correct number to pair with the complex structure moduli. We now consider 1/2 BPS black hole solutions of this four-dimensional $N = 2$ theory. Choosing any S^2 which surrounds the horizon, we can define the charge γ as the class in $H^3(M, \mathbb{Z})$ which reproduces the corresponding magnetic charges

$$Q_i = \frac{1}{2\pi}\int_{S^2} \mathrm{d}A_i \equiv \int_M \omega_i \wedge \gamma$$

Using eqn [10], this includes all charges.

One can show that the mass M of any charged object in supergravity satisfies a BPS bound,

$$M^2 \geq |Z(\gamma; z)|^2 \qquad [11]$$

The quantity $|Z(\gamma; z)|^2$, defined in eqn [9], depends explicitly on γ, and implicitly on the complex structure moduli z through Ω. A 1/2 BPS solution by definition saturates this bound.

We now explain the "attractor paradox." According to Bekenstein and Hawking, the entropy of any black hole is proportional to the area of its event horizon. This area can be found by finding the black hole as an explicit solution of four-dimensional supergravity, which clearly depends on the charge γ. In fact, we must fix boundary conditions for all the fields at infinity, in particular the complex structure moduli, to get a particular black hole solution. Now, normally varying the boundary conditions varies all the data of a solution in a continuous way. On the other hand, if the entropy has any microscopic interpretation as the logarithm of the number of quantum states of the black hole, one would expect e^S to be integrally quantized. Thus, it must remain fixed as the boundary conditions on complex structure moduli are varied, in contradiction with naive expectations for the area of the horizon, and seemingly contradicting Bekenstein and Hawking.

The resolution of this paradox is the attractor mechanism. Let us work in coordinates for which the four-dimensional metric takes the form

$$\mathrm{d}s^2 = -f(r)\,\mathrm{d}t^2 + \mathrm{d}r^2 + \frac{A(r)}{4\pi}\,\mathrm{d}\Omega_{S^2}^2$$

With some work, one can see that in the 1/2 BPS case, the equations of motion imply that as r decreases, the complex structure moduli z follow gradient flow with respect to $|Z(\gamma, z)|^2$ in eqn [11], and the area $A(r)$ of an S^2 at radius r decreases. Finally, at the horizon, z reaches a value z_* at which $|Z(\gamma, z_*)|^2$ is a local minimum, and the area of the event horizon is $A = 4\pi|Z(\gamma, z_*)|^2$. Since z_* is determined by minimization, this area will not change under small variations of the initial z, resolving the paradox.

A little algebra shows that the problem of finding nonzero critical points of $|Z(\gamma, z)|^2$ is equivalent to that of finding critical points $D_i Z = 0$ of the period associated to γ,

$$Z = \int_M \gamma \wedge \Omega \qquad [12]$$

usually called the central charge, with respect to the covariant derivative

$$D_i Z = \partial_i Z + (\partial_i K)Z \qquad [13]$$

Here

$$\mathrm{e}^{-K} \equiv \int \Omega \wedge \bar{\Omega} \qquad [14]$$

The mathematical significance of this rephrasing is that K is a Kähler potential for the Weil–Peterson Kähler metric on $\mathcal{M}_c(M)$, with Kähler form $\omega = \partial\bar{\partial}K$, and eqn [13] is the unique connection on $H^{(3,0)}(M, \mathbb{C})$ regarded as a line bundle over $\mathcal{M}_c(M)$, whose curvature is $-\omega$. These facts can be used to show that $D_i\Omega$ provides a basis for $H^{(2,1)}(M, \mathbb{C})$, so that the critical point condition forces the projection of γ on $H^{(2,1)}$ to vanish. This justifies our original definition eqn [8].

Flux Vacua in IIb String Theory

We will not describe our second problem in as much detail, but just give the analogous final formulation. In this problem, a "choice of flux" is a pair of elements of $H^3(M, \mathbb{Z})$, or equivalently a single element

$$F \in H^3(M, \mathbb{Z} \oplus \tau\mathbb{Z}) \qquad [15]$$

where $\tau \in \mathcal{H} \equiv \{\tau \in \mathbb{C} | \mathrm{Im}\tau > 0\}$ is the so-called "dilaton-axion."

A flux vacuum is then a choice of complex structure J and τ for which

$$F \in H_J^{3,0}(M, \mathbb{C}) \oplus H_J^{1,2}(M, \mathbb{C}) \qquad [16]$$

Now we have $h^{2,1} + h^{0,3} = h^{2,1} + 1$ complex conditions on the joint choice of $h^{2,1}$ complex structure

moduli and τ, so this condition also picks out special points, now in $\mathcal{M}_c \times \mathcal{H}$.

The critical point formulation of this problem is that of finding critical points of

$$W = \int \Omega \wedge F \qquad [17]$$

under the covariant derivatives eqn [13] and

$$D_\tau W = \partial_\tau W + (\partial_\tau W) Z$$

with K the sum of eqn [14] and the Kähler potential $-\log \mathrm{Im}\tau$ for the metric on the upper half-plane of constant curvature –1.

This is a sort of complexified version of the previous problem and arises naturally in IIb compactification by postulating a nonzero value F for a certain 3-form gauge field strength, the flux. The quantity eqn [17] is the superpotential of the resulting $N = 1$ supergravity theory, and it is a standard fact in this context that supersymmetric vacua (critical points of the effective potential) are critical points of W in the sense we just stated.

We can again pose the question of finding the distribution of flux vacua in $\mathcal{M}_c(M) \times \mathcal{H}$. Besides $|W|^2$, which physically is one of the contributions to the vacuum energy, we can also use the "length of the flux"

$$L = \frac{1}{\mathrm{Im}\tau} \int \mathrm{Re}\ F \wedge \mathrm{Im}\ F \qquad [18]$$

as a control parameter, and count flux vacua for which $L \leq L_{\max}$. In fact, this parameter arises naturally in the actual IIb problem, as the "orientifold three-plane charge."

What makes this problem particularly interesting physically is that it (and its analogs in other string theories) may bear on the solution of the cosmological constant problem. This begins with Einstein's famous observation that the equations of general relativity admit a one-parameter generalization,

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = 8\pi T_{\mu\nu} + \Lambda g_{\mu\nu}$$

Physically, the cosmological constant Λ is the vacuum energy, which in our flux problem takes the form $\Lambda = \cdots - 3|W|^2$ (the other terms are inessential for us here).

Cosmological observations tell us that Λ is very small, of the same order as the energy of matter in the present era, about $10^{-122} M^4_{\mathrm{Planck}}$ in Planck units. However, in a generic theory of quantum gravity, including string theory, quantum effects are expected to produce a large vacuum energy, *a priori* of order M^4_{Planck}. Finding an explanation for why the theory of our universe is in this sense nongeneric is the cosmological constant problem.

One of the standard solutions of this problem is the "anthropic solution," initiated in work of Weinberg and others, and discussed in string theory in Bousso and Polchinski (2000). Suppose that we are discussing a theory with a large number of vacuum states, all of which are otherwise candidates to describe our universe, but which differ in Λ. If the number of these vacuum states were sufficiently large, the claim that a few of these states realize a small Λ would not be surprising. But one might still feel a need to explain why our universe is a vacuum with small Λ, and not one of the multitude with large Λ.

The anthropic argument is that, according to accepted models for early cosmology, if the value of $|\Lambda|$ were even 100 times larger than what is observed, galaxies and stars could not form. Thus, the known laws of physics guarantee that we will observe a universe with Λ within this bound; it is irrelevant whether other possible vacuum states "exist" in any sense.

While such anthropic arguments are controversial, one can avoid them in this case by simply asking whether or not any vacuum state fits the observed value of Λ. Given a precise definition of vacuum state, this is a question of mathematics. Still, answering it for any given vacuum state is extremely difficult, as it would require computing Λ to 10^{-122} precision. But it is not out of reach to argue that out of a large number of vacua, some of them are expected to realize small Λ. For example, if we could show that the number of otherwise physically acceptable vacua was larger than 10^{122}, and that the distribution of Λ among these was approximately uniform over the range $(-M^4_{\mathrm{Planck}}, M^4_{\mathrm{Planck}})$, we would have made a good case for this expectation. This style of reasoning can be vastly generalized and, given favorable assumptions about the number of vacua in a theory, could lead to falsifiable predictions independent of any *a priori* assumptions about the choice of vacuum state (Douglas 2003).

Asymptotic Counting Formulas

We have just defined two classes of physically preferred points in the complex structure moduli space of Calabi–Yau 3-folds, the attractor points and the flux vacua. Both have simple definitions in terms of Hodge structure, eqn [8] and eqn [16], and both are also critical points of integral periods of the holomorphic 3-form.

This second phrasing of the problem suggests the following language. We define a random period of the holomorphic 3-form to be the period for a randomly chosen cycle in $H_3(M, \mathbb{Z})$ of the types we

just discussed (real or complex, and with the appropriate control parameters). We are then interested in the expected distribution of critical points for a random period. This brings our problem into the framework of random algebraic geometry. Before proceeding to use this framework, let us first point out some differences with the toy problems we discussed. First, while eqn [12] and eqn [17] are sums of the form eqn [6], we take not an orthonormal basis but instead a basis s_i of integral periods of Ω. Second, the coefficients c_i are not normally distributed but instead drawn from a discrete uniform distribution, that is, correspond to a choice of γ in $H^3(M,\mathbb{Z})$ or F as in eqn [15], satisfying the bounds on $|Z|$ or L. Finally, we do not normalize the distribution (which is thus not a probability measure) but instead take each choice with unit weight.

These choices can of course be modified, but are made in order to answer the question, "how many attractor points (or flux vacua) sit within a specified region of moduli space?" The answer we will get is a density $\mu(Z_{\max})$ or $\mu(L_{\max})$ on moduli space, such that as the control parameter becomes large, the number of critical points within a region R asymptotes to

$$\mathcal{N}(R;Z_{\max}) \sim \int_R \mu(Z_{\max})$$

The key observation is that to get such asymptotics, we can start with a Gaussian random element s of $H^3(M,\mathbb{R})$ or $H^3(M,\mathbb{C})$. In other words, we neglect the integral quantization of the charge or flux. Intuitively, this might be expected to make little difference in the limit that the charge or flux is large, and in fact one can prove that this simplification reproduces the leading large L or $|Z|$ asymptotics for the density of critical points, using standard ideas in lattice point counting.

This justifies starting with a two-point function like eqn [7]. While the integral periods s_i of Ω can be computed in principle (and have been in many examples) by solving a system of linear PDEs, the Picard–Fuchs equations, it turns out that one does not need such detailed results. Rather, one can use the following ansatz for the two-point function,

$$\begin{aligned} G(z_1,\bar{z}_2) &= \sum_{I=1}^{b_3} \eta^{IJ} s_I(z_1) s_J^*(\bar{z}_2) \\ &= \int_M \Omega(z_1)\wedge\bar{\Omega}(\bar{z}_2) \\ &= \exp -K(z_1,\bar{z}_2) \end{aligned}$$

In words, the two-point function is the formal continuation of the Kähler potential on $\mathcal{M}_c(M)$ to independent holomorphic and antiholomorphic variables. This incorporates the quadratic form appearing in eqn [18] and can be used to count sections with such a bound.

We can now follow the same strategy as before, by introducing an expected density of critical points,

$$\mathrm{d}\mu(z) = \mathbb{E}[\delta^{(n)}(D_i s(z))\delta^{(n)}(\bar{D}_{\bar{i}}\bar{s}(\bar{z}))\mid \det_{1\le i,j\le 2n} H_{ij}|] \quad [19]$$

where the "complex Hessian" H is the $2n\times 2n$ matrix of second derivatives

$$H \equiv \begin{pmatrix} \partial_i\bar{D}_{\bar{j}}\bar{s}(z) & \partial_i D_j s(z) \\ \bar{\partial}_{\bar{i}}\bar{D}_{\bar{j}}\bar{s}(z) & \bar{\partial}_{\bar{i}} D_j s(z) \end{pmatrix} \quad [20]$$

(note that $\partial Ds = DDs$ at a critical point). One can then compute this density along the same lines. The holomorphy of s implies that $\partial_i\bar{D}_{\bar{j}}s = \omega_{i\bar{j}}s$, which is one simplification. Other geometric simplifications follow from the fact that eqn [19] depends only on s and a finite number of its derivatives at the point z.

For the attractor problem, using the identity

$$D_iD_j s = \mathcal{F}_{ijk}\omega^{k\bar{k}}\bar{D}_{\bar{k}}s = 0$$

from special geometry of Calabi–Yau 3-folds, the Hessian becomes trivial, and $\det H = |s|^{2n}$. One thus finds (Denef and Douglas 2004) that the asymptotic density of attractor points with large $|Z| \le Z_{\max}$ in a region R is

$$\mathcal{N}(R,|Z|\le Z_{\max}) \sim \frac{2^{n+1}}{(n+1)\pi^n} Z_{\max}^{n+1}\cdot \mathrm{vol}(R)$$

where $\mathrm{vol}(R) = \int_R \omega^n/n!$ is the volume of R in the Weil–Peterson metric. The total volume is known to be finite for Calabi–Yau 3-fold moduli spaces, and thus so is the number of attractor points under this bound.

The flux vacuum problem is complicated by the fact that DDs is nonzero and thus the determinant of the Hessian does not take a definite sign, and implementing the absolute value in eqn [19] is nontrivial. The result (Douglas, *et al.* 2004) is

$$\begin{aligned} \mu(z) \sim \frac{1}{b_3!\sqrt{\det\Lambda(z)}} \int_{\mathcal{H}(z)\times\mathbb{C}} &|\det(HH^* - |x|^2\cdot\mathbf{1})| \\ &\times \mathrm{e}^{H^{\mathrm{t}}\Lambda(z)^{-1}H - |x|^2}\,\mathrm{d}H\,\mathrm{d}x \end{aligned}$$

where $\mathcal{H}(z)$ is the subspace of Hessian matrices eqn [20] obtainable from periods at the point z, and $\Lambda(z)$ is a covariance matrix computable from the period data.

A simpler lower bound for the number of solutions can be obtained by instead computing the index density

$$\mu_I(z) = \mathbb{E}\left[\delta^{(n)}(D_i s)\delta^{(n)}(\bar{D}_{\bar{i}}\bar{s}) \det_{1\le i,j\le 2n} H_{ij}\right] \qquad [21]$$

so-called because it weighs the vacua with a Morse–Witten sign factor. This admits a simple explicit formula (Ashok and Douglas 2004),

$$I_{\rm vac}(R, L \le L_{\max}) \sim \frac{(2\pi L_{\max})^{b_3}}{\pi^{n+1} b_3!}\int_R \det(\mathcal{R} + \omega \cdot 1) \qquad [22]$$

where $\mathcal{R}$ is the $(n+1)\times(n+1)$-dimensional matrix of curvature 2-forms for the Weil–Peterson metric.

One might have guessed this density by the following reasoning. If s had been a single-valued section on a compact $\mathcal{M}_c$ (it is not), topological arguments determine the total index to be $[c_{n+1}(\mathcal{L}\otimes T^*\mathcal{M})]$, and this is the simplest density constructed solely from the metric and curvatures in the same cohomology class.

It is not in general known whether this integral over Calabi–Yau moduli space is finite, though this is true in examples studied so far. One can also control $|W|^2$ as well as other observables, and one finds that the distribution of $|W|^2$ among flux vacua is to a good approximation uniform. Considering explicit examples, the prefactor in eqn [22] is of order 10^{100}–10^{300}, so assuming that this factor dominates the integral, we have justified the Bousso–Polchinski solution to the cosmological constant problem in these models.

The finite L corrections to these formulas can be estimated using van der Corput techniques, and are suppressed by better than the naive $L^{-1/2}$ or $|Z|^{-1}$ one might have expected. However the asymptotic formulas for the numbers of flux vacuum break down in certain limits of moduli space, such as the large complex structure limit. This is because eqn [18] is an indefinite quadratic form, and the fact that it bounds the number of solutions at all is somewhat subtle. These points are discussed at length in (Douglas *et al.* 2005).

Similar results have been obtained for a wide variety of flux vacuum counting problems, with constraints on the value of the effective potential at the minimum, on the masses of scalar fields, on scales of supersymmetry breaking, and so on. And in principle, this is just the tip of an iceberg, as the study of more or less any class of superstring vacua leads to similar questions of counting and distribution, less well understood at present. Some of these are discussed in Douglas (2003), Acharya *et al.* (2005), Denef and Douglas (2005), Blumenhagen *et al.* (2005).

See also: Black Hole Mechanics; Chaos and Attractors; Compactification of Superstring Theory; Supergravity.

Further Reading

Acharya BS, Denef F, and Valandro R (2005) Statistics of M theory vacua. *Journal of High Energy Physics* 0506: 056.

Ashok S and Douglas MR (2004) Counting flux vacua. *Journal of High Energy Physics* 0401: 060.

Blumenhagen R, Gmeiner F, Honecker G, Lust D, and Weigand T (2005) The statistics of supersymmetric D-brane models. *Nuclear Physics B* 713: 83.

Bousso R and Polchinski J (2000) Quantization of four-form fluxes and dynamical neutralization of the cosmological constant. *Journal of High Energy Physics* 06: 06.

Cox DA and Katz S (1999) *Mirror Symmetry and Algebraic Geometry*. Providence, RI: American Mathematical Society.

Denef F and Douglas MR (2004) Distributions of flux vacua. *Journal of High Energy Physics* 0405: 072.

Denef F and Douglas MR (2005) Distributions of nonsupersymmetric flux vacua. *Journal of High Energy Physics* 0503: 061.

Douglas MR (2003) The statistics of string/M theory vacua. *Journal of High Energy Physics* 5: 046.

Douglas MR, Shiffman B, and Zelditch S (2004) Critical points and supersymmetric vacua I. *Communications in Mathematical Physics* 252(1–3): 325–358; II: Asymptotics (to appear).

Douglas MR, Shiffman B, and Zelditch S (2005) Critical points and supersymmetric vacua, III: String/M models. Submitted to *Communications in Mathematical Physics*. http://arxiv./org/abs/math-ph/0506015.

Edelman A and Kostlan E (1995) How many zeros of a random polynomial are real? *Bull. Amer. Math. Soc. (N.S.)* 32: 1–37.

Ferrara S, Kallosh R, and Strominger A (1995) $N=2$ extremal black holes, *Physical Review D* 52: 5412.

Giddings SB, Kachru S, and Polchinski J (2002) Hierarchies from fluxes in string compactifications. *Physical Review D (3)* 66(10): 106006.

Gross M, Huybrechts D, and Joyce D (2003) *Calabi–Yau Manifolds and Related Geometries*. New York: Springer.

Moore GW (2004) Les Houches lectures on strings and arithmetic, arXiv:hep-th/0401049.

Shiffman B and Zelditch S (2003) Equilibrium distribution of zeros of random polynomials. *Int. Math. Res. Not.* 1: 25–49.

Zelditch S (2001) From random polynomials to symplectic geometry. In: *XIIIth International Congress of Mathematical Physics*, pp. 367–376. Beijing: Higher Education Press.

Zelditch S (2005) Random complex geometry and vacua, or: How to count universes in string/M theory, 2005 preprint.

Random Dynamical Systems

V Araújo, Universidade do Porto, Porto, Portugal

Introduction

The concept of random dynamical system is a comparatively recent development combining ideas and methods from the well-developed areas of probability theory and dynamical systems.

Let us consider a mathematical model of some physical process given by the iterates $T_0^k = T_0 \circ \overset{k}{\cdots} \circ T_0$, $k \geq 1$, of a smooth transformation $T_0 : M \circlearrowleft$ of a manifold into itself. A realization of the process with initial condition x_0 is modeled by the sequence $(T_0^k(x_0))_{k\geq 1}$, the orbit of x_0.

Due to our inaccurate knowledge of the particular physical system or due to computational or theoretical limitations (e.g., lack of sufficient computational power, inefficient algorithms, or insufficiently developed mathematical or physical theory), the mathematical models never correspond exactly to the phenomenon they are meant to model. Moreover, when considering practical systems, we cannot avoid either external noise or measurement or inaccuracy errors, so every realistic mathematical model should allow for small errors along orbits not to disturb the long-term behavior too much. To be able to cope with unavoidable uncertainty about the "correct" parameter values, observed initial states and even the specific mathematical formulation involved, let randomness be embedded within the model to begin with.

This article presents the most basic classes of models, defines the general concept, and presents some developments and examples of applications.

Dynamics with Noise

To model random perturbations of a transformation T_0, we may consider a transition from the image $T_0(x)$ to some point according to a given probability law, obtaining a Markov chain, or, if T_0 depends on a parameter p, we may choose p at random at each iteration, which also can be seen as a Markov chain but whose transitions are strongly correlated.

Random Noise

Given $T_0 : M \circlearrowleft$ and a family $\{p(\cdot \mid x) : x \in M\}$ of probability measures on M such that the support of $p(\cdot \mid x)$ is close to $T_0(x)$, the random orbits are sequences $(x_k)_{k\geq 1}$ where each x_{k+1} is a random variable with law $p(\cdot \mid x_k)$. This is a Markov chain with state space M and transition probabilities $\{p(\cdot \mid x)\}_{x\in M}$. To extend the concept of invariant measure of a transformation to this setting, a probability measure μ is said to be "stationary" if $\mu(A) = \int p(A \mid x)\, \mathrm{d}\mu(x)$ for every measurable (Borel) subset A. This can be conveniently translated by saying that the skew-product measure $\mu \times p^{\mathbb{N}}$ on $M \times M^{\mathbb{N}}$ given by

$$\begin{aligned}&\mathrm{d}(\mu \times p^{\mathbb{N}})(x_0, x_1, \ldots, x_n, \ldots)\\ &\quad = \mathrm{d}\mu(x_0)p(\mathrm{d}x_1 \mid x_0) \cdots p(\mathrm{d}x_{n+1} \mid x_n) \cdots\end{aligned}$$

is invariant by the shift map $\mathcal{S} : M \times M^{\mathbb{N}} \circlearrowleft$ on the space of orbits. Hence, we may use the ergodic theorem and get that time averages of all continuous observables $\varphi : M \to \mathbb{R}$, that is, writing $\underline{x} = (x_k)_{k\geq 0}$ and

$$\begin{aligned}\tilde{\varphi}(\underline{x}) &= \lim_{n\to+\infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi(x_k)\\ &= \lim_{n\to+\infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi(\pi_0(\mathcal{S}^k(\underline{x})))\end{aligned}$$

exist for $\mu \times p^{\mathbb{N}}$-almost all sequences $\underline{x}$, where $\pi_0 : M \times M^{\mathbb{N}} \to M$ is the natural projection on the first coordinate. It is well known that stationary measures always exist if the transition probabilities $p(\cdot \mid x)$ depend continuously on x.

A function $\varphi : M \to \mathbb{R}$ is invariant if $\varphi(x) = \int \varphi(z)p(\mathrm{d}z \mid x)$ for μ-almost every x. We then say that μ is ergodic if every invariant function is constant μ-almost everywhere. Using the ergodic theorem again, if μ is ergodic, then $\tilde{\varphi} = \int \varphi\, \mathrm{d}\mu$, μ-almost everywhere.

Stationary measures are the building blocks for more sophisticated analysis involving, for example, asymptotic sojourn times, Lyapunov exponents, decay of correlations, entropy and/or dimensions, exit/entrance times from/to subsets of M, to name just a few frequent notions of dynamical and probabilistic/statistical nature.

Example 1 (Random jumps). Given $\epsilon > 0$ and $T_0 : M \to M$, let us define

$$p^{\epsilon}(A \mid x) = \frac{m(A \cap B(T_0(x), \epsilon))}{m(B(T_0(x), \epsilon))}$$

where m denotes some choice of Riemannian volume form on M. Then $p^{\epsilon}; (\cdot \mid x)$ is the normalized volume restricted to the ϵ-neighborhood of $T_0(x)$.

This defines a family of transition probabilities allowing the points to "jump" from $T_0(x)$ to any point in the ϵ-neighborhood of $T_0(x)$ following a uniform distribution law.

Random Maps

Alternatively, we may choose maps $T_1, T_2, \ldots, T_k$ independently at random near T_0 according to a probability law ν on the space $T(M)$ of maps, whose support is close to T_0 in some topology, and consider sequences $x_k = T_k \circ \cdots \circ T_1(x_0)$ obtained through random iteration, $k \geq 1$, $x_0 \in M$.

This is again a Markov chain whose transition probabilities are given for any $x \in M$ by

$$p(A \mid x) = \nu(\{T \in T(M): T(x) \in A\})$$

so this model may be reduced to the first one. However, in the random-maps setting, we may associate, with each random orbit, a sequence of maps which are iterated, enabling us to use "robust properties" of the transformation T_0 (i.e., properties which are known to hold for T_0 and for every nearby map T) to derive properties of the random orbits.

Under some regularity conditions on the map $x \mapsto p(A \mid x)$ for every Borel subset A, it is possible to represent random noise by random maps on suitably chosen spaces of transformations. In fact, the transition probability measures obtained in the random-maps setting exhibit strong spatial correlation: $p(\cdot \mid x)$ is close to $p(\cdot \mid y)$ as x is near y.

If we have a parametrized family $T: \mathcal{U} \times M \to M$ of maps, we can specify the law ν by giving a probability θ on $\mathcal{U}$. Then with every sequence $T_1, \ldots, T_k, \ldots$ of maps of the given family, we associate a sequence $\omega_1, \ldots, \omega_k, \ldots$ of parameters in $\mathcal{U}$ since

$$T_k \circ \cdots \circ T_1 = T_{\omega_k} \circ \cdots \circ T_{\omega_1} = T^k_{\omega_1, \ldots, \omega_k}$$

for all $k \geq 1$, where we write $T_\omega(x) = T(\omega, x)$. In this setting, the shift map $\mathcal{S}$ becomes a skew-product transformation

$$\mathcal{S}: M \times \mathcal{U}^{\mathbb{N}} \circlearrowleft \quad (x, \underline{\omega}) \mapsto (T_{\omega_1}(x), \sigma(\underline{\omega}))$$

to which many of the standard methods of dynamical systems and ergodic theory can be applied, yielding stronger results that can be interpreted in random terms.

Example 2 (Parametric noise). Let $T: P \times M \to M$ be a smooth map where P, M are finite-dimensional Riemannian manifolds. We fix $p_0 \in P$, denote by m some choice of Riemannian volume form on P, set $T_w(x) = T(w, x)$, and for every $\epsilon > 0$ write $\theta_\epsilon = (m(B(p_0, \epsilon))^{-1} \cdot (m \mid B(p_0, \epsilon))$, the normalized restriction of m to the ϵ-neighborhood of p_0. Then $(T_w)_{w \in P}$, together with θ_ϵ, defines a random perturbation of T_{p_0}, for every small enough $\epsilon > 0$.

Example 3 (Global additive perturbations). Let M be a homogeneous space, that is, a compact connected Lie group admitting an invariant Riemannian metric. Fixing a neighborhood $\mathcal{U}$ of the identity $e \in M$, we can define a map $T: \mathcal{U} \times M \to M$, $(u, x) \mapsto L_u(T_0(x))$, where $L_u(x) = u \cdot x$ is the left translation associated with $u \in M$. The invariance of the metric means that left (and also right) translations are isometries, hence fixing $u \in \mathcal{U}$ and taking any $(x, v) \in TM$, we get

$$\begin{aligned} \|DT_u(x) \cdot v\| &= \|DL_u(T_0(x))(DT_0(x) \cdot v)\| \\ &= \|DT_0(x) \cdot v\| \end{aligned}$$

In the particular case of $M = \mathbb{T}^d$, the d-dimensional torus, we have $T_u(x) = T_0(x) + u$, and this simplest case suggests the name "additive random perturbations" for random perturbations defined using families of maps of this type.

For the probability measure on $\mathcal{U}$, we may take θ_ϵ, any probability measure supported in the ϵ-neighborhood of e and absolutely continuous with respect to the Riemannian metric on M, for any $\epsilon > 0$ small enough.

Example 4 (Local additive perturbations). If $M = \mathbb{R}^d$ and U_0 is a bounded open subset of M strictly invariant under a diffeomorphism T_0, that is, closure $(T_0(U_0)) \subset U_0$, then we can define an isometric random perturbation setting:

(i) $V = T_0(U_0)$ (so that closure $(V) =$ closure $(T_0(U_0)) \subset U_0$);
(ii) $G \simeq \mathbb{R}^d$ the group of translations of $\mathbb{R}^d$; and
(iii) $\mathcal{V}$ a small enough neighborhood of 0 in G.

Then for $v \in \mathcal{V}$ and $x \in V$, we set $T_v(x) = x + v$, with the standard notation for vector addition, and clearly T_v is an isometry. For θ_ϵ, we may take any probability measure on the ϵ-neighborhood of 0, supported in $\mathcal{V}$ and absolutely continuous with respect to the volume in $\mathbb{R}^d$, for every small enough $\epsilon > 0$.

Random Perturbations of Flows

In the continuous-time case, the basic model to start with is an ordinary differential equation $\mathrm{d}X_t = f(t, X_t)\mathrm{d}t$, where $f: [0, +\infty) \to \mathcal{X}(M)$ and $\mathcal{X}(M)$ is the family of vector fields in M. We embed randomness in the differential equation

basically through "diffusion," the perturbation is given by white noise or Brownian motion "added" to the ordinary solution.

In this setting, assuming for simplicity that $M = \mathbb{R}^n$, the random orbits are solutions of stochastic differential equations

$$\begin{aligned} dX_t &= f(t, X_t)dt + \epsilon \cdot \sigma(t, X_t)dW_t, \\ 0 \le t \le T,\ X_0 &= Z \end{aligned}$$

where Z is a random variable, $\epsilon, T > 0$ and both $f : [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$ and $\sigma : [0, T] \times \mathbb{R}^n \to \mathcal{L}(\mathbb{R}^k, \mathbb{R}^n)$ are measurable functions. The space of linear maps $\mathbb{R}^k \to \mathbb{R}^n$ is written on $\mathcal{L}(\mathbb{R}^k, \mathbb{R}^n)$ and W_t is the white-noise process on $\mathbb{R}^k$. The solution of this equation is a stochastic process:

$$X : \mathbb{R} \times \Omega \to M \quad (t, \omega) \mapsto X_t(\omega)$$

for some (abstract) probability space Ω, given by

$$X_t = Z + \int_0^T f(s, X_s)ds + \int_0^T \epsilon \cdot \sigma(s, X_s)dW_s$$

where the last term is a stochastic integral in the sense of Itô. Under reasonable conditions on f and σ, there exists a unique solution with continuous paths, that is,

$$[0, +\infty) \ni t \mapsto X_t(\omega)$$

is continuous for almost all $\omega \in \Omega$ (in general these paths are nowhere differentiable).

Setting $Z = \delta_{x_0}$, the probability measure concentrated on the point x_0, the initial point of the path is x_0 with probability 1. We write $X_t(\omega)x_0$ for paths of this type. Hence, $x \mapsto X_t(\omega)x$ defines a map $X_t(\omega) : M \circlearrowleft$ which can be shown to be a homeomorphism and even diffeomorphisms under suitable conditions on f and σ. These maps satisfy a cocycle property

$$\begin{aligned} X_0(\omega) &= \mathrm{Id}_M \quad (\text{identity map of } M) \\ X_{t+s}(\omega) &= X_t(\theta(s)(\omega)) \circ X_s(\omega) \end{aligned}$$

for $s, t \ge 0$ and $\omega \in \Omega$, for a family of measure-preserving transformations $\theta(s) : (\Omega, \mathbb{P}) \circlearrowleft$ on a suitably chosen probability space $(\Omega, \mathbb{P})$. This enables us to write the solution of this kind of equations also as a skew product.

The Abstract Framework

The illustrative particular cases presented can all be written in skew-product form as follows.

Let $(\Omega, \mathbb{P})$ be a given probability space, which will be the model for the noise, and let $\mathbb{T}$ be time, which usually means $\mathbb{Z}_+, \mathbb{Z}$ (discrete, resp. invertible system) or $\mathbb{R}_+, \mathbb{R}$ (continuous, resp. invertible system). A random dynamical system is a skew product

$$\mathcal{S}_t : \Omega \times M \circlearrowleft, (\omega, x) \mapsto (\theta(t)(\omega), \varphi(t, \omega)(x))$$

for all $t \in \mathbb{T}$, where $\theta : \mathbb{T} \times \Omega \to \Omega$ is a family of measure-preserving maps $\theta(t) : (\Omega, \mathbb{P}) \circlearrowleft$ and $\varphi : \mathbb{T} \times \Omega \times M \to M$ is a family of maps $\varphi(t, \omega) : M \circlearrowleft$ satisfying the cocycle property: for $s, t \in \mathbb{T}$, $\omega \in \Omega$,

$$\begin{aligned} \varphi(0, \omega) &= \mathrm{Id}_M \\ \varphi(t + s, \omega) &= \varphi(t, \theta(s)(\omega)) \circ \varphi(s, \omega) \end{aligned}$$

In this general setting an invariant measure for the random dynamical system is any probability measure μ on $\Omega \times M$ which is $\mathcal{S}_t$-invariant for all $t \in \mathbb{T}$ and whose marginal is $\mathbb{P}$, that is, $\mu(\mathcal{S}_t^{-1}(U)) = \mu(U)$ and $\mu(\pi_\Omega^{-1}(U)) = \mathbb{P}(U)$ for every measurable $U \subset \Omega \times M$, respectively, with $\pi_\Omega : \Omega \times M \to \Omega$ the natural projection.

Example 5 In the setting of the previous examples of random perturbations of maps, the product measure $\eta = \mathbb{P} \times \mu$ on $\Omega \times M$, with $\Omega = \mathcal{U}^{\mathbb{N}}$, $\mathbb{P} = \theta_\epsilon^{\mathbb{N}}$ and μ any stationary measure, is clearly invariant. However, not all invariant measures are product measures of this type.

Naturally an invariant measure is ergodic if every $\mathcal{S}_t$-invariant function is μ-almost everywhere constant. That is, if $\psi : \Omega \times M \to \mathbb{R}$ satisfies $\psi \circ \mathcal{S}_t = \psi$ μ-almost everywhere for every $t \in \mathbb{T}$, then ψ is μ-almost everywhere constant.

Applications

The well-established applications of both probability or stochastic differential equations (solution of boundary value problems, optimal stopping, stochastic control etc.) and dynamical systems (all kinds of models of physical, economic or biological phenomena, solutions of differential equations, control systems etc.) will not be presented here. Instead, this section focuses on topics where the subject sheds new light on these areas.

Products of Random Matrices and the Multiplicative Ergodic Theorem

The following celebrated result on products of random matrices has far-reaching applications on dynamical systems theory.

Let $(X_n)_{n \ge 0}$ be a sequence of independent and identically distributed random variables on the probability space (Ω, P) with values in $\mathcal{L}(\mathbb{R}^k, \mathbb{R}^k)$ such that $E(\log^+ \|X_1\|) < +\infty$, where $\log^+ x = \max\{0, \log x\}$ and $\|\cdot\|$ is a given norm on

$\mathcal{L}(\mathbb{R}^k, \mathbb{R}^k)$. Writing $\varphi_n(\omega) = X_n(\omega) \circ \cdots \circ X_1(\omega)$ for all $n \geq 1$ and $\omega \in \Omega$ we obtain a cocycle. If we set

$$B = \left\{ (\omega, y) \in \Omega \times \mathbb{R}^k : \lim_{n \to +\infty} \frac{1}{n} \log \|\varphi_n(\omega) y\| \right.$$
$$\left. \text{exists and is finite or is } -\infty \right\},$$
$$\Omega' = \{\omega \in \Omega : (\omega, y) \in B \text{ for all } y \in \mathbb{R}^k\}$$

then Ω' contains a subset Ω'' of full probability and there exist random variables (which might take the value $-\infty$) $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ with the following properties.

1. Let $I = \{k+1 = i_1 > i_2 > \cdots > i_{l+1} = 1\}$ be any $(l+1)$-tuple of integers and then we define

$$\Omega_I = \{\omega \in \Omega'' : \lambda_i(\omega) = \lambda_j(\omega), i_h > i, j \geq i_{h+1},$$
$$\text{and } \lambda_{i_h}(\omega) > \lambda_{i_{h+1}}(\omega) \text{ for all } 1 < h < l\}$$

the set of elements where the sequence λ_i jumps exactly at the indexes in I. Then for $\omega \in \Omega_I$, $1 < h \leq l$,

$$\Sigma_{I,h}(\omega) = \left\{ y \in \mathbb{R}^k : \lim_{n \to +\infty} \frac{1}{n} \log \|\varphi_n(\omega)\| \leq \lambda_{i_h}(\omega) \right\}$$

is a vector subspace with dimension $i_{h-1} - 1$.
2. Setting $\Sigma_{I,k+1}(\omega) = \{0\}$, then

$$\lim_{n \to +\infty} \frac{1}{n} \log \|\varphi_n(\omega)\| = \lambda_{i_h}(\omega)$$

for every $y \in \Sigma_{I,h}(\omega) \setminus \Sigma_{I,h+1}(\omega)$.
3. For all $\omega \in \Omega''$ there exists the matrix

$$A(\omega) = \lim_{n \to +\infty} [(\varphi_n(\omega))^* \varphi_n(\omega)]^{1/2n}$$

whose eigenvalues form the set $\{e^{\lambda_i} : i = 1, \ldots, k\}$.

The values of λ_i are the random Lyapunov characteristics and the corresponding subspaces are analogous to random eigenspaces. If the sequence $(X_n)_{n \geq 0}$ is ergodic, then the Lyapunov characteristics become nonrandom constants, but the Lyapunov subspaces are still random.

We can easily deduce the multiplicative ergodic theorem for measure-preserving differentiable maps (T_0, μ) on manifolds M from this result. For simplicity, we assume that $M \subset \mathbb{R}^k$ and set $p(A \mid x) = \delta_{T_0(x)}(A) = 1$ if $T_0(x) \in A$ and 0 otherwise. Then the measure $\mu \times p^{\mathbb{N}}$ on $M \times M^{\mathbb{N}}$ is σ-invariant (as defined earlier) and we have that $\pi_0 \circ \sigma = T_0 \circ \pi_0$, where $\pi_0 : M^{\mathbb{N}} \to M$ is the projection on the first coordinate, and also $(\pi_0)_*(\mu \times p^{\mathbb{N}}) = \mu$. Then, setting for $n \geq 1$

$$X : M \to \mathcal{L}(\mathbb{R}^k, \mathbb{R}^k) \quad \text{and} \quad X_n = X \circ \pi_0 \circ \sigma^n$$
$$x \mapsto DT_0(x)$$

we obtain a stationary sequence to which we can apply the previous result, obtaining the existence of Lyapunov exponents and of Lyapunov subspaces on a full measure subset for any C^1 measure-preserving dynamical system.

By a standard extension of the previous setup, we obtain a random version of the multiplicative ergodic theorem. We take a family of skew-product maps $\mathcal{S}_t : \Omega \times M \circlearrowleft$ as in the section "The abstract framework" with an invariant probability measure μ and such that $\varphi(t, \omega) : M \circlearrowleft$ is (for simplicity) a local diffeomorphism. We then consider the stationary family

$$X_t : \Omega \to \mathcal{L}(TM),\ \omega \mapsto D\varphi(t, \omega) : TM \circlearrowleft \quad t \in \mathbb{T}$$

where $D\varphi(t, \omega)$ is the tangent map to $\varphi(t, \omega)$. This is a cocycle since for all $t, s \in \mathbb{T}$, $\omega \in \Omega$ we have

$$X(s + t, \omega) = X(s, \theta(t)\omega) \circ X(t, \omega)$$

If we assume that

$$\sup_{0 \leq t \leq 1} \sup_{x \in M} \left(\log^+ \|D\varphi(t, \omega)(x)\| \right) \in L^1(\Omega, \mathbb{P})$$

where $\|\cdot\|$ denotes the norm on the corresponding space of linear maps given by the induced norm (from the Riemannian metric) on the appropriate tangent spaces, then we obtain a sequence of random variables (which might take the value $-\infty$) $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$, with k being the dimension of M, such that

$$\lim_{t \to +\infty} \frac{1}{t} \log \|X_t(\omega, x) y\| = \lambda_i(\omega, x)$$

for every $y \in E_i\omega, x) = \Sigma_i(\omega, x) \setminus \Sigma_{i+1}(\omega, x)$ and $i = 1, \ldots, k+1$, where $(\Sigma_i(\omega, x))_i$ is a sequence of vector subspaces in $T_x M$ as before, measurable with respect to (ω, x). In this setting, the subspaces $E_i(\omega, x)$ and the Lyapunov exponents are invariant, that is, for all $t \in \mathbb{T}$ and μ-almost every $(\omega, x) \in \Omega \times M$, we have

$$\lambda_i(\mathcal{S}_t(\omega, x)) = \lambda_i(\omega, x) \quad \text{and} \quad E_i(\mathcal{S}_t(\omega, x)) = E_i(\omega, x)$$

The dependence of Lyapunov exponents on the map T_0 has been a fruitful and central research program in dynamical systems for decades extending to the present day. The random multiplicative ergodic theorem sets the stage for the study of the stability of Lyapunov exponents under random perturbations.

Stochastic Stability of Physical Measures

The development of the theory of dynamical systems has shown that models involving expressions as simple as quadratic polynomials (as the logistic family or Hénon attractor), or autonomous ordinary

differential equations with a hyperbolic singularity of saddle type, as the Lorenz flow, exhibit sensitive dependence on initial conditions, a common feature of chaotic dynamics: small initial differences are rapidly augmented as time passes, causing two trajectories originally coming from practically indistinguishable points to behave in a completely different manner after a short while. Long-term predictions based on such models are unfeasible, since it is not possible to both specify initial conditions with arbitrary accuracy and numerically calculate with arbitrary precision.

Physical measures Inspired by an analogous situation of unpredictability faced in the field of statistical mechanics/thermodynamics, researchers focused on the statistics of the data provided by the time averages of some observable (a continuous function on the manifold) of the system. Time averages are guaranteed to exist for a positive-volume subset of initial states (also called an observable subset) on the mathematical model if the transformation, or the flow associated with the ordinary differential equation, admits a smooth invariant measure (a density) or a physical measure.

Indeed, if μ_0 is an ergodic invariant measure for the transformation T_0, then the ergodic theorem ensures that for every μ-integrable function $\varphi : M \to \mathbb{R}$ and for μ-almost every point x in the manifold M, the time average $\tilde{\varphi}(x) = \lim_{n\to+\infty} n^{-1} \sum_{j=0}^{n-1} \varphi(T_0^j(x))$ exists and equals the space average $\int \varphi \, \mathrm{d}\mu_0$. A physical measure μ is an invariant probability measure for which it is required that time averages of every continuous function φ exist for a positive Lebesgue measure (volume) subset of the space and be equal to the space average $\mu(\varphi)$.

We note that if μ is a density, that is, absolutely continuous with respect to the volume measure, then the ergodic theorem ensures that μ is physical. However, not every physical measure is absolutely continuous. To see why in a simple example, we consider a singularity p of a vector field which is an attracting fixed point (a sink), then the Dirac mass δ_p concentrated on p is a physical probability measure, since every orbit in the basin of attraction of p will have asymptotic time averages for any continuous observable φ given by $\varphi(p) = \delta_p(\varphi)$.

Physical measures need not be unique or even exist in general but, when they do exist, it is desirable that the set of points whose asymptotic time averages are described by physical measures (such a set is called the basin of the physical measures) be of full Lebesgue measure – only an exceptional set of points with zero volume would not have a well-defined asymptotic behavior. This is yet far from being proved for most dynamical systems, in spite of much recent progress in this direction.

There are robust examples of systems admitting several physical measures whose basins together are of full Lebesgue measure, where "robust" means that there are whole open sets of maps of a manifold in the C^2 topology exhibiting these features. For typical parametrized families of one-dimensional unimodal maps (maps of the circle or of the interval with a unique critical point), it is known that the above scenario holds true for Lebesgue almost every parameter. It is known that there are systems admitting no physical measure, but the only known cases are not robust, that is, there are systems arbitrarily close which admit physical measures.

It is hoped that conclusions drawn from models admitting physical measures to be effectively observable in the physical processes being modeled. In order to lend more weight to this expectation, researchers demand stability properties from such invariant measures.

Stochastic stability There are two main issues concerning a mathematical model, both from theoretical and practical standpoints. The first one is to describe the asymptotic behavior of most orbits, that is, to understand what happens to orbits when time tends to infinity. The second and equally important one is to ascertain whether the asymptotic behavior is stable under small changes of the system, that is, whether the limiting behavior is still essentially the same after small changes to the law of evolution. In fact, since models are always simplifications of the real system (we cannot ever take into account the whole state of the universe in any model), the lack of stability considerably weakens the conclusions drawn from such models, because some properties might be specific to it and not in any way resembling the real system.

Random dynamical systems come into play in this setting when we need to check whether a given model is stable under small random changes to the law of evolution.

In more precise terms, we suppose that there is a dynamical system (a transformation or a flow) admitting a physical measure μ_0 and we take any random dynamical system obtained from this one through the introduction of small random perturbations on the dynamics, as in Examples 1–4 or in the section on "Random perturbations of flows," with the noise level $\epsilon > 0$ close to zero.

In this setting if, for any choice μ_ϵ of invariant measure for the random dynamical system for all $\epsilon > 0$ small enough, the set of accumulation points of

the family $(\mu_\epsilon)_{\epsilon>0}$, when ϵ tends to 0 – also known as zero-noise limits – is formed by physical measures or, more generally, by convex linear combinations of physical measures, then the original unperturbed dynamical system is stochastically stable.

This intuitively means that the asymptotic behavior measured through time averages of continuous observables for the random system is close to the behavior of the unperturbed system.

Recent progress in one-dimensional dynamics has shown that, for typical families $(f_t)_{t\in(0,1)}$ of maps of the circle or of the interval having a unique critical point, a full Lebesgue measure subset T of the set of parameters is such that, for $t \in T$, the dynamics of f_t admits a unique stochastically stable (under additive noise type random perturbations) physical measure μ_t whose basin has full measure in the ambient space (either the circle or the interval). Therefore, models involving one-dimensional unimodal maps typically are stochastically stable.

In many settings (e.g., low-dimensional dynamical systems), Lyapunov exponents can be given by time averages of continuous functions – for example, the time average of $\log \|DT_0\|$ gives the biggest exponent. In this case, stochastic stability directly implies stability of the Lyapunov exponents under small random perturbations of the dynamics.

Example 6 (Stochastically stable examples). Let $T_0 : \mathbb{S}^1 \circlearrowleft$ be a map such that λ, the Lebesgue (length) measure on the circle, is T_0-invariant and ergodic. Then λ is physical.

We consider the parametrized family $T_t : \mathbb{S}^1 \times \mathbb{S}^1 \to \mathbb{S}^1$, $(t, x) \mapsto x + t$ and a family of probability measures $\theta_\epsilon = (\lambda(-\epsilon, \epsilon))^{-1} \cdot (\lambda \mid (-\epsilon, \epsilon))$ given by the normalized restriction of λ to the ϵ-neighborhood of 0, where we regard $\mathbb{S}^1$ as the Lie group $\mathbb{R}/\mathbb{Z}$ and use additive notation for the group operation. Since λ is T_t-invariant for every $t \in \mathbb{S}^1$, λ is also an invariant measure for the measure-preserving random system

$$\mathcal{S} : (\mathbb{S}^1 \times \Omega^{\mathbb{N}}, \lambda \times \theta_\epsilon^{\mathbb{N}}) \circlearrowleft$$

for every $\epsilon > 0$, where $\Omega = (\mathbb{S}^1)^{\mathbb{N}}$. Hence, (T_0, λ) is stochastically stable under additive noise perturbations.

Concrete examples can be irrational rotations, $T_0(x) = x + \alpha$ with $\alpha \in \mathbb{R}\setminus\mathbb{Q}$, or expanding maps of the circle, $T_0(x) = b \cdot x$ for some $b \in \mathbb{N}$, $n \geq 2$. Analogous examples exist in higher-dimensional tori.

Example 7 (Stochastic stability depends on the type of noise). In spite of the straightforward method for obtaining stochastic stability in Example 6, for example, an expanding circle map $T_0(x) = 2 \cdot x$, we can choose a continuous family of probability measures θ_ϵ such that the same map T_0 is not stochastically stable.

It is well known that λ is the unique absolutely continuous invariant measure for T_0 and also the unique physical measure. Given $\epsilon > 0$ small, let us define transition probability measures as follows:

$$p_\epsilon(\cdot \mid z) = \frac{\lambda \mid [\phi_\epsilon(z) - \epsilon, \phi_\epsilon(z) + \epsilon]}{\lambda([\phi_\epsilon(z) - \epsilon, \phi_\epsilon(z) + \epsilon])}$$

where $\phi_\epsilon \mid (-\epsilon, \epsilon) \equiv 0$, $\phi_\epsilon \mid [\mathbb{S}^1 \setminus (-2\epsilon, 2\epsilon)] \equiv T_0$, and over $(-2\epsilon, -\epsilon] \cup [\epsilon, 2\epsilon)$, we can define ϕ_ϵ by interpolation in order that it be smooth.

In this setting, every random orbit starting at $(-\epsilon, \epsilon)$ never leaves this neighborhood in the future. Moreover, it is easy to see that every random orbit eventually enters $(-\epsilon, \epsilon)$. Hence, every invariant probability measure μ_ϵ for this Markov chain model is supported in $[-\epsilon, \epsilon]$. Thus, letting $\epsilon \to 0$, we see that the only zero-noise limit is δ_0, the Dirac mass concentrated at 0, which is not a physical measure for T_0.

This construction can be achieved in a random-maps setting, but only in the C^0 topology – it is not possible to realize this Markov chain by random maps that are C^1 close to T_0 for ϵ near 0.

Characterization of Measures Satisfying the Entropy Formula

Significant effort has been put in recent years in extending important results from dynamical systems to the random setting. Among many examples are: the local conjugacy between the dynamics near a hyperbolic fixed point and the action of the derivative of the map on the tangent space, the stable/unstable manifold theorems for hyperbolic invariant sets and the notions and properties of metric and topological entropy, dimensions and equilibrium states for potentials on random (or fuzzy) sets.

The characterization of measures satisfying the entropy formula is one important result whose extension to the setting of iteration of independent and identically distributed random maps has recently had interesting new consequences back into nonrandom dynamical systems.

Metric entropy for random perturbations Given a probability measure μ and a partition ξ of M, except perhaps for a subset of μ-null measure, the entropy of μ with respect to ξ is defined to be

$$H_\mu(\xi) = -\sum_{R\in\xi} \mu(R) \log \mu(R)$$

where the convention that $0 \log 0 = 0$ has been used. Given another finite partition ζ, we write $\xi \vee \zeta$ to indicate the partition obtained through intersection of every element of ξ with every element of ζ, and analogously for any finite number of partitions. If μ is also a stationary measure for a random-maps model (see the section "Random maps"), then for any finite measurable partition ξ of M,

$$h_\mu(\xi) = \inf_{n \geq 1} \frac{1}{n} \int H_\mu \left(\bigvee_{i=0}^{n-1} \left(T^i_{\underline{\omega}} \right)^{-1} (\xi) \right) \mathrm{d}p^{\mathbb{N}}(\underline{\omega})$$

is finite and is called the entropy of the random dynamical system with respect to ξ and to μ.

We define $h_\mu = \sup_\xi h_\mu(\xi)$ as the metric entropy of the random dynamical system, where the supremo is taken over all μ-measurable partitions. An important point here is the following notion: setting $\mathcal{A}$ the Borel σ-algebra of M, we say that a finite partition ξ of M is a random generating partition for $\mathcal{A}$ if

$$\bigvee_{i=0}^{+\infty} (T^i_{\underline{\omega}})^{-1} (\xi) = \mathcal{A}$$

(except μ-null sets) for $p^{\mathbb{N}}$-almost all $\omega \in \Omega = \mathcal{U}^{\mathbb{N}}$. Then a classical result from ergodic theory ensures that we can calculate the entropy using only a random generating partition ξ, that is, $h_\mu = h_\mu(\xi)$.

The entropy formula There exists a general relation ensuring that the entropy of a measure-preserving differentiable transformation (T_0, μ) on a compact Riemannian manifold is bounded from above by the sum of the positive Lyapunov exponents of T_0

$$h_\mu(T_0) \leq \int \sum_{\lambda_i(x) > 0} \lambda_i(x) \, \mathrm{d}\mu(x)$$

The equality (entropy formula) was first shown to hold for diffeomorphisms preserving a measure equivalent to the Riemannian volume, and then the measures satisfying the entropy formula were characterized: for C^2 diffeomorphisms the equality holds if and only if the disintegration of μ along the unstable manifolds is formed by measures absolutely continuous with respect to the Riemannian volume restricted to those submanifolds. The unstable manifolds are the submanifolds of M everywhere tangent to the Lyapunov subspaces corresponding to all positive Lyapunov exponents, analogous to "integrating the distribution of Lyapunov subspaces corresponding to positive exponents" – this particular point is a main subject of smooth ergodic theory for nonuniformly hyperbolic dynamics.

Both the inequality and the characterization of stationary measures satisfying the entropy formula were extended to random iterations of independent and identically distributed C^2 maps (noninjective and admitting critical points), and the inequality reads

$$h_\mu \leq \iint \sum_{\lambda_i(x,\omega) > 0} \lambda_i(x, \omega) \, \mathrm{d}\mu(x) \, \mathrm{d}p^{\mathbb{N}}(\omega)$$

where the functions λ_i are the random variables provided by the random multiplicative ergodic theorem.

Construction of Physical Measures as Zero-Noise Limits

The characterization of measures which satisfy the entropy formula enables us to construct physical measures as zero-noise limits of random invariant measures in some settings, outlined in the following, obtaining in the process that the physical measures so constructed are also stochastically stable.

The physical measures obtained in this manner arguably are natural measures for the system, since they are both stable under (certain types of) random perturbations and describe the asymptotic behavior of the system for a positive-volume subset of initial conditions. This is a significant contribution to the state-of-the-art of present knowledge on dynamics from the perspective of random dynamical systems.

Hyperbolic measures and the entropy formula The main idea is that an ergodic invariant measure μ for a diffeomorphism T_0 which satisfies the entropy formula and whose Lyapunov exponents are everywhere nonzero (known as hyperbolic measure) necessarily is a physical measure for T_0. This follows from standard arguments of smooth nonuniformly hyperbolic ergodic theory.

Indeed μ satisfies the entropy formula if and only if μ disintegrates into densities along the unstable submanifolds of T_0. The unstable manifolds $W^{\mathrm{u}}(x)$ are tangent to the subspace corresponding to every positive Lyapunov exponent at μ-almost every point x, they are an invariant family, that is, $T_0(W^{\mathrm{u}}(x)) = W^{\mathrm{u}}(x)$ for μ-almost every x, and distances on them are uniformly contracted under iteration by T_0^{-1}.

If the exponents along the complementary directions are nonzero, then they must be negative and smooth ergodic theory ensures that there exist stable manifolds, which are submanifolds $W^{\mathrm{s}}(x)$ of

M everywhere tangent to the subspace of negative Lyapunov exponents at μ-almost every point x, form a T_0-invariant family ($T_0(W^s(x)) = W^s(x)$, μ-almost everywhere), and distances on them are uniformly contracted under iteration by T_0.

We still need to understand that time averages are constant along both stable and unstable manifolds, and that the families of stable and unstable manifolds are absolutely continuous, in order to realize how a hyperbolic measure is a physical measure.

Given $y \in W^s(x)$, the time averages of x and y coincide for continuous observables simply because dist $(T_0^n(x), T_0^n(y)) \to 0$ when $n \to +\infty$. For unstable manifolds, the same holds when considering time averages for T_0^{-1}. Since forward and backward time averages are equal μ-almost everywhere, the set of points having asymptotic time averages given by μ has positive Lebesgue measure if the set

$$B = \bigcup \{ W^s(y) \colon y \in W^u(x) \cap \operatorname{supp}(\mu) \}$$

has positive volume in M, for some x whose time averages are well defined.

Now, stable and unstable manifolds are transverse everywhere where they are defined, but they are only defined μ-almost everywhere and depend measurably on the base point, so we cannot use transversality arguments from differential topology, in spite of $W^u(x) \cap \operatorname{supp}(\mu)$ having positive volume in $W^u(x)$ by the existence of a smooth disintegration of μ along the unstable manifolds. However, it is known for smooth (C^2) transformations that the families of stable and unstable manifolds are absolutely continuous, meaning that projections along leaves preserve sets of zero volume. This is precisely what is needed for measure-theoretic arguments to show that B has positive volume.

Zero-noise limits satisfying the entropy formula Using the extension of the characterization of measures satisfying the entropy formula for the random-maps setting, we can build random dynamical systems, which are small random perturbations of a map T_0, having invariant measures μ_ϵ satisfying the entropy formula for all sufficiently small $\epsilon > 0$. Indeed, it is enough to construct small random perturbations of T_0 having absolutely continuous invariant probability measures μ_ϵ for all small enough $\epsilon > 0$.

In order to obtain such random dynamical systems, we choose families of maps $T : \mathcal{U} \times M \to M$ and of probability measures $(\theta_\epsilon)_{\epsilon>0}$ as in Examples 3 and 4, where we assume that $o \in \mathcal{U}$, so that T_0 belongs to the family. Letting $T_x(u) = T(u, x)$ for all $(u, x) \in \mathcal{U} \times M$, we then have that $T_x(\theta_\epsilon)$ is absolutely continuous. This means that sets of perturbations of positive θ_ϵ-measure send points of M onto positive-volume subsets of M. Such a perturbation can be constructed for every continuous map of any manifold.

In this setting, any invariant probability measure for the associated skew-product map $\mathcal{S} : \Omega \times M \circlearrowleft$ of the form $\theta_\epsilon^{\mathbb{N}} \times \mu_\epsilon$ is such that μ_ϵ is absolutely continuous with respect to volume on M. Then the entropy formula holds:

$$h_{\mu_\epsilon} = \iint \sum_{\lambda_i(x,\omega)>0} \lambda_i(x, \omega) \, \mathrm{d}\mu_\epsilon(x) \, \mathrm{d}\theta_\epsilon^{\mathbb{N}}(\omega)$$

Having this and knowing the characterization of measures satisfying the entropy formula, it is natural to look for conditions under which we can guarantee that the above inequality extends to any zero-noise limit μ_0 of μ_ϵ when $\epsilon \to 0$. In this case, μ_0 satisfies the entropy formula for T_0.

If, in addition, we are able to show that μ_0 is a hyperbolic measure, then we obtain a physical measure for T_0 which is stochastically stable by construction.

These ideas can be carried out completely for hyperbolic diffeomorphisms, that is, maps admitting a continuous invariant splitting of the tangent space into two sub-bundles $E \oplus F$ defined everywhere with bounded angles, whose Lyapunov exponents are negative along E and positive along F. Recently, maps satisfying weaker conditions were shown to admit stochastically stable physical measures following the same ideas.

These ideas also have applications to the construction and stochastic stability of physical measure for strange attractors and for all mathematical models involving ordinary differential equations or iterations of maps.

See also: Dynamical Systems in Mathematical Physics: An Illustration from Water Waves; Homeomorphisms and Diffeomorphisms of the Circle; Lyapunov Exponents and Strange Attractors; Nonequilibrium Statistical Mechanics (Stationary): Overview; Random Walks in Random Environments; Stochastic Differential Equations.

Further Reading

Arnold L (1998) *Random Dynamical Systems*. Berlin: Springer.

Billingsley P (1965) *Ergodic Theory and Information*. New York: Wiley.

Billingsley P (1985) *Probability and Measure*, 3rd edn. New York: Wiley.

Bonatti C, Díaz L, and Viana M (2004) *Dynamics Beyond Hyperbolicity: A Global Geometric and Probabilistic Perspective*. Berlin: Springer.

Bonatti C, Díaz L, and Viana M (2005) *Dynamics Beyond Uniform Hyperbolicity. A Global Geometric and Probabilistic Perspective*, Encyclopaedia of Mathematical Sciences, 102 Mathematical Physics III. Berlin: Springer.
Doob J (1953) *Stochastic Processes*. New York: Wiley.
Fathi A, Herman M-R, and Yoccoz J-C (1983) A proof of Pesin's stable manifold theorem. In: Palis J (ed.) *Geometric Dynamics (Rio de Janeiro, 1981)*, Lecture Notes in Mathematics, vol. 1007, 177–215. Berlin: Springer.
Kifer Y (1986) *Ergodic Theory of Random Perturbations*. Boston: Birkhäuser.
Kifer Y (1988) *Random Perturbations of Dynamical Systems*. Boston: Birkhäuser.
Kunita H (1990) *Stochastic Flows and Stochastic Differential Equations*. Cambridge: Cambridge University Press.
Ledrappier F and Young L-S (1998) Entropy formula for random transformations. *Probability Theory and Related Fields* 80(2): 217–240.
Liu P-D and Qian M (1995) *Smooth Ergodic Theory of Random Dynamical Systems*, Lecture Notes in Mathematics, vol. 1606. Berlin: Springer.
Øskendal B (1992) *Stochastic Differential Equations. Universitext*, 3rd edn. Berlin: Springer.
Viana M (2000) What's new on Lorenz strange attractor. *Mathematical Intelligencer* 22(3): 6–19.
Walters P (1982) *An Introduction to Ergodic Theory*. Berlin: Springer.

Random Matrix Theory in Physics

T Guhr, Lunds Universitet, Lund, Sweden

Introduction

We wish to study energy correlations of quantum spectra. Suppose the spectrum of a quantum system has been measured or calculated. All levels in the total spectrum having the same quantum numbers form one particular *subspectrum*. Its energy levels are at positions x_n, $n = 1, 2, \ldots, N$, say. We assume that N, the number of levels in this subspectrum, is large. With a proper *smoothing* procedure, we obtain the level density $R_1(x)$, that is, the probability density of finding a level at the energy x. As indicated in the top part of Figure 1, the level density $R_1(x)$ increases with x for most physics systems. In the present context, however, we are not so interested in the level density. We want to measure the spectral correlations independently of it. Hence, we have to remove the level density from the subspectrum. This is referred to as *unfolding*. We introduce a new dimensionless energy scale ξ such that $\mathrm{d}\xi = R_1(x)\,\mathrm{d}x$. By construction, the resulting subspectrum in ξ has level density unity, as shown schematically in the bottom part of Figure 1. It is always understood that the energy correlations are analyzed in the *unfolded subspectra*.

Surprisingly, a remarkable *universality* is found in the spectral correlations of a large class of systems, including nuclei, atoms, molecules, quantum chaotic and disordered systems, and even quantum chromodynamics on the lattice. Consider the *nearest-neighbor spacing distribution* $p(s)$. It is the probability density of finding two adjacent levels in the distance s. If the positions of the levels are uncorrelated, the nearest-neighbor spacing distribution can be shown to follow the *Poisson law*

$$p^{(\mathrm{P})}(s) = \exp(-s) \qquad [1]$$

While this is occasionally found, many more systems show a rather different nearest-neighbor spacing distribution, the *Wigner surmise*

$$p^{(\mathrm{W})}(s) = \frac{\pi}{2} s \exp\left(-\frac{\pi}{4} s^2\right) \qquad [2]$$

As shown in Figure 2, the Wigner surmise excludes degeneracies, $p^{(\mathrm{W})}(0) = 0$, the levels repel each other. This is only possible if they are correlated. Thus, the Poisson law and the Wigner surmise reflect the absence or the presence of energy correlations, respectively.

Now, the question arises: if these correlation patterns are so frequently found in physics, is there some simple, phenomenological model? – Yes, random matrix theory (RMT) is precisely this. To describe the absence of correlations, we choose, in view of what has been said above, a diagonal Hamiltonian

$$H = \mathrm{diag}(x_1, \ldots, x_N) \qquad [3]$$

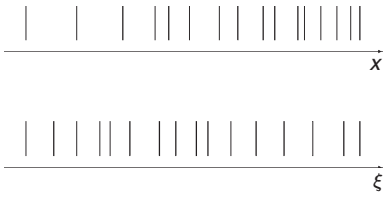


Figure 1 Original (top) and unfolded (bottom) spectrum.

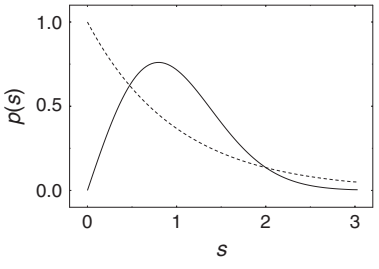


Figure 2 Wigner surmise (solid) and Poisson law (dashed).

whose elements, the eigenvalues x_n, are uncorrelated random numbers. To model the presence of correlations, we insert off-diagonal matrix elements,

$$H = \begin{bmatrix} H_{11} & \cdots & H_{1N} \\ \vdots & & \vdots \\ H_{N1} & \cdots & H_{NN} \end{bmatrix} \qquad [4]$$

We require that H is real symmetric, $H^{\mathrm{T}} = H$. The independent elements H_{nm} are random numbers. The random matrix H is diagonalized to obtain the energy levels $x_n, n = 1, 2, \ldots, N$. Indeed, a numerical simulation shows that these two models yield, after unfolding, the Poisson law and the Wigner surmise for large N, that is, the absence or presence of correlations. This is the most important insight into the phenomenology of RMT.

In this article, we set up RMT in a more formal way; we discuss analytical calculations of correlation functions, demonstrate how this relates to supersymmetry and stochastic field theory and show the connection to chaos, and we briefly sketch the numerous applications in many-body physics, in disordered and mesoscopic systems, in models for interacting fermions, and in quantum chromodynamics. We also mention applications in other fields, even beyond physics.

Random Matrix Theory

Classical Gaussian Ensembles

For now, we consider a system whose energy levels are correlated. The $N \times N$ matrix H modeling it has no fixed zeros but random entries everywhere. There are three possible *symmetry classes* of random matrices in standard Schrödinger quantum mechanics. They are labeled by the Dyson index β. If the system is not time-reversal invariant, H has to be Hermitian and the random entries H_{nm} are complex ($\beta = 2$). If time-reversal invariance holds, two possibilities must be distinguished: if either the system is rotational symmetric, or it has integer spin and rotational symmetry is broken, the Hamilton matrix H can be chosen to be real symmetric ($\beta = 1$). This is the case in eqn [4]. If, on the other hand, the system has half-integer spin and rotational symmetry is broken, H is self-dual ($\beta = 4$) and the random entries H_{nm} are 2×2 quaternionic. The Dyson index β is the dimension of the number field over which H is constructed.

As we are interested in the eigenvalue correlations, we diagonalize the random matrix, $H = U^{-1} x U$. Here, $x = \mathrm{diag}(x_1, \ldots, x_N)$ is the diagonal matrix of the N eigenvalues. For $\beta = 4$, every eigenvalue is doubly degenerate. This is *Kramers' degeneracy*. The diagonalizing matrix U is in the orthogonal group O(N) for $\beta = 1$, in the unitary group U(N) for $\beta = 2$ and in the unitary–symplectic group USp($2N$) for $\beta = 4$. Accordingly, the three symmetry classes are referred to as *orthogonal, unitary*, and *symplectic*.

We have not yet chosen the probability densities for the random entries H_{nm}. To keep our assumptions about the system at a minimum, we treat all entries on equal footing. This is achieved by *rotational invariance* of the probability density $P_N^{(\beta)}(H)$, not to be confused with the rotational symmetry employed above to define the symmetry classes. No basis for the matrices is preferred in any way if we construct $P_N^{(\beta)}(H)$ from matrix invariants, that is, from traces and determinants, such that it depends only on the eigenvalues, $P_N^{(\beta)}(H) = P_N^{(\beta)}(x)$. A particularly convenient choice is the Gaussian

$$P_N^{(\beta)}(H) = C_N^{(\beta)} \exp\left(-\frac{\beta}{4v^2} \mathrm{tr}\, H^2\right) \qquad [5]$$

where the constant v sets the energy scale and the constant $C_N^{(\beta)}$ ensures normalization. The three symmetry classes together with the probability densities [5] define the *Gaussian ensembles*: the Gaussian orthogonal (GOE), unitary (GUE) and symplectic (GSE) ensemble for $\beta = 1, 2, 4$.

The phenomenology of the three Gaussian ensembles differs considerably. The higher β, the stronger the level repulsion between the eigenvalues x_n. Numerical simulation quickly shows that the nearest-neighbor spacing distribution behaves like $p^{(\beta)}(s) \sim s^{\beta}$ for small spacings s. This also becomes obvious by working out the differential probability $P_N^{(\beta)}(H)\mathrm{d}[H]$ of the random matrices H in eigenvalue–angle coordinates x and U. Here, $\mathrm{d}[H]$ is the invariant measure or volume element in the matrix space. When writing $\mathrm{d}[\cdot]$, we always mean the product of all differentials of independent variables for the quantity in the square brackets. Up to constants, we have

$$\mathrm{d}[H] = |\Delta_N(x)|^{\beta}\, \mathrm{d}[x]\, \mathrm{d}\mu(U) \qquad [6]$$

where $\mathrm{d}\mu(U)$ is, apart from certain phase contributions, the invariant or Haar measure on O(N), U(N), or USp($2N$), respectively. The Jacobian of the transformation is the modulus of the Vandermonde determinant

$$\Delta_N(x) = \prod_{n<m} (x_n - x_m) \qquad [7]$$

raised to the power β. Thus, the differential probability $P_N^{(\beta)}(H)\,\mathrm{d}[H]$ vanishes whenever any two eigenvalues x_n degenerate. This is the level

repulsion. It immediately explains the behavior of the nearest-neighbor spacing distribution for small spacings.

Additional symmetry constraints lead to new random matrix ensembles relevant in physics, the *Andreev* and the *chiral Gaussian ensembles*. If one refers to the classical Gaussian ensembles, one usually means the three ensembles introduced above.

Correlation Functions

The probability density to find k energy levels at positions $x_1, \ldots, x_k$ is the k-level correlation function $R_k^{(\beta)}(x_1, \ldots, x_k)$. We find it by integrating out $N-k$ levels in the N-level differential probability $P_N^{(\beta)}(H)\,\mathrm{d}[H]$. We also have to average over the bases, that is, over the diagonalizing matrices U. Due to rotational invariance, this simply yields the group volume. Thus, we have

$$R_k^{(\beta)}(x_1, \ldots, x_k) = \frac{N!}{(N-k)!}\int_{-\infty}^{+\infty} \mathrm{d}x_{k+1} \cdots \int_{-\infty}^{+\infty} \mathrm{d}x_N |\Delta_N(x)|^\beta P_N^{(\beta)}(x) \quad [8]$$

Once more, we used rotational invariance which implies that $P_N^{(\beta)}(x)$ is invariant under permutation of the levels x_n. Since the same then also holds for the correlation functions [8], it is convenient to normalize them to the combinatorial factor in front of the integrals. A constant ensuring this has been absorbed into $P_N^{(\beta)}(x)$.

Remarkably, the integrals in eqn [8] can be done in closed form. The GUE case ($\beta = 2$) is mathematically the simplest, and one finds the determinant structure

$$R_k^{(2)}(x_1, \ldots, x_k) = \det[K_N^{(2)}(x_p, x_q)]_{p,q=1,\ldots,k} \quad [9]$$

All entries of the determinant can be expressed in terms of the kernel $K_N^{(2)}(x_p, x_q)$, which depends on two energy arguments (x_p, x_q). Analogous but more complicated formulae are valid for the GOE ($\beta = 1$) and the GSE ($\beta = 4$), involving quaternion determinants and integrals and derivatives of the kernel.

As argued in the Introduction, we are interested in the energy correlations on the unfolded energy scale. The level density is formally the one-level correlation function. For the three Gaussian ensembles it is, to leading order in the level number N, the *Wigner semicircle*

$$R_1^{(\beta)}(x_1) = \frac{1}{2\pi v^2}\sqrt{4Nv^2 - x_1^2} \quad [10]$$

for $|x_1| \le 2\sqrt{N}v$ and zero for $|x_1| > 2\sqrt{N}v$. None of the common systems in physics has such a level density. When unfolding, we also want to take the limit of infinitely many levels $N \to \infty$ to remove cutoff effects due to the finite dimension of the random matrices. It suffices to stay in the center of the semicircle where the mean level spacing is $D = 1/R_1^{(\beta)}(0) = \pi v/\sqrt{N}$. We introduce the dimensionless energies $\xi_p = x_p/D, p = 1, \ldots, k$, which have to be held fixed when taking the limit $N \to \infty$. The *unfolded correlation functions* are given by

$$X_k^{(\beta)}(\xi_1, \ldots, \xi_k) = \lim_{N\to\infty} D^k R_k^{(\beta)}(D\xi_1, \ldots, D\xi_k) \quad [11]$$

As we are dealing with probability densities, the Jacobians $\mathrm{d}x_p/\mathrm{d}\xi_p$ enter the reformulation in the new energy variables. This explains the factor D^k. Unfolding makes the correlation functions *translation invariant*; they depend only on the differences $\xi_p - \xi_q$. The unfolded correlation functions can be written in a rather compact form. For the GUE ($\beta = 2$), they read

$$X_k^{(2)}(\xi_1, \ldots, \xi_k) = \det\left[\frac{\sin\pi(\xi_p - \xi_q)}{\pi(\xi_p - \xi_q)}\right]_{p,q=1,\ldots,k} \quad [12]$$

There are similar, but more complicated, formulae for the GOE ($\beta = 1$) and the GSE ($\beta = 4$). By construction, one has $X_1^{(\beta)}(\xi_1) = 1$.

It is useful to formulate the case where correlations are absent, that is, the *Poisson case*, accordingly. The level density $R_1^{(\mathrm{P})}(x_1)$ is simply N times the (smooth) probability density chosen for the entries in the diagonal matrix [4]. Lack of correlations means that the k-level correlation function only involves one-level correlations,

$$R_k^{(\mathrm{P})}(x_1, \ldots, x_k) = \frac{N!}{(N-k)!N^k}\prod_{p=1}^{k} R_1^{(\mathrm{P})}(x_p) \quad [13]$$

The combinatorial factor is important, since we always normalize to $N!/(N-k)!$. Hence, one finds

$$X_k^{(\mathrm{P})}(\xi_1, \ldots, \xi_k) = 1 \quad [14]$$

for all unfolded correlation functions.

Statistical Observables

The unfolded correlation functions yield all statistical observables. The two-level correlation function $X_2(r)$ with $r = \xi_1 - \xi_2$ is of particular interest in applications. If we do not write the superscript (β) or (P), we mean either of the functions. For the Gaussian ensembles, $X_2^{(\beta)}(r)$ is shown in **Figure 3**. One often writes $X_2(r) = 1 - Y_2(r)$. The two-level *cluster function* $Y_2(r)$ nicely measures the deviation from the uncorrelated Poisson case, where one has $X_2^{(\mathrm{P})}(r) = 1$ and $Y_2^{(\mathrm{P})}(r) = 0$.

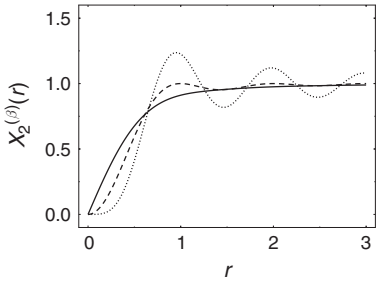


Figure 3 Two-level correlation function $X_2^{(\beta)}(r)$ for GOE (solid), GUE (dashed) and GSE (dotted).

By construction, the average level number in an interval of length L in the unfolded spectrum is L. The level *number variance* $\Sigma^2(L)$ is shown to be an average over the two-level cluster function,

$$\Sigma^2(L) = L - 2\int_0^L (L-r)Y_2(r)\mathrm{d}r \qquad [15]$$

We find $L \pm \sqrt{\Sigma^2(L)}$ levels in an interval of length L. In the uncorrelated Poisson case, one has $\Sigma^{2(\mathrm{P})}(L) = L$. This is just *Poisson's error law*. For the Gaussian ensembles $\Sigma^{2(\beta)}(L)$ behaves logarithmically for large L. The spectrum is said to be more *rigid* than in the Poisson case. As Figure 4 shows, the level number variance probes longer distances in the spectrum, in contrast to the nearest-neighbor spacing distribution.

Many more observables, also sensitive to higher order, $k > 2$ correlations, have been defined. In practice, however, one is often restricted to analyzing two-level correlations. An exception is, to some extent, the nearest-neighbor spacing distribution $p(s)$. It is the two-level correlation function with the *additional* requirement that the two levels in question are adjacent, that is, that there are no levels between them. Thus, *all* correlation functions are needed if one wishes to calculate the exact nearest-neighbor spacing distribution $p^{(\beta)}(s)$ for the Gaussian ensembles. These considerations explain that we have $X_2^{(\beta)}(s) \simeq p^{(\beta)}(s)$ for small s. But while $X_2^{(\beta)}(s)$ saturates for large s, $p^{(\beta)}(s)$ quickly goes to zero in a Gaussian fashion. Thus, although the nearest-neighbor spacing distribution mathematically involves all correlations, it makes in practice only a meaningful statement about the two-level correlations. Luckily, $p^{(\beta)}(s)$ differs only very slightly from the heuristic Wigner surmise [2] (corresponding to $\beta = 1$), respectively from its extensions (corresponding to $\beta = 2$ and $\beta = 4$).

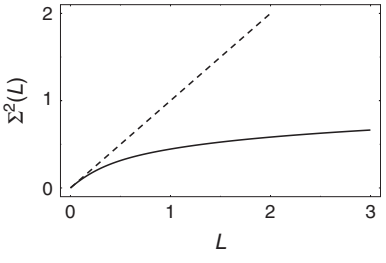


Figure 4 Level number variance $\Sigma^2(L)$ for GOE (solid) and Poisson case (dashed).

Ergodicity and Universality

We constructed the correlation functions as averages over an ensemble of random matrices. But this is not how we proceeded in the data analysis sketched in the Introduction. There, we started from *one single* spectrum with very many levels and obtained the statistical observable just by sampling and, if necessary, smoothing. Do these two averages, the *ensemble average* and the *spectral average*, yield the same? Indeed, one can show that the answer is affirmative, if the level number N goes to infinity. This is referred to as *ergodicity* in RMT.

Moreover, as already briefly indicated in the Introduction, very many systems from different areas of physics are well described by RMT. This seems to be at odds with the Gaussian assumption [5]. There is hardly any system whose Hamilton matrix elements follow a Gaussian probability density. The solution for this puzzle lies in the unfolding. Indeed, it has been shown that almost all functional forms of the probability density $P_N^{(\beta)}(H)$ yield the same unfolded correlation functions, if no new scale comparable to the mean level spacing is present in $P_N^{(\beta)}(H)$. This is the mathematical side of the empirically found *universality*.

Ergodicity and universality are of crucial importance for the applicability of RMT in data analysis.

Wave Functions

By modeling the Hamiltonian of a system with a random matrix H, we do not only make an assumption about the statistics of the energies, but also about those of the wave functions. Because of the eigenvalue equation $Hu_n = x_n u_n, n = 1, \ldots, N$, the wave function belonging to the eigenenergy x_n is modeled by the eigenvector u_n. The columns of the diagonalizing matrix $U = [u_1\, u_2 \cdots u_N]$ are these eigenvectors. The probability density of the components u_{nm} of the eigenvector u_n can be calculated rather easily. For large N it approaches a Gaussian. This is equivalent to the *Porter–Thomas distribution*. While wave functions are often not accessible in an experiment, one can measure transition amplitudes and widths, giving information about the matrix elements of a transition operator and a

projection of the wave functions onto a certain state in Hilbert space. If the latter are represented by a fixed matrix A or a fixed vector a, respectively, one can calculate the RMT prediction for the probability densities of the matrix elements $u_n^\dagger A u_m$ or the widths $a^\dagger u_n$ from the probability density of the eigenvectors.

Scattering Systems

It is important that RMT can be used as a powerful tool in scattering theory, because the major part of the experimental information about quantum systems comes from scattering experiments. Consider an example from compound nucleus scattering. In an accelerator, a proton is shot on a nucleus, with which it forms a compound nucleus. This then decays by emitting a neutron. More generally, the ingoing *channel* ν (the proton in our example) connects to the *interaction region* (the nucleus), which also connects to an outgoing channel μ (the neutron). There are Λ channels with channel wave functions which are labeled $\nu = 1, \ldots, \Lambda$. The interaction region is described by an $N \times N$ Hamiltonian matrix H whose eigenvalues x_n are bound-state energies labeled $n = 1, \ldots, N$. The dimension N is a cutoff which has to be taken to infinity at the end of a calculation. The $\Lambda \times \Lambda$ *scattering matrix* S contains the information about how the ingoing channels are transformed into the outgoing channels. The scattering matrix S is unitary. Under certain and often justified assumptions, a scattering matrix element can be cast into the form

$$S_{\nu\mu} = \delta_{\nu\mu} - \mathrm{i}2\pi W_\nu^\dagger G^{-1} W_\mu \qquad [16]$$

The couplings $W_{n\nu}$ between the bound states n and the channels ν are collected in the $N \times \Lambda$ matrix W, W_ν is its νth column. The propagator G^{-1} is the inverse of

$$G = z1_N - H + \mathrm{i}\pi \sum_{\nu \text{ open}} W_\nu W_\nu^\dagger \qquad [17]$$

Here, z is the scattering energy and the summation is only over channels which are open, that is, accessible. Formula [16] has a clear intuitive interpretation. The scattering region is entered through channel ν, the bound states of H become *resonances* in the scattering process according to eqn [17], the interaction region is left through channel μ. This formulation applies in many areas of physics. All observables such as transmission coefficients, cross sections, and others can be calculated from the scattering matrix S.

We have not made any statistical assumptions yet. Often, one can understand generic features of a scattering system by assuming that the Hamiltonian H is a random matrix, taken from one of the three classical ensembles. This is one RMT approach used in scattering theory.

Another RMT approach is based on the scattering matrix itself, S is modeled by a $\Lambda \times \Lambda$ unitary random matrix. Taking into account additional symmetries, one arrives at the *three circular ensembles*, circular orthogonal (COE), unitary (CUE) and symplectic (CSE). They correspond to the three classical Gaussian ensembles and are also labeled with the Dyson index $\beta = 1, 2, 4$. The eigenphases of the random scattering matrix correspond to the eigenvalues of the random Hamiltonian matrix. The unfolded correlation functions of the circular ensembles are identical to those of the Gaussian ensembles.

Supersymmetry

Apart from the symmetries, random matrices contain nothing but random numbers. Thus, a certain type of redundancy is present in RMT. Remarkably, this redundancy can be removed, without losing any piece of information by using *supersymmetry*, that is, by a reformulation of the random matrix model involving commuting and anticommuting variables. For the sake of simplicity, we sketch the main ideas for the GUE, but they apply to the GOE and the GSE accordingly.

One defines the k-level correlation functions by using the resolvent of the Schrödinger equation,

$$\widehat{R}_k^{(2)}(x_1, \ldots, x_k) = \frac{1}{\pi^k} \int P_N^{(2)}(H) \prod_{p=1}^{k} \mathrm{tr} \frac{1}{x_p^\pm - H} \mathrm{d}[H] \qquad [18]$$

The energies carry an imaginary increment $x_p^\pm = x_p \pm \mathrm{i}\varepsilon$ and the limit $\varepsilon \to 0$ has to be taken at the end of the calculation. The k-level correlation functions $R_k^{(2)}(x_1, \ldots, x_k)$ as defined in eqn [8] can always be obtained from the functions [18] by constructing a linear combination of the $\widehat{R}_k^{(2)}(x_1, \ldots, x_k)$ in which the signs of the imaginary increments are chosen such that only the imaginary parts of the traces contribute. Some trivial δ-distributions have to be removed. The k-level correlation functions [18] can be written as the k-fold derivative

$$\widehat{R}_k^{(2)}(x_1, \ldots, x_k) = \frac{1}{(2\pi)^k} \frac{\partial^k}{\prod_{p=1}^{k} \partial J_p} Z_k^{(2)}(x + J) \Big|_{J=0} \qquad [19]$$

of the generating function

$$Z_k^{(2)}(x+J) = \int P_N^{(2)}(H) \prod_{p=1}^{k} \frac{\det\left(x_p^{\pm} + J_p - H\right)}{\det\left(x_p^{\pm} - J_p - H\right)} \mathrm{d}[H] \quad [20]$$

which depends on the energies and k new source variables $J_p, p = 1, \ldots, k$, ordered in $2k \times 2k$ diagonal matrices

$$x = \mathrm{diag}(x_1, x_1, \ldots, x_k, x_k)$$
$$J = \mathrm{diag}(+J_1, -J_1, \ldots, +J_k, -J_k) \quad [21]$$

We notice the normalization $Z_k^{(2)}(x) = 1$ at $J = 0$. The generating function [20] is an integral over an ordinary $N \times N$ matrix H. It can be exactly rewritten as an integral over a $2k \times 2k$ *supermatrix* σ containing commuting and anticommuting variables,

$$Z_k^{(2)}(x+J) = \int Q_k^{(2)}(\sigma) \mathrm{sdet}^{-N}(x^{\pm} + J - \sigma) \mathrm{d}[\sigma] \quad [22]$$

The integrals over the commuting variables are of the ordinary *Riemann–Stiltjes* type, while those over the anticommuting variables are *Berezin* integrals. The Gaussian probability density [5] is mapped onto its counterpart in superspace

$$Q_k^{(2)}(\sigma) = c_k^{(2)} \exp\left(-\frac{1}{2v^2} \mathrm{str}\, \sigma^2\right) \quad [23]$$

where $c_k^{(2)}$ is a normalization constant. The supertrace str and the superdeterminant sdet generalize the corresponding invariants for ordinary matrices. The total number of integrations in eqn [22] is drastically reduced as compared to eqn [20]. Importantly, it is independent of the level number N which now only appears as the negative power of the superdeterminant in eqn [22], that is, as an explicit parameter. This most convenient feature makes it possible to take the limit of infinitely many levels by means of a saddle point approximation to the generating function.

Loosely speaking, the supersymmetric formulation can be viewed as an irreducible representation of RMT which yields a clearer insight into the mathematical structures. The same is true for applications in scattering theory and in models for crossover transitions to be discussed below. This explains why supersymmetry is so often used in RMT calculations.

It should be emphasized that the rôle of supersymmetry in RMT is quite different from the one in high-energy physics, where the commuting and anticommuting variables represent physical particles, bosons and fermions, respectively. This is not so in the RMT context. The commuting and anticommuting variables have no direct physics interpretation; they appear simply as helpful mathematical devices to cast the RMT model into an often much more convenient form.

Crossover Transitions

The RMT models discussed up to now describe four extreme situations, the absence of correlations in the Poisson case and the presence of correlations as in the three fully rotational invariant models GOE, GUE, and GSE. A real physics system, however, is often between these extreme situations. The corresponding RMT models can vary considerably, depending on the specific situation. Nevertheless, those models in which the random matrices for two extreme situations are simply added with some weight are useful in so many applications that they acquired a rather generic standing. One writes

$$H(\alpha) = H^{(0)} + \alpha H^{(\beta)} \quad [24]$$

where $H^{(0)}$ is a random matrix drawn from an ensemble with a completely arbitrary probability density $P_N^{(0)}(H^{(0)})$. The case of a fixed matrix is included, because one may choose a product of δ-distributions for the probability density. The matrix $H^{(\beta)}$ is random and drawn from the classical Gaussian ensembles with probability density $P_N^{(\beta)}(H^{(\beta)})$ for $\beta = 1, 2, 4$. One requires that the group diagonalizing $H^{(0)}$ is a subgroup of the one diagonalizing $H^{(\beta)}$. The model [24] describes a *crossover transition*. The weight α is referred to as *transition parameter*. It is useful to choose the spectral support of $H^{(0)}$ and $H^{(\beta)}$ equal. One can then view α as the root-mean-square matrix element of $H^{(\beta)}$. At $\alpha = 0$, one has the arbitrary ensemble. The Gaussian ensembles are formally recovered in the limit $\alpha \to \infty$, to be taken in a proper way such that the energies remain finite.

We are always interested in the unfolded correlation functions. Thus, α has to be measured in units of the mean level spacing D such that $\lambda = \alpha/D$ is the physically relevant transition parameter. It means that, depending on the numerical value of D, even a small effect on the original energy scale can have sizeable impact on the spectral statistics. This is referred to as *statistical enhancement*. The nearest-neighbor spacing distribution is already very close to $p^{(\beta)}(s)$ for the Gaussian ensembles if λ is larger than 0.5 or so. In the long-range observables such as the level number variance $\Sigma^2(L)$, the deviation from the Gaussian ensemble statistics becomes visible at interval lengths L comparable to λ.

Crossover transitions can be interpreted as *diffusion processes*. With the fictitious time $t = \alpha^2/2$, the probability density $P_N(x,t)$ of the eigenvalues x of the total Hamilton matrix $H = H(t) = H(\alpha)$ satisfies the diffusion equation

$$\Delta_x P_N(x,t) = \frac{4}{\beta}\frac{\partial}{\partial t} P_N(x,t) \qquad [25]$$

where the probability density for the arbitrary ensemble is the initial condition $P_N(x,0) = P_N^{(0)}(x)$. The Laplacian

$$\Delta_x = \sum_{n=1}^{N} \frac{\partial^2}{\partial x_n^2} + \sum_{n<m} \frac{\beta}{x_n - x_m}\left(\frac{\partial}{\partial x_n} - \frac{\partial}{\partial x_m}\right) \qquad [26]$$

lives in the curved space of the eigenvalues x. This diffusion process is *Dyson's Brownian motion* in slightly simplified form. It has a rather general meaning for *harmonic analysis on symmetric spaces*, connecting to the spherical functions of Gelfand and Harish-Chandra, Itzykson–Zuber integrals, and to *Calogero–Sutherland models of interacting particles*. All this generalizes to superspace. In the supersymmetric version of Dyson's Brownian motion the generating function of the correlation functions is propagated,

$$\Delta_s Z_k(s,t) = \frac{4}{\beta}\frac{\partial}{\partial t} Z_k(s,t) \qquad [27]$$

where the initial condition $Z_k(s,0) = Z_k^{(0)}(s)$ is the generating function of the correlation functions for the arbitrary ensemble. Here, s denotes the eigenvalues of some supermatrices, not to be confused with the spacing between adjacent levels. Since the Laplacian Δ_s lives in this curved eigenvalue space, this diffusion process establishes an intimate connection to harmonic analysis on superspaces. Advantageously, the diffusion [27] is the same on the original and on the unfolded energy scales.

Fields of Application

Many-Body Systems

Numerous studies apply RMT to nuclear physics which is also the field of its origin. If the total number of nucleons, that is, protons and neutrons, is not too small, nuclei show *single-particle* and *collective* motion. Roughly speaking, the former is decoherent out-of-phase motion of the nucleons confined in the nucleus, while the latter is coherent in-phase motion of all nucleons or of large groups of them such that any additional individual motion of the nucleons becomes largely irrelevant. It has been shown empirically that the single-particle excitations lead to GOE statistics, while collective excitations produce different statistics, often of the Poisson type. Mixed statistics as described by crossover transitions are then of particular interest to investigate the character of excitations. For example, one applies the model [24] with $H^{(0)}$ drawn from a Poisson ensemble and $H^{(\beta)}$ from a GOE. Another application of crossover transitions is breaking of *time-reversal invariance* in nuclei. Here, $H^{(0)}$ is from a GOE and $H^{(\beta)}$ from a GUE. Indeed, a fit of spectral data to this model yields an upper bound for the *time-reversal invariance* violating root-mean-square matrix element in nuclei. Yet another application is breaking of symmetries such as parity or isospin. In the case of two quantum numbers, positive and negative parity, say, one chooses $H^{(0)} = \mathrm{diag}(H^{(+)}, H^{(-)})$ block-diagonal with $H^{(+)}$ and $H^{(-)}$ drawn from two uncorrelated GOE and $H^{(\beta)}$ from a third uncorrelated GOE which breaks the block structure. Again, root-mean-square matrix elements for symmetry breaking have been derived from the data.

Nuclear excitation spectra are extracted from scattering experiments. An analysis as described above is only possible if the resonances are isolated. Often, this is not the case and the resonance widths are comparable to or even much larger than the mean level spacing, making it impossible to obtain the excitation energies directly from the cross sections. One then analyzes the latter and their fluctuations as measured and applies the concepts sketched above for scattering systems. This approach has also been successful for crossover transitions.

Due to the complexity of the nuclear many-body problem, one has to use effective or phenomenological interactions when calculating spectra. Hence, one often studies whether the statistical features found in the experimental data are also present in the calculated spectra which result from the various models for nuclei.

Other many-body systems, such as complex atoms and molecules, have also been studied with RMT concepts, but the main focus has always been on nuclei.

Quantum Chaos

Originally, RMT was intended for modeling systems with many degrees of freedom such as nuclei. Surprisingly, RMT proved useful for systems with few degrees of freedom as well. Most of these studies aim at establishing a link between RMT and *classical chaos*. Consider as an example the classical motion of a point-like particle in a rectangle billiard. Ideal reflection at the boundaries and absence of friction are assumed, implying that the particle is reflected infinitely many times. A second billiard is built by taking a rectangle and replacing one corner with a quarter circle as shown in Figure 5. The motion of the particle in this *Sinai*

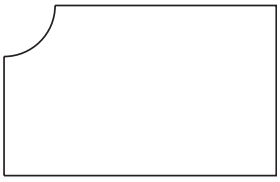


Figure 5 The Sinai billiard.

billiard is very different from the one in the rectangle. The quarter circle acts like a convex mirror which spreads out the rays of light upon reflection. This effect accumulates, because the vast majority of the possible trajectories hit the quarter circle infinitely many times under different angles. This makes the motion in the Sinai billiard *classically chaotic*, while the one in the rectangle is *classically regular*. The rectangle is *separable* and *integrable*, while this feature is destroyed in the Sinai billiard. One now quantizes these billiard systems, calculates the spectra, and analyzes their statistics. Up to certain scales, the rectangle (for irrational squared ratio of the side lengths) shows Poisson behavior, the Sinai billiard yields GOE statistics.

A wealth of such empirical studies led to the Bohigas–Giannoni–Schmit conjecture. We state it here not in its original, but in a frequently used form: *spectra of systems whose classical analogues are fully chaotic show correlation properties as modeled by the Gaussian ensembles*. The Berry–Tabor conjecture is complementary: *spectra of systems whose classical analogs are fully regular show correlation properties which are often those of the Poisson type*. As far as concrete physics applications are concerned, these conjectures are well-posed. From a strict mathematical viewpoint, they have to be supplemented with certain conditions to exclude exceptions such as Artin's billiard. Due to the defnition of this system on the hyperbolic plane, its quantum version shows Poisson-like statistics, although the classical dynamics is chaotic. Up to now, no general and mathematically rigorous proofs could be given. However, *semiclassical* reasoning involving periodic orbit theory and, in particular, the *Gutzwiller trace formula*, yields at least a heuristic understanding.

Quantum chaos has been studied in numerous systems. An especially prominent example is the Hydrogen atom put in a strong magnetic field, which breaks the integrability and drives the correlations towards the GOE limit.

Disordered and Mesoscopic Systems

An electron moving in a probe, a piece of wire, say, is scattered many times at impurities in the material. This renders the motion diffusive. In a statistical model, one writes the Hamilton operator as a sum of the kinetic part, that is, the Laplacian, and a white-noise disorder potential $V(\boldsymbol{r})$ with second moment

$$\langle V(\boldsymbol{r})V(\boldsymbol{r}')\rangle = c_V\delta^{(d)}(\boldsymbol{r}-\boldsymbol{r}') \qquad [28]$$

Here, $\boldsymbol{r}$ is the position vector in d dimensions. The constant c_V determines the mean free time between two scattering processes in relation to the density of states. It is assumed that phase coherence is present such that quantum effects are still significant. This defines the *mesoscopic regime*. The average over the disorder potential can be done with supersymmetry. In fact, this is the context in which supersymmetric techniques in statistical physics were developed, before they were applied to RMT models. In the case of weak disorder, the resulting field theory in superspace for two-level correlations acquires the form

$$\int \mathrm{d}\mu(Q)f(Q)\exp(-S(Q)) \qquad [29]$$

where $f(Q)$ projects out the observable under consideration and where $S(Q)$ is the *effective Lagrangian*

$$S(Q) = -\int \mathrm{str}\left(\mathcal{D}(\nabla Q(\boldsymbol{r}))^2 + \mathrm{i}2rMQ(\boldsymbol{r})\right)\mathrm{d}^d r \qquad [30]$$

This is the *supersymmetric nonlinear σ model*. It is used to study level correlations, but also to obtain information about the conductance and conductance fluctuations when the probe is coupled to external leads. The supermatrix field $Q(\boldsymbol{r})$ is the remainder of the disorder average, its matrix dimension is four or eight, depending on the symmetry class. This field is a Goldstone mode. It does not directly represent a particle as often the case in high-energy physics. The matrix $Q(\boldsymbol{r})$ lives in a coset space of certain supergroups. A tensor M appears in the calculation, and r is the energy difference on the unfolded scale, not to be confused with the position vector $\boldsymbol{r}$.

The first term in the effective Lagrangian involving a gradient squared is the *kinetic term*, it stems from the Laplacian in the Hamiltonian. The constant $\mathcal{D}$ is the classical diffusion constant for the motion of the electron through the probe. The second term is the ergodic term. In the limit of zero dimensions, $d \to 0$, the kinetic term vanishes and the remaining ergodic term yields precisely the unfolded two-level correlations of the Gaussian ensembles. Thus, RMT can be viewed as the zero-dimensional limit of field theory for disordered systems. For $d > 0$, there is a competition between the two terms. The diffusion constant $\mathcal{D}$ and the system size determine an energy scale, the *Thouless*

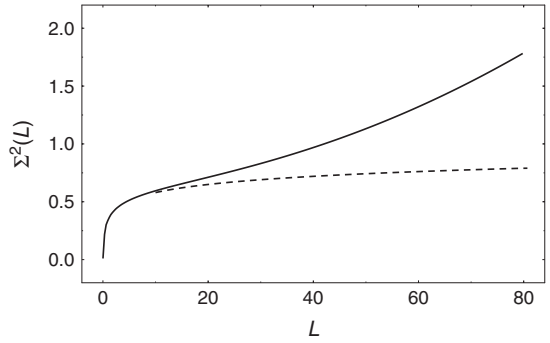


Figure 6 Level number variance $\Sigma^2(L)$. In this example, the Thouless energy is $E_c \approx 10$ on the unfolded scale. The Gaussian ensemble behavior is dashed.

energy E_c, within which the spectral statistics is of the Gaussian ensemble type and beyond which it approaches the Poisson limit. In Figure 6, this is schematically shown for the level number variance $\Sigma^2(L)$, which bends from Gaussian ensemble to Poisson behavior when $L > E_c$. This relates to the crossover transitions in RMT. Gaussian ensemble statistics means that the electron states extend over the probe, while Poisson statistics implies their spatial localization. Hence, the Thouless energy is directly the dimensionless conductance.

A large number of issues in disordered and mesoscopic systems have been studied with the supersymmetric nonlinear σ model. Most results have been derived for *quasi-one-dimensional* systems. Through a proper discretization, a link is established to models involving chains of random matrices. As the conductance can be formulated in terms of the scattering matrix, the experience with RMT for scattering systems can be applied and indeed leads to numerous new results.

Quantum Chromodynamics

Quarks interact by exchanging gluons. In quantum chromodynamics, the gluons are described by gauge fields. Relativistic quantum mechanics has to be used. Analytical calculations are only possible after some drastic assumptions and one must resort to *lattice gauge theory*, that is, to demanding numerics, to study the full problem.

The massless Dirac operator has *chiral symmetry*, implying that all nonzero eigenvalues come in pairs $(-\lambda_n, +\lambda_n)$ symmetrically around zero. In chiral RMT, the Dirac operator is replaced with block off-diagonal matrices

$$W = \begin{bmatrix} 0 & W_b \\ W_b^\dagger & 0 \end{bmatrix} \qquad [31]$$

where W_b is a random matrix without further symmetries. By construction, W has chiral symmetry. The assumption underlying chiral RMT is that the gauge fields effectively randomize the motion of the quark. Indeed, this simple schematic model correctly reproduces low-energy sum rules and spectral statistics of lattice gauge calculations. Near the center of the spectrum, there is a direct connection to the partition function of quantum chromodynamics. Furthermore, a similarity to disordered systems exists and an analog of the Thouless energy could be found.

Other Fields

Of the wealth of further investigations, we can mention but a few. RMT is in general useful for wave phenomena of all kinds, including classical ones. This has been shown for *elastomechanical* and *electromagnetic* resonances.

An important field of application is *quantum gravity* and matrix model aspects of *string theory*. We decided not to go into this, because the reason for the emergence of RMT concepts there is very different from everything else discussed above.

RMT is also successful beyond physics. Not surprisingly, it always received interest in mathematical statistics, but, as already said, it also relates to harmonic analysis. A connection to number theory exists as well. The high-lying zeros of the Riemann ζ function follow the GUE predictions over certain interval lengths. Unfortunately, a deeper understanding is still lacking.

As the interest in statistical concepts grows, RMT keeps finding new applications. Recently, one even started using RMT for risk management in finance.

See also: Arithmetic Quantum Chaos; Chaos and Attractors; Determinantal Random Fields; Free Probability Theory; Growth Processes in Random Matrix Theory; Hyperbolic Billiards; Integrable Systems in Random Matrix Theory; Integrable Systems: Overview; Number Theory in Physics; Ordinary Special Functions; Quantum Chromodynamics; Quantum Mechanical Scattering Theory; Random Partitions; Random Walks in Random Environments; Semi-Classical Spectra and Closed Orbits; Supermanifolds; Supersymmetry Methods in Random Matrix Theory; Symmetry Classes in Random Matrix Theory.

Further Reading

Beenakker CWJ (1997) Random-matrix theory of quantum Transport. *Reviews of Modern Physics* 69: 731–808.

Bohigas O (1984) Chaotic motion and random matrix theory. In: Dehesa JS, Gomez JMG, and Polls A (eds.) *Mathematical and Computational Methods in Physics*, Lecture Notes in Physics, vol. 209. Berlin: Springer.

Brody TA, Flores J, French JB, Mello PA, Pandey A, and Wong SSM (1981) Random-matrix physics: spectrum and strength fluctuations. *Reviews of Modern Physics* 53: 385–479.
Efetov K (1997) *Supersymmetry in Disorder and Chaos*. Cambridge: Cambridge University Press.
Forrester PJ, Snaith NC, and Verbaarschot JJM (eds.) (2003) Special issue: random matrix theory. *Journal of Physics A* 36.
Guhr T, Müller-Groeling A, and Weidenmüller HA (1998) Random-matrix theories in quantum physics: common concepts. *Physics Reports* 299: 189–428.
Haake F (2001) *Quantum Signatures of Chaos*, 2nd edn. Berlin: Springer.
Mehta ML (2004) *Random Matrices*, 3rd edn. New York: Academic Press.
Stöckmann HJ (1999) *Quantum Chaos: An Introduction*. Cambridge: Cambridge University Press.
Verbaarschot JJM and Wettig T (2000) Random matrix theory and chiral symmetry in QCD. *Annual Review of Nuclear and Particle Science* 50: 343–410.

Random Partitions

A Okounkov, Princeton University, Princeton, NJ, USA

Partitions

A partition of n is a monotone sequence of nonnegative integers,

$$\lambda = (\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq 0)$$

with sum n. The number n is also denoted by $|\lambda|$ and is called the size of n. The number of nonzero terms in λ is called the length of λ and often denoted by $\ell(\lambda)$. It is convenient to make the sequence λ infinite by adding a string of zeros at the end.

A geometric object associated to partition is its diagram. The diagram of $\lambda = (4, 2, 2, 1)$ is shown in Figure 1. A larger diagram, flipped and rotated by 135°, can be seen in Figure 2. Flipping the diagram introduces an involution on the set of partitions of n known as transposition. The transposed partition is denoted by λ'.

Partitions serve as natural combinatorial labels for many basic objects in mathematics and physics. For example, partitions of n index both conjugacy classes and irreducible representations of the symmetric group $S(n)$. Partitions λ with $\ell(\lambda) \leq n$ index irreducible polynomial representations of the general linear group $\mathrm{GL}(n)$. More generally, the highest weight of a rational representation of $\mathrm{GL}(n)$ can be naturally viewed as two partitions of total length $\leq n$.

For an even more basic example, partitions λ with $\lambda_1 \leq m$ and $\ell(\lambda) \leq n$ are the same as upright lattice paths making n steps up and m steps to the right (just follow the boundary of λ). In particular, there are $\binom{n+m}{n}$ of such. By a variation on this theme, partitions label the standard basis of fermionic Fock space (Miwa *et al.* 2000). They also label a standard basis of the bosonic Fock space.

In most instances, partitions naturally occur together with some weight function. For example, the dimension, $\dim \lambda$, of an irreducible representation of $S(n)$, or some power of it, is what always appears in harmonic analysis on $S(n)$. By a theorem of Burnside,

$$\mathfrak{M}_{\mathrm{Planch}}(\lambda) = \frac{(\dim \lambda)^2}{n!} \qquad [1]$$

is a probability measure on the set of partitions of n; it is known as the Plancherel measure. Besides harmonic analysis, there are many other contexts in which it appears, for example, by a theorem of Schensted (see Sagan (2001) and Stanley (1999)), the distribution of the first part λ_1 of a Plancherel random partition λ is the same as the distribution of the longest increasing subsequence in a uniformly random permutation of $\{1, 2, \ldots, n\}$.

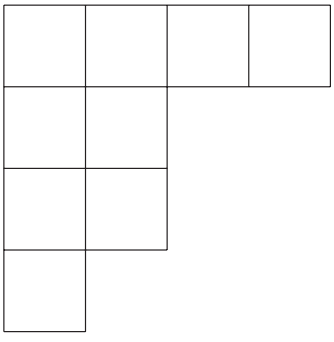


Figure 1 Diagram of a partition.

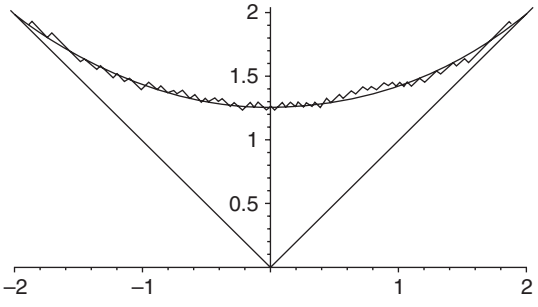


Figure 2 A Plancherel-random partition of 1000 and the limit shape.

Partitions of n being just a finite set, one is often interested in letting $n \to \infty$. Even if the original problem was not of a probabilistic origin, one can still often benefit from adopting a probabilistic viewpoint because of the intuition and techniques that it brings. This is best illustrated by concrete examples, which is what we now turn to. These examples are not meant to be a panorama of random partitions. This is an old and still rapidly growing field and a simple list of all major contributions will take more space than is allowed. The books Kerov (2003), Pitman (n.d.) Sagan (2001), and Stanley (1999) offer much more information on the topics discussed below.

Plancherel Measure

Dimension of a Diagram

There are several formulas and interpretations for the number $\dim \lambda$ in [1]; see Sagan (2001) and Stanley (1999). The one that often appears in the context of growth processes is the following: $\dim \lambda$ is the number of ways to grow the diagram λ from the empty diagram $\emptyset$ by adding a square at a time. That is, $\dim \lambda$ is number of chains of the form

$$\emptyset = \lambda^{(0)} \subset \lambda^{(1)} \subset \cdots \subset \lambda^{(n-1)} \subset \lambda^{(n)} = \lambda$$

where $|\lambda^{(k)}| = k$ and $\mu \subset \lambda$ means inclusion of diagrams.

From the classical formula

$$\dim \lambda = \frac{|\lambda|!}{\prod(\lambda_i + k - i)!} \prod_{i \le j \le k} (\lambda_i - \lambda_j + j - i) \qquad [2]$$

where k is any number such that $\lambda_{k+1} = 0$, one sees that the Plancherel measure is a discrete analog of the eigenvalue density

$$e^{-(1/2)\sum x_i^2} \prod_{i<j} (x_i - x_j)^2$$

of a GUE random matrix (Mehta 1991). Indeed, the first factor in [2], which looks like a multinomial coefficient, is the analog of the Gaussian weight. Kerov (2003) and Johansson were among the first to recognize the analogy between Plancherel measure and GUE. One comes across many partition sums that are discrete analogs of random matrix integrals.

The most compact formula for $\dim \lambda$ is the hook formula

$$\frac{\dim \lambda}{|\lambda|!} = \prod_{\approx \in \lambda} h(\approx)^{-1} \qquad [3]$$

Here the product is over all squares $\approx$ in the diagram of λ and

$$h(\approx) = 1 + a(\approx) + l(\approx)$$

where $a(\approx)$ and $l(\approx)$ is the number of squares to the right of the square $\approx$ and below it, respectively. (These are known as arm-length and leg-length.)

Limit Shape and Edge Scaling

When the diagram of λ is very large, the logarithm of the hook product approximates a double integral. The analysis of the corresponding integral plays the central role, (see Kerov (2003), chapter 3) in the proof of the following law of large numbers for the Plancherel measure.

Take the diagram of λ, flip and rotate it as in **Figure 1** and rescale by a factor of $\sqrt{n}$ so that it has unit area. In this way one obtains a measure on continuous and, in fact, Lipschitz functions. By a result of Logan and Shepp and, independently, Vershik and Kerov these measures converge as $n \to \infty$ to the δ-measure on a single function $\Omega(x)$. This limit shape for the Plancherel measure, is also plotted in **Figure 2**. Explicitly,

$$\Omega(x) = \begin{cases} \frac{2}{\pi}\left(x \arcsin(x/2) + \sqrt{4 - x^2}\right), & |x| \le 2 \\ |x|, & |x| > 2 \end{cases}$$

This is an analog of Wigner's semicircle law (Mehta 1991) for spectra of random matrices. The Gaussian correction to the limit shape was also found by Kerov (2003).

The limit shape result can be refined to show that $\lambda_1/\sqrt{n} \to 2$ in probability. Together with Schensted's theorem, this answers the question posed by Ulam about the longest increasing subsequence in a random permutation. Further progress came in the work of Baik, Deift, and Johansson (see Deift (2000)), who conjectured (and proved for $i = 1$ and 2) that as $n \to \infty$ the joint distribution

$$\frac{\lambda_i - 2\sqrt{n}}{n^{1/6}}, \quad i = 1, 2, \ldots$$

becomes exactly the same as the distribution of largest eigenvalues of a GUE random matrix. In particular, the longest increasing subsequence, suitably scaled, is distributed exactly like the largest eigenvalue. The distribution of the latter is known as the Tracy–Widom distribution; it is given in terms of a particular solution of the Painlevé II equation. For more information about the proof of the full conjecture, see Aldous and Diaconis (1999), Deift (2000), and Okounkov (2002).

Correlation Functions

One way to prove the full BDJ conjecture is to use the following exact formula first obtained in a more general setting by Borodin and Olshanski (see Olshanski (2003), and Okounkov (2002) for further generalizations). Look at the downsteps of the zig-zag curve in Figure 2. The x-coordinates of their midpoints are the numbers

$$\mathfrak{S}(\lambda) = \{\lambda_i - i + \tfrac{1}{2}\} \subset \mathbb{Z} + \tfrac{1}{2} \qquad [4]$$

The map $\lambda \mapsto \mathfrak{S}(\lambda)$ makes a random partition a random subset of $\mathbb{Z} + \frac{1}{2}$, that is, a random point field on a lattice. These random points should be treated like eigenvalues of a random matrix. In particular, it is natural to consider their correlations, that is, the probability that $X \subset \mathfrak{S}(\lambda)$ for some fixed $X \subset \mathbb{Z} + \frac{1}{2}$.

Many formulas work better if we replace the Plancherel measures $\mathfrak{M}_{\mathrm{Planch},n}$ on partitions of a fixed number n by their Poisson average,

$$\mathfrak{M}_\xi = e^{-\xi} \sum_{n \geq 0} \frac{\xi^n}{n!} \mathfrak{M}_{\mathrm{Planch},n}$$

Here $\xi > 0$ is a parameter. It equals the expected size of λ. For any finite set X, we have

$$\mathrm{Prob}_\xi(X \subset \mathfrak{S}(\lambda)) = \det\left[\mathbb{K}_{\mathrm{Bessel}}(x_i, x_j; \xi)\right]_{x_i, x_j \in X} \qquad [5]$$

where $\mathbb{K}_{\mathrm{Bessel}}$ is the discrete Bessel kernel given by

$$\mathbb{K}_{\mathrm{Bessel}}(x,y;\xi) = \sqrt{\xi}\,\frac{J_{x-1/2}(2\sqrt{\xi})J_{y+1/2}(2\sqrt{\xi}) - J_{x+1/2}(2\sqrt{\xi})J_{y-1/2}(2\sqrt{\xi})}{x-y}$$

Note that only Bessel function of integral order enter this formula.

For large argument $\xi, J_n(2\sqrt{\xi})$ has sine asymptotics if $n \ll 2\sqrt{\xi}$ and Airy function asymptotics if $n \approx 2\sqrt{\xi}$. Consequently, one gets the random matrix behavior near the edge of the limit shape and discrete sine kernel asymptotics of correlations in the bulk of the limit shape.

Permutation Enumeration

A basic combinatorial problem is to count permutations $\sigma_1, \ldots, \sigma_p \in S(n)$ of given cycle types $\mu^{(1)}, \ldots, \mu^{(p)}$ such that

$$\sigma_1 \cdots \sigma_p = 1 \qquad [6]$$

A geometric interpretation of this problem is to count covers of the sphere $S^2 = \mathbb{CP}^1$ branched over p given points with monodromy $\mu^{(1)}, \ldots, \mu^{(p)}$. Elementary character theory of $S(n)$ gives (Jones 1998)

$$\#\{\sigma_i \in C_{\mu^{(i)}}, \prod \sigma_i = 1\} = \left\langle \prod f_{\mu^{(i)}} \right\rangle_{\mathrm{Planch}} \qquad [7]$$

where C_μ is the conjugacy class with cycle type μ and

$$f_\mu(\lambda) = |C_\mu| \frac{\chi^\lambda_\mu}{\dim \lambda}$$

is the central character of the irreducible representation λ. Here χ^λ_μ is the character of any $\sigma \in C_\mu$ in the representation λ.

Let μ be of the form $(\bar{\mu}, 1, 1, \ldots)$ with $\bar{\mu}$ fixed. By a result of Kerov and Olshanski, $\binom{n}{|\bar{\mu}|}^{-1} f_\mu(\lambda)$, is a polynomial in λ of degree $|\bar{\mu}|$. See [11] for the simplest example $\bar{\mu} = (2)$, that is, for the central character of a transposition. We thus recognize in [7] a discrete analog of the GUE expectation of a polynomial in traces of a random matrix. This analogy becomes even clearer in the Gromov–Witten theory of $\mathbb{CP}^1$, which can be viewed as taking into account contributions of certain degenerate covers, see Okounkov (2002).

There is a generalization, due to Burnside, of [7] to counting branched covers of surfaces of any genus g; see Jones (1998). The only modification required is that a representation λ is now counted with the weight $(\dim \lambda)^{2-2g}$. For example, covers of the torus correspond to a uniform measure on partitions. In particular, the probability that two random permutation from $S(n)$ commute is $p(n)/n!$, where $p(n)$ is the number of partitions of n.

Generalizations of Plancherel Measure

Schur Functions and Cauchy Identity

Schur functions $s_\lambda(x_1, \ldots, x_n)$, where λ is a partition with at most n parts, form a distinguished linear basis of the algebra of symmetric polynomials in $x_1, \ldots, x_n$. Various definitions and many remarkable properties of these function are discussed in, for example, Sagan (2001) and Stanley (1999). One of them is that $s_\lambda(x)$ is the trace of a matrix with eigenvalues $\{x_i\}$ in an irreducible $\mathrm{GL}(n)$ module with highest weight λ. The following stability of s_λ,

$$s_\lambda(x_1, \ldots, x_n, 0) = s_\lambda(x_1, \ldots, x_n), \quad \ell(\lambda) \leq n$$

allows one to define Schur functions in infinitely many variables. The formulas

$$p_\mu = \sum_\lambda \chi^\lambda_\mu s_\lambda, \qquad s_\lambda = \sum_\mu \frac{\chi^\lambda_\mu}{\mathfrak{z}(\mu)} p_\mu$$

where

$$\mathfrak{z}(\mu) = \frac{|\mu|!}{|C_\mu|} = |\mathrm{Aut}(\mu)| \prod \mu_i$$

establish the transition between the basis of Schur function and the basis of power sum functions

$$p_\mu = \prod p_{\mu_i}, \qquad p_k = \sum_i x_i^k$$

In particular, the dimension function $\dim \lambda$ is the following specialization of the Schur function:

$$\frac{\dim \lambda}{|\lambda|!} = s_\lambda\big|_{p_1=1,\, p_2=p_3=\cdots=0}$$

We will discuss other important specializations of Schur functions later.

A typical situation in which a random matrix integral can be reduced to a sum over partition is when one uses the Cauchy identity

$$\frac{1}{\prod(1-x_i y_i)} = \exp\left(\sum \frac{p_k(x)p_k(y)}{k}\right) = \sum_\lambda s_\lambda(x)s_\lambda(y) \qquad [8]$$

to expand the integrand in Schur function and integrate term by term using, for example, the orthogonality of characters or the identity

$$\int_{\mathrm{U}(n)} s_\lambda(AgBg^{-1})\,\mathrm{d}g = \frac{1}{\dim_n \lambda} s_\lambda(A)s_\lambda(B) \qquad [9]$$

Here $s_\lambda(A)$ denotes the Schur function in eigenvalues of a matrix A, $\mathrm{d}g$ is the normalized Haar measure on the unitary group $\mathrm{U}(n)$, and

$$\dim_n \lambda = s_\lambda(\underbrace{1,\ldots,1}_{n \text{ times}})$$

is the dimension of irreducible $\mathrm{GL}(n)$ module V^λ with highest weight λ. The meaning of [9] is that normalized characters are algebra homomorphisms from the center of the group algebra of $\mathrm{U}(n)$ to numbers. This method of converting a random matrix problem to a random partition problem is known as character expansion (see, e.g., Kazakov (2001)).

Inspired by the Cauchy identity, one can generalize Plancherel measure to

$$\mathfrak{M}_{\mathrm{Schur}} = \prod(1-x_i y_i)s_\lambda(x)s_\lambda(y)$$

where x and y, or, equivalently, $p_k(x)$ and $p_k(y)$, are viewed as parameters. This is known as the Schur measure. If $p_1(x) = p_2(y) = \sqrt{\xi}$ and all other p_k's vanish, we get $\mathfrak{M}_{\mathrm{Schur}} = \mathfrak{M}_\xi$. Many properties of the Plancherel measure can be generalized to Schur measure, in particular, exact formulas for correlation functions, description of the limit shape, etc. (Okounkov 2002).

Dimension Functions

We already met the function $\dim_n \lambda$. There is a useful formula

$$\dim_n \lambda = \prod_{\square\in\lambda} \frac{n + c(\square)}{h(\square)} \qquad [10]$$

where $c((i,j)) = j - i$ is the content of the square $\square$ in ith row and jth column. From [10] it is clear that $\dim_n$ makes sense for arbitrary complex values of n. The corresponding specializations of the Schur measure

$$x = \underbrace{\sqrt{\xi},\ldots,\sqrt{\xi}}_{z \text{ times}}, \qquad y = \underbrace{\sqrt{\xi},\ldots,\sqrt{\xi}}_{z' \text{ times}}$$

where ξ, z, z' are parameters, are related to the so-called Z-measures and their theory is much-developed (Olshanski 2003). As $z, z', \xi^{-1} \to \infty$ in such a way that $zz'\xi \to \xi_0$, we get $\mathfrak{M}_{\xi_0}$ in the limit.

The enumerative problems discussed in the section "Permutation enumeration" have analogs for the unitary groups $\mathrm{U}(n)$ and, suitably interpreted, the answers are the same with the dimension $\dim_n \lambda$ replacing $\dim \lambda$. For example, instead of counting the solutions to [6], one may be interested in the volume of the set of p-tuples of unitary matrices with given eigenvalues that multiply to 1. Geometrically, such data arise as the monodromy of a flat unitary connection over $S^2 \setminus \{p \text{ points}\}$, which is a $\mathrm{U}(n)$ analog of a branched cover. The analog of Burnside's formula is Witten's formula for the volumes of moduli spaces of flat connections on a genus g surface with given holonomy around p punctures, (see, e.g., Witten (1991) and Woodward (2004)). It involves summing normalized characters over all representations V^λ, not necessarily polynomial, with the weight $(\dim V^\lambda)^{2-2g}$. If additionally weighted by a Gaussian of the form $\exp(-A(f_2(\lambda) + (n/2)|\lambda|))$, where

$$f_2(\lambda) = \frac{1}{2}\sum_i \left[\left(\lambda_i - i + \tfrac{1}{2}\right)^2 - \left(-i + \tfrac{1}{2}\right)^2\right] = \sum_{\square\in\lambda} c(\square) \qquad [11]$$

this becomes Migdal's formula for the partition function of the 2D Yang–Mills theory, the positive constant A being the area of the surface (see, e.g., Witten (1991) and Woodward (2004)).

A further generalization naturally arising in the theory of quantum groups is the quantum dimension

$$\dim_{n,q} \lambda = s_\lambda(q^{1-n}, q^{3-n}, \ldots, q^{n-3}, q^{n-1}) = \prod_{\square\in\lambda} \frac{q^{n+c(\square)} - q^{-n-c(\square)}}{q^{h(\square)} - q^{-h(\square)}}$$

where q is a parameter (it is more common to use $\dim_{n,q^{1/2}}$ instead). Obviously, $\dim_{n,q} \to \dim_n$ as $q \to 1$. The function $\dim_{n,q}$ is an important building block of, for example, quantum invariants of knots and 3-folds, and various related objects (see, e.g., Bakalov and Kirillov (2001)). The Verlinde formula (Bakalov and Kirillov 2001) can be viewed as an analog of Burnside's formula with weight $\dim_{n,q}$. When q is a root of unity the summation over λ is naturally truncated to a finite sum.

The next level of generalization is obtained by deforming Schur function to Jack and, more generally, Macdonald symmetric functions (Macdonald 1995). In particular, the Jack polynomial analog of the Plancherel measure is

$$\mathfrak{M}_{\mathrm{Jack}}(\lambda) = \prod_{\square\in\lambda} \frac{n!(t_1t_2)^n}{((a(\square)+1)t_1 + l(\square)t_2)(a(\square)t_1 + (l(\square)+1)t_2)}$$

where t_1, t_2 are parameters, and $a(\square)$ and $l(\square)$ denote, as above, the arm- and leg-length of a square $\square$. This measure depends only on the ratio t_2/t_1 which is the usual parameter of Jack polynomials. To continue the analogy with random matrices, this should be viewed as a general β analog of the Plancherel measure.

The measure $\mathfrak{M}_{\mathrm{Jack}}$ naturally arises in Atiyah–Bott localization computations on the Hilbert scheme of n points in $\mathbb{C}^2$. By definition, this Hilbert scheme parametrizes ideals $I \subset \mathbb{C}[x,y]$ of codimension n as linear spaces. The torus $(\mathbb{C}^*)^2$ acts on it by rescaling x and y and the fixed points of this action are

$$I_\lambda = \text{Span of } \{x^{j-1}y^{i-1}\}_{(i,j)\notin\lambda}$$

where λ is a partition of n. The weight of this fixed point in the Atiyah–Bott formula is proportional to $\mathfrak{M}_{\mathrm{Jack}}(\lambda)$, the parameters t_1 and $-t_2$ being the standard torus weights. Corresponding formulas in K-theory involve a Macdonald polynomial analog of $\dim\lambda$.

Nekrasov defines the partition functions of $\mathcal{N}=2$ supersymmetric gauge theories by formally applying the Atiyah–Bott localization formula to (noncompact) instanton moduli spaces. The resulting expression is a sum over partitions with a weight which is a generalization of $\mathfrak{M}_{\mathrm{Jack}}$. In this way, random partitions enter gauge theory. What is more, statistical properties of these random partitions are reflected in the dynamics of gauge theories. For example, the limit shape turns out to be precisely the Seiberg–Witten curve (see Nekrasov and Okounkov (2003), Okounkov (2002), and also Nakajima and Yoshioka (2003)).

Harmonic Functions on Young Graph

Definitions

Partitions form a natural directed graph $\mathbb{Y}$, known as Young graph, in which there is an edge from μ to λ if λ is obtained from μ by adding a square. We will denote this by $\mu\nearrow\lambda$. Let κ be a non-negative function (called multiplicity) on edges of $\mathbb{Y}$. A function ϕ on the vertices of $\mathbb{Y}$ is harmonic if it satisfies

$$\phi(\mu) = \sum_{\lambda\nwarrow\mu} \kappa(\mu,\lambda)\phi(\lambda) \qquad [12]$$

for any μ. For given edge multiplicities κ, non-negative harmonic functions normalized by $\phi(\emptyset)=1$ form a convex compact (with respect to pointwise convergence) set, which we will denote by $\boldsymbol{H}(\kappa)$. The extreme points of $\boldsymbol{H}(\kappa)$ are the indecomposable or ergodic harmonic functions. They are the most important ones. One defines

$$\dim_\kappa \lambda/\mu = \sum_{\mu=\nu_0\nearrow\nu_1\nearrow\cdots\nearrow\nu_{|\lambda|-|\mu|}=\lambda} \prod \kappa(\nu_i,\nu_{i+1})$$

and $\dim_\kappa\lambda = \dim_\kappa\lambda/\emptyset$. For example, if $\kappa\equiv 1$ then $\dim_\kappa\lambda = \dim\lambda$. Any function $\phi\in\boldsymbol{H}(\kappa)$ defines a probability measure on partitions of fixed size $n, n=0,1,2,\ldots$, by

$$\mathfrak{M}_{\phi,n}(\lambda) = \phi(\lambda)\dim_\kappa\lambda, \quad |\lambda| = n \qquad [13]$$

The mean value property [12] implies a certain coherence of these measures for different values of n, which, in general, does not hold for measures like $\mathfrak{M}_{\mathrm{Schur}}$. Two multiplicity functions κ and κ' are gauge equivalent if

$$\kappa'(\mu,\lambda) = f(\mu)\kappa(\mu,\lambda)f(\lambda)^{-1}$$

for some function f. In this case, $\boldsymbol{H}(\kappa)$ and $\boldsymbol{H}(\kappa')$ are naturally isomorphic and the measures $\mathfrak{M}_\phi$ are the same.

First Example: Thoma Theorem

Let F be a central function on the infinite symmetric group $S(\infty)=\bigcup_n S(n)$, normalized by $F(1)=1$. Restricted to $S(n)$, F is a linear combination of irreducible characters

$$F|_{S(n)} = \sum_{|\lambda|=n} \phi(\lambda)\chi^\lambda$$

The branching rule $\chi^\lambda|_{S(n-1)} = \sum_{\mu\nearrow\lambda}\chi^\mu$ implies that the Fourier coefficients ϕ are harmonic with respect to $\kappa\equiv 1$. They are non-negative if and only if F is a positive-definite function on $S(\infty)$, which means that the matrix $(F(g_ig_j^{-1}))$ is non-negative definite for any $\{g_i\}\subset S(\infty)$. The description of all indecomposable positive-definite central functions on $S(\infty)$ was first

obtained by Thoma (see Kerov (2003, 1998) and Olshanski (2003)). Rephrased in our language, it says that the functions

$$\phi(\lambda) = s_\lambda|_{p_1=1, p_k=\sum \alpha_i^k+(-1)^{k+1}\sum \beta_i^k, k>1}$$

are the extreme points of $\boldsymbol{H}(1)$. Here α_i and β_i are parameters satisfying

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq 0, \qquad \beta_1 \geq \beta_2 \geq \cdots \geq 0$$
$$\sum \alpha_i + \beta_i \leq 1$$

This set is known as the Thoma simplex. The origin $\alpha_i = \beta_i = 0$ corresponds to the Plancherel measure.

A general positive-definite central functions on $S(\infty)$ defines a measure on the Thoma simplex. This measure can be interpreted as a point process on the real line, for example, by placing particles at positions $\{\alpha_i\}$ and $\{-\beta_i\}$. Interesting central functions lead to interesting processes (see Olshanski (2003)).

Second Example: Kingman Theorem

Let Π be a partition of the naturals $\mathbb{N}$ into disjoint subsets. For any $n = 1, 2, \ldots, \Pi$ defines the induced partition Π_n of $\{1, \ldots, n\}$ and hence a partition $\lambda(\Pi_n)$ of the number n. A measure $\mathcal{M}$ on partitions Π is called exchangeable if

$$\mathcal{M}(\Pi_n) = \phi(\lambda(\Pi_n)), \quad n = 1, 2, \ldots$$

for some function ϕ on $\mathbb{Y}$. This implies that ϕ is harmonic for

$$\kappa_{\mathrm{K}}(\mu, \lambda) = \rho_k$$

where $\mu = 1^{\rho_1}2^{\rho_2}\cdots$ and $\lambda = 1^{\rho_1}2^{\rho_2}\cdots k^{\rho_k - 1}(k+1)^{\rho_{k+1}+1}\cdots$. The description of all exchangeable measures $\mathcal{M}$ was first obtained by Kingman. In our language, it says that the extreme points of $\boldsymbol{H}(\kappa_{\mathrm{K}})$ are

$$\phi(\lambda) = m_\lambda|_{p_1=1, p_k=\sum \alpha_i^k, k>1}$$

where m_λ is the monomial symmetric function (sum of all monomials with exponents λ) and α_i are parameters as before. The corresponding measure $\mathcal{M}_\alpha$ can be described as follows. Let X_i be a sequence of independent, identically distributed random variables such that $\{\alpha_i\}$ are the measures of atoms of their distribution. This defines a random partition Π of $\mathbb{N}$ by putting i and j in the same block of Π if and only if $X_i = X_j$. A general exchangeable measure $\mathcal{M}$ is then a convex linear combination of $\mathcal{M}_\alpha$, which can be viewed as making the common distribution of X_i also random. See Pitman (n.d.) for a lot more about Kingman's theorem.

The multiplicities κ_{K} are gauge equivalent to multiplicities

$$\kappa_{\mathrm{VP}}(\mu, \lambda) = k\rho_k \qquad [14]$$

which arise in the study of probability measures on virtual permutations $\mathfrak{S}$ (Olshanski 2003). By definition,

$$\mathfrak{S} = \varprojlim S(n)$$

with respect to the maps $S(n) \to S(n-1)$ that delete n from the disjoint cycle decomposition of a permutation $\sigma \in S(n)$. For $n \geq 5$, this is the unique map that commutes with the right and left action of $S(n-1)$. Thus, $\mathfrak{S}$ has a natural $S(\infty) \times S(\infty)$ action; however, it is not a group. A measure $\mathcal{M}$ on $\mathfrak{S}$ is central if it is invariant under the action of the diagonal subgroup in $S(\infty) \times S(\infty)$. Let the push-forward of $\mathcal{M}$ to $S(n)$ give mass $\phi(\lambda)$ to a permutation with cycle type λ. It is then easy to see that ϕ is harmonic with respect to [14]. Thus, Kingman's theorem gives a description of ergodic central measures on $\mathfrak{S}$. For example, $\alpha_i = 0$ corresponds to the δ-measure at the identity.

Ergodic Method

A unified approach to this type of problems was proposed and developed by Vershik and Kerov. It is based on the following ergodic theorem. Let ϕ be an ergodic harmonic function. Then

$$\phi(\mu) = \lim \frac{\dim_\kappa \lambda/\mu}{\dim_\kappa \lambda}, \quad |\lambda| \to \infty \qquad [15]$$

for almost all λ with respect to the measure [13] (Kerov 2003). This is similar to approximating a Gibbs measure in infinite volume by a sequence of finite-volume Gibbs measures with appropriate boundary conditions. The ratio on the RHS of [15] is known as the Martin kernel. Its asymptotics as $|\lambda| \to \infty$ plays the essential role.

Let us call a sequence $\{\lambda(n)\}$ of partitions of n regular if the limit in [15] exists for all μ. For $\kappa \equiv 1$, Vershik and Kerov proved that $\{\lambda(n)\}$ is regular if and only if the following limits exist:

$$\frac{\lambda(n)_i}{n} \to \alpha_i, \qquad \frac{\lambda(n)'_i}{n} \to \beta_i \qquad [16]$$

that is, if the rows and columns of $\lambda(n)$, scaled by n, have a limit. In this case, the limit in [15] is the harmonic function with Thoma parameters α_i and β_i. This simultaneously proves Thoma classification and gives a law of large numbers for the corresponding measures [13]. It also gives a transparent geometric interpretation of Thoma parameters. Note that the behavior [16] is very different from the formation of a smooth limit shape that we saw earlier. For a common generalization of this result and Kingman's theorem see Kerov (1998).

See also: Determinantal Random Fields; Growth Processes in Random Matrix Theory; Integrable Systems in Random Matrix Theory; Random Matrix Theory in Physics; Symmetry Classes in Random Matrix Theory.

Further Reading

Aldous D and Diaconis P (1999) Longest increasing subsequences: from patience sorting to the Baik–Deift–Johansson theorem. *Bulletin of the American Mathematical Society* 36(4): 413–432.

Bakalov B and Kirillov A Jr. (2001) Lectures on Tensor Categories and Modular Functors. *University Lecture Series,* vol. 21. American Mathematical Society.

Deift P (2000) Integrable systems and combinatorial theory. *Notices of the American Mathematical Society* 47(6): 631–640.

Jones G (1998) *Characters and Surfaces: A Survey*, The Atlas of Finite Groups: Ten Years on (Birmingham, 1995), London Mathematical Society Lecture Note Series, vol. 249, pp. 90–118. Cambridge: Cambridge University Press.

Kazakov V (2001) Solvable Matrix Models, *Random Matrices and Their Applications,* vol. 40. MSRI Publications; Cambridge: Cambridge University Press.

Kerov S (2003) *Asymptotic Representation Theory of the Symmetric Group and its Applications in Analysis*. American Mathematical Society.

Kerov S, Okounkov A, and Olshanski G (1998) The boundary of the Young graph with Jack edge multiplicities. *International Mathematics Research Notices* 4: 173–199.

Macdonald IG (1995) *Symmetric Functions and Hall Polynomials*. Oxford: Clarendon.

Mehta ML (1991) *Random Matrices*, 2nd edn. Boston, MA: Academic Press.

Miwa T, Jimbo M, and Date E (2000) *Solitons. Differential Equations, Symmetries and Infinite-Dimensional Algebras*. Cambridge: Cambridge University Press.

Nakajima H and Yoshioka K (2003) Lectures on instanton counting, math.AG/0311058.

Nekrasov N and Okounkov A (2003) Seiberg–Witten theory and random partitions, hep-th/0306238.

Okounkov A (2002) Symmetric Functions and Random Partitions, *Symmetric Functions 2001: Surveys of Developments and Perspectives*, pp. 223–252, NATO Sci. Ser. II Math. Phys. Chem., 74. (math.CO/0309074). Dordrecht: Kluwer Academic.

Olshanski G (2003) An introduction to harmonic analysis on the infinite symmetric group, math.RT/0311369.

Okounkov A (2002) The uses of random partitions, math-ph/0309015.

Pitman J (n.d.) *Combinatorial Stochastic Processes*, Lecture Notes from St. Four Course, available from www.stat.berekeley.edu.

Sagan B (2001) The Symmetric Group. Representations, Combinatorial Algorithms, and Symmetric Functions. *Graduate Texts in Mathematics*, 2nd edn., vol. 203, New York: Springer.

Stanley R (1999) *Enumerative Combinatorics, II*. Cambridge: Cambridge University Press.

Witten E (1991) On quantum gauge theories in two dimensions. *Communications in Mathematical Physics* 141(1): 153–209.

Woodward C (2004) Localization for the norm-square of the moment map and the two-dimensional Yang–Mills integral, math.SG/0404413.

Random Walks in Random Environments

L V Bogachev, University of Leeds, Leeds, UK

Introduction

Random walks provide a simple conventional model to describe various transport processes, for example, propagation of heat or diffusion of matter through a medium (for a general reference see, e.g., Hughes (1995)). However, in many practical cases, the medium where the system evolves is highly irregular, due to factors such as defects, impurities, fluctuations, etc. It is natural to model such irregularities as "random environment," treating the observable sample as a statistical realization of an ensemble, obtained by choosing the local characteristics of the motion (e.g., transport coefficients and driving fields) at random, according to a certain probability distribution.

In the random walks context, such models are referred to as "random walks in random environments" (RWRE). This is a relatively new chapter in applied probability and physics of disordered systems initiated in the 1970s. Early interest in RWRE models was motivated by some problems in biology, crystallography, and metal physics, but later applications have spread through numerous areas (see review papers by Alexander *et al.* (1981), Bouchaud and Georges (1990), and a comprehensive monograph by Hughes (1996)). After 30 years of extensive work, RWRE remain a very active area of research, which has been a rich source of hard and challenging questions and has already led to many surprising discoveries, such as subdiffusive behavior, trapping effects, localization, etc. It is fair to say that the RWRE paradigm has become firmly established in physics of random media, and its models, ideas, methods, results, and general effects have become an indispensable part of the standard tool kit of a mathematical physicist.

One of the central problems in random media theory is to establish conditions ensuring homogenization, whereby a given stochastic system evolving in a random medium can be adequately described, on some spatial–temporal scale, using a suitable effective system in a homogeneous (nonrandom) medium. In particular, such systems would exhibit classical diffusive behavior with effective drift and diffusion coefficient. Such an approximation, called "effective medium approximation" (EMA), may be expected to

be successful for systems exposed to a relatively small disorder of the environment. However, in certain circumstances, EMA may fail due to atypical environment configurations ("large deviations") leading to various anomalous effects. For instance, with small but positive probability a realization of the environment may create "traps" that would hold the particle for an anomalously long time, resulting in the subdiffusive behavior, with the mean square displacement growing slower than linearly in time.

RWRE models have been studied by various nonrigorous methods including Monte Carlo simulations, series expansions, and the renormalization group techniques (see more details in the above references), but only a few models have been analyzed rigorously, especially in dimensions greater than one. The situation is much more satisfactory in the one-dimensional case, where the mathematical theory has matured and the RWRE dynamics has been understood fairly well.

The goal of this article is to give a brief introduction to the beautiful area of RWRE. The principal model to be discussed is a random walk with nearest-neighbor jumps in independent and identically distributed (i.i.d.) random environment in one dimension, although we shall also comment on some generalizations. The focus is on rigorous results; however, heuristics will be used freely to motivate the ideas and explain the approaches and proofs. In a few cases, sketches of the proofs have been included, which should help appreciate the flavor of the results and methods.

Ordinary Random Walks: A Reminder

To put our exposition in perspective, let us give a brief account of a few basic concepts and facts for ordinary random walks, that is, evolving in a nonrandom environment (see further details in Hughes (1995)). In such models, space is modeled using a suitable graph, for example, a d-dimensional integer lattice $\mathbb{Z}^d$, while time may be discrete or continuous. The latter distinction is not essential, and in this article we will mostly focus on the discrete-time case. The random mechanism of spatial motion is then determined by the given transition probabilities (probabilities of jumps) at each site of the graph. In the lattice case, it is usually assumed that the walk is translation invariant, so that at each step distribution of jumps is the same, with no regard to the current location of the walk.

In one dimension ($d=1$), the simple (nearest-neighbor) random walk may move one step to right or to the left at a time, with some probabilities p and $q=1-p$, respectively. An important assumption is that only the current location of the walk determines the random motion mechanism, whereas the past history is not relevant. In terms of probability theory, such a process is referred to as "Markov chain." Thus, assuming that the walk starts at the origin, its position after n steps can be represented as the sum of consecutive displacements, $X_n = Z_1 + \cdots + Z_n$, where Z_i are independent random variables with the same distribution $P\{Z_i=1\}=p$, $P\{Z_i=-1\}=q$.

The strong law of large numbers (LLN) states that almost surely (i.e., with probability 1)

$$\lim_{n\to\infty} \frac{X_n}{n} = EZ_1 = p-q, \quad P\text{-a.s.} \qquad [1]$$

where E denotes expectation (mean value) with respect to P. This result shows that the random walk moves with the asymptotic average velocity close to $p-q$. It follows that if $p-q\neq 0$, then the process X_n, with probability 1, will ultimately drift to infinity (more precisely, $+\infty$ if $p-q>0$ and $-\infty$ if $p-q<0$). In particular, in this case, the random walk may return to the origin (and in fact visit any site on $\mathbb{Z}$) only finitely many times. Such behavior is called "transient." However, in the symmetric case (i.e., $p=q=0.5$) the average velocity vanishes, so the above argument fails. In this case, the walk behavior appears to be more complicated, as it makes increasingly large excursions both to the right and to the left, so that $\overline{\lim}_{n\to\infty} X_n = +\infty$, $\underline{\lim}_{n\to\infty} X_n = -\infty$ (P-a.s.). This implies that a symmetric random walk in one dimension is "recurrent," in that it visits the origin (and indeed any site on $\mathbb{Z}$) infinitely often. Moreover, it can be shown to be "null-recurrent," which means that the expected time to return to the origin is infinite. That is to say, return to the origin is guaranteed, but it takes very long until this happens.

Fluctuations of the random walk can be characterized further via the central limit theorem (CLT), which amounts to saying that the probability distribution of X_n is asymptotically normal, with mean $n(p-q)$ and variance $4npq$:

$$\lim_{n\to\infty} P\left\{\frac{X_n - n(p-q)}{\sqrt{4npq}} \le x\right\} = \Phi(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \mathrm{e}^{-y^2/2}\,\mathrm{d}y \qquad [2]$$

These results can be extended to more general walks in one dimension, and also to higher dimensions. For instance, the criterion of recurrence for a general one-dimensional random walk is that it is unbiased, $E(X_1 - X_0)=0$. In the two-dimensional case, in addition one needs $E|X_1 - X_0|^2 < \infty$. In higher dimensions, any random walk (which does not reduce to lower dimension) is transient.

Random Environments and Random Walks

The definition of an RWRE involves two ingredients: (1) the environment, which is randomly chosen but remains fixed throughout the time evolution, and (2) the random walk, whose transition probabilities are determined by the environment. The set of environments (sample space) is denoted by $\Omega = \{\omega\}$, and we use $\mathbb{P}$ to denote the probability distribution on this space. For each $\omega \in \Omega$, we define the random walk in the environment ω as the (time-homogeneous) Markov chain $\{X_t,\ t = 0, 1, 2, \ldots\}$ on $\mathbb{Z}^d$ with certain (random) transition probabilities

$$p(x, y, \omega) = P^{\omega}\{X_1 = y | X_0 = x\} \qquad [3]$$

The probability measure P^{ω} that determines the distribution of the random walk in a given environment ω is referred to as the "quenched" law. We often use a subindex to indicate the initial position of the walk, so that, for example, $P^{\omega}_x\{X_0 = x\} = 1$.

By averaging the quenched probability P^{ω}_x further, with respect to the environment distribution, we obtain the "annealed" measure $\boldsymbol{P}_x = \mathbb{P} \times P^{\omega}_x$, which determines the probability law of the RWRE:

$$\boldsymbol{P}_x(A) = \int_{\Omega} P^{\omega}_x(A)\ \mathbb{P}(\mathrm{d}\omega) = \mathbb{E}P^{\omega}_x(A) \qquad [4]$$

Expectation with respect to the annealed measure $\boldsymbol{P}_x$ will be denoted by $\boldsymbol{E}_x$.

Equation [4] implies that if some property A of the RWRE holds almost surely with respect to the quenched law P^{ω}_x for almost all environments (i.e., for all $\omega \in \Omega'$ such that $\mathbb{P}(\Omega') = 1$), then this property is also true with probability 1 under the annealed law $\boldsymbol{P}_x$.

Note that the random walk X_n is a Markov chain only conditionally on the fixed environment (i.e., with respect to P^{ω}_x), but the Markov property fails under the annealed measure $\boldsymbol{P}_x$. This is because the past history cannot be neglected, as it tells what information about the medium must be taken into account when averaging with respect to environment. That is to say, the walk learns more about the environment by taking more steps. (This idea motivates the method of "environment viewed from the particle," see related section below.)

The simplest model is the nearest-neighbor one-dimensional walk, with transition probabilities

$$p(x, y, \omega) = \begin{cases} p_x & \text{if } y = x + 1 \\ q_x & \text{if } y = x - 1 \\ 0 & \text{otherwise} \end{cases}$$

where p_x and $q_x = 1 - p_x$ $(x \in \mathbb{Z})$ are random variables on the probability space $(\Omega, \mathbb{P})$. That is to say, given the environment $\omega \in \Omega$, the random walk currently at point $x \in \mathbb{Z}$ will make a one-unit step to the right, with probability p_x, or to the left, with probability q_x. Here the environment is determined by the sequence of random variables $\{p_x\}$. For most of the article, we assume that the random probabilities $\{p_x, x \in \mathbb{Z}\}$ are i.i.d., which is referred to as "i.i.d. environment." Some extensions to more general environments will be mentioned briefly in the section "Some generalizations and variations." The study of RWRE is simplified under the following natural condition called "(uniform) ellipticity:"

$$0 < \delta \le p_x \le 1 - \delta < 1, \quad x \in \mathbb{Z},\ \mathbb{P}\text{-a.s.} \qquad [5]$$

which will be frequently assumed in the sequel.

Transience and Recurrence

In this section, we discuss a criterion for the RWRE to be transient or recurrent. The following theorem is due to Solomon (1975).

Theorem 1 *Set* $\rho_x := q_x/p_x$, $x \in \mathbb{Z}$, *and* $\eta := \mathbb{E} \ln \rho_0$.

(i) *If* $\eta \neq 0$ *then* X_t *is transient* ($\boldsymbol{P}_0$*-a.s.*); *moreover, if* $\eta < 0$ *then* $\lim_{t\to\infty} X_t = +\infty$, *while if* $\eta > 0$ *then* $\lim_{t\to\infty} X_t = -\infty$ ($\boldsymbol{P}_0$*-a.s.*).

(ii) *If* $\eta = 0$ *then* X_t *is recurrent* ($\boldsymbol{P}_0$*-a.s.*); *moreover,*

$$\overline{\lim_{t\to\infty}}\ X_t = +\infty, \quad \underline{\lim_{t\to\infty}}\ X_t = -\infty, \quad \boldsymbol{P}_0\text{-a.s.}$$

Let us sketch the proof. Consider the hitting times $T_x := \min\{t \ge 0 : X_t = x\}$ and denote by f_{xy} the quenched first-passage probability from x to y:

$$f_{xy} := P^{\omega}_x\{1 \le T_y < \infty\}$$

Starting from 0, the first step of the walk may be either to the right or to the left, hence by the Markov property the return probability f_{00} can be decomposed as

$$f_{00} = p_0 f_{10} + q_0 f_{-1,0} \qquad [6]$$

To evaluate f_{10}, for $n \ge 1$ set

$$u_x \equiv u^{(n)}_x := P^{\omega}_x\{T_0 < T_n\}, \quad 0 \le x \le n$$

which is the probability to reach 0 prior to n, starting from x. Clearly,

$$f_{10} = \lim_{n\to\infty} u^{(n)}_1 \qquad [7]$$

Decomposition with respect to the first step yields the difference equation

$$u_x = p_x u_{x+1} + q_x u_{x-1}, \quad 0 < x < n \qquad [8]$$

with the boundary conditions

$$u_0 = 1, \qquad u_n = 0 \qquad [9]$$

Using $p_x + q_x = 1$, eqn [8] can be rewritten as

$$u_{x+1} - u_x = \rho_x(u_x - u_{x-1})$$

whence by iterations

$$u_{x+1} - u_x = (u_1 - u_0)\prod_{j=1}^{x}\rho_j \qquad [10]$$

Summing over x and using the boundary conditions [9] we obtain

$$1 - u_1 = \left(\sum_{x=0}^{n-1}\prod_{j=1}^{x}\rho_j\right)^{-1} \qquad [11]$$

(if $x = 0$, the product over j is interpreted as 1). In view of eqn [7] it follows that $f_{10} = 1$ if and only if the right-hand side of eqn [11] tends to 0, that is,

$$\sum_{x=1}^{\infty}\exp(Y_x) = \infty, \quad Y_x := \sum_{j=1}^{x}\ln\rho_j \qquad [12]$$

Note that the random variables $\ln\rho_j$ are i.i.d., hence by the strong LLN

$$\lim_{x\to\infty}\frac{Y_x}{x} = \mathbb{E}\ln\rho_0 \equiv \eta, \quad \mathbb{P}\text{-a.s.}$$

That is, the general term of the series [12] for large x behaves like $\exp(x\eta)$; hence, for $\eta > 0$ the condition [12] holds true (and so $f_{10} = 1$), whereas for $\eta < 0$ it fails (and so $f_{10} < 1$).

By interchanging the roles of p_x and q_x, we also have $f_{-1,0} < 1$ if $\eta > 0$ and $f_{-1,0} = 1$ if $\eta < 0$. From eqn [6], it then follows that in both cases $f_{00} < 1$, that is, the random walk is transient.

In the critical case, $\eta = 0$, by a general result from probability theory, $Y_x \geq 0$ for infinitely many x ($\mathbb{P}$-a.s.), and so the series in eqn [12] diverges. Hence, $f_{10} = 1$ and, similarly, $f_{-1,0} = 1$, so by eqn [6] $f_{00} = 1$, that is, the random walk is recurrent.

It may be surprising that the critical parameter appears in the form $\eta = \mathbb{E}\ln\rho_0$, as it is probably more natural to expect, by analogy with the ordinary random walk, that the RWRE criterion would be based on the mean drift, $\mathbb{E}(p_0 - q_0)$. In the next section, we will see that the sign of d may be misleading.

A canonical model of RWRE is specified by the assumption that the random variables p_x take only two values, β and $1-\beta$, with probabilities

$$\mathbb{P}\{p_x = \beta\} = \alpha, \quad \mathbb{P}\{p_x = 1-\beta\} = 1-\alpha \qquad [13]$$

where $0 < \alpha < 1$, $0 < \beta < 1$. Here $\eta = (2\alpha - 1)\times\ln(1 + (1-2\beta)/\beta)$, and it is easy to see that, for example, $\eta < 0$ if $\alpha < 1/2$, $\beta < 1/2$ or $\alpha > 1/2$, $\beta > 1/2$. The recurrent region where $\eta = 0$ splits into two lines, $\beta = 1/2$ and $\alpha = 1/2$. Note that the first case is degenerate and amounts to the ordinary symmetric random walk, while the second one (except where $\beta = 1/2$) corresponds to Sinai's problem (see the section "Sinai's localization"). A "phase diagram" for this model, showing various limiting regimes as a function of the parameters α, β, is presented in **Figure 1**.

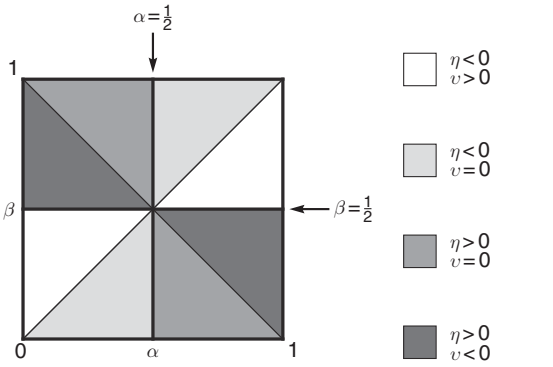


Figure 1 Phase diagram for the canonical model, eqn [13]. In the regions where $\eta < 0$ or $\eta > 0$, the RWRE is transient to $+\infty$ or $-\infty$, respectively. The recurrent case, $\eta = 0$, arises when $\alpha = 1/2$ or $\beta = 1/2$. The asymptotic velocity $v := \lim_{t\to\infty} x_t/t$ is given by eqn [14]. Adapted from Hughes BD (1996) *Random Walks and Random Environments. Volume 2: Random Environments*, Ch. 6, p. 391. Oxford: Clarendon, by permission of Oxford University Press.

Asymptotic Velocity

In the transient case the walk escapes to infinity, and it is reasonable to ask at what speed. For a nonrandom environment, $p_x \equiv p$, the answer is given by the LLN, eqn [1]. For the simple RWRE, the asymptotic velocity was obtained by Solomon (1975). Note that by Jensen's inequality, $(\mathbb{E}\rho_0)^{-1} \leq \mathbb{E}\rho_0^{-1}$.

Theorem 2 *The limit* $v := \lim_{t\to\infty} X_t/t$ *exists* ($\boldsymbol{P}_0$*-a.s.*) *and is given by*

$$v = \begin{cases} \dfrac{1-\mathbb{E}\rho_0}{1+\mathbb{E}\rho_0} & \text{if } \mathbb{E}\rho_0 < 1 \\ -\dfrac{1-\mathbb{E}\rho_0^{-1}}{1+\mathbb{E}\rho_0^{-1}} & \text{if } \mathbb{E}\rho_0^{-1} < 1 \\ 0 & \text{otherwise} \end{cases} \qquad [14]$$

Thus, the RWRE has a well-defined nonzero asymptotic velocity except when $(\mathbb{E}\rho_0)^{-1} \leq 1 \leq \mathbb{E}\rho_0^{-1}$. For instance, in the canonical example eqn [13] (see **Figure 1**), the criterion $\mathbb{E}\rho_0 < 1$ for the velocity v to be positive amounts to the condition that both $(1-\alpha)/\alpha$ and $(1-\beta)/\beta$ lie on the same side of point 1.

The key idea of the proof is to analyze the hitting times T_n first, deducing results for the walk X_t later. More specifically, set $\tau_i = T_i - T_{i-1}$, which is the time to hit i after hitting $i-1$ (providing that $i > X_0$). If $X_0 = 0$ and $n \geq 1$, then $T_n = \tau_1 + \cdots + \tau_n$. Note that in fixed environment ω the random variables $\{\tau_i\}$ are independent, since the quenched random walk "forgets" its past. Although there is no independence with respect to the annealed probability measure $\boldsymbol{P}_0$, one can show that, due to the i.i.d. property of the environment, the sequence $\{\tau_i\}$ is ergodic and therefore satisfies the LLN:

$$\frac{T_n}{n} = \frac{\tau_1 + \cdots + \tau_n}{n} \to \boldsymbol{E}_0\tau_1, \quad \boldsymbol{P}_0\text{-a.s.}$$

In turn, this implies

$$\frac{X_t}{t} \to \frac{1}{\boldsymbol{E}_0\tau_1}, \quad \boldsymbol{P}_0\text{-a.s.} \qquad [15]$$

(the clue is to note that $X_{T_n} = n$).

To compute the mean value $\boldsymbol{E}_0\tau_1$, observe that

$$\tau_1 = \mathbf{1}_{\{X_1=1\}} + \mathbf{1}_{\{X_1=-1\}}(1 + \tau_0' + \tau_1') \qquad [16]$$

where $\mathbf{1}_A$ is the indicator of event A and τ_0', τ_1' are, respectively, the times to get from -1 to 0 and then from 0 to 1. Taking expectations in a fixed environment ω, we obtain

$$E_0^\omega\tau_1 = p_0 + q_0(1 + E_0^\omega\tau_0' + E_0^\omega\tau_1) \qquad [17]$$

and so

$$E_0^\omega\tau_1 = 1 + \rho_0 + \rho_0 E_0^\omega\tau_0' \qquad [18]$$

Note that $E_0^\omega\tau_0'$ is a function of $\{p_x, x < 0\}$ and hence is independent of $\rho_0 = q_0/p_0$. Averaging eqn [18] over the environment and using $\boldsymbol{E}_0\tau_0' = \boldsymbol{E}_0\tau_1$ yields

$$\boldsymbol{E}_0\tau_1 = \begin{cases} \dfrac{1 + \mathbb{E}\rho_0}{1 - \mathbb{E}\rho_0} & \text{if } \mathbb{E}\rho_0 < 1 \\ \infty & \text{if } \mathbb{E}\rho_0 \geq 1 \end{cases} \qquad [19]$$

and by eqn [15] "half" of eqn [14] follows. The other half, in terms of $\mathbb{E}\rho_0^{-1}$, can be obtained by interchanging the roles of p_x and q_x, whereby ρ_0 is replaced with ρ_0^{-1}.

Let us make a few remarks concerning Theorems 1 and 2. First of all, note that by Jensen's inequality $\mathbb{E}\ln\rho_0 \leq \ln\mathbb{E}\rho_0$, with a strict inequality whenever ρ_0 is nondegenerate. Therefore, it may be possible that, with P_0-probability 1, $X_t \to \infty$ but $X_t/t \to 0$ (see **Figure 1**). This is quite unusual as compared to the ordinary random walk (see the subsection "Ordinary random walks: a reminder"), and indicates some kind of slowdown in the transient case.

Furthermore, by Jensen's inequality

$$\mathbb{E}\rho_0 = \mathbb{E}p_0^{-1} - 1 \geq (\mathbb{E}p_0)^{-1} - 1$$

so eqn [14] implies that if $\mathbb{E}\rho_0 < 1$, then

$$0 < v \leq 2\,\mathbb{E}p_0 - 1 = \mathbb{E}\,(p_0 - q_0)$$

and the inequality is strict if p_0 is genuinely random (i.e., does not reduce to a constant). Hence, the asymptotic velocity v is less than the mean drift $\mathbb{E}(p_0 - q_0)$, which is yet another evidence of slowdown. What is even more surprising is that it is possible to have $\mathbb{E}(p_0 - q_0) > 0$ but $\eta = \mathbb{E}\ln\rho_0 > 0$, so that $\mathbb{P}_0$-a.s. $X_t \to -\infty$ (although with velocity $v = 0$). Indeed, following Sznitman (2004) suppose that

$$\mathbb{P}\{p_0 = \beta\} = \alpha, \qquad \mathbb{P}\{p_0 = \gamma\} = 1 - \alpha$$

with $\alpha > 1/2$. Then $\mathbb{E}p_0 \geq \alpha\beta > 1/2$ if $1 > \beta > 1/2\alpha$, hence $\mathbb{E}(p_0 - q_0) = 2\,\mathbb{E}p_0 - 1 > 0$. On the other hand,

$$\mathbb{E}\ln\rho_0 = \alpha\ln\frac{1-\beta}{\beta} + (1-\alpha)\ln\frac{1-\gamma}{\gamma} > 0$$

if γ is sufficiently small.

Critical Exponent, Excursions, and Traps

Extending the previous analysis of the hitting times, one can obtain useful information about the limit distribution of T_n (and hence X_t). To appreciate this, note that from the recursion eqn [16] it follows

$$\tau_1^s = \mathbf{1}_{\{X_1=1\}} + \mathbf{1}_{\{X_1=-1\}}(1 + \tau_0' + \tau_1')^s$$

and, similarly to [17],

$$E_0^\omega\tau_1^s = p_0 + q_0 E_0^\omega(1 + \tau_0' + \tau_1')^s$$

Taking here expectation $\mathbb{E}$, one can deduce that $\boldsymbol{E}_0\tau_1^s < \infty$ if and only if $\mathbb{E}\rho_0^s < 1$. Therefore, it is natural to expect that the root κ of the equation

$$\mathbb{E}\rho_0^\kappa = 1 \qquad [20]$$

plays the role of a critical exponent responsible for the growth rate (and hence, for the type of the limit distribution) of the sum $T_n = \tau_1 + \cdots + \tau_n$. In particular, by analogy with sums of i.i.d. random variables one can expect that if $\kappa > 2$, then T_n is asymptotically normal, with the standard scaling $\sqrt{n}$, while for $\kappa < 2$ the limit law of T_n is stable (with index κ) under scaling $\approx n^{1/\kappa}$.

Alternatively, eqn [20] can be obtained from consideration of excursions of the random walk. Let T_{11}^L be the left-excursion time from site 1, that is the time to return to 1 after moving to the left at the first step. If $\eta = \mathbb{E}\ln\rho_0 < 0$, then $T_{11}^L < \infty$ ($\boldsymbol{P}_0$-a.s.). Fixing an environment ω, let $w_1 = E_1^\omega T_{11}^L$ be the

quenched mean duration of the excursion T_{11}^L and observe that $w_1 = 1 + E_0^\omega \tau_1$, where τ_1 is the time to get back to 1 after stepping to 0.

As a matter of fact, this representation and eqn [19] imply that the annealed mean duration of the left excursion, $\boldsymbol{E}_0 T_{11}^L$, is given by

$$\mathbb{E}w_1 = \begin{cases} \dfrac{2}{1-\mathbb{E}\rho_0} & \text{if } \mathbb{E}\rho_0 < 1 \\ \infty & \text{if } \mathbb{E}\rho_0 \geq 1 \end{cases} \qquad [21]$$

Note that in the latter case (and bearing in mind $\eta < 0$), the random walk starting from 1 will eventually drift to $+\infty$, thus making only a finite number of visits to 0, but the expected number of such visits is infinite.

In fact, our goal here is to characterize the distribution of w_1 under the law $\mathbb{P}$. To this end, observe that the excursion T_{11}^L involves at least two steps (the first and the last ones) and, possibly, several left excursions from 0, each with mean time $w_0 = E_0^\omega T_{00}^L$. Therefore,

$$w_1 = 2 + \sum_{j=1}^{\infty} q_0^j p_0(jw_0) = 2 + \rho_0 w_0 \qquad [22]$$

By the translation invariance of the environment, the random variables w_1 and w_0 have the same distribution. Furthermore, similarly to recursion [22], we have $w_0 = 2 + \rho_{-1} w_{-1}$. This implies that w_0 is a function of p_x with $x \leq -1$ only, and hence w_0 and ρ_0 are independent random variables. Introducing the Laplace transform $\phi(s) = \mathbb{E}\exp(-sw_1)$ and conditioning on ρ_0, from eqn [22] we get the equation

$$\phi(s) = \mathrm{e}^{-2s}\mathbb{E}\phi(s\rho_0) \qquad [23]$$

Suppose that

$$1 - \phi(s) \sim as^\kappa, \quad s \to 0$$

then eqn [23] amounts to

$$1 - as^\kappa + \cdots = (1 - 2s + \cdots)(1 - as^\kappa \mathbb{E}\rho_0^\kappa + \cdots)$$

Expanding the product on the right, one can see that a solution with $\kappa = 1$ is possible only if $\mathbb{E}\rho_0 < 1$, in which case

$$a = \mathbb{E}w_1 = \frac{2}{1-\mathbb{E}\rho_0}$$

We have already obtained this result in eqn [21].

The case $\kappa < 1$ is possible if $\mathbb{E}\rho_0^\kappa = 1$, which is exactly eqn [20]. Returning to w_1, one expects a slow decay of the distribution tail,

$$\mathbb{P}\{w_1 > t\} \sim bt^{-1/\kappa}, \quad t \to \infty$$

In particular, in this case the annealed mean duration of the left excursion appears to be infinite.

Although the above considerations point to the critical parameter κ, eqn [20], which may be expected to determine the slowdown scale, they provide little explanation of a mechanism of the slowdown phenomenon. Heuristically, it is natural to attribute the slowdown effects to the presence of "traps" in the environment, which may be thought of as regions that are easy to enter but hard to leave. In the one-dimensional case, such a trap would occur, for example, between two long series of successive sites where the probabilities p_x are fairly large (on the left) and small (on the right).

Remarkably, traps can be characterized quantitatively with regard to the properties of the random environment, by linking them to certain large-deviation effects (see Sznitman (2002, 2004)). The key role in this analysis is played by the function $F(u) := \ln \mathbb{E}\rho_0^u$, $u \in \mathbb{R}$. Suppose that $\eta = \mathbb{E}\ln\rho_0 < 0$ (so that by Theorem 1 the RWRE tends to $+\infty$, $\boldsymbol{P}_0$-a.s.) and also that $\mathbb{E}\rho_0 > 1$ and $\mathbb{E}\rho_0^{-1} > 1$ (so that by Theorem 2, $\nu = 0$). The latter means that $F(1) > 0$ and $F(-1) > 0$, and since F is a smooth strictly convex function and $F(0) = 0$, it follows that there is the second root $0 < \kappa < 1$, so that $F(\kappa) = 0$, that is, $\mathbb{E}\rho_0^\kappa = 1$ (cf. eqn [20]).

Let us estimate the probability to have a trap in $U = [-L, L]$ where the RWRE will spend anomalously long time. Using eqn [11], observe that

$$P_1^\omega\{T_0 < T_{L+1}\} \geq 1 - \exp\{-LS_L\}$$

where $S_L := L^{-1}\sum_{x=1}^{L} \ln\rho_x \to \eta < 0$ as $L \to \infty$. However, due to large deviations S_L may exceed level $\epsilon > 0$ with probability

$$\mathbb{P}\{S_L > \epsilon\} \sim \exp\{-LI(\epsilon)\}, \quad L \to \infty$$

where $I(x) := \sup_u \{ux - F(u)\}$ is the Legendre transform of F. We can optimize this estimate by assuming that $\epsilon L \geq \ln n$ and minimizing the ratio $I(\epsilon)/\epsilon$. Note that $F(u)$ can be expressed via the inverse Legendre transform, $F(u) = \sup_x \{xu - I(x)\}$, and it is easy to see that if $\kappa := \min_{\epsilon>0} I(\epsilon)/\epsilon$, then $F(\kappa) = 0$, so κ is the second (positive) root of F.

The "left" probability $P_{-1}^\omega\{T_0 < T_{-L-1}\}$ is estimated in a similar fashion, and one can deduce that for some constants $K > 0, c > 0$, and any $\kappa' > \kappa$, for large n

$$\mathbb{P}\left\{P_0^\omega\left\{\max_{k\leq n}|X_k| \leq K\ln n\right\} \geq c\right\} \geq n^{-\kappa'}$$

That is to say, this is a bound on the probability to see a trap centered at 0, of size $\approx \ln n$, which will retain the RWRE for at least time n. It can be shown that, typically, there will be many such traps both in $[-n^{\kappa'}, 0]$ and $[0, n^{\kappa'}]$, which will essentially

prevent the RWRE from moving at distance $n^{\kappa'}$ from the origin before time n. In particular, it follows that $\lim_{n\to\infty} X_n/n^{\kappa'} = 0$ for any $\kappa' > \kappa$, so recalling that $0 < \kappa < 1$, we have indeed a sublinear growth of X_n. This result is more informative as compared to Theorem 2 (the case $\nu = 0$), and it clarifies the role of traps (see more details in Sznitman (2004)). The nontrivial behavior of the RWRE on the precise growth scale, n^{κ}, is characterized in the next section.

Limit Distributions

Considerations in the previous section suggest that the exponent κ, defined as the solution of eqn [20], characterizes environments in terms of duration of left excursions. These heuristic arguments are confirmed by a limit theorem by Kesten *et al.* (1975), which specifies the slowdown scale. We state here the most striking part of their result. Denote $\ln^+ u := \max\{\ln u, 0\}$; by an arithmetic distribution one means a probability law on $\mathbb{R}$ concentrated on the set of points of the form $0, \pm c, \pm 2c, \dots$.

Theorem 3 *Assume that* $-\infty \le \eta = \mathbb{E}\ln\rho_0 < 0$ *and the distribution of* $\ln\rho_0$ *is nonarithmetic (excluding a possible atom at* $-\infty$*). Suppose that the root* κ *of eqn [20] is such that* $0 < \kappa < 1$ *and* $\mathbb{E}\rho_0^{\kappa}\ln^+\rho_0 < \infty$*. Then*

$$\lim_{n\to\infty} \boldsymbol{P}_0\{n^{-1/\kappa}T_n \le t\} = L_\kappa(t)$$

$$\lim_{t\to\infty} \boldsymbol{P}_0\{t^{-\kappa}X_t \le x\} = 1 - L_\kappa(x^{-1/\kappa})$$

where $L_\kappa(\cdot)$ *is the distribution function of a stable law with index* κ*, concentrated on* $[0,\infty)$.

General information on stable laws can be found in many probability books; we only mention here that the Laplace transform of a stable distribution on $[0,\infty)$ with index κ has the form $\phi(s) = \exp\{-Cs^{\kappa}\}$.

Kesten *et al.* (1975) also consider the case $\kappa \ge 1$. Note that for $\kappa > 1$, we have $\mathbb{E}_{\rho_0} < (\mathbb{E}_{\rho_0^{\kappa}})^{1/\kappa} = 1$, so $\nu > 0$ by eqn [14]. For example, if $\kappa > 2$ then, as expected (see the previous section), there exists a nonrandom $\sigma^2 > 0$ such that

$$\lim_{n\to\infty} \boldsymbol{P}_0\left\{\frac{T_n - n/\nu}{\sigma\sqrt{n}} \le t\right\} = \Phi(t)$$

$$\lim_{t\to\infty} \boldsymbol{P}_0\left\{\frac{X_t - t\nu}{\nu^{3/2}\sigma\sqrt{t}} \le x\right\} = \Phi(x)$$

Let us describe an elegant idea of the proof based on a suitable renewal structure. (1) Let U_i^n $(i \le n)$ be the number of left excursions starting from i up to time T_n, and note that $T_n = n + 2\sum_i U_i^n$. Since the walk is transient to $+\infty$, the sum $\sum_{i\le 0} U_i^n$ is finite ($\boldsymbol{P}_0$-a.s.) and so does not affect the limit. (2) Observe that if the environment ω is fixed then the conditional distribution of U_j^n, given $U_{j+1}^n, \dots, U_n^n = 0$, is the same as the distribution of the sum of $1 + U_{j+1}^n$ i.i.d. random variables $V_1, V_2, \dots$, each with geometric distribution $P_0^{\omega}\{V_i = k\} = p_j q_j^k$ $(k = 0, 1, 2, \dots)$. Therefore, the sum $\sum_{i=1}^n U_i^n$ (read from right to left) can be represented as $\sum_{t=0}^{n-1} Z_t$, where $Z_0 = 0, Z_1, Z_2, \dots$ is a branching process (in random environment $\{p_j\}$) with one immigrant at each step and the geometric offspring distribution with parameter p_j for each particle present at time j. (3) Consider the successive "regeneration" times τ_k^*, at which the process Z_t vanishes. The partial sums $W_k := \sum_{\tau_k^* \le t < \tau_{k+1}^*} Z_t$ form an i.i.d. sequence, and the proof amounts to showing that the sum of W_k has a stable limit of index κ. (4) Finally, the distribution of W_0 can be approximated using $M_0 := \sum_{t=1}^{\infty}\prod_{j=0}^{n-1}\rho_j$ (cf. eqn [11]), which is the quenched mean number of total progeny of the immigrant at time $t = 0$. Using Kesten's renewal theorem, it can be checked that $\mathbb{P}\{M_0 > x\} \sim Kx^{-\kappa}$ as $x \to \infty$, so M_0 is in the domain of attraction of a stable law with index κ, and the result follows.

Let us emphasize the significance of the regeneration times τ_i^*. Returning to the original random walk, one can see that these are times at which the RWRE hits a new "record" on its way to $+\infty$, never to backtrack again. The same idea plays a crucial role in the analysis of the RWRE in higher dimensions (see the subsections "Zero–one laws and LLNs" and "Kalikow's condition and Sznitman's condition (T′)").

Finally, note that the condition $-\infty \le \eta < 0$ allows $\mathbb{P}\{p_0 = 1\} > 0$, so the distribution of ρ_0 may have an atom at 0 (and hence $\ln\rho_0$ at $-\infty$). In view of eqn [20], no atom is possible at $+\infty$. The restriction for the distribution of $\ln\rho_0$ to be nonarithmetic is important. This will be illustrated in the section "Diode model," where we discuss the model of random diodes.

Sinai's Localization

The results discussed in the previous section indicate that the less transient the RWRE is (i.e., the critical exponent decreasing to zero), the slower it moves. Sinai (1982) proved a remarkable theorem showing that for the recurrent RWRE (i.e., with $\eta = \mathbb{E}\ln\rho_0 = 0$), the slowdown effect is exhibited in a striking way.

Theorem 4 *Suppose that the environment* $\{p_x\}$ *is i.i.d. and elliptic, eqn [5], and assume that* $\mathbb{E}\ln\rho_0 = 0$, *with* $\mathbb{P}\{\rho_0 = 1\} < 1$. *Denote* $\sigma^2 := \mathbb{E}\ln^2\rho_0, 0 < \sigma^2 < \infty$. *Then there exists a function* $W_n = W_n(\omega)$ *of the random environment such that for any* $\varepsilon > 0$

$$\lim_{n\to\infty} \boldsymbol{P}_0\left\{\left|\frac{\sigma^2 X_n}{\ln^2 n} - W_n\right| > \varepsilon\right\} = 0 \qquad [24]$$

Moreover, W_n *has a limit distribution*:

$$\lim_{n\to\infty} \mathbb{P}\{W_n \le x\} = G(x) \qquad [25]$$

and thus also the distribution of $\sigma^2 X_n/\ln^2 n$ *under* $\boldsymbol{P}_0$ *converges to the same distribution* $G(x)$.

Sinai's theorem shows that in the recurrent case, the RWRE considered on the spatial scale $\ln^2 n$ becomes localized near some random point (depending on the environment only). This phenomenon, frequently referred to as "Sinai's localization," indicates an extremely strong slowdown of the motion as compared with the ordinary diffusive behavior.

Following Révész (1990), let us explain heuristically why X_n is measured on the scale $\ln^2 n$. Rewrite eqn [11] as

$$P_1^\omega\{T_n < T_0\} = \left(1 + \sum_{x=1}^{n-1} \exp(Y_x)\right)^{-1} \qquad [26]$$

where Y_x is defined in eqn [12]. By the CLT, the typical size of $|Y_x|$ for large x is of order of $\sqrt{x}$, and so eqn [26] yields

$$P_1^\omega\{T_n < T_0\} \approx \exp\{-\sqrt{n}\}$$

This suggests that the walk started at site 1 will make about $\exp\{\sqrt{n}\}$ visits to the origin before reaching level n. Therefore, the first passage to site n takes at least time $\approx \exp\{\sqrt{n}\}$. In other words, one may expect that a typical displacement after n steps will be of order of $\ln^2 n$ (cf. eqn [24]). This argument also indicates, in the spirit of the trapping mechanism of slowdown discussed at the end of the section "Critical exponent, excursions, and traps," that there is typically a trap of size $\approx \ln^2 n$, which retains the RWRE until time n.

It has been shown (independently by H Kesten and A O Golosov) that the limit in [25] coincides with the distribution of a certain functional of the standard Brownian motion, with the density function

$$G'(x) = \frac{2}{\pi}\sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\left\{-\frac{(2k+1)^2\pi^2}{8}\,|x|\right\}$$

Environment Viewed from the Particle

This important technique, dating back to Kozlov and Molchanov (1984), has proved to be quite efficient in the study of random motions in random media. The basic idea is to focus on the evolution of the environment viewed from the current position of the walk.

Let θ be the shift operator acting on the space of environments $\Omega = \{\omega\}$ as follows:

$$\omega = \{p_x\} \overset{\theta}{\mapsto} \bar{\omega} = \{p_{x-1}\}$$

Consider the process

$$\omega_n := \theta^{X_n}\omega, \qquad \omega_0 = \omega$$

which describes the state of the environment from the point of view of an observer moving along with the random walk X_n. One can show that ω_n is a Markov chain (with respect to both P_0^ω and $\boldsymbol{P}_0$), with the transition kernel

$$T(\omega, \mathrm{d}\omega') = p_0\,\delta_{\theta\omega}(\mathrm{d}\omega') + q_0\,\delta_{\theta^{-1}\omega}(\mathrm{d}\omega') \qquad [27]$$

and the respective initial law δ_ω or $\mathbb{P}$ (here δ_ω is the Dirac measure, i.e., unit mass at ω).

This fact as it stands may not seem to be of any practical use, since the state space of this Markov chain is very complex. However, the great advantage is that one can find an explicit invariant probability $\mathbb{Q}$ for the kernel T (i.e., such that $\mathbb{Q}T = \mathbb{Q}$), which is absolutely continuous with respect to $\mathbb{P}$.

More specifically, assume that $\mathbb{E}\rho_0 < 1$ and set $\mathbb{Q} = f(\omega)\mathbb{P}$, where (cf. eqn [14])

$$f = \nu\,(1+\rho_0)\sum_{x=0}^{\infty}\prod_{j=1}^{x}\rho_j \qquad \nu = \frac{1-\mathbb{E}\rho_0}{1+\mathbb{E}\rho_0} \qquad [28]$$

Using independence of $\{\rho_x\}$, we note

$$\int_\Omega \mathbb{Q}(\mathrm{d}\omega) = \mathbb{E}f = (1-\mathbb{E}\rho_0)\sum_{x=0}^{\infty}(\mathbb{E}\rho_0)^x = 1$$

hence $\mathbb{Q}$ is a probability measure on Ω. Furthermore, for any bounded measurable function g on Ω we have

$$\begin{aligned}\mathbb{Q}Tg &= \int_\Omega Tg(\omega)\mathbb{Q}(\mathrm{d}\omega) = \mathbb{E}fTg \\ &= \mathbb{E}\{f[p_0\,(g\circ\theta) + q_0\,(g\circ\theta^{-1})]\} \\ &= \mathbb{E}\{g[(p_0 f)\circ\theta^{-1} + (q_0 f)\circ\theta]\}\end{aligned} \qquad [29]$$

By eqn [28],

$$(p_0 f)\circ\theta^{-1} = \nu p_{-1}(1+\rho_{-1})\sum_{x=0}^{\infty}\prod_{j=1}^{x}\rho_{j-1}$$

$$= \nu\left(1+\rho_0\sum_{x=0}^{\infty}\prod_{j=1}^{x}\rho_j\right) = \nu + \frac{\rho_0}{1+\rho_0}f$$

and similarly

$$(q_0 f)\circ\theta = -\nu + \frac{1}{1+\rho_0}f$$

So from eqn [29] we obtain

$$\mathbb{Q}Tg = \mathbb{E}(gf) = \int_\Omega g(\omega)\ \mathbb{Q}(\mathrm{d}\omega) = \mathbb{Q}g$$

which proves the invariance of $\mathbb{Q}$.

To illustrate the environment method, let us sketch the proof of Solomon's result on the asymptotic velocity (see Theorem 2). Set $d(x,\omega) := E_x^\omega(X_1 - X_0) = p_x - q_x$. Noting that $d(x,\omega) = d(0,\theta^x\omega)$, define

$$D_n := \sum_{i=1}^{n} d(X_{i-1},\omega) = \sum_{i=1}^{n} d(0,\theta^{X_{i-1}}\omega)$$

Due to the Markov property, the process $M_n := X_n - D_n$ is a martingale with respect to the natural filtration $\mathcal{F}_n = \sigma\{X_1,\dots,X_n\}$ and the law P_0^ω,

$$E_0^\omega[M_{n+1}\,|\,\mathcal{F}_n] = M_n, \quad P_0^\omega\text{-a.s.}$$

and it has bounded jumps, $|M_n - M_{n-1}| \le 2$. By general results, this implies $M_n/n \to 0$ (P_0^ω-a.s.).

On the other hand, by Birkhoff's ergodic theorem

$$\lim_{n\to\infty}\frac{D_n}{n} = \int_\Omega d(0,\omega)\ \mathbb{Q}(\mathrm{d}\omega), \quad \boldsymbol{P}_0\text{-a.s.}$$

The last integral is easily evaluated to yield

$$\mathbb{E}(p_0 - q_0)f = \nu\mathbb{E}\sum_{x=0}^{\infty}\prod_{j=1}^{x}\rho_j(1-\rho_0)$$

$$= \nu(1-\mathbb{E}\rho_0)\sum_{x=0}^{\infty}(\mathbb{E}\rho_0)^x = \nu$$

and the first part of the formula [14] follows.

The case $\mathbb{E}\rho_0 \ge 1$ can be handled using a comparison argument (Sznitman 2004). Observe that if $p_x \le \tilde{p}_x$ for all x then for the corresponding random walks we have $X_t \le \tilde{X}_t$ (P_0^ω-a.s.). We now define a suitable dominating random medium by setting (for $\gamma > 0$)

$$\tilde{p}_x := \frac{p_x}{1+\gamma} + \frac{\gamma}{1+\gamma} \ge p_x$$

Then $\mathbb{E}\,\tilde{\rho}_0 = \mathbb{E}q_0/(p_0+\gamma) < 1$ if γ is large enough, so by the first part of the theorem, P_0^ω-a.s.,

$$\overline{\lim_{n\to\infty}}\frac{X_n}{n} \le \lim_{n\to\infty}\frac{\tilde{X}_n}{n} = \frac{1-\mathbb{E}\tilde{\rho}_0}{1+\mathbb{E}\tilde{\rho}_0} \qquad [30]$$

Note that $\mathbb{E}\tilde{\rho}_0$ is a continuous function of γ with values in $[0,\mathbb{E}\rho_0] \ni 1$, so there exists γ^* such that $\mathbb{E}\tilde{\rho}_0$ attains the value 1. Passing to the limit in eqn [30] as $\gamma \uparrow \gamma^*$, we obtain $\overline{\lim}_{n\to\infty} X_n/n \le 0$ (P_0^ω-a.s.). Similarly, we get the reverse inequality, which proves the second part of the theorem.

A more prominent advantage of the environment method is that it naturally leads to statements of CLT type. A key step is to find a function $H(x,t,\omega) = x - \nu t + h(x,\omega)$ (called "harmonic coordinate") such that the process $H(X_n,n,\omega)$ is a martingale. To this end, by the Markov property it suffices to have

$$E_{X_n}^\omega H(X_{n+1},n+1,\omega) = H(X_n,n,\omega), \quad P_0^\omega\text{-a.s.}$$

For $\Delta(x,\omega) := h(x+1,\omega) - h(x,\omega)$ this condition leads to the equation

$$\Delta(x,\omega) = \rho_x\Delta(x-1,\omega) + \nu - 1 + (1+\nu)\rho_x$$

If $\mathbb{E}\rho_0 < 1$ (so that $\nu > 0$), there exists a bounded solution

$$\Delta(x,\omega) = \nu - 1 + 2\nu\sum_{k=0}^{\infty}\prod_{i=0}^{k}\rho_{x-i}$$

and we note that $\Delta(x,\omega) = \Delta(0,\theta^x\omega)$ is a stationary sequence with mean $\mathbb{E}\Delta(x,\omega) = 0$. Finally, setting $h(0,\omega) = 0$ we find

$$h(x,\omega) = \begin{cases} \displaystyle\sum_{k=0}^{x-1}\Delta(k,\omega), & x > 0 \\ \displaystyle -\sum_{k=1}^{-x}\Delta(-k,\omega), & x < 0 \end{cases}$$

As a result, we have the representation

$$X_n - n\nu = H(X_n,n,\omega) + h(X_n,\omega) \qquad [31]$$

For a fixed ω, one can apply a suitable CLT for martingale differences to the martingale term in eqn [31], while using that $X_n \sim n\nu$ ($\boldsymbol{P}_0$-a.s.), the second term in eqn [31] is approximated by the sum $\sum_{k=0}^{n\nu}\Delta(k,\omega)$, which can be handled via a CLT for stationary sequences. This way, we arrive at the following result.

Theorem 5 *Suppose that the environment is elliptic, eqn [5], and such that* $\mathbb{E}\rho_0^{2+\varepsilon} < 1$ *for some* $\varepsilon > 0$ (*which implies that* $\mathbb{E}\rho_0 < 1$ *and hence* $\nu > 0$). *Then there exists a nonrandom* $\sigma^2 > 0$ *such that*

$$\lim_{n\to\infty}\boldsymbol{P}_0\left\{\frac{X_n - n\nu}{\sqrt{n\sigma^2}} \le x\right\} = \Phi(x)$$

Note that this theorem is parallel to the result by Kesten *et al.* (1975) on asymptotic normality when $\kappa > 2$ (see the section "Limit distributions"). The moment assumptions in Theorem 5 are more restrictive, but they can be relaxed. On the other hand, Theorem 5 does not impose the nonarithmetic condition on the distribution of the environment (cf. Theorem 3). More importantly, the environment method proves to be quite efficient in more general situations, including non-i.i.d. environments and higher dimensions (at least in some cases, e.g., for random bonds RWRE and balanced RWRE discussed subsequently).

Diode Model

In the preceding sections (except in the section "Limit distributions," where however we were limited to a nonarithmetic case), we assumed that $0 < p_x < 1$ and therefore excluded the situation where there are sites through which motion is permitted in one direction only. Allowing for such a possibility leads to the "diode model" (Solomon 1975). Specifically, suppose that

$$\mathbb{P}\{p_x = \beta\} = \alpha, \qquad \mathbb{P}\{p_x = 1\} = 1 - \alpha \qquad [32]$$

with $0 < \alpha < 1$, $0 < \beta < 1$, so that with probability α a point $x \in \mathbb{Z}$ is a usual two-way site and with probability $1 - \alpha$ it is a repelling barrier ("diode"), through which passage is only possible from left to right. This is an interesting example of statistically inhomogeneous medium, where the particle motion is strongly irreversible due to the presence of special semipenetrable nodes. The principal mathematical advantage of such a model is that the random walk can be decomposed into independent excursions from one diode to the next.

Due to diodes, the RWRE will eventually drift to $+\infty$. If $\beta > 1/2$, then on average it moves faster than in a nonrandom environment with $p_x \equiv \beta$. The situation where $\beta \le 1/2$ is potentially more interesting, as then there is a competition between the local drift of the walk to the left (in ordinary sites) and the presence of repelling diodes on its way. Note that $\mathbb{E}\rho_0 = \alpha\rho$, where $\rho := (1-\beta)/\beta$, so the condition $\mathbb{E}\rho_0 < 1$ amounts to $\beta > \alpha/(1+\alpha)$. In this case (which includes $\beta > 1/2$), formula [14] for the asymptotic velocity applies.

As explained in the section "Critical exponent, excursions, and traps," the quenched mean duration w of the left excursion has Laplace transform given by eqn [23], which now reads

$$\phi(s) = \mathrm{e}^{-2s}\{1 - \alpha + \alpha\,\phi(s\rho)\}$$

This equation is easily solved by iterations:

$$\phi(s) = (1-\alpha)\sum_{k=0}^{\infty} \alpha^k \mathrm{e}^{-st_k}$$
$$t_k := 2\sum_{j=0}^{k} \rho^j \qquad [33]$$

hence the distribution of w is given by

$$\mathbb{P}\{w = t_k\} = (1-\alpha)\alpha^k, \quad k = 0, 1, \ldots$$

This result has a transparent probabilistic meaning. In fact, the factor $(1-\alpha)\alpha^k$ is the probability that the nearest diode on the left of the starting point occurs at distance $k+1$, whereas t_k is the corresponding mean excursion time. Note that formula [33] for t_k easily follows from the recursion $t_k = 2 + \rho t_{k-1}$ (cf. eqn [22]) with the boundary condition $t_0 = 2$.

A self-similar hierarchy of timescales [33] indicates that the process will exhibit temporal oscillations. Indeed, for $\alpha\rho > 1$ the average waiting time until passing through a valley of ordinary sites of length k is asymptotically proportional to $t_k \sim 2\rho^k$, so one may expect the annealed mean displacement $\boldsymbol{E}_0 X_n$ to have a local minimum at $n \approx t_k$. Passing to logarithms, we note that $\ln t_{k+1} - \ln t_k \sim \ln\rho$, which suggests the occurrence of persistent oscillations on the logarithmic timescale, with period $\ln\rho$ (see **Figure 2**). This was confirmed by Bernasconi and Schneider (1985) who showed that for $\alpha\rho > 1$

$$\boldsymbol{E}_0 X_n \sim n^{\kappa} F(\ln n), \quad n \to \infty \qquad [34]$$

where $\kappa = -\ln\alpha/\ln\rho < 1$ is the solution of eqn [20] and the function F is periodic with period $\ln\rho$ (see **Figure 2**).

In contrast, for $\alpha\rho = 1$ one has

$$\boldsymbol{E}_0 X_n \sim \frac{n\ln\rho}{2\ln n}, \quad n \to \infty$$

and there are no oscillations of the above kind.

These results illuminate the earlier analysis of the diode model by Solomon (1975), which in the main has revealed the following. If $\alpha\rho = 1$, then X_n satisfies the strong LLN:

$$\lim_{n\to\infty} \frac{X_n}{n/\ln n} = \frac{\ln\rho}{2}, \quad \boldsymbol{P}_0\text{-a.s.}$$

while in the case $\alpha\rho > 1$ the asymptotic behavior of X_n is quite complicated and unusual: if $n_i \to \infty$ is a sequence of integers such that $\{\ln n_i\} \to \gamma$ (here $\{a\} = a - [a]$ denotes the fractional part of a), then the distribution of $n_i^{-\kappa} X_{n_i}$ under $\boldsymbol{P}_0$ converges to a nondegenerate distribution which depends on γ. Thus, the very existence of the limiting distribution

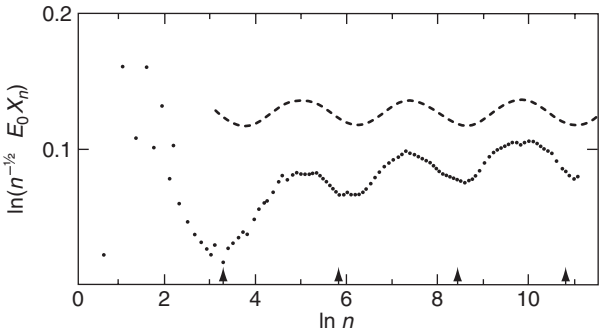


Figure 2 Temporal oscillations for the diode model, eqn [32]. Here $\alpha = 0.3$ and $\rho = 1/0.09$, so that $\alpha\rho > 1$ and $\kappa = 1/2$. The dots represent an average of Monte Carlo simulations over 10 000 samples of the environment with a random walk of 200 000 steps in each realization. The broken curve refers to the exact asymptotic solution [34]. The arrows indicate the simulated locations of the minima t_k, the asymptotic spacing of which is predicted to be $\ln \rho \approx 241$. Reproduced from Bernasconi J and Schneider WR (1982). Diffusion on a one-dimensional lattice with random asymmetric transition rates. *Journal of Physics A: Mathematical and General* 15: L729–L734, by permission of IOP Publishing Ltd.

of X_n and the limit itself heavily depend on the subsequence n_i chosen to approach infinity.

This should be compared with a more "regular" result Theorem 3. Note that almost all the conditions of this theorem are satisfied in the diode model, except that here the distribution of $\ln \rho_0$ is arithmetic (recall that the value $\ln \rho_0 = -\infty$ is permissible), so it is the discreteness of the environment distribution that does not provide enough "mixing" and hence leads to such peculiar features of the asymptotics.

Some Generalizations and Variations

Most of the results discussed above in the simplest context of RWRE with nearest-neighbor jumps in an i.i.d. random environment have been extended to some other cases. One natural generalization is to relax the i.i.d. assumption, for example, by considering stationary ergodic environments (see details in Zeitouni (2004)). In this context, one relies on an ergodic theorem instead of the usual strong LLN. For instance, this way one readily obtains an extension of Solomon's criterion of transience versus recurrence (see Theorem 1). Other examples include an LLN (along with a formula for the asymptotic velocity, cf. Theorem 2), a CLT and stable laws for the asymptotic distribution of X_n (cf. Theorem 3), and Sinai's localization result for the recurrent RWRE (cf. Theorem 4). Usually, however, ergodic theorems cannot be applied directly (like, e.g., to X_n, as the sequence $X_n - X_{n-1}$ is not stationary). In this case, one rather uses the hitting times which possess the desired stationarity (cf. the sections "Asymptotic velocity" and "Critical exponent, excursions, and traps"). In some situations, in addition to stationarity, one needs suitable mixing conditions in order to ensure enough decoupling (e.g., in Sinai's problem). The method of environment viewed from the particle (discussed earlier) is also suited very well to dealing with stationarity.

In the remainder of this section, we describe some other generalizations including RWRE with bounded jumps, RWRE where randomness is attached to bonds rather than sites, and continuous-time (symmetric) RWRE driven by the randomized master equation.

RWRE with Bounded Jumps

The previous discussion was restricted to the case of RWRE with nearest-neighbor jumps. A natural extension is RWRE with bounded jumps. Let L, R be fixed natural numbers, and suppose that from each site $x \in \mathbb{Z}$ jumps are only possible to the sites $x + i$, $i = -L, \ldots, R$, with (random) probabilities

$$p_x(i) \geq 0, \quad \sum_{i=-L}^{R} p_x(i) = 1 \qquad [35]$$

We assume that the random vectors $p_x(\cdot)$ determining the environment are i.i.d. for different $x \in \mathbb{Z}$ (although many results can be extended to the stationary ergodic case).

The study of asymptotic properties of such a model is essentially more complex, as it involves products of certain random matrices and hence must use extensively the theory of Lyapunov exponents (see details and further references in Brémont (2004)). Lyapunov exponents, being natural analogs of logarithms of eigenvalues, characterize the asymptotic action of the product of random matrices along (random) principal directions, as described by Oseledec's multiplicative ergodic theorem. In most situations, however, the Lyapunov spectrum can

only be accessed implicitly, which makes the analysis rather hard.

To explain how random matrices arise here, let us first consider a particular case $R=1, L\geq 1$. Assume that $p_x(-L), p_x(1)\geq \delta>0$ for all $x\in\mathbb{Z}$ (ellipticity condition, cf. eqn [5]), and consider the hitting probabilities $u_n := P_n^{\omega}\{T_0<\infty\}$, where $T_0 := \min\{t\geq 0: X_t\leq 0\}$ (cf. the section "Transience and recurrence"). By decomposing with respect to the first step, for $n\geq 1$ we obtain the difference equation

$$u_n = p_n(1)u_{n+1} + \sum_{i=0}^{L} p_n(-i)u_{n-i} \qquad [36]$$

with the boundary conditions $u_0=\cdots=u_{-L+1}=1$. Using that $1=p_n(1)+\sum_{i=0}^{L} p_n(-i)$, we can rewrite eqn [36] as

$$p_n(1)(u_n-u_{n+1}) = \sum_{i=1}^{L} p_n(-i)(u_{n-i}-u_n)$$

or, equivalently,

$$v_n = \sum_{i=1}^{L} b_n(i)v_{n-i} \qquad [37]$$

where $v_i := u_i - u_{i+1}$ and

$$b_n(i) := \frac{p_n(-i)+\cdots+p_n(-L)}{p_n(1)} \qquad [38]$$

Recursion [37] can be written in a matrix form, $V_n = M_n V_{n-1}$, where $V_n := (v_n,\ldots,v_{n-L+1})^{\top}$,

$$M_n := \begin{pmatrix} b_n(1) & \cdots & \cdots & b_n(L) \\ 1 & \ldots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix} \qquad [39]$$

and by iterations we get (cf. eqn [10])

$$V_n = M_n\cdots M_1 V_0, \qquad V_0 = (1-u_1, 0, \ldots, 0)^{\top}$$

Note that M_n depends only on the transition probability vector $p_n(\cdot)$, and hence $M_n\cdots M_1$ is the product of i.i.d. random (non-negative) matrices. By Furstenberg–Kesten's theorem, the limiting behavior of such a product, as $n\to\infty$, is controlled by the largest Lyapunov exponent

$$\gamma_1 := \lim_{n\to\infty} n^{-1}\ln\|M_n\ldots M_1\| \qquad [40]$$

(by Kingman's subadditive ergodic theorem, the limit exists $\mathbb{P}$-a.s. and is nonrandom). It follows that, $\boldsymbol{P}_0$-a.s., the RWRE X_n is transient if and only if $\gamma_1\neq 0$, and moreover, $\lim_{n\to\infty} X_n = +\infty\,(-\infty)$ when $\gamma_1<0\,(>0)$, whereas $\underline{\lim}_{n\to\infty} X_n = -\infty$, $\overline{\lim}_{n\to\infty} X_n = +\infty$ when $\gamma_1=0$.

For orientation, note that if $p_n(i)=p(i)$ are nonrandom constants, then $\gamma_1=\ln\lambda_1$, where $\lambda_1>0$ is the largest eigenvalue of M_0, and so $\gamma_1<0$ if and only if $\lambda_1<1$. The latter means that the characteristic polynomial $\varphi(\lambda) := \det(M_0-\lambda I)$ satisfies the condition $(-1)^L\varphi(1)>0$. To evaluate $\det(M_0-I)$, replace the first column by the sum of all columns and expand to get $\varphi(1)=(-1)^{L-1}(b_1+\cdots+b_L)$. Substituting expressions [38] it is easy to see that the above condition amounts to $p(1)-\sum_{i=1}^{L} ip\times(-i)>0$, that is, the mean drift of the random walk is positive and hence $X_n\to+\infty$ a.s.

In the general case, $L\geq 1, R\geq 1$, similar considerations lead to the following matrices of order $d := L+R-1$ (cf. eqn [39]):

$$M_n = \begin{pmatrix} a_n(R-1) & \cdots & a_n(1) & b_n(1) & \cdots & b_n(L) \\ 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 & 0 \end{pmatrix}$$

where $b_n(i)$ are given by eqn [38] and

$$a_n(i) := -\frac{p_n(i)+\cdots+p_n(R)}{p_n(R)}$$

Suppose that the ellipticity condition is satisfied in the form $p_n(i)\geq\delta>0$, $i\neq 0$, $-L\leq i\leq R$, and let $\gamma_1\geq\gamma_2\geq\cdots\geq\gamma_d$ be the (nonrandom) Lyapunov exponents of $\{M_n\}$. The largest exponent γ_1 is again given by eqn [40], while other exponents are determined recursively from the equalities

$$\gamma_1+\cdots+\gamma_k = \lim_{n\to\infty} n^{-1}\ln\|\wedge^k(M_n\cdots M_1)\|$$

($1\leq k\leq d$). Here $\wedge$ denotes the external (antisymmetric) product: $x\wedge y=-y\wedge x$ ($x,y\in\mathbb{R}^d$), and $\wedge^k M$ acts on the external product space $\wedge^k\mathbb{R}^d$, generated by the canonical basis $\{e_{i_1}\wedge\cdots\wedge e_{i_k}, 1\leq i_1<\cdots<i_k\leq d\}$, as follows:

$$\wedge^k M(x_1\wedge\cdots\wedge x_k) := M(x_1)\wedge\cdots\wedge M(x_k)$$

One can show that all exponents except γ_R are sign-definite: $\gamma_{R-1}>0>\gamma_{R+1}$. Moreover, it is the sign of γ_R that determines whether the RWRE is transient or recurrent, the dichotomy being the same as in the case $R=1$ above (with γ_1 replaced by γ_R). Let us also mention that an LLN and CLT can be proved here (see Brémont (2004)).

In conclusion, let us point out an alternative approach due to Bolthausen and Goldsheid (2000)

who studied a more general RWRE on a strip $\mathbb{Z}\times\{0,1,\ldots,m-1\}$. The link between these two models is given by the representation $X_n = mY_n + Z_n$, where $m := \max\{L,R\}$, $Y_n \in \mathbb{Z}$, $Z_n \in \{0,\ldots,m-1\}$. Random matrices arising here are constructed indirectly using an auxiliary stationary sequence. Even though these matrices are nonindependent, thanks to their positivity the criterion of transience can be given in terms of the sign of the largest Lyapunov exponent, which is usually much easier to deal with. An additional attractive feature of this approach is that the condition $p_x(R) > 0$ ($\mathbb{P}$-a.s.), which was essential for the previous technique, can be replaced with a more natural condition $\mathbb{P}\{p_x(R) > 0\} > 0$.

Random Bonds RWRE

Instead of having random probabilities of jumps at each site, one could assign random weights to bonds between the sites. For instance, the transition probabilities $p_x = p(x, x+1, \omega)$ can be defined by

$$p_x = \frac{c_{x,x+1}}{c_{x-1,x} + c_{x,x+1}} \qquad [41]$$

where $c_{x,x+1} > 0$ are i.i.d. random variables on the environment space Ω.

The difference between the two models may not seem very prominent, but the behavior of the walk in the modified model [41] appears to be quite different. Indeed, working as in the section "Transience and recurrence," we note that

$$\rho_x = \frac{q_x}{p_x} = \frac{c_{x-1,x}}{c_{x,x+1}}$$

hence, exploiting formulas [11] and [41], we obtain, $\mathbb{P}$-a.s.,

$$\frac{1}{1-u_1} = \sum_{x=0}^{n-1} \frac{c_{01}}{c_{x,x+1}} \sim c_{01}\, n\, \mathbb{E}c_{01}^{-1} \to \infty \qquad [42]$$

since $\mathbb{E}c_{01}^{-1} > 0$. Therefore, $f_{00} = 1$, that is, the random walk is recurrent ($\boldsymbol{P}_0$-a.s.).

The method of environment viewed from the particle can also be applied here (see Sznitman (2004)). Similarly to the section "Environment viewed from the particle," we define a new probability measure $\mathbb{Q} = f(\omega)\,\mathbb{P}$ using the density

$$f(\omega) = Z^{-1}\big(c_{-1,0}(\omega) + c_{01}(\omega)\big)$$

where $Z = 2\mathbb{E}c_{01}$ is the normalizing constant (we assume that $\mathbb{E}c_{01} < \infty$). One can check that $\mathbb{Q}$ is invariant with respect to the transition kernel eqn [41], and by similar arguments as in that section, we obtain that $\lim_{n\to\infty} X_n/n$ exists (P_0^{ω}-a.s.) and is given by

$$\int_{\Omega} d(0,\omega)\ \mathbb{Q}(\mathrm{d}\omega) = Z^{-1}\mathbb{E}\big[c_{01} - c_{-1,0}\big] = 0$$

so the asymptotic velocity vanishes.

Furthermore, under suitable technical conditions on the environment (e.g., c_{01} being bounded away from 0 and ∞, cf. eqn [5]), one can prove the following CLT:

$$\lim_{n\to\infty} \boldsymbol{P}_0\left\{\frac{X_n}{\sqrt{n\sigma^2}} \le x\right\} = \Phi(x) \qquad [43]$$

where $\sigma^2 = (\mathbb{E}c_{01}\cdot\mathbb{E}c_{01}^{-1})^{-1}$. Note that $\sigma^2 \le 1$ (with a strict inequality if c_{01} is not reduced to a constant), which indicates some slowdown in the spatial spread of the random bonds RWRE, as compared to the ordinary symmetric random walk.

Thus, there is a dramatic distinction between the random bonds RWRE, which is recurrent and diffusive, and the random sites RWRE, with a much more complex asymptotics including both transient and recurrent scenarios, slowdown effects, and subdiffusive behavior. This can be explained heuristically by noting that the random bonds RWRE is reversible, that is, $m(x)p(x,y) = m(y)\times p(y,x)$ for all $x, y \in \mathbb{Z}$, with $m(x) := c_{x-1,x} + c_{x,x+1}$ (this property also easily extends to multidimensional versions). Hence, it appears impossible to create extended traps which would retain the particle for a very long time. Instead, the mechanism of the diffusive slowdown in a reversible case is associated with the natural variability of the environment resulting in the occasional occurrence of isolated "screening" bonds with an anomalously small weight $c_{x,x+1}$.

Let us point out that the RWRE determined by eqn [41] can be interpreted in terms of the random conductivity model (see Hughes (1996)). Suppose that each random variable $c_{x,x+1}$ attached to the bond $(x, x+1)$ has the meaning of the conductance of this bond (the reciprocal, $c_{x,x+1}^{-1}$, being its resistance). If a voltage drop V is applied across the system of N successive bonds, say from 0 to N, then the same current I flows in each of the conductors and by Ohm's law we have $I = c_{x,x+1}\ V_{x,x+1}$, where $V_{x,x+1}$ is the voltage drop across the corresponding bond. Hence

$$V = \sum_{x=0}^{N} V_{x,x+1} = I\sum_{x=0}^{N} c_{x,x+1}^{-1}$$

which amounts to saying that the total resistance of the system of consecutive elements is given by the sum of the individual resistances. The effective

conductivity of the finite system, $\bar{c}_N$, is defined as the average conductance per bond, so that

$$\bar{c}_N^{-1} = \frac{1}{N}\sum_{x=0}^{N} c_{x,x+1}^{-1}$$

and by the strong LLN, $\bar{c}_N^{-1} \to \mathbb{E}c_{01}^{-1}$ as $N \to \infty$ ($\mathbb{P}$-a.s.). Therefore, the effective conductivity of the infinite system is given by $\bar{c} = (\mathbb{E}c_{01}^{-1})^{-1}$, and we note that $\bar{c} < \mathbb{E}c_{01}$ if the random medium is nondegenerate.

Returning to the random bonds RWRE, eqn [41], it is easy to see that a site j is recurrent if and only if the conductance $c_{j,\infty}$ between x and ∞ equals zero. Using again Ohm's law, we have (cf. eqn [42])

$$c_{j,+\infty}^{-1} = \sum_{x=j}^{\infty} c_{x,x+1}^{-1} = 0, \quad \mathbb{P}\text{-a.s.}$$

and we recover the result about recurrence.

Continuous-Time RWRE

As in the discrete-time case, a random walk on $\mathbb{Z}$ with continuous time is a homogeneous Markov chain X_t, $t \in [0, \infty)$, with state space $\mathbb{Z}$ and nearest-neighbor (or at least bounded) jumps. The term "Markov" as usual refers to the "lack of memory" property, which amounts to saying that from the entire history of the process development up to a given time, only the current position of the walk is important for the future evolution while all other information is irrelevant.

Since there is no smallest time unit as in the discrete-time case, it is convenient to describe transitions of X_t in terms of transition rates characterizing the likelihood of various jumps during a very short time. More precisely, if $p_{xy}(t) := P\{X_t = y \mid X_0 = x\}$ are the transition probabilities over time t, then for $h \to 0$

$$\begin{aligned} p_{xy}(h) &= c_{xy}h + o(h) \quad (x \neq y) \\ p_{xx}(h) &= 1 - h\sum_{y \neq x} c_{xy} + o(h) \end{aligned} \qquad [44]$$

Equations for the functions $p_{xy}(t)$ can then be derived by adapting the method of decomposition commonly used for discrete-time Markov chains (cf. the section "Transience and recurrence"). Here it is more convenient to decompose with respect to the "last" step, that is, by considering all possible transitions during a small increment of time at the end of the time interval $[0, t+h]$. Using Markov property and eqn [44] we can write

$$\begin{aligned} p_{0x}(t+h) = h\sum_{y \neq x} p_{0y}(t)\, c_{yx} \\ + p_{0x}(t)\left(1 - h\sum_{y \neq x} c_{xy}\right) + o(h) \end{aligned}$$

which in the limit $h \to 0$ yields the master equation (or Chapman–Kolmogorov's forward equation)

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}p_{0x}(t) &= \sum_{y \neq x}\{c_{yx}p_{0y}(t) - c_{xy}p_{0x}(t)\} \\ p_{0x}(0) &= \delta_0(x) \end{aligned} \qquad [45]$$

where $\delta_0(x)$ is the Kronecker symbol.

Continuous-time RWRE are therefore naturally described via the randomized master equation, that is, with random transition rates. The canonical example, originally motivated by Dyson's study of the chain of harmonic oscillators with random couplings, is a symmetric nearest-neighbor RWRE, where the random transition rates c_{xy} are nonzero only for $y = x \pm 1$ and satisfy the condition $c_{x,x+1} = c_{x+1,x}$, otherwise being i.i.d. (see Alexander *et al.* (1981)). In this case, the problem [45] can be formally solved using the Laplace transform, leading to the equations

$$s + G_0^+ + G_0^- = [\hat{p}_0(s)]^{-1} \qquad [46]$$

$$s + G_x^- + G_x^+ = 0 \quad (x \neq 0) \qquad [47]$$

where G_x^-, G_x^+ are defined as

$$G_x^{\pm} := c_{x,x\pm1}\,\frac{\hat{p}_{0x}(s) - \hat{p}_{0,x\pm1}(s)}{\hat{p}_{0x}(s)} \qquad [48]$$

and $\hat{p}_{0x}(s) := \int_0^\infty p_{0x}(t)\, \mathrm{e}^{-st}\mathrm{d}t$. From eqns [47] and [48] one obtains the recursion

$$\begin{aligned} G_x^{\pm} &= \left(\frac{1}{c_{x,x\pm1}} + \frac{1}{s + G_{x\pm1}^{+}}\right)^{-1} \\ x &= 0, \pm1, \pm2, \ldots \end{aligned} \qquad [49]$$

The quantities $G_0^{\pm}$ are therefore expressed as infinite continued fractions depending on s and the random variables $c_{x,x\pm1}, c_{x,x\pm2}, \ldots$. The function $\hat{p}_{00}(s)$ can then be found from eqn [46].

In its generality, the problem is far too hard, and we shall only comment on how one can evaluate the annealed mean

$$\mathbb{E}\hat{p}_{00}(s) = \mathbb{E}\big(s + G_0^+ + G_0^-\big)^{-1}$$

According to eqn [49], the random variables G_0^+, G_0^- are determined by the same algebraic formula, but involve the rate coefficients from different sides of site x, and hence are i.i.d. Furthermore, eqn [49] implies that the random variables G_0^+, G_1^+ have the same distribution and, moreover, G_1^+ and c_{01} are independent. Therefore, eqn [49] may be used as an integral equation for the unknown density function of G_0^+. It can be proved that the suitable solution exists and is unique, and

although an explicit solution is not available, one can obtain the asymptotics of small values of s, thereby rendering information about the behavior of $p_{00}(t)$ for large t. More specifically, one can show that if $c_*:=(\mathbb{E}c_{01}^{-1})^{-1}>0$, then

$$\mathbb{E}\hat{p}_{00}(s)\sim(4c_*s)^{-1/2},\quad s\to 0$$

and so by a Tauberian theorem

$$\mathbb{E}p_{00}(t)\sim(4\pi c_*t)^{-1/2},\quad t\to\infty \qquad [50]$$

Note that asymptotics [50] appears to be the same as for an ordinary symmetric random walk with constant transition rates $c_{x,x+1}=c_{x+1,x}=c_*$, suggesting that the latter provides an EMA for the RWRE considered above.

This is further confirmed by the asymptotic calculation of the annealed mean square displacement, $\boldsymbol{E}_0X_t^2\sim 2c_*t$ as $t\to\infty$ (Alexander *et al.* 1981). Moreover, Kawazu and Kesten (1984) proved that X_t is asymptotically normal:

$$\lim_{t\to\infty}\boldsymbol{P}_0\left\{\frac{X_t}{\sqrt{2c_*t}}\le x\right\}=\Phi(x) \qquad [51]$$

Therefore, if $c_*>0$, then the RWRE has the same diffusive behavior as the corresponding ordered system, with a well-defined diffusion constant $D=c_*$.

In the case where $c_*=0$ (i.e., $\mathbb{E}c_{01}^{-1}=\infty$), one may expect that the RWRE exhibits subdiffusive behavior. For example, if the density function of the transition rates is modeled by

$$f_a(u)=(1-\alpha)\,u^{-\alpha}\mathbf{1}_{\{0<u<1\}}\quad(0<\alpha<1)$$

then, as shown by Alexander *et al.* (1981),

$$\mathbb{E}p_{00}(t)\sim C_\alpha t^{-(1-\alpha)/(2-\alpha)}$$
$$\boldsymbol{E}_0X_t^2\sim C'_\alpha t^{2(1-\alpha)/(2-\alpha)}$$

In fact, Kawazu and Kesten (1984) proved that in this case $t^{-\alpha/(1+\alpha)}X_t$ has a (non-Gaussian) limit distribution as $t\to\infty$.

To conclude the discussion of the continuous-time case, let us point out that some useful information about recurrence of X_t can be obtained by considering an imbedded (discrete-time) random walk $\tilde{X}_n$, defined as the position of X_t after n jumps. Note that continuous-time Markov chains admit an alternative description of their evolution in terms of sojourn times and the distribution of transitions at a jump. Namely, if the environment ω is fixed, then the random sojourn time of X_t in each state x is exponentially distributed with mean $1/c_x$, where $c_x:=\sum_{y\neq x}c_{xy}$, while the distribution of transitions from x is given by the probabilities $p_{xy}=c_{xy}/c_x$.

For the symmetric nearest-neighbor RWRE considered above, the transition probabilities of the imbedded random walk are given by

$$p_x:=p_{x,x+1}=\frac{c_{x,x+1}}{c_{x-1,x}+c_{x,x+1}}$$
$$q_x:=p_{x,x-1}=1-p_x$$

and we recognize here the transition law of a random walk in the random bonds environment considered in the previous subsection (cf. eqn [41]). Recurrence and zero asymptotic velocity established there are consistent with the results discussed in the present section (e.g., note that the CLT for both X_n, eqn [43], and X_t, eqn [51], does not involve any centering). Let us point out, however, that a "naive" discretization of time using the mean sojourn time appears to be incorrect, as this would lead to the scaling $t=n\delta_1$ with $\delta_1:=\mathbb{E}(c_{-1,0}+c_{01})^{-1}$, while from comparing the limit theorems in these two cases, one can conclude that the true value of the effective discretization step is given by $\delta_*:=(2c_*)^{-1}=(1/2)\mathbb{E}c_{01}^{-1}$. In fact, by the arithmetic–harmonic mean inequality we have $\delta_*>\delta_1$, which is a manifestation of the RWRE's diffusive slowdown.

RWRE in Higher Dimensions

Multidimensional RWRE with nearest-neighbor jumps are defined in a similar fashion: from site $x\in\mathbb{Z}^d$ the random walk can jump to one of the $2d$ adjacent sites $x+e\in\mathbb{Z}^d$ (such that $|e|=1$), with probabilities $p_x(e)\ge 0$, $\sum_{|e|=1}p_x(e)=1$, where the random vectors $p_x(\cdot)$ are assumed to be i.i.d. for different $x\in\mathbb{Z}^d$. As usual, we will also impose the condition of uniform ellipticity:

$$\begin{aligned}&p_x(e)\ge\delta>0,\quad \mathbb{P}\text{-a.s.}\\ &|e|=1,\quad x\in\mathbb{Z}^d\end{aligned} \qquad [52]$$

In contrast to the one-dimensional case, theory of RWRE in higher dimensions is far from maturity. Possible asymptotic behaviors of the RWRE for $d\ge 2$ are not understood well enough, and many basic questions remain open. For instance, no definitive classification of the RWRE is available regarding transience and recurrence. Similarly, LLN and CLT have been proved only for a limited number of specific models, while no general sharp results have been obtained. On a more positive note, there has been considerable progress in recent years in the so-called ballistic case, where powerful techniques have been developed (see Sznitman (2002, 2004) and Zeitouni (2003, 2004)). Unfortunately, not much is

known for nonballistic RWRE, apart from special cases of balanced RWRE in $d \geq 2$ (Lawler 1982), small isotropic perturbations of ordinary symmetric random walks in $d \geq 3$ (Bricmont and Kupiainen 1991), and some examples based on combining components of ordinary random walks and RWRE in $d \geq 7$ (Bolthausen *et al.* 2003). In particular, there are no examples of subdiffusive behavior in any dimension $d \geq 2$, and in fact it is largely believed that a CLT is always true in any uniformly elliptic, i.i.d. random environment in dimensions $d \geq 3$, with somewhat less certainty about $d = 2$. A heuristic explanation for such a striking difference with the case $d = 1$ is that due to a less restricted topology of space in higher dimensions, it is much harder to force the random walk to visit traps, and hence the slowdown is not so pronounced.

In what follows, we give a brief account of some of the known results and methods in this fast-developing area (for further information and specific references, see an extensive review by Zeitouni (2004)).

Zero–One Laws and LLNs

A natural first step in a multidimensional context is to explore the behavior of the random walk X_n as projected on various one-dimensional straight lines. Let us fix a test unit vector $\ell \in \mathbb{R}^d$, and consider the process $Z_n^\ell := X_n \cdot \ell$. Then for the events $A_{\pm\ell} := \{\lim_{n\to\infty} Z_n^\ell = \pm\infty\}$ one can show that

$$\boldsymbol{P}_0(A_\ell \cup A_{-\ell}) \in \{0, 1\} \qquad [53]$$

That is to say, for each ℓ the probability that the random walk escapes to infinity in the direction ℓ is either 0 or 1.

Let us sketch the proof. We say that τ is "record time" if $|Z_t^\ell| > |Z_k^\ell|$ for all $k < t$, and "regeneration time" if in addition $|Z_\tau^\ell| \leq |Z_n^\ell|$ for all $n \geq \tau$. Note that by the ellipticity condition [52], $\overline{\lim}_{n\to\infty} |Z_n^\ell| = \infty$ ($\boldsymbol{P}_0$-a.s.), hence there is an infinite sequence of record times $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$. If $\boldsymbol{P}_0(A_\ell \cup A_{-\ell}) > 0$, we can pick a subsequence of record times τ_i', each of which has a positive $\boldsymbol{P}_0$-probability to be a regeneration time (because otherwise $|Z_n^\ell|$ would persistently backtrack towards the origin and the event $A_\ell \cup A_{-\ell}$ could not occur). Since the trials for different record times are independent, it follows that a regeneration time τ^* occurs $\boldsymbol{P}_0$-a.s. Repeating this argument, we conclude that there exists an infinite sequence of regeneration times τ_i^*, which implies that $|Z_n^\ell| \to \infty$ ($\boldsymbol{P}_0$-a.s.), that is, $\boldsymbol{P}(A_\ell \cup A_{-\ell}) = 1$.

Regeneration structure introduced by the sequence $\{\tau_i^*\}$ plays a key role in further analysis of the RWRE and is particularly useful for proving an LLN and a CLT, due to the fact that pieces of the random walk between consecutive regeneration times (and fragments of the random environment involved thereby) are independent and identically distributed (at least starting from τ_1^*). In this vein, one can prove a "directional" version of the LLN, stating that for each ℓ there exist deterministic $v_\ell, v_{-\ell}$ (possibly zero) such that

$$\lim_{n\to\infty} \frac{Z_n^\ell}{n} = v_\ell \mathbf{1}_{A_\ell} + v_{-\ell} \mathbf{1}_{A_{-\ell}}, \quad \boldsymbol{P}_0\text{-a.s.} \qquad [54]$$

Note that if $\boldsymbol{P}_0(A_\ell) \in \{0, 1\}$, then eqn [54] in conjunction with eqn [53] would readily imply

$$\lim_{n\to\infty} \frac{Z_n^\ell}{n} = v_\ell, \quad \boldsymbol{P}_0\text{-a.s.} \qquad [55]$$

Moreover, if $\boldsymbol{P}_0(A_\ell) \in \{0, 1\}$ for any ℓ, then there exists a deterministic v (possibly zero) such that

$$\lim_{n\to\infty} \frac{X_n}{n} = v, \quad \boldsymbol{P}_0\text{-a.s.} \qquad [56]$$

Therefore, it is natural to ask if a zero–one law [53] can be enhanced to that for the individual probabilities $\boldsymbol{P}_0(A_\ell)$. It is known that the answer is affirmative for i.i.d. environments in $d = 2$, where indeed $\boldsymbol{P}(A_\ell) \in \{0, 1\}$ for any ℓ, with counterexamples in certain stationary ergodic (but not uniformly elliptic) environments. However, in the case $d \geq 3$ this is an open problem.

Kalikow's Condition and Sznitman's Condition (T′)

An RWRE is called "ballistic" (ballistic in direction ℓ) if $v \neq 0$ ($v_\ell \neq 0$), see eqns [55] and [56]. In this section, we describe conditions on the random environment which ensure that the RWRE is ballistic.

Let U be a connected strict subset of $\mathbb{Z}^d$ containing the origin. For $x \in U$, denote by

$$g(x, \omega) := E_0^\omega \sum_{n=0}^{T_U} \mathbf{1}_{\{X_n = x\}}$$

the quenched mean number of visits to x prior to the exit time $T_U := \min\{n \geq 0 : X_n \notin U\}$. Consider an auxiliary Markov chain $\widehat{X}_n$, which starts from 0, makes nearest-neighbor jumps while in U, with (nonrandom) probabilities

$$\widehat{p}_x(e) = \frac{\mathbb{E}[g(x, \omega) p_x(e)]}{\mathbb{E}[g(x, \omega)]}, \quad x \in U \qquad [57]$$

and is absorbed as soon as it first leaves U. Note that the expectations in eqn [57] are finite; indeed, if α_x is the probability to return to x before leaving U,

then, by the Markov property, the mean number of returns is given by

$$\sum_{k=1}^{\infty} k\alpha_x^k(1-\alpha_x) = \frac{\alpha_x}{1-\alpha_x} < \infty$$

since, due to ellipticity, $\alpha_x < 1$.

An important property, highlighting the usefulness of $\widehat{X}_n$, is that if $\widehat{X}_n$ leaves U with probability 1, then the same is true for the original RWRE X_n (under the annealed law $\boldsymbol{P}_0$), and moreover, the exit points $\widehat{X}_{\widehat{T}_U}$ and X_{T_U} have the same distribution laws.

Let $\ell \in \mathbb{R}^d, |\ell| = 1$. One says that Kalikow's condition with respect to ℓ holds if the local drift of $\widehat{X}_n$ in the direction ℓ is uniformly bounded away from zero:

$$\inf_U \inf_{x \in U} \sum_{|e|=1} (e \cdot \ell)\, \widehat{p}_x(e) > 0 \qquad [58]$$

A sufficient condition for [58] is, for example, that for some $\kappa > 0$

$$\mathbb{E}\big[(d(0,\omega)\cdot\ell)_+\big] \geq \kappa \mathbb{E}\big[(d(0,\omega)\cdot\ell)_-\big]$$

where $d(0,\omega) = E_0^\omega X_1$ and $u_\pm := \max\{\pm u, 0\}$.

A natural implication of Kalikow's condition [58] is that $\boldsymbol{P}_0(A_\ell) = 1$ and $v_\ell > 0$ (see eqn [55]). Moreover, noting that eqn [58] also holds for all ℓ' in a vicinity of ℓ and applying the above result with d noncollinear vectors from that vicinity, we conclude that under Kalikow's condition there exists a deterministic $v \neq 0$ such that $X_n/n \to v$ as $n \to \infty$ ($\boldsymbol{P}_0$-a.s.). Furthermore, it can be proved that $(X_n - nv)/\sqrt{n}$ converges in law to a Gaussian distribution (see Sznitman (2004)).

It is not hard to check that in dimension $d = 1$ Kalikow's condition is equivalent to $v \neq 0$ and therefore characterizes completely all ballistic walks. For $d \geq 2$, the situation is less clear; for instance, it is not known if there exist RWRE with $\boldsymbol{P}(A_\ell) > 0$ and $v_\ell = 0$ (of course, such RWRE cannot satisfy Kalikow's condition).

Sznitman (2004) has proposed a more complicated transience condition (T′) involving certain regeneration times τ_i^* similar to those described in the previous subsection. An RWRE is said to satisfy Sznitman's condition (T′) relative to direction ℓ if $\boldsymbol{P}_0(A_\ell) = 1$ and for some $c > 0$ and all $0 < \gamma < 1$

$$\boldsymbol{E}_0 \exp\left(c \sup_{n \leq \tau_1^*} |X_n|^\gamma\right) < \infty \qquad [59]$$

This condition provides a powerful control over τ_1^* for $d \geq 2$ and in particular ensures that τ_1^* has finite moments of any order. This is in sharp contrast with the one-dimensional case, and should be viewed as a reflection of much weaker traps in dimensions $d \geq 2$.

Condition [59] can also be reformulated in terms of the exit distribution of the RWRE from infinite thick slabs "orthonormal" to directions ℓ' sufficiently close to ℓ. As it stands, the latter reformulation is difficult to check, but Sznitman (2004) has developed a remarkable "effective" criterion reducing the job to a similar condition in finite boxes, which is much more tractable and can be checked in a number of cases.

In fact, condition (T′) follows from Kalikow's condition, but not the other way around. In the one-dimensional case, condition (T′) (applied to $\ell = 1$ and $\ell = -1$) proves to be equivalent to the transient behavior of the RWRE, which, as we have seen in Theorem 2, may happen with $v = 0$, that is, in a nonballistic scenario. The situation in $d \geq 2$ is quite different, as condition (T′) implies that the RWRE is ballistic in the direction ℓ (with $v_\ell > 0$) and satisfies a CLT (under $\boldsymbol{P}_0$). It is not known whether the ballistic behavior for $d \geq 2$ is completely characterized by condition (T′), although this is expected to be true.

Balanced RWRE

In this section we discuss a particular case of nonballistic RWRE, for which LLN and CLT can be proved. Following Lawler (1982), we say that an RWRE is "balanced" if $p_x(e) = p_x(-e)$ for all $x \in \mathbb{Z}^d, |e| = 1$ ($\mathbb{P}$-a.s.). In this case, the local drift vanishes, $d(x,\omega) = 0$, hence the coordinate processes X_n^i ($i = 1, \ldots, d$) are martingales with respect to the natural filtration $\mathcal{F}_n = \sigma\{X_0, \ldots, X_n\}$. The quenched covariance matrix of the increments $\Delta X_n^i := X_{n+1}^i - X_n^i$ ($i = 1, \ldots, d$) is given by

$$E_0^\omega\big[\Delta X_n^i \Delta X_n^j | \mathcal{F}_n\big] = 2\delta_{ij} p_{X_n}(e_i) \qquad [60]$$

Since the right-hand side of eqn [60] is uniformly bounded, it follows that $X_n/n \to 0$ ($\boldsymbol{P}_0$-a.s.). Further, it can be proved that there exist deterministic positive constants $a_1, \ldots, a_d$ such that for $i = 1, \ldots, d$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} p_{X_k}(e_i) = \frac{a_i}{2}, \quad \boldsymbol{P}_0\text{-a.s.} \qquad [61]$$

Once this is proved, a multidimensional CLT for martingale differences yields that $X_n/\sqrt{n}$ converges in law to a Gaussian distribution with zero mean and the covariances $b_{ij} = \delta_{ij} a_i$.

The proof of [61] employs the method of environment viewed from the particle. Namely, define a Markov chain $\omega_n := \theta^{X_n}\omega$ with the transition kernel

$$T(\omega, \mathrm{d}\omega') = \sum_{i=1}^{d} [p_0(e_i)\delta_{\theta_\omega}(\mathrm{d}\omega') + p_0(-e_i)\delta_{\theta^{-1}\omega}(\mathrm{d}\omega')]$$

(cf. eqn [27]). The next step is to find a probability measure $\mathbb{Q}$ on Ω invariant under T and absolutely continuous with respect to $\mathbb{P}$. Unlike the one-dimensional case, however, an explicit form of $\mathbb{Q}$ is not available, and $\mathbb{Q}$ is constructed indirectly as the limit of invariant measures of certain periodic modifications of the RWRE. Birkhoff's ergodic theorem then yields, $\boldsymbol{P}_0$-a.s.,

$$\frac{1}{n}\sum_{k=0}^{n-1} p_{X_k}(e_i,\omega) = \frac{1}{n}\sum_{k=0}^{n-1} p_0(e_i,\omega_k) \to \int_\Omega p_0(e_i,\omega)\ \mathbb{Q}(\mathrm{d}\omega) \geq \delta$$

by the ellipticity condition [52], and eqn [61] follows.

With regard to transience, balanced RWREs admit a complete and simple classification. Namely, it has been proved (see Zeitouni (2004)) that any balanced RWRE is transient for $d \geq 3$ and recurrent for $d = 2$ ($\boldsymbol{P}_0$-a.s.). It is interesting to note, however, that these answers may be false for certain balanced random walks in a fixed environment ($\mathbb{P}$-probability of such environments being zero, of course). Indeed, examples can be constructed of balanced random walks in $\mathbb{Z}^2$ and in $\mathbb{Z}^d$ with $d \geq 3$, which are transient and recurrent, respectively (Zeitouni 2004).

RWRE Based on Modification of Ordinary Random Walks

A number of partial results are known for RWRE constructed on the basis of ordinary random walks via certain randomization of the environment. A natural model is obtained by a small perturbation of a simple symmetric random walk. To be more precise, suppose that: (1) $|p_x(e) - 1/2d| < \varepsilon$ for all $x \in \mathbb{Z}^d$ and any $|e| = 1$, where $\varepsilon > 0$ is small enough; (2) $\mathbb{E}p_x(e) = 1/2d$; (3) vectors $p_x(\cdot)$ are i.i.d. for different $x \in \mathbb{Z}^d$; and (4) the distribution of the vector $p_x(\cdot)$ is isotropic, that is, invariant with respect to permutations of its coordinates. Then for $d \geq 3$ Bricmont and Kupiainen (1991) have proved an LLN (with zero asymptotic velocity) and a quenched CLT (with nondegenerate covariance matrix). The proof is based on the renormalization group method, which involves decimation in time combined with a suitable spatial–temporal scaling. This transformation replaces an RWRE by another RWRE with weaker randomness, and it can be shown that iterations converge to a Gaussian fixed point.

Another class of examples is also built using small perturbations of simple symmetric random walks, but is anisotropic and exhibits ballistic behavior, providing that the annealed local drift in some direction is strong enough (see Sznitman (2004)). More precisely, suppose that $d \geq 3$ and $\eta \in (0,1)$. Then there exists $\varepsilon_0 = \varepsilon_0(d,\eta) > 0$ such that if $|p_x(e) - 1/2d| < \varepsilon$ ($x \in \mathbb{Z}^d, |e| = 1$) with $0 < \varepsilon < \varepsilon_0$, and for some e_0 one has $\mathbb{E}[d(x,\omega)\cdot e_0] \geq \varepsilon^{2.5-\eta}$ ($d = 3$) or $\geq \varepsilon^{3-\eta}$ ($d \geq 4$), then Sznitman's condition (T′) is satisfied with respect to e_0 and therefore the RWRE is ballistic in the direction e_0 (cf. the subsection "Kalikow's condition and Sznitman's condition (T′)").

Examples of a different type are constructed in dimensions $d \geq 6$ by letting the first $d_1 \geq 5$ coordinates of the RWRE X_n behave according to an ordinary random walk, while the remaining $d_2 = d - d_1$ coordinates are exposed to a random environment (see Bolthausen *et al.* (2003)). One can show that there exists a deterministic v (possibly zero) such that $X_n/n \to v$ ($\boldsymbol{P}_0$-a.s.). Moreover, if $d_1 \geq 13$, then $(X_n - nv)/\sqrt{n}$ satisfies both quenched and annealed CLT. Incidentally, such models can be used to demonstrate the surprising features of the multidimensional RWRE. For instance, for $d \geq 7$ one can construct an RWRE X_n such that the annealed local drift does not vanish, $\mathbb{E}d(x,\omega) \neq 0$, but the asymptotic velocity is zero, $X_n/n \to 0$ ($\boldsymbol{P}_0$ - a.s.), and furthermore, if $d \geq 15$, then in this example $X_n/\sqrt{n}$ satisfies a quenched CLT. (In fact, one can construct such RWRE as small perturbations of a simple symmetric walk.) On the other hand, there exist examples (in high enough dimensions) where the walk is ballistic with a velocity which has an opposite direction to the annealed drift $\mathbb{E}d(x,\omega) \neq 0$. These striking examples provide "experimental" evidence of many unusual properties of the multidimensional RWRE, which, no doubt, will be discovered in the years to come.

See also: Averaging Methods; Growth Processes in Random Matrix Theory; Lagrangian Dispersion (Passive Scalar); Random Dynamical Systems; Random Matrix Theory in Physics; Stochastic Differential Equations; Stochastic Loewner Evolutions.

Further Reading

Alexander S, Bernasconi J, Schneider WR, and Orbach R (1981) Excitation dynamics in random one-dimensional systems. *Reviews of Modern Physics* 53: 175–198.

Bernasconi J and Schneider WR (1985) Random walks in one-dimensional random media. *Helvetica Physica Acta* 58: 597–621.

Bolthausen E and Goldsheid I (2000) Recurrence and transience of random walks in random environments on a strip. *Communications in Mathematical Physics* 214: 429–447.

Bolthausen E, Sznitman A-S, and Zeitouni O (2003) Cut points and diffusive random walks in random environments. *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques* 39: 527–555.

Bouchaud J-P and Georges A (1990) Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications. *Physical Reports* 195: 127–293.

Brémont J (2004) Random walks in random medium on $\mathbb{Z}$ and Lyapunov spectrum. *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques* 40: 309–336.

Bricmont J and Kupiainen A (1991) Random walks in asymmetric random environments. *Communications in Mathematical Physics* 142: 345–420.

Hughes BD (1995) *Random Walks and Random Environments. Volume 1: Random Walks*. Oxford: Clarendon.

Hughes BD (1996) *Random Walks and Random Environments. Volume 2: Random Environments*. Oxford: Clarendon.

Kawazu K and Kesten H (1984) On birth and death processes in symmetric random environment. *Journal of Statistical Physics* 37: 561–576.

Kesten H, Kozlov MV, and Spitzer F (1975) A limit law for random walk in a random environment. *Compositio Mathematica* 30: 145–168.

Kozlov SM and Molchanov SA (1984) On conditions for applicability of the central limit theorem to random walks on a lattice. *Soviet Mathematics Doklady* 30: 410–413.

Lawler GF (1982) Weak convergence of a random walk in a random environment. *Communications in Mathematical Physics* 87: 81–87.

Molchanov SA (1994) Lectures on random media. In: Bernard P (ed.) *Lectures on Probability Theory*, Ecole d'Eté de Probabilités de Saint-Flour XXII-1992, Lecture Notes in Mathematics, vol. 1581, pp. 242–411. Berlin: Springer.

Révész P (1990) *Random Walk in Random and Non-Random Environments*. Singapore: World Scientific.

Sinai YaG (1982) The limiting behavior of a one-dimensional random walk in a random medium. *Theory of Probability and Its Applications* 27: 256–268.

Solomon F (1975) Random walks in a random environment. *The Annals of Probability* 3: 1–31.

Sznitman A-S (2002) Lectures on random motions in random media. In: Bolthausen E and Sznitman A-S. *Ten Lectures on Random Media*, DMV Seminar, vol. 32. Basel: Birkhäuser.

Sznitman A-S (2004) Topics in random walks in random environment. In: Lawler GF (ed.) *School and Conference on Probability Theory* (Trieste, 2002), ICTP Lecture Notes Series, vol. XVII, pp. 203–266 (Available at http://www.ictp.trieste.it/~pub_off/lectures/vol17.html).

Zeitouni O (2003) Random walks in random environments. In: Tatsien Li (ed.) *Proceedings of the International Congress of Mathematicians* (Beijing, 2002), vol. III, pp. 117–127. Beijing: Higher Education Press.

Zeitouni O (2004) Random walks in random environment. In: Picard J (ed.) *Lectures on Probability Theory and Statistics*, Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001, Lecture Notes in Mathematics, vol. 1837, pp. 189–312. New York: Springer.

Recursion Operators in Classical Mechanics

F Magri, Università di Milano Bicocca, Milan, Italy
M Pedroni, Università di Bergamo, Dalmine (BG), Italy

Introduction

One of the tasks of classical mechanics has always been to identify those Hamiltonian systems which, by their peculiar properties, are considered solvable. The integrable systems of Liouville and the separable systems of Jacobi can serve as representative examples here. The bi-Hamiltonian geometry, a branch of Poisson geometry dealing with a special kind of deformation of Poisson bracket, suggests two further classes of Hamiltonian systems – the bi-Hamiltonian systems and the cyclic systems of Levi-Civita. The purpose of this article is to investigate the second class of systems mentioned above, and to explain why they are relevant for classical mechanics. (*see* Bi-Hamiltonian Methods in Soliton Theory and Multi-Hamiltonian Systems for further details).

To define a cyclic system of Levi-Civita, one must consider a symplectic manifold (S, ω) endowed with a tensor field of type $(1, 1)$, seen as an endomorphism $N : TS \to TS$ that obeys two conditions. The first condition is that the vector-valued 2-form

$$T_N(X, Y) = [NX, NY] - N[NX, Y] - N[X, NY] + N^2[X, Y]$$

(called the Nijenhuis torsion of N) vanishes identically. In this case N is termed a "recursion operator." The second condition is that

$$\omega'(X, Y) = \omega(NX, Y)$$

is a closed 2-form. The manifolds where these conditions are fulfilled are called ωN manifolds. On these manifolds, each Hamiltonian vector field X_h is embedded into the distribution

$$D_h = \langle X_h, NX_h, N^2X_h, \ldots \rangle$$

which is the minimal invariant distribution containing X_h. This can be called the Levi-Civita distribution generated by X_h. Experience has shown that D_h is seldom integrable. The cyclic systems of Levi-Civita are, by definition, the generators of the integrable Levi-Civita distributions. Even though this notion is new in classical mechanics, many interesting classical systems display this property.

The aim of this article is to show that the cyclic systems of Levi-Civita are closely related to

separable systems of Jacobi. To this end, the article is organized in four sections, of which the first three clarify the above-mentioned concepts. In the section "ωN manifolds," the idea of ωN manifolds is explained from the viewpoint of bi-Hamiltonian geometry. The section "Cotangent bundles" shows that cotangent bundles provide a large class of ωN manifolds, proving that such manifolds are not rare. Next, two basic examples of cyclic systems of Levi-Civita are presented. Finally, the relation between cyclic systems of Levi-Civita and separable systems of Jacobi is explained briefly.

ωN Manifolds

Let us consider a symplectic manifold (S, ω) with its Hamiltonian vector fields X_h defined by

$$\omega(X_h, \cdot) = -\mathrm{d}h$$

and with the Poisson bracket

$$\{f, g\} = \omega(X_f, X_g)$$

Both the Hamiltonian vector fields and the functions on S form a Lie algebra, and these algebras are homomorphic, since

$$[X_f, X_g] = X_{\{f,g\}}$$

The bi-Hamiltonian geometry is the study of the deformations of the Lie algebras which preserve the above morphism.

We start from the deformations of the Poisson algebra of functions, by replacing the bracket $\{f, g\}$ with the linear pencil

$$\{f, g\}_\epsilon = \{f, g\} + \epsilon\{f, g\}', \quad \epsilon \in \mathbb{R}$$

The problem is to find $\{f, g\}'$ in such a way that the linear pencil satisfies the Jacobi identity for any value of the parameter ϵ. To solve this problem it is convenient to represent the bracket $\{f, g\}'$ in the form

$$\{f, g\}' = \omega'(X_f, X_g)$$

(which is analogous to the standard representation of the Poisson bracket of S) and then to notice that there exists a unique (1, 1) tensor field $N : TS \to TS$ such that

$$\omega'(X_f, X_g) = \omega(NX_f, X_g)$$

Due to the skew-symmetry of ω', the tensor field N must satisfy the condition

$$\omega(NX_f, X_g) = \omega(X_f, NX_g)$$

To the first order in ϵ, the Jacobi identity on $\{f, g\}_\epsilon$ gives

$$\{\{f, g\}, h\}' + \{\{f, g\}', h\} + \text{cyclic permutations} = 0$$

This condition entails a constraint on ω'. One can readily check that ω' must be a closed 2-form:

$$\mathrm{d}\omega' = 0$$

In turn, this constraint imposes a condition on N. The translation of the closure of ω' on N is

$$[NX_f, X_g] + [X_f, NX_g] - N[X_f, X_g] = X_{\{f,g\}'}$$

To the second order in ϵ, the Jacobi identity on $\{f, g\}_\epsilon$ gives

$$\{\{f, g\}', h\}' + \text{cyclic permutations} = 0$$

entailing the condition

$$[NX_f, NX_g] = NX_{\{f,g\}'}$$

on N. Thus, the Jacobi identity is satisfied at any order in ϵ if and only if N is torsion free and ω' is a closed 2-form. Hence, according to the definition given in the "Introduction," the manifold S is an ωN manifold.

It may be of interest to notice that the bracket

$$[X, Y]_N = [NX, Y] + [X, NY] - N[X, Y]$$

is a new (deformed) commutator on vector fields, since the torsion of N vanishes. The same is also true for

$$[X, Y]_\epsilon = [X, Y] + \epsilon[X, Y]_N$$

since the torsion of $(\mathrm{Id} + \epsilon N)$ vanishes too. Therefore, one can write

$$[X_f, X_g]_\epsilon = X_{\{f,g\}_\epsilon}$$

This formula shows that this process of deformation is rigid. For each change of the Poisson bracket, there is a deformation of the commutator of vector fields such that the basic correspondence between functions and Hamiltonian vector fields, established by the symplectic form ω, remains a Lie algebra morphism.

The same phenomenon can be observed in connection with the definition of Hamiltonian vector field. If one introduces the pencil of 2-forms

$$\omega_\epsilon = \omega + \epsilon\omega'$$

and the pencil of derivations

$$\mathrm{d}_\epsilon = \mathrm{d} + \epsilon\mathrm{d}_N$$

where d_N is the derivation of type d and degree 1 canonically associated with N according to the

theory of graded derivations of Frölicher and Nijenhuis, one can prove that

$$\mathrm{d}_\epsilon^2 = 0, \qquad \mathrm{d}_\epsilon \omega_\epsilon = 0$$

and that

$$\omega_\epsilon(X_h, \cdot) = -\mathrm{d}_\epsilon h$$

This means that, on an ωN manifold, the symplectic form ω and the de Rham differential d are deformed in such a way that the basic relation between functions and Hamiltonian vector fields established by ω holds true.

Cotangent Bundles

Cotangent bundles are a source of examples of ωN manifolds. The construction begins on the base manifold Q. For any $(1,1)$ tensor field $L : TQ \to TQ$ with vanishing Nijenhuis torsion, one constructs the deformed Liouville 1-form

$$\theta' = \sum_{i=1}^{n} y_i L^*(\mathrm{d}x^i)$$

and its exterior derivative

$$\omega' = \mathrm{d}\theta'$$

It can be proved that ω' satisfies the conditions explained in the previous section, and conclude that T^*Q, endowed with the pencil of 2-form $\omega_\epsilon = \omega +_\epsilon \omega'$, is an ωN manifold.

A subclass of these structures merits attention. It is related to the polynomials

$$s(\lambda) = \lambda^n - \left(s_1 \lambda^{n-1} + s_2 \lambda^{n-2} + \cdots + s_n\right)$$

the coefficients of which are functions on Q satisfying the condition

$$\mathrm{d}s_1 \wedge \mathrm{d}s_2 \wedge \cdots \wedge \mathrm{d}s_n \neq 0$$

(almost) everywhere on Q. Moreover, it is convenient to assume that the roots $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ of $s(\lambda)$ are distinct and real, so that they are functionally independent and can be used as coordinates on Q. Therefore, the choice of $s(\lambda)$ is equivalent to fix a special system of coordinates on Q, as it happens in $\mathbb{R}^3$ when one introduces the elliptical coordinates as the roots of the polynomial

$$s(\lambda) = (\lambda - a)(\lambda - b)(\lambda - c) \times \left(1 + \frac{x^2}{\lambda - a} + \frac{y^2}{\lambda - b} + \frac{z^2}{\lambda - c}\right)$$

The peculiarity of this situation is that there exists a unique recursion operator $L : TQ \to TQ$ whose characteristic polynomial is $s(\lambda)$. Thus, the choice of $s(\lambda)$ also determines an ωN structure on T^*Q according to the previous prescription. The conclusion is that there is a relation between pencils of Poisson brackets on T^*Q and coordinate systems on Q. This relation is the clue to understand the geometry of separable systems of Jacobi.

Cyclic Systems of Levi-Civita

The systems of coupled harmonic oscillators are the first example of cyclic systems of Levi-Civita. Let us consider, for simplicity, a system formed by only two particles, with masses m_1 and m_2, moving on a line under the action of an internal elastic force. The Lagrangian of the system is

$$L = \tfrac{1}{2}\left(m_1 \dot{x}_1^2 + m_2 \dot{x}_2^2\right) - \tfrac{1}{2} k (x_1 - x_2)^2$$

and the equations of motion are

$$M\ddot{x} + Kx = 0, \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

where

$$M = \begin{pmatrix} m_1 & 0 \\ 0 & m_2 \end{pmatrix}, \qquad K = \begin{pmatrix} k & -k \\ -k & k \end{pmatrix}$$

Under a change of coordinates, the entries of the matrices M and K obey the transformation law of the components of a second-order covariant tensor. Therefore, the entries of the matrix $L = M^{-1}K$ are the components of a tensor field of type $(1,1)$ on $\mathbb{R}^2$. The defining equations of the associated endomorphism $L : T\mathbb{R}^2 \to \mathbb{R}^2$ are

$$L^*(\mathrm{d}x_1) = \omega_1^2(\mathrm{d}x_2 - \mathrm{d}x_1)$$
$$L^*(\mathrm{d}x_2) = \omega_2^2(\mathrm{d}x_1 - \mathrm{d}x_2)$$

if $\omega_1^2 = k/m_1$ and $\omega_2^2 = k/m_2$, and these equations clearly show that L is torsion free. The same argument holds for any system of coupled harmonic oscillators. Therefore, the cotangent bundle associated with any system of coupled harmonic oscillators is an ωN manifold.

To compute the tensor field N in our example, one has to follow the prescription, passing from

$$\theta' = \left(\omega_1^2 y_1 - \omega_2^2 y_2\right)(\mathrm{d}x_2 - \mathrm{d}x_1)$$

to

$$\omega' = \left(\omega_1^2\, \mathrm{d}y_1 - \omega_2^2\, \mathrm{d}y_2\right) \wedge (\mathrm{d}x_2 - \mathrm{d}x_1)$$

and to

$$N\left(\frac{\partial}{\partial x_1}\right) = \omega_1^2 \frac{\partial}{\partial x_1} - \omega_2^2 \frac{\partial}{\partial x_2}$$
$$N\left(\frac{\partial}{\partial x_2}\right) = -\omega_1^2 \frac{\partial}{\partial x_1} + \omega_2^2 \frac{\partial}{\partial x_2}$$
$$N\left(\frac{\partial}{\partial y_1}\right) = \omega_1^2 \left(\frac{\partial}{\partial y_1} - \frac{\partial}{\partial y_2}\right)$$
$$N\left(\frac{\partial}{\partial y_2}\right) = \omega_2^2 \left(-\frac{\partial}{\partial y_1} + \frac{\partial}{\partial y_2}\right)$$

The Levi-Civita distribution D_h is therefore spanned by the vector fields

$$X_h = k\left[\frac{y_1}{\omega_1^2}\frac{\partial}{\partial x_1} + \frac{y_2}{\omega_2^2}\frac{\partial}{\partial x_2} + (x_2 - x_1)\left(\frac{\partial}{\partial y_1} - \frac{\partial}{\partial y_2}\right)\right]$$
$$NX_h = k\left[\left(y_1 - \frac{\omega_1^2}{\omega_2^2}y_2\right)\frac{\partial}{\partial x_1} + \left(y_2 - \frac{\omega_2^2}{\omega_1^2}y_1\right)\frac{\partial}{\partial x_2} + (\omega_1^2 + \omega_2^2)(x_2 - x_1)\left(\frac{\partial}{\partial y_1} - \frac{\partial}{\partial y_2}\right)\right]$$

related to the Hamiltonian

$$h = \frac{y_1^2}{2m_1} + \frac{y_2^2}{2m_2} + \frac{1}{2}k(x_1 - x_2)^2$$

of the system of coupled oscillators. Since $[X_h, NX_h] = 0$, the distribution is integrable; therefore, the system is a cyclic system of Levi-Civita. This property holds for any system of coupled harmonic oscillators. It will be apparent at the end of this article that this result is due to the eigenvectors of L defining the separation coordinates of the coupled oscillators.

The second and final example of cyclic systems of Levi-Civita is the Neumann system, that is, the anisotropic harmonic oscillator on the sphere S^2, whose Lagrangian is

$$L = \tfrac{1}{2}m(\dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2) - \tfrac{1}{2}(a_1x_1^2 + a_2x_2^2 + a_3x_3^2)$$

with the constraint

$$x_1^2 + x_2^2 + x_3^2 = 1$$

This constraint can be avoided by using the first two Cartesian coordinates (x_1, x_2) as local coordinates on S^2. The Hamiltonian of the system can then be written in the form

$$h = \tfrac{1}{2}(1 + x_1^2)y_1^2 - x_1x_2y_1y_2 + \tfrac{1}{2}(1 + x_2^2)y_2^2 + \tfrac{1}{2}(a_1 - a_3)x_1^2 + \tfrac{1}{2}(a_2 - a_3)x_2^2$$

where, for simplicity, $m = 1$. Formally one is back in $\mathbb{R}^2$ as in the previous example, but the nonlinearity of the equations of motion hinders us to readily see the appropriate recursion operator $L : T\mathbb{R}^2 \to T\mathbb{R}^2$ to be used to construct the ωN structure on $T^*\mathbb{R}^2$. Let us however recall that according to Neumann, the system is separable in elliptical spherical (also called spheroconical) coordinates, defined as the roots of the restriction to S^2 of the polynomial

$$s(\lambda) = (\lambda - a)(\lambda - b)(\lambda - c)\left(\frac{x_1^2}{\lambda - a} + \frac{x_2^2}{\lambda - b} + \frac{x_3^2}{\lambda - c}\right)$$
$$= \lambda^2 - (s_1\lambda + s_2)$$

Let us, therefore, use this polynomial to construct the unique recursion operator L having $s(\lambda)$ as its characteristic polynomial. It is given by

$$L^*(\mathrm{d}s_1) = \mathrm{d}s_2 + s_1\,\mathrm{d}s_1$$
$$L^*(\mathrm{d}s_2) = s_2\,\mathrm{d}s_1$$

or, after a brief computation, by

$$L^*(\mathrm{d}x_1) = a_1\,\mathrm{d}x_1 - x_1\,\mathrm{d}\left[\tfrac{1}{2}(a_1 - a_3)x_1^2 + \tfrac{1}{2}(a_2 - a_3)x_2^2\right]$$
$$L^*(\mathrm{d}x_2) = a_2\,\mathrm{d}x_2 - x_2\,\mathrm{d}\left[\tfrac{1}{2}(a_1 - a_3)x_1^2 + \tfrac{1}{2}(a_2 - a_3)x_2^2\right]$$

The situation stays the same as in the previous example. Accordingly, the recursion operator N on $T^*\mathbb{R}^2$ is now given by

$$N^*\,\mathrm{d}x_1 = a_1\,\mathrm{d}x_1 - x_1\,\mathrm{d}f$$
$$N^*\,\mathrm{d}x_2 = a_2\,\mathrm{d}x_2 - x_2\,\mathrm{d}f$$
$$N^*\,\mathrm{d}y_1 = a_1\,\mathrm{d}y_1 - (a_1 - a_3)x_1\,\mathrm{d}g + y_1\,\mathrm{d}f$$
$$N^*\,\mathrm{d}y_2 = a_2\,\mathrm{d}y_2 - (a_2 - a_3)x_2\,\mathrm{d}g + y_2\,\mathrm{d}f$$

where the shorthand notations

$$f = \tfrac{1}{2}(a_1 - a_3)x_1^2 + \tfrac{1}{2}(a_2 - a_3)x_2^2$$
$$g = x_1y_1 + x_2y_2$$

have been used. The derivation d_N, associated with N, is accordingly defined by

$$\mathrm{d}_N x_1 = N^*\,\mathrm{d}x_1 = \left[a_1 + (a_3 - a_1)x_1^2\right]\mathrm{d}x_1 + (a_3 - a_2)x_1x_2\,\mathrm{d}x_2$$
$$\mathrm{d}_N x_2 = N^*\,\mathrm{d}x_2 = (a_3 - a_1)x_1x_2\,\mathrm{d}x_1 + \left[a_2 + (a_3 - a_2)x_2^2\right]\mathrm{d}x_2$$
$$\mathrm{d}_N y_1 = N^*\,\mathrm{d}y_1 = \left[(a_3 - a_1)x_1y_2 - (a_3 - a_2)x_2y_1\right]\mathrm{d}x_2 + \left[a_1 + (a_3 - a_1)x_1^2\right]\mathrm{d}y_1 + (a_3 - a_1)x_1x_2\,\mathrm{d}y_2$$
$$\mathrm{d}_N y_2 = N^*\,\mathrm{d}y_2 = \left[(a_3 - a_2)x_2y_1 - (a_3 - a_1)x_1y_2\right]\mathrm{d}x_1 + (a_3 - a_2)x_1x_2\,\mathrm{d}y_1 + \left[a_2 + (a_3 - a_2)x_2^2\right]\mathrm{d}y_2$$

on the coordinate functions. Recalling that d_N anticommutes with d, one can then easily check the condition

$$\mathrm{d}\mathrm{d}_N h = \mathrm{d}s_1 \wedge \mathrm{d}h$$

where s_1 is the first coefficient of the polynomial defining the elliptical spherical coordinates, and h is the Hamiltonian of the Neumann system. By the Frobenius theorem, this equation alone entails the integrability of the distribution D_h, without the need of computing X_h, NX_h, and their commutator $[X_h, NX_h]$. Thus, it can be concluded that the Neumann system too is a cyclic system of Levi-Civita, and that the recursion operator N, generating the distribution D_h, is closely related to the polynomial defining the separation coordinates of the Neumann system.

Separable System of Jacobi

In 1838, Jacobi noticed that the Hamilton–Jacobi equation

$$h\left(x_1, x_2, \ldots, x_n, \frac{\partial W}{\partial x_1}, \ldots, \frac{\partial W}{\partial x_n}\right) = e$$

of many Hamiltonian systems splits owing to an appropriate choice of coordinates in a set of ordinary differential equations. On account of this property, these systems have been called separable. In 1904, Levi-Civita gave a first partial characterization of separable Hamiltonians by means of his separability conditions. In a letter addressed to Stäckel, he proved that h is separable in a preassigned system of canonical coordinates if and only if the conditions

$$\frac{\partial^2 h}{\partial x_j \partial x_k}\frac{\partial h}{\partial y_j}\frac{\partial h}{\partial y_k} - \frac{\partial^2 h}{\partial x_j \partial y_k}\frac{\partial h}{\partial y_j}\frac{\partial h}{\partial x_k} - \frac{\partial^2 h}{\partial y_j \partial x_k}\frac{\partial h}{\partial x_j}\frac{\partial h}{\partial y_k} + \frac{\partial^2 h}{\partial y_j \partial y_k}\frac{\partial h}{\partial x_j}\frac{\partial h}{\partial x_k} = 0$$

are satisfied by h. One must notice the nontensorial character of these conditions; they hold only in a specific coordinate system, and if the coordinates are changed, it is not possible to reconstruct the form of the separability conditions in the new coordinates. The nontensorial character is the major drawback of the separability conditions of Levi-Civita, making them practically useless in the search of separation coordinates.

The contact between the theory of separable system of Jacobi and the theory of cyclic systems of Levi-Civita rests on two occurrences. The first is the form of the integrability conditions of the distribution D_h generated by any vector field X_h on an ωN manifold. Exploiting the Frobenius integrability conditions and the properties of the differential operator d_N associated with the recursion operator N, it can be proved that D_h is integrable if and only if the 2-form $\mathrm{dd}_N h$ vanishes on D_h:

$$\mathrm{dd}_N h = 0 \quad \text{on} \qquad D_h$$

Suppose now that the dimension of D_h is maximal, that is, equal to $n = (1/2)\dim S$. Then D_h is spanned by the n vector fields $(X_h, NX_h, \ldots, N^{n-1}X_h)$, and the vanishing condition of $\mathrm{dd}_N h$ on D_h turns out to be equivalent to

$$\mathrm{dd}_N h(N^j X_h, N^k X_h) = 0$$

for any value of j and k from 0 to $n-1$. Thus, the number of separability conditions of h and the number of integrability conditions of D_h are equal. This circumstance strongly suggests that the two sets of conditions are related. The nontensorial character of the Levi-Civita conditions, compared with the tensorial character of the integrability conditions of D_h, further suggests that the former should be the evaluation of the latter in a specific system of coordinates. These coordinates are the "normal coordinates" of an ωN manifold, that will be introduced in the following.

Assume that the minimal polynomial of N has real and distinct roots $(l_1, \ldots, l_n)$. In this case, the ωN manifold is said to be semisimple. A two-dimensional eigenspace is associated with each root l_k. Let us consider the distribution E_k spanned by all the eigenvectors of N, except those associated with l_k. Since N is torsion free, each distribution E_k is integrable. Let us fix the attention on one of these distributions. It turns out that its leaves are symplectic submanifolds of codimension 2. So they are the level surfaces of a pair of (local) functions which are not in involution. By collecting together the pairs of functions associated with the n distributions $(E_1, \ldots, E_n)$, one obtains, at the end, a coordinate system $(\lambda_1, \mu_1, \lambda_2, \mu_2, \ldots, \lambda_n, \mu_n)$ on S. Moreover, these functions can be chosen in such a way to form a system of canonical coordinates. The final result is that, on a semisimple ωN manifold, one can construct a coordinate system such that

$$\omega = \sum_{j=1}^{n} \mathrm{d}\mu_j \wedge \mathrm{d}\lambda_j$$

and

$$N^*(\mathrm{d}\lambda_j) = l_j \mathrm{d}\lambda_j$$
$$N^*(\mathrm{d}\mu_j) = l_j \mathrm{d}\mu_j$$

These coordinates are called the normal coordinates (or sometimes, the Darboux–Nijenhuis coordinates) of the ωN manifold. One can prove that the separability conditions of Levi-Civita are the integrability

conditions of D_h, written in normal coordinates. This result allows us to claim that the cyclic systems of Levi-Civita on semisimple ωN manifolds are all separable.

The reverse is also true. As has already been shown in the example of the Neumann system, a given separable system of Jacobi can be associated with a recursion operator N in such a way that its phase space (with the possible exclusion of a singular locus) becomes an ωN manifold, and the Hamiltonian vector field X_h becomes a cyclic system of Levi-Civita. A new interpretation of the process of separation of variables follows from this result. Indeed, to find separation coordinates for a given system on a symplectic manifold S is equivalent to deforming the Poisson bracket of S into a pencil

$$\{f,g\}_\epsilon = \{f,g\} + \epsilon\{f,g\}'$$

in such a way that the recursion operator N defining the pencil $\{f,g\}_\epsilon$ generates, with X_h, an integrable distribution D_h. Therefore, classical mechanics is deeply entangled with the theory of recursion operators, even if the insistence on the use of separation coordinates has hidden this factor for a long time.

See also: Bi-Hamiltonian Methods in Soliton Theory; Classical *r*-Matrices, Lie Bialgebras, and Poisson Lie Groups; Integrable Systems and Algebraic Geometry; Integrable Systems and Recursion Operators on Symplectic and Jacobi Manifolds; Integrable Systems: Overview; Multi-Hamiltonian Systems; Separation of Variables for Differential Equations; Solitons and Kac–Moody Lie Algebras.

Further Reading

Dubrovin BA, Krichever IM, and Novikov SP (2001) Integrable systems I. In: Arnol'd VI (ed.) *Encyclopaedia of Mathematical Sciences. Dynamical Systems IV*, pp. 177–332. Berlin: Springer.

Jacobi CGJ (1996) *Vorlesungen ber analytische Mechanik*, Deutsche Mathematiker Vereinigung, Freiburg. Braunschweig: Friedrich Vieweg and Sohn.

Ivan K, Michor PW, and Slovák J (1993) *Natural Operations in Differential Geometry*. Berlin: Springer.

Kalnins EG (1986) *Separation of Variables for Riemannian Spaces of Constant Curvature*. New York: Wiley.

Krasilshchik IS and Kersten PHM (2000) *Symmetries and Recursion Operators for Classical and Supersymmetric Differential Equations*. Dordrecht: Kluwer.

Magri F, Falqui G, and Pedroni M (2003) The method of Poisson pairs in the theory of nonlinear PDEs. In: Conte R, Magri F, Musette M, Satsuma J, and Winternitz P (eds.) *Direct and Inverse Methods in Nonlinear Evolution Equations*, Lecture Notes in Physics, vol. 632, pp. 85–136. Berlin: Springer.

Miller W (1977) *Symmetry and Separation of Variables*. Reading, MA–London–Amsterdam: Addison-Wesley.

Olver PJ (1993) *Applications of Lie Groups to Differential Equations*, 2nd edn. New York: Springer.

Pars LA (1965) *A Treatise on Analytical Dynamics*. London: Heinemann.

Vaisman I (1994) *Lectures on the Geometry of Poisson Manifolds*. Basel: Birkhäuser.

Vilasi G (2001) *Hamiltonian Dynamics*. River Edge, NJ: World Scientific.

Yano K and Ishihara S (1973) *Tangent and Cotangent Bundles: Differential Geometry*. New York: Dekker.

Reflection Positivity and Phase Transitions

Y Kondratiev, Universität Bielefeld, Bielefeld, Germany
Y Kozitsky, Uniwersytet Marii Curie-Sklodowskiej, Lublin, Poland

Phase Transitions in Lattice Systems

Introduction

Phase transitions are among the main objects of equilibrium statistical mechanics, both classical and quantum. There exist several approaches to the description of these phenomena. Their common point is that the macroscopic behavior of a statistical mechanical model can be different at the same values of the model parameters. This corresponds to the multiplicity of equilibrium phases, each of which has its own properties. In the mathematical formulation, models are defined by interaction potentials and equilibrium phases appear as states – positive linear functionals on algebras of observables. In the classical case the states are defined by means of the probability measures which satisfy equilibrium conditions, formulated in terms of the interaction potentials. Such measures are called Gibbs measures and the corresponding states are called Gibbs states. The observables are then integrable functions. In the quantum case the states mostly are introduced by means of the Kubo–Martin–Schwinger condition – a quantum analog of the equilibrium conditions used for classical models. The quantum observables constitute noncommutative von Neumann algebras.

Infinite systems of particles studied in statistical mechanics fall into two main groups. These are continuous systems and lattice systems. In the latter case, particles are attached to the points of various crystalline lattices. In view of the specifics of our subject, in this article we will deal with lattice systems only.

One of the main problems of the mathematical theory of phase transitions is to prove that the Gibbs states of a given model can be multiple, that is, that this model undergoes a phase transition. To solve this problem one has to elaborate corresponding mathematical tools. Typically, at high temperatures (equivalently, for weak interactions), a model, which undergoes a phase transition, has only one Gibbs state. This state inherits all the symmetries possessed by the interaction potentials. At low temperatures this model has multiple Gibbs states, which may lose the symmetries. In this case the phase transition is accompanied by a symmetry breaking. Among the symmetries important in the theory of lattice systems, there is the invariance with respect to the lattice translations. If the Gibbs state of a translation invariant lattice model is unique, it ought to be ergodic with respect to the group of lattice translations. This means in particular that the spacial correlations in this state decay to zero at long distances. Therefore, the lack of the latter property may indicate a phase transition. In a number of lattice models, phase transitions can be established by means of their special property – reflection positivity. The most important consequence of reflection positivity are chessboard (another name checkerboard) estimates, being extended versions of Hölder's inequalities. The proof of a phase transition is then performed either by means of a combination of such estimates and contour methods, or by means of infrared estimates obtained from the chessboard estimates.

In this article we show how to prove phase transitions by means of the infrared estimates for some simple reflection positive models, both classical and quantum. The details on the reflection positivity method in all its versions may be found in the literature listed at the end of the article. There we also provide short bibliographic comments.

Nonergodicity and Infrared Estimates

The following heuristic arguments should give an idea how to establish the nonergodicity of a Gibbs state by means of infrared estimates. Let us consider a classical ferromagnetic translation-invariant model. (Of course, we assume that it possesses Gibbs states, which for models with unbounded spins is a nontrivial property. A particular case of this model is described in more detail in the subsection "Gaussian domination.") This model describes the system of interacting N-dimensional spins $x_\ell \in \mathbb{R}^N$, indexed by the elements $\ell \in \mathbb{Z}^d$ of the d-dimensional simple cubic lattice. The interaction is pairwise, attractive, nearest-neighbor, and invariant with respect to the rotations in $\mathbb{R}^N$. Consider a translation-invariant Gibbs state of this model, which always exists. Let $K(\ell, \ell'), \ell, \ell' \in \mathbb{Z}^d$, be the expectation of the scalar product $(x_\ell, x_{\ell'})$ of spins in this state. Then $K(\ell, \ell')$ is also translation invariant and hence may be written as

$$K(\ell,\ell') = \frac{1}{(2\pi)^d}\int_{(-\pi,\pi]^d} \widehat{K}(p)\mathrm{e}^{\mathrm{i}(p,\ell-\ell')}\,\mathrm{d}p, \quad \mathrm{i}=\sqrt{-1} \qquad [1]$$

where the generalized function $\widehat{K}$ is defined by the Fourier series

$$\widehat{K}(p) = \sum_{\ell'\in\mathbb{Z}^d} K(\ell,\ell')\mathrm{e}^{-\mathrm{i}(p,\ell-\ell')}, \quad p \in (-\pi,\pi]^d \qquad [2]$$

As the model is ferromagnetic, $K(\ell, \ell') \geq 0$. The Gibbs state is nonergodic if $K(\ell, \ell')$ does not tend to zero as $|\ell - \ell'| \to \infty$. In this case $\widehat{K}$ should be singular at $p = 0$. Set

$$\widehat{K}(p) = (2\pi)^d \lambda \delta(p) + g(p) \qquad [3]$$

where $\delta(p)$ is the Dirac δ-function and $g(p)$ is regular at $p = 0$. Then the Gibbs state is nonergodic if $\lambda \neq 0$. Suppose we know that $g(p) \geq 0$ and that the following two estimates hold. The first one is

$$g(p) \leq \gamma / J|p|^2, \quad p \neq 0 \qquad [4]$$

where $\gamma > 0$ is a constant and $J > 0$ is the interaction intensity multiplied by the inverse temperature β. This is the infrared estimate. The second estimate is

$$K(\ell,\ell) \geq \varkappa > 0 \qquad [5]$$

where $\varkappa$ is independent of J. By these estimates and [1], [2], we get

$$\lambda \geq \varkappa - \frac{\gamma}{(2\pi)^d J}\int_{(-\pi,\pi]^d} \frac{\mathrm{d}p}{|p|^2} \qquad [6]$$

For $d \geq 3$, the latter integral exists; hence, $\lambda > 0$ for J large enough, which means that the state we consider is nonergodic.

The quantum case is more involved. The infrared bounds are obtained not for functions like $\widehat{K}(p)$ but for the so-called Duhamel two-point functions. Then one has to prove a number of additional statements, which finally lead to the proof of the result desired. In the section on reflection positivity in quantum systems we indicate how to do this for a simple quantum spin model.

Reflection Positivity and Phase Transitions in Classical Systems

We begin by studying reflection positive (RP) functionals. Gibbs states of RP models are such functionals.

Reflection Positive Functionals

Let Λ be a finite set of indices consisting of an even number $|\Lambda|$ of elements, which label real variables $x_\ell, \ell \in \Lambda$. For $\Lambda' \subseteq \Lambda$, we write $x_{\Lambda'} = (x_\ell)_{\ell \in \Lambda'} \in \mathbb{R}^{|\Lambda'|}$. Suppose we are given a bijection $\rho : \Lambda \to \Lambda, \rho \circ \rho = \mathrm{id}$, such that the set Λ falls into two disjoint parts $\Lambda_\pm$ with the property $\rho : \Lambda_+ \to \Lambda_-$. Therefore, $|\Lambda_+| = |\Lambda_-|$, and the map ρ may be regarded as a reflection. For $x_\Lambda \in \mathbb{R}^{|\Lambda|}$, we set $\rho(x_\Lambda) = (x_{\rho(\ell)})_{\ell \in \Lambda}$. Now let $\mathcal{A}$ be an algebra of functions $A : \mathbb{R}^{|\Lambda|} \to \mathbb{R}$. Then we define the map $\vartheta : \mathcal{A} \to \mathcal{A}$ by setting

$$\vartheta(A)(x_\Lambda) = A(\rho(x_\Lambda)) \tag{7}$$

Clearly, for all $A, B \in \mathcal{A}$ and $\xi, \eta \in \mathbb{R}$,

$$\begin{aligned} \vartheta(\xi A + \eta B) &= \xi\vartheta(A) + \eta\vartheta(B) \\ \vartheta(A \cdot B) &= \vartheta(A) \cdot \vartheta(B) \end{aligned} \tag{8}$$

By $\mathcal{A}^+$ (respectively, $\mathcal{A}^-$), we denote the subalgebra of $\mathcal{A}$ consisting of functions dependent on x_{Λ_+} (respectively, x_{Λ_-}). Then $\vartheta(\mathcal{A}^+) = \mathcal{A}^-$ and $\vartheta \circ \vartheta = \mathrm{id}$.

Definition 1 A linear functional $\phi : \mathcal{A} \to \mathbb{R}$ is called RP with respect to the maps ρ and ϑ, if

$$\forall A \in \mathcal{A}^+ : \phi[A\vartheta(A)] \geq 0 \tag{9}$$

Example 2 Let χ be a Borel measure on the real line (not necessarily positive), with respect to which all real polynomials are integrable. Let also $\mathcal{A}$ be the algebra of all real-valued polynomials on $\mathbb{R}^{|\Lambda|}$, $|\Lambda|$ being even. Finally, let ρ and ϑ be any of the maps with the properties described above. Then the functional

$$\begin{aligned} \phi(A) &= \int_{\mathbb{R}^{|\Lambda|}} A(x_\Lambda)\, \mathrm{d}\chi_\Lambda(x_\Lambda) \\ \mathrm{d}\chi_\Lambda(x_\Lambda) &= \prod_{\ell \in \Lambda} \mathrm{d}\chi(x_\ell) \end{aligned} \tag{10}$$

is RP. Indeed, let $F : \mathbb{R}^{|\Lambda|/2} \to \mathbb{R}$ be such that $A(x_\Lambda) = F(x_{\Lambda_+})$. Then

$$\begin{aligned} \phi[A\vartheta(A)] &= \int F(x_{\Lambda_+}) \prod_{\ell \in \Lambda_+} \mathrm{d}\chi(x_\ell) \cdot \int F(x_{\Lambda_-}) \prod_{\ell \in \Lambda_-} \mathrm{d}\chi(x_\ell) \\ &= \left[\int F(x_{\Lambda_+}) \prod_{\ell \in \Lambda_+} \mathrm{d}\chi(x_\ell) \right]^2 \geq 0 \end{aligned}$$

In the above example the multiplicative structure of the measure χ_Λ is crucial. It results in the positivity of ϕ with respect to all reflections. If one has just one such reflection, the measure which defines ϕ may be decomposable onto two measures only. Let $\Lambda, \mathcal{A}, \rho$, and ϑ be as above. Consider a Borel measure ν on $\mathbb{R}^{|\Lambda|/2}$ such that every real-valued polynomial on $\mathbb{R}^{|\Lambda|/2}$ is ν-integrable.

Proposition 3 *The functional*

$$\phi(A) = \int_{\mathbb{R}^{|\Lambda|}} A(x_\Lambda)\, \mathrm{d}\nu(x_{\Lambda_+})\, \mathrm{d}\nu(x_{\Lambda_-}) \tag{11}$$

is RP.

In both these examples the states are symmetric, that is,

$$\phi[A\vartheta(B)] = \phi[B\vartheta(A)], \quad \text{for all } A, B \in \mathcal{A}^+ \tag{12}$$

In the sequel we shall suppose that all RP functionals possess this property. Therefore, RP functionals obey a Cauchy–Schwarz type inequality.

Lemma 4 *If ϕ is RP, then for any $A, B \in \mathcal{A}^+$,*

$$\{\phi[A\vartheta(B)]\}^2 \leq \phi[A\vartheta(A)] \cdot \phi[B\vartheta(B)] \tag{13}$$

Proof For $\xi \in \mathbb{R}$, by [8] we have

$$\begin{aligned} &\phi[(A + \xi B)\vartheta(A + \xi B)] \\ &\quad = \phi[(A + \xi B)(\vartheta(A) + \xi\vartheta(B))] \geq 0 \end{aligned}$$

Since ϕ is linear, the latter can be written as a 3-nomial, whose positivity for all $\xi \in \mathbb{R}$ is equivalent to [13]. □

Now let an RP functional ϕ be such that for

$$A, B, C_1, \ldots, C_m, D_1, \ldots, D_m \in \mathcal{A}^+$$

there exists

$$\phi\left[\exp\left(A + \vartheta(B) + \sum_{i=1}^m C_i\vartheta(D_i)\right)\right]$$

and that the series

$$\begin{aligned} &\sum_{n_1, \ldots, n_m = 0}^{\infty} \frac{1}{n_1! \cdots n_m!} \cdot \phi\{[C_1\vartheta(C_1)]^{n_1} \cdots [C_m\vartheta(C_m)]^{n_m} \\ &\quad \times \exp[A + \vartheta(B)]\} \end{aligned} \tag{14}$$

as well as the one with all C_is replaced by D_is converge absolutely.

Lemma 5 *Let the functional ϕ and the functions $A, B, C_i, D_i, i = 1, \ldots, m$, be as above. Then*

$$\begin{aligned} &\left\{\phi\left[\exp\left(A + \vartheta(B) + \sum_{i=1}^m C_i\vartheta(D_i)\right)\right]\right\}^2 \\ &\quad \leq \phi\left[\exp\left(A + \vartheta(A) + \sum_{i=1}^m C_i\vartheta(C_i)\right)\right] \\ &\quad \times \phi\left[\exp\left(B + \vartheta(B) + \sum_{i=1}^m D_i\vartheta(D_i)\right)\right] \end{aligned} \tag{15}$$

Proof By the above assumptions

$$\phi\left[\exp\left(A+\vartheta(B)+\sum_{i=1}^{m}C_i\vartheta(D_i)\right)\right]$$
$$=\phi\left[F\vartheta(G)\exp\left(\sum_{i=1}^{m}C_i\vartheta(D_i)\right)\right]$$
$$=\sum_{n_1,\dots,n_m=0}^{\infty}\frac{1}{n_1!\cdots n_m!}\cdot\phi[F\vartheta(G)[C_1\vartheta(D_1)]^{n_1}\cdots$$
$$\times[C_m\vartheta(D_m)]^{n_m}] \qquad [16]$$

where $F=\mathrm{e}^A$, $G=\mathrm{e}^B$. Then by [13] and the Cauchy–Schwarz inequality for sums we get

RHS[16]
$$\le\sum_{n_1,\dots,n_m=0}^{\infty}\left\{\frac{1}{n_1!\cdots n_m!}\cdot\phi[F\vartheta(F)[C_1\vartheta(C_1)]^{n_1}\cdots[C_m\vartheta(C_m)]^{n_m}]\right\}^{1/2}$$
$$\times\left\{\frac{1}{n_1!\cdots n_m!}\cdot\phi[G\vartheta(G)[D_1\vartheta(D_1)]^{n_1}\cdots[D_m\vartheta(D_m)]^{n_m}]\right\}^{1/2}$$
$$\le\left\{\sum_{n_1,\dots,n_m=0}^{\infty}\frac{1}{n_1!\cdots n_m!}\cdot\phi[F\vartheta(F)[C_1\vartheta(C_1)]^{n_1}\cdots[C_m\vartheta(C_m)]^{n_m}]\right\}^{1/2}$$
$$\times\left\{\sum_{n_1,\dots,n_m=0}^{\infty}\frac{1}{n_1!\cdots n_m!}\cdot\phi[G\vartheta(G)[D_1\vartheta(D_1)]^{n_1}\cdots[D_m\vartheta(D_m)]^{n_m}]\right\}^{1/2}$$
$$=\left\{\phi\left[\exp\left(A+\vartheta(A)+\sum_{i=1}^{m}C_i\vartheta(C_i)\right)\right]\right\}^{1/2}$$
$$\times\left\{\phi\left[\exp\left(B+\vartheta(B)+\sum_{i=1}^{m}D_i\vartheta(D_i)\right)\right]\right\}^{1/2}$$

which yields [15]. □

Main Estimate

Let Δ be a finite set and Δ' be its nonempty subset. Let also μ and ν be finite Borel measures on $\mathbb{R}^{N|\Delta|}, N\in\mathbb{N}$. For vectors $b,c\in\mathbb{R}^N$, by (b,c) and $|b|,|c|$ we denote their scalar product $\sum_{k=1}^{N}b^{(k)}c^{(k)}$ and the corresponding norms, respectively. By x_Δ we denote $(x_\ell)_{\ell\in\Delta}, x_\ell\in\mathbb{R}^N$; hence, $x_\Delta\in\mathbb{R}^{N|\Delta|}$.

Lemma 6 *Let the sets Δ,Δ' and the measures μ,ν be as above. Then for every $(a_\ell)_{\ell\in\Delta'}\in\mathbb{R}^{N|\Delta'|}$ and $J\ge 0$,*

$$\left[\int_{\mathbb{R}^{2N|\Delta|}}\exp\left(-\frac{J}{2}\sum_{\ell\in\Delta'}|x_\ell-y_\ell-a_\ell|^2\right)\mathrm{d}\mu(x_\Delta)\mathrm{d}\nu(y_\Delta)\right]^2$$
$$\le\int_{\mathbb{R}^{2N|\Delta|}}\exp\left(-\frac{J}{2}\sum_{\ell\in\Delta'}|x_\ell-y_\ell|^2\right)\mathrm{d}\mu(x_\Delta)\,\mathrm{d}\mu(y_\Delta)$$
$$\times\int_{\mathbb{R}^{2N|\Delta|}}\exp\left(-\frac{J}{2}\sum_{\ell\in\Delta'}|x_\ell-y_\ell|^2\right)\mathrm{d}\nu(x_\Delta)\,\mathrm{d}\nu(y_\Delta) \qquad [17]$$

Proof Take two copies of Δ and denote them by $\Lambda_\pm$. Furthermore, by $\Lambda'_\pm\subset\Lambda_\pm$ we denote the subsets consisting of the elements of $\Delta'\subset\Delta$. For an $\ell\in\Lambda_+$, by $\rho(\ell)$ we denote its counterpart in Λ_-. Then ρ is a reflection and $\rho(\Lambda'_+)=\Lambda'_-$. Let $\Lambda=\Lambda_+\cup\Lambda_-, \Lambda'=\Lambda'_+\cup\Lambda'_-$, and $\mathcal{A}$ be the algebra of all polynomials of $(x_{\Lambda'},y_{\Lambda'})\in\mathbb{R}^{2N|\Lambda'|}$. Note that $x_{\Lambda'}$ may be regarded as the pair $(x_{\Lambda'_+},x_{\Lambda'_-})$. Let $\mathcal{A}^+$ (respectively, $\mathcal{A}^-$) be the subalgebra of $\mathcal{A}$ consisting of the polynomials which depend on $x_{\Lambda'_+},y_{\Lambda'_+}$ (respectively, $x_{\Lambda'_-},y_{\Lambda'_-}$) only. Introduce the measures

$$\mathrm{d}\tilde\mu(x_\Delta)=\exp\left(-\frac{J}{2}\sum_{\ell\in\Delta'}|x_\ell|^2\right)\mathrm{d}\mu(x_\Delta)$$
$$\mathrm{d}\tilde\nu(x_\Delta)=\exp\left(-\frac{J}{2}\sum_{\ell\in\Delta'}|x|_\ell^2\right)\mathrm{d}\nu(x_\Delta)$$

and define the following functional on $\mathcal{A}$:

$$\phi(F)=\int_{\mathbb{R}^{2N|\Delta|}}F(x_{\Lambda'},y_{\Lambda'})\,\mathrm{d}\tilde\mu(x_{\Lambda_+})$$
$$\times\,\mathrm{d}\tilde\nu(y_{\Lambda_+})\,\mathrm{d}\tilde\mu(x_{\Lambda_-})\,\mathrm{d}\tilde\nu(y_{\Lambda_-}) \qquad [18]$$

It has the same structure as the one described by Proposition 3, hence is RP with respect to the map ϑ defined by the reflection ρ. Set

$$\Upsilon_\mu=\int_{\mathbb{R}^{N|\Delta|}}\mathrm{d}\tilde\mu(x_\Delta),\qquad \Upsilon_\nu=\int_{\mathbb{R}^{N|\Delta|}}\mathrm{d}\tilde\nu(y_\Delta) \qquad [19]$$

and

$$A\equiv 0,\qquad B=-J\sum_{\ell\in\Lambda'_+}\left[\frac{1}{2}|a_\ell|^2+(a_\ell,y_\ell)\right]$$
$$C_\ell^{(k)}=\sqrt{J}x_\ell^{(k)},\qquad D_\ell^{(k)}=\sqrt{J}\left(y_\ell^{(k)}+a_\ell^{(k)}\right) \qquad [20]$$
$$\ell\in\Lambda'_+,\quad k=1,\dots,N$$

Then the left-hand side of [17] is

LHS [17]
$$=\frac{1}{(\Upsilon_\mu\Upsilon_\nu)^2}\left|\phi\left[\exp\left(A+\vartheta(B)\right.\right.\right.$$
$$\left.\left.\left.+\sum_{\ell\in\Lambda'_+}\sum_{k=1}^{N}C_\ell^{(k)}\vartheta\left(D_\ell^{(k)}\right)\right)\right]\right|^2 \qquad [21]$$

with ϕ given by [18]. Applying [15] and taking into account [19], we arrive at

LHS [17]
$$\le\frac{1}{(\Upsilon_\mu\Upsilon_\nu)^2}\int_{\mathbb{R}^{2N|\Delta|}}\exp\left(J\sum_{\ell\in\Lambda'_+}x_\ell x_{\rho(\ell)}\right)$$
$$\times\,\mathrm{d}\tilde\mu(x_{\Lambda_+})\,\mathrm{d}\tilde\mu(x_{\Lambda_-})\,\mathrm{d}\tilde\nu(y_{\Lambda_+})\,\mathrm{d}\tilde\nu(y_{\Lambda_-})$$
$$\times\int_{\mathbb{R}^{2N|\Lambda|}}\exp\left(J\sum_{\ell\in\Lambda'_+}y_\ell y_{\rho(\ell)}\right)$$
$$\times\,\mathrm{d}\tilde\mu(x_{\Lambda_+})\,\mathrm{d}\tilde\mu(x_{\Lambda_-})\,\mathrm{d}\tilde\nu(y_{\Lambda_+})\,\mathrm{d}\tilde\nu(y_{\Lambda_-})=\text{RHS [17]}$$

which completes the proof. □

Gaussian Domination

Let Λ be a finite set, $|\Lambda|$ even, and E be a set of unordered pairs of elements of Λ, such that the graph (Λ, E) is connected. If $e \in E$ connects given $\ell, \ell' \in \Lambda$, we write $e = \langle \ell, \ell' \rangle$. We suppose that E contains no loops $\langle \ell, \ell \rangle$. With each $\ell \in \Lambda$ we associate a random N-component vector x_ℓ, called spin. The joint probability distribution of the spins $(x_\ell)_{\ell \in \Lambda}$ is defined by means of the local Gibbs measure

$$d\mu_\Lambda(x_\Lambda) = \frac{1}{Z_\Lambda} \exp\left(-\frac{J}{2} \sum_{\langle \ell, \ell' \rangle \in E} |x_\ell - x_{\ell'}|^2\right) d\chi_\Lambda(x_\Lambda),$$
$$x_\Lambda \in \mathbb{R}^{N|\Lambda|} \qquad [22]$$

Here the measure

$$d\chi_\Lambda(x_\Lambda) = \prod_{\ell \in \Lambda} d\chi(x_\ell) \qquad [23]$$

describes the system if the interaction intensity J equals zero. In general, $J \geq 0$, that is, the model [22], [23] is ferromagnetic. The single-spin measure χ is a probability measure on $\mathbb{R}^N$ and

$$Z_\Lambda = \int_{\mathbb{R}^{N|\Lambda|}} \exp\left(-\frac{J}{2} \sum_{\langle \ell, \ell' \rangle \in E} |x_\ell - x_{\ell'}|^2\right) d\chi_\Lambda(x_\Lambda) \qquad [24]$$

is the partition function. Set

$$Z_\Lambda(h) = \int_{\mathbb{R}^{N|\Lambda|}} \exp\left(-\frac{J}{2} \sum_{\langle \ell, \ell' \rangle \in E} |x_\ell - x_{\ell'} - h_{\ell\ell'}|^2\right) \times d\chi_\Lambda(x_\Lambda) \qquad [25]$$

where $h_{\ell\ell'} = h_{\ell'\ell} \in \mathbb{R}^N$, $\langle \ell, \ell' \rangle \in E$.

Definition 7 The model [22]–[23] admits Gaussian domination if for all $h = (h_{\ell\ell'})_{\langle \ell, \ell' \rangle \in E}$,

$$Z_\Lambda(h) \leq Z_\Lambda(0) \qquad [26]$$

We prove that our model admits Gaussian domination if the graph satisfies the following:

Assumption 8 The set of edges E can be decomposed

$$E = \bigcup_{n=1}^{m} E_n, \qquad E_n \bigcap E_{n'} = \emptyset, \quad \text{if } n \neq n' \qquad [27]$$

in such a way that for every $n = 1, \ldots, m$, the graph $(\Lambda, E \backslash E_n)$ is disconnected and falls into two connected components, $(\Lambda_+^{(n)}, E_+^{(n)})$ and $(\Lambda_-^{(n)}, E_-^{(n)})$, which are isomorphic. This means that there exists a bijection $\rho_n : \Lambda \to \Lambda$, $\rho_n \circ \rho_n = \mathrm{id}$, such that $\rho_n(\Lambda_+^{(n)}) = \Lambda_-^{(n)}$ and $\langle \rho_n(\ell), \rho_n(\ell') \rangle \in E_-^{(n)}$ whenever $\langle \ell, \ell' \rangle \in E_+^{(n)}$. Finally, we assume that if $\langle \ell, \ell' \rangle \in E_n$ and $\ell \in \Lambda_+^{(n)}$, then $\rho_n(\ell) = \ell'$.

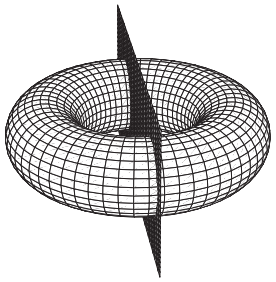


Figure 1 The torus.

By this assumption if $\langle \ell, \ell' \rangle \in E_n$, then no other elements of E_n can be of the form $\langle \ell, \ell'' \rangle$ or $\langle \ell'', \ell' \rangle$. The basic example here is the torus which one obtains from a rectangular box $\Lambda \subset \mathbb{Z}^d$, $|\Lambda|$ even, by imposing periodic conditions on its boundaries. The set of edges is $E = \{\langle \ell, \ell' \rangle | |\ell - \ell'|_\Lambda = 1\}$, where $|\ell - \ell'|_\Lambda$ is the periodic distance on Λ (see the next subsection). Then every plane which contains the center of the torus and its axis cuts it out along a family of edges onto two subgraphs with the property desired (see **Figure 1**).

Theorem 9 *The model [22]–[23] defined on the graph obeying Assumption 8 admits Gaussian domination.*

Proof For $\sigma = \pm 1$, $h = (h_{\ell\ell'})_{\langle \ell, \ell' \rangle \in E}$, and $n = 1, \ldots, m$, we define the map

$$\left(T_n^\sigma h\right)_{\ell\ell'} = \begin{cases} h_{\ell\ell'}, & \text{if } \langle \ell, \ell' \rangle \in E_\sigma^{(n)} \\ h_{\rho_n(\ell)\rho_n(\ell')}, & \text{if } \langle \ell, \ell' \rangle \in E_{-\sigma}^{(n)} \\ 0, & \text{if } \langle \ell, \ell' \rangle \in E_n \end{cases} \qquad [28]$$

According to Assumption 8

$$Z_\Lambda(h) = \int_{\mathbb{R}^{N|\Lambda|}} \exp\left(-\frac{J}{2} \sum_{\langle \ell, \ell' \rangle \in E_1} |x_\ell - x_{\ell'} - h_{\ell\ell'}|^2\right) \times d\nu^+_{\Lambda_+^{(1)}}\left(x_{\Lambda_+^{(1)}}\right) d\nu^-_{\Lambda_-^{(1)}}\left(x_{\Lambda_-^{(1)}}\right) \qquad [29]$$

where

$$d\nu^\sigma_{\Lambda_\sigma^{(1)}}\left(x_{\Lambda_\sigma^{(1)}}\right) = \exp\left(-\frac{J}{2} \sum_{\langle \ell, \ell' \rangle \in E_\sigma^{(1)}} |x_\ell - x_{\ell'} - h_{\ell\ell'}|^2\right) d\chi_{\Lambda_\sigma^{(1)}}\left(x_{\Lambda_\sigma^{(1)}}\right),$$
$$\sigma = \pm 1$$

Set

$$\mathrm{d}\nu^{-}_{\Lambda^{(1)}_{+}}(x_{\Lambda^{(1)}_{+}}) = \exp\left(-\frac{J}{2}\sum_{\langle \ell,\ell'\rangle\in E^{(1)}_{+}}|x_\ell - x_{\ell'} - h_{\rho(\ell)\rho(\ell')}|^2\right) \times \mathrm{d}\chi_{\Lambda^{(1)}_{+}}(x_{\Lambda^{(1)}_{+}})$$

$$\mathrm{d}\nu^{+}_{\Lambda^{(1)}_{-}}\left(x_{\Lambda^{(1)}_{-}}\right) = \exp\left(-\frac{J}{2}\sum_{\langle \ell,\ell'\rangle\in E^{(1)}_{+}}|x_{\rho(\ell)} - x_{\rho(\ell')} - h_{\ell\ell'}|^2\right) \times \mathrm{d}\chi_{\Lambda^{(1)}_{-}}(x_{\Lambda^{(1)}_{-}})$$

Then we apply here Lemma 6, with $\Lambda'_{+} = \{\ell \in \Lambda^{(1)}_{+} | \langle \ell, \ell'\rangle \in E_1\}$, and obtain

$$\begin{aligned}[Z_\Lambda(h)]^2 &\le \int_{\mathbb{R}^{N|\Lambda|}} \exp\left(-\frac{J}{2}\sum_{\langle \ell,\ell'\rangle\in E_1}|x_\ell - x_{\ell'}|^2\right) \times \mathrm{d}\nu^{+}_{\Lambda^{(1)}_{+}}\left(x_{\Lambda^{(1)}_{+}}\right)\mathrm{d}\nu^{+}_{\Lambda^{(1)}_{-}}\left(x_{\Lambda^{(1)}_{-}}\right) \\ &\times \int_{\mathbb{R}^{N|\Lambda|}} \exp\left(-\frac{J}{2}\sum_{\langle \ell,\ell'\rangle\in E_1}|x_\ell - x_{\ell'}|^2\right) \times \mathrm{d}\nu^{-}_{\Lambda^{(1)}_{-}}\left(x_{\Lambda^{(1)}_{-}}\right)\mathrm{d}\nu^{-}_{\Lambda^{(1)}_{+}}\left(x_{\Lambda^{(1)}_{+}}\right) \\ &= Z_\Lambda(T_1^{+}h)Z_\Lambda(T_1^{-}h)\end{aligned}$$

Next we estimate both $Z_\Lambda(T_1^{\pm}h)$ employing E_2 and T_2^{σ}. Repeating this procedure due times we finally get

$$[Z_\Lambda(h)]^{2^m} \le \prod_{\sigma_1,\dots,\sigma_m=\pm 1} Z_\Lambda(T_m^{\sigma_m}\cdots T_1^{\sigma_1}h) = [Z_\Lambda(0)]^{2^m} \quad [30]$$

Note that $T_m^{\sigma_m}\cdots T_1^{\sigma_1}h = 0$ for any $h \in \mathbb{R}^{N|E|}$ and any sequence $\sigma_1,\dots,\sigma_m = \pm 1$, which follows from [27] and [28]. □

As might be clear from the proof given above, the local Gibbs state

$$\phi_\Lambda(A) = \int_{\mathbb{R}^{N|\Lambda|}} A(x_\Lambda)\,\mathrm{d}\mu_\Lambda(x_\Lambda) \quad [31]$$

defined by means of the measure [22], is RP with respect to all reflections $\rho_n, n = 1,\dots,m$. Indeed, the functional defined by the product measure

$$\mathrm{d}\tilde{\chi}_\Lambda(x_\Lambda) \stackrel{\mathrm{def}}{=} \exp\left(-\frac{J}{2}\sum_{\ell\in\Lambda}|x_\ell|^2\right)\mathrm{d}\chi_\Lambda(x_\Lambda) \quad [32]$$

is RP (see Example 2). The Gibbs measure [22] can be written as

$$\mathrm{d}\mu_\Lambda(x_\Lambda) = \frac{1}{Z_\Lambda(0)}\exp\left(\sum_{n=1}^{m}\sum_{k=1}^{N}\sum_{\ell\in\Lambda'_{+,n}} C_\ell^{(k)}\vartheta_n\left(C_\ell^{(k)}\right)\right) \times \mathrm{d}\tilde{\chi}_\Lambda(x_\Lambda) \quad [33]$$

where $C_\ell^{(k)}, k = 1,\dots,N$, are the same as in [20] and $\Lambda'_{+,n} \stackrel{\mathrm{def}}{=} \{\ell \in \Lambda^{(n)}_{+} | \langle \ell, \ell'\rangle \in E_n\}$. Then the reflection positivity of the Gibbs state [31] can be obtained along the line of arguments used for proving Lemma 6. It appears that this is the only possible way to construct an RP functional from another RP functional.

Repeated application of the estimate [15] also yields

$$\phi_\Lambda\left(\prod_{\ell\in\Lambda}F_\ell(x_\ell)\right) \le \prod_{\ell\in\Lambda}\left[\phi_\Lambda\left(\prod_{\ell'\in\Lambda}F_\ell(x_{\ell'})\right)\right]^{1/|\Lambda|} \quad [34]$$

which holds for any family of functions $\{F_\ell : \mathbb{R}^N \to [0,+\infty)\}_{\ell\in\Lambda}$, for which the above expressions make sense. The estimate [34] is a chessboard estimate, which is a very important element of the theory of phase transitions in RP models. The estimate [26] may be obtained from [34].

Infrared Bound

Let us show now how to derive the infrared estimates from the Gaussian domination [26]. Consider the system of N-dimensional spins indexed by the elements of $\mathbb{Z}^d$ with the nearest-neighbor ferromagnetic interaction and the single-spin measure χ. To construct the periodic local Gibbs measure of this system, we take the box

$$\Lambda = (-L, L]^d \bigcap \mathbb{Z}^d, \quad L \in \mathbb{N} \quad [35]$$

and impose periodic conditions on its boundaries. This defines the periodic distance

$$|\ell - \ell'|_\Lambda = \left[\sum_{j=1}^{d}|\ell_j - \ell'_j|_L^2\right]^{1/2}, \quad \ell,\ell' \in \Lambda$$
$$|\ell_j - \ell'_j|_L = \min\{|\ell_j - \ell'_j|;\ L - |\ell_j - \ell'_j|\} \quad [36]$$

and hence the set of edges E, being unordered pairs $\langle \ell, \ell'\rangle$ such that $|\ell - \ell'|_\Lambda = 1$. Thus, we have the graph (Λ, E) and the measure [22]. This is the periodic local Gibbs measure of our model. By [31] it defines the periodic local Gibbs state ϕ_Λ. We have included the inverse temperature β into J

and assumed that the single-spin measure χ is rotation invariant. Let us introduce the Fourier transformation

$$\hat{x}(p) = \frac{1}{\sqrt{|\Lambda|}} \sum_{\ell \in \Lambda} x_\ell e^{i(\ell,p)}$$
$$x_\ell = \frac{1}{\sqrt{|\Lambda|}} \sum_{p \in \Lambda_*} \hat{x}(p) e^{-i(\ell,p)} \qquad [37]$$

$$\Lambda_* = \Big\{ p = (p_1, \dots, p_d) \,|\, p_j = -\pi + \frac{\pi}{L}\kappa_j, \quad \kappa_j = 1, \dots, 2L,\ j = 1, \dots, d \Big\} \qquad [38]$$

Then we can set

$$\widehat{K}_\Lambda^{(k)}(p) = \phi_\Lambda \left[\hat{x}^{(k)}(p) \hat{x}^{(k)}(-p) \right]$$
$$\widehat{K}_\Lambda(p) = \sum_{k=1}^{N} \widehat{K}_\Lambda^{(k)}(p) \qquad [39]$$

Thereby, cf. [1], [2],

$$K_\Lambda(\ell, \ell') \stackrel{\text{def}}{=} \phi_\Lambda[(x_\ell, x_{\ell'})] = \frac{1}{|\Lambda|} \sum_{p \in \Lambda_*} \widehat{K}_\Lambda(p) e^{i(p, \ell - \ell')} \qquad [40]$$

By construction, for any $\ell_0 \in \Lambda$,

$$K_\Lambda(\ell, \ell') = K_\Lambda(\ell + \ell_0, \ell' + \ell_0) \qquad [41]$$

where addition is componentwise modulo $2L$. This means that $K_\Lambda(\ell, \ell')$ is invariant with respect to the translations on the corresponding torus. One can show that $K_\Lambda(\ell, \ell')$ converges, as $L \to +\infty$, to $K(\ell, \ell')$ discussed in the Introduction. The corresponding Gibbs state of the whole model is called the periodic Gibbs state. By construction, it is translation invariant. Set

$$\mathcal{E}(p) = \sum_{j=1}^{d} [1 - \cos p_j], \quad p \in (-\pi, \pi]^d \qquad [42]$$

Theorem 10 *For all* $p \in \Lambda_* \setminus \{0\}$,

$$\widehat{K}_\Lambda(p) \le \frac{N}{2J\mathcal{E}(p)} \qquad [43]$$

Proof Consider the function $f(\xi) = Z_\Lambda(\xi h), \xi \in \mathbb{R}$, where $Z_\Lambda(h)$ is defined by [25]. By Theorem 9 it has a maximum at $\xi = 0$; hence,

$$f''(0) \le 0 \qquad [44]$$

Obviously, $f''(0)$ depends on $h = (h_{\ell\ell'})_{\langle \ell, \ell' \rangle \in E}$, $h_{\ell\ell'} \in \mathbb{R}^N$. Let us choose h such that only the first components $h^{(1)}_{\ell\ell'}$ are nonzero. Then [44] holds if

$$J \sum_{\langle \ell_1, \ell_1' \rangle \in E} \sum_{\langle \ell_2, \ell_2' \rangle \in E} \phi_\Lambda \left[\left(x^{(1)}_{\ell_1} - x^{(1)}_{\ell_1'} \right) \left(x^{(1)}_{\ell_2} - x^{(1)}_{\ell_2'} \right) \right] h^{(1)}_{\ell_1 \ell_1'} h^{(1)}_{\ell_2 \ell_2'}$$
$$\le \sum_{\langle \ell, \ell' \rangle \in E} \left[h^{(1)}_{\ell\ell'} \right]^2 \qquad [45]$$

This means that the eigenvalues of the matrix of the real quadratic form (with respect to h) defined by the left-hand side of [45] do not exceed one. The same ought to be true for the extension of this form to the complex case. Let us show that the complex eigenvectors $h^{(1)}_{\ell\ell'}(p)$ of this matrix and the corresponding eigenvalues $\lambda(p)$ are

$$h^{(1)}_{\ell\ell'}(p) = (e^{i(p,\ell)} - e^{i(p,\ell')})/\sqrt{|\Lambda|}$$
$$\lambda(p) = 2J\mathcal{E}(p)\widehat{K}^{(1)}_\Lambda(p) \qquad p \in \Lambda_* \qquad [46]$$

For $j = 1, \dots, d$, let $\theta_j \in \mathbb{Z}^d$ be the unit vector with the jth component equal to 1. Then for $\langle \ell, \ell' \rangle \in E$, there exists θ_j such that $\ell - \ell' = \pm\theta_j$. Since the edge $\langle \ell, \ell' \rangle$ is an unordered set, let us fix $\ell' = \ell + \theta_j$. Thereby,

$$\frac{1}{|\Lambda|^{1/2}} \sum_{\langle \ell, \ell' \rangle \in E} \left(x^{(1)}_\ell - x^{(1)}_{\ell'} \right) \left(e^{i(p,\ell)} - e^{i(p,\ell')} \right)$$
$$= \frac{2}{|\Lambda|^{1/2}} \sum_{\ell \in \Lambda} \sum_{j=1}^{d} \left[x^{(1)}_\ell e^{i(p,\ell)} - x^{(1)}_\ell e^{i(p,\ell)} \cos(p, \theta_j) \right]$$
$$= 2\hat{x}^{(1)}(p)\mathcal{E}(p)$$

In view of [41], one has

$$\phi_\Lambda[\hat{x}^{(1)}(p)\hat{x}^{(1)}(p')] = \delta_{0,p+p'} \widehat{K}^{(1)}_\Lambda(p)$$

Then employing the latter two facts and [37], we get

$$J \sum_{\langle \ell_2, \ell_2' \rangle \in E} \phi_\Lambda \left[\left(x^{(1)}_{\ell_1} - x^{(1)}_{\ell_1'} \right) \left(x^{(1)}_{\ell_2} - x^{(1)}_{\ell_2'} \right) h_{\ell_2 \ell_2'}(p) \right]$$
$$= 2J\mathcal{E}(p)\phi_\Lambda \left[\left(x^{(1)}_{\ell_1} - x^{(1)}_{\ell_1'} \right) \hat{x}^{(1)}(p) \right]$$
$$= 2J\mathcal{E}(p) \cdot \frac{1}{|\Lambda|^{1/2}} \sum_{p' \in \Lambda_*} \phi_\Lambda \left[\hat{x}^{(1)}(p') \hat{x}^{(1)}(p) \right]$$
$$\times \left(e^{-i(p', \ell_1)} - e^{-i(p', \ell_1')} \right)$$
$$= 2J\mathcal{E}(p) \widehat{K}^{(1)}_\Lambda(p) h_{\ell_1 \ell_1'}(p)$$

which proves [46]. Then by [45] $\widehat{K}^{(1)}_\Lambda(p) \le 1/2J\mathcal{E}(p)$, for $p \ne 0$. The same holds for $\widehat{K}^{(k)}_\Lambda(p), k = 2, \dots, N$, which by [39] yields [43]. □

The result just proved and the convergence of $K_\Lambda(\ell, \ell') \to K(\ell, \ell')$, as $L \to +\infty$, imply the infrared bound [4]. It turns out that the estimate [43]

may be used directly to prove the phase transition. Consider

$$P_\Lambda \stackrel{\text{def}}{=} \frac{1}{|\Lambda|^2}\sum_{\ell_1,\ell_2\in\Lambda}\phi_\Lambda[(x_{\ell_1},x_{\ell_2})] = \phi_\Lambda\left(\left|\frac{1}{|\Lambda|}\sum_{\ell\in\Lambda}x_\ell\right|^2\right)\geq 0 \qquad [47]$$

where Λ is the box [35]. By [40] and [41], we have

$$P_\Lambda = \frac{1}{|\Lambda|}\widehat{K}_\Lambda(0) \qquad [48]$$

One can show that if $P \stackrel{\text{def}}{=} \lim_{L\to+\infty} P_\Lambda$ is positive, then there exist multiple Gibbs states. By [40], [41], and [48], we get that for any $\ell\in\Lambda$,

$$K_\Lambda(\ell,\ell) = P_\Lambda + \frac{1}{|\Lambda|}\sum_{p\in\Lambda_*\setminus\{0\}}\widehat{K}(p) \qquad [49]$$

Suppose that, cf. [5],

$$K_\Lambda(\ell,\ell)\geq \varkappa > 0 \qquad [50]$$

with $\varkappa$ independent of Λ and J. Employing in [49] this estimate and [43], and passing to the limit $L\to+\infty$, we get

$$P \geq \varkappa - \mathcal{I}(d)N/2J \qquad [51]$$

where

$$\mathcal{I}(d) \stackrel{\text{def}}{=} \frac{1}{(2\pi)^d}\int_{(-\pi,\pi]^d}\frac{\mathrm{d}p}{\mathcal{E}(p)} \qquad [52]$$

which is finite for $d\geq 3$. Thereby, we have proved the following:

Theorem 11 *For the spin model* [22], [23], *there exist multiple Gibbs states, and hence multiple phases, if* $d\geq 3$ *and* $J > \mathcal{I}(d)N/2\varkappa$.

Finally, let us pay some attention to the estimate [50], which is closely related with the properties of the single-spin measure χ (note that χ played no role in obtaining [26] and [43]). If it is the uniform measure on the unit sphere $S_{N-1}\subset\mathbb{R}^N$, then $K_\Lambda(\ell,\ell)=1$ and [50] is trivial. In general, one has to employ some technique to obtain such an estimate.

Reflection Positivity and Phase Transitions in Quantum Systems

As in the classical case, the way of proving the phase transition for appropriate models leads from an estimate like [17] to Gaussian domination and then to the infrared bound. However, here this way is much more complicated, so in the frames of this article we can only sketch its main elements basing on the original paper by Dyson *et al.* (1978), where the interested reader can find the details. As above, we start by studying reflection positive functionals.

Reflection Positivity in Nonabelian Case

Again we consider a finite set Λ, $|\Lambda|$ being even. For every $\ell\in\Lambda$, let a complex Hilbert space $\mathcal{H}_\ell$ be given. This is the single-spin physical Hilbert space for our quantum system. We suppose that all $\mathcal{H}_\ell$, $\ell\in\Lambda$, are the copies of a certain finite-dimensional space $\mathcal{H}$. The physical Hilbert space $\mathcal{H}_\Lambda$ corresponding to Λ is the tensor product of $\mathcal{H}_\ell$, $\ell\in\Lambda$. Let $\mathcal{A}_\Lambda$ be the algebra of all linear operators defined on $\mathcal{H}_\Lambda$. This is the algebra of observables in our case; it is noncommutative (nonabelian) and contains the unit element I – the identity operator. As above, Λ splits into two subsets $\Lambda_\pm$, which are the mirror images of each other, that is, we are given a reflection $\rho:\Lambda\to\Lambda$, such that $\rho(\Lambda_+)=\Lambda_-$. This allows us to introduce the corresponding subalgebras $\mathcal{A}_\Lambda^\pm$ by setting the elements of $\mathcal{A}_\Lambda^+$ to be of the form $A\otimes I$, where $A:\mathcal{H}_{\Lambda_+}\to\mathcal{H}_{\Lambda_+}$ is a linear operator and I is the identity operator on $\mathcal{H}_{\Lambda_-}$. Respectively, the elements of $\mathcal{A}_\Lambda^-$ are to be of the form $I\otimes A$. Then we define the map $\vartheta:\mathcal{A}_\Lambda^+\to\mathcal{A}_\Lambda^-$ as

$$\vartheta(A\otimes I) = I\otimes\bar{A} \qquad [53]$$

where $A\mapsto\bar{A}$ is complex (not Hermitian) conjugation; it may be realized as transposing and taking Hermitian conjugation. For $A_1,\dots,A_n\in\mathcal{A}$, one has $\bar{A}_1\cdots\bar{A}_n = \overline{A_1,\cdots A_n}$. We also suppose that ϑ possesses the properties [8]. A linear functional $\phi:\mathcal{A}_\Lambda\to\mathbb{R}$ is called RP (with respect to the pair ρ,ϑ) if it has the property [9].

Definition 12 A functional ϕ is called generalized reflection positive (GRP) if for any $A_1,\dots,A_n\in\mathcal{A}_\Lambda^+$,

$$\phi[A_1\vartheta(A_1)\cdots A_n\vartheta(A_n)]\geq 0 \qquad [54]$$

In principle, this notion differs from the reflection positivity only in the nonabelian case. However, if the algebras $\mathcal{A}_\Lambda^\pm$ commute (they do commute in our case), a functional ϕ is RP if and only if it is GRP.

Example 13 Let

$$\phi(A) = \mathrm{trace}(A),\quad A\in\mathcal{A}_\Lambda \qquad [55]$$

Since the space $\mathcal{H}_\Lambda$ is finite dimensional, this ϕ is well defined. It is GRP. Indeed, as the algebras $\mathcal{A}_\Lambda^\pm$ commute, we have

$$\begin{aligned}&\phi[A_1\otimes I\cdot\vartheta(A_1\otimes I)\cdots A_n\otimes I\cdot\vartheta(A_n\otimes I)]\\ &\quad=\phi[A_1\otimes I\cdots A_n\otimes I\cdot\vartheta(A_1\otimes I)\cdots\vartheta(A_n\otimes I)]\\ &\quad=\phi[A_1\otimes I\cdots A_n\otimes I\cdot\vartheta(A_1\otimes I\cdots A_n\otimes I)]\\ &\quad=\mathrm{trace}[A_1\cdots A_n]\cdot\mathrm{trace}[\bar{A}_1\cdots\bar{A}_n]\\ &\quad=|\mathrm{trace}[A_1\cdots A_n]|^2\geq 0\end{aligned}$$

The Cauchy–Schwarz inequality [13] obviously holds also in the quantum case. By means of this inequality and the Trotter product formula

$$\exp(A+B)=\lim_{n\to+\infty}[\exp(A/n)\,\exp(B/n)]^n \qquad [56]$$

one can prove that every RP functional obeys an estimate like [17]. Thereby, we have the following analog of Lemma 6:

Lemma 14 *Let $A, B, C_1, \ldots, C_n \in \mathcal{A}_\Lambda^+$ be any self-adjoint operators possessing real matrix representation and $a_1, \ldots, a_m$ be any real numbers. Then*

$$\left[\operatorname{trace}\left\{\exp\left(A+\vartheta(B)-\sum_{n=1}^{m}[C_n-\vartheta(C_n)-a_n]^2\right)\right\}\right]^2 \le \operatorname{trace}\left\{\exp\left(A+\vartheta(A)-\sum_{n=1}^{m}[C_n-\vartheta(C_n)]^2\right)\right\} \times \operatorname{trace}\left\{\exp\left(B+\vartheta(B)-\sum_{n=1}^{m}[C_n-\vartheta(C_n)]^2\right)\right\} \qquad [57]$$

Gaussian Domination and Phase Transitions

To proceed further we need a concrete model with finite-dimensional physical Hilbert spaces. As every quantum model, it is defined by its Hamiltonian. Let $\Lambda \subset \mathbb{Z}^d$ be the box [35] and (Λ, E) be the same graph as in the subsection "Infrared bound." The periodic Hamiltonian of our model is

$$H_\Lambda=\sum_{\ell\in\Lambda}Q_\ell+\frac{1}{2}\sum_{\langle\ell,\ell'\rangle\in E}|S_\ell-S_{\ell'}|^2 \qquad [58]$$

where at each $\ell \in \Lambda$ we have the copies Q_ℓ, $S_\ell^{(1)}, \ldots, S_\ell^{(N)}$ of $N+1$ basic operators, acting in the Hilbert space $\mathcal{H}_\ell$, and

$$|S_\ell-S_{\ell'}|^2=\sum_{k=1}^{N}\left(S_\ell^{(k)}-S_{\ell'}^{(k)}\right)^2$$

The only condition we impose so far is that all these operators can simultaneously be chosen as real matrices. For $h=(h_{\ell\ell'})_{\langle\ell,\ell'\rangle\in E}\in\mathbb{R}^{N|E|}$, we set

$$Z_\Lambda(h)=\operatorname{trace}\left\{\exp\left(-\beta\sum_{\ell\in\Lambda}Q_\ell -\frac{\beta}{2}\sum_{\langle\ell,\ell'\rangle\in E}|S_\ell-S_{\ell'}-h_{\ell\ell'}|^2\right)\right\} \qquad [59]$$

where $\beta>0$ is the inverse temperature.

Theorem 15 *For the model [58] and any $h=(h_{\ell\ell'})_{\langle\ell,\ell'\rangle\in E}\in\mathbb{R}^{N|E|}$,*

$$Z_\Lambda(h)\le Z_\Lambda(0) \qquad [60]$$

The proof is performed by means of Lemma 14.

The periodic local Gibbs state of the model [58] at the inverse temperature β, analogous to the state [31], is

$$\phi_\Lambda(A)=\operatorname{trace}\{A\exp(-\beta H_\Lambda)\}/Z_\Lambda(0),\quad A\in\mathcal{A}_\Lambda \qquad [61]$$

As in the classical case, one can define the parameter [47]. However, now the fact that $\lim_{L\to+\infty}P_\Lambda>0$ does not yet imply the phase transition. One has to prove a more general fact

$$\lim_{L'\to+\infty}\left\{\lim_{L\to+\infty}\phi_\Lambda\left(\left|\frac{1}{|\Lambda'|}\sum_{\ell\in\Lambda'}S_\ell\right|^2\right)\right\}>0 \qquad [62]$$

where Λ' is the box [35] of side $2L'$. Furthermore, in the quantum case the Gaussian domination [60] does not lead directly to the estimate [43], which yields [51]. Instead, one can get a bound like [43] but for the Duhamel two-point function (DTF). Given $A, B \in \mathcal{A}_\Lambda$, their DTF is

$$(A,B)=\int_0^1\phi_\Lambda(A\mathrm{e}^{-\xi\beta H_\Lambda}B\mathrm{e}^{\xi\beta H_\Lambda})\mathrm{d}\xi \qquad [63]$$

By means of [56] one can show that

$$(A,B)=\frac{1}{Z_\Lambda(0)} \times\left\{\frac{\partial^2}{\partial\xi\partial\eta}\operatorname{trace}[\exp(\xi A+\eta B-\beta H_\Lambda)]\right\}_{\xi=\eta=0} \qquad [64]$$

Let $\hat{S}(p)=(\hat{S}^{(1)}(p),\ldots,\hat{S}^{(N)}(p)), p\in\Lambda_*$, be the Fourier image of S_ℓ, defined by [37], [38]. Then

$$\left(\hat{S}(p),\hat{S}(-p)\right)=\sum_{k=1}^{N}\left(\hat{S}^{(k)}(p),\hat{S}^{(k)}(-p)\right)$$

Theorem 16 *For all $p\in\Lambda_*\setminus\{0\}$, it follows that*

$$\left(\hat{S}(p),\hat{S}(-p)\right)\le\frac{N}{2\beta\mathcal{E}(p)} \qquad [65]$$

To prove this statement one has to use the Gaussian bound [60] exactly as in the case of Theorem 10. The second derivative with respect to ξ gives the corresponding DTF (see [64]).

Now let us indicate how the infrared bound [65] leads to the phase transition. To this end we use the simplest quantum spin model with the Hamiltonian [58], for which $Q_\ell=0, N=2$, and $S_\ell^{(k)}, k=1,2$, being the copies of the Pauli matrices

$$S^{(1)}=\begin{pmatrix}0&1\\1&0\end{pmatrix},\qquad S^{(2)}=\begin{pmatrix}1&0\\0&-1\end{pmatrix}$$

Then

$$K_\Lambda^{(k)}(\ell,\ell)=\phi_\Lambda\left(S_\ell^{(k)}\cdot S_\ell^{(k)}\right)=1 \quad \text{for all } \ell\in\Lambda,\quad k=1,2 \qquad [66]$$

which gives the bound $\varkappa$ (see [50]). For $A, B \in \mathcal{A}_\Lambda$, by $[A, B]$ we denote the commutator $AB - BA$. Set

$$\Sigma_\Lambda^{(k)}(p) = \phi_\Lambda\Big(\Big[\hat{S}^{(k)}(p), \Big[H_\Lambda, \hat{S}^{(k)}(-p)\Big]\Big]\Big) \quad k = 1, 2 \qquad [67]$$

The phase transition in the model we consider can be established by means of the following statement (see Dyson 1978, Theorem 5.1).

Proposition 17 *Suppose there exist* $\Sigma^{(k)}(p), k = 1, 2$, $p \in (-\pi, \pi]^d$ *such that, for all* $L \in \mathbb{N}$,

$$\Sigma_\Lambda^{(k)}(p) \le \Sigma^{(k)}(p), \quad k = 1, 2, \ p \in \Lambda_* \qquad [68]$$

Then the model undergoes a phase transition at a certain finite β *if* $d \ge 3$ *and*

$$\frac{1}{(2\pi)^d} \int_{(-\pi,\pi]^d} \left[\frac{\Sigma^{(k)}(p)}{8\mathcal{E}(p)}\right]^{1/2} \mathrm{d}p < 1 \qquad [69]$$

for a certain, and hence for both, $k = 1, 2$.

Thus to prove the phase transition we have to estimate $\Sigma_\Lambda^{(k)}(p), k = 1, 2$. By means of the Cauchy–Schwarz inequality, the estimate [69] may be transformed into the following:

$$\frac{1}{(2\pi)^d} \int_{(-\pi,\pi]^d} \Big[\Sigma^{(1)}(p) + \Sigma^{(2)}(p)\Big] \mathrm{d}p < 16/\mathcal{I}(d)$$

where $\mathcal{I}(d)$ is the same as in [52]. The integral on the left-hand side can be estimated from above by $8\sqrt{d(d+1)}$; hence, the latter inequality holds if

$$\mathcal{I}(d)\sqrt{d(d+1)} < 2$$

which holds for all $d \ge 3$. In particular, $\mathcal{I}(3) \approx 0.505$.

Bibliographic Notes

As the original sources on the RP method in the theory of phase transitions we mention the papers Fröhlich *et al.* (1976) (classical case), Dyson *et al.* (1978) (quantum case), and Fröhlich and Lieb (1978) (both cases). In a unified way and with many examples, this method is described in Fröhlich *et al.* (1978, 1980). A detailed analysis of the method, especially in its applications to classical models with unbounded spins, was given in Shlosman (1986). The techniques based on the chessboard and contour estimates are described in Fröhlich and Lieb (1978) and Shlosman (1986). As was mentioned above, the quantum case is much more complicated; it gets even more complicated if one deals with quantum models employing infinite-dimensional physical Hilbert spaces and unbounded operators, such as quantum crystals. The adaptation of the RP method to such models was made in Driessler *et al.* (1979), Pastur and Khoruzhenko (1987), Barbulyak and Kondratiev (1992), and Kondratiev (1994). In the latter two papers a general version of the quantum crystal was studied in the framework of the Euclidean approach, based on functional integrals (see Albeverio *et al.* (2002)). In this approach the quantum crystal is represented as a lattice spin model with unbounded infinite-dimensional spins. Like in the case of classical models with unbounded spins, here establishing the estimate [5] becomes a highly nontrivial task. In particular cases, for example, for ϕ^4-models, one applies special tools like the Bogoliubov inequalities (see Driessler *et al.* (1979) and Pastur and Khoruzhenko (1987)). In the general case quasiclassical asymptotics allow us to get the lower bound [5] (see Barbulyak and Kondratiev (1992) and Kondratiev (1994)). There is one more technique based on reflection positivity (see Lieb (1989)). It employs reflections in spin spaces, whereas the properties of the index sets (lattices) play no role. This technique proved to be useful in the theory of strongly correlated electron systems, see Tian (2004). Finally, we mention the books of Georgii (1988), Prum (1986), and Sinai (1982) where different aspects of the RP method are described. In Georgii (1988), one can also find extended bibliographical and historical comments on this subject.

See also: Phase Transition Dynamics; Phase Transitions in Continuous Systems; Quantum Spin Systems; Renormalization: Statistical Mechanics and Condensed Matter.

Further Reading

Albeverio S, Kondratiev YG, Kozitsky Y, and Röckner M (2002) Euclidean Gibbs states of quantum lattice systems. *Reviews in Mathematical Physics* 14: 1–67.

Barbulyak VS and Kondratiev YG (1992) The quasiclassical limit for the Schrödinger operator and phase transitions in quantum statistical physics. *Functional Analysis and its Applications* 26(2): 61–64.

Driessler W, Landau L, and Fernando-Perez J (1979) Estimates of critical lengths and critical temperatures for classical and quantum lattice systems. *Journal of Statistical Physics* 20: 123–162.

Dyson FJ, Lieb EH, and Simon B (1978) Phase transitions in quantum spin systems with isotropic and nonisotropic interactions. *Journal of Statistical Physics* 18: 335–383.

Fröhlich J, Israel R, Lieb EH, and Simon B (1978) Phase transitions and reflection positivity, I. General theory and long range lattice models. *Communications in Mathematical Physics* 62: 1–34.

Fröhlich J, Israel R, Lieb EH, and Simon B (1980) Phase transitions and reflection positivity, II. Lattice systems with short-range and Coulomb interactions. *Journal of Statistical Physics* 22: 297–347.

Fröhlich J and Lieb EH (1978) Phase transitions in anisotropic lattice spin systems. *Communications in Mathematical Physics* 60: 233–267.

Fröhlich J, Simon B, and Spenser T (1976) Infrared bounds, phase transitions and continuous symmetry breaking. *Communications in Mathematical Physics* 50: 79–85.

Georgii H-O (1988) *Gibbs Measures and Phase Transitions*. Studies in Mathematics, vol.9. Berlin: Walter de Gruyter.

Kondratiev JG (1994) Phase transitions in quantum models of ferroelectrics. In: *Stochastic Processes, Physics and Geometry*, vol. II, pp. 465–475. Singapure: World Scientific.

Lieb EH (1989) Two Theorems on the Hubbard Model. *Physical Review Letters* 62: 1201–1204.

Pastur LA and Khoruzhenko BA (1987) Phase transitions in quantum models of rotators and ferroelectrics. *Theoretical and Mathematical Physics* 73: 111–124.

Prum P (1986) *Processus sur un Réseau et Mesures de Gibbs*. Applications. Techniques Stochastiques. Paris: Masson.

Prum B and Fort J-C (1991) *Stochastic Processes on a Lattice and Gibbs Measures* (translated from the French by Bertram Eugene Schwarzbach and revised by the authors). Mathematical Physics Studies, vol.11. Dordrecht: Kluwer Academic Publishers Group.

Shlosman SB (1986) The method of reflection positivity in the mathematical theory of first-order phase transitions. *Russian Mathematical Surveys* 41: 83–134.

Sinai Ya (1982) *Theory of Phase Transitions. Rigorous Results*. Oxford: Pergamon.

Tian G-S (2004) Lieb's spin-reflection-positivity method and its applications to strongly correlated electron systems. *Journal of Statistical Physics* 116: 629–680.

Regularization for Dynamical ζ-Functions

V Baladi, Institut Mathématique de Jussieu, Paris, France

Introduction

If A is a finite, say $N \times N$, matrix with complex coefficients, the following easy equality gives an expression for the polynomial $\prod_{k=1}^{N}(1 - z\lambda_k) = \det(\mathrm{Id} - zA)$:

$$\det(\mathrm{Id} - zA) = \exp\left(-\sum_{n=1}^{\infty} \frac{z^n}{n} \operatorname{tr} A^n\right) \qquad [1]$$

(here, Id denotes the identity matrix and tr is the trace of a matrix). Even in this trivial finite-dimensional case, the z-radius of convergence of the logarithm of the right-hand side only gives information about the spectral radius (the modulus of the largest eigenvalue) of A. The zeros of the left-hand side (i.e., the inverses $z = 1/\lambda_k$ of the nonzero eigenvalues of A) can only be located after extending holomorphically the right-hand side. The purpose of this article is to discuss some dynamical situations in which A is replaced by a linear bounded operator $\mathcal{L}$, acting on an infinite-dimensional space, and for which a dynamical determinant (or dynamical ζ-function), constructed from periodic orbits, takes the part of the right-hand side. In the examples presented, $\mathcal{L}$ will be a transfer operator associated to a weighted discrete-time dynamical system: given a transformation $f : M \to M$ on a compact manifold M and a function $g : M \to \mathbb{C}$, we set

$$\mathcal{L}\varphi = g \cdot \varphi \circ f^{-1} \qquad [2]$$

(If f is not inversible, it is understood, e.g., that f has at most finitely many inverse branches, and that the right-hand side of [2] is the sum over these inverse branches, see the next section.) We let $\mathcal{L}$ act on a Banach space of functions or distributions φ on M. For suitable g (in particular $g = |\det Tf^{-1}|$ when this Jacobian makes sense), the spectrum of $\mathcal{L}$ is related to the fine statistical properties of the dynamics f: existence and uniqueness of equilibrium states (related to the maximal eigenvector of $\mathcal{L}$), decay of correlations (related to the spectral gap), limit laws, entropies, etc: see, for example, Baladi (1998) or Cvitanović *et al.* (2005). The operator $\mathcal{L}$ is not always trace-class, indeed, it sometimes is not compact on any reasonable space. Even worse, its essential spectral radius may coincide with its spectral radius. (Recall that the essential spectral radius of a bounded linear operator $\mathcal{L}$ acting on a Banach space is the infimum of those $\rho > 0$, such that the spectrum of $\mathcal{L}$ outside of the disk of radius ρ is a finite set of eigenvalues of finite algebraic multiplicity.) However, various techniques allow us to prove that a suitable dynamically defined replacement for the right-hand side of [1] extends holomorphically to a disk in which its zeros describe at least part of the spectrum of $\mathcal{L}$. Some of these techniques have a "regularization" flavor, and we shall concentrate on them.

In the following section, we present the simplest case: analytic expanding or hyperbolic dynamics, for which no regularization is necessary and the Grothendieck–Fredholm theory can be applied. Next, we consider analytic situations where finitely many neutral periodic orbits introduce branch cuts in the dynamical determinant, and see how to "regularize" them. Finally, we discuss a

kneading operator regularization approach, inspired by the work of Milnor and Thurston, and applicable to dynamical systems with finite smoothness.

Despite the terminology, none of the regularization techniques discussed below match the following "ζ-regularization" formula:

$$\prod_{k=1}^{\infty} a_k = \exp\left(-\frac{\mathrm{d}}{\mathrm{d}s}\sum_{k=1}^{\infty} a_k^{-s}|_{s=0}\right) \qquad [3]$$

(For information about the above ζ-regularization and its applications to physics, we refer, e.g., to Elizalde 1995. See also Voros (1987) and Fried (1986) for more geometrical approaches and further references, e.g., to the work of Ray and Singer.)

We do not cover all aspects of dynamical ζ-functions here. For more information and references, we refer to our survey Baladi (1998), to the more recent surveys by Pollicott (2001) and Ruelle (2002), and also to the exhaustive account by Cvitanović *et al.* (2005), which contains a rich array of physical applications.

The Grothendieck–Fredholm Case

Let M be a real analytic compact manifold (e.g., the circle or the d-torus), and let $f: M \to M$ be real analytic and $g: M \to \mathbb{C}$ be analytic.

First suppose that f is uniformly expanding, that is, there is $\lambda > 1$ so that $\|Tf(v)\| \geq \lambda \|v\|$. (For example, $f(z) = z^2$ on the unit circle, or a small analytic perturbation thereof.) Consider

$$\mathcal{L}_{f,g}\varphi(x) = \sum_{y:\, f(y)=x} g(y)\varphi(y) \qquad [4]$$

(For example, with $g(y) = 1/|\det Tf(y)|$ or $1/|\det Tf(y)|^s$.) Ruelle (1976) proved that an operator $\mathcal{L}_0$, which is essentially the same as $\mathcal{L}_{f,g}$ (the difference, if any, arises from the use of Markov partitions, especially in higher dimensions), acting on a Banach space of holomorphic and bounded functions, is not only compact, but is in fact a nuclear operator in the sense of Grothendieck. In particular, the traces of all its powers are well defined, and the Grothendieck–Fredholm (Gohberg *et al.* 2000) determinant

$$d_0(z) = \exp\left(-\sum_{n=1}^{\infty} \frac{z^n}{n} \operatorname{tr} \mathcal{L}_0^n\right) \qquad [5]$$

extends to an entire function of finite order, the zeros of which are exactly the inverses of the nonzero eigenvalues of $\mathcal{L}_0$. (The order of the zero coincides with the algebraic multiplicity of the eigenvalue.) Ruelle also proved that the traces can be written as sums over periodic orbits:

$$\operatorname{tr} \mathcal{L}_0^n = \sum_{x:\, f^n(x)=x}^{*} \frac{\prod_{k=0}^{n-1} g(f^k x)}{|\det(\mathrm{Id} - Tf_x^{-n})|}$$

where $\sum^*$ means that the fixed points of f^n lying in the intersection of two or more elements of the Markov partition must be counted two or more times. (Note that if $f^n(x) = x$, then this closed orbit gives a natural inverse branch for f^{-n}.) Taking into account the periodic orbits on the boundaries of the Markov partition, Ruelle expresses the following "dynamical determinant":

$$d_{f,g}(z) = \exp\left[-\sum_{n=1}^{\infty} \frac{z^n}{n} \sum_{x:\, f^n(x)=x} \frac{\prod_{k=0}^{n-1} g(f^k x)}{|\det(\mathrm{Id} - Tf_x^{-n})|}\right] \qquad [6]$$

as an alternated product of determinants $d_0(z)$ as in [5].

The expression [6] is sometimes also called a "dynamical ζ-function," but we prefer to reserve this terminology for the following power series:

$$\zeta_{f,g}(z) = \exp\left[+\sum_{n=1}^{\infty} \frac{z^n}{n} \sum_{x:\, f^n(x)=x} \prod_{k=0}^{n-1} g(f^k x)\right] \qquad [7]$$

It is not difficult to write $\zeta_{f,g}(z)$ as (Baladi 1998) an alternated product of determinants d_{f,g_i}, for $i = 0, \ldots, d$, and appropriate weights g_i.

In fact, the results just described hold in more generality, for example, for piecewise bijective and analytic interval maps. Such maps, f, appear naturally, for example, when considering Schottky subgroups of $PSL(2, \mathbb{Z})$. We mention the recent work of Guillopé–Lin–Zworski (2004), who let the transfer operator associated to such f and weights $g_s(y) = 1/|f'(y)|^s$ act (as trace-class operators) on suitable Hilbert spaces of holomorphic functions. This allows them to obtain precise estimates for the number of zeros of $s \mapsto d_{f,g_s}$ [1] in the complex plane: these zeros are the resonances (in the sense of the spectrum of the Laplacian).

Note that the nuclearity properties extend also to the Gauss map $f(x) = \{1/x\}$, which has infinitely many inverse branches, if the weight g has summability properties over the branches (e.g., $g_s(y) = |1/f'(y)|^s$, where s is a complex parameter, with $\Re s > 1/2$). The dynamical determinant $d_{f,g_s}(z)$ for the transfer operator of the Gauss map is related to the Selberg ζ-function (see e.g., Chang and Mayer (2001) and references therein).

Next, assume that M and g are as before, but f is a uniformly hyperbolic real analytic diffeomorphism.

For example, M is the 2-torus and f is a small real analytic perturbation of the linear automorphism

$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

More generally, we may assume that f is a real analytic Anosov diffeomorphism, that is, there are $C \geq 1$ and $\lambda > 1$ such that the tangent bundle decomposes as $TM = E^u \oplus E^s$, where the dynamical bundles E^u and E^s are Tf-invariant, with $\|Tf^n|_{E^s}\| \leq C\lambda^{-n}$ and $\|Tf^{-n}|_{E^u}\| \leq C\lambda^{-n}$ for all $n \in \mathbb{Z}_+$. In general, the smoothness of $x \mapsto E^u(x)$ and $E^s(x)$ is only Hölder. Under the very strong additional assumption that $E^u(x)$ and $E^s(x)$ are real analytic, Ruelle (1976) (see also Fried (1986)) showed that the power series $d_{f,g}(z)$ can again be written as a finite alternated product (this product being again an artifact of the Markov partition) of entire functions of finite order. For this, he constructed auxiliary transfer operators associated to the expanding (and analytic!) quotiented dynamics acting on holomorphic functions on disks. The analyticity assumption on the dynamical bundles was later lifted by Rugh (1996) (see also Fried (1995)), who let their transfer operators act on Banach topological tensor products of spaces of holomorphic functions on a disk with the dual of such a space. In all these cases, the transfer operator is a nuclear operator in the sense of Grothendieck and no regularization is needed. (More recent work of Kitaev (1999), when applied to this analytic setting, shows that the "meromorphic" function $d_{f,g}(z)$ in fact does not have poles.)

Regularization and Intermittency

Consider the interval $M = [0, 1]$, and f defined on M by $f(x) = f_1(x) = x/(1-x)$ on $[0, 1/2]$, and $f(x) = f_2(x) = (1-x)/x$ on $[1/2, 1]$. (This is the Farey map, which appears naturally when considering continued fractions.) Each of the two branches is an analytic bijection onto $[0, 1]$. The second branch is expanding, but the first one, f_1, has a (parabolic) neutral fixed point at $x = 0$ (the expansion is $f(x) = x + x^2 + x^3 + \cdots$). Let $g = g_s$ be an analytic weight of the form $g(y) = 1/|f'(y)|^s$ for $\Re s \geq 1/2$. We are interested in the spectrum of the operator $\mathcal{L}_{f,g}$ associated with the pair (f, g) by [4]. Clearly, the expression [6] is not a good candidate for an analog of the Fredholm determinant of $\mathcal{L}_{f,g}$. Rugh (1996) introduced a Banach space $\mathcal{B}$ of functions in a complex neighborhood of M, having a controlled singularity at 0, and such that the spectral radius of $\mathcal{L}_{f,g}$ on $\mathcal{B}$ is equal to 1, and such that the following regularized determinant

$$d_{f,g}(z) = \exp\left[-\sum_{n=1}^{\infty} \frac{z^n}{n} \sum_{x \in (0,1]:\, f^n(x) = x} \frac{\prod_{k=0}^{n-1} g_s(f^k x)}{1 - Tf_x^{-n}} \right] \qquad [8]$$

is a holomorphic function in the cut complex plane $\{z \in \mathbb{C} \mid z \notin [1, \infty)\}$. Furthermore, its zeros z in this cut plane are in bijection with the spectrum of $\mathcal{L}_{f,g}|_{\mathcal{B}}$ outside of the unit interval $[0, 1]$, and this spectrum consists of eigenvalues $1/z$ of finite multiplicities. Finally, these eigenvalues can only accumulate at 0 or 1, although each point in the unit interval belongs to the spectrum of $\mathcal{L}_{f,g}$. In particular, the essential spectral radius of $\mathcal{L}_{f,g}$ on $\mathcal{B}$ coincides with its spectral radius.

Let us define the Banach space $\mathcal{B}$ and explain the key ideas in the proof of the above result (Rugh's claim is in fact more general than the statement above and applies to a class of maps f with neutral fixed points). The starting point is the decomposition

$$\mathcal{L}_{f,g} = \mathcal{L}_1 + \mathcal{L}_2$$

where $\mathcal{L}_i \varphi = \varphi \circ f_i^{-1} \cdot |(f_i^{-1})'|^s$. The operator $\mathcal{L}_2$ is of the type discussed in the previous section, and it is nuclear when acting, for example, on bounded holomorphic functions in a complex neighborhood of M. Since f_1 is not expanding (because of the parabolic fixed point at 0), other ideas must be used to handle the operator $\mathcal{L}_1$. The change of coordinates (this idea goes back to Fatou) $w = 1/x$ replaces the weak contraction f_1^{-1} by the translation $w \mapsto w + 1$ in a suitable domain containing a half-plane $\Re w > w_0$. In order to take into account the weight g_s, it is convenient to use the change of variables $\Psi(w) = \varphi(1/w) \cdot w^{-2s}$. Indeed, in the new coordinates the operator $\mathcal{L}_1$ reads as

$$\mathcal{M}_1 \Psi(w) = \Psi(w + 1)$$

The next step consists in letting $\mathcal{M}_1$ act on the Banach space $\mathcal{B}_w$ of Laplace transforms of $L^1(\mathbb{R}^+, \text{Lebesgue})$, that is, functions

$$\Psi(w) = \int_0^{\infty} \mathrm{e}^{-(w - w_0)t} \psi(t)\, \mathrm{d}t$$

with the induced norm $\|\Psi\|_{\mathcal{B}_w} = \int |\psi(t)|\, \mathrm{d}t$. Since $\mathcal{M}_1$ maps ψ to $\mathrm{e}^{-t}\psi(t)$, it is not difficult to see that the spectrum of $\mathcal{M}_1$ on $\mathcal{B}_w$ (and thus of $\mathcal{L}_1$ on the pullback $\mathcal{B}$ of $\mathcal{B}_w$ by Φ, which consists of functions in a complex neighborhood of [0,1], holomorphic in a sector at 0, and with a possible, but controlled, singularity at 0) is the closed unit interval. One can check that $\mathcal{L}_2$ is nuclear on $\mathcal{B}$. Composing a bounded operator with a

nuclear operator gives a nuclear operator. If $1/z \notin [0,1]$, the resolvent $(1-z\mathcal{L}_1)^{-1}$ is a bounded operator, and therefore, for such z, the operator

$$\mathcal{P}(z) := z\mathcal{L}_2(1-z\mathcal{L}_1)^{-1} \qquad [9]$$

is nuclear on $\mathcal{B}$. We view $\mathcal{P}(z)$ as a "regularized" version of $\mathcal{L}_{f,g} = \mathcal{L}_1 + \mathcal{L}_2$. Now, since

$$\begin{aligned}(1-z\mathcal{L}_{f,g})^{-1} &= (1-z(\mathcal{L}_1+\mathcal{L}_2))^{-1} \\ &= (1-z\mathcal{L}_1)^{-1}\left(1-z\mathcal{L}_2(1-z\mathcal{L}_1)^{-1}\right)^{-1}\end{aligned}$$

it is not surprising that one can prove (Rugh 1996) that the Fredholm determinant

$$u \mapsto \det\left(1-\mathcal{L}_2(u-\mathcal{L}_1)^{-1}\right)$$

(which is holomorphic in $u \notin [0,1]$) has as its zero set $\mathrm{sp}(\mathcal{L}_{f,g|\mathcal{B}}) \setminus [0,1]$, and that this set consists in isolated eigenvalues of finite multiplicity (equal to the order of the corresponding zero) for $\mathcal{L}_{f,g}$. Formally,

$$(1-z\mathcal{L}_1)^{-1} = \sum_{k=0}^{\infty} z^k \mathcal{L}_1^k \qquad [10]$$

so that the regularization we just described can be viewed as mirroring an induction (or renormalization) procedure, where the dynamics f is replaced by the first-return map to the "chaotic" part of the phase space [0, 1/2]. (For the Farey map, the induced map is just the Gauss map.) The formal equality [10] is also behind the fact that (Rugh 1996)

$$\mathrm{tr}\,\mathcal{P}(z)^n = \sum_{x \neq 0:\, f^n(x)=x} \frac{\prod_{k=0}^{n-1} g_s(f^k x)}{1 - Tf_x^{-n}}$$

An extension of this theory to the two-dimensional setting has been obtained by Baladi, Pujals, and Sambarino.

Regularization and Kneading Determinants

Up to now we have only discussed analytic dynamical systems, for which hyperbolicity (or uniform expansion) guaranteed that the transfer operator (or a regularized version thereof) was compact, even nuclear, on a natural Banach space. When considering hyperbolic invertible (or expanding noninvertible) maps f, and weights g with "finite smoothness," say C^r for some finite $r > 1$, the transfer operator defined by [2] or [4] is usually not compact on any infinite-dimensional space. However, one can often prove a "Lasota–Yorke" type inequality (see e.g., Baladi (1998)) which ensures that the essential spectral radius $\rho_{\mathrm{ess}}(\mathcal{L}_{f,g})$, defined in the "Introduction," is strictly smaller than the spectral radius. Then, the goal is to prove that the dynamical determinant [6] defines a holomorphic function in the disk of radius $1/\rho_{\mathrm{ess}}$, and that its zeros in this disk are exactly the inverses of the eigenvalues of $\mathcal{L}_{f,g}$. For uniformly expanding C^r maps f on compact manifolds, and C^r weights, denoting by $\lambda > 1$ the expansion coefficient as in the section "The Grothendieck–Fredholm case," this goal was essentially attained by Ruelle (1990). For $\mathcal{L}_{f,g}$ acting on the Banach space of C^r functions on M, Ruelle proved $\rho_{\mathrm{ess}}(\mathcal{L}_{f,g}) \leq \lambda^{-r}$ and was able to extend $d_{f,g}(z)$ (and interpret its zeros) in the disk of radius λ^r.

For C^r Anosov diffeomorphisms f, and C^r weights g, Pollicott, Ruelle, Haydn, and others obtained important results using the symbolic dynamics description (for which the maximal smoothness which can be used is $r \leq 1$, because of the metric-space model). Later, Kitaev (1999) was able to show that $d_{f,g}(z)$ extends to a holomorphic function in the disk of radius $\lambda^{-r/2}$, but did not give any spectral interpretation of the zeros of $d_{f,g}(z)$. More recently, Liverani (2005) was able to give such an interpretation, in a smaller disk however.

All the works mentioned in the previous paragraph are based on some approximation scheme (Taylor expansion style). In the early 1990s, a new approach, with a regularization flavor, was launched (see e.g., Baladi and Ruelle (1996)), initially for piecewise monotone interval maps. We present it next.

Consider a finite set of local homeomorphisms $\psi_\omega : U_\omega \to \psi_\omega(U_\omega)$, where each U_ω is a bounded open interval of $\mathbb{R}$, and of associated weight functions g_ω which are continuous, of bounded variation, and have support inside U_ω. For example, the ψ_ω can be the inverse branches of a single piecewise monotone interval map f, and g_ω can be $g \circ \psi_\omega$ for a single g. (No contraction assumption is required on the ψ_ω: their graph can even coincide with the diagonal on a segment.) The transfer operator is now

$$\mathcal{M}\varphi = \sum_\omega g_\omega \cdot (\varphi \circ \psi_\omega)$$

Ruelle obtained an estimate, noted $\widehat{R}$, for the essential spectral radius of $\mathcal{M}$ acting on the Banach space BV of functions of bounded variation. The main result of Baladi and Ruelle (1996) links the eigenvalues of $\mathcal{M} : \mathrm{BV} \to \mathrm{BV}$ outside of the disk of radius $\widehat{R}$, with the zeros of the following "sharp determinant":

$$\det{}^{\#}(\mathrm{Id} - z\mathcal{M}) = \exp\left(-\sum_{n=1}^{\infty} \frac{z^n}{n} \mathrm{tr}^{\#}\mathcal{M}^n\right) \qquad [11]$$

where (with the understanding that $y/|y| = 0$ if $y = 0$)

$$\mathrm{tr}^{\#}\mathcal{M} = \sum_\omega \int \frac{1}{2}\frac{\psi_\omega(x) - x}{|\psi_\omega(x) - x|}\, \mathrm{d}g_\omega(x)$$

If the ψ_ω are strict contractions which form the set of inverse branches of a piecewise monotone interval map f, and $g_\omega = g \circ \psi_\omega$, then integration by parts together with the key property that

$$\mathrm{d}\frac{x}{2|x|} = \delta, \text{the Dirac delta at the origin of } \mathbb{R}$$

show that $\det^{\#}(\mathrm{Id} - z\mathcal{M}) = 1/\zeta_{f,g}(z)$ (recall [7]). If one assumes instead only that the graph of each admissible composition $\psi^n_{\boldsymbol{\omega}}$ of n successive ψ_ω's (with $n \geq 1$) intersects the diagonal transversally, then

$$\det{}^{\#}(\mathrm{Id} - z\mathcal{M}) = \exp\left[-\sum_{n=1}^{\infty}\frac{z^n}{n}\sum_{\text{admissible }\psi^n_{\boldsymbol{\omega}}}\sum_{x:\psi^n_{\boldsymbol{\omega}}(x)=x} L(x,\psi^n_{\boldsymbol{\omega}}) \times \prod_{k=0}^{n-1} g_{\omega_k}\left(\psi^k_{\boldsymbol{\omega}}(x)\right)\right] \quad [12]$$

where $L(x,\psi) \in \{-1,1\}$ is the Lefschetz number of a transversal fixed point $x = \psi(x)$ (if ψ is C^1 this is just $\mathrm{sgn}\,(1-\psi'(x))$). Therefore, we call the sharp determinant $\det^{\#}(\mathrm{Id} - z\mathcal{M})$ a Ruelle–Lefschetz (dynamical) determinant. For a class of "unimodal" interval maps f and constant weight $g = 1$, the expression [12] with Lefschetz numbers, coming from the additional transversality assumption, gives that $\det^{\#}(\mathrm{Id} - z\mathcal{M})$ is just $1/\zeta^-(z)$, where the "negative ζ-function"

$$\zeta^-(z) = \exp\left[+\sum_{n=1}^{\infty}\frac{z^n}{n}(2\#\mathrm{Fix}^-(f^n) - 1)\right] \quad [13]$$

is defined by counting (twice) the sets

$$\mathrm{Fix}^-(f^n) = \{x | f^n(x) = x,\ f \text{ strictly decreasing in a neighborhood of } x\}$$

of "negative fixed points." This negative ζ-function was studied by Milnor and Thurston, who proved the remarkable identity

$$(\zeta^-(z))^{-1} = \det(1 + \widehat{D}(z))$$

where $\widehat{D}(z)$ is a 1×1 "matrix," which is just a power series in z with coefficients in $\{-1, 0, +1\}$, given by the signed itinerary of the image of the turning point (the so-called "kneading" data).

Returning now to the general setup ψ_ω, g_ω, the crucial step in the proof of the spectral interpretation of the zeros of this Ruelle–Lefschetz determinant consists in establishing the following continuous version of the Milnor–Thurston identity:

$$\det{}^{\#}(\mathrm{Id} - z\mathcal{M}) = \det{}^{*}(\mathrm{Id} + \widehat{\mathcal{D}}(z)) \quad [14]$$

where the "kneading operator" $\widehat{\mathcal{D}}(z)$ replaces (formally) the finite kneading matrix of Milnor and Thurston. In a suitable z-disk, one proves that this operator $\widehat{\mathcal{D}}(z)$ is a Hilbert–Schmidt operator on an L^2 space (its kernel is bounded and compactly supported), thus allowing the use of regularized determinants of order 2 (see e.g., Gohberg *et al.* (2000)). By definition, $\det^{*}(\mathrm{Id} + \widehat{\mathcal{D}}(z))$ is the product of this regularized determinant with the exponential of the average of the kernel of $\widehat{\mathcal{D}}(z)$ along the diagonal, which is well defined. Another kneading operator, $\mathcal{D}(z)$, is essential. If $1/z$ is not in the spectrum of $\mathcal{M}$ (on BV), then $\mathcal{D}(z)$ is also Hilbert–Schmidt, and one can show $\det^{*}(\mathrm{Id} + \widehat{\mathcal{D}}(z)) = \det^{*}(\mathrm{Id} + \mathcal{D}(z))^{-1}$. The initial definitions of $\widehat{\mathcal{D}}(z)$ and $\mathcal{D}(z)$ were technical and we shall not give them here. However, a more conceptual definition of the $\mathcal{D}(z)$ was later implemented:

$$\mathcal{D}(z) = \mathcal{N}(\mathrm{Id} - z\mathcal{M})^{-1}\mathcal{S} \quad [15]$$

where $\mathcal{N}$ is an auxiliary transfer operator and $\mathcal{S}$ is the convolution

$$\mathcal{S}\varphi(x) = \int \frac{1}{2}\frac{x-y}{|x-y|}\varphi(y)\,\mathrm{d}\mu$$

where μ is an auxiliary non-negative finite measure. From [15], it becomes clear that the kneading operator is a regularized (through the convolution $\mathcal{S}$) object which describes the inverse spectrum of the transfer operator: the resolvent $(\mathrm{Id} - z\mathcal{M})^{-1}$ in [15] means that poles can only appear if $1/z$ is an eigenvalue. Since $\det^{*}(\mathrm{Id} + \widehat{\mathcal{D}}(z)) = \det^{*}(\mathrm{Id} + \mathcal{D}(z))^{-1}$, this can be translated into a statement for zeros of $\det^{*}(\mathrm{Id} + \widehat{\mathcal{D}}(z))$. The Milnor–Thurston identity [14] then implies that any zero of $\det^{\#}(\mathrm{Id} - z\mathcal{M})$ is an inverse eigenvalue of $\mathcal{M}$.

The one-dimensional kneading regularization we just presented is well understood. The higher-dimensional theory is not as developed yet. Let U_ω be now finitely many bounded open subsets of $\mathbb{R}^d$, $\psi_\omega : U_\omega \to \psi_\omega(U_\omega)$ be local C^r homeomorphisms or diffeomorphisms, while $g_\omega : U_\omega \to \mathbb{C}$ are compactly supported C^r functions, for $r \geq 1$.

In 1995, A Kitaev wrote a two-page sketch proving a higher-dimensional Milnor–Thurston formula, under an additional transversality assumption. This assumption guarantees that the set of fixed points of each fixed period m is finite, so that the Ruelle–Lefschetz determinant $\det^{\#}(\mathrm{Id} - z\mathcal{M})$ can be defined through [12]. Inspired by Kitaev's unpublished note, Baillif (2004) proved the following Milnor–Thurston formula:

$$\det{}^{\#}(\mathrm{Id} - z\mathcal{M}) = \prod_{k=0}^{d-1}\det{}^{\flat}(\mathrm{Id} + \mathcal{D}_k(z))^{(-1)^{k+1}} \quad [16]$$

Here, the $\mathcal{D}_k(z)$ are kernel operators acting on $(k+1)$-forms, constructed with the resolvent $(\mathrm{Id} - z\mathcal{M}_k)^{-1}$, together with a convolution operator $\mathcal{S}_k$, mapping

$(k+1)$-forms to k-forms and which satisfies the homotopy equation $\mathrm{d}\mathcal{S}+\mathcal{S}\mathrm{d}=1$. The kernel $\sigma_k(x,y)$ of $\mathcal{S}_k$ has singularities of the form $(x-y)/\|x-y\|^d$. The transversality assumption allows Baillif to interpret the determinant obtained by integrating the kernels along the diagonal as a flat determinant in the sense of Atiyah and Bott, whence the notation $\det^\flat$ in the right-hand side of [16].

Baillif (2004) did not give a spectral interpretation of zeros or poles of the sharp determinant [16], but he noticed that for $|z|$ very small, suitably high iterates of the $\mathcal{D}_k(z)$ are trace-class on $L^2(\mathbb{R}^d)$, showing that the corresponding regularized determinant has a nonzero radius of convergence under weak assumptions. The spectral interpretation of the sharp determinant [12] in arbitrary dimension, but under additional assumptions, was subsequently carried out by Baillif and the author of the present article, giving a new proof of some of the results in Ruelle (1990).

See also: Chaos and Attractors; Dynamical Systems and Thermodynamics; Ergodic Theory; Hyperbolic Dynamical Systems; Number Theory in Physics; Quantum Ergodicity and Mixing of Eigenfunctions; Quillen Determinant; Semi-Classical Spectra and Closed Orbits; Spectral Theory for Linear Operators.

Further Reading

Baillif M (2004) Kneading operators, sharp determinants, and weighted Lefschetz zeta functions in higher dimensions. *Duke Mathematical Journal* 124: 145–175.

Baladi V (1998) *Periodic Orbits and Dynamical Spectra, Ergodic Theory Dynam. Systems*, vol. 18, pp. 255–292 (with an addendum by Dolgopyat D and Pollicott M, pp. 293–301.)

Baladi V and Ruelle D (1996) Sharp determinants. *Inventiones Mathematicae* 123: 553–574.

Chang CH and Mayer DH (2001) An extension of the thermodynamic formalism approach to Selberg's zeta function for general modular groups. In: Fiedler B (ed.) *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, pp. 523–562. Springer: Berlin.

Cvitanović P, Artuso R, Mainieri R, Tanner G, and Vattay G (2005) *Chaos: Classical and Quantum*, ChaosBook.org. Copenhagen: Niels Bohr Institute.

Elizalde E (1995) Ten Physical Applications of Spectral Zeta Functions, *Lecture Notes in Physics*, New Series m:35. Springer: Berlin.

Fried D (1986) The zeta functions of Ruelle and Selberg. I. *Ann. Sci. École Norm. Sup* 19: 491–517.

Fried D (1986) Analytic torsion and closed geodesics on hyperbolic manifolds. *Inventiones Mathematicae* 84: 523–540.

Fried D (1995) Meromorphic zeta functions for analytic flows. *Communications in Mathematical Physics* 174: 161–190.

Gohberg I, Goldberg S, and Krupnik N (2000) *Traces and Determinants of Linear Operators*. Basel: Birkhäuser.

Guillopé L, Lin K, and Zworski M (2004) The Selberg zeta function for convex co-compact Schottky groups. *Communications in Mathematical Physics* 245: 149–176.

Kitaev AY (1999) Fredholm determinants for hyperbolic diffeomorphisms of finite smoothness. *Nonlinearity* 12: 141–179.

Liverani C (2005) Fredholm determinants, Anosov maps and Ruelle resonances. *Discrete and Continuous Dynamical Systems* 13: 1203–1215.

Pollicott M (2001) Dynamical zeta functions. In: Katok A, de la Llave R, Pesin Y, and Weiss H (eds.) *Smooth Ergodic Theory and Its Applications (Seattle, WA, 1999)*, Proc. Sympos. Pure Math., vol. 69, pp. 409–427. Providence, RI: American Mathematical Society.

Ruelle D (1976) Zeta functions for expanding maps and Anosov flows. *Inventiones Mathematicae* 34: 231–242.

Ruelle D (1990) *An Extension of the Theory of Fredholm Determinants, Inst. Hautes Études Sci. Publ. Math.* 175–193.

Ruelle D (2002) *Dynamical Zeta Functions and Transfer Operators, Notices American Mathematical Society*: 887–895.

Rugh HH (1996) *Generalized Fredholm Determinants and Selberg Zeta Functions for Axiom A Dynamical Systems, Ergodic Theory Dynam. Systems*. 805–819.

Rugh HH (1999) Intermittency and regularized Fredholm determinants. *Inventiones Mathematicae* 135: 1–25.

Voros A (1987) Spectral functions, special functions and the Selberg zeta function. *Communications in Mathematical Physics* 110: 439–465.

Relativistic Wave Equations Including Higher Spin Fields

R Illge and V Wünsch, Friedrich-Schiller-Universität Jena, Jena, Germany

Introduction

The description of phenomena at high energies requires the investigation of relativistic wave equations, that is, equations which are invariant under Lorentz transformations. Our discussion will be given classically (i.e., nonquantum). A classification of the wave equations may be based on the spin of the particles (or physical fields), which was discovered for the electron by Goudsmith and Uhlenbeck in 1925. For the greater part of physics, the three spin numbers $s=0, 1/2$, and 1 are sufficient; the respective equations named after their discoverers Klein–Gordon, Dirac, and Proca for massive fields and D'Alembert, Weyl, and Maxwell for massless fields, respectively (see the following section).

In their original form, these equations look rather different. However, their translation into spinor form shows that the wave equations for bosons and fermions

have the same structure, if $s > 0$. Therefore, most of the equations dealt with in this article are formulated for spinor fields. (Strictly speaking, the exclusive use of 2-spinors restricts the relativistic invariance to the proper Lorentz group $SO^+(1,3)$. However, all the results presented here can be "translated back" into tensor or bispinor form, respectively (Illge 1993).) Relativistic wave equations for free fields with arbitrary spin $s > 0$ in Minkowski spacetime are discussed in the section "Higher spin in Minkowski spacetime"; they were first given by Dirac (1936).

In the subsequent section, we explain how the field theory can be extended to curved spacetimes. If a Lagrangian is known, then there exists a well-known mathematical procedure ("Lagrange formalism") to obtain the field equations, the energy–momentum tensor, etc. All field equations for "low" spin $s \leq 1$ arise from an action principle. Consequently, they can be extended to curved spacetime by simply replacing the flat metric and connection with their curved versions.

If $s > 1$, then the wave equations do not follow from a variation principle without supplementary conditions. Nevertheless, one can try to generalize the equations of the section "Higher spin in Minkowski spacetime" to curved spacetime by the "principle of minimal coupling," too. However, the arising equations are not satisfactory, since there is an algebraic consistency condition in curved space if $s > 1$ (Buchdahl 1962), and another for charged fields in the presence of electromagnetism if $s > 1/2$ (Fierz and Pauli 1939).

There have been numerous attempts to avoid these inconsistencies. As a rule, the alternative theories require an extended spacetime structure or additional new fields or they give up some important principle. An extensive literature is devoted to just this problem – unfortunately, a survey article or book is missing.

Finally, we present a possibility to describe fields with arbitrary spin $s > 0$ within the framework of Einstein's general relativity without any auxiliary fields and subsidiary conditions in a uniform manner. The approach is based on irreducible representations of type $D(s,0)$ and $D(s-1/2,1/2)$ instead of $D(s/2,s/2)$ in the Fierz theory for bosons and $D(s/2+1/4,s/2-1/4)$ in the Rarita–Schwinger theory for fermions. It was first pointed out by Buchdahl (1982) that this type of field equations can be generalized to a curved spacetime if the mass is positive. After a short time Wünsch (1985) simplified them to their final form:

$$\begin{aligned} \nabla^{A}_{P'}\varphi_{AB\ldots E} + m_1\chi_{B\ldots EP'} &= 0 \\ \nabla^{P'}_{(A}\chi_{B\ldots E)P'} - m_2\varphi_{AB\ldots E} &= 0 \end{aligned} \qquad [1]$$

This system contains the well-known wave equations for low spin $s = 1/2$ and $s = 1$ as special cases. By iteration we obtain second-order wave equations of normal hyperbolic type. Further, Cauchy's initial-value problem is well posed and a Lagrangian is known. For zero mass, we state the wave equations

$$\nabla^{A}_{(A'}\Theta_{|A|B'\ldots E')} = 0 \qquad [2]$$

which are just the curved versions of the equations for the potential of a massless field. They are consistent in curved spacetime, too, and the Cauchy problem is well posed (Illge 1988).

Last but not least, let us mention the esthetic aspect. Equations [1] and [2] satisfy Dirac's demand: "Physical laws should have mathematical beauty."

In the following, we assume that the spacetime and all the spinor and tensor fields are of class C^∞. All considerations are purely local. We will call a symmetric ("irreducible") spinor to be of type (n,k) if and only if it has n unprimed and k primed indices (irrespective of their position). Moreover, we use the notations and conventions of Penrose and Rindler (1984), especially for the curvature spinors Ψ_{ABCD} and $\Phi_{ABA'B'}$.

Wave Equations for Low Spin in Minkowski Spacetime

The spin (or intrinsic angular momentum) of a particle is found to be quantized. Its projection on any fixed direction is an integer or half-integer multiple of Planck's constant $\hbar$; the only possible values are

$$-s\hbar,\ (-s+1)\hbar,\ \ldots,\ (s-1)\hbar,\ s\hbar$$

The spin quantum number s so defined can have one of the values $s = 0, 1/2, 1, 3/2, 2, \ldots$ and is a characteristic for all elementary particles along with their mass m and electric charge e. The particles with integer s are called "bosons," those with half-integer s "fermions." The three numbers $s = 0, 1/2$, and 1 are referred to as "low" spin; they are sufficient for the greater part of physics.

The principle of first quantization associates a type of field and a field equation to each type of elementary particles. Massive particles, with rest mass $m > 0$, and massless particles, with rest mass $m = 0$, are to be distinguished. Accordingly, we obtain six linear wave equations for $s \leq 1$, which read as follows in units such that $c = \hbar = 1$ (see **Table 1**):

For the sake of simplicity, we consider only free fields in **Table 1**; no source terms or interaction terms appear here. The associated "free" Lagrangians are given in **Table 2**.

Since the electromagnetic field tensor F_{ab} satisfies the first part of Maxwell's equations $\partial_{[c}F_{ab]} = 0$, it follows

Table 1 Relativistic wave equations for low spin $s = 0, 1/2$, and 1

Spin, mass	*Wave equation*	*Associated particles*
$s=0,\ m>0$	Klein–Gordon eqn. $(\approx + m^2)u = 0$	Scalar mesons $\pi, \eta, K, \ldots$
$s=0,\ m=0$	D'Alembert eqn. $\approx u = 0$	–
$s=1/2,\ m>0$	Dirac eqn. $\partial^A_{A'}\varphi_A + \frac{im}{\sqrt{2}}\chi_{A'} = 0$ $\partial^{A'}_{A}\chi_{A'} - \frac{im}{\sqrt{2}}\varphi_A = 0$	Leptons e, μ, τ Baryons $p, n, \Lambda, \Xi, \Sigma, \ldots$
$s=1/2,\ m=0$	Weyl eqn. $\partial^A_{A'}\nu_A = 0$	Massless(?) neutrinos ν_e, ν_μ, ν_τ
$s=1,\ m>0$	Proca eqn. $H_{ab} = \partial_a U_b - \partial_b U_a$ $\partial^c H_{ca} + m^2 U_a = 0$	Vector mesons $\rho, \omega, \psi, \Phi, \ldots$
$s=1,\ m=0$	Maxwell eqn. $\partial_{[a}F_{bc]} = 0$ $\partial_a F^{ab} = 0$	Photon γ

Table 2 The Lagrangian densities for free (i.e., noninteracting) fields with low spin

Field	*Lagrangian density*
Scalar field	$\mathfrak{L} = \frac{1}{2}\{(\partial^a u)(\partial_a u) - m^2 u^2\}$
Dirac field	$\mathfrak{L} = \frac{i}{\sqrt{2}}(\bar{\chi}_A \partial^{AA'}\chi_{A'} + \bar{\varphi}^{B'}\partial_{BB'}\varphi^B - \varphi^B\partial_{BB'}\bar{\varphi}^{B'} - \chi_{A'}\partial^{AA'}\bar{\chi}_A) + m(\bar{\chi}_A\varphi^A + \bar{\varphi}^{A'}\chi_{A'})$
Weyl field	$\mathfrak{L} = \frac{i}{\sqrt{2}}(\bar{\nu}_{A'}\partial^{AA'}\nu_A - \nu_A\partial^{AA'}\bar{\nu}_{A'})$
Proca field	$\mathfrak{L} = \frac{1}{4}H_{ab}H^{ab} - H^{ab}\partial_{[a}U_{b]} + \frac{m^2}{2}U_a U^a$
Maxwell field	$\mathfrak{L} = -\frac{1}{4}F_{ab}F^{ab} = -(\partial_{[a}A_{b]})(\partial^{[a}A^{b]})$

that a vector field A_a exists such that $F_{ab} = \partial_a A_b - \partial_b A_a$. This vector field is called the "electromagnetic 4-potential." It is not uniquely determined by the field F_{ab}; the freedom in A_a is $A_a \to A_a + \partial_a\tau$ where $\tau = \tau(x)$ is a real-valued function. This gauge transformation of A_a can be used, for example, to obtain the Lorentz gauge condition $\partial^a A_a = 0$.

The wave equations listed in Table 1 look rather different, but this formal disadvantage can be overcome. To begin with, we remark that fermions require spinors for their description. The Dirac and Weyl equations are not describable by linear equations for tensor fields. On the other hand, bosons can be described by spinors as well. All tensor equations can be "translated" into spinor form using the mixed spinor–tensor $\sigma^a_{AA'}$. We will demonstrate this procedure for the Proca field in some detail.

The (possibly complex) skew-symmetric tensor H_{ab} and the vector U_a have the spinor equivalents

$$H_{ab}\sigma^a_{AA'}\sigma^b_{BB'} = \varphi_{AB}\varepsilon_{A'B'} + \xi_{A'B'}\varepsilon_{AB}$$
$$U_a\sigma^a_{AA'} = \chi_{AA'}$$

where φ and ξ are both symmetric spinors: $\varphi_{AB} = \varphi_{(AB)}$, $\xi_{A'B'} = \xi_{(A'B')}$. After a straightforward calculation the Proca equation yields

$$\partial^{C'}_{(A}\chi_{B)C'} + \varphi_{AB} = 0, \qquad \partial^{C'}_{(A'}\chi_{B')C} + \xi_{A'B'} = 0$$
$$\partial^C_{A'}\varphi_{CA} + \partial^{C'}_{A}\xi_{C'A'} + m^2\chi_{AA'} = 0$$

Further, from the equation $\partial_{[c}H_{ab]} = 0$, we obtain $\partial^{C'}_{A}\xi_{A'C'} = \partial^C_{A'}\varphi_{AC}$; thus, the first and second summand in the third equation are equal. Consequently, we find the following spinor form of the Proca equations:

$$\partial^C_{A'}\varphi_{CA} + \frac{m^2}{2}\chi_{AA'} = 0, \qquad \partial^{C'}_{(A}\chi_{B)C'} + \varphi_{AB} = 0$$
$$\partial^{C'}_{A}\xi_{C'A'} + \frac{m^2}{2}\chi_{AA'} = 0, \qquad \partial^{C'}_{(A'}\chi_{B')C} + \xi_{A'B'} = 0 \qquad [3]$$

If the tensor fields H and U are real, then we have $\xi_{A'B'} = \bar{\varphi}_{A'B'}$, $\chi_{AA'} = \bar{\chi}_{AA'}$, and the second pair of equations is just the complex conjugate of the first.

Now it is readily seen that the Dirac and Proca equations have the same structure. They are coupled first-order systems of differential equations for pairs of spinor fields. The only decisive difference is that the spinors have one index if $s = 1/2$ and two indices if $s = 1$.

We obtain a similar result for Maxwell fields. The real tensor F_{ab} has the spinor equivalent

$$F_{ab}\sigma^a_{AA'}\sigma^b_{BB'} = \varphi_{AB}\varepsilon_{A'B'} + \bar{\varphi}_{A'B'}\varepsilon_{AB}$$

with a symmetric spinor φ_{AB}. The spinor form of Maxwell's equations is (Penrose and Rindler 1984)

$$\partial^A_{A'}\varphi_{AB} = 0 \qquad [4]$$

and has the same structure as the Weyl equation.

Here we found an example for the power and utility of spinor techniques since they allow the formulation of the wave equations for bosons and fermions in a uniform manner. Only the cases $m > 0$ and $m = 0$ are to be distinguished. Moreover, the above results suggest the way for generalizing the wave equations to higher spin. Therefore, we can already end the discussion of the fields with low spin and take them as special cases of those with arbitrary spin.

Higher Spin in Minkowski Spacetime

Massive Fields

Relativistic wave equations for particles with arbitrary spin were first considered by Dirac (1936). His equations read

$$\partial^A_{P'}\varphi_{AB\ldots DQ'\ldots T'} + m_1\chi_{B\ldots DP'Q'\ldots T'} = 0$$
$$\partial^{P'}_{A}\chi_{B\ldots DP'Q'\ldots T'} - m_2\varphi_{AB\ldots DQ'\ldots T'} = 0 \qquad [5]$$

where the spinors φ and χ are of type (n, k) and $(n-1, k+1)$, respectively (corresponding to irreducible representations of the restricted Lorentz group $SO^+(1,3)$). The constants m_1 and m_2 are mass parameters ($m^2 = -2m_1m_2$) and the spin s is one half of the total number of indices of each spinor, $s = (1/2)(n+k)$. As in the preceding section, we assume that electromagnetism and other interactions are absent. We should mention that equations for higher spin were not motivated by observations or empirical facts in that period of time, because only a few elementary particles were known (proton, neutron, electron, positron, and photon), and all of them have low spin (see **Table 1**). Since that time, particles with $s > 1$ were found in nature, for example, resonances in scattering experiments.

The system [5] allows a uniform description of free fields with arbitrary spin $s > 0$, including Dirac and Proca fields, as we know from the preceding section. (Remark: The symmetrization in eqns [3] can be omitted since the vector field U is divergence-free as a consequence of the second Proca equation.) Various other field equations proposed subsequently can be comprehended as its special cases (Corson 1953). Examples are the Rarita–Schwinger equations for fermions: if they are written in terms of 2-spinors, then one obtains just the system [5] where the spinor φ is of type $(s+1/2, s-1/2)$ and the spinor χ is of type $(s-1/2, s+1/2)$.

If we apply $\partial_E^{P'}$ to the first of the equations in [5] and use the second, we obtain

$$(\Box + m^2)\varphi_{AB\ldots DQ'\ldots T'} = 0 \qquad [6a]$$

since the second derivatives commute in flat spacetimes. Similarly,

$$(\Box + m^2)\chi_{B\ldots DP'Q'\ldots T'} = 0 \qquad [6b]$$

so both fields φ and χ satisfy a Klein–Gordon type equation. Moreover, eqns [5] imply that each of φ and χ is divergence-free

$$\partial^{AQ'}\varphi_{AB\ldots DQ'\ldots T'} = 0 = \partial^{BP'}\chi_{B\ldots DP'Q'\ldots T'} \qquad [7]$$

if they have at least one index of each kind.

In a sense, this procedure can be reversed. Let a symmetric spinor field φ be given that satisfies [6a] and [7]. (Remark: A significant example is the Fierz system

$$(\Box + m^2)U_{ab\ldots d} = 0, \qquad \partial^a U_{ab\ldots d} = 0$$

for a symmetric, tracefree tensor field U, since the spinor equivalent of U is of type (k, k).)

Define

$$\chi_{B\ldots DP'Q'\ldots T'} := \partial^A_{P'}\varphi_{AB\ldots DQ'\ldots T'}$$

Then χ is symmetric in all its indices since φ is divergence-free. Further, we obtain

$$\begin{aligned}\partial_E^{P'}\chi_{B\ldots DP'Q'\ldots T'} &= \partial_E^{P'}\partial^A_{P'}\varphi_{AB\ldots DQ'\ldots T'} \\ &\equiv -\frac{1}{2}\Box\varphi_{EB\ldots DQ'\ldots T'} \\ &= \frac{m^2}{2}\varphi_{EB\ldots DQ'\ldots T'}\end{aligned}$$

since φ satisfies the Klein–Gordon equation [6a]. Consequently, the pair (φ, χ) satisfies a system [5]. Obviously, this procedure can be continued: define

$$\eta_{C\ldots DO'P'Q'\ldots T'} := \partial^B_{O'}\chi_{B\ldots DP'Q'\ldots T'}$$

etc. We obtain a sequence of spinors of type $(0, 2s), (1, 2s-1), \ldots, (2s, 0)$ each of which is obtainable from its immediate neighbors by a differentiation contracted on one index. Together, these spinors form an invariant exact set (Penrose and Rindler 1984).

The just given arguments show that there is an ambiguity in the system [5]. The spin s fixes only the total number of indices of φ and χ. However, their partition into primed and unprimed ones is not *a priori* fixed. Therefore, we can choose a "convenient" partition for the respective needs.

Massless Fields

If $m = 0$, then the Dirac system [5] is decoupled. Therefore, we have to state a single equation for a single field. Let φ be a spinor field of type $(n, 0)$. The massless free-field equation for spin $(1/2)n$ is then taken to be

$$\partial^A_{A'}\varphi_{AB\ldots E} = 0 \qquad [8]$$

More precisely, the solutions of [8] represent left-handed massless particles with helicity $-(1/2)n\hbar$, whereas the solutions of the complex-conjugate form of this equation are right-handed particles (helicity $+(1/2)n\hbar$). Recall that the Weyl equation ($n = 1$) and the source-free Maxwell equation ($n = 2$) have this form. (Remark: The Bianchi identity in Einstein spaces also falls in this category, with the Weyl spinor Ψ_{ABCD} taking the place of $\varphi\ldots$. Moreover, we may think of [8] with $n = 4$ as the gauge-invariant equation for the weak vacuum gravitational field.)

The massless field equation [8] can be solved using methods of twistor geometry. Moreover, there is an explicit integral formula for representing massless free fields in terms of arbitrarily chosen null data on a light cone (Penrose and Rindler 1984, 1986, Ward and Wells 1990). We do not discuss eqns [8] in detail since they are generally inconsistent in curved spacetimes if $n > 2$ (see the next

section). We only indicate that each solution of [8] satisfies the second-order wave equation

$$\Box\varphi_{AB\ldots E} = 0$$

Maxwell's equations imply the existence of an electromagnetic potential (cf. section "Wave equations for low spin in Minkowski spacetime"). This concept can be generalized to higher spin. A "potential" for a spinor field $\varphi_{AB\ldots E}$ of type $(n, 0)$ is a spinor field $\Theta_{AB'\ldots E'}$ of type $(1, n-1)$ such that

$$\partial^A_{(A'}\Theta_{|A|B'\ldots E')} = 0 \qquad [9]$$

and

$$\varphi_{AB\ldots E} = \partial^{B'}_{(B}\cdots\partial^{E'}_{E}\Theta_{A)B'\ldots E'} \qquad [10]$$

One can check in a straightforward manner that a spinor field φ that is given by [9] and [10] satisfies the massless equation [8]. If $n > 1$, there is a gauge freedom in these potentials; it turns out to be

$$\Theta_{AB'\ldots E'} \to \Theta_{AB'\ldots E'} + \partial_{A(B'}\omega_{C'\ldots E')}$$

for any spinor field ω of type $(0, n-2)$. Furthermore, the general massless field φ can locally be expressed in this way (Penrose and Rindler 1986).

Wave Equations in Curved Spacetimes, Consistency Conditions

First of all we emphasize that Hamilton's principle of stationary action is extremely important in field theories (see, e.g., Schmutzer (1968)). Assume that the Lagrangian $\mathfrak{L}$ contains at most first derivatives of a field $\psi_\Sigma : \mathfrak{L} = \mathfrak{L}(\psi_\Sigma(x), \partial_a\psi_\Sigma(x))$. "Special relativity" states that $\mathfrak{L}$ is invariant under Lorentz transformations. The Euler–Lagrange equations with respect to variation of ψ_Σ read

$$\frac{\partial\mathfrak{L}}{\partial\psi_\Sigma} - \partial_a\frac{\partial\mathfrak{L}}{\partial(\partial_a\psi_\Sigma)} = 0 \qquad [11]$$

and these are the field equations that ψ_Σ is required to satisfy.

In "general relativity," the Lagrangian $\mathfrak{L}$ has to be generally covariant. So we have $\mathfrak{L} = \mathfrak{L}(\psi_\Sigma(x), \nabla_a\psi_\Sigma(x))$ and the Euler–Lagrange equations

$$\frac{\partial\mathfrak{L}}{\partial\psi_\Sigma} - \nabla_a\frac{\partial\mathfrak{L}}{\partial(\nabla_a\psi_\Sigma)} = 0 \qquad [12]$$

emerge. If we assume that the Lagrangian $\mathfrak{L}$ does not contain the curvature tensors and their derivatives explicitly and compare [11] and [12], then it is easily seen how the wave equations in curved spacetime can be obtained: by simply replacing the flat metric and connection with their curved versions. This procedure is called the "principle of minimal coupling."

All equations for low spin in Minkowski spacetime are the Euler–Lagrange equations of a variation principle (see Table 2). Consequently, they can be extended to curved spacetime by simply using the principle of minimal coupling. The arising equations are perfectly acceptable. No complications arise, and so we do not repeat them in this section.

If $s > 1$, then neither the massive nor the massless wave equations follow from a variation principle without supplementary conditions. Nevertheless, we can try to generalize the equations of the previous section to a curved spacetime by formally replacing the flat metric and connection with their curved versions, too. However, serious problems arise:

Let us first consider massless fields of helicity $-(1/2)n\hbar$. The principle of minimal coupling yields

$$\nabla^A_{A'}\varphi_{AB\ldots E} = 0 \qquad [13]$$

If we apply $\nabla^{A'}_F$ to this equation, we obtain

$$\nabla^{A'}_F\nabla^A_{A'}\varphi_{AB\ldots E} = 0$$

Since the covariant derivatives do not commute with each other, the term on the left-hand side is not completely symmetric in the unprimed indices. Therefore, this equation can be decomposed into two nontrivial irreducible parts if $n > 1$: symmetrization yields the covariant D'Alembert equation

$$\nabla^a\nabla_a\varphi_{B\ldots EF} = 0$$

as required, while antisymmetrization yields by use of the spinor Ricci identities

$$(n-2)\Psi^{KLM}{}_{(C}\varphi_{D\ldots E)KLM} = 0 \qquad [14]$$

where Ψ_{ABCD} is the Weyl spinor. If $n > 2$ and the spacetime is not conformally flat, then this algebraic consistency condition effectively renders eqn [13] useless as physical field equations.

If $m > 0$, the situation is not better. In somewhat similar way, we obtain the algebraic consistency conditions

$$\begin{aligned}&(n-2)\Psi^{KLM}{}_{(C}\varphi_{D\ldots E)KLMQ'P'\ldots T'}\\&\quad + k\Phi^{KLX'}{}_{(Q'}\varphi_{|KLC\ldots E|P'\ldots T')X'} = 0 \quad (n > 1)\\&(k-1)\bar\Psi^{X'Y'Z'}{}_{(S'}\chi_{|B\ldots DX'Y'Z'|T'\ldots U')}\\&\quad + (n-1)\Phi_{(B}{}^{KX'Y'}\chi_{C\ldots D)KX'Y'S'T'\ldots U'} = 0 \quad (k > 0)\end{aligned} \qquad [15]$$

if the spinor field φ is of type (n, k) (Buchdahl 1962).

We remark that similar consistency conditions occur if we have no gravitation, but an interaction with an electromagnetic field. Then the partial

derivative is to be replaced by $D_a = \partial_a - ieA_a$ and we obtain consistency conditions like [14] and [15], where the curvature spinors are to be replaced by the electromagnetic spinor (Fierz and Pauli 1939).

So far one is left with the problem: "Find the 'correct' laws for arbitrary spin, that means field equations which coincide with the well-known approved ones for low spin and which remain consistent even for higher spin when electromagnetism and/or gravitation is coupled!"

An extensive literature is devoted to just this problem. Let us briefly sketch some means by which the authors tried to solve it:

- derivation of the desired field equations from a variation principle where the original spinor fields are supplemented by auxiliary fields;
- extension of the four-dimensional spacetime geometry to a richer one: higher number of dimensions, complexification, addition of torsion, nonmetrical connection, ...;
- replacement of the algebra of spinors by some richer algebra;
- disclaim of the principle of minimal coupling; and
- supergravity theories.

Some of these attempts are able to solve the problem, at least partially. But, as a rule, they pay a price of new difficulties. In the next section, we offer "good" equations for arbitrary $s > 0$ within the conventional framework of the minimal coupling principle and of a curved spacetime background.

Wave Equations for Arbitrary Spin without Consistency Conditions

Massive Fields

The ansatz which leads to the desired result is surprisingly simple. We avoid the ambiguity in the Dirac system [5] that has been discussed earlier as well as any consistency condition if we state the wave equations

$$\begin{aligned} \nabla^A_{P'}\varphi_{AB\ldots E} + m_1\chi_{B\ldots EP'} &= 0 \\ \nabla^{P'}_{(A}\chi_{B\ldots E)P'} - m_2\varphi_{AB\ldots E} &= 0 \end{aligned} \qquad [16]$$

This system was first proposed by Wünsch (1985); it is equivalent to a pair of equations given by Buchdahl (1982) which contains the Weyl spinor explicitly. As before, φ and χ are symmetric spinor fields, φ has n unprimed indices (and no one else!) and the constants m_1, m_2 are mass parameters ($m^2 = -2m_1m_2$). We assume $m_1 \neq 0$ in this section. Obviously, the Dirac and Proca equations are special cases of [16], choose $n=1$ and $n=2$, respectively. (Remark: An electromagnetic field can be included in [16] by $\nabla_a \to D_a = \nabla_a - ieA_a$, and the equations remain consistent (Illge 1993).)

First of all, we remark that eqns [16] are the Euler–Lagrange equations of an action principle. The existence of a Lagrangian is plausible since the number of equations and the number of degrees of freedom are equal. We do not state the Lagrangian, the energy–momentum tensor, and the current vector in this article and refer the reader to Illge (1993).

If $n > 1$, we can apply $\nabla^{BP'}$ to the first equation of [16] and obtain using the spinor Ricci identities:

$$\begin{aligned} \nabla^{BP'}\chi_{BC\ldots EP'} &= -\frac{1}{m_1}\nabla^{BP'}\nabla^A_{P'}\varphi_{ABC\ldots E} \\ &= -\frac{n-2}{m_1}\Psi^{KLM}{}_{(C}\varphi_{D\ldots E)KLM} \end{aligned} \qquad [17]$$

Hence the divergence of χ vanishes if $n=2$ or if the spacetime is conformally flat. These are exactly the cases where the symmetrization in the second equation of [16] can be omitted.

Now we are going to derive the second-order equations for φ and χ. Substituting

$$\chi_{BC\ldots EP'} = -\frac{1}{m_1}\nabla^A_{P'}\varphi_{AB\ldots E} \qquad [18]$$

into the second equation of [16], we obtain, after a bit of algebra,

$$\begin{aligned} &\nabla^a\nabla_a\varphi_{AB\ldots E} - 2(n-1)\Psi^{KL}{}_{(AB}\varphi_{C\ldots E)KL} \\ &\quad + \left(\frac{n+2}{12}R + m^2\right)\varphi_{AB\ldots E} = 0 \end{aligned} \qquad [19]$$

This is a linear second-order equation of normal hyperbolic type for the spinor field φ. It can be used to solve Cauchy's problem for the system [16].

Similarily, we get a second-order equation for χ:

$$\begin{aligned} &\nabla^a\nabla_a\chi_{B\ldots EP'} - 2(n-1)\Phi_{(B}{}^K{}_{P'}{}^{W'}\chi_{C\ldots E)KW'} \\ &\quad + \left(\frac{R}{4} + m^2\right)\chi_{B\ldots EP'} \\ &\quad = 2\frac{n-1}{n}\nabla_{(BP'}\nabla^{KW'}\chi_{C\ldots E)KW'} \end{aligned} \qquad [20]$$

Seemingly this is not an equation of hyperbolic type if $n > 1$. However, the second derivatives of χ on the right-hand side of [20] can be eliminated using [17]. Therefore, if the spinor field φ is already known by solving [19], then [20] is an equation of Klein–Gordon type, too. However, it is generally inhomogeneous if $n > 2$. A wave equation that contains the spinor field χ alone exists only if $n=1, n=2$, or the spacetime is conformally flat.

Now we are going to discuss the "Cauchy problem" for the wave equations [16] (for details see Wünsch (1985)). Let a spacelike hypersurface S be given and let n^a denote the future-directed unit normal vector on S and $\nabla_n = n^a \nabla_a$. The local Cauchy problem is to find a solution (φ, χ) of [16] with given Cauchy data φ^0, χ^0 on S.

In general, the initial data φ^0 and χ^0 cannot be prescribed arbitrarily. Suppose that a solution (φ, χ) of [16] does exist. Then the differential equations have to be satisfied on S, too. Thus, we obtain

$$(\nabla_n \varphi_{AB\ldots E})|_S = 2n_A^{A'}\left(\tilde{\nabla}^F_{A'}\varphi_{B\ldots EF} + m_1 \chi_{B\ldots EA'}\right)|_S \quad [21]$$

where the differential operator $\tilde{\nabla}_{AA'} = \nabla_{AA'} - n_{AA'}\nabla_n$ is just the tangential part of $\nabla_{AA'}$ with respect to S. Therefore, the right-hand side of [21] is completely determined by the initial data. Now the symmetry of the solution $\varphi_{AB\ldots E}$ implies the symmetry of $\nabla_n \varphi_{AB\ldots E}$. Consequently, the right-hand side of [21] has to be symmetric with respect to the unprimed indices and so we obtain the following constraints for the initial data if φ has at least two indices:

$$n^{BA'}\left(\tilde{\nabla}^F_{A'}\varphi^0{}_{B\ldots EF} + m_1 \chi^0{}_{B\ldots EA'}\right)|_S = 0 \quad [22]$$

Now we can state:

Theorem 1 *If the Cauchy data φ^0 and χ^0 satisfy the constraints [22], then the Cauchy problem has a unique solution in a neighborhood of S.*

For each differential equation of hyperbolic type we can ask the question whether the wave propagation is "sharp," that is, free of tails. If this property is valid we say that the equation satisfies "Huygens' principle" (for an exact definition, see, e.g., Wünsch (1994)). Using invariant Taylor expansions of the parallel propagator and of the Riesz kernels in normal coordinates we can prove (Wünsch 1985):

Theorem 2 *The massive wave equations [16] for spin $s > 0$ satisfy Huygens' principle if and only if the spacetime is of constant curvature and $R = -(6m^2/s)$.*

Massless Fields

In the preceding section, we have seen that the premise $m_1 \neq 0$ is decisive for the consistency of [16] if $s > 1$. This fact agrees with the result of the previous section, that eqn [13] is inconsistent if $s > 1$ and the spacetime is not conformally flat. On the other hand, $m_2 = 0$ is possible. Therefore we state the wave equations

$$\nabla^A_{(A'}\Theta_{|A|B'\ldots E')} = 0 \quad [23]$$

for a spinor field Θ of type $(1, n-1)$. This is just eqn [9] for the potential of a massless field. We will show that [23] is a satisfactory equation in a generally curved spacetime (Illge 1988). Unfortunately, no Lagrangian has been found if $n > 1$.

To begin with, we remark that there is a gauge freedom in curved spacetimes, too, since the solution Θ of [23] cannot be uniquely determined if $n > 1$. We use this freedom to prescribe the divergence of Θ. So let an arbitrary spinor field ω of type $(0, n-2)$ be given. We consider eqns [23] and

$$\nabla^{AB'}\Theta_{AB'C'\ldots E'} = \omega_{C'\ldots E'}$$

or, together,

$$\nabla^A_{A'}\Theta_{AB'\ldots E'} = -\frac{n-1}{n}\varepsilon_{A'(B'}\omega_{C'\ldots E')} \quad [24]$$

If we apply $\nabla^{A'}_B$ to this equation, we obtain using the spinor Ricci identities

$$\nabla^a\nabla_a\Theta_{BB'\ldots E'} - 2(n-1)\Phi_B{}^K{}_{(B'}{}^{W'}\Theta_{|K|C'\ldots E')W'} + \frac{R}{4}\Theta_{BB'\ldots E'} = \frac{2(n-1)}{n}\nabla_{B(B'}\omega_{C'\ldots E')} \quad [25]$$

This is a linear second-order equation of normal hyperbolic type for the spinor field Θ (cf. [20]).

Now let us discuss some particular cases. If $n = 1$, then [23] is just the Weyl equation itself. Therefore, the equations for the field and its potential are identical and there is no gauge freedom. If $n = 2$, then the spinor field $\Theta_{AA'}$ is a (complex) vector field and eqn [23] yields

$$\nabla^A_{(A'}\Theta_{|A|B')} = 0$$

The gauge field ω is just a scalar function, especially we can choose $\omega = 0$ (Lorentz gauge). As in eqn [10] we define the field spinor as

$$\varphi_{AB} = \nabla^{B'}_{(A}\Theta_{B)B'}$$

Since we have the identity

$$\nabla^A_{B'}\nabla^{A'}_{(B}\eta_{A)A'} = \nabla^{A'}_B\nabla^A_{(B'}\eta_{|A|A')}$$

for arbitrary spinor fields $\eta_{AA'}$ (which must not have additional free indices!), the spinor field φ_{AB} satisfies the massless free-field equation

$$\nabla^A_{B'}\varphi_{AB} = 0$$

If $n > 2$, we can define a field $\varphi_{AB\ldots E}$ via the relation [10], too, replacing the partial with the covariant derivatives. But the field equation for $\varphi_{AB\ldots E}$ becomes more complicated than [13]. This

fact is not surprising, since eqn [23] is a consistent one, whereas [13] is inconsistent.

We continue with some remarks on "conformal rescalings of the metric." The equations for massless fields have to be invariant with respect to such transformations. Therefore, the "curved space" scalar wave equation is

$$\left(\Box + \frac{R}{6}\right)\varphi = 0 \qquad [26]$$

Further, the equations

$$\nabla^{A}_{(A'}\eta_{|AB\ldots E|B'\ldots F')} = 0 \qquad [27]$$

for any spinor field η of type (n, k) are conformally invariant (Penrose and Rindler 1984). Especially, eqns [23] for the massless potential and [13] for the massless field have this property.

We mention a further special case of [27]. If η is of type $(k+1, k)$, then these equations are consistent, too (Frauendiener and Sparling 1999). The Cauchy problem is well posed and a Lagrangian is known. Unfortunately, the solutions do not satisfy a wave equation of second order if $k > 0$.

We conclude with the discussion of the Cauchy problem for eqn [24]. As in the preceding section, let a spacelike hypersurface S and initial data Θ^0 on S be given. We can state:

Theorem 3 *If a symmetric spinor field ω of type $(0, n-2)$ is given, then there exists a neighborhood of S in which eqn [24] has one and only one solution satisfying $\Theta_{|S} = \Theta^0$.*

The proof is given in Illge (1988). We emphasize that there are no constraints on the Cauchy data for the massless equation [24].

In contrast to massive fields we are far away from an answer to the question whether Huygens principle is valid for the massless equations. A particular result is Wünsch (1994):

Theorem 4 *Huygen's principle for the conformally invariant scalar wave equation [26], the Weyl, and the Maxwell equations is valid only for conformally flat and plane wave metrics within the classes of centrally symmetric, recurrent, $(2,2)$-decomposable, Petrov type N, III or D spacetimes as well as those with $\nabla_{[a}R_{b]c} = 0$.*

See also: Clifford Algebras and Their Representations; Dirac Fields in Gravitation and Nonabelian Gauge Theory; Euclidean Field Theory; Evolution Equations: Linear and Nonlinear; Spinors and Spin Coefficients; Standard Model of Particle Physics; Twistors.

Further Reading

Buchdahl HA (1962) On the compatibility of relativistic wave equations in Riemann spaces. *Nuovo Cimento* 25: 486–496.

Buchdahl HA (1982) On the compatibility of relativistic wave equations in Riemann spaces II. *Journal of Physics A* 15: 1–5.

Corson EM (1953) *Introduction to Tensors, Spinors, and Relativistic Wave-Equations*. London and Glasgow: Blackie and Son Ltd.

Dirac PAM (1936) Relativistic wave equations. *Proceedings of the Royal Society London Series A* 155: 447–459.

Fierz M and Pauli W (1939) On relativistic wave equations for particles of arbitrary spin in an electromagnetic field. *Proceedings of the Royal Society London Series A* 173: 211–232.

Frauendiener J and Sparling GAJ (1999) On a class of consistent higher spin equations on curved manifolds. *Journal of Geometry and Physics* 30: 54–101.

Greiner W (1997) *Relativistic Quantum Mechanics – Wave Equations*, 2nd edn. Berlin: Springer.

Illge R (1988) On potentials for several classes of spinor and tensor fields in curved spacetimes. *General Relativity and Gravitation* 20: 551–564.

Illge R (1993) Massive fields of arbitrary spin in curved spacetimes. *Communications in Mathematical Physics* 158: 433–457.

Penrose R and Rindler W (1984) *Spinors and Space-Time, Two-Spinor Calculus and Relativistic Fields*, vol. 1. Cambridge: Cambridge University Press.

Penrose R and Rindler W (1986) *Spinors and Space-Time, Spinor and Twistor Methods in Space-Time Geometry*, vol. 2. Cambridge: Cambridge University Press.

Schmutzer E (1968) *Relativistische Physik*. Leipzig: Teubner-Verlag.

Ward RS and Wells RO (1990) *Twistor Geometry and Field Theory*. Cambridge: Cambridge University Press.

Wünsch V (1985) Cauchy's problem and Huygens' principle for relativistic higher spin wave equations in an arbitrarily curved space-time. *General Relativity and Gravitation* 17: 15–38.

Wünsch V (1994) Moments and Huygens' principle for conformally invariant field equations in curved space-times. *Annales de l'Institute Henri Poincaré – Physique théorique* 60: 433–455.

Renormalization: General Theory

J C Collins, Penn State University, University Park, PA, USA

Introduction

Quantum field theories (QFTs) provide a natural framework for quantum theories that obey the principles of special relativity. Among their most striking features are ultraviolet (UV) divergences, which at first sight invalidate the existence of the theories. The divergences arise from Fourier modes of very high wave number, and hence from the structure of the theories at very short distances. In the very restricted class of theories called "renormalizable," the divergences may be removed by a singular redefinition of the parameters of the theory. This is the process of renormalization that defines a QFT as a nontrivial limit of a theory with a UV cutoff.

A very important QFT is the standard model, an accurate and successful theory for all the known interactions except gravity. Calculations using renormalization and related methods are vital to the theory's success.

The basic idea of renormalization predates QFT. Suppose we treat an observed electron as a combination of a bare electron of mass m_0 and the associated classical electromagnetic field down to a radius a. The observed mass of the electron is its bare mass plus the energy in the field (divided by c^2). The field energy is substantial, for example, 0.7 MeV when $a = 10^{-15}$ m, and it diverges when $a \to 0$. The observed mass, 0.5 MeV, is the sum of the large (or infinite) field contribution compensated by a negative and large (or infinite) bare mass. This calculation needs replacing by a more correct version for short distances, of course, but it remains a good motivation.

In this article, we review the theory of renormalization in its classic form, as applied to weak-coupling perturbation theory, or Feynman graphs. It is this method, rather than the Wilsonian approach (*see* Exact Renormalization Group), that is typically used in practice for perturbative calculations in the standard model, especially its QCD part.

Much of the emphasis is on weak-coupling perturbation theory, where there are well-known algorithmic rules for performing calculations and renormalization. Applications (*see* Quantum Chromodynamics for some important nontrivial examples) involve further related results, such as the operator product expansion, factorization theorems, and the renormalization group (RG), to go far beyond simple fixed-order perturbation theory. The construction of fully rigorous mathematical treatments for the exact theory is a topic of future research.

Formulation of QFT

A QFT is specified by its Lagrangian density. A simple example is ϕ^4 theory:

$$\mathcal{L} \stackrel{?}{=} \frac{(\partial\phi)^2}{2} - \frac{m^2\phi^2}{2} - \frac{\lambda\phi^4}{4!} \qquad [1]$$

where $\phi(x) = \phi(t, \boldsymbol{x})$ is a single component Hermitian field. The Lagrangian density and the resulting equation of motion, $\partial^2\phi + m^2\phi + (1/6)\lambda\phi^3 = 0$, are local; they involve only products of fields at the same spacetime point. Such locality is characteristic of relativistic theories, where otherwise it is difficult or impossible to preserve causality, but it is also the source of the UV divergences. The question mark over the equality symbol in eqn [1] is a reminder that renormalization of UV divergences will force us to modify the equation.

The Feynman rules for perturbation theory are given by a free propagator $\mathrm{i}/(p^2 - m^2 + \mathrm{i}0)$ and an interaction vertex $-\mathrm{i}\lambda$. Although we will usually work in four spacetime dimensions, it is useful also to consider the theory in a general spacetime dimensionality n, where the coupling has energy dimension $[\lambda] = E^{4-n}$. We use "natural units," that is, with $\hbar = c = 1$. The "i0" in the propagator $\mathrm{i}/(p^2 - m^2 + \mathrm{i}0)$ symbolizes the location of the pole relative to the integration contour; it is often written as $\mathrm{i}\epsilon$.

The primary targets of calculations are the vacuum expectation values of time-ordered products of ϕ; in QFT these are called the Green functions of the theory. From these can be reconstructed the scattering matrix, scattering cross sections, and other measurable quantities.

One-Loop Calculations

Low-order graphs for the connected and amputated four-point Green function are shown in **Figure 1**. Each one-loop graph has the form

$$-\mathrm{i}\lambda^2 I(p^2) \stackrel{?}{=} \frac{\lambda^2}{2}\int \frac{\mathrm{d}^4k}{(2\pi)^4}\,\frac{1}{(k^2 - m^2 + \mathrm{i}0)[(p-k)^2 - m^2 + \mathrm{i}0]} \qquad [2]$$

where p is a combination of external momenta. There is a divergence from where the loop

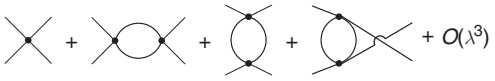


Figure 1 One-loop approximation to connected and amputated four-point function, before renormalization.

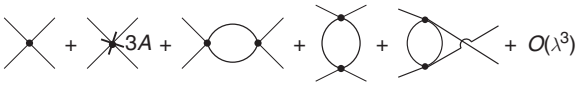


Figure 2 One-loop approximation to renormalized connected and amputated four-point function, with counter-term.

momentum k goes to infinity. We define the degree of divergence, Δ, by counting powers of k at large k, to get $\Delta = 0$. In an n-dimensional spacetime we would have $\Delta = n - 4$. The integral is divergent whenever $\Delta \geq 0$. Comparing the dimensions of the one-loop and tree graphs shows that Δ equals the negative of the energy dimension of the coupling λ. Thus, the dimensionlessness of λ at the physical spacetime dimension is equivalent to the integral being just divergent.

The infinity in the integral implies that the theory *in its naive formulation* is not defined. With the aid of RG methods, it has been shown that the problem is with the complete theory, not just perturbation theory.

The divergence only arises because we use a continuum spacetime. So suppose that we formulate the theory initially on a lattice of spacing a (in space or spacetime). Our loop graph is now

$$-\mathrm{i}\lambda^2 I(p;m,a) = \frac{-\lambda^2}{32\pi^4}\int \mathrm{d}^4k\, S(k,m;a)\, S(p-k,m;a) \qquad [3]$$

where the free propagator $S(k,m;a)$ approaches the usual value $\mathrm{i}/(k^2 - m^2 + \mathrm{i}0)$ when k is much smaller than $1/a$, and it falls off more rapidly for large k. The basic observation that propels the renormalization program is that the divergence as $a \to 0$ is independent of p. This is most easily seen by differentiating once with respect to p, after which the integral is convergent when $a = 0$, because the differentiated integral has degree of divergence –1.

Thus we can cancel the divergence in eqn [2] by replacing the coupling in the first term in Figure 1, by the so-called bare coupling

$$\lambda_0 = \lambda + 3A(a)\lambda^2 + O(\lambda^3) \qquad [4]$$

Here $A(a)$ is chosen so that the renormalized value of our one-loop graph,

$$-\mathrm{i}\lambda^2 I_R(p^2, m^2) = -\mathrm{i}\lambda^2 \lim_{a\to 0}\left[I(p;m,a) + A(a)\right] \qquad [5]$$

exists, at $a = 0$, with $A(a)$ in fact being real valued. The factor 3 multiplying $A(a)$ in eqn [4] is because there are three one-loop graphs, with equal divergent parts. The replacement for the coupling is made in the tree graph in Figure 1, but not yet at the vertices of the other graphs, because at the moment we are only doing a calculation accurate to order λ^2; the appropriate expansion parameter of the theory is the finite renormalized coupling λ, held fixed as $a \to 0$. We call the extra term in eqn [5] a counter-term. The diagrams for the correct renormalized calculation are represented in Figure 2, which has a counter-term graph compared with Figure 1.

In the physics terminology, used here, the cutting-off of the divergence by using a modified theory is called a regularization. This contrasts with the mathematics literature, where "regularized integral" usually means the same as a physicist's "renormalized integral."

There is always freedom to add a finite term to a counter-term. When we discuss the RG, we will see that this corresponds to a reorganization of the perturbation expansion and provides a powerful tool for improving perturbatively based calculations, especially in QCD. Contrary to the impression given in some parts of the literature, it is not necessary that a renormalized mass equal a corresponding physical particle mass, with similar statements for coupling and field renormalization. While such a prescription is common and natural in a simple theory like QED, it is by no means required and certainly may not always be best. If nothing else, the correspondence between fields and stable particles may be poor or nonexistent (as in QCD).

One classic possibility is to subtract the value of the graph at $p = 0$, a prescription associated with Bogoliubov, Parasiuk, and Hepp (BPH), which leads to

$$-\mathrm{i}\lambda^2 I_{\mathrm{R,\,BPH}}(p^2) = \frac{-\mathrm{i}\lambda^2}{32\pi^2}\int_0^1 \mathrm{d}x \ln\left[1 - p^2 x(1-x)/m^2\right] \qquad [6]$$

In obtaining this from [2], we used a standard Feynman parameter formula,

$$\frac{1}{AB} = \int_0^1 \mathrm{d}x \frac{1}{[Ax + B(1-x)]^2} \qquad [7]$$

to combine the propagator denominators, after which the integral over the momentum variable k is elementary. We then obtain the renormalized one-loop (four-point and amputated) Green function

$$-\mathrm{i}\lambda - \mathrm{i}\lambda^2[I_R(s) + I_R(t) + I_R(u)] + O(\lambda^3) \qquad [8]$$

where s, t, and u are the three standard Mandelstam invariants for the Green function. (For a $2 \to 2$

scattering process, or a corresponding off-shell Green function, in which particles of momenta p_1 and p_2 scatter to particles of momenta p_1' and p_2', the Mandelstam variables are defined as $s=(p_1+p_2)^2$, $t=(p_1-p_1')^2$, and $u=(p_1-p_2')^2$.)

In the general case, with a nonzero degree of divergence, the divergent part of an integral is a polynomial in p and m of degree D, where D is the smallest positive integer less than or equal to Δ. In a higher spacetime dimension, this implies that renormalization of the original, momentum-independent, interaction vertex is not sufficient to cancel the divergences. We would need higher derivative terms, and this is evidence that the theory is not renormalizable in higher than 4 spacetime dimensions. Even so, the terms needed would be local, because of the polynomiality in p.

Complete Formulation of Renormalization Program

The full renormalization program motivated by example calculations is:

- the theory is regulated to cut off the divergences;
- the numerical value of each coefficient in $\mathcal{L}$ is allowed to depend on the regulator parameter (e.g., a); and
- these dependences are adjusted so that finite results for Green functions are obtained after removal of the regulator.

In ϕ^4 theory, we therefore replace $\mathcal{L}$ by

$$\mathcal{L}=\frac{Z}{2}(\partial\phi)^2-\frac{Zm_0^2}{2}\phi^2-\frac{Z^2\lambda_0}{4!}\phi^4 \qquad [9]$$

with the bare parameters, Z, m_0 and λ_0, having a regulator dependence such that Green functions of ϕ are finite at $a=0$.

The slightly odd labeling of the coefficients in eqn [9] arises because observables like cross sections are invariant under a redefinition of the field by a factor. In terms of the bare field $\phi_0 \stackrel{\text{def}}{=} \sqrt{Z}\phi$, we have

$$\mathcal{L}=\frac{1}{2}(\partial\phi_0)^2-\frac{m_0^2}{2}\phi_0^2-\frac{\lambda_0}{4!}\phi_0^4 \qquad [10]$$

The unit coefficient of $(1/2)(\partial\phi_0)^2$ implies that ϕ_0 has canonical commutation relations (in the regulated theory). This provides a natural standard for the normalization of the bare mass m_0 and the bare coupling λ_0.

All terms in $\mathcal{L}$ have coefficients with dimension zero or larger. This is commonly characterized by saying that the terms $\mathcal{L}$ "have dimension 4 or less," which refers to the products of field operators and derivatives in each term. A generalization of the power-counting analysis shows that if we start with a theory whose $\mathcal{L}$ only has terms of dimension 4 or less, then no terms of higher dimension are needed as counter-terms, at least not in perturbation theory. This is a very powerful restriction on self-contained QFTs, and was critical in the discovery of the standard model.

Sometimes it is found that the description of some piece of physics appears to need higher-dimension operators, as was the case originally with weak-interaction physics. The lack of renormalizability of such theories indicates that they cannot be complete, and an upper bound on the scale of their applicability can be computed, for example, a few hundred GeV for the four-fermion theory of weak interactions. Eventually, this theory was superseded by the renormalizable Weinberg–Salam theory of weak interactions, now a part of the standard model, to which the four-fermion theory provides a low-energy approximation for charged current weak interactions.

Certain operators of allowed dimensions are missing in eqn [9]: the unit operator, and ϕ and ϕ^3. Symmetry under the transformation $\phi\to-\phi$ implies that Green functions with an odd number of fields vanish, so that no ϕ and ϕ^3 counter-terms are needed. Divergences with the unit operator do appear, but not for ordinary Green functions. In gravitational physics, the coefficient of the unit operator gives renormalization of the cosmological constant.

To implement renormalized perturbation theory, we partition $\mathcal{L}$ (nonuniquely) as

$$\mathcal{L}=\mathcal{L}_{\text{free}}+\mathcal{L}_{\text{basic interaction}}+\mathcal{L}_{\text{counter-term}} \qquad [11]$$

where the free, the basic interaction, and the counter-term Lagrangians are

$$\mathcal{L}_{\text{free}}=\frac{1}{2}(\partial\phi)^2-\frac{m^2}{2}\phi^2 \qquad [12]$$

$$\mathcal{L}_{\text{basic interaction}}=-\frac{\lambda}{4!}\phi^4 \qquad [13]$$

$$\mathcal{L}_{\text{counter-term}}=\frac{Z-1}{2}(\partial\phi)^2-\frac{(Zm_0^2-m^2)}{2}\phi^2-\frac{(Z^2\lambda_0-\lambda)}{4!}\phi^4 \qquad [14]$$

The renormalized coupling and mass, λ and m, are to be fixed and finite when the UV regulator is removed. Both the basic interaction and the counter-terms are treated as interactions. First we compute "basic graphs" for Green functions using only the basic

interaction. The counter-terms are expanded in powers of λ, and then all graphs involving counter-term vertices at the chosen order in λ are added to the calculation. The counter-terms are arranged to cancel all the divergences, so that the UV regulator can be removed, with m and λ held fixed. The counter-terms cancel the parts of the basic Feynman graphs associated with large loop momenta. An algorithmic specification of the otherwise arbitrary finite parts of the counter-terms is called a renormalization prescription or a renormalization scheme. Thus, it gives a definite relation between the renormalized and bare parameters, and hence a definite specification of the partitioning of $\mathcal{L}$ into its three parts.

It has been proved that this procedure works to all orders in λ, with corresponding results for other theories. Even in the absence of fully rigorous nonperturbative proofs, it appears clear that the results extend beyond perturbation theory, at least in asymptotically free theories like QCD: see the discussion on Wilsonian RG (*see* Exact Renormalization Group).

Dimensional Regularization and Minimal Subtraction

The final result for renormalized graphs does not depend on the particular regularization procedure. A particularly convenient procedure, especially in QCD, is dimensional regularization, where divergences are removed by going to a low spacetime dimension n. To make a useful regularization method, n is treated as a continuous variable, $n = 4 - 2\epsilon$.

Great advantages of the method are that it preserves Poincaré invariance and many other symmetries (including the gauge symmetry of QCD), and that Feynman graph calculations are minimally more complicated than for finite graphs at $n = 4$, particularly when all the lines are massless, as in many QCD calculations.

Although there is no such object as a genuine vector space of finite noninteger dimension, it is possible to construct an operation that behaves as if it were an integration over such a space. The operation was proved unique by Wilson, and explicit constructions have been made, so that consistency is assured at the level of all Feynman graphs. Whether a satisfactory definition beyond perturbation theory exists remains to be determined.

It is convenient to arrange that the renormalized coupling is dimensionless in the regulated theory. This is done by changing the normalization of λ with the aid of an extra parameter, the unit of mass μ:

$$\lambda_0 = \mu^{2\epsilon}(\lambda + \text{counter-terms}) \qquad [15]$$

with λ and μ being held fixed when $\epsilon \longrightarrow 0$. (Thus, the basic interaction in eqn [13] is changed to $-\lambda\mu^{2\epsilon}\phi^4/4!$.) Then for the one-loop graph of eqn [2], dimensionally regularized Feynman parameter methods give

$$-\mathrm{i}\lambda^2 I(p; m, \epsilon) = \frac{\mathrm{i}\lambda^2}{32\pi^2}(4\pi)^{\epsilon}\Gamma(\epsilon) \times \int_0^1 \mathrm{d}x \left[\frac{m^2 - p^2x(1-x) - \mathrm{i}0}{\mu^2}\right]^{-\epsilon} \qquad [16]$$

A natural renormalization procedure is to subtract the pole at $\epsilon = 0$, but it is convenient to accompany this with other factors to remove some universally occurring finite terms. So $\overline{\mathrm{MS}}$ renormalization ("modified minimal subtraction") is defined by using the counter-term

$$-\mathrm{i}A(\epsilon)\lambda^2 = -\mathrm{i}\frac{\lambda^2 S_\epsilon}{32\pi^2\epsilon} \qquad [17]$$

where $S_\epsilon \stackrel{\text{def}}{=} (4\pi\,\mathrm{e}^{-\gamma_\mathrm{E}})^{\epsilon}$, with $\gamma_\mathrm{E} = 0.5772\ldots$ being the Euler constant. This gives a renormalized integral (at $\epsilon = 0$)

$$-\frac{\mathrm{i}\lambda^2}{32\pi^2}\int_0^1 \mathrm{d}x \ln\left[\frac{m^2 - p^2x(1-x)}{\mu^2}\right] \qquad [18]$$

which can be evaluated easily. A particularly simple result is obtained at $m = 0$:

$$\frac{\mathrm{i}\lambda^2}{32\pi^2}\left[-\ln\frac{-p^2}{\mu^2} + 2\right] \qquad [19]$$

This formula symptomizes important and very useful algorithmic simplifications in the higher-order massless calculations common in QCD.

The $\overline{\mathrm{MS}}$ scheme amounts to a *de facto* standard for QCD. At higher orders a factor of $S_\epsilon{}^L$ is used in the counter-terms, with L being the number of loops.

Coordinate Space

Quantum fields are written as if they are functions of x, but they are in fact distributions or generalized functions, with quantum-mechanical operator values. This indicates that using products of fields is dangerous and in need of careful definition. The relation with ordinary distribution theory is simplest in the coordinate-space version of Feynman graphs. Indeed in the 1950s, Bogoliubov and Shirkov formulated renormalization as a problem of defining products of the singular numeric-valued distributions in coordinate-space Feynman graphs; theirs was perhaps the best treatment of renormalization in that era.

For example, the coordinate-space version of eqn [5] is

$$-\lambda^2 \lim_{a\to 0} \int \mathrm{d}^4x\, \mathrm{d}^4y\, f(x,y) \times \left[\tfrac{1}{2}\tilde{S}(x-y;m,a)^2 + \mathrm{i}A(a)\delta^{(4)}(x-y)\right] \quad [20]$$

where x and y are the coordinates for the interaction vertices, $f(x,y)$ is the product of external-line free propagators, and $\tilde{S}(x-y;m,a)$ is the coordinate-space free propagator, which at $a=0$ has a singularity

$$\frac{1}{4\pi^2[-(x-y)^2+\mathrm{i}0]} \quad [21]$$

as $(x-y)^2 \to 0$. We see in eqn [20] a version of the Hadamard finite part of a divergent integral, and renormalization theory generalizes this to particular kinds of arbitrarily high-dimension integrals. The physical realization and justification of the use of the finite-part procedure is in terms of renormalization of parameters in the Lagrangian; this also gives the procedure a significance that goes beyond the integrals themselves and involves the full nonperturbative formulation of QFT.

General Counter-Term Formulation

We have written $\mathcal{L}$ as a basic Lagrangian density plus counter-terms, and have seen in an example how to cancel divergences at one-loop order. In this section, we will see how the procedure works to all orders. The central mathematical tool is Bogoliubov's R-operation. Here the counter-terms are expanded as a sum of terms, one for each basic one-particle irreducible (1PI) graph with a non-negative degree of divergence. To each basic graph for a Green function is added a set of counter-term graphs associated with divergences for subgraphs. The central theorem of renormalization is that this procedure does in fact remove all the UV divergences, with the form of the counter-terms being determined by the simple computation of the degree of divergence for 1PI graphs.

To see the essential difficulty to be solved, consider a two-loop graph like the first one in Figure 3. Its divergence is not a polynomial in external momenta, and is therefore not canceled by an allowed counter-term. This is shown by differentiation with respect to external momenta, which does not produce a finite result because of the divergent one-loop subgraph. But for consistency of the theory, the one-loop counter-terms already computed must be themselves put into loop graphs. Among others, this gives the second graph of Figure 3, where the cross denotes that a counter-term contribution is used. The contribution used here is actually 2/3 of the total one-loop counter-term, for reasons of symmetry factors that are not fully evident at first sight. The remainder of the one-loop coupling renormalization cancels a subdivergence in another two-loop graph. It is readily shown that the divergence of the sum of the first two graphs in Figure 3 is momentum independent, and thus can be canceled by a vertex counter-term.

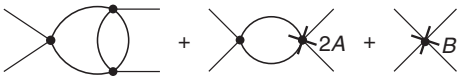


Figure 3 A two-loop graph and its counter-terms. The label B indicates that it is the two-loop overall counter-term for this graph.

This method is fully general, and is formalized in the Bogoliubov R-operation, which gives a recursive specification of the renormalized value $R(G)$ of a graph G:

$$R(G) \stackrel{\text{def}}{=} G + \sum_{\{\gamma_1,\ldots,\gamma_n\}} G|_{\gamma_i \to C(\gamma_i)} \quad [22]$$

The sum is over all sets of nonintersecting 1PI subgraphs of G, and the notation $G|_{\gamma_i \to C(\gamma_i)}$ denotes G with all the subgraphs γ_i replaced by associated counter-terms $C(\gamma_i)$. The counter-term $C(\gamma)$ of a 1PI graph γ has the form

$$C(\gamma) \stackrel{\text{def}}{=} -T(\gamma + \text{counter-terms for subdivergences}) \quad [23]$$

Here T is an operation that extracts the divergent part of its argument and whose precise definition gives the renormalization scheme. For example, in minimal subtraction we define

$$T(\Gamma) = \text{pole part at } \epsilon = 0 \text{ of } \Gamma \quad [24]$$

We formalize the term inside parentheses in eqn [23] as

$$\bar{R}(\gamma) \stackrel{\text{def}}{=} \gamma + \text{counterterms for subdivergences} = \gamma + \sum_{\{\gamma_1,\ldots,\gamma_n\}}{}' G|_{\gamma_i \to C(\gamma_i)} \quad [25]$$

where the prime on the $\sum'$ denotes that we sum over all sets of nonintersecting 1PI subgraphs except for the case that there is a single γ_i equal to the whole graph (i.e., the term with $n=1$ and $\gamma_1=\gamma$ is omitted).

Note that, for the $\overline{\text{MS}}$ scheme, we define the T operation to be applied to a factor of constant dimension obtained by taking the appropriate power of μ^ϵ outside of the pole-part operation. Moreover, it is not a strict pole-part operation; instead each

pole is to be multiplied by $S_\epsilon{}^L$, where L is the number of loops, and S_ϵ is defined after eqn [17].

Equations [22]–[25] give a recursive construction of the renormalization of an arbitrary graph. The recursion starts on one-loop graphs, since they have no subdivergences, that is, $C(\gamma) = -T(\gamma)$ for a one-loop 1PI graph.

Each counter-term $C(\gamma)$ is implemented as a contribution to the counter-term Lagrangian. The Feynman rules ensure that once $C(\gamma)$ has been computed, it appears as a vertex in bigger graphs in such a way as to give exactly the counter-terms for subdivergences used in the R-operation. It has been proved that the R-operation does in fact give finite results for Feynman graphs, and that basic power counting in exactly the same fashion as at one-loop determines the relevant operators.

In early treatments of renormalization, a problem was caused by graphs like Figure 4. This graph has three divergent subgraphs which overlap, rather than being nested. Within the R-operation approach, such cases are no harder to deal with than merely nested divergences.

The recursive specification of R-operation can be converted to a nonrecursive formulation by the forest formula of Zavyalov and Stepanov, later rediscovered by Zimmerman. It is normally the recursive formulation that is suited to all-orders proofs.

Whether these results, proved to all orders of perturbation theory, genuinely extend to the complete theory is not so easy to answer, certainly in a realistic four-dimensional QFT. One illuminating case is of a nonrelativistic quantum mechanics model with a delta-function potential in a two-dimensional space. Renormalization can be applied just as in field theory, but the model can also be treated exactly, and it has been shown that the results agree with perturbation theory.

Perturbation series in relativistic QFTs can at best be expected to be asymptotic, not convergent. So instead of a radius of convergence, we should talk about a region of applicability of a weak-coupling expansion. In a direct calculation of counter-terms, etc., the radius of applicability shrinks to zero as the regulator is removed. However, we can deduce the expansion for a renormalized quantity, whose expansion is expected to have a nonzero range of applicability. We can therefore appeal to the uniqueness of power series expansions to allow the calculation, at intermediate stages, to use bare quantities that are divergent as the regulator is removed.

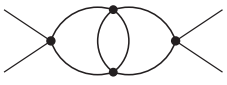


Figure 4 Graph with overlapping divergent subgraphs.

Renormalizability, Non-Renormalizability, and Super-Renormalizability

The basic power-counting method shows that if a theory with conventional fields (at $n = 4$) has only operators of dimension 4 or less in its $\mathcal{L}$, then the necessary counter-term operators are also of dimension 4 or less. So if we start with a Lagrangian with all possible such operators, given the field content, then the theory is renormalizable. This is not the whole story, as we will see in the discussion of gauge theories.

If we start with a Lagrangian containing operators of dimension higher than 4, then renormalization requires operators of ever higher dimension as counter-terms when one goes to higher orders in perturbation theory. Therefore, such a theory is said to be perturbatively non-renormalizable. Some very powerful methods of cancelation or some nonperturbative effects are needed to evade this result.

In the case of dimension-4 interactions, there is only a finite set of operators given the set of basic fields, but divergences occur at arbitrarily high orders in perturbation theory. If, instead, all the operators have at most dimension 3, then only a finite number of graphs need counter-terms. Such theories are called super-renormalizable. The divergent graphs also occur as subgraphs inside bigger graphs, of course. There is only one such theory in a four-dimensional spacetime: ϕ^3 theory, which suffers from an energy density that is unbounded from below, so it is not physical. In lower spacetime dimension, where the requirements on operator dimension are different, there are many more known super-renormalizable theories, some with a very rigorous proof of existence.

All the above characterizations rely primarily on perturbative analysis, so they are subject to being not quite accurate in an exact theory, but they form a guide to the relevant issues.

Renormalization and Symmetries: Gauge Theories

In most physical applications, we are interested in QFTs whose Lagrangian is restricted to obey certain symmetry requirements. Are these symmetries preserved by renormalization? That is, is the Lagrangian with all necessary counter-terms still invariant under the symmetry?

We first discuss nonchiral symmetries; these are symmetries in which the left-handed and right-handed parts of Dirac fields transform identically.

For Poincaré invariance and simple global internal symmetries, it is simplest to use a regulator, like dimensional regularization, which respects the symmetries. Then it is easily shown that the symmetries are preserved under renormalization. This holds even if the internal symmetries are spontaneously broken (as happens with a "wrong-sign mass term," e.g., negative m^2 in eqn [1]).

The case of local gauge symmetries is harder. But their preservation is more important, because gauge theories contain vector fields which, without a gauge symmetry, generally give unphysical features to the theory. For perturbation theory, BRST quantization is usually used, in which, instead of gauge symmetry, there is a BRST supersymmetry. This is manifested at the Green function level by Slavnov–Taylor identities that are more complicated, in general, than the Ward identities for simple global symmetries and for abelian local symmetries.

Dimensional regularization preserves these symmetries and the Slavnov–Taylor identities. Moreover, the R-operation still produces finite results with local counter-terms, but cancelations and relations occur between divergences for different graphs in order to preserve the symmetry. A simple example is QED, which has an abelian U(1) gauge symmetry, and whose gauge-invariant Lagrangian is

$$\mathcal{L} = -\tfrac{1}{4}\left(\partial_\mu A_\nu^{(0)} - \partial_\nu A_\mu^{(0)}\right)^2 + \bar{\psi}_0\left(\mathrm{i}\gamma^\mu\partial_\mu - e_0 A_\mu^{(0)} - m_0\right)\psi_0 \qquad [26]$$

At the level of individual divergent 1PI graphs, we get counter-terms proportional to ${A_\mu}^2$ and to $({A_\mu}^2)^2$, operators not present in the gauge-invariant Lagrangian. The Ward identities and Slavnov–Taylor identities show that these counter-terms cancel when they are summed over all graphs at a given order of renormalized perturbation theory. Moreover, the renormalization of coupling and the gauge field are inverse, so that $e_0 A_\mu^{(0)}$ equals the corresponding object with renormalized quantities, $\mu^\epsilon e A_\mu$. Naturally, sums of contributions to a counter-term in $\mathcal{L}$ can only be quantified with use of a regulator.

In nonabelian theories, the gauge-invariance properties are not just the absence of certain terms in $\mathcal{L}$ but quantitative relations between the coefficients of terms with different numbers of fields. Even so, the argument with Slavnov–Taylor identities generalizes appropriately and proves renormalizability of QCD, for example. But note that the relation concerning the product of the coupling and the gauge field does not generally hold; the form of the gauge transformation is itself renormalized, in a certain sense.

Anomalies

Chiral symmetries, as in the weak-interaction part of the gauge symmetry of the standard model, are much harder to deal with. Chiral symmetries are ones for which the left-handed and right-handed components of Dirac field transform independently under different components of the symmetry group, local or global as the case may be. Occasionally, some or other of the left-handed or right-handed components may not even be present.

In general, chiral symmetries are not preserved by regularization, at least not without some other pathology. At best one can adjust the finite parts of counter-terms such that in the limit of the removal of the regulator, the Ward or Slavnov–Taylor identities hold. But in general, this cannot be done consistently, and the theory is said to suffer from an anomaly. In the case of chiral gauge theories, the presence of an anomaly prevents the (candidate) theory from being valid. A dramatic and nontrivial result (Adler–Bardeen theorem and some nontrivial generalizations) is that if chiral anomalies cancel at the one-loop level, then they cancel at all orders.

Similar results, but more difficult ones, hold for supersymmetries.

The anomaly cancelation conditions in the standard model lead to constraints that relate the lepton content to the quark content in each generation. For example, given the existence of the b quark, and the τ and ν_τ leptons (of masses around 4.5 GeV, 1.8 GeV, and zero respectively), it was strongly predicted on the grounds of anomaly cancelation that there must be a t quark partner of the b to complete the third generation of quark doublets. This prediction was much later vindicated by the discovery of the much heavier top quark with $m_t \simeq 175$ GeV.

Renormalization Schemes

A precise definition of the counter-terms entails a specification of the renormalization prescription (or scheme), so that the finite parts of the counter-terms are determined. This apparently induces extra arbitrariness in the results. However, in the ϕ^4 Lagrangian (for example), there are really only two independent parameters. (A scaling of the field does not affect any observables, so we do not count Z as a parameter here.) Thus, at fixed regulator parameter a or ϵ, renormalization actually just gives a reparametrization of a two-parameter collection of theories. A renormalization prescription gives the

change of variables between bare and renormalized parameters, a rather singular transformation when the regulator is removed. If we have two different prescriptions, we can deduce a transformation between the renormalized parameters in the two schemes. The renormalized mass and coupling m_1 and λ_1 in one scheme can be obtained as functions of their values m_2 and λ_2 in the other scheme, with the bare parameters, and hence the physics, being the same in both schemes. Since these are renormalized parameters, the removal of the regulator leaves the transformation well behaved.

Generalization to all renormalizable theories is immediate.

Renormalization Group and Applications and Generalizations

One part of the choice of renormalization scheme is that of a scale parameter such as the unit of mass μ of the $\overline{\text{MS}}$ scheme. The physical predictions of the theory are invariant if a change of μ is accompanied by a suitable change of the renormalized parameters, now considered as μ-dependent parameters $\lambda(\mu)$ and $m(\mu)$. These are called the effective, or running, coupling and mass. The transformation of the parametrization of the theory is called an RG transformation.

The bare coupling and mass λ_0 and m_0 are RG invariant, and this can be used to obtain equations for the RG evolution of the effective parameters from the perturbatively computed counter-terms. For example, in ϕ^4 theory, we have (in the renormalized theory after removal of the regulator)

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\ln\mu^2} = \beta(\lambda) \qquad [27]$$

with $\beta(\lambda) = 3\lambda^2/(16\pi^2) + O(\lambda^3)$. As exemplified in eqns. [18] and [19], Feynman diagrams depend logarithmically on μ. By choosing μ to be comparable to the physical external momentum scale, we remove possible large logarithms in this and higher orders. Thus, provided that the effective coupling at this scale is weak, we get an effective perturbation expansion.

This is a basic technique for exploiting perturbation theory in QCD, for the strong interactions, where the interactions are not automatically weak. In this theory the RG β function is negative so that the coupling decreases to zero as $\mu \to \infty$; this is the asymptotic freedom of QCD.

A closely related method is that associated with the Callan–Symanzik equation, which is a formulation of a Ward identity for anomalously broken scale invariance. However, RG methods are the actually used ones, normally, even if sometimes an RG equation is incorrectly labeled as a Callan–Symanzik equation.

The elementary use of the RG is not sufficient for most interesting processes, which involve a set of widely different scales. Then more powerful theorems come into play. Typical are the factorization theorems of QCD (*see* Quantum Chromodynamics). These express differential cross sections for certain important reactions as a product of quantities that involve a single scale:

$$\mathrm{d}\sigma = C(Q, \mu, \lambda(\mu)) \otimes f(m, \mu, \lambda(\mu)) + \text{small correction} \qquad [28]$$

The product is typically a matrix or a convolution product. The factors obey nontrivial RG equations, and these enable different values of μ to be used in the different factors. Predictions arise because some factors and the kernels of the RG equation are perturbatively calculable, with a weak effective coupling. Other factors, such as f in eqn [28], are not perturbative. These are quantities with names like "parton distribution functions," and they are universal between many different processes. Thus, the nonperturbative functions can be measured in a limited set of reactions and used to predict cross sections for many other reactions with the aid of calculations of the perturbative factors.

Ultimately, this whole area depends on physical phenomena associated with renormalization.

Concluding Remarks

The actual ability to remove the divergences in certain QFTs to produce consistent, finite, and nontrivial theories is a quite dramatic result. Moreover, associated with the integrals that give the divergences is behavior of the kind that is analyzed with RG methods and generalizations. So the properties of QFTs associated with renormalization get tightly coupled to many interesting consequences of the theories, most notably in QCD.

QFTs are actually very abstruse and difficult theories; only certain aspects currently lend themselves to practical calculations. So the reader should not assume that all aspects of their rigorous mathematical treatment are perfect. Experience, both within the theories and in their comparison with experiment, indicates, nevertheless, that we have a good approximation to the truth.

When one examines the mathematics associated with the R-operation and its generalizations with factorization theorems, there are clearly present some interesting mathematical structures that are not yet formulated in their most general terms. Some

indications of this can be seen in the work by Connes and Kreimer (*see* Hopf Algebra Structure of Renormalizable Quantum Field Theory), where it is seen that renormalization is associated with a Hopf algebra structure for Feynman graphs.

With such a deep subject, it is not surprising that it lends itself to other approaches, notably the Connes–Kreimer one and the Wilsonian one (*see* Exact Renormalization Group). Readers new to the subject should not be surprised if it is difficult to get a fully unified view of these different approaches.

Notes on Bibliography

Reliable textbooks on quantum field theory are Sterman (1993) and Weinberg (1995). A clear account of the foundations of perturbative QCD methods is given by Sterman (1996).

A pedagogical account of renormalization and related subjects may be found in Collins (1984). The best account of renormalization theory before the 1970s is given by Bogoliubov and Shirkov (1959); the viewpoint is very modern, including a coordinate-space distribution-theoretic view. A full account of the Wilsonian method as applied to renormalization is given by Polchinski (1984).

Manuel and Tarrach (1994) give an excellent account of renormalization for a theory with a non-relativistic delta-function potential in 2 space dimensions, which provides a fully tractable model.

Tkachov (1994) reviews a systematic application of distribution theoretic methods to asymptotic problems in QFT. Finally, Weinzierl (1999) provides a construction of dimensional regularization with the aid of K-theory using an underlying vector space of the physical integer dimension. Other constructions, referred to in this paper, follow Wilson and use an infinite-dimensional underlying space.

Acknowledgments

The author would like to thank Professor K Goeke for hospitality and support at the Ruhr-Universität-Bochum. This work was also supported by the US DOE.

See also: Anomalies; BRST Quantization; Effective Field Theories; Electroweak Theory; Euclidean Field Theory; Exact Renormalization Group; High T_c Superconductor Theory; Holomorphic Dynamics; Hopf Algebra Structure of Renormalizable Quantum Field Theory; Lattice Gauge Theory; Operator Product Expansion in Quantum Field Theory; Perturbation Theory and its Techniques; Perturbative Renormalization Theory and BRST; Quantum Chromodynamics; Quantum Field Theory: A Brief Introduction; Singularities of the Ricci Flow; Standard Model of Particle Physics; Supergravity.

Further Reading

Bogoliubov NN and Shirkov DV (1959) *Introduction to the Theory of Quantized Fields*. New York: Wiley-Interscience.

Collins JC (1984) *Renormalization*. Cambridge: Cambridge University Press.

Kraus E (1998) Renormalization of the electroweak standard model to all orders. *Annals of Physics* 262: 155–259.

Manuel C and Tarrach R (1994) Perturbative renormalization in quantum mechanics. *Physics Letters B* 328: 113–118.

Polchinski J (1984) Renormalization and effective Lagrangians. *Nuclear Physics B* 231: 269–295.

Sterman G (1993) *An Introduction to Quantum Field Theory*. Cambridge: Cambridge University Press.

Sterman G (1996) Partons, factorization and resummation. In: Soper DE (ed.) *QCD and Beyond*, pp. 327–406. (hep-ph/9606312). Singapore: World Scientific.

Tkachov FV (1994) Theory of asymptotic operation. A summary of basic principles. *Soviet Journal of Particles and Nuclei* 25: 649.

Weinberg S (1995) *The Quantum Theory of Fields, Vol. I. Foundations*. Cambridge: Cambridge University Press.

Weinzierl S (1999) Equivariant dimensional regularization, hep-ph/9903380.

Zavyalov OI (1990) *Renormalized Quantum Field Theory*. Dordrecht: Kluwer.

Renormalization: Statistical Mechanics and Condensed Matter

M Salmhofer, Universität Leipzig, Leipzig, Germany

Renormalization Group and Condensed Matter

Statistical mechanical systems at critical points exhibit scaling laws of order parameters, susceptibilities, and other observables. The exponents of these laws are universal, that is, independent of most details of the system. For example, the liquid–gas transition for real gases has the same exponents as the magnetization transition in the three-dimensional Ising model.

The renormalization group (RG) was developed by Kadanoff, Wilson, and Wegner, to understand these critical phenomena (Domb and Green 1976). The central idea is that the system becomes scale invariant at the critical point, which makes it natural to average over degrees of freedom on increasing length scales successively in the

calculation of the partition function. This leads to a map between effective interactions associated to different length scales. Thus, the focus shifts from the analysis of a single interaction to that of a flow on a space of interactions. This space is in general much larger than the original formulation of the model would suggest: the description of long-distance or low-energy properties may be in terms of variables that were not even present in the original formulation of the system. Phenomenologically, this corresponds to the emergence of collective degrees of freedom.

Condensed matter theory is itself already an effective theory, and its "microscopic" formulation gets inputs from the underlying theories, which determine in particular the statistics of the particles and their interactions at the scale of atomic energies. At much lower-energy scales, which are relevant for low-temperature phenomena in condensed matter, collective excitations of different, sometimes exotic, statistics may emerge, but the starting point is given naturally in terms of fermionic and bosonic particles. For this reason, the discussion given below will be split in these two cases.

A major difference between high-energy and condensed matter systems is that the latter have a well-defined Hamiltonian which can be used to define the finite-volume ensembles of quantum statistical mechanics and which determines the time evolution, as well as various analyticity properties.

The relevant spatial dimensions in condensed matter are $d \leq 3$, but some results in higher dimensions relevant for the development of the method will also be discussed below. The cases $d = 1$ and $d = 2$ have always been of mathematical interest but in recent years have become important for the theory of new materials.

Some interesting topics cannot be covered here due to space restrictions, notably the application of renormalization methods to membrane theory (see Wiese (2001)) and renormalization methods for operators (see Bach *et al.* (1998)).

The Renormalization Group

In this section we briefly describe the setup of two important versions of the RG, namely the block spin RG and the RG based on scale decompositions of singular covariances.

Block spin RG

Let Λ be a finite lattice, for example, a finite subset of $\mathbb{Z}^d$. For the following, it is convenient to take Λ to be a cube of side-length L^K for $L > 1$ and some large K. Let T be a set and $\Phi_\Lambda = \{\phi : \Lambda \to T\}$ be the set of spin configurations. Common examples for the target space T are $T = \{-1, 1\}$ for the Ising model, $T = S^{N-1}$ for the O(N) model, and $T = \mathbb{R}^n$ for unbounded spins. Let $S_\Lambda : \Phi_\Lambda \to \mathbb{R}, \phi \mapsto S_\Lambda(\phi)$ be an interaction and

$$Z(\Lambda, S_\Lambda) = \int \prod_{x \in \Lambda} \mathrm{d}\phi(x) \mathrm{e}^{-S_\Lambda(\phi)} \qquad [1]$$

In the unbounded case, S_Λ is assumed to grow sufficiently fast for $|\phi| \to \infty$, so that Z exists; for the case of a finite set T, the integral is replaced by a sum. Denote the corresponding Boltzmann factor by $\rho(\Lambda, S_\Lambda)$,

$$\rho(\Lambda, S_\Lambda)(\phi) = \frac{1}{Z(\Lambda, S_\Lambda)} \mathrm{e}^{-S_\Lambda(\phi)} \qquad [2]$$

The block spin transformation consists of an integration step and a rescaling step. Divide the lattice into cubic blocks of side-length L and define a new lattice Λ' by associating one lattice site of the new lattice to each L-block of the old lattice. For any $\phi' : \Lambda' \to T$, let

$$\rho'(\phi') = \int \prod_{x \in \Lambda} \mathrm{d}\phi(x) P(\phi', \phi) \mathrm{e}^{-S_\Lambda(\phi)} \qquad [3]$$

where $P(\phi', \phi) \geq 0$ and $\int \prod_{x' \in \Lambda'} \mathrm{d}\phi'(x') P(\phi', \phi) = 1$ for all ϕ, so that ρ' remains a probability distribution. Since ρ' is positive, one defines

$$S'_{\Lambda'}(\phi') = -\log \rho'(\phi') \qquad [4]$$

By construction, the partition function is invariant: $Z(\Lambda', S'_{\Lambda'}) = Z(\Lambda, S_\Lambda)$. The new lattice Λ' has spacing L; now rescale to make it a unit lattice. This completes the RG step in finite volume.

In an algorithmic sense, the "blocking rule" $P(\phi', \phi)$ can be viewed as a transition probability of a configuration ϕ to a configuration ϕ'. P may be deterministic, that is, simply fix ϕ' as a function of ϕ. From the intuition of averaging over local fluctuations, ϕ' is often taken to be some average of $\phi(x)$ at x in a block around x', hence the name.

Obviously, the thus defined RG transformation often cannot be iterated arbitrarily, since in every application, the number of points of the lattice shrinks by a factor L^d, so that after K iterations, a lattice with only a single point is left over. It is necessary to take the infinite-volume limit $L \to \infty$ to obtain a map that operates from a space to itself. However, [4] can become problematic in that limit: Gibbs measures ρ can map to measures ρ' whose large-deviation properties differ from those of Gibbs measures. The discussion of this problem and its solution is reviewed in Bricmont and Kupiainen (2001). The problem can be

solved in different ways, relaxing conditions on Gibbs measures or, in the Ising model, changing the description from the spins to the contours. The crucial point is that the difficulties arise only because [4] is applied globally, that is, to every ϕ'. The set of bad ϕ' has very small probability.

Block spin methods have been used in mathematical construction of quantum field theories, for example, in the work of Gawedzki and Kupiainen (1985) and Balaban (1988) (see the subsection "Field theory and statistical mechanics"). The above-mentioned problem was avoided there by not taking a logarithm in the so-called large-field region (which has very small probability).

Scale Decomposition RG

The generating functionals of quantum field theory and quantum statistical mechanics can be cast into the form

$$Z(C, V, \phi) = \int \mathrm{d}\mu_C(\phi')\, \mathrm{e}^{-V(\phi'+\phi)} \qquad [5]$$

Here $\mathrm{d}\mu_C$ denotes the Gaussian measure with covariance C, and V is the two-body interaction between the particles. The field variables are real or complex for bosons and Grassmann-valued for fermions. Differentiating $\log Z$ with respect to the external field ϕ generates the connected amputated correlation functions. The covariance determines the free propagation of particles; the interaction their collisions.

In most cases, such functional integrals are *a priori* ill-defined, even if V is small (and bounded from below) because the covariance C is singular. That is, the integral kernel $C(X, X')$ of the operator C either diverges as $|x - x'| \to 0$ (ultraviolet (UV) problem) or $C(X, X')$ has a slow decay as $|x - x'| \to \infty$ (infrared (IR) problem). In our notational convention, X may, in addition to the configuration variable x, also contain discrete indices of the fields, such as a spin or color index. The dependence of C on x and x' is assumed to be of the form $x - x'$. A typical example is the massless Gaussian field in d dimensions, where C is the inverse Fourier transform of $\hat{C}(k) = 1/k^2$, $k \in \mathbb{R}^d$, which has both a UV and an IR problem, or its lattice analog,

$$\hat{D}(k) = \left(\frac{2}{a^2} \sum_{i=1}^{d} (1 - \cos(ak_i)) \right)^{-1}$$

with a the lattice constant, which has only an IR problem. A typical interaction is of the type

$$V(\phi) = \int \mathrm{d}X\, \mathrm{d}Y \bar{\phi}(X) \phi(X) v(X, Y) \bar{\phi}(Y) \phi(Y) \qquad [6]$$

Again, we assume that the potential v depends on x and y only via $x - y$, so that translation invariance holds. In both UV and IR cases, naive perturbation theory fails even as a formal power series. That is, writing $V = \lambda V_0$, with a coupling constant λ which is treated as a formal expansion parameter, the singularity of C leads to termwise divergences in the series. The theory is called perturbatively renormalizable if all divergences can be removed by posing counterterms of certain types, which are fixed by physically sensible renormalization conditions. Identifying the UV renormalizable theories was a breakthrough in high-energy physics. The IR renormalization problem is different, and in some respects harder, because there is almost no freedom to put counter-terms: the microscopic model is given from the start. This will be discussed in more detail below for an example.

A much more ambitious, and largely open, project is to do this renormalization nonperturbatively, that is, to treat λ as a real (typically, small) parameter. Some results will be discussed below.

The RG is set up by a scale decomposition $C = \sum_j C_j$. In the example of the massless Gaussian field, one would take each $\hat{C}_j$ to be a C^∞ function supported in the region $\{k \in \mathbb{R}^d : M^j \le k^2 \le M^{j+1}\}$, where $M > 1$ is a fixed constant, and the summation over j runs over $\mathbb{Z}$.

The scale decomposition of C leads to a representation of [5] by an iteration of Gaussian convolution integrals with covariances C_j, hence a sequence of effective interactions V_j, defined recursively by

$$\mathrm{e}^{-V_j(\phi)} = \int \mathrm{d}\mu_{C_{j+1}}(\phi')\, \mathrm{e}^{-V_{j+1}(\phi'+\phi)}, \quad V_0 = V \qquad [7]$$

For a singular covariance, the scale decomposition is an infinite sum. A formal object like [5] is now regularized by starting with a finite sum, that is, imposing a UV and IR cutoff, which is mathematically well defined, and then taking limits of the thus defined objects. Again, in condensed matter applications, imposing an IR cutoff is an operation that needs to be justified, for example, by showing that taking the limit as the cutoff is removed commutes with the infinite-volume limit.

Note that the RG map, which is the iteration $V_j \mapsto V_{j-1}$, goes to lower and lower j, corresponding to longer and longer length scales. The convention that the iteration starts at some fixed j, for example, $j = 0$, is appropriate for IR problems. In UV problems, the iteration would start at some large J_{UV}, which defines a UV cutoff and is taken to infinity, to remove the cutoff, at the end.

A variant using a continuous scale decomposition, $C = \int \mathrm{d}s \dot{C}_s$, originally due to Wegner and Houghton, became very popular after Polchinski (1984) used it

to give a short argument for perturbative renormalizability. Polchinski's equation, the analog of the recursion [7], reads

$$\frac{\partial V}{\partial s} = -\frac{1}{2} e^{V} \Delta_{\dot{C}_s} e^{-V} = \frac{1}{2} \Delta_{\dot{C}_s} V - \frac{1}{2} \left(\frac{\delta V}{\delta \phi}, \dot{C}_s \frac{\delta V}{\delta \phi} \right) \quad [8]$$

Here

$$\Delta_C = \left(\frac{\delta}{\delta \phi}, C \frac{\delta}{\delta \phi} \right)$$

denotes the Laplacian in field space associated to the covariance C. Polchinski's argument has been developed into a mathematical tool that applies to many models. For an introduction to perturbative renormalization using this method, see Salmhofer (1998). Equations of the type [8] have also been very useful beyond perturbation theory: much work has been done based on the beautiful representation of Mayer expansions found in Brydges and Kennedy (1987) using RG equations.

Mathematical Structure and Difficulties

The RG flow is thus, depending on the implementation, either a sequence or a continuous flow of interactions. Setting up this flow in mathematical terms is not easy and indeed part of the mathematical RG analysis is to find a suitable space of interactions that is left invariant by the successive convolutions, and then to control the RG iteration. A serious problem is the proliferation of interactions: already a single application of the RG transformation [7] maps a simple interaction, such as [6], to a nonlocal functional of the fields,

$$V_j(\phi) = \sum_{m \geq 0} \int dX_1 \cdots dX_m \times v_m^{(j)}(X_1, \ldots, X_m)\, \phi(X_1) \cdots \phi(X_m) \quad [9]$$

Already for perturbative renormalization, one needs to extract local terms, calculate their flow more explicitly, and control the power counting of the remainder. The convergence of the series is not an issue in formal perturbation theory because in every finite order r in λ, the sum over m is finite.

For nonperturbative renormalization, however, the problem is much more serious. For bosonic systems, the expansion in powers of the fields in [9] is divergent, and one needs a split into small-field and large-field regions and cluster expansions to obtain a well-defined sequence of effective actions (Gawedzki and Kupiainen 1985, Feldman *et al.* 1987, Rivassean 1993). That is, the local parts are extracted and treated explicitly only in the small-field region, and this is combined with estimates on the rareness of large-field regions using cluster expansions. For fermions, the expansion in powers of the fields can be proved to converge for regular, summable covariances, which leads to substantial technical simplifications.

The spatial proliferation of interactions is absent only in certain one-dimensional and in specially constructed higher-dimensional models, the so-called "hierarchical models." In these models, the search for an RG fixed point is still a nonlinear fixed-point problem, whose treatment leads to interesting mathematical results.

This article will be restricted to the mathematical use of the RG both in perturbative and nonperturbative quantum field theory of condensed matter systems. Many nonrigorous but very interesting applications have also come out of this method, showing that it also works well in practice, but they will not be reviewed here. Before discussing condensed matter systems, the pioneering works done on the mathematical RG, which were largely motivated by high-energy physics, will be reviewed briefly, as they laid the foundation of much of the technique used later in the condensed matter case.

Field Theory and Statistical Mechanics

Because of the close connection between quantum field theory and statistical mechanics given by formulas of the Feynman–Kac type, a significant amount of work on the mathematical RG focused on models of classical statistical mechanics in connection with field theories and gauge theories. Here we mention some of the pioneering results in that field.

The scale decomposition method was developed in a mathematical form and applied to perturbative UV renormalization of scalar field theories, as well as nonperturbative analysis of some models, by Gallavotti and Nicolò (Gallavotti 1985).

Infrared ϕ^4 theory in four dimensions was constructed using block spin methods (Gawedzki and Kupiainen 1985) and scale decomposition RG (Feldman *et al.* 1987). An essential feature of the ϕ^4_4 model is its IR asymptotic freedom, meaning that the local part of the effective quartic interaction tends to zero in the IR limit.

Block spin methods were used by Balaban (1988) to construct gauge theories in three and four dimensions. For gauge theories, the block spin RG has the major advantage that it allows to define a gauge-invariant RG flow. The scale decomposition violates gauge invariance, which creates substantial technical problems (Rivasseau 1993).

Condensed Matter: Fermions

Starting with the seminal work of Feldman and Trubowitz (1990, 1991) and Benfatto and Gallavotti (1995), this field has become one of the most successful applications of the mathematical RG. We use this example to discuss the scale decomposition method in a bit more detail.

We shall mainly focus on models in $d \geq 2$ dimensions (the case $d = 1$ is described in detail in Benfatto and Gallavotti (1995)). The system is put into a finite (very large) box Λ of side-length L. For simplicity we take periodic boundary conditions. The Hilbert space for spin-1/2 electrons is the fermionic Fock space $\mathcal{F} = \bigoplus_{n\geq 0} \bigwedge^n L^2(\Lambda, \mathbb{C}^2)$. The grand canonical ensemble in finite volume is given by the density operator $\rho = Z^{-1} e^{-\beta(H-\mu N)}$, with the Hamiltonian H and the number operator N, in the usual second quantized form. The parameter $\beta = T^{-1}$ is the inverse temperature and the chemical potential μ is an auxiliary parameter used to fix the average particle number.

The grand canonical trace defining the ensemble can be rewritten in functional-integral form. It takes the form [5], but now $d\mu_C$ stands for a Grassmann Gaussian "measure," which is really only a linear functional (for definitions, see, e.g., Salmhofer (1998, chapter 4 and appendix B)). A two-body interaction corresponds to a quartic interaction polynomial V, as in [6]. The covariance is (in the infinite-volume limit $L \to \infty$)

$$C(\tau, x) = \frac{1}{\beta} \sum_{\omega \in \mathbb{M}_F} \int \frac{dk}{(2\pi)^d} e^{i(k\cdot x - \omega\tau)} \hat{C}(\omega, k)$$
$$\hat{C}(\omega, k) = \frac{1}{i\omega - e(k)} \qquad [10]$$

where $\tau \in (0, \beta]$ is a Euclidian time variable and k is the spatial momentum. The summation over ω runs over the set of fermionic Matsubara frequencies $\mathbb{M}_F = \pi T(2\mathbb{Z} + 1)$. The function $e(k) = \varepsilon(k) - \mu$, where $\varepsilon(k)$ is the band function given by the single-particle term in the Hamiltonian. For a lattice system, $k \in \mathcal{B}_d$, the momentum space torus (e.g., for the lattice $\mathbb{Z}^d$, $\mathcal{B}_d = \mathbb{R}^d/2\pi\mathbb{Z}^d$); for a continuous system, $k \in \mathbb{R}^d$, hence there is a spatial UV problem. Electrons in a crystal have a natural spatial UV cutoff (see Salmhofer (1998, chapter 4) for a discussion) so we assume in the following that there is either a UV cutoff or that the system is on a lattice. A nonperturbative definition of the functional integral involves a limit from discrete times (by the Trotter product formula); see, for example, Salmhofer (1998) or Feldman *et al.* (2003, 2004).

Perturbative Renormalization

Renormalization of the Fermi surface at zero temperature In the limit $T \to 0$, the Matsubara frequency ω becomes a real variable, hence the propagator has a singularity at $\omega = 0$ and $k \in S$, where $S = \{k : e(k) = 0\}$, a codimension-1 subset of $\mathcal{B}_d$, is the Fermi surface. The existence of a Fermi surface which does not degenerate to a point is a characteristic feature of systems showing metallic behavior.

The singularity implies that $\hat{C} \notin L^p(\mathbb{R} \times \mathcal{B}_d)$ for any $p \geq 2$. Because terms of the type

$$\int d\omega \int dk F(\omega, k) \hat{C}(\omega, k) \times \prod_{i=1}^{p-1} \Big(T_i(\omega, k) \hat{C}(\omega, k) \Big) \qquad [11]$$

appear for all $p \geq 1$ in the formal perturbation expansion, with functions T_i and F that do not vanish on the singularity set of C, the perturbation expansion for observables is termwise divergent. The deeper reason for these problems is that the interaction shifts the Fermi surface so that the true propagator has a singularity of the form $G(\omega, k) = (i\omega - e(k) - \sigma(\omega, k))^{-1}$. If the self-energy σ is a sufficiently regular function, G has the same integrability properties as C, but the singularity of G is on the set $\tilde{S} = \{k : e(k) + \sigma(0, k) = 0\}$ (the singularity in ω remains at $\omega = 0$).

Let $1 = \sum_{j\leq 0} \chi_j(\omega, k)$ be a C^∞ partition of unity such that

$$\text{for } j < 0 \quad \text{supp } \chi_j \subset \{(\omega, k) : \epsilon_0 M^{j-2} \leq |i\omega - e(k)| \leq \epsilon_0 M^j\} \qquad [12]$$

where $M > 1$ and ϵ_0 is a fixed constant (an energy scale determined by the global properties of the function e; see Salmhofer (1998, chapter 4)). The corresponding covariances $\hat{C}_j = \hat{C}_{\chi_j}$ have the properties that for $j < 0$, $\|\hat{C}_j\|_1 \leq \text{const.} M^j$ and $\|\hat{C}_j\|_\infty \leq \text{const.} M^{-j}$. Using these bounds and expanding $\nu_m^{(j)} = \sum_{r\geq 1} \nu_{m,r}^{(j)} \lambda^r$, one can derive estimates for the coefficient functions $\nu_{m,r}^{(j)}$.

Of course, the scale decomposition by itself does not solve the problem of the moving singularity. It only allows us to pinpoint the problematic terms in the expansion. To construct the self-energy σ, as well as all higher Green functions, a two-step method is used (Feldman and Trubowitz 1990, 1991, Feldman *et al.* 1996, 2000). First, a counterterm function K which modifies e is introduced, so that all two-point insertions T_i get subtracted on the Fermi surface, hence replaced by $\tilde{T}_i(\omega, k) = T_i(\omega, k) - T_i(0, k')$, with k' obtained from k by a

projection to the Fermi surface (Feldman and Trubouitz 1990, 1991). Consequently, the $\tilde{T}_i$ vanish linearly on the Fermi surface, so that the integral over k in [11] converges. The effect of the counter-term function K can be described less technically: it fixes the Fermi surface to be S, the zero set of e. Thus, K forces S to be the Fermi surface of the interacting system. To achieve this, K must be chosen a function of e, k, and λV. In contrast to the situation for covariances with point singularities, the function K will, for a nontrivial Fermi surface, be very different from the original e. It can, however, be constructed to all orders in perturbation theory for a large class of Fermi surfaces. More precisely, one can prove: if $e \in C^2(\mathcal{B}_d, \mathbb{R})$, $\hat{v} \in C^2(\mathcal{B}_d, \mathbb{R})$, and the Fermi surface S contains no points k with $\nabla e(k) = 0$ and no flat sides, then $K = \sum_r \lambda^r K_r$ exists as a formal power series in λ and the map $e \mapsto e + K$ is locally injective on this set of e's (Feldman *et al.* 1996, 2000). With this counter-term, the order-r m-point functions on scale j satisfy the bounds

$$\left\| \hat{v}^{(j)}_{m,r} \right\|_\infty \leq w_{m,r} M^{(4-m)j/2} |j|^r$$

and

$$\left\| \hat{v}^{(j)}_{m,r} \right\|_1 \leq \tilde{w}_{m,r} \qquad [13]$$

with constants $w_{m,r}$ and $\tilde{w}_{m,r}$. Here $\hat{v}^{(j)}_{m,r}$ is the Fourier transform of $v^{(j)}_{m,r}$ (see [9], with the momentum conservation delta function from translation invariance removed.

Equation [13] implies that in the RG sense, the two-point function is relevant, the four-point function is marginal, and all higher m-point functions are irrelevant.

In one dimension, the Fermi "surface" reduces to two points which are related by a symmetry, so the counter-term function K is just a constant, that is, an adjustment of the chemical potential μ, which is justified because μ is only an auxiliary parameter used to fix the average value of the particle number. The counter-term function is a constant also in higher dimensions in the special case $e(k) = k^2 - \mu$: there, rotational symmetry implies that K can be chosen independent of k (if v is also rotationally symmetric). However, in the generic case of non-spherical Fermi surfaces, K depends nontrivially on k, and an inversion problem arises: adding the counter-term changes the model. To obtain the Green functions of a model with a given dispersion relation and interaction (E, V), one needs to show that given E in a suitable set, the equation

$$e(k) + K(\lambda, e, V)(k) = E(k) \qquad [14]$$

has a unique solution. If this is done, the procedure for renormalization is as follows. For a model given by dispersion relation and interaction (E, V), solve [14], then add and subtract e in the kinetic term. This automatically puts $K = E - e$ as a counter-term, and the expansion is now set up automatically with the right counter-term. The function K describes the shift from the Fermi surface of the free system (the zero set of E) to that of the interacting system (the zero set of e). Proving that K is sufficiently regular and solving [14] is nontrivial. Uniqueness of the solution follows from the above stated properties of K as a function of e. Existence was shown for a class of Fermi surfaces with strictly positive curvature in Feldman *et al.* (1996, 2000), to every order in perturbation theory. This implies a bijective relation between the Fermi surfaces of the free and the interacting model.

Positive temperature and the zero-limit temperature

One advantage of the functional-integral approach is that the setup at positive temperatures is identical to that at zero temperature, save for the discreteness of the set $\mathbb{M}_\mathrm{F}$ at $T > 0$. Because $0 \notin \mathbb{M}_\mathrm{F}$, the temperature effectively provides an IR cutoff, so that all term-by-term divergences are regularized in a natural way. However, renormalization is still necessary because the temperature is a physical parameter and unrenormalized expansions give disastrous bounds for the behavior of observables as functions of the temperature. Renormalization carries over essentially unchanged (the counter-term function is constructed slightly differently).

Because $|\omega| \geq \pi/\beta$ for all $\omega \in \mathbb{M}_\mathrm{F}$, [12] implies $\operatorname{supp} \chi_j = \emptyset$ for $j < -J_\beta$, where

$$J_\beta = \log_M \frac{\beta \epsilon_0}{\pi} \qquad [15]$$

Thus, the scale decomposition is now a finite sum over $0 \geq j \geq -J_\beta$. This restriction is inessential for the problem of renormalizing the Fermi surface, but it puts a cutoff on the marginal growth of the four-point function: [15] and [13] imply that

$$\| \hat{v}^{(j)}_{m,r} \|_1 \leq \tilde{w}_{m,r} \left(\log \frac{\beta \epsilon_0}{\pi} \right)^r \qquad [16]$$

If one can show that $\tilde{w}_{m,r} \leq AB^r$ with constants A and B, this implies that perturbation theory converges for $|\lambda| \log(\beta \epsilon_0/\pi) < B^{-1}$. Such a bound has been shown using constructive methods (Disertori and Rivasseau 2000, Feldman *et al.* 2003, 2004) (see below). The logarithm of β is due to the Cooper instability (see Feldman and Trubowitz (1990, 1991) and Salmhofer (1998, section 4.5)).

The application of renormalization at positive temperature also led to the solution of a longstanding puzzle in solid-state physics, namely the (seeming) discontinuity of the results of perturbation theory as a function of the temperature claimed in the early literature. When renormalization is done correctly, there is no discontinuity in the temperature.

Nonperturbative Renormalization for Fermions

It is a remarkable feature of fermionic field theories that for a covariance for which $\|\hat{C}\|_1$ and $\|C\|_1$ are both finite, the effective action defined in [7] exists and is analytic in the fields and in the original interaction V, thanks to determinant bounds. For a V as in [6], with $\lambda\nu$ weak and of short range, the skeleton functions (where all relevant m-point functions are projected back to their initial values in the RG iteration) satisfy

$$\|\hat{v}_m^{(j)}\|_\infty \leq \text{const.}\|\hat{C}_j\|_1^{-(m/2)+1}\|C_j\|_1^{-1} \qquad [17]$$

For the many-electron covariance [10], with a positively curved C^d Fermi surface and with the scale decomposition [12], $\|\hat{C}_j\|_1$ is of order M^j and $\|C_j\|_1$ is of order $M^{-j(d+1)/2}$. The right-hand side of [17] then contains $M^{(d+3-m)j/2}$, which agrees (up to logarithms) with the perturbative power-counting bounds [13] only for $d=1$. In dimension $d=2$, the method has been refined by dividing the Fermi surface into angular sectors. The corresponding sectorized propagators have a better decay bound $\|C_j\|_1$, but the trade-off is sector sums at every vertex. Momentum conservation restricts these sector sums sufficiently in two dimensions to allow for good power-counting bounds. This has allowed for the construction of an interesting class of interacting fermionic models.

The major results obtained with the RG method are as follows.

Luttinger liquid behavior at zero temperature was proved for one-dimensional models with a repulsive interaction (Benfatto and Gallavotti 1995).

Fermi liquid behavior in the region where $|\lambda|\log(\beta\epsilon_0) \ll 1$ was proved for the two-dimensional model with $e(k)=k^2-1$, a local potential V, and a UV cutoff both on k and the Matsubara frequencies ω in Disertori and Rivasseau (2000).

A two-dimensional model with a band function $e(k)$ that is nonsymmetric under $k\to -k$ and a general short-range interaction was proved to be a Fermi liquid at zero temperature (Feldman *et al.* 2003, 2004). Due to the asymmetry under $k\to -k$, the Cooper instability can be proved to be absent. In Feldman *et al.* (2003, 2004), a counter-term function as in Feldman *et al.* (1996, 2000) was used. The nonperturbative proof of the corresponding inversion theorem remains open.

In $d=3$, the proof of Fermi liquid behavior remains an open problem, despite some partial results.

Condensed Matter: Bosons

Recent advances in quantum optice, in particular the trapping of ultracold atoms, have led to the experimental realization of Bose–Einstein condensation (BEC), which caused a surge of theoretical and mathematical works. For bosons, the definition of the ensembles is similar to, but more involved than in, the fermionic case. On a formal level, the functional-integral representation is analogous to fermions, except that the fields are not Grassmann fields but complex fields, and the covariance is given by a sum as in [10], but now the summation over ω runs over the bosonic Matsubara frequencies $\mathbb{M}_B = 2\pi T\mathbb{Z}$. The existence of even the free partition function in finite volume restricts the chemical potential (for free particles, $\mu < \inf_k \varepsilon(k)$ must hold). Note that C is complex and Gaussian measures with complex covariances exist in infinite dimensions only under rather restricted conditions, which are not satisfied by [10]. This is inessential for perturbative studies, where everything can be reduced to finite-dimensional integrals involving the covariance, but a nonperturbative definition of functional integrals for such systems requires again a carefully regularized (e.g., discrete-time) definition of the functional integral.

Bose–Einstein Condensation

The problem was treated to all orders in perturbation theory at positive particle density $\rho > 0$ by Benfatto (Benfatto and Gallavotti 1995). The initial interaction is again quartic, $\varepsilon(k)=k^2$, and one considers the problem at zero temperature, in the limit $\mu\to 0^-$, which is the limit in which BEC occurs for free particles. The interaction is expected to change the value of μ, given the density, so a chemical potential term is included in the action, to give the interaction

$$V(\phi) = \int \mathrm{d}\tau\mathrm{d}x\mathrm{d}y|\phi(\tau,x)|^2\nu(x-y)|\phi(\tau,y)|^2 + \nu\int \mathrm{d}\tau\mathrm{d}x|\phi(\tau,x)|^2 \qquad [18]$$

After writing $\phi(\tau,x)=\xi+\varphi(\tau,x)$, where ξ is independent of τ and x, the density condition becomes $\rho=|\xi|^2$. ν now needs to be chosen such that the free energy has a minimum at $\xi=\sqrt{\rho}$. This can be reformulated in terms of the self-energy of the boson.

Benfatto uses the RG to prove that the propagator of the interacting system no longer has the singularity structure $(i\omega - k^2)^{-1}$ but instead $(\omega^2 + c^2k^2)^{-1}$, where c is a constant. This requires a nontrivial analysis of Ward identities in the RG flow.

BEC has been proved in the Gross–Pitaevskii limit (Lieb *et al.* 2002). In the present formulation, this limit corresponds to an infinite-volume limit $L \to \infty$ where the density ρ is taken to zero as an inverse power of L. A nonperturbative proof of BEC at fixed positive particle density remains an open problem.

Superconductivity

Superconductivity (SC) occurs in fermionic systems, but it happens at energy scales where the relevant excitations have bosonic character: the Cooper pairs are bosons. In the RG framework, they arise naturally when the fermionic RG flow discussed above is stopped before it leaves the weak-coupling region and the dominant Cooper pairing term is rewritten by a Hubbard–Stratonovich transformation. The fermions can then be integrated over, resulting in the typical Mexican hat potential of an O(2) nonlinear sigma model. Effectively, one now has to deal with a problem similar to the one for BEC, but the action is considerably more complicated.

The Nonlinear Sigma Models

The prototypical model, into whose universality class both examples mentioned above fall, is that of O(N) nonlinear sigma models: both BEC and SC can be reformulated as spontaneous symmetry breaking (SSB) in the O(2) model in dimensions $d \geq 3$. For $d = 2$, long-range order is possible only at zero temperature because only then does the time direction truly represent a third dimension, preventing the Mermin–Wagner theorem from applying.

SSB has been proved for lattice O(N) models by reflection positivity and Gaussian domination methods (Fröhlich *et al.* 1976). The elegance and simplicity of this method is unsurpassed, but only very special actions satisfy reflection positivity, so that the method cannot be used for the effective actions obtained in condensed matter models. Results in the direction of proving SSB in O(N) models for $d \geq 3$ by RG methods, which apply to much more general actions, have been obtained by Balaban (1995).

See also: Bose–Einstein Condensates; Fermionic Systems; High T_c Superconductor Theory; Holomorphic Dynamics; Operator Product Expansion in Quantum Field Theory; Perturbative Renormalization Theory and BRST; Phase Transition Dynamics; Reflection Positivity and Phase Transitions.

Further Reading

Bach V, Fröhlich J, and Sigal IM (1998) Renormalization group analysis of spectral problems in quantum field theory. *Advances in Mathematics* 137: 205–298.

Balaban T (1988) Convergent renormalization expansions for lattice gauge theories. *Communications in Mathematical Physics* 119: 243–285.

Balaban T (1995) A low-temperature expansion for classical N-vector models I. A renormalization group flow. *Communications in Mathematical Physics* 167: 103–154.

Benfatto G and Gallavotti G (1995) *Renormalization Group*. Princeton: Princeton University Press.

Bricmont J and Kupiainen A (2001) Renormalizing the renormalization group pathologies. *Physics Reports* 348: 5–31.

Brydges DC and Kennedy T (1987) Mayer expansions and the Hamilton–Jacobi equation. *Journal of Statistical Physics* 48: 19–49.

Disertori M and Rivasseau V (2000) Interacting Fermi liquid in two dimensions at finite temperature I. Convergent contributions. *Communications in Mathematical Physics* 215: 251–290.

Domb C and Green M (eds.) (1976) *Phase Transitions and Critical Phenomena*, vol. 6. London: Academic Press.

Feldman J, Magnen J, Rivasseau V, and Sénéor R (1987) Construction of infrared ϕ^4_4 by a phase space expansion. *Communications in Mathematical Physics* 109: 437.

Feldman J and Trubowitz E (1990) Perturbation theory for many-fermion systems. *Helvetica Physica Acta* 63: 157.

Feldman J and Trubowitz E (1991) The flow of an electron-phonon system to the superconducting state. *Helvetica Physica Acta* 64: 213.

Feldman J, Salmhofer M, and Trubowitz E (1996) Perturbation theory around non-nested Fermi surfaces I. Keeping the Fermi surface fixed. *Journal of Statistical Physics* 84: 1209–1336.

Feldman J, Salmhofer M, and Trubowitz E (2000) An inversion theorem in Fermi surface theory. *Communications on Pure and Applied Mathematics* 53: 1350–1384.

Feldman J, Knörrer H, and Trubowitz E (2003) A class of Fermi liquids. *Reviews in Mathematical Physics* 15: 949–1169.

Feldman J, Knörrer H, and Trubowitz E (2004) *Communications in Mathematical Physics* 247: 1–319.

Fröhlich J, Simon B, and Spencer T (1976) Infrared bounds, phase transitions, and continuous symmetry breaking. *Communications in Mathematical Physics* 50: 79.

Gallavotti G (1985) Renormalization theory and ultraviolet stability via renormalization group methods. *Reviews of Modern Physics* 57: 471–569.

Gawedzki K and Kupiainen A (1985) Massless lattice ϕ^4_4 theory: Rigorous control of a renormalizable asymptotically free model. *Communications in Mathematical Physics* 99: 197–252.

Lieb E, Seiringer R, Solovej JP, and Yngvason J (2002) The ground state of the Bose gas. In: *Current Developments in Mathematics, 2001*, pp. 131–178. Cambridge: International Press.

Polchinski J (1984) Renormalization and effective Lagrangians. *Nuclear Physics B* 231: 269.

Rivasseau V (1993) *From Perturbative to Constructive Renormalization*. Princeton, NJ: Princeton University Press.

Salmhofer M (1998) *Renormalization: An Introduction*, Springer Texts and Monographs in Physics. Heidelberg: Springer.

Wiese KJ (2001) Polymerized membranes, a review. In: Domb C and Lebowitz J (eds.) *Phase Transitions and Critical Phenomena*, vol. 19. Academic Press.

Resonances

N Burq, Université Paris-Sud, Orsay, France

Introduction

In quantum mechanics and wave propagation, eigenvalues (and eigenfunctions) appear naturally as they describe the behavior of a quantum system (or the vibration of a structure). There are however some cases where these simple notions do not suffice and one has to appeal to the more subtle notion of resonances. For example, if the vibration of a drum is well understood in terms of eigenvalues (the audible frequencies) and eigenfunctions (the corresponding vibrating modes), the notion of resonances is necessary to understand the propagation of waves in the exterior of a bounded obstacle. Another example (taken from Zworski (2002)) which allows us to understand both the similarities of resonances with eigenvalues and their differences is the following: consider the motion of a classical particle submitted to a force field deriving from the potential $V_1(x)$ on a bounded interval as shown in **Figure 1a**. If the classical momentum is denoted by ξ, then the classical energy is given by

$$E = |\xi|^2 + V_1(x)$$

and the classical motion is given by the relations of Hamiltonian mechanics:

$$\dot{x} = \frac{\partial E}{\partial \xi} = 2\xi, \qquad \dot{\xi} = -\frac{\partial E}{\partial x} = -V'(x)$$

Since energy is conserved, if the initial energy is smaller than the top of the barrier, then the classical particle bounces forever in the well. Now we can consider the same example with the potential $V_2(x)$ on $\mathbb{R}$ as shown in **Figure 1b**. Of course, if the particle is initially inside the well (with the same energy as before), the classical motion remains the same.

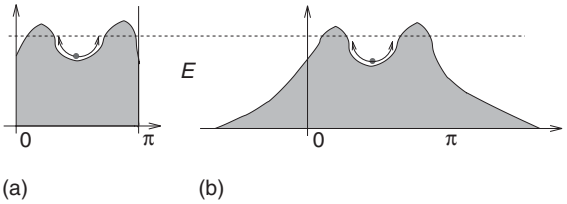


Figure 1a, b A particle trapped in a well.

On the quantum mechanics point of view, both systems are described by the Hamiltonians

$$H_i = -h^2 \frac{\mathrm{d}^2}{\mathrm{d}x^2} + V_i(x)$$

acting on $L^2([-1,1])$ (with boundary conditions) and $L^2(\mathbb{R})$, respectively. In the first case, H_1 has a discrete spectrum, $\lambda_{j,h} \in \mathbb{R}$ with eigenfunctions $e_{j,h}(x), j \in \mathbb{N}$, and the time evolution of the system is given by

$$\mathrm{e}^{\mathrm{i}tH_1}u = \sum_j \mathrm{e}^{\mathrm{i}t\lambda_{j,h}} u_{j,h} \times e_{j,h} \qquad [1]$$

where $u_{j,h} \times e_{j,h}$ is the orthogonal projection of u on the eigenspace $\mathbb{C}e_{j,h}$. In the second case, H_2 has no square integrable eigenfunction, and no simple description as [1] can consequently hold. However as $h \to 0$, the correspondence principle tells us that quantum mechanics should get close to classical mechanics. Since for both quantum problems the classical limit is the same (at least for initial states confined in the well with energy E), we expect that for the second potential there should exist a quantum state corresponding to the classical one. In fact, this is indeed the case and one can show that there exist resonant states $e_{j,h}$ associated to resonances $E_{j,h}$ which are solution of the equation

$$H_2 e_{j,h} - E_{j,h} e_{j,h}, \quad E_{j,h} \sim E$$

are not square integrable, but still have moderate growth at infinity and are confined in the interior of the well (see sections "Definition" and "Location of resonances"). On the other hand, the first quantum system is confined, whereas the second one is not and we know that even for initial states confined in the well, tunneling effect allows the quantum particle to escape to infinity. This fact should be described by the theory as a main difference between eigenvalues and resonances. This is indeed the case as the resonances $E_{j,h}$ are not real (contrarily to eigenvalues of self-adjoint operators) but have a nonvanishing imaginary part (see section "Resonance-free regions")

$$\mathrm{Im}\, E_{j,h} \sim \mathrm{e}^{-C/h}$$

If we assume that a similar description as [1] still holds for the second system, at least locally in space (see section "Resonances and time asymptotics"), then, for time $t >> \mathrm{e}^{C/h}$, the factor $\mathrm{e}^{\mathrm{i}tE_h}$ becomes very small (the quantum particle has left the well due to tunneling effect).

There have been several studies on resonances and scattering theory and the presentation here cannot be complete. For a more in-depth presentation, one can

consult the books by Lax and Phillips (1989) and Hislop and Sigal (1987), or the reviews on resonances by Vodev (2001) and Zworski (1994) for example.

Definition

There are different (equivalent) definitions of resonances. The most elegant is certainly the Helffer and Sjöstrand (1986) definition (see also the presentation of complex scaling by Combes *et al.* (1984) and the very general "black box" framework by Sjöstrand and Zworski (1991)). However, it requires a few prerequisites and we preferred to stick to the more elementary (but less general) resolvent point of view. The starting point for this definition of resonances is the fact that the eigenvalues of a (self-adjoint) operator P are the points where P is not injective. The more general resonances will be the points where the operator is not invertible (on suitable spaces).

More precisely, consider a perturbation of the Laplace operator on $\mathbb{R}^n$, $P_0(h) = -h^2\Delta$ in the following sense: let $\Theta \subset \mathbb{R}^d$ be a (possibly empty) smooth obstacle whose complementary, $\Omega = \Theta^c$, is connected. Consider a classical self-adjoint operator defined on $L^2(\Omega)$:

$$P_h u = (-h^2\Delta + V(x))u \qquad [2]$$

with boundary conditions (Dirichlet)

$$u\,|_{\partial\Theta} = 0 \qquad [3]$$

(Neumann boundary conditions could be used too). This setting contains both the Schrödinger operator ($P_h = -h^2\Delta + V(x) - E$ on $\Omega = \mathbb{R}^n$) and the Helmoltz equation with Dirichlet conditions, in the exterior of an obstacle (waves at large frequencies: $P = -\Delta - \tau^2$; in this case, define $h = \tau^{-1}$ and $P_h = -h^2\Delta$), which we shall define as acoustical scattering.

We assume that P is a perturbation of P_0, that is, $V \to 0, |x| \to +\infty$ sufficiently fast (see Sjöstrand and Zworski (1991) for the very general black box assumptions). For example, this perturbation assumption is fulfilled if V has compact support. Then the resolvent $P_h(z) = (P_h - z)^{-1}$ is well defined for $\mathrm{Im}\, z \neq 0$ as a bounded operator from $L^2(\Omega)$ to

$$H^2(\Omega) \cap H_0^1(\Omega)$$

(because the operator P_h is self-adjoint). However, it is not bounded for $z > 0$ on $L^2(\Omega)$ because the essential spectrum of P_h is precisely the semiaxis $z > 0$, but it admits a meromorphic continuation from $\mathbf{Im}\, z > 0$ toward the lower half-plane:

$$R_h(z) : L^2(\Omega)_{\mathrm{comp}} \to L^2(\Omega)_{\mathrm{loc}}$$

The poles of this resolvent R_h are by definition the semiclassical resonances, $\mathrm{Res}_{\mathrm{sc}}(P_h)$.

Remark 1 In the case of acoustical scattering ($P = -\Delta - \tau^2, \tau = h^{-1}$), the introduction of the additional parameter z is pointless and one works directly with the parameter $\tau = h^{-1}\sqrt{z}$. In that case the resolvent $R(\tau)(-\Delta - \tau^2)^{-1}$ is well defined for $\mathbf{Im}\, \tau < 0$, the essential spectrum is precisely the axis $\tau \in \mathbb{R}$ and the resolvent admits a meromorphic continuation from $\mathbf{Im}\, z < 0$ toward the upper half-plane (with possibly a cut at 0):

$$R(\tau) : L^2(\Omega)_{\mathrm{comp}} \to L^2(\Omega)_{\mathrm{loc}}$$

The acoustic resonances are by definition the poles of this meromorphic continuation. They are related to semiclassical resonances by the relation

$$\mathrm{Res}_{\mathrm{sc}} = h\sqrt{\mathrm{Res}_{\mathrm{ac}}}$$

It can also be shown that if z is a resonance, there exists an associated resonant state e_z such that

$$(P_h - z)e_z = 0$$

the function e_z satisfies Sommerfeld radiation conditions (in polar coordinates $(r, \theta) \in [0, +\infty) \times \mathbb{S}^{n-1}$)

$$|h\partial_r e - i\sqrt{z}e| \leq C|\mathrm{e}^{\mathrm{i}\sqrt{z}r}|/r^{1+n/2}$$

and the function

$$\frac{e_z}{1 + r^{(1/2)+\epsilon}}\ \mathrm{e}^{\mathrm{i}\sqrt{z}r}$$

is square integrable.

Resonance-Free Regions

The very first result about resonance-free regions is based on Rellich uniqueness theorem (uniqueness for solutions of elliptic second-order equations) and says that there are no real resonances (except possibly 0). The more precise determination of resonance-free regions (originally in acoustical scattering) has been a subject of study from the 1960s and it has motivated a large range of works from the multiplier methods of Morawetz (1975) to the general propagation of singularity theorem of Melrose and Sjöstrand (1978). To state the main result in this direction, we need the notion of nontrapping perturbation.

Definition 1 A generalized bicharacteristic at energy $E(x(s), \xi(s))$ is an integral curve of the Hamiltonian field

$$H_p = \frac{\partial p}{\partial \xi}\frac{\partial}{\partial x} - \frac{\partial p}{\partial x}\frac{\partial}{\partial \xi}$$

of the principal symbol $p(x, \xi) = |\xi|^2 + V(x)$ of the operator P, included in the characteristic set $p(x, \xi) = E$ and which, when hitting the boundary of the obstacle, reflects according to the laws of geometric optics (see (Melrose and Sjöstrand 1978)).

The operator P (or by extension the obstacle in the case of acoustic scattering) is said to be nontrapping at energy E if all generalized bicharacteristics go to the infinity:

$$\lim_{s\to\pm\infty} |x(s)| = +\infty$$

The operator P (or by extension the obstacle in the case of acoustic scattering) is said to be nontrapping near energy E if P is nontrapping at energy E' for E' in a neighborhood of E.

The following result was obtained in different generalities by Morawetz (1975), Melrose and Sjöstrand (1978), and others.

Theorem 1 *Assume that the operator P is nontrapping near energy E. Then for any $N > 0$ there exist $h_0 > 0$ such that for $0 < h < h_0$ there are no resonances in the set*

$$\{z; |\mathrm{Im}\, z| \leq -Nh\log(h)\}$$

In the case of analytic geometries (and coefficients), this result (see Bardos *et al.* 1987) can be improved to

Theorem 2 *Assume that the operator P is non trapping. Then there exist $\epsilon > 0, N_0 > 0$ and $h_0 > 0$ such that for $0 < h < h_0$ there are no resonances in the set*

$$\{z; |\mathrm{Im}\, z| \leq N_0 h^{1-(1/3)}\} \cap \{|z - E| \leq \epsilon\}$$

Remark 2 In the case of acoustical scattering, with the new definition of resonances, $\tau = h^{-1}\sqrt{z}$, the resonance-free zones have respectively the forms

$$\{z; |\mathrm{Im}\, z| \leq -N\log(|z|), |z| >> 1\}$$
$$\{z; |\mathrm{Im}\, z| \leq N_0|z|^{1/3}, |z| >> 1\}$$

In the case of trapping perturbations, the first result was obtained by Burq (1998).

Theorem 3 *There exist $C > 0$ and $h_0 > 0$ such that for $0 < h < h_0$ there are no resonances in the set*

$$\{z; |\mathrm{Im}\, z| \leq N_0\ \mathrm{e}^{-C/h}\} \cap \{|z - E| \leq \epsilon\}$$

Resonances and Time Asymptotics

The relationship between eigenfunctions/eigenvalues and time asymptotics is straightforward. This is no longer the case for resonances. For nontrapping problems however, this question has been studied in the late 1960s by Lax and Phillips (1989) and Vainberg (1968). In particular, this approach was decisive to study the local energy decay in acoustical scattering. As a consequence of Theorem 1, we have

Theorem 4 *If the acoustical problem is nontrapping, then there exist $C, \alpha > 0$ such that for any solution of the wave equation*

$$\Box u = 0, \quad u|_{t=0} = u_0,\ \partial_t u|_{t=0} = u_1,\ u|_{\Gamma_D} = 0,\ \frac{\partial u}{\partial_n}|_{\Gamma_N} = 0$$

with compactly supported initial data (u_0, u_1) (in a fixed compact), one has

$$E_{\mathrm{loc}}(u) = \int_{\Omega\cap\{|x|\leq C} |\nabla u|^2 + |\partial_t u|^2 \leq \begin{cases} \mathrm{Ce}^{-\alpha t} & \text{if the space dimension is even} \\ \dfrac{C}{t^d} & \text{if the space dimension is odd} \end{cases} \quad [4]$$

Trapping perturbations were investigated more recently. In that case, the local energy decays, but the rate cannot be uniform. The first trapping example in acoustic scattering was studied by Ikawa (1983): the obstacle is the union of a finite number (and at least two) convex bodies. In that case, one has

Theorem 5 *For any $\epsilon > 0$ there exists $C > 0$ such that for any initial data supported in a fixed compact set*

$$E_{\mathrm{loc}}(u)(t) \leq \mathrm{Ce}^{-\alpha t}\|(u_0, u_1)\|^2_{D((1-\Delta)^{(1+\epsilon)/2})}$$

where $D((1-\Delta)^{(1+\epsilon)/2})$ is the domain of the operator $(1-\Delta)^{(1+\epsilon)/2}$. Remark that the norm in $D((1-\Delta)^{1/2})$ is the natural energy and consequently the estimate above exhibits a loss of ϵ derivatives. For strongly trapping perturbations, the results are worse. They are consequences of Theorem 3.

Theorem 6 *For any k there exists $C_k > 0$ such that for any initial data supported in a fixed compact set*

$$E_{\mathrm{loc}}(u)(t) \leq \frac{C_k}{\log(t)^{2k}}\|(u_0, u_1)\|^2_{D((1-\Delta)^{(1+k)/2})}$$

One can also obtain real asymptotic expansions in terms of resonances (see the work by Tang and Zworski (2000)).

Theorem 7 *Let $\chi \in C_c^\infty(\mathbb{R}^n)$ and $\psi \in C_c^\infty((0,\infty))$ and let* chsupp $\psi = [a, b]$. *There exists $0 < \delta < c(h) < 2\delta$ such that for every $M > M_0$ there exists $L = L(M)$, and we have*

$$\chi \mathrm{e}^{-\mathrm{i}t\mu(P)/h}\chi\psi(P) = \sum_{z\in\Omega(h)\cap\mathrm{Res}(P)} \chi\mathrm{Res}(\mathrm{e}^{-\mathrm{i}t\mu(\bullet)/h} \times R(\bullet, h), z)\chi\psi(P) + \mathcal{O}_{\mathcal{H}\to\mathcal{H}}(h^\infty), \quad \text{for } t > h^{-L} \quad [5]$$

$$\Omega(h) = (a - c(h), b + c(h)) - \mathrm{i}[0, h^M)$$

where $\mathrm{Res}(f(\bullet), z)$ *denotes the residue of a meromorphic family of operators,* f*, at* z*.*

The function $c(h)$ depends on the distribution of resonances: roughly speaking we cannot "cut" through a dense cloud of resonances. Even in the very well understood case of the modular surface there is, currently at least, a need for some nonexplicit grouping of terms. The same ideas can be applied to acoustic scattering.

Trace Formulas

Trace formulas provide a description of the classical/quantum correspondence: one side is given by the trace of a certain function of the operator $f(P_h)$, whereas the other side is described in terms of classical objects (closed orbits of the classical flow). In the case of discrete eigenvalues, the question is relatively simple and can be solved by using the spectral theorem. In the case of continuous spectrum, the problem is much more subtle (self-adjoint operators with continuous spectrum behave in some ways as non-normal operators). It has been studied by Lax and Phillips (1989), Bardos *et al.* (1982), and Melrose (1982). More recently, Sjöstrand (1997) introduced a local notion of trace formulas.

Let $W \subset \Omega$ be an open precompact subsets of $e^{i[-2\theta_0,0]}]0, +\infty[$. Assume that the intersections I and J of W and Ω with the real axis are intervals and that Ω is simply connected.

Theorem 8 *Let* $f(z,h)$ *be a family of holomorphic functions on* $z \in \Omega$ *such that* $|f|_{\Omega\setminus W}| \leq 1$*. Let* $\chi \in C_0^\infty(\mathbb{R})$ *equal to 1 on a neighborhood of* $\overline{I}$*. Then*

$$\mathrm{Trace}\big((\chi f)(P_h) - (\chi f)(-h^2\Delta)\big) = \sum_{\lambda \text{ a resonance of } P_h \cap \Omega} f(\lambda, h) + \mathcal{O}(h^{-n})$$

The use of this result with a clever choice of functions f allows Sjöstrand to show that an analytic singularity of the function $E \mapsto \mathrm{Vol}(\{x; V(x) \geq E\})$ (observe that if V is bounded, this function vanishes for large E and consequently it has analytic singularities) gives a lower bound for Ω a neighborhood of E

$$\sharp \mathrm{Res}(P_h) \cap \Omega \geq ch^{-n}$$

which coincides with the upper bound (see Zworski (2002) and the references given there).

Location of Resonances

In some particular cases, one can expect to have a precise description of the location of resonances. This is the case in Ikawa's example in acoustic scattering where the obstacle is the union of two disjoint convex bodies. In this case, the line minimizing the distance, d, between the bodies is trapped. However, this trapped trajectory is isolated and of hyperbolic type (unstable). Ikawa (1983) and Gérard (1988) have obtained:

Theorem 9 *There exist geometric positive constants* $k_p \to +\infty$ *as* $p \to +\infty$ *such that all resonances located above the line* $\mathrm{Im}\, z \geq -C$ *(*C *arbitrary large but fixed) have an asymptotic expansion*

$$\lambda \sim \lambda_{j,p} + \sum_l a_{l,p} \lambda_{j,p}^{-l/2} + \mathcal{O}(\lambda_{j,p}^{-\infty}), \quad j \to +\infty$$

where the approximate resonances

$$\lambda_{j,p} = j\frac{\pi}{d} - \mathrm{i}k_p$$

are located on horizontal lines.

Another example is when the obstacle is convex. This example is nontrapping and Sjöstrand and Zworski (1999) are able to prove that the resonances in any region $\mathrm{Im}\, z \geq N|z|^{1/3}$ (N arbitrary large) are asymptotically distributed near cubic curves

$$\mathcal{C}_j = \{z \in \mathbb{C}; \mathrm{Im}\, z = -c_j|z|^{1/3}\}$$

Finally, the last main example where one can give a precise asymptotic for resonances is when there exists a stable (elliptic) periodic trajectory for the Hamiltonian flow. In that case it had been known from the 1960s (see the works by Babič (1968)) that one can construct quasimodes, that is, compactly supported approximate solutions of the eigenfunctions equation:

$$(P_h - E_h)e_j = \mathcal{O}(h^\infty)$$

It is only recently that Tang and Zworski (1998) and Stefanov (1999) proved that these quasimodes constructions imply the existence of resonances asymptotic to $E_h, h \to 0$.

See also: h-Pseudodifferential Operators and Applications; Semi-Classical Spectra and Closed Orbits.

Further Reading

Babič VM (1968) Eigenfunctions which are concentrated in the neighborhood of a closed geodesic. *Zapiski Nauchnykh Seminanov Leningradkogo Otdeleniya Matematicheskogo Instituta Imeni V.A. Steklova* 9: 15–63.

Bardos C, Guillot JC, and Ralston J (1982) La relation de Poisson pour l'équation des ondes dans un ouvert non borné. Application à la théorie de la diffusion. *Communications in Partial Differential Equations* 7: 905–958.

Bardos C, Lebeau G, and Rauch J (1987) Scattering frequencies and Gevrey 3 singularities. *Inventiones Mathematicae* 90: 77–114.

Burq N (1998) Décroissance de l'énergie locale de l'équation des ondes pour le problème extérieur et absence de résonance au voisinage du réel. *Acta Mathematica* 180: 1–29.

Combes J-M, Duclos P, and Seiler R (1984) On the shape resonance. In: *Resonances – Models and Phenomena* (*Bielefeld, 1984*), Lecture Notes in Physics, vol. 211, pp. 64–77. Berlin: Springer.

Gérard C (1988) Asymptotique des pôles de la matrice de scattering pour deux obstacles strictement convexes. *Supplément au Bulletin de la Société Mathématique de France* 116: 146 pp.

Helffer B and Sjöstrand J (1986) Resonances en limite semi-classique. *Mémoire de la S.M.F* 114(24–25): 228 pp.

Hislop PD and Sigal IM (1987) Shape resonances in quantum mechanics. In: *Differential Equations and Mathematical Physics* (*Birmingham, Ala., 1986*), Lecture Notes in Mathematics, vol. 1285, pp. 180–196. Berlin: Springer.

Ikawa M (1983) On the poles of the scattering matrix for two convex obstacles. *Journal of Mathematics of the Kyoto University* 23: 127–194.

Lax PD and Phillips RS (1989) *Scattering Theory*, Pure and Applied Mathematics, 2nd edn., vol. 26. Boston: Academic Press.

Melrose RB (1982) Scattering theory and the trace of the wave group. *Journal of Functional Analysis* 45: 429–440.

Melrose RB and Sjöstrand J (1978) Singularities of boundary value problems. I. *Communications in Pure and Applied Mathematics* 31: 593–617.

Melrose RB and Sjöstrand J (1982) Singularities of boundary value problems. II. *Communications in Pure and Applied Mathematics* 35: 129–168.

Morawetz CS (1975) Decay for solutions of the exterior problem for the wave equation. *Communication in Pure Applied Mathematics* 28: 229–264.

Sjöstrand J (1997) A trace formula and review of some estimates for resonances. In: Rodino L (ed.) *Microlocal Analysis and Spectral Theory* (*Lucca, 1996*), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., vol. 490, pp. 377–437. Dordrecht: Kluwer Academic.

Sjöstrand J and Zworski M (1991) Complex scaling and the distribution of scattering poles. *Journal of the American Mathematical Society* 4(4): 729–769.

Sjöstrand J and Zworski M (1999) Asymptotic distribution of resonances for convex obstacles. *Acta Mathematica* 183(2): 191–253.

Stefanov P (1999) Quasimodes and resonances: sharp lower bounds. *Duke Mathematical Journal* 99(1): 75–92.

Tang S-H and Zworski M (1998) From quasimodes to reasonances. *Mathematical Research Letters* 5(3): 261–272.

Tang SH and Zworski M (2000) Resonance expansions of scattered waves. *Communication in Pure and Applied Mathematics* 53(10): 1305–1334.

Vainberg BR (1968) On the analytical properties of the resolvent for a certain class of operator pencils. *Mathematics of the USSR Sbornik* 6(2): 241–273.

Vodev G (2001) *Resonances in euclidean scattering. Cubo Matematica Educacional* 3: 317–360. http://www.math.sciences.univ-nantes.fr.

Zworski M (1994) Counting scattering poles. In: Ikawa M (ed.) *Spectral and Scattering Theory (Sanda, 1992)*, Lecture Notes in Pure and Applied Mathematics, vol. 161, pp. 301–331. New York: Dekker.

Zworski M (2002) Quantum resonances and partial differential equations. In: Li Ta Tsien (ed.) *Proceedings of the International Congress of Mathematicians* (*Beijing, 2002*), vol. III, pp. 243–252. Beijing: Higher Education Press.

Ricci Flow *see* Singularities of the Ricci Flow

Riemann Surfaces

K Hulek, Universität Hannover, Hannover, Germany

Introduction

Riemann surfaces were first studied as the natural domain of definition of (multivalued) holomorphic or meromorphic functions. They were the starting point for the development of the theory of real and complex manifolds (see Weyl (1997)). Nowadays, Riemann surfaces are simply defined as one-dimensional complex manifolds (see the next section). Compact Riemann surfaces can be embedded into projective spaces and are thus, by virtue of Chow's theorem, algebraic curves. By uniformization theory, the universal cover of a connected Riemann surface is either the unit disk, the complex plane, or the Riemann sphere (see the section "Uniformization").

This article discusses the basic theory of compact Riemann surfaces, such as their topology, their periods, and the definition of the Jacobian variety. Studying the zeros and poles of meromorphic functions leads to the notion of divisors and linear systems. In modern language this can be rephrased in terms of line bundles, resp. locally free sheaves (see the section "Divisors, linear systems, and line bundles"). One of the fundamental results is the Riemann–Roch theorem which expresses the difference between the dimension of a linear system and that of its adjoint system in terms of the degree of the linear system and the genus of the curve. This theorem has been vastly generalized and is truly one of the cornerstones of algebraic geometry. A formulation of this result and a discussion of some of its applications are also discussed.

A study of the subsets of the Jacobians parametrizing linear systems of given degree and dimension leads to Brill–Noether theory, which is discussed in the section "Brill–Noether theory." This is followed by a brief introduction to the theory of equations and syzygies of canonical curves.

Moduli spaces play a central role in the theory of complex variables and in algebraic geometry. Arguably, the most important of these is the moduli space of curves of genus g. This and related moduli problems are treated in the section "Moduli of compact Riemann surfaces." In particular, the space of stable maps is closely related to quantum cohomology. Finally, we present a brief discussion of the Verlinde formula and conformal blocks.

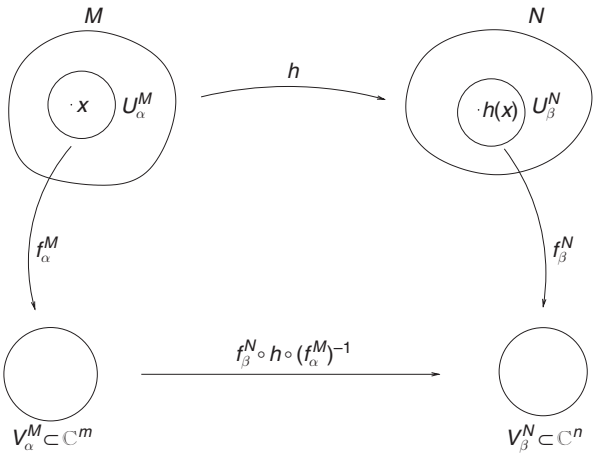


Figure 2 Holomorphic map between manifolds.

Basic Definitions

Riemann surfaces are one-dimensional complex manifolds. An n-dimensional complex manifold M is a topological Hausdorff space (i.e., for any two points $x \neq y$ on M, there are disjoint open neighborhoods containing x and y), which has a countable basis for its topology, together with a complex atlas $\mathcal{A}$. The latter is an open covering $(U_\alpha)_{\alpha\in\mathcal{A}}$ together with homeomorphisms $f_\alpha : U_\alpha \to V_\alpha \subset \mathbb{C}^n$, where the U_α are open subsets of M and the V_α are open sets in $\mathbb{C}^n$. The main requirement is that these charts are holomorphically compatible, that is, for $U_\alpha \cap U_\beta \neq \emptyset$, the map shown in Figure 1,

$$f_\beta \circ f_\alpha^{-1}|_{f_\alpha(U_\alpha\cap U_\beta)} : f_\alpha(U_\alpha \cap U_\beta) \to f_\beta(U_\alpha \cap U_\beta) \subset \mathbb{C}^n$$

is biholomorphic. A map $h : M \to N$ between two complex manifolds is holomorphic if it is so with respect to the local charts. This means the following:

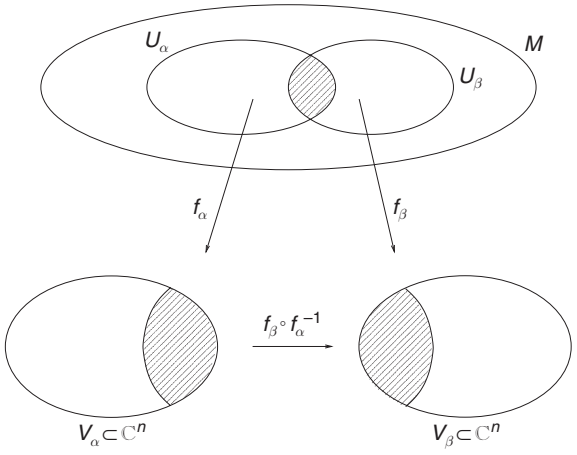


Figure 1 Charts of a complex manifold.

for each point $x \in M$, there are charts $f_\alpha^M : U_\alpha^M \to V_\alpha^M \subset \mathbb{C}^n$ near x and $f_\beta^N : U_\beta^N \to V_\beta^N \subset \mathbb{C}^m$ near $h(x)$ with $h(U_\alpha^M) \subset U_\beta^N$ such that the map shown in Figure 2

$$f_\beta^N \circ h \circ (f_\alpha^M)^{-1} : V_\alpha^M \to V_\beta^N \subset \mathbb{C}^m$$

is holomorphic (one checks easily that this does not depend on the choice of the charts).

A Riemann surface is a one-dimensional complex manifold. Trivial examples are given by open sets in $\mathbb{C}$ (where one chart suffices). Another example is the Riemann sphere $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, which can be covered by the two charts given by $z \neq \infty$ and $z \neq 0$. Both of these charts are homeomorphic to $\mathbb{C}$ with the transition function given by $z \mapsto 1/z$. Historically, Riemann surfaces were viewed as (branched) coverings of $\mathbb{C}$ or of the sphere, where they appear as the natural domain of definition of multivalued holomorphic or meromorphic functions.

Uniformization

If M is a Riemann surface, then its universal covering $\tilde{M}$ is again a Riemann surface. The connected and simply connected Riemann surfaces can be fully classified. Let

$$\mathbb{E} = \{z \in \mathbb{C}; |z| < 1\}$$

be the unit disk and $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ the Riemann sphere. The latter can be identified with the complex projective line $\mathbb{P}^1_{\mathbb{C}}$.

Theorem 1 (Generalized Riemann mapping theorem). *Every connected and simply connected Riemann surface is biholomorphically equivalent*

to the unit disk $\mathbb{E}$, *the complex plane* $\mathbb{C}$, *or the Riemann sphere* $\hat{\mathbb{C}}$.

This theorem was proved rigorously by Koebe and Poincaré at the beginning of the twentieth century.

Compact Riemann Surfaces

The topological structure of a compact Riemann surface C is determined by its genus g (Figure 3). Topologically, a Riemann surface of genus g is a sphere with g handles or, equivalently, a torus with g holes.

Analytically, the genus can be characterized as the maximal number of linearly independent holomorphic forms on C (see also the section "The Riemann–Roch theorem and applications").

There exists a very close link with algebraic geometry: every compact Riemann surface C can be embedded into some projective space $\mathbb{P}^n_{\mathbb{C}}$ (in fact already into $\mathbb{P}^3_{\mathbb{C}}$). By Chow's theorem, C is then a (projective) algebraic variety, that is, it can be described by finitely many homogeneous equations. It should be noted that such a phenomenon is special to complex dimension 1. The crucial point is that one can always construct a non-constant meromorphic function on a Riemann surface (e.g., by Dirichlet's principle). Given such a function, it is not difficult to find a projective embedding of a compact Riemann surface C. On the other hand, it is easy to construct a compact two-dimensional torus $T = \mathbb{C}^2/L$ for some suitably chosen lattice L, which cannot be embedded into any projective space $\mathbb{P}^n_{\mathbb{C}}$.

The dichotomy Riemann surface/algebraic curve arises from different points of view: analysts think of a real two-dimensional surface with a Riemannian metric which, via isothermal coordinates, defines a holomorphic structure, whereas algebraic geometers think of a complex one-dimensional object.

In this article, the expressions compact Riemann surface and (projective) algebraic curve are both used interchangeably. The choice depends on which expression is more commonly used in the part of the theory which is discussed in the relevant section.

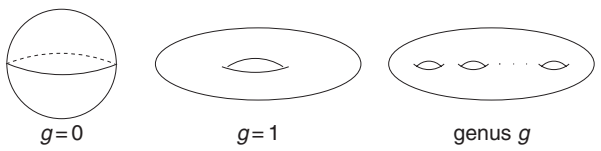


Figure 3 Genus of Riemann surfaces.

Periods and the Jacobian

On a compact Riemann surface C of genus g, there exist $2g$ homologically independent paths, that is, $H_1(C, \mathbb{Z}) \cong \mathbb{Z}^{2g}$.

Let $\gamma_1, \ldots, \gamma_{2g}$ be a basis of $H_1(C, \mathbb{Z})$ and let $\omega_1, \ldots, \omega_g$ be a basis of the space of holomorphic 1-forms on C. Integrating these forms over the paths $\gamma_1, \ldots, \gamma_{2g}$ defines the period matrix

$$\Omega = \begin{pmatrix} \int_{\gamma_1} \omega_1 & \cdots & \int_{\gamma_{2g}} \omega_1 \\ \vdots & & \vdots \\ \int_{\gamma_1} \omega_g & \cdots & \int_{\gamma_{2g}} \omega_g \end{pmatrix}$$

If $Q = (\gamma_i, \gamma_j)$ is the intersection matrix of the paths $\gamma_1, \ldots, \gamma_{2g}$, then Ω satisfies the Riemann bilinear relations

$$\Omega Q \Omega^t = 0, \quad \sqrt{-1}\, \Omega Q \overline{\Omega}^t > 0 \qquad [1]$$

where the latter condition means positive definite. One can choose (see Figure 4) $\gamma_1, \ldots, \gamma_{2g}$ such that

$$Q = J = \begin{pmatrix} 0 & \mathbf{1}_g \\ -\mathbf{1}_g & 0 \end{pmatrix}$$

where $\mathbf{1}_g$ is the $g \times g$ unit matrix. Moreover, $\omega_1, \ldots, \omega_g$ can be chosen such that

$$\Omega = \begin{pmatrix} 1 & \cdots & 0 & \tau_{11} & \cdots & \tau_{1g} \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & \tau_{g1} & \cdots & \tau_{gg} \end{pmatrix}$$

Let

$$\Omega_0 = (\tau_{ij})_{1 \le i,j \le g}$$

Then the Riemann bilinear relations [1] become

$$\Omega_0 = \Omega_0^t, \quad \operatorname{Im} \Omega_0^t > 0$$

that is, Ω_0 is an element of the Siegel upper half-space

$$\mathbb{H}_g = \{\tau \in \operatorname{Mat}(g \times g, \mathbb{C}); \tau = \tau^t, \operatorname{Im} \tau > 0\}$$

The matrix Ω_0 is defined by the Riemann surface C only up to the action of the symplectic group

$$\operatorname{Sp}(2g, \mathbb{Z}) = \{M \in \operatorname{Mat}(2g \times 2g, \mathbb{Z}); MJM^t = J\}$$

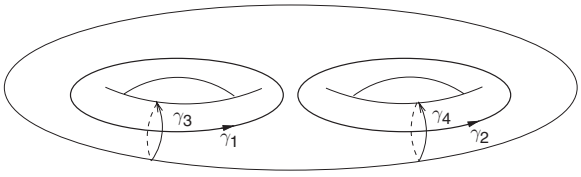


Figure 4 Homology of a compact Riemann surface.

which acts on the Siegel space $\mathbb{H}_g$ by

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} : \tau \mapsto (A\tau + B)(C\tau + D)^{-1}$$

Here $A, \ldots, D$ are $g \times g$ blocks.

The rows of the matrix Ω define a rank-$2g$ lattice L_Ω in $\mathbb{C}^g$ and the Jacobian of C is the torus

$$J(C) = \mathbb{C}^g / L_\Omega$$

More intrinsically, one can define $J(C)$ as follows. Let $H^0(C, \omega_C)$ be the space of holomorphic differential forms on C. Then, integration over cycles defines a monomorphism

$$H_1(C, \mathbb{Z}) \to H^0(C, \omega_C)^*$$
$$\gamma \mapsto \int_\gamma$$

and

$$J(C) = H^0(C, \omega_C)^* / H_1(C, \mathbb{Z})$$

For a fixed base point $P_0 \in C$, the Abel–Jacobi map is defined by

$$u : C \to J(C)$$
$$P \mapsto \left(\int_{P_0}^{P} \omega_1, \ldots, \int_{P_0}^{P} \omega_g \right)$$

Here, the integration is taken over some path from P_0 to P. Obviously, the integral depends on the choice of this path, but since $J(C)$ was obtained by dividing out the periods given by integrating over a basis of $H_1(C, \mathbb{Z})$, the map is well defined.

Let C^d be the dth Cartesian product of C, that is, the set of all ordered d-tuples $(P_1, \ldots, P_d)$. Then, u defines a map

$$u^d : C^d \to J(C)$$
$$(P_1, \ldots, P_d) \mapsto u(P_1) + \cdots + u(P_d)$$

where $+$ is the usual addition on the torus $J(C)$. If $d = g - 1$, then

$$\Theta = \mathrm{Im}(u^{g-1}) \subset J(C)$$

is a hypersurface (i.e., has codimension 1 in $J(C)$) and is called a theta divisor. A different choice of the base point P_0 results in a translation of the theta divisor. Using the theta divisor, one can show that $J(C)$ is an abelian variety, that is, $J(C)$ can be embedded into some projective space $\mathbb{P}^n_{\mathbb{C}}$. The pair $(J(C), \Theta)$ is a principally polarized abelian variety and Torelli's theorem states that C can be reconstructed from its Jacobian $J(C)$ and the theta divisor Θ.

Divisors, Linear Systems, and Line Bundles

A divisor D on C is a formal sum

$$D = n_1 P_1 + \cdots + n_k P_k, \quad P_i \in C, \ n_i \in \mathbb{Z}$$

The degree of D is defined as

$$\deg D = n_1 + \cdots + n_k$$

and D is called "effective" if all $n_i \geq 0$. Every meromorphic function $f \neq 0$ defines a divisor

$$(f) = f_0 - f_\infty$$

where f_0 are the zeros of f and f_∞ the poles (each counted with multiplicity). Divisors of the form (f) are called principal divisors and the degree of any principal divisor is 0 (see the next section). Two divisors D_1 and D_2 are called linearly equivalent ($D_1 \sim D_2$) if their difference is a principal divisor, that is,

$$D_1 - D_2 = (f)$$

for some meromorphic function $f \neq 0$. This defines an equivalence relation on the group $\mathrm{Div}(C)$ of all divisors on C. Since principal divisors have degree 0, the notion of degree also makes sense for classes of linearly equivalent divisors. We define the divisor class group of C by

$$\mathrm{Cl}(C) = \mathrm{Div}(C) / \sim$$

The degree map defines an exact sequence

$$0 \to \mathrm{Cl}^0(C) \to \mathrm{Cl}(C) \xrightarrow{\deg} \mathbb{Z} \to 0$$

where $\mathrm{Cl}^0(C)$ is the subgroup of $\mathrm{Cl}(C)$ of divisor classes of degree 0.

Let C_d be the set of unordered d-tuples of points on C, that is,

$$C_d = C^d / S_d$$

where the symmetric group S_d acts on the Cartesian product C^d by permutation. This is again a smooth projective variety and the Abel–Jacobi map $u^d : C^d \to J(C)$ clearly factors through a map

$$u_d : C_d \to J(C)$$

The fibers of this map are of particular interest.

Theorem 2 (Abel). *Two effective divisors D_1 and D_2 on C of the same degree d are linearly equivalent if and only if $u_d(D_1) = u_d(D_2)$.*

One normally denotes the inverse image of $u_d(D)$ by

$$|D| = u_d^{-1}(u_d(D)) = \{D'; D' \geq 0, D' \sim D\}$$

Note that the latter description also makes sense if D itself is not necessarily effective. One calls $|D|$ the

complete linear system defined by the divisor D. If $\deg D < 0$, then automatically $|D| = \emptyset$, but the converse is not necessarily true. Let $\mathcal{M}_C$ be the field of meromorphic (or equivalently rational) functions on C. Then, one defines

$$L(D) = \{f \in \mathcal{M}_C; (f) \geq -D\}$$

This is a $\mathbb{C}$-vector space and it is not difficult to see that $L(D)$ has finite dimension. To every function $0 \neq f \in L(D)$, one can associate the effective divisor

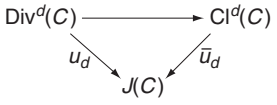
$$D_f = (f) + D \geq 0$$

Clearly, $D_f \sim D$ and every effective divisor with this property arises in this way. This gives a bijection

$$\mathbb{P}(L(D)) = |D|$$

showing that the complete linear system $|D|$ has the structure of a projective space. A linear system is a projective subspace of some complete linear system $|D|$.

Clearly, the map $u_d : C_d \to J(C)$ can be extended to the set $\mathrm{Div}^d(C)$ of degree d divisors and Abel's theorem then states that this map factors through $\mathrm{Cl}^d(C)$, that is, that we have a commutative diagram



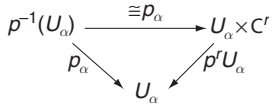
where $\overline{u}_d$ is injective.

Theorem 3 (Jacobi's Inversion Theorem). *The map u_d is surjective and hence induces an isomorphism*

$$\overline{u}_d : \mathrm{Cl}^d(C) \cong J(C)$$

It should be noted that the definition of the maps u_d depends on the choice of a base point $P_0 \in C$. Hence, the maps $\overline{u}_d$ are not canonical, with the exception of the isomorphism $\overline{u}_0 : \mathrm{Cl}^0(C) \cong J(C)$ where the choice of P_0 drops out.

The concepts of divisors and linear systems can be rephrased in the language of line bundles. A (holomorphic) vector bundle on a complex manifold M is a complex manifold E together with a projection $p : E \to M$ which is a locally trivial $\mathbb{C}^r$-bundle. This means that an open covering $(U_\alpha)_{\alpha \in A}$ of M and local trivializations



exist, such that the transition maps

$$\varphi_\beta \circ \varphi_\alpha^{-1}|_{(U_\alpha \cap U_\beta) \times \mathbb{C}^r} : (U_\alpha \cap U_\beta) \times \mathbb{C}^r \to (U_\alpha \cap U_\beta) \times \mathbb{C}^r$$

are fiberwise linear isomorphisms. If M is connected, then r is constant and is called the rank of the vector bundle. A line bundle is simply a rank-1 vector bundle.

Alternatively, one can view vector bundles as locally free $\mathcal{O}_M$-modules, where $\mathcal{O}_M$ denotes the structure sheaf of holomorphic (or in the algebro-geometric setting regular) functions on M. An $\mathcal{O}_M$-module $\mathcal{E}$ is called locally free of rank r, if an open covering $(U_\alpha)_{\alpha \in A}$ of M exists such that $\mathcal{E}|_{U_\alpha} \cong \mathcal{O}_{U_\alpha}^{\oplus r}$. The transition functions of a locally free sheaf can be used to define a vector bundle and vice versa, and hence the concepts of vector bundles and locally free sheaves can be used interchangeably. The open coverings U_α can be viewed either in the complex topology, or, if M is an algebraic variety, in the Zariski topology, thus leading to either holomorphic vector bundles (locally free sheaves in the $\mathbb{C}$-topology) or algebraic vector bundles (locally free sheaves in the Zariski topology). Clearly, every algebraic vector bundle defines a holomorphic vector bundle. Conversely, on a projective variety M, Serre's GAGA theorem (géométrie algébriques et géométrie analytique), a vast generalization of Chow's theorem, states that there exists a bijection between the equivalence classes of algebraic and holomorphic vector bundles (locally free sheaves).

The Picard group $\mathrm{Pic}\, M$ is the set of all isomorphism classes of line bundles on M. The tensor product defines a group structure on $\mathrm{Pic}\, M$ where the neutral element is the trivial line bundle $\mathcal{O}_M$ and the inverse of a line bundle $\mathcal{L}$ is its dual bundle $\mathcal{L}^*$, which is also denoted by $\mathcal{L}^{-1}$. For this reason, locally free sheaves of rank 1 are also called invertible sheaves.

We now return to the case of a compact Riemann surface (algebraic curve) C. The concept of line bundles and divisors can be translated into each other. If $D = \sum n_i P_i$ is a divisor on C and U an open set, then we denote by D_U the restriction of D to U, that is, the divisor consisting of all points $P_i \in U$ with multiplicity n_i. One then defines a locally free sheaf (line bundle) $\mathcal{L}(D)$ by

$$\mathcal{L}(D)(U) = \{f \in \mathcal{M}_C(U); (f) \geq -D_U\}$$

To see that this is locally free, it is enough to consider for each point P_i a neighborhood U_i on which a holomorphic function t_i exists, which vanishes only at P_i and there of order 1 (i.e., it is a local parameter near the point P_i). Then,

$$\mathcal{L}(D)(U_i) = t_i^{-n_i} \mathcal{O}_{U_i} \cong \mathcal{O}_{U_i}$$

This correspondence defines a map

$$\begin{aligned} \mathrm{Div}\, C &\to \mathrm{Pic}\, C \\ D &\mapsto \mathcal{L}(D) \end{aligned}$$

It is not hard to show that:

1. every line bundle $\mathcal{L} \in \operatorname{Pic} C$ is of the form $\mathcal{L} = \mathcal{L}(D)$ for some divisor D on the curve C;
2. $D_1 \sim D_2 \iff \mathcal{L}(D_1) \cong \mathcal{L}(D_2)$;
3. $\mathcal{L}(D_1) \otimes \mathcal{L}(D_2) \cong \mathcal{L}(D_1 + D_2)$; and
4. $\mathcal{L}(-D) \cong \mathcal{L}(D)^{-1}$.

Hence, there is an isomorphism of abelian groups

$$\operatorname{Cl}(C) \cong \operatorname{Pic} C$$

This correspondence allows to define the degree of a line bundle $\mathcal{L}$. In the complex analytic setting this can also be interpreted as follows. Let $\mathcal{O}_C^*$ be the sheaf of nowhere-vanishing functions. Using cocycles, one easily identifies

$$H^1(C, \mathcal{O}_C^*) \cong \operatorname{Pic} C$$

and the exponential sequence

$$0 \to \mathbb{Z} \to \mathcal{O}_C \xrightarrow{\exp} \mathcal{O}_C^* \to 0$$

induces an exact sequence

$$\begin{aligned} 0 \to H^1(C, \mathbb{Z}) \to H^1(C, \mathcal{O}_C) \\ \to H^1(C, \mathcal{O}_C^*) = \operatorname{Pic} C \to H^2(C, \mathbb{Z}) \end{aligned}$$

The last map in this exact sequence associates to each line bundle $\mathcal{L}$ its first Chern class $c_1(\mathcal{L}) \in H^2(C, \mathbb{Z}) \cong \mathbb{Z}$, which can be identified with the degree of $\mathcal{L}$. Hence, the subgroup $\operatorname{Pic}^0 C$ of degree 0 line bundles on C is isomorphic to

$$\operatorname{Pic}^0 C \cong H^1(C, \mathcal{O}_C)/H^1(C, \mathbb{Z})$$

Altogether there are identifications

$$\operatorname{Pic}^0 C \cong \operatorname{Cl}^0 C \cong J(C)$$

The Riemann–Roch Theorem and Applications

For every divisor D on a compact Riemann surface C, the discussion of the preceding section shows that there is an identification of finite-dimensional vector spaces

$$L(D) = H^0(C, \mathcal{L}(D))$$

where $H^0(C, \mathcal{L}(D))$ is the space of global sections of the line bundle $\mathcal{L}(D)$. One defines

$$l(D) = \dim_{\mathbb{C}} L(D)$$

It is a crucial question in the theory of compact Riemann surfaces to study the dimension $l(D)$ as D varies.

The canonical bundle ω_C of C is defined as the dual of the tangent bundle of C. Its global sections are holomorphic 1-forms. Every divisor K_C on C with $\omega_C = \mathcal{L}(K_C)$ is called (a) canonical divisor. The canonical divisors are the divisors of the meromorphic 1-forms on C, whereas the effective canonical divisors correspond to the divisors of holomorphic 1-forms (here, we simply write a 1-form locally as $f(z)\,dz$ and define a divisor by taking the zeros, resp. poles of $f(z)$). By abuse of notation, we also denote the divisor class corresponding to canonical divisors by K_C. There is a natural identification

$$\mathbb{P}(H^0(C, \omega_C)) = |K_C|$$

For a divisor D, the index of speciality is defined by

$$i(D) = l(K_C - D) = \dim_{\mathbb{C}} L(K_C - D)$$

The linear system $|K_C - D|$ is called the adjoint system of $|D|$. A crucial role is played by the

Theorem 4 (Riemann–Roch). *For any divisor D on a compact Riemann surface C of genus g, the equality*

$$l(D) - i(D) = \deg D + 1 - g \qquad [2]$$

holds.

This can also be written in terms of line bundles. If $\mathcal{L}$ is any line bundle, then we denote the dimension of the space of global sections by

$$h^0(\mathcal{L}) = \dim_{\mathbb{C}} H^0(C, \mathcal{L})$$

Then, the Riemann–Roch theorem can be written as

$$h^0(\mathcal{L}) - h^0(\omega_C \otimes \mathcal{L}^{-1}) = \deg \mathcal{L} + 1 - g \qquad [3]$$

This can be written yet again in a different way, if we use sheaf cohomology. By Serre duality, there is an isomorphism of cohomology groups

$$H^1(C, \mathcal{L}) \cong H^0(C, \omega_C \otimes \mathcal{L}^{-1})^*$$

and hence if we set

$$h^1(\mathcal{L}) = \dim_{\mathbb{C}} H^1(C, \mathcal{L})$$

then [3] reads

$$h^0(\mathcal{L}) - h^1(\mathcal{L}) = \deg \mathcal{L} + 1 - g \qquad [4]$$

Whereas [2] is the classical formulation of the Riemann–Roch theorem, formula [4] is the formulation which is more suitable for generalizations. From this point of view, the classical Riemann–Roch theorem is a combination of the cohomological formulation [4] together with Serre duality.

The Riemann–Roch theorem has been vastly generalized. This was first achieved by Hirzebruch who proved what is nowadays called the Hirzebruch–Riemann–Roch theorem for vector bundles on projective manifolds. A further generalization is due to Grothendieck, who proved a "relative" version involving maps between varieties. Nowadays, theorems like the Hirzebruch–Riemann–Roch theorem can be

viewed as special cases of the Atiyah–Singer index theorem for elliptic operators. The latter also contains the Gauss–Bonnet theorem from differential geometry as a special case. Moreover, Serre duality holds in much greater generality, namely for coherent sheaves on projective varieties.

Applying the Riemann–Roch theorem [3] to the zero divisor $D=0$, resp. the trivial line bundle $\mathcal{O}_C$, one obtains

$$h^0(\omega_C) = g \qquad [5]$$

that is, the number of independent global holomorphic 1-forms equals the genus of the curve C. Similarly, for $D=K_C$, resp. $\mathcal{L}=\omega_C$, we find from [3] and [5] that

$$\deg K_C = 2g-2$$

These relations show, how the Riemann–Roch theorem links analytic, resp. algebraic, invariants with the topology of the curve C.

Finally, if $\deg D > 2g-2$, then $\deg(K_C - D) < 0$ and hence $i(D) = l(K_C - D) = 0$ and [2] becomes

$$l(D) = \deg D + 1 - g \quad \text{if } \deg D > 2g-2$$

which is Riemann's original version of the theorem.

Classically, linear series arose in the study of projective embeddings of algebraic curves. For a nonzero effective divisor

$$D = \sum_{i=1}^{k} n_i P_i, \quad n_i > 0$$

the support of D is defined by

$$\operatorname{supp}(D) = \{P_1, \ldots, P_k\}$$

A complete linear system $|D|$ is called base point free, if no point P exists which is in the support of every divisor $D' \in |D|$. This is the same as saying that for every $P \in C$ a section $s \in H^0(C, \mathcal{L}(D))$ exists which does not vanish at P. Let $|D|$ be base point free and let $s_0, \ldots, s_n \in H^0(C, \mathcal{L}(D))$ be a basis of the space of sections. Then, one obtains a map

$$\varphi_{|D|} : C \to \mathbb{P}(H^0(C, \mathcal{L}(D))) = \mathbb{P}^n$$
$$P \mapsto (s_0(P) : \ldots : s_n(P))$$

The divisors $D' \in |D|$ are then exactly the pullbacks of the hyperplanes H of $\mathbb{P}^n$ under the map $\varphi_{|D|}$. Note that the map $\varphi_{|D|}$ as defined here depends on the choice of the basis $s_0, \ldots, s_n$, but any two such choices only differ by an automorphism of $\mathbb{P}^n$. We say that $|D|$, resp. the associated line bundle $\mathcal{L} = \mathcal{L}(D)$, is very ample if $\varphi_{|D|}$ defines an embedding. Using the Riemann–Roch theorem, it is not difficult to prove:

Proposition 1 *Let D be a divisor of degree d on the curve C. Then*

(i) *$|D|$ is base point free if $d \geq 2g$ and*
(ii) *$|D|$ is very ample if $d \geq 2g+1$.*

If the genus $g(C) \geq 2$, then one can prove that $|K_C|$ is base point free and consider the canonical map

$$\varphi_{|K_C|} : C \to \mathbb{P}^{g-1}$$

A curve C is called hyperelliptic if there exists a surjective map $f : C \to \mathbb{P}^1$ which is a covering of degree 2. In genus 2 every curve is hyperelliptic, whereas for genus $g \geq 3$ hyperelliptic curves are special. The connection with the canonical map is given by

Theorem 5 (Clifford). *Let C be a curve of genus $g \geq 2$. Then the canonical map is an embedding if and only if C is not hyperelliptic.*

We end this section by stating Hurwitz's theorem: Let $f : C \to D$ be a surjective holomorphic map between compact Riemann surfaces (if f is not constant then it is automatically surjective). Then, near a point $P \in C$ the map f is given in local analytic coordinates by $f(t) = t^{n_P}$ and we call f "ramified" of order n_P if $n_P > 1$. The ramification divisor of f is defined as

$$R = \sum_{P \in C} (n_P - 1)P$$

Note that this is a finite sum. If we define

$$f^*(Q) = \sum_{P \in f^{-1}(Q)} n_P P$$

then one can show that

$$\deg f = \deg f^*(Q) = \sum_{P \in f^{-1}(Q)} n_P$$

is independent of the point Q. This number is called the degree of the map f. (This should not be confused with the degree $\deg(f)$ of the principal divisor (f) defined by f.) In fact, applying the above equality to the map $f : C \to \mathbb{P}^1$ associated to a nonconstant meromorphic function f shows that the degree of the principal divisor (f) is zero, since

$$\deg(f) = \deg f^*(0) - \deg f^*(\infty) = 0$$

Theorem 6 (Hurwitz). *Let $f : C \to D$ be a surjective holomorphic map between compact Riemann surfaces of genus $g(C)$ and $g(D)$, respectively. Then,*

$$2g(C) - 2 = \deg f \cdot (2g(D) - 2) + \deg R$$

where R is the ramification divisor.

Brill–Noether Theory

In this section, we state the main results of Brill–Noether theory. For a divisor D on a curve C we denote by

$$r(D) = l(D) - 1$$

the projective dimension of the complete linear system $|D|$. The principal objects of Brill–Noether theory are the sets $W^r_d \subset \mathrm{Cl}^d(C) = \mathrm{Pic}^d(C)$ given by

$$W^r_d(C) = \{D; \deg D = d, r(D) \geq r\}$$

These sets are subvarieties of $\mathrm{Cl}^d(C) = \mathrm{Pic}^d(C)$.

We denote by g^r_d a linear system (not necessarily complete) of degree d and projective dimension r. Closely related to the varieties W^r_d are the sets

$$G^r_d(C) = \{\delta; \delta \text{ is a } g^r_d \text{ on } C\}$$

These sets also have a natural structure as a projective variety. Clearly, there are maps $G^r_d(C) \to W^r_d(C)$.

If $g = g(C)$ is the genus of the curve C, then the Brill–Noether number is defined as

$$\rho(g, r, d) = g - (r+1)(g - d + r)$$

Its significance is that it is the expected dimension of the varieties $G^r_d(C)$. The two basic results of Brill–Noether theory are:

Theorem 7 (Existence Theorem). *Let C be a curve of genus g. Let d, r be integers such that $d \geq 1$, $r \geq 0$, and $\rho(g, r, d) \geq 0$. Then $G^r_d(C)$ and hence $W^r_d(C)$ are nonempty and every component of $G^r_d(C)$ has dimension at least ρ. If $r \geq d - g$, then the same is true for $W^r_d(C)$.*

Theorem 8 (Connectedness Theorem). *Let C be a curve of genus g and d, r integers such that $d \geq 1$, $r \geq 0$, and $\rho(g, r, d) \geq 1$. Then $G^r_d(C)$ and hence also $W^r_d(C)$ are connected.*

The above theorems hold for all curves C. There are other theorems which only hold for general curves (where general means outside a countable union of proper subvarieties in the moduli space, see the section "Moduli of compact Riemann surfaces").

Theorem 9 (Dimension Theorem). *Let C be a general curve of genus g and $d \geq 1$, $r \geq 0$ integers. If $\rho(g, r, d) < 0$, then $G^r_d(C) = \emptyset$. If $\rho \geq 0$, then every component of $G^r_d(C)$ has dimension ρ.*

Theorem 10 (Smoothness Theorem). *Let C be a general curve of genus g and $d \geq 1$, $r \geq 0$. Then, $G^r_d(C)$ is smooth of dimension ρ. If $\rho \geq 1$, then $G^r_d(C)$ and hence $W^r_d(C)$ are irreducible.*

Brill–Noether theory started with a paper of Brill and Noether in 1873. It was, however, only from the 1970s onwards that the main theorems could be proved rigorously, due to the work of Griffiths, Harris, Kleiman, Mumford, and many others. For an extensive treatment of the theory, as well as a list of references, the reader is referred to Arbarello *et al.* (1985).

Green's Conjecture

In recent years, much progress was achieved in understanding the equations of canonical curves. If the curve C is not hyperelliptic, then the canonical map $\varphi_{|K_C|} : C \to \mathbb{P}^{g-1}$ defines an embedding. We shall, in this case, identify C with its image in $\mathbb{P}^{g-1}$ and call this a canonical curve. The Clifford index (for a precise definition see Lazarsfeld (1989)) is a first measure of how special a curve C is with respect to the canonical map. Hyperelliptic curves, where the canonical map fails to be an embedding, have, by definition, Clifford index 0. The two next special cases are plane quintic curves (they have a g^2_5) and trigonal curves. A curve C is called trigonal, if there is a $3:1$ map $C \to \mathbb{P}^1$, in which case C has a g^1_3. More generally, the gonality of a curve C is the minimal degree of a surjective map $C \to \mathbb{P}^1$. Plane quintics and trigonal curves are precisely the curves which have Clifford index 1.

Theorem 11 (Enriques–Babbage). *If $C \subset \mathbb{P}^{g-1}$ is a canonical curve, then C is either defined by quadratic equations, or it is trigonal or isomorphic to a plane quintic curve (i.e., it has Clifford index 1).*

One can now ask more refined questions about the equations defining canonical curves and the relations (syzygies) among these equations. This leads to looking at the minimal free resolution of a canonical curve C, which is of the form

$$0 \leftarrow \mathcal{I}_C \leftarrow \oplus_j \mathcal{O}_{\mathbb{P}^{g-1}}(-j)^{\beta_{0j}} \leftarrow \cdots \leftarrow \oplus_j \mathcal{O}_{\mathbb{P}^{g-1}}(-j)^{\beta_{kj}} \leftarrow 0$$

Here, $\mathcal{I}_C$ is the ideal sheaf of C and $\mathcal{O}_{\mathbb{P}^{g-1}}(n)$ is the nth power of the dual of the Hopf bundle (or tautological sub-bundle) on $\mathbb{P}^{g-1}$ if $n \geq 0$, resp. the $|n|$th power of the Hopf bundle if $n < 0$. The $\beta_{ij}(C)$ are called the Betti numbers of C. The Green conjecture predicts a link between the nonvanishing of certain Betti numbers and geometric properties of the canonical curve, such as the existence of multisecants. Recently, C Voisin and M Teixidor have proved the Green conjecture for general curves of given gonality (see Beauville (2003)).

Moduli of Compact Riemann Surfaces

As a set, the moduli space of compact Riemann surfaces of genus g is defined as

$$\mathcal{M}_g = \{C; C \text{ is a compact Riemann surface of genus } g\}/\cong$$

For genus $g=0$, the only Riemann surface is the Riemann sphere $\hat{\mathbb{C}} = \mathbb{P}^1$ and hence $\mathcal{M}_0$ consists of one point only. Every Riemann surface of genus 1 is a torus

$$E = \mathbb{C}/L$$

for some lattice L, which can be written in the form

$$L_\tau = \mathbb{Z}\tau + \mathbb{Z}, \quad \operatorname{Im}\tau > 0$$

Two elliptic curves $E_\tau = \mathbb{C}/L_\tau$ and $E_{\tau'} = \mathbb{C}/L_{\tau'}$ are isomorphic if and only if a matrix

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z})$$

exists with

$$\tau' = \frac{a\tau + b}{c\tau + d}$$

This proves that

$$\mathcal{M}_1 = \mathbb{H}_1/\mathrm{SL}(2,\mathbb{Z})$$

and this construction also shows that $\mathcal{M}_1$ can itself be given the structure of a Riemann surface. Using the j-function, one obtains that

$$\mathcal{M}_1 \cong \mathbb{C}$$

The situation is considerably more complicated for genus $g \geq 2$. The space of infinitesimal deformations of a curve C is given by $H^1(C, T_C)$ where T_C is the tangent bundle. By Serre duality

$$H^1(C, T_C) \cong H^0(C, \omega_C^{\otimes 2})^*$$

and by Riemann's theorem it then follows that

$$\dim H^1(C, T_C) = \dim H^0(C, \omega_C^{\otimes 2}) = 3g - 3$$

This shows that a curve of genus g depends on $3g-3$ parameters or moduli, a dimension count which was first performed by Riemann.

In genus 2 every curve has the hyperelliptic involution, and for a general curve of genus 2 this is the only automorphism. In genus $g \geq 3$ the general curve has no automorphisms, but some curves do. The order of the automorphism group is bounded by $84(g-1)$. The existence of automorphism for some curves means that $\mathcal{M}_g$ is not a manifold, but has singularities. The singularities are, however, fairly mild. Locally, $\mathcal{M}_g$ always looks like $\mathbb{C}^{3g-3}/G$ near the origin, where G is a finite group acting linearly on $\mathbb{C}^{3g-3}$. One expresses this by saying that $\mathcal{M}_g$ has only finite quotient singularities. A space with this property is also sometimes referred to as a V-manifold or an orbifold. Moreover, $\mathcal{M}_g$ is a quasiprojective variety, that is, a Zariski-open subset of a projective variety. As the above parameter count implies, the dimension of $\mathcal{M}_g$ is $3g-3$. At this point it can also be clarified what is meant by a general curve in the context of Brill–Noether theory: a property is said to hold for the general curve in Brill–Noether theory if it holds outside a countable number of proper subvarieties of $\mathcal{M}_g$.

It is often useful to work with projective, rather than quasiprojective, varieties. This means that one wants to compactify $\mathcal{M}_g$ to a projective variety $\overline{\mathcal{M}}_g$, preferably in such a way that the points one adds still correspond to geometric objects. The crucial concept in this context is that of a stable curve. A stable curve of genus g is a one-dimensional projective variety with the following properties:

1. C is connected (but not necessarily irreducible),
2. C has at most nodal singularities (i.e., two local analytic branches meet transversally),
3. the arithmetic genus $p_a(C) = h^1(C, \mathcal{O}_C) = g$, and
4. the automorphism group $\mathrm{Aut}(C)$ of C is finite.

The last of these conditions is equivalent to the following: if a component of C is an elliptic curve, then this must either meet another component or have a node, and if a component is a rational curve, then this component must either have at least two nodes or one node and intersect another component, or it is smooth and has at least three points of intersection with other components.

It should be noted that, in contrast to the previous illustrations, Figure 5 is drawn from the complex point of view, that is, the curves appear as one-dimensional objects.

The concept of stable curves leads to what is generally known as the Deligne–Mumford compactification of $\mathcal{M}_g$:

$$\overline{\mathcal{M}}_g = \{C; C \text{ is a stable curve of genus } g\}/\cong$$

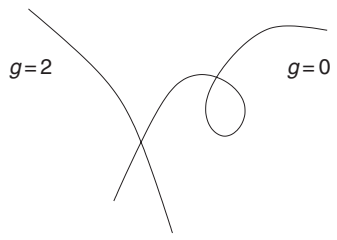


Figure 5 An example of a stable curve of genus 3.

Theorem 12 (Deligne–Mumford, Knudsen). *$\overline{\mathcal{M}}_g$ is an irreducible, projective variety of dimension $3g-3$ with only finite quotient singularities.*

The spaces $\overline{\mathcal{M}}_g$ have been studied intensively over the last 30 years. From the point of view of classification, an important question is to determine the Kodaira dimension of these spaces.

Theorem 13 (Harris–Mumford, Eisenbud–Harris). *The moduli spaces $\overline{\mathcal{M}}_g$ are of general type for $g > 23$.*

On the other hand, it is known that $\overline{\mathcal{M}}_g$ is rational for $g \leq 6$, unirational for $g \leq 14$, and has negative Kodaira dimension for $g \leq 16$.

A further topic is to understand the cohomology of $\overline{\mathcal{M}}_g$, resp. the Chow ring, and to compute the intersection theory on $\overline{\mathcal{M}}_g$. For these topics we refer the reader to Vakil (2003).

Closely related is the moduli problem of stable n-pointed curves. A stable n-pointed curve (Figure 6) is an $(n+1)$-tuple $(C, x_1, \ldots, x_n)$, where C is a connected nodal curve and $x_1, \ldots, x_n$ are smooth points of C with the stability condition that the automorphism group of $(C, x_1, \ldots, x_n)$ is finite. These curves can be parametrized by a coarse moduli space $\overline{\mathcal{M}}_{g,n}$. These spaces share many properties of the spaces $\overline{\mathcal{M}}_g$: they are irreducible, projective varieties with finite quotient singularities and of dimension $3g-3+n$.

A further development, which has become very important in recent years, is that of moduli spaces of stable maps. These were introduced by Kontsevich in the context of quantum cohomology. To define stable maps, one first fixes a projective variety X and then considers $(n+2)$-tuples $(C, x_1, \ldots, x_n, f)$ where $(C, x_1, \ldots, x_n)$ is an n-pointed curve of genus g and $f : C \to X$ a map. The stability condition is, that this object allows only finitely many automorphisms $\varphi : C \to C$, fixing the marked points $x_1, \ldots, x_n$, such that $f \circ \varphi = f$. In order to obtain meaningful moduli spaces, one also fixes a class $\gamma \in H_2(X, \mathbb{Z})$. One then asks for a space parametrizing all stable $(n+2)$-tuples $(C, x_1, \ldots, x_n, f)$ with the additional property that $f_*[C] = \gamma$. This construction is best treated in the language of stacks, and one can show that this moduli problem gives rise to a proper Deligne–Mumford stack $\overline{\mathcal{M}}_{g,n}(X, \gamma)$. In general, this stack is very complicated, it need not be connected, can be very singular, and may have several components of different dimensions. Its expected dimension is

$$\exp.\dim \overline{\mathcal{M}}_{g,n}(X,\gamma) = (\dim X - 3)(1-g) + n + \int_\gamma c_1(T_X)$$

Quantum cohomology can now be rephrased as intersection theory on the stack $\overline{\mathcal{M}}_{g,n}(X, \gamma)$. In general, these stacks do not have the expected dimension. For this reason, Behrend and Fantechi (1997) have constructed a virtual fundamental class of the right dimension, which is the correct tool for the intersection theory which gives the algebro-geometric definition of quantum cohomology. In addition to this, there is also a symplectic formulation. It was shown by B Siebert that both approaches coincide.

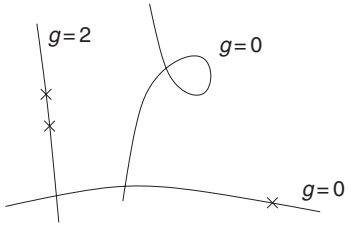


Figure 6 An example of marked stable curve.

Verlinde Formula and Conformal Blocks

The study of vector bundles (locally free sheaves) on a compact Riemann surface is an area of research in its own right. For a rank-r bundle $\mathcal{E}$, the slope of $\mathcal{E}$ is defined by

$$\mu(\mathcal{E}) = \frac{\deg \mathcal{E}}{r}$$

where the degree of $\mathcal{E}$ is defined as the degree of the line bundle $\bigwedge^r \mathcal{E} = \det \mathcal{E}$. The bundle $\mathcal{E}$ is called stable, resp. semistable, if

$$\mu(\mathcal{F}) < \mu(\mathcal{E}), \quad \text{resp.} \quad \mu(\mathcal{F}) \leq \mu(\mathcal{E})$$

for every proper sub-bundle $\{0\} \subsetneq \mathcal{F} \subsetneq \mathcal{E}$. Let C be a compact Riemann surface of genus $g \geq 2$ and let $\mathrm{SU}_C(r)$ be the moduli space of semistable rank-r vector bundles with trivial determinant $\det \mathcal{E} = \mathcal{O}_C$. This is a projective variety of dimension $(r^2-1)(g-1)$. It contains a smooth open set, whose points correspond to the isomorphism classes of stable vector bundles. The complement of this set is in general the singular locus of $\mathrm{SU}_C(r)$ and its points correspond to direct sums of line bundles of degree 0. These are the so-called graded objects of the semistable, but not stable, bundles. By a theorem of Narasimhan and Seshadri, the points of $\mathrm{SU}_C(r)$ are also in one-to-one correspondence with the isomorphism classes of representations $\pi_1(C) \to \mathrm{SU}(r)$.

Let $L \in \mathrm{Pic}^{g-1}(C)$ be any line bundle of degree $g-1$ on C. Then, the set

$$\Theta_L = \{\mathcal{E} \in \mathrm{SU}_C(r); \dim H^0(C, \mathcal{E} \otimes L) > 0\}$$

is a Cartier divisor on $SU_C(r)$ and thus defines a line bundle $\mathcal{L}$ on $SU_C(r)$. This is a natural generalization of the construction of the classical theta divisor. The line bundle $\mathcal{L}$ generates the Picard group of the moduli space $SU_C(r)$.

Theorem 14 (Verlinde Formula). *If C has genus g and k is a positive integer, then*

$$\dim H^0(\mathrm{SU}_C(r), \mathcal{L}^k) = \left(\frac{r}{r+k}\right)^g \sum_{\substack{S\sqcup T=\{1,\ldots,r+k\} \\ |S|=r}} \prod_{\substack{s\in S \\ t\in T}} \left|\sin \pi \frac{s-t}{r+k}\right|^{g-1}$$

This formula was first found by Verlinde in the context of conformal field theory. Due to this relationship, the spaces $H^0(\mathrm{SU}_C(r), \mathcal{L}^k)$ are also called conformal blocks. These spaces can also be defined for principal bundles. Rigorous proofs for the general case of the Verlinde formula are due to Beauville–Laszlo and Faltings. For a survey, see Beauville (1995).

See also: Characteristic Classes; Cohomology Theories; Index Theorems; Mirror Symmetry: a Geometric Survey; Moduli Spaces: An Introduction; Polygonal Billiards; Several Complex Variables: Basic Geometric Theory; Several Complex Variables: Compact Manifolds; Topological Gravity, Two-Dimensional.

Further Reading

Arbarello E, Cornalba M, Griffiths Ph, and Harris J (1985) *Geometry of Algebraic Curves*, vol. I. New York: Springer.

Beauville A (1995) Vector bundles on curves and generalized theta functions: recent results and open problems. In: Boutet de Montel A and Morchenko V (eds.) *Current Topics in Complex Algebraic Geometry*, (Berkeley, CA, 1992/93), Math. Sci. Res. Inst. Publ., vol. 28, pp. 17–33. Cambridge: Cambridge University Press.

Beauville A (2003) La conjecture de green générique [d'après C. Voisin]. Exposé 924 du Séminaire Bourbaki.

Behrend K and Fantechi B (1997) The intrinsic normal cone. *Inventiones Mathematicae* 128: 45–88.

Faber C and Looijenga E (1999) Remarks on moduli of curves. In: Faber C and Looijenga E (eds.) *Moduli of Curves and Abelian Varieties*, Aspects Math. E33, pp. 23–45. Braunschweig: Vieweg.

Farkas H and Kra I (1992) *Riemann Surfaces*, 2nd edn. New York: Springer.

Forster O (1991) *Lectures on Riemann Surfaces* (translated from the 1977 German Original by Bruce Gilligan), Reprint of the 1981 English Translation. New York: Springer.

Griffiths Ph and Harris J (1994) *Principles of Algebraic Geometry*, Reprint of the 1978 Edition, Wiley Classics Library. New York: Wiley.

Hartshorne R (1977) *Algebraic Geometry*. Heidelberg: Springer.

Jost J (1997) *Compact Riemann Surfaces. An Introduction to Contemporary Mathematics* (translated from the German Manuscript by Simha RR). Berlin: Springer.

Kirwan F (1992) *Complex Algebraic Curves*. Cambridge: Cambridge University Press.

Lazarsfeld R (1989) A sampling of vector bundle techniques in the study of linear series. In: Cornalba M, Gomez-Mont X, and Verjovsk (eds.) *Lectures on Riemann Surfaces*, pp. 500–559. Teaneck, NJ: World Scientific.

Miranda R (1995) *Algebraic Curves and Riemann Surfaces*. Providence: American Mathematical Society.

Mumford D (1995) *Algebraic Geometry. I. Complex Projective Varieties*, Reprint of the 1976 Edition. Berlin: Springer.

Vakil R (2003) The moduli space of curves and its tautological ring. *Notices of the American Mathematical Society* 50(6): 647–658.

Weyl H (1997) *Die Idee der Riemannschen Fläche* (Reprint of the 1913 German Original, With Essays by Reinhold Remmert, Michael Schneider, Stefan Hildebrandt, Klaus Hulek and Samuel Patterson. Edited and with a Preface and a Biography of Weyl by Remmert). Stuttgart: Teubner.

Riemann–Hilbert Methods in Integrable Systems

D Shepelsky, Institute for Low Temperature Physics and Engineering, Kharkov, Ukraine

Introduction

The Riemann–Hilbert (RH) method in mathematical physics and analysis consists in reducing a particular problem to the problem of reconstruction of an analytic, scalar- or matrix-valued function in the complex plane from a prescribed jump across a given curve. More precisely, let an oriented contour Σ be given in the complex λ-plane. The contour Σ may have points of self-intersections, and it may consist of several connected components; typical contours appearing in applications to integrable systems are shown in Figure 1.

The orientation of an arc in Σ defines the $+$ and the $-$ side of Σ. Suppose in addition that we are given a map $v: \Sigma \to GL(N, \mathbb{C})$ with $v, v^{-1} \in L^\infty(\Sigma)$. The (normalized) RH problem determined by the pair (Σ, v) consists in finding an $N \times N$

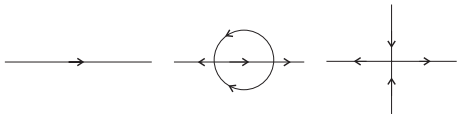


Figure 1 Typical contours for RH problems.

matrix-valued function $m(\lambda)$ with the following properties:

$$m(\lambda) \text{ is analytic in } \mathbb{C}\backslash\Sigma \qquad [1a]$$

$$m_+(\lambda) = m_-(\lambda)v(\lambda) \text{ for } \lambda \in \Sigma \text{ where } m_+(\lambda)(m_-(\lambda)) \text{ is the limit of } m \text{ from the } + (-) \text{ side of } \Sigma \qquad [1b]$$

$$m(\lambda) \to I \text{ (identity matrix) as } \lambda \to \infty \qquad [1c]$$

The precise sense in which the limit at ∞ and the boundary values $m_\pm$ are attained are technical matter that should be specified for each given RH problem (Σ, v).

Concerning the name RH problem we note that in literature (particularly, in the theory of boundary values of analytic functions), the problem of reconstructing a function from its jump across a curve is often called the Hilbert boundary-value problem. The closely related problem of analytic matrix factorization (given Σ and v, find $G(\lambda)$ analytic and nondegenerate in $\mathbb{C}\backslash\Sigma$ such that $G_+G_- = v$ on Σ) is sometimes called the Riemann problem. The name "RH problem" is also attributed to the reconstruction of a Fuchsian system with given poles and a given monodromy group.

In applications, the jump matrix v also depends on certain parameters, in which the original problem at hand is naturally formulated (e.g., $v = v(\lambda; x, t)$ in applications to the integrable nonlinear differential equations in dimension $1 + 1$, with x being the space variable and t the time variable), and the main concern is the behavior of the solution of the RH problem, $m(\lambda; x, t)$, as a function of x and t. Particular interest is in the behavior of $m(\lambda; x, t)$ as x and t become large.

In the scalar case, $N = 1$, rewriting the original multiplicative jump condition in the additive form

$$\log m_+(\lambda) = \log m_+(\lambda) + \log v(\lambda)$$

and using the Cauchy–Plemelj–Sokhotskii formula give an explicit integral representation for the solution

$$m(\lambda) = \exp\left\{\frac{1}{2\pi i}\int_\Sigma \frac{\log v(\mu)}{\mu - \lambda}\mathrm{d}\mu\right\} \qquad [2]$$

(in the case of nonzero index, $\Delta \log v|_\Sigma \neq 0$, formula [2] admits a suitable modification).

A generic (nonabelian) matrix RH problem cannot be solved explicitly in terms of contour integrals; however, it can always be reduced to a system of linear singular-integral equations, thus linearizing an originally nonlinear system.

The main benefit of reducing an originally nonlinear problem to the analytic factorization of a given matrix function arises in asymptotic analysis. Typically, the dependence of the jump matrix on the external parameters (say, x and t) is oscillatory. In analogy of asymptotic evaluation of oscillatory contour integrals via the classical method of steepest descent, in the asymptotic evaluation of the solution $m(\lambda; x, t)$ of the matrix RH problem as $x, t \to \infty$, the nonlinear steepest-descent method examines the analytic structure of the jump matrix $v(\lambda; x, t)$ in order to deform the contour Σ to contours where the oscillatory factors become exponentially small as $x, t \to \infty$, and hence the original RH problem reduces to a collection of local RH problems associated with the relevant points of stationary phase. Although the method has (in the matrix case) noncommutative and nonlinear elements, the final result of the analysis is as efficient as the asymptotic evaluation of the oscillatory integrals.

Dressing Method

The RH method allows describing the solution of a differential system independently of the theory of differential equations. The solution might be explicit, that is, given in terms of elementary or elliptic or abelian functions and contour integrals of such functions. In general (transcendental) case, the solution can be represented in terms of the solution of certain linear singular integral equations.

In the modern theory of integrable systems, a system of nonlinear differential equations is often called integrable if it can be represented as a compatibility condition of an auxiliary overdetermined linear system of differential equations called a Lax pair of the given nonlinear system (actually it might involve more than two linear equations). In order that the compatibility condition represents a nontrivial nonlinear system of equations, the Lax pair is required to depend rationally on an auxiliary parameter (called a spectral parameter). The RH problem formulated in the complex plane of the spectral parameter allows, given a particular solution of the compatibility equations, to construct directly new solutions of the compatibility system by "dressing" the initial one.

For example, let $D(x, \lambda), x \in \mathbb{R}^n, \lambda \in \mathbb{C}$ be an $N \times N$ diagonal, polynomial in λ with smooth coefficients, function such that $a_j := \partial D/\partial x_j$ are polynomials in λ of degree d_j. Then $\Psi_0 := \exp D(x, \lambda)$ solves the system of linear equations $\partial\Psi_0/\partial x_j = a_j\Psi_0$, whose compatibility conditions $\partial^2\Psi_0/\partial x_j\partial x_k = \partial^2\Psi_0/\partial x_k\partial x_j$ are trivially satisfied. Given a contour Σ and a smooth function v, consider the matrix RH problem [1]

with the jump matrix $\tilde{v}(\lambda;x) := \exp D(x,\lambda) v(\lambda) \exp -D(x,\lambda)$. Let $m(\lambda;x)$ be the solution of this RH problem. Then $(D_j m)_+ = (D_j m)_- \tilde{v}$, where $D_j f := \partial f/\partial x_j + [a_j, f]$ with $[a,b] := ab - ba$. The Liouville theorem implies that $(D_j m)m^{-1}$ is an entire function which is $o(\lambda^{d_j})$ as $\lambda \to \infty$. Setting $\Psi(x,\lambda) := m(\lambda;x) \exp D(x,\lambda)$ gives the system of linear equations

$$\frac{\partial \Psi}{\partial x_j} = a_j + \sum_{k<d_j} \lambda^k q_{jk}(x) \equiv R_j(x,\lambda)\Psi \qquad [3]$$

the compatibility conditions for which are

$$\frac{\partial R_k}{\partial x_j} - \frac{\partial R_j}{\partial x_k} = [R_j, R_k] \qquad [4]$$

Equating coefficients of various powers of λ in [4] gives a (generally) nonlinear system of partial differential equations for the coefficient matrices q_{jk}. Thus, given $D(x,\lambda)$, the RH problem, if it is solvable, maps the pair (Σ, v) to solutions of [4].

Specializing to $n = 2$ with variables $(x,t) \in \mathbb{R}^2$, the overdetermined system of linear equations and the corresponding compatibility conditions are

$$\Psi_x = U\Psi, \qquad \Psi_t = V\Psi \qquad [5]$$

and

$$U_t - V_x + [U,V] = 0 \qquad [6]$$

respectively. Conditions [6] are sometimes called the zero-curvature conditions.

Equations [5] and [6] with U and V depending rationally on the spectral parameter λ represents the integrable nonlinear systems in $1+1$ dimension. A typical example of such a system is the (defocusing) nonlinear Schrödinger (NLS) equation

$$iq_t + q_{xx} - 2|q|^2 q = 0 \qquad [7]$$

Starting from the RH problem with the 2×2 jump matrix

$$v(\lambda;x,t) = e^{i\theta\sigma_3/2} v(\lambda) e^{-i\theta\sigma_3/2} \qquad [8]$$

where $\theta(\lambda;x,t) = -t\lambda^2 + x\lambda$, $\sigma_3 = \mathrm{diag}\{1,-1\}$, and $v(\lambda)$ satisfies the involution $\sigma_3 v^*(\lambda)\sigma_3 = v(\bar{\lambda})$, expanding out the limit of the solution of the RH problem as $\lambda \to \infty$

$$m(\lambda;x,t) = I + \frac{m_1(x,t)}{\lambda} + o\left(\frac{1}{\lambda}\right) \qquad [9]$$

and arguing as above gives [5], with

$$U = \frac{i\lambda\sigma_3}{2} + \begin{pmatrix} 0 & q \\ \bar{q} & 0 \end{pmatrix} \qquad [10]$$

and $q = -i(m_1)_{12}$, whereas the compatibility condition [6] reduces to [7].

The relation between the RH problem and the differential equations [5] is local in x and t; it is based only on the unique solvability of the RH problem, the Liouville theorem, and the explicit dependence of the jump matrix in x and t. The uniqueness of the solution of an RH problem is basically provided by the Liouville theorem: the ratio $m^{(1)}(m^{(2)})^{-1}$ of any two solutions is analytic in $\mathbb{C}\backslash\Sigma$ and continuous across Σ and is therefore identically equal to I by the normalization condition [1c].

On the other hand, there are no completely general effective criteria for the solvability. Nevertheless, many RH problems seen in applications to integrable systems satisfy the following sufficient condition: if Σ is symmetric with respect to $\mathbb{R}$ and contains $\mathbb{R}$, and if, in addition, $v^*(\lambda) = v(\bar{\lambda})$ for $\lambda \in \Sigma\backslash\mathbb{R}$ and $\mathrm{Re}\, v(\lambda) > 0$ for $\lambda \in \mathbb{R}$, then the RH problem is solvable.

For nonlinear equations supporting solitons, the RH problem appears naturally in a more general setting, as a meromorphic factorization problem, where m in [1] is sought to be a (piecewise) meromorphic function, with additionally prescribed poles and respective residue conditions. Alternatively, in the Riemann factorization problem $G_+ G_- = v$, one assumes that G degenerates at some given points $\lambda_1, \ldots, \lambda_n \in \Omega^+$ and $\mu_1, \ldots, \mu_n \in \Omega^-$, where $\mathbb{C} = \Omega^+ \cup \Omega^- \cup \Sigma$, and prescribes two sets of subspaces, $\mathrm{Im}\, G|_{\lambda=\lambda_j}$ and $\mathrm{Ker}\, G|_{\lambda=\mu_j}$. In the case $v \equiv I$, the solution of the factorization problem with zeros (meromorphic RH problem) is purely algebraic, and gives formulas describing multisoliton solutions. In the general case, $v \not\equiv I$, the meromorphic RH problem can be algebraically converted to a holomorphic RH problem, by subsequently removing the poles with the help of the Blaschke–Potapov factors.

Alternatively, a meromorphic RH problem can be converted to a holomorphic one by adding to Σ an additional contour Σ_{aux} enclosing all the poles, interpolating the constants involved in the residue conditions inside the region surrounded by Σ_{aux}, and defining a new jump matrix on Σ_{aux} using the interpolant and the Blaschke–Potapov factors.

RH problems formulated on the complex plane $\mathbb{C}$ correspond typically to solutions of relevant nonlinear problems decaying at infinity. For other types of boundary conditions (e.g., nonzero constants or periodic or quasiperiodic boundary conditions), the corresponding RH problem is naturally formulated on a Riemann surface. For example, the RH problem associated with finite density conditions $q(x,t) \to \rho e^{i\phi_\pm}$ as $x \to \pm\infty$ for the NLS equation [7] is naturally formulated on the two-sheet Riemann surface of the function $k(\lambda) = \sqrt{\lambda^2 - 4\rho^2}$ with

the contour Σ consisting of the points (λ, ε), where $|\lambda| \geq 2\rho$ and $\varepsilon = \pm 1$ marks the surface sheet.

Inverse-Scattering Transform

The inverse-scattering transform method for solving initial-value problems for integrable nonlinear equations written as the compatibility conditions [6] for linear equations [5] consists in the following: starting from the given initial data, solve the direct problem, that is, determine appropriate eigenfunctions (solutions of the differential x-equation in the Lax pair [5]) having well-controlled analytic properties as functions of the auxiliary (spectral) parameter λ and the associated spectral functions of λ; then, by virtue of the t-equations in the Lax pair [5], the associated functions evolve in a simple, explicit way. Finally, using the explicit evolution of the spectral functions, solve the inverse problem of finding the associated coefficients in the x-equation, which, by [5], evolve according to the given nonlinear equation and thus solve the Cauchy problem for this equation. The last step in this procedure, the inverse-scattering problem, can be effectively solved by reformulating it as an RH problem, which in turn can be related to a system of singular integral equations. The classical Gelfand–Levitan–Marchenko integral equation of the inverse-scattering problem is the Fourier transform of some special cases of these singular integral equations.

To fix ideas, consider the initial-value problem for the NLS equation [7], where the data $q(x, t=0) = q_0(x)$ have sufficient smooth and decay as $|x| \to \infty$. For each $\lambda \in \mathbb{C}\backslash\mathbb{R}$, one constructs solutions $\Psi(x, \lambda)$ of $\Psi_x = U\Psi$ with U given by [10], having the properties

$$m(x, \lambda) := \Psi(x, \lambda) \exp\left(\frac{-\mathrm{i}x\lambda\sigma_3}{2}\right) \to I \quad \text{as } x \to -\infty$$

and $m(x, \lambda)$ is bounded as $x \to \infty$. For each fixed x, the 2×2 matrix function $m(x, \lambda)$ solves the RH problem in λ, where $\Sigma = \mathbb{R}$ and the jump matrix is

$$v = v(\lambda; x) = \begin{pmatrix} 1 - |r(\lambda)|^2 & r(\lambda)\,\mathrm{e}^{\mathrm{i}\lambda x} \\ -\bar{r}(\lambda)\,\mathrm{e}^{-\mathrm{i}\lambda x} & 1 \end{pmatrix} \qquad [11]$$

Here $r(\lambda)$ is the reflection coefficient of $q_0(x)$.

The direct scattering map $\mathcal{R}$ is described by mapping $q \mapsto r$,

$$q \mapsto m(x, \lambda) = m(x, \lambda; q) \mapsto v(\lambda; x) \mapsto r = \mathcal{R}(q)$$

By virtue of the t-equations in [5], if $q(t) = q(x, t)$ solves the NLS equation, then $r(t) = \mathcal{R}(q(\,\cdot\,, t))$ evolves as $r(t) = r(t, \lambda) = \mathrm{e}^{-\mathrm{i}t\lambda^2} r_0(\lambda)$, where $r_0 = \mathcal{R}(q_0)$. Given r, the inverse-scattering map $\mathcal{R}^{-1}$ is obtained by solving the normalized RH problem (RHP) with the jump matrix [11] and evaluating its solution $m(x, \lambda)$ as $\lambda \to \infty$ [9]:

$$\begin{aligned} r \mapsto v \mapsto \mathrm{RHP} &\mapsto m(x, \lambda) \\ &= m(x, \lambda; r) \mapsto m_1(x) \mapsto q(x) \\ &= -\mathrm{i}(m_1(x))_{12} \end{aligned}$$

and thus

$$q(x, t) = \mathcal{R}^{-1}\left(\mathrm{e}^{\mathrm{i}x(\cdot) - \mathrm{i}t(\cdot)^2} r(\cdot)\right) \qquad [12]$$

The mathematical rigor to this scheme is provided by the general theory of analytic matrix factorization making use of the relation between the factorization problem and certain singular integral equations; this relation can be established with the help of the Cauchy operators

$$Ch(\lambda) = \int_\Sigma \frac{h(\mu)}{\mu - \lambda} \frac{\mathrm{d}\mu}{2\pi\mathrm{i}}, \quad \lambda \in \mathbb{C}\backslash\Sigma$$

and

$$C^\pm h(\lambda) = \lim_{\substack{\lambda' \to \lambda \\ \lambda' \in (\pm)\text{-side of } \Sigma}} (Ch)(\lambda')$$

For a very general class of contours, the Cauchy operators $C^\pm : L^p \to L^p, 1 < p < \infty$, are bounded, $C^+ - C^- = I$, and $C^+ + C^- = -H$, where

$$Hh(\lambda) := \lim_{\varepsilon \to 0} \int_{\substack{\Sigma \\ |\mu - \lambda| > \varepsilon}} \frac{h(\mu)}{\lambda - \mu} \frac{\mathrm{d}\mu}{\pi\mathrm{i}}$$

is the Hilbert transform.

The map $\mathcal{R}$ is often considered as a nonlinear Fourier-type map; this point of view is supported by the fact that $\mathcal{R}$ is a bijection between the corresponding Schwartz spaces of functions. Making use of the L^p or Hölder theory of the Cauchy operators and the related factorization problems, it is possible to analyze the action of $\mathcal{R}$ and $\mathcal{R}^{-1}$ in various functional spaces. This also requires making more precise the definition of the RH problem: for fixed $1 < p < \infty$, given Σ and v such that $v, v^{-1} \in L^\infty(\Sigma \to GL(N, \mathbb{C}))$, we say that $m_\pm$ solves an RH L^p-problem if $m_\pm \in I + \partial C(L^p)$ and $m_+(\lambda) = m_-(\lambda)v(\lambda)$ for $\lambda \in \Sigma$. Here a pair of $L^p(\Sigma)$-functions $f_\pm \in \partial C(L^p)$ if there exists a unique function $h \in L^p(\Sigma)$ such that $f_\pm(\lambda) = (C^\pm h)(\lambda)$. Then $f(\lambda) = Ch(\lambda)$, $\lambda \in \mathbb{C}\backslash\Sigma$, is called the extension of $f_\pm$ off Σ.

Given a factorization of $v = (v^-)^{-1}v^+ = (I - w^-)^{-1}(I + w^+)$ on Σ with $v^\pm, (v^\pm)^{-1} \in L^p$, the basic associated singular integral operator is defined by

$$C_w h := C^+(hw^-) + C^-(hw^+)$$

If the operator $I - C_w$ is invertible on $L^p(\Sigma)$, with $\mu \in I + L^p(\Sigma)$, solving $(I - C_w)m = I$, then $m(\lambda) = I + (C(\mu(w_+ + w^-)))(\lambda)$ is the unique solution of the

RH problem (Σ, v). Although the operator C_w need not be compact, in many cases it is Fredholm with zero index. Then the existence of $(I - C_w)^{-1}$ is equivalent to the solvability of the RH problem (Σ, v), and the normalized RH problem ($m \to I$ as $\lambda \to \infty$) has a unique solution if and only if the corresponding homogeneous RH problem (with $m \to 0$ as $\lambda \to \infty$) has only the trivial solution (vanishing lemma).

The most complete theory for RH problem relative to simple contours is the theory when v is in an inverse, closed, decomposing Banach algebra $\mathcal{A}$, that is, the algebra of continuous functions with the Hilbert transform bounded in it such that if $f \in \mathcal{A}$, then $f^{-1} \in \mathcal{A}$. For contours with self-intersections, the RH factorization theory is formulated in terms of a pair of decomposing algebras: choosing the orientation of the contour in such a way that it divides the λ-plane into two disjoint regions, Ω^+ and Ω^-, and each arc of Σ forms part of the positively oriented boundary of Ω^+, the functions in the $+(-)$ algebra are continuous up to the boundary in each connected component of Ω^+ (Ω^-).

The choice of functional spaces in the RH problem should be based on the integrable system at hand. For example, an integrable flow connected to the scattering problem for $\Psi_x = U\Psi$, with U defined by [10], has in general the form $e^{it\lambda^p\sigma_3} v(\lambda) e^{-it\lambda^p\sigma_3}$ (Ablowitz–Kaup–Newell–Segur (AKNS) hierarchy) in the scattering space (for the NLS equation, $p = 2$), so that appropriate spaces are $L^2((1+x^2)\,dx) \cap H^{p-1}$ for $q(\cdot, t)$ and $L^2((1+|\lambda|^{2p-2})|d\lambda|) \cap H^1$ as the scattering space. Deift and Zhou showed that in this case the scattering map $\mathcal{R}$ and the inverse-scattering map $\mathcal{R}^{-1}$ indeed involve no "loss" of smoothness or decay.

A generalization of the inverse-scattering transform method to the initial boundary-value problems for integrable nonlinear equations (on the half-line or on a finite interval with respect to the space variable x) can be also developed on the basis of the RH problem formalism. It this case, the construction of the corresponding RH problem involves simultaneous spectral analysis of the both linear equations in the Lax pair [5]. The boundary values generate an additional set of spectral functions, which generally makes the construction of the associated RH problem more complicated than in the case of the corresponding initial-value problem (particularly, the contour is to be enhanced by adding the part coming from the spectral analysis of the t-equation); however, this RH problem again depends explicitly on x and t, which makes it possible to develop relevant techniques (such as the nonlinear steepest-descent method for the asymptotic analysis) in the same spirit as in the case of initial-value problems.

An RH problem may be viewed as a special case in a more general setting of problems of reconstructing an analytic function from the known structure of its singularities. The departure from analyticity of a function m of the complex variable λ can be described in terms of the "d-bar" derivative, $\partial m/\partial\bar{\lambda}$. If $\partial m/\partial\bar{\lambda}$ can be linearly related to m itself, then the use of the extension of Cauchy's formula

$$m(\lambda) = \frac{1}{2\pi i}\int_D d\mu \wedge d\bar{\mu}\frac{1}{\mu-\lambda}\frac{\partial m}{\partial\bar{\mu}} + \frac{1}{2\pi i}\int_{\partial D} d\mu \frac{m(\mu)}{\mu-\lambda}$$

leads to a linear integral equation for m. This is the case for some multidimensional $(2+1)$ nonlinear integrable equations. For example, for the Kadomtsev–Petviashvili-I equation (the two-dimensional generalization of the Korteweg–de Vries equation) $(q_t + 6qq_x + q_{xxx})_x = 3q_{yy}$, the appropriate eigenfunctions are still sectionally meromorphic, but their jumps across a contour are connected nonlocally to m on the contour, which leads to nonlocal RH problem of the type

$$m_+(\lambda) = m_-(\lambda) + \int_\Sigma d\mu m_-(\mu) f(\mu, \lambda), \quad \lambda \in \Sigma$$

with given $f(\mu, \lambda)$ (analogue of scattering data). Contrarily, the eigenfunctions for the Kadomtsev–Petviashvili-II equation $(q_t + 6qq_x + q_{xxx})_x = -3q_{yy}$ are nowhere analytic, with $\partial m/\partial\bar{\lambda}$ related to m by

$$\frac{\partial m}{\partial\bar{\lambda}}(\lambda) = F(\mathrm{Re}\,\lambda, \mathrm{Im}\,\lambda) m(-\bar{\lambda}), \quad \lambda \in \mathbb{C}$$

Nonlinear Steepest-Descent Method

The nonlinear steepest-descent method is based on a direct asymptotic analysis of the relevant RH problem; it is general and algorithmic in the sense that it does not require *a priori* information (anzatz) about the form of the solution of the asymptotic problem. However, the noncommutativity of the matrix setting requires developing rather sophisticated technical ideas, which, in particular, enable an explicit solution of the associated local RH problems.

To fix ideas, let us again consider the NLS equation. The dependence of the jump matrix $v(\lambda; x, t)$ on x and t is oscillatory; it is the same as in the integral

$$q(x,t) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{i(x\lambda - t\lambda^2)}\hat{q}_0(\lambda) d\lambda \qquad [13]$$

which solves the initial-value problem for the linearized version of [7]:

$$iq_t + q_{xx} = 0, \quad q(x,0) = q_0(x) \qquad [14]$$

(here $\hat{q}_0(\lambda)$ is the Fourier transform of the initial data q_0). The main contribution to [13] as $|x|$ and t tend to ∞ comes from the point of stationary phase of $e^{i(x\lambda - t\lambda^2)}$, that is, the point $\lambda = \lambda_0 = x/2t$, for which

$$\frac{d}{d\lambda}(x\lambda - t\lambda^2) = 0$$

If $\hat{q}_0(\lambda)$ is analytic in a strip $|\mathrm{Im}\,\lambda| < \varepsilon$, then one can use Cauchy's theorem to deform [13] to an integral on a contour Σ_ε such that $|e^{i(x\lambda - t\lambda^2)}|$ decreases rapidly on Σ_ε away from $\lambda = \lambda_0$. Hence, as $t \to \infty$, the problem localizes to a neighborhood of $\lambda = \lambda_0$; this constitutes the standard method of steepest descent.

In the spirit of the oscillatory contour integral case, the nonlinear steepest-descent method for an oscillatory RH problem introduced by Deift and Zhou consists in the following: deform the contour and (rationally) approximate the jump matrix in order to obtain an RH problem with a jump matrix that decays to the identity away from stationary phase points; then, rescaling the problem near the stationary phase points, obtain a (local) RH problem with a piecewise constant jump matrix, which can be solved in closed form, usually in terms of certain special functions.

The contour deformation means the following. Suppose that the jump matrix of an RH problem (Σ, v) has a factorization $v = b_-^{-1} v_1 b_+$ between two points on Σ, where $b_+(b_-)$ has holomorphic and nondegenerating continuation to the part $\Omega^+(\Omega^-)$ of a disk Ω supported by these points, see **Figure 2a**. Then the contour Σ may be deformed to the contour $\Sigma' = \Sigma \cup \partial\Omega$, and the jump matrices across Σ' may be defined as indicated in **Figure 2b**. If m solves the RH problem (Σ, v), then m' defined by $m' = m b_\pm^{-1}$ in $\Omega^\pm$ and $m' = m$ outside Ω solves the deformed RH problem associated with Σ'.

The appropriate factorization of v given by [8] and the contour deformation are to be chosen in accordance with signature table; for the NLS equation, it is given in **Figure 3**. The key step is to move algebraically the factors $e^{\pm i\theta}$ in $v(\lambda; x, t)$ into regions of the complex plane, where they are exponentially decreasing as $t \to \infty$. The jump matrix admits two algebraic factorizations:

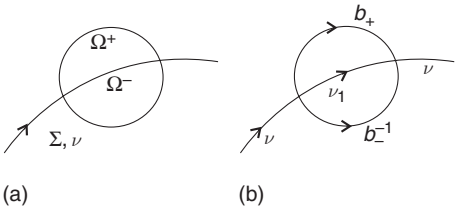


Figure 2 Deformation of an RH problem.

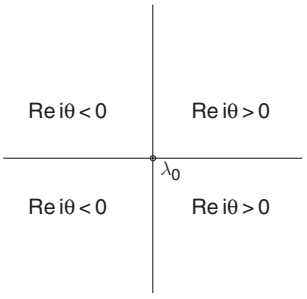


Figure 3 Signature table.

$$\begin{aligned} v &= \begin{pmatrix} 1-|r|^2 & re^{i\theta} \\ -\bar{r}e^{-i\theta} & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & re^{i\theta} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\bar{r}e^{-i\theta} & 1 \end{pmatrix} \quad (\lambda > \lambda_0) \\ &= \begin{pmatrix} 1 & 0 \\ -\dfrac{\bar{r}e^{-i\theta}}{1-|r|^2} & 1 \end{pmatrix} \begin{pmatrix} 1-|r|^2 & 0 \\ 0 & \dfrac{1}{1-|r|^2} \end{pmatrix} \\ &\quad \times \begin{pmatrix} 1 & \dfrac{re^{i\theta}}{1-|r|^2} \\ 0 & 1 \end{pmatrix} \quad (\lambda < \lambda_0) \end{aligned}$$

The diagonal factors $(1-|r|^2)^{\pm 1}$ can be removed by conjugating v by $\delta_\pm^{\sigma_3}$, where $\delta(\lambda)$ solves the scalar, normalized RH problem on $\mathbb{R}: \delta_+ = \delta_-(1-|r|^2)$ for $\lambda < \lambda_0$ and $\delta_+ = \delta_-$ for $\lambda > \lambda_0$; the solution of the latter can be written in a closed form:

$$\delta(\lambda) = \exp\left\{ \frac{1}{2\pi i} \int_{-\infty}^{\lambda_0} \frac{\log(1-|r(\mu)|^2)}{\mu - \lambda}\, d\mu \right\}$$

Then $\tilde{m} := m\delta^{-\sigma_3}$ solves the RH problem across $\Sigma = \mathbb{R}$, with the jump matrix

$$\begin{aligned} \tilde{v} &= \begin{pmatrix} 1 & r\delta^2 e^{i\theta} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\bar{r}\delta^{-2} e^{-i\theta} & 1 \end{pmatrix} \quad (\lambda > \lambda_0) \\ &= \begin{pmatrix} 1 & 0 \\ -\dfrac{\bar{r}\delta_-^{-2} e^{-i\theta}}{1-|r|^2} & 1 \end{pmatrix} \begin{pmatrix} 1 & \dfrac{r\delta_+^2 e^{i\theta}}{1-|r|^2} \\ 0 & 1 \end{pmatrix} \quad (\lambda < \lambda_0) \end{aligned}$$

Replacing $r, \bar{r}$, etc., by appropriate rational approximations $[r], [\bar{r}]$, matching at $\lambda = \lambda_0$,

$$\tilde{m}_+ \begin{pmatrix} 1 & 0 \\ -[\bar{r}]\delta^{-2} e^{-i\theta} & 1 \end{pmatrix}$$

can be continued to the sector above $\mathbb{R}_+ + \lambda_0$ and

$$\tilde{m}_- \begin{pmatrix} 1 & [r]\delta^2 e^{i\theta} \\ 0 & 1 \end{pmatrix}$$

can be continued to the sector below $\mathbb{R}_+ + \lambda_0$, where the factors $e^{\pm i\theta}$ are exponentially decreasing. Doing the same for the appropriate factors on $\mathbb{R}_- + \lambda_0$, we obtain an RH problem on a cross, say, $(\lambda_0 + e^{i\pi/4}\mathbb{R}) \cup (\lambda_0 + e^{-i\pi/4}\mathbb{R})$. As $t \to \infty$, the RH problem then localizes at λ_0.

Performing an appropriate scaling, a straightforward computation shows that, as $t \to \infty$, the problem reduces to an RH problem with the jump matrix that does not depend on λ (it is determined by $r(\lambda_0)$), which make it possible to solve this problem explicitly (in terms of the parabolic cylinder functions, in the case of the NLS equation). Using explicit asymptotics for these functions and controlling the error terms, it is possible to obtain the uniform (for all $x \in \mathbb{R}$) asymptotics for the solution of the initial-value problem for the NLS equation with $q_0 \in L^2((1+x^2)\,dx) \cap H^1$ of the form

$$q(x,t) = t^{-1/2}\alpha(\lambda_0)\exp(ix^2/(4t) - i\nu(\lambda_0)\log 2t) + O(t^{-(1/2+\kappa)})$$

for any fixed $0 < \kappa < 1/4$, where α and ν are given in terms of $r = \mathcal{R}(q_0)$:

$$\nu(\lambda) = -\frac{1}{2\pi}\log(1 - |r(\lambda)|^2)$$
$$|\alpha(\lambda)|^2 = \frac{\nu(\lambda)}{2}$$

and

$$\arg\alpha(\lambda) = \frac{1}{\pi}\int_{-\infty}^{\lambda} \log(\lambda - \mu)\,d(\log(1 - |r(\mu)|^2)) + \frac{\pi}{4} + \arg\Gamma(i\nu(\lambda)) + \arg r(\lambda)$$

The method can be used to obtain asymptotic expansions to all orders. Also, for nonlinear equations supporting solitons, the soliton part of the asymptotics can be incorporated via the dressing method.

Further applications include long-time asymptotics for near-integrable systems, such as the perturbed NLS equation $iq_t + q_{xx} - 2|q|^2 q - \varepsilon|q|^l q = 0$ for $l > 2$ and $\varepsilon > 0$, and the small-dispersion limits of integrable equations (e.g., for the Korteweg–de Vries equation $q_t - 6qq_x + \varepsilon^2 q_{xxx} = 0$ with small dispersion $\varepsilon \searrow 0$).

The RH formalism makes possible a comprehensive global asymptotic analysis of the Painlevé transcendents (which, due to their increasing role in the modern mathematical physics, should be considered as new nonlinear special functions), including explicit connection formulas, as x approaches relevant critical points along different directions in the complex plane.

The development of the RH method in the theory of integrable systems caused emerging new analytic and algebraic ideas for other branches of mathematics and theoretical physics. The recent examples are the study of the asymptotics in the theory of orthogonal polynomials and random matrices and in combinatories (random permutations).

See also: Boundary-Value Problems for Integrable Equations; $\bar{\delta}$ Approach to Integrable Systems; Integrable Systems and Algebraic Geometry; Integrable Systems and the Inverse Scattering Method; Integrable Systems: Overview; Nonlinear Schrödinger Equations; Painlevé Equations; Twistor Theory: Some Applications [in Integrable Systems, Complex Geometry and String Theory]; Riemann–Hilbert Problem.

Further Reading

Ablowitz MJ and Clarkson PA (1991) Solitons, Nonlinear Evolution Equations and Inverse Scatting, *London Math. Soc.*, Lecture Notes Series, vol. 149. Cambridge: Cambridge University Press.

Beals R, Deift PA, and Tomei C (1988) Direct and Inverse Scattering on the Line. *Mathematical Surveys and Monographs 28*. Providence, RI: American Mathematical Society.

Belokolos ED, Bobenko AI, Enol'skii VZ, and Its AR (1994) Algebro-Geometric Approach to Nonlinear Integrable Equations. *Springer Series in Nonlinear Dynamics*. Berlin: Springer.

Deift PA (1999) Orthogonal Polynomials and Random Matrices: A Riemann–Hilbert Approach. *Courant Lecture Notes in Mathematics*, vol. 3. New York: CIMS.

Deift PA and Zhou X (2003) Long-time asymptotics for solutions of the NLS equation with initial data in a weighted Sobolev space. *Communications on Pure and Applied Mathematics* 56(8): 1029–1077.

Deift PA, Its AR, and Zhou X (1993) Long-time asymptotics for integrable nonlinear wave equations. In: Fokas AS and Zakharov VE (eds.) *Important Developments in Soliton Theory*, pp. 181–204. Berlin: Springer.

Faddeev LD and Takhtajan LA (1987) *Hamiltonian Methods in the Theory of Solitons*. Berlin: Springer.

Fokas AS (2000) On the integrability of linear and nonlinear partial differential equations. *Journal of Mathematical Physics* 41: 4188–4237.

Its AR (2003) The Riemann–Hilbert problem and integrable systems. *Notices of the AMS* 50(11): 1389–1400.

Novikov SP, Manakov SV, Pitaevskii LP, and Zakharov VE (1984) *Theory of Solitons. The Inverse Scattering Method*. New York: Consultants Bureau.

Zhou X (1989) The Riemann–Hilbert problem and inverse scattering. *Journal on Mathematical Analysis. Society for Industrial and Applied Mathematics (SIAM)* 20: 966–986.

Riemann–Hilbert Problem

V P Kostov, Université de Nice Sophia Antipolis, Nice, France

Regular and Fuchsian Linear Systems on the Riemann Sphere

Consider a system of ordinary linear differential equations with time belonging to the Riemann sphere $CP^1 = \mathbf{C} \cup \infty$:

$$\mathrm{d}X/\mathrm{d}t = A(t)X \qquad [1]$$

The $n \times n$ matrix A is meromorphic on CP^1, with poles at $a_1, \ldots, a_{p+1}$; the dependent variables X form an $n \times n$ matrix. One can assume that ∞ is not among the poles a_j and it is not a pole of the 1-form $A(t)\mathrm{d}t$ (this can be achieved by a fractionally-linear transformation of t).

P Deligne has introduced a terminology of meromorphic connections and sections which is often preferred in modern literature to the one of meromorphic linear systems and their solutions, and there is a one-to-one correspondence between the two languages.

Definition 1 System [1] is regular at the pole a_j if its solutions have a moderate (or polynomial) growth rate there, that is, for every sector S centered at a_j and not containing other poles of the system and for every solution X restricted to S there exists $N_j \in \mathbf{R}$ such that $\|X(t-a_j)\| = O(|t-a_j|^{N_j})$ for all $t \in S$. System [1] is regular if it is regular at all poles a_j. System [1] is Fuchsian if its poles are logarithmic (i.e., of first order). Every Fuchsian system is regular.

Remark 2 The opening of the sector S might be $> 2\pi$. Restricting to a sector is necessary because the solutions are, in general, ramified at the poles a_j and by turning around the poles much faster than approaching them one can obtain any growth rate.

A Fuchsian system can be presented in the form

$$\mathrm{d}X/\mathrm{d}t = \left(\sum_{j=1}^{p+1} A_j/(t-a_j)\right)X, \quad A_j \in \mathrm{gl}(n,\mathbf{C}) \qquad [2]$$

The sum of its *matrices-residua* A_j is 0, that is,

$$A_1 + \cdots + A_{p+1} = 0 \qquad [3]$$

(recall that ∞ is not a pole of the system).

Remark 3 The linear equation (with meromorphic coefficients) $\sum_{j=0}^{n} a_j(t)x^{(j)} = 0$ is Fuchsian if a_j has poles of order only $\leq n-j$. A linear equation is Fuchsian if and only if it is regular. The best-studied Fuchsian equations are the hypergeometric one and its generalizations and the Jordan–Pochhammer equation.

The linear change of the dependent variables

$$X \mapsto W(t)X \qquad [4]$$

(where W is meromorphic on CP^1) makes system [2] undergo the gauge transformation

$$A \to -W^{-1}(\mathrm{d}W/\mathrm{d}t) + W^{-1}AW \qquad [5]$$

(Most often one requires W to be holomorphic and holomorphically invertible for $t \neq a_j, j = 1, \ldots, p+1$, so that no new singular points appear in the system.) This transformation preserves regularity but not necessarily being Fuchsian. The only invariant under the group of linear transformations [4] is the monodromy group of the system.

Definition 4 Set $\Sigma = CP^1 \setminus \{a_1, \ldots, a_{p+1}\}$. Fix a base point $a_0 \in \Sigma$ and a matrix $B \in \mathrm{GL}(n,\mathbf{C})$. Consider a closed contour γ with base point a_0 and bypassing the poles of the system. The monodromy operator of system [1] defined by this contour is the linear operator M acting on the solution space of the system which maps the solution X with $X|_{t=a_0} = B$ into the value of its analytic continuation along γ. Notation: $X \overset{\gamma}{\mapsto} XM$. The monodromy operator depends only on the class of homotopy equivalence of γ.

The monodromy group is the subgroup of $\mathrm{GL}(n,\mathbf{C})$ generated by all monodromy operators. It is defined only up to conjugacy due to the freedom to choose a_0 and B.

Definition 5 Define the product (concatenation) $\gamma_1\gamma_2$ of two paths γ_1, γ_2 in Σ (where the end of γ_1 coincides with the beginning of γ_2) as the path obtained by running γ_1 first and γ_2 next.

Remark 6 The monodromy group is an antirepresentation of the fundamental group $\pi_1(\Sigma)$ into $\mathrm{GL}(n,\mathbf{C})$ because one has

$$X \overset{\gamma_1}{\mapsto} XM_1 \overset{\gamma_2}{\mapsto} XM_2M_1 \qquad [6]$$

that is, the concatenation $\gamma_1\gamma_2$ of the two contours defines the monodromy operator M_2M_1. In the text, the monodromy group is referred to as to a representation, not an antirepresentation.

One usually chooses a standard set of generators of $\pi_1(\Sigma)$ (see **Figure 1**) defined by contours $\gamma_j, j = 1, \ldots, p+1$, where γ_j consists of a segment

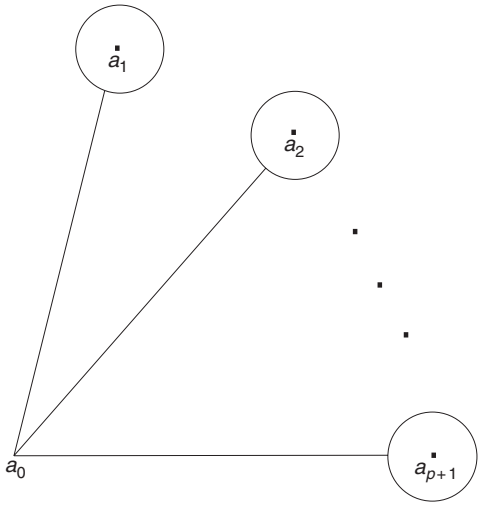


Figure 1 The standard set of generators.

$[a_0, a_j']$ (a_j' being a point close to a_j), of a small circumference run counterclockwise (centered at a_j, passing through a_j' and containing inside no pole of the system other than a_j), and of the segment $[a_j', a_0]$. Thus, γ_j is freely homotopic to a small loop circumventing counterclockwise a_j (and no other pole a_i). The indices of the poles are chosen such that the indices of the contours increase from 1 to $p+1$ when one turns around a_0 clockwise.

For the standard choice of the contours the generators M_j satisfy the relation

$$M_1 \dots M_{p+1} = I \qquad [7]$$

Indeed, the concatenation of contours $\gamma_{p+1} \dots \gamma_1$ is homotopy equivalent to 0 and equality [7] results from Remark 6.

Remarks 7

(i) If the matrix-residuum A_j of a Fuchsian system has no eigenvalues differing by a nonzero integer, then the monodromy operator M_j defined as above is conjugate to $\exp(2\pi i A_j)$. It is always true that the eigenvalues $\sigma_{k,j}$ of M_j equal $\exp(2\pi i \lambda_{k,j})$, where $\lambda_{k,j}$ are the eigenvalues of A_j.

(ii) If the generators M_j of the monodromy group are defined after a standard set of contours γ_j, then they are conjugate to the corresponding operators L_j of local monodromy, that is, when the poles a_j are circumvented counterclockwise along small loops. The operators L_j of a regular system can be computed (up to conjugacy) algorithmically – one first makes the system Fuchsian at a_j by means of a change [4] and then carries out the computation. Thus, $M_j = Q_j^{-1} L_j Q_j$ for some $Q_j \in GL(n, \mathbf{C})$ and the difficulty when computing the monodromy group of system [1] consists in computing the matrices Q_j which is a transcendental problem.

(iii) As will be noted in Theorem 9, every component of every solution to a regular linear system is a function of the class of Nilsson, that is, representable as a convergent (on sectors) series $\sum_{k \in N,\, 1 \le i \le n,\, 0 \le \nu \le n-1} a_{i,k,\nu} t^{\alpha_i + k} \ln^{\nu} t$, $\alpha_i \in \mathbf{C}$, $a_{i,k,\nu} \in \mathbf{C}$.

Example 8 The Fuchsian system $dX/dt = (A/t)X$, $A \in \mathrm{gl}(n, \mathbf{C})$, has two poles – at 0 and at ∞, with matrices-residua A and $-A$. Any solution is of the form $X = \exp(A \ln t)G$, $G \in \mathrm{GL}(n, \mathbf{C})$. To compute the local monodromy around 0, change the argument of t by $2\pi i$. This results in $\ln t \mapsto \ln t + 2\pi i$ and $X \mapsto XG^{-1} \exp(2\pi i A)G$, that is the monodromy operator at 0 equals $G^{-1} \exp(2\pi i A)G$ (and in the same way the one at ∞ equals $G^{-1} \exp(-2\pi i A)G$).

Formulation and History of the Problem

The Riemann–Hilbert problem (or Hilbert's twenty-first problem) is formulated as follows:

Prove that for any set of points $a_1, \dots, a_{p+1} \in \mathbf{C}P^1$ *and for any set of matrices* $M_1, \dots, M_p \in \mathrm{GL}(n, \mathbf{C})$ *there exists a Fuchsian linear system with poles at and only at* $a_1, \dots, a_{p+1}$ *for which the corresponding monodromy operators are* $M_1, \dots, M_p$, $M_{p+1} = (M_1 \dots M_p)^{-1}$.

Historically, the Riemann–Hilbert problem was first stated for Fuchsian equations, not for systems – Riemann mentions in a note at the end of the 1850s the problem how to reconstruct a Fuchsian equation from its monodromy representation and Hilbert includes it in 1900 as the twenty-first problem on his list in a formulation mentioning equations and not systems. However, the number of parameters necessary to parametrize a Fuchsian equation is, in general, smaller than the one necessary to parametrize a monodromy group generated by p matrices. Therefore, one has to allow the presence of additional apparent singularities in the equation, that is, singularities the monodromy around which is trivial.

It had been believed for a long time that the Riemann–Hilbert problem has a positive solution for any $n \in \mathbf{N}$, after J. Plemelj in 1908 gave a proof with a gap. In his proof, Plemelj tries to reduce the Riemann–Hilbert problem to the so-called homogeneous Hilbert boundary-value problem of the theory of singular integral equations. It follows from the correct part of the proof that if one of

the monodromy operators of system [1] is diagonalizable, then system [1] is equivalent to a Fuchsian one; this is due to Yu S Il'yashenko. (In particular, if one allows just one additional apparent singularity, then the Riemann–Hilbert problem is positively solvable. The author has shown that the result still holds if one of the monodromy operators has one Jordan block of size 2 and $n-2$ Jordan blocks of size 1. The result is sharp – it would be false if one allows one Jordan block of size ≥ 3 or two blocks of size 2.) It also follows that any finitely generated subgroup of $\mathrm{GL}(n, \mathbf{C})$ is the monodromy group of a regular system with prescribed poles which is Fuchsian at all the poles with the possible exception of one (where the system is regular) which can be chosen among them at random.

After the publication of Plemelj's result, the interest shifted basically towards the question how to construct a Fuchsian system given the monodromy operators M_j. At the end of the 1920s IA Lappo-Danilevskii expressed the solutions to a Fuchsian system as series of the monodromy operators. These series are convergent for monodromy operators close to the identity matrix and for such operators one can express the residua A_j of the Fuchsian system as convergent series of the monodromy operators.

In 1956 BL Krylov proved that the Riemann–Hilbert problem is solvable for $n=p=2$ by constructing a Fuchsian system after its monodromy group. In 1983 NP Erugin did the same in the case $n=2, p=3$, and established a connection between the Riemann–Hilbert problem and Painlevé's equations.

In 1957 H Röhrl reformulated the problem in terms of fibre bundles. His approach is more geometric; however, it does not require the system realizing a given monodromy group to be Fuchsian, but only regular.

In 1978 W Dekkers considered the particular case $n=2$ of the Riemann–Hilbert problem, and gave a positive answer to it. The gap in Plemelj's proof was detected in the 1980s by AT Kohn and YuS Il'yashenko.

It was proved by AA Bolibrukh in 1989 that, for $n \geq 3$, the problem has a negative answer. For $n=3$, the answer is negative precisely for those couples (monodromy group, set of poles) for which each monodromy operator $M_1, \ldots, M_{p+1}$ is conjugate to a Jordan block of size 3, the monodromy group is reducible, with an invariant subspace or factor-space of dimension 2, the monodromy sub- or factor-representation corresponding to it is irreducible and cannot be realized by a Fuchsian system having all its matrices-residua conjugate to Jordan blocks of size 2. In Bolibrukh's work, the last condition is formulated in a different (but equivalent) way using the notion of Fuchsian weight.

The New Setting of the Problem

After the negative answer to the Riemann–Hilbert problem for $n \geq 3$, it is reasonable to reformulate it as follows:

Find necessary and/or sufficient conditions for the choice of the monodromy operators $M_1, \ldots, M_p$ and the points $a_1, \ldots, a_{p+1}$ so that there should exist a Fuchsian system with poles at and only at the given points and whose monodromy operators M_j should be the given ones.

In the new setting of the Riemann–Hilbert problem, the answer is positive if the monodromy group is irreducible (for any positions of the poles a_j). This has been first proved by Bolibrukh for $n=3$ and then independently by the author and by him for any n.

Bolibrukh found many examples of couples (reducible monodromy group, poles) for which the answer to the Riemann–Hilbert problem is negative. For $n=3$, the negative answer is due to possible "bad position" of the poles and a small shift from this position while keeping the same monodromy group leads to a couple for which the answer is positive. For $n \geq 4$, there are couples where the negative answer is due to arithmetic properties of the eigenvalues of the matrices-residua and the corresponding monodromy groups are not realizable by Fuchsian systems for any position of the poles. During the last years of his life, Bolibrukh studied upper-triangular monodromy representations and found other examples with negative answer to the Riemann–Hilbert problem.

Bolibrukh also found some sufficient conditions for the positive resolvability of the Riemann–Hilbert problem in the case of a reducible monodromy group. For example, suppose that the monodromy group is a semidirect sum:

$$M_j = \begin{pmatrix} M_j^1 & * \\ 0 & M_j^2 \end{pmatrix}$$

where the matrices M_j^i (of size $l_i \times l_i, i=1,2$) define the representations χ_i. Suppose that the representation χ_2 is realizable by a Fuchsian system, that the representation χ_1 is irreducible, and that one of the matrices M_j is block-diagonal, with left upper block of size $s \times s$, where $s \leq l_1$. Then for any choice of the poles a_j the monodromy group can be realized by some Fuchsian system.

Bolibrukh also gave an estimation upon the number m of additional apparent singularities in a Fuchsian equation which are sufficient to realize a given irreducible monodromy group. It follows from his result that

$$m \leq \frac{n(n-1)(p-1)}{2} + 1 - n$$

One can ask the question what the codimension of the subset in the space (monodromy group, poles) is which provides the negative answer to the Riemann–Hilbert problem in its initial setting. The (author's) answer for $p \geq 3$ is $2p(n-1)$, and for $n \geq 7$ this codimension is attained only at couples (monodromy group, poles) for which every monodromy operator M_j is conjugate to a Jordan block of size n, the group has an invariant subspace or factor-space of dimension $n-1$, the corresponding sub- or factor-representation is irreducible and cannot be realized by a Fuchsian system in which all matrices-residua are conjugate to Jordan blocks of size $n-1$. For $n \leq 6$ there are examples where the same codimension is attained (but cannot be decreased) on other couples as well.

Levelt's Result and Bolibrukh's Method

In 1961, AHM Levelt described the form of the solution to a regular system at its pole. His result is in the core of Bolibrukh's method for solving the Riemann–Hilbert problem.

Theorem 9 *In the neighborhood of a pole, the solution to a regular linear system is representable in the form*

$$X = U_j(t-a_j)(t-a_j)^{D_j}(t-a_j)^{E_j}G_j \qquad [8]$$

where the matrix U_j is holomorphic in a neighborhood of 0, $D_j = \mathrm{diag}(\varphi_{1,j}, \ldots, \varphi_{n,j}), \varphi_{n,j} \in \mathbf{Z}$, $\det G_j \neq 0$. The matrix E_j is in upper-triangular form and the real parts of its eigenvalues belong to $[0,1)$ (by definition, $(t-a_j)^{E_j} = \mathrm{e}^{E_j \ln(t-a_j)}$). The numbers $\varphi_{k,j}$ satisfy the condition [10] formulated below. They are valuations in the eigenspaces of the monodromy operator M_j (i.e., in the maximal subspaces invariant for M_j on which it acts as an operator with a single eigenvalue).

A regular system is Fuchsian at a_j if and only if

$$\det U_j(0) \neq 0 \qquad [9]$$

The condition on $\varphi_{k,j}$ can be formulated as follows: let E_j have one and the same eigenvalue in the rows with indices $s_1 < s_2 < \cdots < s_q$. Then one has

$$\varphi_{s_1,j} \geq \varphi_{s_2,j} \geq \cdots \geq \varphi_{s_q,j} \qquad [10]$$

Remark 10 Denote by $\beta_{k,j}$ the diagonal entries (i.e., the eigenvalues) of the matrix E_j. Then the sums $\beta_{k,j} + \varphi_{k,j}$ are the eigenvalues of the matrix-residuum A_j at a_j.

In proving that the Riemann–Hilbert problem is positively solved in the case of an irreducible monodromy group, Bolibrukh (or the author) uses the correct part of Plemelj's proof – namely, that the given monodromy group can be realized by a regular system which is Fuchsian at all poles but one. After this, a suitable change [4] is sought which makes the system Fuchsian at the last pole. The criterium to be Fuchsian is provided by the above theorem; one checks how the matrices D_j, that is, the exponents $\varphi_{k,j}$ and the matrices U_j change as a result of the transformation [4]. This is easier (one has only to multiply to the left by $W(t)$) than to see how the matrix $A(t)$ of system [1] changes because one has conjugation in rule [5]. This idea is also due to Bolibrukh.

When Bolibrukh obtains the negative answer to the Riemann–Hilbert problem in some case of reducible monodromy group, he often uses the following two propositions:

Proposition 11 *The sum $\sum \beta_{k,j} + \varphi_{k,j}$ relative to a subspace of the solution space invariant for all monodromy operators is a non-positive integer.*

In particular, the sum of all exponents $\beta_{k,j} + \varphi_{k,j}$ is a non-positive integer which is 0 if and only if the system is Fuchsian.

Proposition 12 *If some component of some column of some matrix solution to a regular system is identically equal to 0, then the monodromy group of the system is reducible.*

A reducible monodromy group can be conjugated to a block upper-triangular form, with the diagonal blocks defining irreducible representations. Thus, the Riemann–Hilbert problem for reducible monodromy groups makes necessary the answer to the question "given the set of poles a_j, for which sets of exponents $\varphi_{k,j}$ can a given irreducible monodromy group be realized by such a Fuchsian system?" For $n \geq 2$, an irreducible monodromy group can be *a priori* realized by infinitely many Fuchsian systems, with different sets of exponents $\varphi_{k,j}$. Consider the case when these exponents are fixed for $j \neq 1$; suppose that $a_1 = 0$. The author has shown that then infinitely many of the *a priori* possible choices of the exponents $\varphi_{k,1}$ cannot be realized by Fuchsian systems if and only if the given monodromy group is realized by a Fuchsian system which is obtained from another one via the change of time $t \mapsto t^k/(b_k t^k + b_{k-1}t^{k-1} + \cdots + b_0)$, $b_i \in \mathbf{C}, b_0 \neq 0, k \in \mathbf{N}^*, k > 1$. This change increases the number of poles.

Further Developments – The Deligne–Simpson Problem

The Riemann–Hilbert problem can be generalized for irregular systems as follows. One asks whether for given poles a_j there exists a linear system of ordinary differential equations on the Riemann sphere with these and only these poles which is Fuchsian at the regular singular points, which has prescribed formal normal forms, formal monodromies and Stokes multipliers at the irregular singular points, and which has a prescribed global monodromy.

The Riemann–Hilbert problem has been considered in some papers (of H Esnault, E Vieweg, and C Hertling) in the context of algebraic curves of higher genus instead of CP^1.

The study of the so-called Riemann–Hilbert correspondence between the category of holonomic $\mathcal{D}$-modules and the one of perverse sheaves with constructible cohomology has been initiated in the works of J Bernstein in the algebraic aspect and of M Sato, T Kawai, and M Kashiwara in the analytic one. This has been done in the case of a variety of arbitrary dimension (not necessarily CP^1), with codimension one pole divisor. Perversity has been defined by P Deligne, M Goresky, and R MacPherson. Regularity has been defined by M Kashiwara in the analytic aspect and by Z Mebkhout in the geometric one. Important contributions in the domain are due to Ph Maisonobe, M Merle, N Nitsure, C Sabbah, and the list is far from being exhaustive. The Riemann–Hilbert correspondence plays an important role in other trends of mathematics as well.

The Deligne–Simpson problem is formulated like this: *Give necessary and sufficient conditions upon the choice of the conjugacy classes* $c_j \subset \mathrm{gl}(n, \mathbf{C})$ *or* $C_j \subset \mathrm{GL}(n, \mathbf{C})$ *so that there should exist an irreducible* (*i.e.*, *without proper invariant subspace*) $(p+1)$*-tuple of matrices* $A_j \in c_j$ *satisfying* [3] *or of matrices* M_j *satisfying* [7].

The problem was stated in the 1980s by P Deligne for matrices M_j and in the 1990s by the author for matrices A_j. C Simpson was the first to obtain results towards its resolution in the case of matrices M_j. The problem admits the following geometric interpretation in the case of matrices M_j: *For which* $(p+1)$*-tuples of local monodromies does there exist an irreducible global monodromy with such local monodromies?*

For generic eigenvalues the problem has found a complete solution in the author's papers in the form of a criterium upon the Jordan normal forms defined by the conjugacy classes. The author has treated the case of nilpotent matrices A_j and the one of unipotent matrices M_j as well. For matrices A_j, the problem has been completely solved (for any eigenvalues) by W Crawley-Boevey. The case of matrices A_j with $p = 2$ has been treated by O Gleizer using results of A Klyachko. The case when the matrices M_j are unitary is considered in papers of S Agnihotri, P Belkale, I Biswas, C Teleman, and C Woodward. Several cases of finite groups have been considered by M Dettweiler, S Reiter, K Strambach, J Thompson, and H Völklein. The important rigid case has been studied by NM Katz. Y Haraoka has considered the problem in the context of linear systems in Okubo's normal form. One can find details in an author's survey on the Deligne–Simpson problem (Kostov, 2004).

See also: Affine Quantum Groups; Bicrossproduct Hopf Algebras and Non-Commutative Spacetime; Einstein Equations: Exact Solutions; Holonomic Quantum Fields; Integrable Systems: Overview; Isomonodromic Deformations; Leray–Schauder Theory and Mapping Degree; Painlevé Equations; Riemann–Hilbert Methods in Integrable Systems; Twistors; WDVV Equations and Frobenius Manifolds.

Further Reading

Anosov DV and Bolibruch AA (1994) *The Riemann–Hilbert Problem*, A Publication from the Moscow Institute of Mathematics, Aspects of Mathematics, Vieweg.

Arnol'd VI and Ilyashenko YuS (1988) Ordinary differential equations. In: *Dynamical Systems I*, Encyclopedia of Mathematical Sciences, t. 1. Berlin: Springer.

Beukers F and Heckman G (1989) Monodromy for the hypergeometric function ${}_nF_{n-1}$. *Inventiones Mathematicae* 95: 325–354.

Bolibrukh AA (1990) The Riemann–Hilbert problem. *Russian Mathematical Surveys* 45(2): 1–49.

Bolibrukh AA (1992) Sufficient conditions for the positive solvability of the Riemann–Hilbert problem. *Mathematical Notes* 51(1–2): 110–117.

Crawley-Boevey W (2003) On matrices in prescribed conjugacy classes with no common invariant subspace and sum zero. *Duke Mathematical Journal* 118(2): 339–352.

Dekkers W (1979) The matrix of a connection having regular singularities on a vector bundle of rank 2 on $\boldsymbol{P}^1(C)$. *Lecture Notes in Mathematics* 712: 33–43.

Deligne P (1970) *Equations différentielles à points singuliers réguliers*, Lecture Notes in Mathematics, vol. 163, pp. 133. Berlin: Springer.

Dettweiler M and Reiter S (1999) On rigid tuples in linear groups of odd dimension. *Journal of Algebra* 222(2): 550–560.

Esnault H and Viehweg E (1999) Semistable bundles on curves and irreducible representations of the fundamental group. *Algebraic geometry: Hirzebruch 70 (Warsaw, 1998), Contemporary Mathematics AMS, Providence, RI* 241: 129–138.

Esnault H and Hertling C (2001) Semistable bundles on curves and reducible representations of the fundamental group. *International Journal of Mathematics* 12(7): 847–855.

Haraoka Y (1994) Finite monodromy of Pochhammer equation. *Annales de l'Institut Fourier* 44(3): 767–810.

Katz NM (1995) Rigid Local systems. *Annnals of Mathematics, Studies Series, Study*, vol. 139. Princeton: Princeton University Press.

Kohn A and Treibich (1983) *Un résultat de Plemelj, Mathematics and Physics (Paris 1979/1982)*. In: Progr. Math., vol. 37, pp. 307–312. Boston: Birkhäuser.

Kostov VP (1992) Fuchsian linear systems on CP^1 and the Riemann–Hilbert problem. *Comptes Rendus de l'Académie des Sciences à Paris*, 143–148.

Kostov VP (1999) The Deligne–Simpson problem. *C.R. Acad. Sci. Paris, t. 329 Série I*, 657–662.

Kostov VP (2004) The Deligne–Simpson problem – a survey. *Journal of Algebra* 281: 83–108.

Levelt AHM (1961) Hypergeometric functions. *Indagationes Mathematicae* 23: 361–401.

Maisonobe Ph and Narváez-Macarro L (eds.) (2004) *Eléments de la théorie des systèmes différentiels géométriques*. Cours du C.I.M.P.A. Ecole d'été de Séville. Séminaires et Congrès 8, xx + 430 pages.

Völklein H (1998) Rigid generators of classical groups. *Mathematische Annalen* 311(3): 421–438.

Wasow WR (1976) *Asymptotic Expansions for Ordinary Differential Equations*. New York: Huntington.

Riemannian Holonomy Groups and Exceptional Holonomy

D D Joyce, University of Oxford, Oxford, UK

Riemannian Holonomy Groups

Let (M, g) be a Riemannian n-manifold. The holonomy group $\mathrm{Hol}(g)$ is a Lie subgroup of $\mathrm{O}(n)$, a global invariant of g which measures the constant tensors S on M preserved by the Levi-Civita connection ∇ of g. The most well-known examples of metrics with special holonomy are Kähler metrics, with $\mathrm{Hol}(g) \subseteq U(m) \subset \mathrm{O}(2m)$. A Kähler manifold (M, g) also carries a complex structure J and Kähler 2-form ω with $\nabla J = \nabla\omega = 0$.

The classification of Riemannian holonomy groups gives a list of interesting special Riemannian geometries such as Calabi–Yau manifolds and the exceptional holonomy groups G_2 and Spin(7), all of which are important in physics. These geometries have many features in common with Kähler geometry, and are characterized by the existence of constant exterior forms.

General Properties of Holonomy Groups

Let M be a connected manifold of dimension n and g a Riemannian metric on M, with Levi-Civita connection ∇, regarded as a connection on the tangent bundle TM of M. Suppose $\gamma : [0, 1] \to M$ is a smooth path, with $\gamma(0) = x$ and $\gamma(1) = y$. Let s be a smooth section of $\gamma^*(TM)$, so that $s : [0, 1] \to TM$ with $s(t) \in T_{\gamma(t)}M$ for each $t \in [0, 1]$. Then we say that s is parallel if $\nabla_{\dot\gamma(t)} s(t) = 0$ for all $t \in [0, 1]$, where $\dot\gamma(t)$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\gamma(t) \in T_{\gamma(t)}M$$

For each $v \in T_xM$, there is a unique parallel section s of $\gamma^*(TM)$ with $s(0) = v$. Define a map $P_\gamma : T_xM \to T_yM$ by $P_\gamma(v) = s(1)$. Then P_γ is well defined and linear, and is called the parallel transport map along γ. This easily generalizes to continuous, piecewise-smooth paths γ. As $\nabla g = 0$, we see that $P_\gamma : T_xM \to T_yM$ is orthogonal with respect to the metric g on T_xM and T_yM.

Definition 1 Fix a point $x \in M$. γ is said to be loop based at x if $\gamma : [0, 1] \to M$ is a continuous, piecewise-smooth path with $\gamma(0) = \gamma(1) = x$. If γ is a loop based at x, then the parallel transport map P_γ lies in $\mathrm{O}(T_xM)$, the group of orthogonal linear transformations of T_xM. Define the (Riemannian) holonomy group $\mathrm{Hol}_x(g)$ of g based at x to be

$$\begin{aligned}\mathrm{Hol}_x(g) &= \{P_\gamma : \gamma \text{ is a loop based at } x\} \\ &\subseteq \mathrm{O}(T_xM)\end{aligned} \quad [1]$$

Here are some elementary properties of $\mathrm{Hol}_x(g)$. The only difficult part is showing that $\mathrm{Hol}_x(g)$ is a (closed) Lie subgroup.

Theorem 2 *$\mathrm{Hol}_x(g)$ is a Lie subgroup of $\mathrm{O}(T_xM)$, which is closed and connected if M is simply connected, but need not be closed or connected otherwise. Let $x, y \in M$, and suppose $\gamma : [0, 1] \to M$ is a continuous, piecewise-smooth path with $\gamma(0) = x$ and $\gamma(1) = y$, so that $P_\gamma : T_xM \to T_yM$. Then*

$$P_\gamma \,\mathrm{Hol}_x(g) P_\gamma^{-1} = \mathrm{Hol}_y(g) \quad [2]$$

By choosing an orthonormal basis for T_xM we can identify $\mathrm{O}(T_xM)$ with the Lie group $\mathrm{O}(n)$, and so identify $\mathrm{Hol}_x(g)$ with a Lie subgroup of $\mathrm{O}(n)$. Changing the basis changes the subgroups by conjugation by an element of $\mathrm{O}(n)$. Thus, $\mathrm{Hol}_x(g)$ may be regarded as a Lie subgroup of $\mathrm{O}(n)$ defined up to conjugation. Equation [2] shows that in this sense, $\mathrm{Hol}_x(g)$ is independent of the base point x. Therefore, we omit the subscript x and write $\mathrm{Hol}(g)$ for the holonomy group of g, regarded as a subgroup of $\mathrm{O}(n)$ defined up to conjugation.

It is significant that $\mathrm{Hol}(g)$ is a global invariant of g, that is, it does not vary from point to point like local invariants of g such as the curvature. Generic metrics g on M have $\mathrm{Hol}(g) = \mathrm{SO}(n)$ if M is orientable, and $\mathrm{Hol}(g) = \mathrm{O}(n)$ otherwise. But some special metrics g can have $\mathrm{Hol}(g)$ a proper

subgroup of $\mathrm{SO}(n)$ or $\mathrm{O}(n)$. Then M carries some extra geometric structures compatible with g.

Broadly, the smaller $\mathrm{Hol}(g)$ is as a subgroup of $\mathrm{O}(n)$, the more special g is, and the more extra geometric structures there are. Therefore, understanding and classifying the possible holonomy groups gives a family of interesting special Riemannian geometries, such as Kähler geometry. All of these special geometries have cropped up in physics.

Define the holonomy algebra $\mathfrak{hol}(g)$ to be the Lie algebra of $\mathrm{Hol}(g)$, regarded as a Lie subalgebra of $\mathfrak{o}(n)$, defined up to the adjoint action of $\mathrm{O}(n)$. Define $\mathfrak{hol}_x(g)$ to be the Lie algebra of $\mathrm{Hol}_x(g)$, as a Lie subalgebra of $\mathfrak{o}(T_xM) \cong \Lambda^2 T_x^*M$. The holonomy algebra $\mathfrak{hol}(g)$ is intimately connected with the Riemann curvature tensor $R_{abcd} = g_{ae}R^e{}_{bcd}$ of g.

Theorem 3 *The Riemann curvature tensor R_{abcd} lies in $S^2\mathfrak{hol}_x(g)$ at x, where $\mathfrak{hol}_x(g)$ is regarded as a subspace of $\Lambda^2 T_x^*M$. It also satisfies the first and second Bianchi identities*

$$R_{abcd} + R_{adbc} + R_{acdb} = 0 \qquad [3]$$

$$\nabla_e R_{abcd} + \nabla_c R_{abde} + \nabla_d R_{abec} = 0 \qquad [4]$$

A related result is the Ambrose–Singer holonomy theorem, which, roughly speaking, says that $\mathfrak{hol}_x(g)$ may be reconstructed from $R_{abcd}|_y$ for all $y \in M$, moved to x by parallel transport.

If (M, g) and (N, h) are Riemannian manifolds, the product $M \times N$ carries a product metric $g \times h$. It is easy to show that $\mathrm{Hol}(g \times h) = \mathrm{Hol}(g) \times \mathrm{Hol}(h)$. A Riemannian manifold (M, g) is called reducible if every point has an open neighborhood isometric to a Riemannian product and irreducible otherwise.

Theorem 4 *Let (M, g) be Riemannian n-manifold. Then the natural representation of $\mathrm{Hol}(g)$ on $\mathbb{R}^n$ is reducible if and only if g is reducible.*

There is a class of Riemannian manifolds called the "Riemannian symmetric spaces" which are important in the theory of Riemannian holonomy groups. A Riemannian symmetric space is a special kind of Riemannian manifold with a transitive isometry group. The theory of symmetric spaces was worked out by Élie Cartan in the 1920s, who classified them completely, using his own classification of Lie groups and their representations.

A Riemannian metric g is called "locally symmetric" if $\nabla_e R_{abcd} \equiv 0$, and "nonsymmetric" otherwise. Every locally symmetric metric is locally isometric to a Riemannian symmetric space. The relevance of symmetric spaces to holonomy groups is that many possible holonomy groups are the holonomy group of a Riemannian symmetric space, but are not realized by any nonsymmetric metric. Therefore, by restricting attention to nonsymmetric metrics, one considerably reduces the number of possible Riemannian holonomy groups.

A tensor S on M is constant if $\nabla S = 0$. An important property of $\mathrm{Hol}(g)$ is that it determines the constant tensors on M.

Theorem 5 *Let (M, g) be a Riemannian manifold, with Levi-Civita connection ∇. Fix $x \in M$, so that $\mathrm{Hol}_x(g)$ acts on T_xM, and so on the tensor powers $\bigotimes^k T_xM \otimes \bigotimes^l T_x^*M$. Suppose $S \in C^\infty(\bigotimes^k TM \otimes \bigotimes^l T^*M)$ is a constant tensor. Then $S|_x$ is fixed by the action of $\mathrm{Hol}_x(g)$. Conversely, if $S|_x \in \bigotimes^k T_xM \otimes \bigotimes^l T_x^*M$ is fixed by $\mathrm{Hol}_x(g)$, it extends to a unique constant tensor $S \in C^\infty(\bigotimes^k TM \otimes \bigotimes^l T^*M)$.*

The main idea in the proof is that if S is a constant tensor and $\gamma : [0, 1] \to M$ is a path from x to y, then $P_\gamma(S|_x) = S|_y$, that is, "constant tensors are invariant under parallel transport." In particular, they are invariant under parallel transport around closed loops based at x, and so under elements of $\mathrm{Hol}_x(g)$.

Berger's Classification of Holonomy Groups

Berger classified Riemannian holonomy groups in 1955.

Theorem 6 *Let M be a simply connected, n-dimensional manifold, and g an irreducible, nonsymmetric Riemannian metric on M. Then*

(i) $\mathrm{Hol}(g) = \mathrm{SO}(n)$,
(ii) $n = 2m$ *and* $\mathrm{Hol}(g) = \mathrm{SU}(m)$ *or* $\mathrm{U}(m)$,
(iii) $n = 4m$ *and* $\mathrm{Hol}(g) = \mathrm{Sp}(m)$ *or* $\mathrm{Sp}(m)\mathrm{Sp}(1)$,
(iv) $n = 7$ *and* $\mathrm{Hol}(g) = G_2$, *or*
(v) $n = 8$ *and* $\mathrm{Hol}(g) = \mathrm{Spin}(7)$.

To simplify the classification, Berger makes three assumptions: M is simply connected, g is irreducible, and g is nonsymmetric. We can make M simply connected by passing to the "universal cover." The holonomy group of a reducible metric is a product of holonomy groups of irreducible metrics, and the holonomy groups of locally symmetric metrics follow from Cartan's classification of Riemannian symmetric spaces. Thus, these three assumptions can easily be removed.

Here is a sketch of Berger's proof of Theorem 6. As M is simply connected, Theorem 2 shows $\mathrm{Hol}(g)$ is a closed, connected Lie subgroup of $\mathrm{SO}(n)$, and since g is irreducible, Theorem 4 shows the representation of $\mathrm{Hol}(g)$ on $\mathbb{R}^n$ is irreducible. So, suppose that H is a closed, connected subgroup of

$\mathrm{SO}(n)$ acting irreducibly on $\mathbb{R}^n$, with Lie algebra $\mathfrak{h}$. The classification of all such H follows from the classification of Lie groups (and is of considerable complexity). Berger's method was to take the list of all such groups H, and to apply two tests to each possibility to find out if it could be a holonomy group. The only groups H which passed both tests are those in the theorem.

Berger's tests are algebraic and involve the curvature tensor. Suppose that R_{abcd} is the Riemann curvature of a metric g with $\mathrm{Hol}(g) = H$. Then Theorem 3 gives $R_{abcd} \in S^2\mathfrak{h}$, and the first Bianchi identity [3] applies. But if $\mathfrak{h}$ has large codimension in $\mathfrak{o}(n)$, then the vector space $\mathfrak{R}^H$ of elements of $S^2\mathfrak{h}$ satisfying [3] will be small, or even zero. However, the "Ambrose–Singer holonomy theorem" shows that $\mathfrak{R}^H$ must be big enough to generate $\mathfrak{h}$. For many of the candidate groups H, this does not hold, and so H cannot be a holonomy group. This is the first test.

Now $\nabla_e R_{abcd}$ lies in $(\mathbb{R}^n)^* \otimes \mathfrak{R}^H$, and also satisfies the second Bianchi identity, eqn [4]. Frequently, these imply that $\nabla R = 0$, so that g is locally symmetric. Therefore, we may exclude such H, and this is Berger's second test.

Berger's proof does not show that the groups on his list actually occur as Riemannian holonomy groups – only that no others do. It is now known, though this took another thirty years to find out, that all possibilities in Theorem 6 do occur.

The Groups on Berger's List

Here are some brief remarks about each group on Berger's list.

(i) $\mathrm{SO}(n)$ is the holonomy group of generic Riemannian metrics.

(ii) Riemannian metrics g with $\mathrm{Hol}(g) \subseteq \mathrm{U}(m)$ are called "Kähler metrics." Kähler metrics are a natural class of metrics on complex manifolds, and generic Kähler metrics on a given complex manifold have holonomy $\mathrm{U}(m)$.

Metrics g with $\mathrm{Hol}(g) = \mathrm{SU}(m)$ are called Calabi–Yau metrics. Since $\mathrm{SU}(m)$ is a subgroup of $\mathrm{U}(m)$, all Calabi–Yau metrics are Kähler. If g is Kähler and M is simply connected, then $\mathrm{Hol}(g) \subseteq \mathrm{SU}(m)$ if and only if g is Ricci-flat. Thus, Calabi–Yau metrics are locally more or less the same as Ricci-flat Kähler metrics.

If (M, J) is a compact complex manifold with trivial canonical bundle admitting Kähler metrics, then Yau's solution of the Calabi conjecture gives a unique Ricci-flat Kähler metric in each canonical class. This gives a way to construct many examples of Calabi–Yau manifolds, and explains why these have been named after them.

(iii) Metrics g with $\mathrm{Hol}(g) = \mathrm{Sp}(m)$ are called "hyper-Kähler." As $\mathrm{Sp}(m) \subseteq \mathrm{SU}(2m) \subset \mathrm{U}(2m)$, hyper-Kähler metrics are Ricci-flat and Kähler.

Metrics g with holonomy group $\mathrm{Sp}(m)\mathrm{Sp}(1)$ for $m \geq 2$ are called "quaternionic Kähler." (Note that quaternionic Kähler metrics are not in fact Kähler.) They are Einstein, but not Ricci-flat.

(iv), (v) G_2 and Spin(7) are the exceptional cases, so they are called the "exceptional holonomy groups." Metrics with these holonomy groups are Ricci-flat.

The groups can be understood in terms of the four division algebras: the real numbers $\mathbb{R}$, the complex numbers $\mathbb{C}$, the quaternions $\mathbb{H}$, and the octonions or Cayley numbers $\mathbb{O}$.

- $\mathrm{SO}(n)$ is a group of automorphisms of $\mathbb{R}^n$.
- $\mathrm{U}(m)$ and $\mathrm{SU}(m)$ are groups of automorphisms of $\mathbb{C}^m$.
- $\mathrm{Sp}(m)$ and $\mathrm{Sp}(m)\ \mathrm{Sp}(1)$ are automorphism groups of $\mathbb{H}^m$.
- G_2 is the automorphism group of $\mathrm{Im}\,\mathbb{O} \cong \mathbb{R}^7$. Spin(7) is a group of automorphisms of $\mathbb{O} \cong \mathbb{R}^8$, preserving part of the structure on $\mathbb{O}$.

The Exceptional Holonomy Groups

For some time after Berger's classification, the exceptional holonomy groups remained a mystery. In 1987, Bryant used the theory of exterior differential systems to show that locally there exist many metrics with these holonomy groups, and gave some explicit, incomplete examples. Then in 1989, Bryant and Salamon found explicit, complete metrics with holonomy G_2 and Spin(7) on non-compact manifolds. In 1994–95, the author constructed the first examples of metrics with holonomy G_2 and Spin(7) on compact manifolds. For more information on exceptional holonomy, see Joyce (2000, 2002).

The Holonomy Group G_2

Let $(x_1, \ldots, x_7)$ be coordinates on $\mathbb{R}^7$. Write $\mathrm{d}x_{ij\ldots l}$ for the exterior form $\mathrm{d}x_i \wedge \mathrm{d}x_j \wedge \cdots \wedge \mathrm{d}x_l$ on $\mathbb{R}^7$. Define a metric g_0, a 3-form φ_0, and a 4-form $*\varphi_0$ on $\mathbb{R}^7$ by

$$\begin{aligned} g_0 &= \mathrm{d}x_1^2 + \cdots + \mathrm{d}x_7^2 \\ \varphi_0 &= \mathrm{d}x_{123} + \mathrm{d}x_{145} + \mathrm{d}x_{167} + \mathrm{d}x_{246} \\ &\quad - \mathrm{d}x_{257} - \mathrm{d}x_{347} - \mathrm{d}x_{356} \\ *\varphi_0 &= \mathrm{d}x_{4567} + \mathrm{d}x_{2367} + \mathrm{d}x_{2345} + \mathrm{d}x_{1357} \\ &\quad - \mathrm{d}x_{1346} - \mathrm{d}x_{1256} - \mathrm{d}x_{1247} \end{aligned} \qquad [5]$$

The subgroup of $\mathrm{GL}(7, \mathbb{R})$ preserving φ_0 is the exceptional Lie group G_2. It also preserves $g_0, *\varphi_0$,

and the orientation on $\mathbb{R}^7$. It is a compact, semisimple, 14-dimensional Lie group, a subgroup of SO(7).

A G_2-structure on a 7-manifold M is a principal sub-bundle of the frame bundle of M, with structure group G_2. Each G_2-structure gives rise to a 3-form φ and a metric g on M, such that every tangent space of M admits an isomorphism with $\mathbb{R}^7$ identifying φ and g with φ_0 and g_0, respectively. By an abuse of notation, (φ, g) can be referred to as a G_2-structure.

Proposition 7 *Let M be a 7-manifold and (φ, g) a G_2-structure on M. Then the following are equivalent*:

(i) $\mathrm{Hol}(g) \subseteq G_2$, *and φ is the induced 3-form*;
(ii) $\nabla\varphi = 0$ *on M, where ∇ is the Levi-Civita connection of g; and*
(iii) $\mathrm{d}\varphi = \mathrm{d}(*\varphi) = 0$ *on M.*

The equations $\mathrm{d}\varphi = \mathrm{d}(*\varphi) = 0$ look like linear partial differential equations on φ. However, it is better to consider them as nonlinear, for the following reason. The 3-form φ determines the metric g, and g gives the Hodge star $*$ on M. So $*\varphi$ is a nonlinear function of φ, and $\mathrm{d}(*\varphi) = 0$ a nonlinear equation. Thus, constructing and studying G_2-manifolds come down to studying solutions of nonlinear elliptic partial differential equations.

Note that $\mathrm{Hol}(g) \subseteq G_2$ if and only if $\nabla\varphi = 0$ follows from Theorem 5. We call $\nabla\varphi$ the "torsion" of the G_2-structure (φ, g), and when $\nabla\varphi = 0$ the G_2-structure is "torsion-free." A triple (M, φ, g) is called a G_2-manifold if M is a 7-manifold and (φ, g) a torsion-free G_2-structure on M. If g has holonomy $\mathrm{Hol}(g) \subseteq G_2$, then g is Ricci-flat.

Theorem 8 *Let M be a compact 7-manifold, and suppose that (φ, g) is a torsion-free G_2-structure on M. Then $\mathrm{Hol}(g) = G_2$ if and only if $\pi_1(M)$ is finite. In this case, the moduli space of metrics with holonomy G_2 on M, up to diffeomorphisms isotopic to the identity, is a smooth manifold of dimension $b^3(M)$.*

The Holonomy Group Spin(7)

Let $\mathbb{R}^8$ have coordinates $(x_1, \ldots, x_8)$. Define a 4-form Ω_0 on $\mathbb{R}^8$ by

$$\begin{aligned}\Omega_0 = {} & \mathrm{d}x_{1234} + \mathrm{d}x_{1256} + \mathrm{d}x_{1278} + \mathrm{d}x_{1357} - \mathrm{d}x_{1368} \\ & - \mathrm{d}x_{1458} - \mathrm{d}x_{1467} - \mathrm{d}x_{2358} - \mathrm{d}x_{2367} - \mathrm{d}x_{2457} \\ & + \mathrm{d}x_{2468} + \mathrm{d}x_{3456} + \mathrm{d}x_{3478} + \mathrm{d}x_{5678} \end{aligned} \quad [6]$$

The subgroup of $\mathrm{GL}(8, \mathbb{R})$ preserving Ω_0 is the holonomy group Spin(7). It also preserves the orientation on $\mathbb{R}^8$ and the Euclidean metric $g_0 = \mathrm{d}x_1^2 + \cdots + \mathrm{d}x_8^2$. It is a compact, semisimple, 21-dimensional Lie group, a subgroup of SO(8).

A Spin(7)-structure on an 8-manifold M gives rise to a 4-form Ω and a metric g on M, such that each tangent space of M admits an isomorphism with $\mathbb{R}^8$ identifying Ω and g with Ω_0 and g_0, respectively. By an abuse of notation, the pair (Ω, g) is referred to as a Spin(7)-structure.

Proposition 9 *Let M be an 8-manifold and (Ω, g) a* Spin(7)*-structure on M. Then the following are equivalent*:

(i) $\mathrm{Hol}(g) \subseteq \mathrm{Spin}(7)$ *and Ω is the induced 4-form*;
(ii) $\nabla\Omega = 0$ *on M, where ∇ is the Levi-Civita connection of g; and*
(iii) $\mathrm{d}\Omega = 0$ *on M.*

We call $\nabla\Omega$ the torsion of the Spin(7)-structure (Ω, g), and (Ω, g) torsion free if $\nabla\Omega = 0$. A triple (M, Ω, g) is called a Spin(7)-manifold if M is an 8-manifold and (Ω, g) a torsion-free Spin(7)-structure on M. If g has holonomy $\mathrm{Hol}(g) \subseteq \mathrm{Spin}(7)$, then g is Ricci-flat.

Here is a result on compact 8-manifolds with holonomy Spin(7).

Theorem 10 *Let (M, Ω, g) be a compact* Spin(7)*-manifold. Then,* $\mathrm{Hol}(g) = \mathrm{Spin}(7)$ *if and only if M is simply connected, and $b^3(M) + b^4_+(M) = b^2(M) + 2b^4_-(M) + 25$. In this case, the moduli space of metrics with holonomy* Spin(7) *on M, up to diffeomorphisms isotopic to the identity, is a smooth manifold of dimension $1 + b^4_-(M)$.*

The inclusions between the holonomy groups $\mathrm{SU}(m)$, G_2, Spin(7) are

$$\begin{array}{ccccc} \mathrm{SU}(2) & \longrightarrow & \mathrm{SU}(3) & \longrightarrow & G_2 \\ \downarrow & & \downarrow & & \downarrow \\ \mathrm{SU}(2) \times \mathrm{SU}(2) & \longrightarrow & \mathrm{SU}(4) & \longrightarrow & \mathrm{Spin}(7) \end{array} \quad [7]$$

The meaning of the above equation is illustrated by using the inclusion $\mathrm{SU}(3) \hookrightarrow G_2$. As SU(3) acts on $\mathbb{C}^3$, it also acts on $\mathbb{R} \oplus \mathbb{C}^3 \cong \mathbb{R}^7$, taking the SU(3)-action on $\mathbb{R}$ to be trivial. Thus, we embed SU(3) as a subgroup of $\mathrm{GL}(7, \mathbb{R})$. It turns out that SU(3) is contained in the subgroup G_2 of $\mathrm{GL}(7, \mathbb{R})$ defined in the section "The holonomy group G_2."

Constructing Compact G_2- and Spin(7)-Manifolds

The author's method of constructing compact 7-manifolds with holonomy G_2 is based on the

Kummer construction for Calabi–Yau metrics on the $K3$ surface and may be divided into four steps.

Step 1. Let T^7 be the 7-torus and (φ_0, g_0) a flat G_2-structure on T^7. Choose a finite group Γ of isometries of T^7 preserving (φ_0, g_0). Then the quotient T^7/Γ is a singular, compact 7-manifold, an orbifold.

Step 2. For certain special groups Γ, there is a method to resolve the singularities of T^7/Γ in a natural way, using complex geometry. We get a nonsingular, compact 7-manifold M, together with a map $\pi : M \to T^7/\Gamma$, the resolving map.

Step 3. On M, we explicitly write down a one-parameter family of G_2-structures (φ_t, g_t) depending on $t \in (0, \epsilon)$. They are not torsion free, but have small torsion when t is small. As $t \to 0$, the G_2-structure (φ_t, g_t) converges to the singular G_2-structure $\pi^*(\varphi_0, g_0)$.

Step 4. We prove using analysis that for sufficiently small t, the G_2-structure (φ_t, g_t) on M, with small torsion, can be deformed to a G_2-structure $(\tilde{\varphi}_t, \tilde{g}_t)$, with zero torsion. Finally, it is shown that $\tilde{g}_t$ is a metric with holonomy G_2 on the compact 7-manifold M.

We explain the first two steps in greater detail. For Step 1, an example of a suitable group Γ is given here.

Example 11 Let $(x_1, \ldots, x_7)$ be coordinates on $T^7 = \mathbb{R}^7/\mathbb{Z}^7$, where $x_i \in \mathbb{R}/\mathbb{Z}$. Let (φ_0, g_0) be the flat G_2-structure on T^7 defined by [5]. Let α, β, and γ be the involutions of T^7 defined by

$$\begin{aligned}\alpha : (x_1, \ldots, x_7) \\ \mapsto (x_1, x_2, x_3, -x_4, -x_5, -x_6, -x_7)\end{aligned} \quad [8]$$

$$\begin{aligned}\beta : (x_1, \ldots, x_7) \\ \mapsto (x_1, -x_2, -x_3, x_4, x_5, \tfrac{1}{2} - x_6, -x_7)\end{aligned} \quad [9]$$

$$\begin{aligned}\gamma : (x_1, \ldots, x_7) \\ \mapsto \left(-x_1, x_2, -x_3, x_4, \tfrac{1}{2} - x_5, x_6, \tfrac{1}{2} - x_7\right)\end{aligned} \quad [10]$$

By inspection, α, β, and γ preserve (φ_0, g_0), because of the careful choice of exactly which signs to change. Also, $\alpha^2 = \beta^2 = \gamma^2 = 1$, and α, β, and γ commute. Thus, they generate a group $\Gamma = \langle \alpha, \beta, \gamma \rangle \cong \mathbb{Z}_2^3$ of isometries of T^7 preserving the flat G_2-structure (φ_0, g_0).

Having chosen a lattice Λ and finite group Γ, the quotient T^7/Γ is an orbifold, a singular manifold with only quotient singularities. The singularities of T^7/Γ come from the fixed points of nonidentity elements of Γ. We now describe the singularities in the example.

Lemma 12 *In Example 11, $\beta\gamma$, $\gamma\alpha$, $\alpha\beta$, and $\alpha\beta\gamma$ have no fixed points on T^7. The fixed points of α, β, γ are each 16 copies of T^3. The singular set S of T^7/Γ is a disjoint union of 12 copies of T^3, 4 copies from each of α, β, γ. Each component of S is a singularity modeled on that of $T^3 \times \mathbb{C}^2/\{\pm 1\}$.*

The most important consideration in choosing Γ is that we should be able to resolve the singularities of T^7/Γ within holonomy G_2, in Step 2. We have no idea how to resolve general orbifold singularities of G_2-manifolds. However, after fifty years of hard work we understand well how to resolve orbifold singularities of Calabi–Yau manifolds, with holonomy $\mathrm{SU}(m)$. This is done by a combination of algebraic geometry, which produces the underlying complex manifold by a crepant resolution, and Calabi–Yau analysis, which produces the Ricci-flat Kähler metric on this complex manifold.

Now the holonomy groups SU(2) and SU(3) are subgroups of G_2, as in [7]. Our tactic in Step 2 is to ensure that all of the singular set S of T^7/Γ can locally be resolved with holonomy SU(2) or SU(3), and then use Calabi–Yau geometry to do this. In particular, suppose each connected component of S is isomorphic to either

1. $T^3 \times \mathbb{C}^2/G$, for G a finite subgroup of SU(2); or
2. $\mathcal{S}^1 \times \mathbb{C}^3/G$, for G a finite subgroup of SU(3) acting freely on $\mathbb{C}^3 \setminus \{0\}$.

One can use complex algebraic geometry to find a crepant resolution X of $\mathbb{C}^2/G$ or Y of $\mathbb{C}^3/G$. Then $T^3 \times X$ or $\mathcal{S}^1 \times Y$ gives a local model for how to resolve the corresponding component of S in T^7/Γ. Thus we construct a nonsingular, compact 7-manifold M by using the patches $T^3 \times X$ or $\mathcal{S}^1 \times Y$ to repair the singularities of T^7/Γ. In the case of Example 11, this means gluing 12 copies of $T^3 \times X$ into T^7/Γ, where X is the blow-up of $\mathbb{C}^2/\{\pm 1\}$ at its singular point.

By considering different groups Γ acting on T^7, and also by finding topologically distinct resolutions $M_1, \ldots, M_k$ of the same orbifold T^7/Γ, we can construct many compact Riemannian 7-manifolds with holonomy G_2. A good number of examples are given in Joyce (2000, chapter 12). **Figure 1** displays the 252 different sets of Betti numbers of compact, simply connected 7-manifolds with holonomy G_2 constructed there together with 5 more sets from Kovalev. It seems likely to the author that the Betti numbers given in **Figure 1** are only a small proportion of

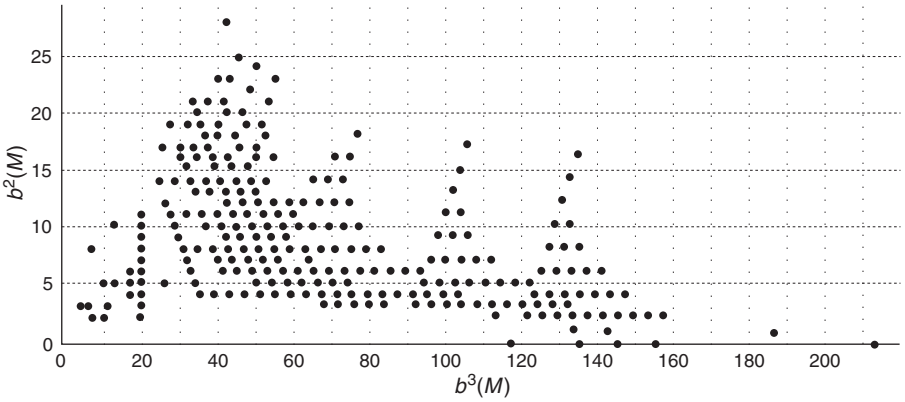


Figure 1 Betti numbers (b^2, b^3) of compact G_2-manifolds. (From Joyce (2000) and Kovalev (2003).)

the Betti numbers of all compact 7-manifolds with holonomy G_2.

A different construction of compact 7-manifolds with holonomy G_2 was given by Kovalev (2003), involving gluing together asymptotically cylindrical Calabi–Yau 3-folds. Compact 8-manifolds with holonomy Spin(7) were constructed by the author using two different methods: first, by resolving singularities of torus orbifolds T^8/Γ in a similar way to the G_2 case (though the details are different and more difficult), and second, by resolving $Y/\langle\sigma\rangle$ for Y a Calabi–Yau 4-orbifold with singularities of a special kind, and σ an antiholomorphic isometric involution of Y. Details can be found in Joyce (2000).

See also: Calibrated Geometry and Special Lagrangian Submanifolds.

Further Reading

Bryant RL (1987) Metrics with exceptional holonomy. *Annals of Mathematics* 126: 525–576.

Bryant RL and Salamon SM (1989) On the construction of some complete metrics with exceptional holonomy. *Duke Mathematical Journal* 58: 829–850.

Gross M, Huybrechts D, and Joyce D (2003) Calabi–Yau *Manifolds and Related Geometries*, Universitext Series. Berlin: Springer.

Joyce DD (2002) Constructing compact manifolds with exceptional holonomy, math.DG/0203158, 2002 (a survey paper).

Joyce DD (2004) Constructing compact manifolds with exceptional holonomy. In: Douglas M, Gauntlett J, and Gross M (eds.) *Strings and Geometry*, Clay Mathematics Proceedings 3. pp. 177–191. Providence, RI: American Mathematical Society. math.DG/0203158.

Kovalev AG (2003) Twisted connected sums and special Riemannian holonomy. *Journal für die Reine und Angewandte Mathematik* 565: 125–160, math.DG/0012189.

S

Saddle Point Problems

M Schechter, University of California at Irvine, Irvine, CA, USA

Introduction

Many problems arising in science and engineering call for the solving of the Euler equations of functionals, that is, equations of the form

$$G'(u) = 0 \qquad [1]$$

where $G(u)$ is a C^1-functional (usually representing the energy) arising from the given data. As an illustration, the equation

$$-\Delta u(x) = f(x, u(x))$$

is the Euler equation of the functional

$$G(u) = \frac{1}{2}\|\nabla u\|^2 - \int F(x, u(x))\,dx$$

on an appropriate space, where

$$F(x,t) = \int_0^t f(x,s)\,ds \qquad [2]$$

and the norm is that of L^2. The solving of the Euler equations is tantamount to finding critical points of the corresponding functional. The classical approach was to look for maxima or minima. If one is looking for a minimum, it is not sufficient to know that the functional is bounded from below, as is easily checked. However, one can show that there is a sequence satisfying

$$G(u_k) \to a, \qquad G'(u_k) \to 0 \qquad [3]$$

for $a = \inf G$. If the sequence has a convergent subsequence, this will produce a minimum.

However, when extrema do not exist, there is no clear way of obtaining critical points. In particular, this happens when the functional is not bounded from either above or below. Until recently, there was no organized procedure for producing critical points which are not extrema. We shall describe an approach which is very useful in such cases.

To illustrate the technique, we consider the problem of finding a solution of

$$-u''(x) + u(x) = f(x, u(x)) \qquad [4]$$

$x \in I = [0, 2\pi]$, under the conditions

$$u(0) = u(2\pi), \qquad u'(0) = u'(2\pi) \qquad [5]$$

We assume that the function $f(x,t)$ is continuous in $I \times \mathbb{R}$ and is periodic in x with period 2π. The approach begins by asking the question, "does there exist a differentiable function G from a space H to $\mathbb{R}$ such that [4], [5] are equivalent to [1]?" It is hoped that one can mimic the methods of calculus to find critical points and thus solve [1].

Actually, we are asking the following: does there exist a mapping G from a space H to $\mathbb{R}$ such that G has a critical point u satisfying $G'(u) = -u'' + u - f(x, u(x))$?

In order to solve the problem one has to

1. find $G(u)$ such that
$$(G'(u), v)_H = (u, v)_H - (f(\cdot, u), v) \qquad [6]$$
holds for each $u, v \in H$,
2. show that there is a function $u(x)$ such that $G'(u) = 0$,
3. show that u'' exists in I,
4. show that [1] implies [4].

We used the notation

$$(u, v) = \int_0^{2\pi} u(x)v(x)\,dx$$

In order to carry out the procedure, we assume that for each $R > 0$ there is a constant C_R such that

$$|f(x,t)| \le C_R, \quad x \in I,\ t \in \mathbb{R},\ |t| \le R \qquad [7]$$

This assumption is used to carry out step (1). We define

$$G(u) = \frac{1}{2}\|u\|_H^2 - \int_0^{2\pi} F(x, u(x))\,dx \qquad [8]$$

where $F(x,t)$ is given by [2] and we take H to be the completion of $C^1(I)$ with respect to the norm

$$\|u\|_H = (\|u'\|^2 + \|u\|^2)^{1/2} \qquad [9]$$

where $\|u\|^2 = (u, u)$. We have

Theorem 1 *If $f(x,t)$ satisfies [7], then $G(u)$ given by [8] is continuously differentiable and satisfies [6].*

Once we have reduced the problem to solving [1], we can search for critical points. The easiest type to locate are "saddle points" which are local minima in some directions and local maxima in all others. For instance, we obtain theorems such as

Theorem 2 *Assume that*

$$\begin{aligned} &|f(x,t)| \le C(|t|+1), \quad x \in I,\ t \in \mathbb{R} \\ &2F(x,t)/t^2 \to \beta(x) \quad \text{a.e. as } |t| \to \infty \end{aligned} \tag{10}$$

with $\beta(x)$ satisfying

$$\begin{aligned} &1+n^2 \le \beta(x) \le 1+(n+1)^2 \\ &1+n^2 \not\equiv \beta(x) \not\equiv 1+(n+1)^2 \end{aligned} \tag{11}$$

and n an integer ≥ 0. If $G(u)$ is given by [8], then there is a $u_0 \in H$ such that

$$G'(u_0) = 0 \tag{12}$$

In particular, u_0 is a solution of [4] and [5] in the usual sense.

In proving this theorem, we shall make use of

Theorem 3 *Let M, N be closed subspaces of a Hilbert space E such that $M = N^{\perp}$. Assume that at least one of these subspaces is finite dimensional. Let G be a continuously differentiable functional on E satisfying*

$$m_0 = \sup_{v \in N} \inf_{w \in M} G(v+w) \ne -\infty \tag{13}$$

and

$$m_1 = \inf_{w \in M} \sup_{v \in N} G(v+w) \ne \infty \tag{14}$$

Then there is a sequence $\{u_k\} \subset E$ such that

$$G(u_k) \to c, \quad m_0 \le c \le m_1, \qquad G'(u_k) \to 0 \tag{15}$$

Theorem 3 allows us to obtain solutions if we can find subspaces of H such that [13] and [14] hold. We use it to give the proof of Theorem 2.

Proof. Note that

$$\|u\|_H^2 = \sum (1+k^2)|\alpha_k|^2, \quad u \in H \tag{16}$$

where the α_k are given by

$$\alpha_k = (u, \bar{\varphi}_k), \quad k = 0, \pm 1, \pm 2, \ldots \tag{17}$$

and

$$\varphi_k(x) = \frac{1}{\sqrt{2\pi}} e^{ikx}, \quad k = 0, \pm 1, \pm 2, \ldots \tag{18}$$

Let

$$N = \{u \in H : \alpha_k = 0 \text{ for } |k| > n\}$$

Thus,

$$\begin{aligned} \|u\|_H^2 &= \sum_{|k| \le n} (1+k^2)|\alpha_k|^2 \\ &\le (1+n^2)\|u\|^2, \quad u \in N \end{aligned} \tag{19}$$

Let

$$M = \{u \in H : \alpha_k = 0 \text{ for } |k| \le n\}$$

In this case,

$$\begin{aligned} \|u\|_H^2 &= \sum_{|k| \ge n+1} (1+k^2)|\alpha_k|^2 \\ &\ge (1+(n+1)^2)\|u\|^2, \quad u \in M \end{aligned} \tag{20}$$

Note that M, N are closed subspaces of H and that $M = N^{\perp}$. Note also that N is finite dimensional. If we consider the functional [8], it is not difficult to show that [11] implies

$$\inf_M G > -\infty; \qquad \sup_N G < \infty \tag{21}$$

We are now in a position to apply Theorem 3. This produces a saddle point satisfying [1]. □

Minimax

Theorem 3 is very useful when extrema do not exist, but it is not always applicable. One is then forced to search for other ways of obtaining critical points. Again, one is faced with the fact that there is no systematic method of finding them. A useful idea is to try to find sets that separate the functional. By this we mean the following:

Definition 1 Two sets A, B separate the functional $G(u)$ if

$$a_0 := \sup_A G \le b_0 := \inf_B G \tag{22}$$

We would like to find sets A and B such that [22] will imply

$$\exists u : G(u) \ge b_0, \quad G'(u) = 0 \tag{23}$$

This is too much to expect since even semiboundedness does not imply the existence of an extremum. Consequently, we weaken our requirements and look for sets A, B such that [22] implies

$$G(u_k) \to a, \qquad G'(u_k) \to 0 \tag{24}$$

with $a \ge b_0$. This leads to

Definition 2 We shall say that the set A links the set B if [22] implies [24] with $a \ge b_0$ for every C^1 functional $G(u)$.

Of course, [24] is a far cry from [23], but if, for example, the sequence [24] has a convergent subsequence, then [24] implies [23]. Whether or not [24] implies [23] is a property of the functional $G(u)$. We state this as

Definition 3 We say that $G(u)$ satisfies the Palais–Smale (PS) condition if [24] always implies [23].

The usual way of verifying this is to show that every sequence satisfying [24] has a convergent subsequence (there are other ways).

All of this leads to

Theorem 4 *If G satisfies the PS condition and is separated by a pair of linking sets, then it has a critical point satisfying* [23].

This theorem cannot be applied until one knows if there are linking sets and functionals that satisfy the PS condition. Fortunately, they exist. Examples and sufficient conditions for A to link B are found in the literature. Obviously, the weaker the conditions, the more pairs will qualify. To date, the conditions described in the next section allow all known examples.

The Details

Let E be a Banach space, and let Φ be the set of all continuous maps $\Gamma = \Gamma(t)$ from $E \times [0,1]$ to E such that

1. $\Gamma(0) = I$, the identity map;
2. for each $t \in [0,1)$, $\Gamma(t)$ is a homeomorphism of E onto E and $\Gamma^{-1}(t) \in C(E \times [0,1), E)$;
3. $\Gamma(1)E$ is a single point in E and $\Gamma(t)A$ converges uniformly to $\Gamma(1)E$ as $t \to 1$ for each bounded set $A \subset E$; and
4. for each $t_0 \in [0,1)$ and each bounded set $A \subset E$,

$$\sup_{0 \le t \le t_0, u \in A} \{\|\Gamma(t)u\| + \|\Gamma^{-1}(t)u\|\} < \infty \qquad [25]$$

We have the following

Theorem 5 *A sufficient condition for A to link B is*

(i) $A \cap B = \phi$ and
(ii) *for each $\Gamma \in \Phi$ there is a $t \in (0,1]$* such that

$$\Gamma(t)A \cap B \neq \phi$$

Theorem 6 *Let G be a C^1-functional on E, and let A, B be subsets of E such that A, B satisfy* [22] *and the hypotheses of Theorem 5. Assume that*

$$a := \inf_{\Gamma \in \Phi} \sup_{0 \le s \le 1, u \in A} G(\Gamma(s)u) \qquad [26]$$

is finite. Let $\psi(t)$ be a positive, locally Lipschitz continuous function on $[0,\infty)$ such that

$$\int_0^\infty \psi(r)\,dr = \infty \qquad [27]$$

Then there is a sequence $\{u_k\} \subset E$ such that

$$G(u_k) \to a, \qquad G'(u_k)/\psi(\|u_k\|) \to 0 \qquad [28]$$

If $a = b_0$, then we can also require that

$$d(u_k, B) \to 0 \qquad [29]$$

Corollary 1 *Under the hypotheses of Theorem 6 there is a sequence $\{u_k\} \subset E$ such that*

$$G(u_k) \to a, \qquad (1 + \|u_k\|)G'(u_k) \to 0 \qquad [30]$$

Proof. We merely take $\psi(u) = 1/(1 + \|u\|)$ in Theorem 6. □

A useful criterion for finding linking subsets is

Theorem 7 *Let F be a continuous map from a Banach space E to $\mathbb{R}^n$, and let $Q \subset E$ be such that $F_0 = F|_Q$ is a homeomorphism of Q onto the closure of a bounded open subset Ω of $\mathbb{R}^n$. If $p \in \Omega$, then $F_0^{-1}(\partial\Omega)$ links $F^{-1}(p)$.*

Some Examples

The following are examples of sets that link.

Example 1 Let M, N be closed subspaces such that $E = M \oplus N$ (with one finite dimensional). Let

$$B_R = \{u \in E : \|u\| < R\}$$

and take $A = \partial B_R \cap N, B = M$. Then A links B. To see this, we identify N with some $\mathbb{R}^n$ and take $\Omega = B_R \cap N, Q = \bar{\Omega}$. For $u \in E$, we write

$$u = v + w, \quad v \in N,\ w \in M \qquad [31]$$

and take F to be the projection

$$Fu = v$$

Since $F|_Q = I$ and $M = F^{-1}(0)$, we see from Theorem 7 that A links B.

Example 2 We take M, N as in Example 1. Let $w_0 \neq 0$ be an element of M, and take

$$A = \{v \in N : \|v\| \le R\} \cup \{sw_0 + v : v \in N, s \ge 0, \|sw_0 + v\| = R\}$$
$$B = \partial B_\delta \cap M, \quad 0 < \delta < R.$$

Then A links B. Again we identify N with some $\mathbb{R}^n$, and we may assume $\|w_0\| = 1$. Let

$$Q = \{sw_0 + v : v \in N, s \ge 0, \|sw_0 + v\| \le R\}$$

Then $A = \partial Q$ in $\mathbb{R}^{n+1}$. If u is given by [31], we define

$$Fu = v + \|w\| w_0$$

Then $F|_Q = I$ and $B = F^{-1}(\delta w_0)$. We can now apply Theorem 7 to conclude that A links B.

Example 3 Take M, N as before and let $v_0 \neq 0$ be an element of N. We write $N = \{v_0\} \oplus N'$. We take

$$\begin{aligned} A &= \{v' \in N' : \|v'\| \leq R\} \\ &\quad \cup \{sv_0 + v' : v' \in N', s \geq 0, \|sv_0 + v'\| = R\} \\ B &= \{w \in M : \|w\| \geq \delta\} \\ &\quad \cup \{sv_0 + w : w \in M, s \geq 0, \|sv_0 + w\| = \delta\} \end{aligned}$$

where $0 < \delta < R$. Then A links B. To see this, we let

$$Q = \{sv_0 + v' : v' \in N', s \geq 0, \|sv_0 + v'\| \leq R\}$$

and reason as before. For simplicity, we assume that $\|v_0\| = 1$, E is a Hilbert space and that the splitting $E = N' \oplus \{v_0\} \oplus M$ is orthogonal. If

$$u = v' + w + sv_0, \quad v' \in N', \ w \in M, \ s \in \mathbb{R} \qquad [32]$$

we define

$$\begin{aligned} F(u) &= v' + \left(s + \delta - \sqrt{\delta^2 - \|w\|^2}\right) v_0, \quad \|w\| \leq \delta \\ &= v' + (s + \delta) v_0, \quad \|w\| > \delta \end{aligned}$$

Note that $F|_Q = I$ while $F^{-1}(\delta v_0)$ is precisely the set B. Hence we can conclude via Theorem 7 that A links B.

Example 4 This is the same as Example 3 with A replaced by $A = \partial B_R \cap N$. The proof is the same with Q replaced by $Q = \bar{B}_R \cap N$.

Example 5 Let M, N be as in Example 1. Take $A = \partial B_\delta \cap N$, and let v_0 be any element in $\partial B_1 \cap N$. Take B to be the set of all u of the form

$$u = w + sv_0, \quad w \in M$$

satisfying any of the following:

(i) $\|w\| \leq R, s = 0$,
(ii) $\|w\| \leq R, s = 2R_0$, and
(iii) $\|w\| = R, 0 \leq s \leq 2R_0$

where $0 < \delta < \min(R, R_0)$. Then A links B. To see this, take $N = \{v_0\} \oplus N'$. Then any $u \in E$ can be written in the form [32]. Define

$$F(u) = v' + \left(R_0 - \max\left\{\frac{R_0}{R}\|w\|, |s - R_0|\right\}\right) v_0$$

and $Q = \bar{B}_\delta \cap N$. Again we may identify N with some $\mathbb{R}^n$. Then $F \in C(E, N)$ and $F|_Q = I$. Moreover, $A = F^{-1}(0)$. Hence, A links B by Theorem 7.

Example 6 Let M, N be as in Example 1. Let v_0 be in $\partial B_1 \cap N$ and write $N = \{v_0\} \oplus N'$. Let $A = \partial B_\delta \cap N$, $Q = \bar{B}_\delta \cap N$, and

$$\begin{aligned} B &= \{w \in M : \|w\| \leq R\} \\ &\quad \cup \{w + sv_0 : w \in M, s \geq 0, \|w + sv_0\| = R\} \end{aligned}$$

where $0 < \delta < R$. Then A links B. To see this, write $u = w + v' + sv_0$, $w \in M$, $v' \in N'$, $s \in \mathbb{R}$ and take

$$F(u) = (cR - \max\{c\|w + sv_0\|, |cR - s|\})v_0 + v'$$

where $c = \delta/(R - \delta)$. Then F is the identity operator on Q, and $F^{-1}(0) = B$. Apply Theorem 7.

Some Applications

Many elliptic semilinear problems can be described in the following way. Let Ω be a domain in $\mathbb{R}^n$, and let A be a self-adjoint operator on $L^2(\Omega)$. We assume that $A \geq \lambda_0 > 0$ and that

$$C_0^\infty(\Omega) \subset D := D(A^{1/2}) \subset H^{m,2}(\Omega) \qquad [33]$$

for some $m > 0$, where $C_0^\infty(\Omega)$ denotes the set of test functions in Ω (i.e., infinitely differentiable functions with compact supports in Ω), and $H^{m,2}(\Omega)$ denotes the Sobolev space. If m is an integer, the norm in $H^{m,2}(\Omega)$ is given by

$$\|u\|_{m,2} := \left(\sum_{|\mu| \leq m} \|D^\mu u\|^2\right)^{1/2} \qquad [34]$$

Here D^μ represents the generic derivative of order $|\mu|$ and the norm on the right-hand side of [34] is that of $L^2(\Omega)$. We shall not assume that m is an integer.

Let q be any number satisfying

$$\begin{aligned} &2 \leq q \leq 2n/(n - 2m), \quad 2m < n \\ &2 \leq q < \infty, \qquad\qquad n \leq 2m \end{aligned}$$

and let $f(x, t)$ be a continuous function on $\Omega \times \mathbb{R}$. We make the following assumptions.

Assumption A The function $f(x, t)$ satisfies

$$|f(x, t)| \leq V_0(x)^q |t|^{q-1} + V_0(x) W_0(x) \qquad [35]$$

and

$$f(x, t)/V_0(x)^q = o(|t|^{q-1}) \text{ as } |t| \to \infty \qquad [36]$$

where $V_0(x) > 0$ is a function in $L^q(\Omega)$ such that

$$\|V_0 u\|_q \leq C\|u\|_D, \quad u \in D \qquad [37]$$

and W_0 is a function in $L^{q'}(\Omega)$. Here

$$\|u\|_q := \left(\int_\Omega |u(x)|^q \, dx\right)^{1/q} \quad [38]$$

$$\|u\|_D := \|A^{1/2}u\| \quad [39]$$

and $q' = q/(q-1)$. With the norm [39], D becomes a Hilbert space. Define G and F by [8] and [2]. It follows that G is a continuously differentiable functional on the whole of D.

We assume further that

$$H(x,t) = 2F(x,t) - tf(x,t) \geq -W_1(x) \in L^1(\Omega), \quad x \in \Omega, t \in \mathbb{R} \quad [40]$$

and

$$H(x,t) \to \infty \quad \text{a.e. as } |t| \to \infty \quad [41]$$

Moreover, we assume that there are functions $V(x), W(x) \in L^2(\Omega)$ such that multiplication by $V(x)$ is a compact operator from D to $L^2(\Omega)$ and

$$F(x,t) \leq C(V(x)^2|t|^2 + V(x)W(x)|t|) \quad [42]$$

We wish to obtain a solution of

$$Au = f(x,u), \quad u \in D \quad [43]$$

By a solution of [43] we shall mean a function $u \in D$ such that

$$(u,v)_D = (f(\cdot,u),v), \quad v \in D \quad [44]$$

If $f(x,u)$ is in $L^2(\Omega)$, then a solution of [44] is in $D(A)$ and solves [43] in the classical sense. Otherwise we call it a weak or semistrong solution. We have

Theorem 8 *Let A be a self-adjoint operator in $L^2(\Omega)$ such that $A \geq \lambda_0 > 0$ and [33] holds for some $m > 0$. Assume that λ_0 is an eigenvalue of A with eigenfunction φ_0. Assume also*

$$2F(x,t) \leq \lambda_0 t^2, \quad |t| \leq \delta \text{ for some } \delta > 0 \quad [45]$$

and

$$2F(x,t) \geq \lambda_0 t^2 - W_0(x), \quad t > 0, x \in \Omega \quad [46]$$

where $W_0 \in L^1(\Omega)$. Assume that $f(x,t)$ satisfies [35], [36], [40], [41], and [42]. Then [43] has a solution $u \neq 0$.

Proof. Under the hypotheses of the theorem, it is known that the following alternative holds: either

(i) there is an infinite number of $y(x) \in D(A)\backslash\{0\}$ such that

$$Ay = f(x,y) = \lambda_0 y \quad [47]$$

or

(ii) for each $\rho > 0$ sufficiently small, there is an $\varepsilon > 0$ such that

$$G(u) \geq \varepsilon, \|u\|_D = \rho \quad [48]$$

We may assume that option (ii) holds, for otherwise we are done. By [46] we have

$$G(R\varphi_0) \leq R^2(\|\varphi_0\|_D^2 - \lambda_0\|\varphi_0\|^2) + \int_\Omega W_0(x)\,dx = \int_\Omega W_0(x)\,dx$$

By Theorem 6, there is a sequence satisfying [28]. Taking $\psi(r) = 1/(r+1)$, we conclude that there is a sequence $\{u_k\} \subset D$ such that

$$G(u_k) \to c, \quad m_0 \leq c \leq m_1, \quad (1 + \|u_k\|_D)G'(u_k) \to 0 \quad [49]$$

In particular, we have

$$\|u_k\|_D^2 - 2\int_\Omega F(x,u_k)\,dx \to c \quad [50]$$

and

$$\|u_k\|_D^2 - (f(\cdot,x_k),u_k) \to 0 \quad [51]$$

Consequently,

$$\int_\Omega H(x,u_k)\,dx \to -c \quad [52]$$

These imply

$$\int_\Omega H(x,u_k)\,dx \leq K \quad [53]$$

If $\rho_k = \|u_k\|_D \to \infty$, let $\tilde{u}_k = u_k/\rho_k$. Then $\|\tilde{u}_k\|_D = 1$. Consequently, there is a renamed subsequence such that $\tilde{u}_k \to \tilde{u}$ weakly in D, strongly in $L^2(\Omega)$, and a.e. in Ω. We have from [42]

$$1 \leq (m_1 + \delta)/\rho_k^2 + 2C\int_\Omega \{V(x)^2\tilde{u}_k^2 + V(x)W(x)|\tilde{u}_k|\rho_k^{-1}\}\,dx$$

Consequently,

$$1 \leq 2C\int_\Omega V(x)^2\tilde{u}^2\,dx \quad [54]$$

This shows that $\tilde{u} \not\equiv 0$. Let Ω_0 be the subset of Ω on which $\tilde{u} \neq 0$. Then

$$|u_k(x)| = \rho_k|\tilde{u}_k(x)| \to \infty, \quad x \in \Omega_0 \quad [55]$$

If $\Omega_1 = \Omega \backslash \Omega_0$, then we have

$$\int_\Omega H(x,u_k)\,dx = \int_{\Omega_0} + \int_{\Omega_1} \geq \int_{\Omega_0} H(x,u_k)\,dx - \int_{\Omega_1} W_1(x)\,dx \to \infty \quad [56]$$

This contradicts [53], and we see that $\rho_k = \|u_k\|_D$ is bounded. Once we know that the ρ_k are bounded, we can apply well-known theorems to obtain the desired conclusion. □

Remark 1 It should be noted that the crucial element in the proof of Theorem 8 was [51]. If we had been dealing with an ordinary Palais–Smale sequence, we could only conclude that

$$\|u_k\|_D^2 - (f(\cdot,u_k),u_k) = o(\rho_k)$$

which would imply only

$$\int_\Omega H(x,u_k)\,dx = o(\rho_k)$$

This would not contradict [56], and the argument would not go through.

As another application, we wish to solve

$$-x''(t) = \nabla_x V(t,x(t)) \quad [57]$$

where

$$x(t) = (x_1(t),\ldots,x_n(t)) \quad [58]$$

is a map from $I = [0,2\pi]$ to $\mathbb{R}^n$ such that each component $x_j(t)$ is a periodic function in H^1 with period 2π, and the function

$$V(t,x) = V(t,x_1,\ldots,x_n)$$

is continuous from $\mathbb{R}^{n+1}$ to $\mathbb{R}$ with a gradient

$$\nabla_x V(t,x) = (\partial V/\partial x_1,\ldots,\partial V/\partial x_n) \in C(\mathbb{R}^{n+1},\mathbb{R}^n) \quad [59]$$

For each $x \in \mathbb{R}^n$, the function $V(t,x)$ is periodic in t with period 2π. We shall study this problem under the following assumptions:

1. $$0 \leq V(t,x) \leq C(|x|^2+1)$$
$$t \in I, x \in \mathbb{R}^n$$

2. There are constants $m > 0, \alpha \leq 3m^2/2\pi^2$ such that
$$V(t,x) \leq \alpha, \quad |x| \leq m, t \in I, x \in \mathbb{R}^n$$

3. There are constants $\beta > 1/2$ and C such that
$$V(t,x) \geq \beta|x|^2$$
when
$$|x| > C, t \in I, x \in \mathbb{R}^n$$

4. The function given by
$$H(t,x) = 2V(t,x) - \nabla_x V(t,x)\cdot x \quad [60]$$
satisfies
$$H(t,x) \leq W(t) \in L^1(I), \quad |x| \geq C \quad [61]$$
$t \in I$, $x \in \mathbb{R}^n$, and
$$H(t,x) \to -\infty \quad \text{as } |x| \to \infty \quad [62]$$

We have

Theorem 9 *Under the above hypotheses, the system* [57] *has a nonconstant solution.*

Proof. Let X be the set of vector functions $x(t)$ described above. It is a Hilbert space with norm satisfying

$$\|x\|_X^2 = \sum_{j=1}^n \|x_j\|_{H^1}^2$$

We also write

$$\|x\|^2 = \sum_{j=1}^n \|x_j\|^2$$

where $\|\cdot\|$ is the $L^2(I)$ norm. Let

$$N = \{x(t) \in X : x_j(t) \equiv \text{constant}, 1 \leq j \leq n\}$$

and $M = N^\perp$. The dimension of N is n, and $X = M \oplus N$. The following is easily proved.

Lemma 1 *If $x \in M$, then*

$$\|x\|_\infty^2 \leq \frac{\pi}{6}\|x'\|^2$$

and

$$\|x\| \leq \|x'\|$$

We define

$$G(x) = \|x'\|^2 - 2\int_I V(t,x(t))\,dt, \quad x \in X \quad [63]$$

For each $x \in X$ write $x = v + w$, where $v \in N$, $w \in M$. For convenience, we shall use the following equivalent norm for X:

$$\|x\|_X^2 = \|w'\|^2 + \|v\|^2$$

If $x \in M$ and

$$\|x'\|^2 = \rho^2 = \frac{6}{\pi}m^2$$

then Lemma 1 implies that $\|x\|_\infty \leq m$, and we have by Hypothesis 2 that $V(t,x) \leq \alpha$.

Hence,

$$G(x) \geq \|x'\|^2 - 2\int_{|x|<m} \alpha \, dt \geq \rho^2 - 2\alpha(2\pi) \geq 0 \quad [64]$$

Note that Hypothesis 3 is equivalent to

$$V(t,x) \geq \beta|x|^2 - C, \quad t \in I,\ x \in \mathbb{R}^n \quad [65]$$

for some constant C. Next, let

$$y(t) = v + sw_0$$

where $v \in N$, $s \geq 0$, and

$$w_0 = (\sin t, 0, \ldots, 0)$$

Then $w_0 \in M$, and

$$\|w_0\|^2 = \|w_0'\|^2 = \pi$$

Note that

$$\|y\|^2 = \|v\|^2 + s^2\pi = 2\pi|v|^2 + \pi s^2$$

Consequently,

$$\begin{aligned} G(y) &= s^2\|w_0'\|^2 - 2\int_I V(t,y(t))\,dt \\ &\leq \pi s^2 - 2\beta\int_I |y(t)|^2\,dt + 2\pi C \\ &= \pi s^2 - 2\beta(\|v\|^2 + \pi s^2) + 2\pi C \\ &\leq (1-2\beta)\pi s^2 - 4\beta\pi|v|^2 + 2\pi C \\ &\to -\infty \text{ as } s^2 + |v|^2 \to \infty \end{aligned}$$

We also note that Hypothesis 1 implies

$$G(v) \leq 0, \quad v \in N \quad [66]$$

Take

$$\begin{aligned} A &= \{v \in N : \|v\| \leq R\} \\ &\quad \cup \{sw_0 + v : v \in N, s \geq 0, \|sw_0 + v\|_X = R\} \\ B &= \partial B_\rho \cap M, 0 < \rho = 6m^2/\pi < R \end{aligned}$$

where

$$B_\sigma = \{x \in X : \|x\|_X < \sigma\}$$

By Example 2, A links B. Moreover, if R is sufficiently large,

$$\sup_A G = 0 \leq \inf_B G \quad [67]$$

Hence, we may conclude that there is a sequence $\{x^{(k)}\} \subset X$ such that

$$G(x^{(k)}) \to c \geq 0, \quad (1 + \|x^{(k)}\|_X)\, G'(x^{(k)}) \to 0$$

Hence,

$$G(x^{(k)}) = \|[x^{(k)}]'\|^2 - 2\int_I V(t,x^{(k)}(t))\,dt \to c \geq 0 \quad [68]$$

$$(G'(x^{(k)}), z)/2 = ([x^{(k)}]', z') - \int_I \nabla_x V(t,x^{(k)}) \cdot z(t)\,dt \to 0, \quad z \in X \quad [69]$$

and

$$(G'(x^{(k)}), x^{(k)})/2 = \|[x^{(k)}]'\|^2 - \int_I \nabla_x V(t,x^{(k)}) \cdot x^{(k)}\,dt \to 0 \quad [70]$$

If

$$\rho_k = \|x^{(k)}\|_X \leq C$$

then there is a renamed subsequence such that $x^{(k)}$ converges to a limit $x \in X$ weakly in X and uniformly on I. From [69] we see that

$$(G'(x), z)/2 = (x', z') - \int_I \nabla_x V(t,x(t)) \cdot z(t)\,dt = 0, \quad z \in X$$

from which we conclude easily that x is a solution of [57]. From [68], we see that

$$G(x) \geq c \geq 0$$

showing that $x(t)$ is not a constant. For if $c > 0$ and $x \in N$, then

$$G(x) = -2\int_I V(t,x(t))\,dt \leq 0$$

If $c = 0$, we see that $x \in B$ by Theorem 6. Hence, $x \in M$. If

$$\rho_k = \|x^{(k)}\|_X \to \infty$$

let $\tilde{x}^{(k)} = x^{(k)}/\rho_k$. Then, $\|\tilde{x}^{(k)}\|_X = 1$. Let $\tilde{x}^{(k)} = \tilde{w}^{(k)} + \tilde{v}^{(k)}$, where $\tilde{w}^{(k)} \in M$ and $\tilde{v}^{(k)} \in N$. There is a renamed subsequence such that $\tilde{x}^{(k)}$ converges uniformly in I to a limit $\tilde{x}$ and $\|[\tilde{x}^{(k)}]'\| \to r$ and $\|\tilde{x}^{(k)}\| \to \tau$, where $r^2 + \tau^2 = 1$. From [68] and [70], we obtain

$$\|[\tilde{x}^{(k)}]'\|^2 - 2\int_I V(t,x^{(k)}(t))\,dt/\rho_k^2 \to 0$$

and

$$\|[\tilde{x}^{(k)}]'\|^2 - \int_I \nabla_x V(t,x^{(k)}) \cdot x^{(k)}\,dt/\rho_k^2 \to 0$$

Thus,

$$2\int_I V(t,x^{(k)}(t))\,dt/\rho_k^2 \to r^2 \quad [71]$$

and

$$\int_I \nabla_x V(t, x^{(k)}) \cdot x^{(k)} \, dt/\rho_k^2 \to r^2 \qquad [72]$$

Hence,

$$\int_I H(t, x^{(k)}(t)) \, dt/\rho_k^2 \to 0 \qquad [73]$$

By Hypothesis 3, the left-hand side of [71] is

$$\geq 2\beta \|\tilde{x}^{(k)}\|^2 - 4\pi C/\rho_k^2$$

Thus,

$$r^2 \geq 2\beta r^2 = 2\beta(1 - r^2)$$

showing that $r > 0$. Hence, $\tilde{x}(t) \not\equiv 0$. Let $\Omega_0 \subset I$ be the set on which $[\tilde{x}(t)] \neq 0$. The measure of Ω_0 is positive. Thus, $|x^{(k)}(t)| \to \infty$ as $k \to \infty$ for $t \in \Omega_0$. Hence,

$$\int_I H(t, x^{(k)}(t)) \, dt \leq \int_{\Omega_0} H(t, x^{(k)}(t)) \, dt + \int_{I \setminus \Omega_0} W(t) \, dt \to -\infty$$

contrary to Hypothesis 4. Thus, the ρ_k are bounded, and the proof is complete. □

Superlinear Problems

Consider the problem

$$-\Delta u = f(x, u), \quad x \in \Omega; \; u = 0 \text{ on } \partial\Omega \qquad [74]$$

where $\Omega \subset \mathbb{R}^n$ is a bounded domain whose boundary is a smooth manifold, and $f(x, t)$ is a continuous function on $\bar{\Omega} \times \mathbb{R}$. This semilinear Dirichlet problem has been studied by many authors. It is called "sublinear" if there is a constant C such that

$$|f(x, t)| \leq C(|t| + 1), \quad x \in \Omega, t \in \mathbb{R}$$

Otherwise, it is called "superlinear". Assume

(a_1) There are constants $c_1, c_2 \geq 0$ such that

$$|f(x, t)| \leq c_1 + c_2 |t|^s$$

where $0 \leq s < (n + 2)/(n - 2)$ if $n > 2$.

(a_2) $f(x, t) = o(|t|)$ as $t \to 0$.

(a_3) Either

$$F(x, t)/t^2 \to \infty \text{ as } t \to \infty$$

or

$$F(x, t)/t^2 \to \infty \text{ as } t \to -\infty.$$

We have

Theorem 10 *Under hypotheses* (a_1)–(a_3) *the boundary-value problem*

$$-\Delta u = \beta f(x, u), \; x \in \Omega; \quad u = 0 \text{ on } \partial\Omega \qquad [75]$$

has a nontrivial solution for almost every positive β.

Unfortunately, this theorem does not give any information for any specific β. It still leaves open the problem of solving [74]. For this purpose, we add the assumption

(a_4) There are constants $\mu > 2, r \geq 0$ such that

$$\mu F(x, t) - t f(x, t) \leq C(t^2 + 1), \quad |t| \geq r \qquad [76]$$

We have

Theorem 11 *Under hypotheses* (a_1)–(a_4) *problem* [74] *has a nontrivial solution.*

We also have

Theorem 12 *If we replace hypothesis* (a_4) *with*

(a_4') *The function* $-H(x, t)$ *is convex in* t,

then the problem [74] *has at least one nontrivial solution.*

Weak Linking

It is not clear if it is possible for A to link B if neither is contained in a finite-dimensional manifold. For instance, if $E = M \oplus N$, where M, N are closed infinite-dimensional subspaces of E and B_R is the ball centered at the origin of radius R in E, it is unknown if the set $A = M \cap \partial B_R$ links $B = N$. (If either M or N is finite dimensional, then A does link B.) Unfortunately, this is the situation which arises in some important applications including Hamiltonian systems, the wave equation and elliptic systems, to name a few.

We now consider linking when both M and N are infinite dimensional and G' has some additional continuity property. A property that is very useful is that of weak-to-weak continuity:

$$u_k \to u \text{ weakly in } E \implies G'(u_k) \to G'(u) \text{ weakly} \qquad [77]$$

We make the following definition:

Definition 3 A subset A of a Banach space E links a subset B of E "weakly" if for every $G \in C^1(E, \mathbb{R})$ satisfying [77] and

$$a_0 := \sup_A G \leq b_0 := \inf_B G \qquad [78]$$

there is a sequence $\{u_k\} \subset E$ and a constant c such that

$$b_0 \leq c < \infty \qquad [79]$$

and

$$G(u_k) \to c, \qquad G'(u_k) \to 0 \tag{80}$$

We have the following counterpart of Theorem 7.

Theorem 13 *Let E be a separable Hilbert space, and let G be a continuous functional on E with a continuous derivative satisfying [77]. Let N be a closed subspace of E, and let Q be a bounded open subset of N containing the point p. Let F be a continuous map of E onto N such that*

(i) $F|_Q = I$, *and*
(ii) *For each finite-dimensional subspace $S \neq \{0\}$ of E containing p, there is a finite-dimensional subspace $S_0 \neq \{0\}$ of N containing p such that*

$$v \in \bar{Q} \cap S_0, \qquad w \in S \Longrightarrow F(v+w) \in S_0 \tag{81}$$

Set $A = \partial Q, B = F^{-1}(p)$. If

$$a_1 = \sup_{\bar{Q}} G < \infty \tag{82}$$

and [22] holds, then there is a sequence $\{u_k\} \subset E$ such that [24] holds with $a \leq a_1$.

Theorem 13 states that if Q, F, p satisfy the hypotheses of that theorem, then $A = \partial Q$ links $B = F^{-1}(p)$ weakly. It follows from this theorem that all sets A, B known to link when one of the subspaces M, N is finite dimensional will link weakly even when M, N are both infinite dimensional.

Now we give some applications of Theorem 13 to semilinear boundary-value problems. Let Ω be a domain in $\mathbb{R}^n$ and let A be a self–adjoint operator in $L^2(\Omega)$ having 0 in its resolvent set (thus, there is an interval (a, b) in its resolvent set satisfying $a < 0 < b$). Let $f(x, t)$ be a continuous function on $\Omega \times \mathbb{R}$ such that

$$|f(x,t)| \leq V(x)^2|t| + W(x)V(x) \tag{83}$$

$x \in \Omega$, $t \in \mathbb{R}$, and

$$f(x,t)/t \to \alpha_\pm(x) \quad \text{as } t \to \pm\infty \tag{84}$$

where $V, W \in L^2(\Omega)$, and multiplication by $V(x) > 0$ is a compact operator from $D = D(|A|^{1/2})$ to $L^2(\Omega)$. Let

$$M = \int_b^\infty dE(\lambda)D, \qquad N = \int_{-\infty}^a dE(\lambda)D$$

where $\{E(\lambda)\}$ is the spectral measure of A. Then M, N are invariant subspaces for A and $D = M \oplus N$. If

$$\alpha(u,v) = \int_\Omega (\alpha_+ u^+ - \alpha_- u^-)v \, dx \tag{85}$$

$\alpha(u) = \alpha(u, u)$, then we assume that

$$\alpha(v) \geq (Av, v), \quad v \in N \tag{86}$$

$$(Aw, w) \geq \alpha(w), \quad w \in M \tag{87}$$

We also assume that the only solution of

$$Au = \alpha_+ u^+ - \alpha_- u^- \tag{88}$$

is $u \equiv 0$, where $u^\pm = \max\{\pm u, 0\}$. We have

Theorem 14 *Under the above hypotheses there is at least one solution of*

$$Au = f(x,u), \quad u \in D(A) \tag{89}$$

Next, we consider an application concerning radially symmetric solutions for the problem

$$u_{tt} - \Delta u = f(t,x,u), \quad t \in \mathbb{R},\ x \in B_R \tag{90}$$

$$u(t,x) = 0, \quad t \in \mathbb{R},\ x \in \partial B_R \tag{91}$$

$$u(t+T,x) = u(t,x), \quad t \in \mathbb{R},\ x \in B_R \tag{92}$$

where $B_R = \{x \in \mathbb{R}^n : |x| < R\}$. We assume that the ratio R/T is rational. Let

$$8R/T = a/b \tag{93}$$

where a, b are relatively prime positive integers. It can be shown that

$$n \not\equiv 3 \pmod{(4,a)} \tag{94}$$

implies that the linear problem corresponding to [90]–[92] has no essential spectrum. If

$$n \equiv 3 \pmod{(4,a)} \tag{95}$$

then the essential spectrum of the linear operator consists of precisely one point

$$\lambda_0 = -(n-3)(n-1)/4R^2 \tag{96}$$

Consider the case

$$f(t,r,s) = \mu s + p(t,r,s) \tag{97}$$

where μ is a point in the resolvent set, $r = |x|$, and

$$|p(t,r,s)| \leq C(|s|^\theta + 1), \quad s \in \mathbb{R} \tag{98}$$

for some number $\theta < 1$. We then have

Theorem 15 *If [94] holds, then [90]–[92] have a weak rotationally invariant solution. If [95] holds and $\lambda_0 < \mu$, assume in addition that $p(t,r,s)$ is nondecreasing in s. If $\mu < \lambda_0$, assume that $p(t,r,s)$ is nonincreasing in s. Then [90]–[92] have a weak rotationally invariant solution.*

See also: Combinatorics: Overview; Homoclinic Phenomena; Ljusternik–Schnirelman Theory; Minimax Principle in the Calculus of Variations.

Further Reading

Ambrosetti A and Prodi G (1993) A primer of nonlinear analysis. *Cambridge Studies in Advanced Mathematics* 34. Cambridge: Cambridge University Press.

Chang KC (1993) *Infinite Dimensional Morse Theory and Multiple Solution Problems*. Boston: Birkhäuser.

Ekeland I and Temam R (1976) *Convex Analysis and Variational Problems*. Amsterdam: North-Holland.

Ghoussoub N (1993) *Duality and Perturbation Methods in Critical Point Theory*. Cambridge: Cambridge University Press.

Mawhin J and Willem M (1989) *Critical Point Theory and Hamiltonian Systems*. Berlin: Springer.

Rabinowitz PR (1986) *Minimax Methods in Critical Point Theory with Applications to Differential Equations*, Conf. Board of Math. Sci. Reg. Conf. Ser. in Math. No. 65. Providence, RI: American Mathematical Society.

Schechter M (1999) *Linking Methods in Critical Point Theory*. Boston: Birkhäuser.

Schechter M (1986) *Spectra of Partial Differential Operators*, 2nd edn. Amsterdam: North-Holland.

Struwe M (1996) *Variational Methods*. Berlin: Springer.

Willem M (1996) *Minimax Theorems*. Boston: Birkhäuser.

Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools

D Buchholz, Universität Göttingen, Göttingen, Germany
S J Summers, University of Florida, Gainesville, FL, USA

Physical Motivation and Mathematical Setting

The primary connection of relativistic quantum field theory to experimental physics is through scattering theory, that is, the theory of the collision of elementary (or compound) particles. It is therefore a central topic in quantum field theory and has attracted the attention of leading mathematical physicists. Although a great deal of progress has been made in the mathematically rigorous understanding of the subject, there are important matters which are still unclear, some of which will be indicated below.

In the paradigmatic scattering experiment, several particles, which are initially sufficiently distant from each other that the idealization that they are not mutually interacting is physically reasonable, approach each other and interact (collide) in a region of microscopic extent. The products of this collision then fly apart until they are sufficiently well separated that the approximation of noninteraction is again reasonable. The initial and final states of the objects in the scattering experiment are therefore to be modeled by states of noninteracting, that is, free, fields, which are mathematically represented on Fock space. Typically, what is measured in such experiments is the probability distribution (cross section) for the transitions from a specified state of the incoming particles to a specified state of the outgoing particles.

It should be mentioned that until the late 1950s, the scattering theory of relativistic quantum particles relied upon ideas from nonrelativistic quantum-mechanical scattering theory (interaction representation, adiabatic limit, etc.), which were invalid in the relativistic context. Only with the advent of axiomatic quantum field theory did it become possible to properly formulate the concepts and mathematical techniques which will be outlined here.

Scattering theory can be rigorously formulated either in the context of quantum fields satisfying the Wightman axioms (Streater and Wightman 1964) or in terms of local algebras satisfying the Haag–Kastler–Araki axioms (Haag 1992). In brief, the relation between these two settings may be described as follows: in the Wightman setting, the theory is formulated in terms of operator-valued distributions ϕ on Minkowski space, the quantum fields, which act on the physical state space. These fields, integrated with test functions f having support in a given region $\mathcal{O}$ of spacetime (only four-dimensional Minkowski space $\mathbb{R}^4$ will be treated here), $\phi(f) = \int \mathrm{d}^4x\, f(x)\phi(x)$, form under the operations of addition, multiplication, and Hermitian conjugation a polynomial *-algebra $\mathcal{P}(\mathcal{O})$ of unbounded operators. In the Haag–Kastler–Araki setting, one proceeds from these algebras to algebras $\mathcal{A}(\mathcal{O})$ of bounded operators which, roughly speaking, are formed by the bounded functions A of the operators $\phi(f)$. This step requires some mathematical care, but these subtleties will not be discussed here. As the statements and proofs of the results in these two frameworks differ only in technical details, the theory is presented here in the more convenient setting of algebras of bounded operators (C*-algebras).

Central to the theory is the notion of a particle, which, in fact, is a quite complex concept, the full nature of which is not completely understood, cf.

below. In order to maintain the focus on the essential points, we consider in the subsequent sections primarily a single massive particle of integer spin s, that is, a boson. In standard scattering theory based upon Wigner's characterization, this particle is simply identified with an irreducible unitary representation U_1 of the identity component $\mathcal{P}_+^\uparrow$ of the Poincaré group with spin s and mass $m > 0$. The Hilbert space $\mathcal{H}_1$ upon which $U_1(\mathcal{P}_+^\uparrow)$ acts is called the one-particle space and determines the possible states of a single particle, alone in the universe. Assuming that configurations of several such particles do not interact, one can proceed by a standard construction to a Fock space describing freely propagating multiple particle states,

$$\mathcal{H}_{\mathrm{F}} = \bigoplus_{n\in\mathbb{N}_0} \mathcal{H}_n$$

where $\mathcal{H}_0 = \mathbb{C}$ and $\mathcal{H}_n$ is the n-fold symmetrized direct product of $\mathcal{H}_1$ with itself. This space is spanned by vectors $\Phi_1 \otimes \cdots \otimes \Phi_n$, where $\otimes$ denotes the symmetrized tensor product, representing an n-particle state wherein the kth particle is in the state $\Phi_k \in \mathcal{H}_1$, $k = 1, \ldots, n$. The representation $U_1(\mathcal{P}_+^\uparrow)$ induces a unitary representation $U_{\mathrm{F}}(\mathcal{P}_+^\uparrow)$ on $\mathcal{H}_{\mathrm{F}}$ by

$$U_{\mathrm{F}}(\lambda)(\Phi_1 \otimes \cdots \otimes \Phi_n) \doteq U_1(\lambda)\Phi_1 \otimes \cdots \otimes U_1(\lambda)\Phi_n \quad [1]$$

In interacting theories, the states in the corresponding physical Hilbert space $\mathcal{H}$ do not have such an *a priori* interpretation in physical terms, however. It is the primary goal of scattering theory to identify in $\mathcal{H}$ those vectors which describe, at asymptotic times, incoming, respectively, outgoing, configurations of freely moving particles. Mathematically, this amounts to the construction of certain specific isometries (generalized Møller operators), Ω^{in} and Ω^{out}, mapping $\mathcal{H}_{\mathrm{F}}$ onto subspaces $\mathcal{H}^{\mathrm{in}} \subset \mathcal{H}$ and $\mathcal{H}^{\mathrm{out}} \subset \mathcal{H}$, respectively, and intertwining the unitary actions of the Poincaré group on $\mathcal{H}_{\mathrm{F}}$ and $\mathcal{H}$. The resulting vectors

$$(\Phi_1 \otimes \cdots \otimes \Phi_n)^{\mathrm{in/out}} \doteq \Omega^{\mathrm{in/out}}(\Phi_1 \otimes \cdots \otimes \Phi_n) \in \mathcal{H} \quad [2]$$

are interpreted as incoming and outgoing particle configurations in scattering processes wherein the kth particle is in the state $\Phi_k \in \mathcal{H}_1$.

If, in a theory, the equality $\mathcal{H}^{\mathrm{in}} = \mathcal{H}^{\mathrm{out}}$ holds, then every incoming scattering state evolves, after the collision processes at finite times, into an outgoing scattering state. It is then physically meaningful to define on this space of states the scattering matrix, setting $S = \Omega^{\mathrm{in}}\Omega^{\mathrm{out}*}$. Physical data such as collision cross sections can be derived from S and the corresponding transition amplitudes $\langle(\Phi_1 \otimes \cdots \otimes \Phi_m)^{\mathrm{in}}, (\Phi'_1 \otimes \cdots \otimes \Phi'_n)^{\mathrm{out}}\rangle$, respectively, by a standard procedure. It should be noted, however, that neither the above physically mandatory equality of state spaces nor the more stringent requirement that every state has an interpretation in terms of incoming and outgoing scattering states, that is, $\mathcal{H} = \mathcal{H}^{\mathrm{in}} = \mathcal{H}^{\mathrm{out}}$ (asymptotic completeness), has been fully established in any interacting relativistic field theoretic model so far. This intriguing problem will be touched upon in the last section of this article.

Before going into details, let us state the few physically motivated postulates entering into the analysis. As discussed, the point of departure is a family of algebras $\mathcal{A}(\mathcal{O})$, more precisely a net, associated with the open subregions $\mathcal{O}$ of Minkowski space and acting on $\mathcal{H}$. Restricting attention to the case of bosons, we may assume that this net is local in the sense that if $\mathcal{O}_1$ is spacelike separated from $\mathcal{O}_2$, then all elements of $\mathcal{A}(\mathcal{O}_1)$ commute with all elements of $\mathcal{A}(\mathcal{O}_2)$. (In the presence of fermions, these algebras contain also fermionic operators which anticommute.) This is the mathematical expression of the principle of Einstein causality. The unitary representation U of $\mathcal{P}_+^\uparrow$ acting on $\mathcal{H}$ is assumed to satisfy the relativistic spectrum condition (positivity of energy in all Lorentz frames) and, in the sense of equality of sets, $U(\lambda)\mathcal{A}(\mathcal{O})U(\lambda)^{-1} = \mathcal{A}(\lambda\mathcal{O})$ for all $\lambda \in \mathcal{P}_+^\uparrow$ and regions $\mathcal{O}$, where $\lambda\mathcal{O}$ denotes the Poincaré transformed region. It is also assumed that the subspace of $U(\mathcal{P}_+^\uparrow)$-invariant vectors is spanned by a single unit vector Ω, representing the vacuum, which has the Reeh–Schlieder property, that is, each set of vectors $\mathcal{A}(\mathcal{O})\Omega$ is dense in $\mathcal{H}$. These standing assumptions will subsequently be amended by further conditions concerning the particle content of the theory.

Haag–Ruelle Theory

Haag and Ruelle were the first to establish the existence of scattering states within this general framework (Jost 1965); further substantial improvements are due to Araki and Hepp (Araki 1999). In all of these investigations, the arguments were given for quantum field theories with associated particles (in the Wigner sense) which have strictly positive mass $m > 0$ and for which m is an isolated eigenvalue of the mass operator (upper and lower mass gap). Moreover, it was assumed that states of a single particle can be created from the vacuum by local operations. In physical terms, these assumptions allow only for theories with short-range interactions and particles carrying strictly localizable charges.

In view of these limitations, Haag–Ruelle theory has been developed in a number of different directions. By now, the scattering theory of massive particles is under complete control, including also

particles carrying nonlocalizable (gauge or topological) charges and particles having exotic statistics (anyons, plektons) which can appear in theories in low spacetime dimensions. Due to constraints of space, these results must go without further mention; we refer the interested reader to the articles Buchholz and Fredenhagen (1982) and Fredenhagen *et al.* (1996). Theories of massless particles and of particles carrying charges of electric or magnetic type (infraparticles) will be discussed in subsequent sections.

We outline here a recent generalization of Haag–Ruelle scattering theory presented in Dybalski (2005), which covers massive particles with localizable charges without relying on any further constraints on the mass spectrum. In particular, the scattering of electrically neutral, stable particles fulfilling a sharp dispersion law in the presence of massless particles is included (e.g., neutral atoms in their ground states). Mathematically, this assumption can be expressed by the requirement that there exists a subspace $\mathcal{H}_1 \subset \mathcal{H}$ such that the restriction of $U(\mathcal{P}_+^\uparrow)$ to $\mathcal{H}_1$ is a representation of mass $m > 0$. We denote by P_1 the projection in $\mathcal{H}$ onto $\mathcal{H}_1$.

To establish notation, let $\mathcal{O}$ be a bounded spacetime region and let $A \in \mathcal{A}(\mathcal{O})$ be any operator such that $P_1 A\Omega \neq 0$. The existence of such localized (in brief, local) operators amounts to the assumption that the particle carries a localizable charge. That the particle is stable, that is, completely decouples from the underlying continuum states, can be cast into a condition first stated by Herbst: for all sufficiently small $\mu > 0$

$$\|E_\mu(1 - P_1)A\Omega\| \leq c\mu^\eta \qquad [3]$$

for some constants $c, \eta > 0$, where E_μ is the projection onto the spectral subspace of the mass operator corresponding to spectrum in the interval $(m - \mu, m + \mu)$. In the case originally considered by Haag and Ruelle, where m is isolated from the rest of the mass spectrum, this condition is certainly satisfied.

Setting $A(x) \doteq U(x)AU(x)^{-1}$, where $U(x)$ is the unitary implementing the spacetime translation $x = (x_0, \boldsymbol{x})$ (the velocity of light and Planck's constant are set equal to 1 in what follows), one puts, for $t \neq 0$,

$$A_t(f) = \int \mathrm{d}^4x\, g_t(x_0) f_{x_0}(\boldsymbol{x}) A(x) \qquad [4]$$

Here $x_0 \mapsto g_t(x_0) \doteq g((x_0 - t)/|t|^\kappa)/|t|^\kappa$ induces a time averaging about t, g being any test function which integrates to 1 and whose Fourier transform has compact support, and $1/(1 + \eta) < \kappa < 1$ with η as above. The Fourier transform of f_{x_0} is given by $\widetilde{f_{x_0}}(\boldsymbol{p}) \doteq \widetilde{f}(\boldsymbol{p})\, \mathrm{e}^{-\mathrm{i}x_0\omega(\boldsymbol{p})}$, where f is some test function on $\mathbb{R}^3$ with $\widetilde{f}(\boldsymbol{p})$ having compact support, and $\omega(\boldsymbol{p}) = (\boldsymbol{p}^2 + m^2)^{1/2}$. Note that $(x_0, \boldsymbol{x}) \mapsto f_{x_0}(\boldsymbol{x})$ is a solution of the Klein–Gordon equation of mass m.

With these assumptions, it follows by a straightforward application of the harmonic analysis of unitary groups that in the sense of strong convergence $A_t(f)\Omega \to P_1 A(f)\Omega$ and $A_t(f)^*\Omega \to 0$ as $t \to \pm\infty$, where $A(f) = \int \mathrm{d}^3x\, f(\boldsymbol{x}) A(0, \boldsymbol{x})$. Hence, the operators $A_t(f)$ may be thought of as creation operators and their adjoints as annihilation operators. These operators are the basic ingredients in the construction of scattering states. Choosing local operators A_k as above and test functions $f^{(k)}$ with disjoint compact supports in momentum space, $k = 1, \ldots, n$, the scattering states are obtained as limits of the Haag–Ruelle approximants

$$A_{1t}(f^{(1)}) \cdots A_{nt}(f^{(n)})\Omega \qquad [5]$$

Roughly speaking, the operators $A_{kt}(f^{(k)})$ are localized in spacelike separated regions at asymptotic times t, due to the support properties of the Fourier transforms of the functions $f^{(k)}$. Hence they commute asymptotically because of locality and, by the clustering properties of the vacuum state, the above vector becomes a product state of single-particle states. In order to prove convergence, one proceeds, in analogy to Cook's method in quantum-mechanical scattering theory, to the time derivatives,

$$\begin{aligned} &\partial_t A_{1t}(f^{(1)}) \cdots A_{nt}(f^{(n)})\Omega \\ &\quad = \sum_{k \neq l} A_{1t}(f^{(1)}) \cdots [\partial_t A_{kt}(f^{(k)}), A_{lt}(f^{(l)})] \cdots A_{nt}(f^{(n)})\Omega \\ &\qquad + \sum_k A_{1t}(f^{(1)}) \cdots \overset{k}{\vee} \cdots A_{nt}(f^{(n)}) \partial_t A_{kt}(f^{(k)})\Omega \end{aligned} \qquad [6]$$

where $\overset{k}{\vee}$ denotes omission of $A_{kt}(f^{(k)})$. Employing techniques of Araki and Hepp, one can prove that the terms in the first summation on the right-hand side (RHS) of [6], involving commutators, decay rapidly in norm as t approaches infinity because of locality, as indicated above. By applying condition [3] and the fact that the vectors $\partial_t A_{kt}(f^{(k)})\Omega$ do not have a component in the single-particle space $\mathcal{H}_1$, the terms in the second summation on the RHS of [6] can be shown to decay in norm like $|t|^{-\kappa(1+\eta)}$. Thus, the norm of the vector [6] is integrable in t, implying the existence of the strong limits

$$\begin{aligned} &\left(P_1 A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1 A_n(f^{(n)})\Omega\right)^{\text{in/out}} \\ &\quad \doteq \lim_{t \to \mp\infty} A_{1t}(f^{(1)}) \cdots A_{nt}(f^{(n)})\Omega \end{aligned} \qquad [7]$$

As indicated by the notation, these limits depend only on the single-particle vectors $P_1A_k(f^{(k)})\Omega \in \mathcal{H}_1$, $k = 1, \ldots, n$, but not on the specific choice of operators and test functions. In order to establish their Fock structure, one employs results on clustering properties of vacuum correlation functions in theories without strictly positive minimal mass. Using this, one can compute inner products of arbitrary asymptotic states and verify that the maps

$$\Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1A_n(f^{(n)})\Omega\Big) \mapsto \Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} \quad [8]$$

extend by linearity to isomorphisms $\Omega^{\text{in/out}}$ from the Fock space $\mathcal{H}_F$ onto the subspaces $\mathcal{H}^{\text{in/out}} \subset \mathcal{H}$ generated by the collision states. Moreover, the asymptotic states transform under the Poincaré transformations $U(\mathcal{P}_+^\uparrow)$ as

$$U(\lambda)\Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} = \Big(U_1(\lambda)P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes U_1(\lambda)P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} \quad [9]$$

Thus, the isomorphisms $\Omega^{\text{in/out}}$ intertwine the action of the Poincaré group on $\mathcal{H}_F$ and $\mathcal{H}^{\text{in/out}}$. We summarize these results, which are vital for the physical interpretation of the underlying theory, in the following theorem.

Theorem 1 *Consider a theory of a particle of mass $m > 0$ which satisfies the standing assumptions and the stability condition [3]. Then there exist canonical isometries $\Omega^{\text{in/out}}$, mapping the Fock space $\mathcal{H}_F$ based on the single-particle space $\mathcal{H}_1$ onto subspaces $\mathcal{H}^{\text{in/out}} \subset \mathcal{H}$ of incoming and outgoing scattering states. Moreover, these isometries intertwine the action of the Poincaré transformations on the respective spaces.*

Since the scattering states have been identified with Fock space, asymptotic creation and annihilation operators act on $\mathcal{H}^{\text{in/out}}$ in a natural manner. This point will be explained in the following section.

LSZ Formalism

Prior to the results of Haag and Ruelle, an axiomatic approach to scattering theory was developed by Lehmann, Symanzik, and Zimmermann (LSZ), based on time-ordered vacuum expectation values of quantum fields. The relative advantage of their approach with respect to Haag–Ruelle theory is that useful reduction formulas for the S-matrix greatly facilitate computations, in particular in perturbation theory. Moreover, these formulas are the starting point of general studies of the momentum space analyticity properties of the S-matrix (dispersion relations), as outlined in Dispersion Relations (cf. also Iagolnitzer (1993)). Within the present general setting, the LSZ method was established by Hepp.

For simplicity of discussion, we consider again a single particle type of mass $m > 0$ and integer spin s, subject to condition [3]. According to the results of the preceding section, one then can consistently define asymptotic creation operators on the scattering states, setting

$$\begin{aligned} &A(f)^{\text{in/out}}\Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} \\ &\doteq \lim_{t \to \mp\infty} A_t(f)\Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} \\ &= \Big(P_1A(f)\Omega \otimes P_1A_1(f^{(1)})\Omega \otimes \cdots \\ &\quad \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} \end{aligned} \quad [10]$$

Similarly, one obtains the corresponding asymptotic annihilation operators,

$$\begin{aligned} &A(f)^{\text{in/out}*}\Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} \\ &= \lim_{t \to \mp\infty} A_t(f)^*\Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \\ &\quad \otimes P_1A_n(f^{(n)})\Omega\Big)^{\text{in/out}} = 0 \end{aligned} \quad [11]$$

where the latter equality holds if the Fourier transforms of the functions $f, f^{(1)}, \ldots, f^{(n)}$, have disjoint supports. We mention as an aside that, by replacing the time-averaging function g in the definition of $A_t(f)$ by a delta function, the above formulas still hold. But the convergence is then to be understood in the weak Hilbert space topology. In this form, the above relations were anticipated by LSZ (asymptotic condition).

It is straightforward to proceed from these relations to reduction formulas. Let B be any local operator. Then one has, in the sense of matrix elements between outgoing and incoming scattering states,

$$\begin{aligned} &BA(f)^{\text{in}} - A(f)^{\text{out}}B = \lim_{t \to \infty}(BA(f_{-t}) - A(f_t)B) \\ &= \lim_{t \to \infty}\left(\int \mathrm{d}^4x f_{-t}(x)BA(x) - \int \mathrm{d}^4x f_t(x)A(x)B\right) \end{aligned} \quad [12]$$

$f_t(x) \doteq g_t(x_0)f(x_0)(\text{vec}(x))$. Because of the (essential) support properties of the functions $f_{\pm t}$, the contributions to the latter integrals arise, for asymptotic t, from spacetime points x where the localization

regions of $A(x)$ and B have a negative timelike (first term), respectively, positive timelike (second term) distance. One may therefore proceed from the products of these operators to the time-ordered products $T(BA(x))$, where $T(BA(x)) = A(x)B$ if the localization region of $A(x)$ lies in the future of that of B, and $T(BA(x)) = BA(x)$ if it lies in the past. It is noteworthy that a precise definition of the time ordering for finite x is irrelevant in the present context – any reasonable interpolation between the above relations will do. Similarly, one can define time-ordered products for an arbitrary number of local operators. The preceding limit can then be recast into

$$\lim_{t\to\infty} \int \mathrm{d}^4x (f_{-t}(x) - f_t(x)) T(BA(x)) \qquad [13]$$

The latter expression has a particularly simple form in momentum space. Proceeding to the Fourier transforms of $f_{\pm t}$ and noticing that, in the limit of large t,

$$\left(\widetilde{f_{-t}}(p) - \widetilde{f_t}(p)\right)/(p_0 - \omega(\boldsymbol{p})) \longrightarrow -2\pi\mathrm{i}\widetilde{f}(\boldsymbol{p})\,\delta(p_0 - \omega(\boldsymbol{p})) \qquad [14]$$

one gets

$$BA(f)^{\mathrm{in}} - A(f)^{\mathrm{out}}B = -2\pi\mathrm{i}\int \mathrm{d}^3p \widetilde{f}(\boldsymbol{p})(p_0 - \omega(\boldsymbol{p})) \times T(B\widetilde{A}(-p))\Big|_{p_0=\omega(\boldsymbol{p})} \qquad [15]$$

Here $T(B\widetilde{A}(p))$ denotes the Fourier transform of $T(BA(x))$, and it can be shown that the restriction of $(p_0 - \omega(\boldsymbol{p}))T(B\widetilde{A}(-p))$ to the manifold $\{p \in \mathbb{R}^4 \colon p_0 = \omega(\boldsymbol{p})\}$ (the "mass shell") is meaningful in the sense of distributions on $\mathbb{R}^3$. By the same token, one obtains

$$A(f)^{\mathrm{out}*}B - BA(f)^{\mathrm{in}*} = -2\pi\mathrm{i}\int \mathrm{d}^3p \overline{\widetilde{f}(\boldsymbol{p})}\,(p_0 - \omega(\boldsymbol{p}))T(\widetilde{A^*}(p)B)\Big|_{p_0=\omega(\boldsymbol{p})} \qquad [16]$$

Similar relations, involving an arbitrary number of asymptotic creation and annihilation operators, can be established by analogous considerations. Taking matrix elements of these relations in the vacuum state and recalling the action of the asymptotic creation and annihilation operators on scattering states, one arrives at the following result, which is central in all applications of scattering theory.

Theorem 2 *Consider the theory of a particle of mass $m > 0$ subject to the conditions stated in the preceding sections and let $f^{(1)}, \ldots, f^{(n)}$ be any family of test functions whose Fourier transforms have compact and nonoverlapping supports. Then*

$$\begin{aligned}
&\Big\langle \Big(P_1A_1(f^{(1)})\Omega \otimes \cdots \otimes P_1A_k(f^{(k)})\Omega\Big)^{\mathrm{out}}, \\
&\quad \Big(P_1A_{k+1}(f^{(k+1)})\Omega \otimes \cdots \otimes P_1A_n(f^{(n)})\Omega\Big)^{\mathrm{in}}\Big\rangle \\
&= (-2\pi\mathrm{i})^n \int \cdots \int \mathrm{d}^3p_1 \cdots \mathrm{d}^3p_n \overline{\widetilde{f^{(1)}}(\boldsymbol{p}_1)} \cdots \\
&\quad \times \overline{\widetilde{f^{(k)}}(\boldsymbol{p}_k)}\widetilde{f^{(k+1)}}(\boldsymbol{p}_{k+1}) \cdots \widetilde{f^{(n)}}(\boldsymbol{p}_n) \\
&\quad \times \prod_{i=1}^{n}(p_{i_0} - \omega(\boldsymbol{p}_i))\Big\langle \Omega, T\Big(\widetilde{A_1^*}(p_1) \cdots \\
&\quad \times \widetilde{A_k^*}(p_k)\widetilde{A_{k+1}}(-p_{k+1}) \cdots \\
&\quad \times \widetilde{A_n}(-p_n)\Omega\Big)\Big\rangle\Big|_{p_{i_0}=\omega(\boldsymbol{p}_i)}^{j=1,\ldots,n}
\end{aligned} \qquad [17]$$

in an obvious notation.

Thus, the kernels of the scattering amplitudes in momentum space are obtained by restricting the (by the factor $\prod_{i=1}^{n}(p_{i_0} - \omega(\boldsymbol{p}_i))$) amputated Fourier transforms of the vacuum expectation values of the time-ordered products to the positive and negative mass shells, respectively. These are the famous LSZ reduction formulas, which provide a convenient link between the time-ordered (Green's) functions of a theory and its asymptotic particle interpretation.

Asymptotic Particle Counters

The preceding construction of scattering states applies to a significant class of theories; but even if one restricts attention to the case of massive particles, it does not cover all situations of physical interest. For an essential input in the construction is the existence of local operators interpolating between the vacuum and the single-particle states. There may be no such operators at one's disposal, however, either because the particle in question carries a nonlocalizable charge, or because the given family of operators is too small. The latter case appears, for example, in gauge theories, where in general only the observables are fixed by the principle of local gauge invariance, and the physical particle content as well as the corresponding interpolating operators are not known from the outset. As observables create from the vacuum only neutral states, the above construction of scattering states then fails if charged particles are present. Nevertheless, thinking in physical terms, one would expect that the observables contain all relevant information in order to determine the features of scattering states, in particular their collision cross section. That this is indeed the case was first shown by Araki and Haag (Araki 1999).

In scattering experiments, the measured data are provided by detectors (e.g., particle counters) and

coincidence arrangements of detectors. Essential features of detectors are their lack of response in the vacuum state and their macroscopic localization. Hence, within the present mathematical setting, a general detector is represented by a positive operator C on the physical Hilbert space $\mathcal{H}$ such that $C\Omega = 0$. Because of the Reeh–Schlieder theorem, these conditions cannot be satisfied by local operators. However, they can be fulfilled by "almost-local" operators. Examples of such operators are easy to produce, putting $C = L^*L$ with

$$L = \int \mathrm{d}^4x\, f(x)\, A(x) \qquad [18]$$

where A is any local operator and f any test function whose Fourier transform has compact support in the complement of the closed forward light cone (and hence in the complement of the energy momentum spectrum of the theory). In view of the properties of f and the invariance of Ω under translations, it follows that $C = L^*L$ annihilates the vacuum and can be approximated with arbitrary precision by local operators. The algebra generated by these operators C will be denoted by $\mathcal{C}$.

When preparing a scattering experiment, the first thing one must do with a detector is to calibrate it, that is, test its response to sources of single-particle states. Within the mathematical setting, this amounts to computing the matrix elements of C in states $\Phi \in \mathcal{H}_1$:

$$\langle \Phi, C\Phi \rangle = \int\int \mathrm{d}^3p\, \mathrm{d}^3q\, \overline{\Phi(\boldsymbol{p})}\, \Phi(\boldsymbol{q})\, \langle \boldsymbol{p}|C|\boldsymbol{q}\rangle \qquad [19]$$

Here $\boldsymbol{p} \mapsto \Phi(\boldsymbol{p})$ is the momentum space wave function of Φ, $\langle \cdot|C|\cdot\rangle$ is the kernel of C in the single-particle space $\mathcal{H}_1$, and we have omitted (summations over) indices labeling internal degrees of freedom of the particle, if any. The relevant information about C is encoded in its kernel. As a matter of fact, one only needs to know its restriction to the diagonal, $\boldsymbol{p} \mapsto \langle \boldsymbol{p}|C|\boldsymbol{p}\rangle$. It is called the sensitivity function of C and can be shown to be regular under quite general circumstances (Araki 1999, Buchholz and Fredenhagen 1982).

Given a state $\Psi \in \mathcal{H}$ for which the expectation value $\langle \Psi, C(x))\Psi \rangle$ differs significantly from 0, one concludes that this state deviates from the vacuum in a region about x. For finite x, this does not mean, however, that Ψ has a particle interpretation at x. For that spacetime point may, for example, be just the location of a collision center. Yet, if one proceeds to asymptotic times, one expects, in view of the spreading of wave packets, that the probability of finding two or more particles in the same spacetime region is dominated by the single-particle contributions. It is this physical insight which justifies the expectation that the detectors $C(x)$ become particle counters at asymptotic times. Accordingly, one considers for asymptotic t the operators

$$C_t(h) \doteq \int \mathrm{d}^3x\, h(\boldsymbol{x}/t) C(t, \boldsymbol{x}) \qquad [20]$$

where h is any test function on $\mathbb{R}^3$. The role of the integral is to sum up all single-particle contributions with velocities in the support of h in order to compensate for the decreasing probability of finding such particles at asymptotic times t about the localization center of the detector. That these ideas are consistent was demonstrated by Araki and Haag, who established the following result (Araki 1999).

Theorem 3 *Consider, as before, the theory of a massive particle. Let $C^{(1)}, \ldots, C^{(n)} \in \mathcal{C}$ be any family of detector operators and let $h^{(1)}, \ldots, h^{(n)}$ be any family of test functions on $\mathbb{R}^3$. Then, for any state $\Psi^{\mathrm{out}} \in \mathcal{H}^{\mathrm{out}}$ of finite energy,*

$$\begin{aligned} &\lim_{t\to\infty} \left\langle \Psi^{\mathrm{out}}, C_t^{(1)}(h^{(1)}) \cdots C_t^{(n)}(h^{(n)}) \Psi^{\mathrm{out}} \right\rangle \\ &\quad = \int \cdots \int \mathrm{d}^3p_1 \cdots \mathrm{d}^3p_n\, \langle \Psi^{\mathrm{out}}, \rho^{\mathrm{out}}(\boldsymbol{p}_1) \cdots \rho^{\mathrm{out}}(\boldsymbol{p}_n) \Psi^{\mathrm{out}} \rangle \\ &\qquad \times \prod_{k=1}^{n} h(\boldsymbol{p}_k/\omega(\boldsymbol{p}_k)) \langle \boldsymbol{p}_k|C^{(k)}|\boldsymbol{p}_k\rangle \end{aligned} \qquad [21]$$

where $\rho^{\mathrm{out}}(\boldsymbol{p})$ is the momentum space density (the product of creation and annihilation operators) of outgoing particles of momentum $\boldsymbol{p}$, and (summations over) possible indices labeling internal degrees of freedom of the particle are omitted. An analogous relation holds for incoming scattering states at negative asymptotic times.

This result shows, first of all, that the scattering states have indeed the desired interpretation with regard to the observables, as anticipated in the preceding sections. Since the assertion holds for all scattering states of finite energy, one may replace in the above theorem the outgoing scattering states by any state of finite energy, if the theory is asymptotically complete, that is, $\mathcal{H} = \mathcal{H}^{\mathrm{in}} = \mathcal{H}^{\mathrm{out}}$. Then choosing, in particular, any incoming scattering state and making use of the arbitrariness of the test functions $h^{(k)}$ as well as the knowledge of the sensitivity functions of the detector operators, one can compute the probability distributions of outgoing particle momenta in this state, and thereby the corresponding collision cross sections.

The question of how to construct certain specific incoming scattering states by using only local observables was not settled by Araki and Haag,

however. A general method to that effect was outlined in Buchholz *et al.* (1991). As a matter of fact, for that method only the knowledge of states in the subspace of neutral states is required. Yet in this approach one would need for the computation of, say, elastic collision cross sections of charged particles the vacuum correlation functions involving at least eight local observables. This practical disadvantage of increased computational complexity of the method is offset by the conceptual advantage of making no appeal to quantities which are *a priori* nonobservable.

Massless Particles and Huygens' Principle

The preceding general methods of scattering theory apply only to massive particles. Yet taking advantage of the salient fact that massless particles always move with the speed of light, Buchholz succeeded in establishing a scattering theory also for such particles (Haag 1992). Moreover, his arguments lead to a quantum version of Huygens' principle.

As in the case of massive particles, one assumes that there is a subspace $\mathcal{H}_1 \subset \mathcal{H}$ corresponding to a representation of $U(\mathcal{P}_+^\uparrow)$ of mass $m = 0$ and, for simplicity, integer helicity; moreover, there must exist local operators interpolating between the vacuum and the single-particle states. These assumptions cover, in particular, the important examples of the photon and of Goldstone particles. Picking any suitable local operator A interpolating between Ω and some vector in $\mathcal{H}_1$, one sets, in analogy to [4],

$$A_t \doteq \int \mathrm{d}^4x\, g_t(x_0) \times (-1/2\pi)\varepsilon(x_0)\,\delta(x_0^2 - \boldsymbol{x}^2)\partial_0 A(x) \qquad [22]$$

Here $g_t(x_0) \doteq (1/|\ln t|)\, g((x_0 - t)/|\ln t|)$ with g as in [4], and the solution of the Klein–Gordon equation in [4] has been replaced by the fundamental solution of the wave equation; furthermore, $\partial_0 A(x)$ denotes the derivative of $A(x)$ with respect to x_0. Then, once again, the strong limit of $A_t\Omega$ as $t \to \pm\infty$ is $P_1 A\Omega$, with P_1 the projection onto $\mathcal{H}_1$.

In order to establish the convergence of A_t as in the LSZ approach, one now uses the fact that these operators are, at asymptotic times t, localized in the complement of some forward, respectively, backward, light cone. Because of locality, they therefore commute with all operators which are localized in the interior of the respective cones. More specifically, let $\mathcal{O} \subset \mathbb{R}^4$ be the localization region of A and let $\mathcal{O}_\pm \subset \mathbb{R}^4$ be the two regions having a positive, respectively, negative, timelike distance from all points in $\mathcal{O}$. Then, for any operator B which is compactly localized in $\mathcal{O}_\pm$, respectively, one obtains $\lim_{t\to\pm\infty} A_t B\Omega = \lim_{t\to\pm\infty} BA_t\Omega = BP_1 A\Omega$. This relation establishes the existence of the limits

$$A^{\mathrm{in/out}} = \lim_{t\to\mp\infty} A_t \qquad [23]$$

on the (by the Reeh–Schlieder property) dense sets of vectors $\{B\Omega : B \in \mathcal{A}(\mathcal{O}_\mp)\} \subset \mathcal{H}$. It requires some more detailed analysis to prove that the limits have all of the properties of a (smeared) free massless field, whose translates $x \mapsto A^{\mathrm{in/out}}(x)$ satisfy the wave equation and have c-number commutation relations. From these free fields, one can then proceed to asymptotic creation and annihilation operators and construct asymptotic Fock spaces $\mathcal{H}^{\mathrm{in/out}} \subset \mathcal{H}$ of massless particles and a corresponding scattering matrix as in the massive case. The details of this construction can be found in the original article, cf. Haag (1992).

It also follows from these arguments that the asymptotic fields $A^{\mathrm{in/out}}$ of massless particles emanating from a region $\mathcal{O}$, that is, for which the underlying interpolating operators A are localized in $\mathcal{O}$, commute with all operators localized in $\mathcal{O}_\mp$, respectively. This result may be understood as an expression of Huygens' principle. More precisely, denoting by $\mathcal{A}^{\mathrm{in/out}}(\mathcal{O})$ the algebras of bounded operators generated by the asymptotic fields $A^{\mathrm{in/out}}$, respectively, one arrives at the following quantum version of Huygens' principle.

Theorem 4 *Consider a theory of massless particles as described above and let* $\mathcal{A}^{\mathrm{in/out}}(\mathcal{O})$ *be the algebras generated by massless asymptotic fields* $A^{\mathrm{in/out}}$ *with* $A \in \mathcal{A}(\mathcal{O})$. *Then*

$$\mathcal{A}^{\mathrm{in}}(\mathcal{O}) \subset \mathcal{A}(\mathcal{O}_-)' \quad \text{and} \quad \mathcal{A}^{\mathrm{out}}(\mathcal{O}) \subset \mathcal{A}(\mathcal{O}_+)' \qquad [24]$$

Here the prime denotes the set of bounded operators commuting with all elements of the respective algebras (*i.e.*, *their commutants*).

Beyond Wigner's Concept of Particle

There is by now ample evidence that Wigner's concept of particle is too narrow in order to cover all particle-like structures appearing in quantum field theory. Examples are the partons which show up in nonabelian gauge theories at very small spacetime scales as constituents of hadrons, but which do not appear at large scales due to the confining forces. Their mathematical description

requires a quite different treatment, which cannot be discussed here. But even at large scales, Wigner's concept does not cover all stable particle-like systems, the most prominent examples being particles carrying an abelian gauge charge, such as the electron and the proton, which are inevitably accompanied by infinite clouds of ("on-shell") massless particles.

The latter problem was discussed first by Schroer, who coined the term "infraparticle" for such systems. Later, Buchholz showed in full generality that, as a consequence of Gauss' law, pure states with an abelian gauge charge can neither have a sharp mass nor carry a unitary representation of the Lorentz group, thereby uncovering the simple origin of results found by explicit computations, notably in quantum electrodynamics (Steinmann 2000). Thus, one is faced with the question of an appropriate mathematical characterization of infraparticles which generalizes the concept of particle invented by Wigner. Some significant steps in this direction were taken by Fröhlich, Morchio, and Strocchi, who based a definition of infraparticles on a detailed spectral analysis of the energy–momentum operators. For an account of these developments and further references, cf. Haag (1992).

We outline here an approach, originated by Buchholz, which covers all stable particle-like structures appearing in quantum field theory at asymptotic times. It is based on Dirac's idea of improper particle states with sharp energy and momentum. In the standard (rigged Hilbert space) approach to giving mathematical meaning to these quantities, one regards them as vector-valued distributions, whereby one tacitly assumes that the improper states can coherently be superimposed so as to yield normalizable states. This assumption is valid in the case of Wigner particles but fails in the case of infraparticles. A more adequate method of converting the improper states into normalizable ones is based on the idea of acting on them with suitable localizing operators. In the case of quantum mechanics, one could take as a localizing operator any sufficiently rapidly decreasing function of the position operator. It would map the improper "plane-wave states" of sharp momentum into finitely localized states which thereby become normalizable. In quantum mechanics, these two approaches can be shown to be mathematically equivalent. The situation is different, however, in quantum field theory.

In quantum field theory, the appropriate localizing operators L are of the form [18]. They constitute a (nonclosed) left ideal $\mathcal{L}$ in the C*-algebra $\mathcal{A}$ generated by all local operators. Improper particle states of sharp energy–momentum p can then be defined as linear maps $|\cdot\rangle_p : \mathcal{L} \to \mathcal{H}$ satisfying

$$U(x)|L\rangle_p = \mathrm{e}^{\mathrm{i}px}|L(x)\rangle_p, \quad L \in \mathcal{L} \qquad [25]$$

It is instructive to (formally) replace L here by the identity operator, making it clear that this relation indeed defines improper states of sharp energy–momentum.

In theories of massive particles, one can always find localizing operators $L \in \mathcal{L}$ such that their images $|L\rangle_p \in \mathcal{H}$ are states with a sharp mass. This is the situation covered in Wigner's approach. In theories with long-range forces there are, in general, no such operators, however, since the process of localization inevitably leads to the production of low-energy massless particles. Yet improper states of sharp momentum still exist in this situation, thereby leading to a meaningful generalization of Wigner's particle concept.

That this characterization of particles covers all situations of physical interest can be justified in the general setting of relativistic quantum field theory as follows. Picking g_t as in [4] and any vector $\Psi \in \mathcal{H}$ with finite energy, one can show that the functionals $\rho_t, t \in \mathbb{R}$, given by

$$\rho_t(L^*L) \doteq \int \mathrm{d}^4x\, g_t(x_0)\, \langle \Psi, (L^*L)(x)\Psi \rangle, \quad L \in \mathcal{L} \qquad [26]$$

are well defined and form an equicontinuous family with respect to a certain natural locally convex topology on the algebra $\mathcal{C} = \mathcal{L}^*\mathcal{L}$. This family of functionals therefore has, as $t \to \pm\infty$, weak-* limit points, denoted by σ. The functionals σ are positive on $\mathcal{C}$ but not normalizable. (Technically speaking, they are weights on the underlying algebra $\mathcal{A}$.) Any such σ induces a positive-semidefinite scalar product on the left ideal $\mathcal{L}$ given by

$$\langle L_1 \,|\, L_2 \rangle \doteq \sigma(L_1^* L_2), \quad L_1, L_2 \in \mathcal{L} \qquad [27]$$

After quotienting out elements of zero norm and taking the completion, one obtains a Hilbert space and a linear map $L \mapsto |L\rangle$ from $\mathcal{L}$ into that space. Moreover, the spacetime translations act on this space by a unitary representation satisfying the relativistic spectrum condition.

It is instructive to compute these functionals and maps in theories of massive particles. Making use of relation [21] one obtains, with a slight change of notation,

$$\langle L_1 \,|\, L_2 \rangle = \int \mathrm{d}\mu(p)\, \langle p \,|\, L_1^* L_2 | p \,\rangle \qquad [28]$$

where μ is a measure giving the probability density of finding at asymptotic times in state Ψ a particle of energy–momentum p. Once again, possible summations over different particle types and internal degrees of freedom have been omitted here. Thus,

setting $|L\rangle_p \doteq L|p\rangle$, one concludes that the map $L \mapsto |L\rangle$ can be decomposed into a direct integral of improper particle states of sharp energy–momentum, $|\cdot\rangle = \int_{\oplus} \mathrm{d}\mu(p)^{1/2}|\cdot\rangle_p$. It is crucial that this result can also be established without any *a priori* input about the nature of the particle content of the theory, thereby providing evidence of the universal nature of the concept of improper particle states of sharp momentum, as outlined here.

Theorem 5 *Consider a relativistic quantum field theory satisfying the standing assumptions. Then the maps $L \mapsto |L\rangle$ defined above can be decomposed into improper particle states of sharp energy–momentum p,*

$$|\cdot\rangle = \int_{\oplus} \mathrm{d}\mu(p)^{1/2}|\cdot\rangle_p \qquad [29]$$

where μ is some measure depending on the state Ψ and the respective time limit taken.

It is noteworthy that whenever the space of improper particle states corresponding to fixed energy–momentum p is finite dimensional (finite particle multiplets), then in the corresponding Hilbert space there exists a continuous unitary representation of the little group of p. This implies that improper momentum eigenstates of mass $m = (p^2)^{1/2} > 0$ carry definite (half)integer spin, in accordance with Wigner's classification. However, if $m = 0$, the helicity need not be quantized, in contrast to Wigner's results.

Though a general scattering theory based on improper particle states has not yet been developed, some progress has been made in Buchholz *et al.* (1991). There it is outlined how inclusive collision cross sections of scattering states, where an undetermined number of low-energy massless particles remains unobserved, can be defined in the presence of long-range forces, in spite of the fact that a meaningful scattering matrix may not exist.

Asymptotic Completeness

Whereas the description of the asymptotic particle features of any relativistic quantum field theory can be based on an arsenal of powerful methods, the question of when such a theory has a complete particle interpretation remains open to date. Even in concrete models there exist only partial results, cf. Iagolnitzer (1993) for a comprehensive review of the current state of the art. This situation is in striking contrast to the case of quantum mechanics, where the problem of asymptotic completeness has been completely settled.

One may trace the difficulties in quantum field theory back to the possible formation of superselection sectors (Haag 1992) and the resulting complex particle structures, which cannot appear in quantum-mechanical systems with a finite number of degrees of freedom. Thus, the first step in establishing a complete particle interpretation in a quantum field theory has to be the determination of its full particle content. Here the methods outlined in the preceding section provide a systematic tool. From the resulting data, one must then reconstruct the full physical Hilbert space of the theory comprising all superselection sectors. For theories in which only massive particles appear, such a construction has been established in Buchholz and Fredenhagen (1982), and it has been shown that the resulting Hilbert space contains all scattering states. The question of completeness can then be recast into the familiar problem of the unitarity of the scattering matrix. It is believed that phase space (nuclearity) properties of the theory are of relevance here (Haag 1992).

However, in theories with long-range forces, where a meaningful scattering matrix may not exist, this strategy is bound to fail. Nonetheless, as in most high-energy scattering experiments, only some very specific aspects of the particle interpretation are really tested – one may think of other meaningful formulations of completeness. The interpretation of most scattering experiments relies on the existence of conservation laws, such as those for energy and momentum. If a state has a complete particle interpretation, it ought to be possible to fully recover its energy, say, from its asymptotic particle content, that is, there should be no contributions to its total energy which do not manifest themselves asymptotically in the form of particles. Now the mean energy–momentum of a state $\Psi \in \mathcal{H}$ is given by $\langle \Psi, P\Psi \rangle$, P being the energy–momentum operators, and the mean energy–momentum contained in its asymptotic particle content is $\int \mathrm{d}\mu(p)p$, where μ is the measure appearing in the decomposition [29]. Hence, in case of a complete particle interpretation, the following should hold:

$$\langle \Psi, P\Psi \rangle = \int \mathrm{d}\mu(p)p \qquad [30]$$

Similar relations should also hold for other conserved quantities which can be attributed to particles, such as charge, spin, etc. It seems that such a weak condition of asymptotic completeness suffices for a consistent interpretation of most scattering experiments. One may conjecture that relation [30] and its generalizations hold in all theories admitting a local stress–energy tensor and local currents corresponding to the charges.

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Dispersion Relations; Perturbation Theory and its Techniques; Quantum Chromodynamics; Quantum Field Theory in Curved

Spacetime; Quantum Mechanical Scattering Theory; Scattering, Asymptotic Completeness and Bound States; Scattering in Relativistic Quantum Field Theory: The Analytic Program.

Further Reading

Araki H (1999) *Mathematical Theory of Quantum Fields*. Oxford: Oxford University Press.

Buchholz D and Fredenhagen K (1982) Locality and the structure of particle states. *Communications in Mathematical Physics* 84: 1–54.

Buchholz D, Porrmann M, and Stein U (1991) Dirac versus Wigner: towards a universal particle concept in quantum field theory. *Physics Letters B* 267: 377–381.

Dybalski W (2005) Haag–Ruelle scattering theory in presence of massless particles. *Letters in Mathematical Physics* 72: 27–38.

Fredenhagen K, Gaberdiel MR, and Ruger SM (1996) Scattering states of plektons (particles with braid group statistics) in $(2+1)$ dinemsional quantum field theory. *Communications in Mathematical Physics* 175: 319–336.

Haag R (1992) *Local Quantum Physics*. Berlin: Springer.

Iagolnitzer D (1993) *Scattering in Quantum Field Theories*. Princeton, NJ: Princeton University Press.

Jost R (1965) *General Theory of Quantized Fields*. Providence, RI: American Mathematical Society.

Steinmann O (2000) *Perturbative Quantum Electrodynamics and Axiomatic Field Theory*. Berlin: Springer.

Streater RF and Wightman AS (1964) *PCT, Spin and Statistics, and All That*. Reading, MA: Benjamin/Cummings.

Scattering in Relativistic Quantum Field Theory: The Analytic Program

J Bros, CEA/DSM/SPhT, CEA Saclay, Gif-sur-Yvette, France

Introduction to the Analytic Structures of Quantum Field Theory

The importance of complex variables and of the concept of analyticity in theoretical physics finds one of its best illustrations in the analytic structure of relativistic quantum field theory (QFT). The latter have been investigated from several viewpoints in the last 50 years, according to the successive progress in QFT.

In the two main axiomatic frameworks of QFT, namely the one based on Wightman axioms (for a short presentation, *see* Dispersion Relations and also Axiomatic Quantum Field Theory) and the Haag, Kastler, and Araki theory of "local observables" (*see* Algebraic Approach to Quantum Field Theory), there are general justifications of analyticity properties for relevant "N-point structure functions" both in complexified spacetime variables and in complexified energy–momentum variables.

In the Wightman framework, relativistic quantum fields are operator-valued distributions $\Phi_j(x)$ on four-dimensional Minkowski spacetime that transform covariantly under a unitary representation of the Poincaré group in the Hilbert space of states. The basic quantities of QFT are (tempered) distributions on $\boldsymbol{R}^{4N}$ of the form $<\Psi, \Phi(x_1)\cdots\Phi(x_N)\Psi'>$, which depend on pairs of states Ψ, Ψ', belonging to the Hilbert space of the QFT considered: they can be called N-point structure functions of the field Φ "in x-space," namely in Minkowski spacetime (here, for brevity, we assume that the system is defined in terms of a single quantum field). In parallel, it is important to consider the Fourier transform $\tilde{\Phi}(p) = \int \mathrm{e}^{\mathrm{i}p\cdot x}\Phi(x)\,\mathrm{d}x$ of the field in the Minkowskian energy–momentum space ($p\cdot x \doteq p_0x_0 - \boldsymbol{p}\cdot\boldsymbol{x}$ denoting the Minkowskian scalar product). The corresponding quantities $<\Psi, \tilde{\Phi}(p_1)\cdots\tilde{\Phi}(p_N)\Psi'>$, can then be called N-point structure functions of the field Φ "in p-space," namely in energy–momentum space.

In the algebraic QFT framework, each basic local observable B affiliated to a certain bounded region of spacetime $\mathcal{O}$ generates a Haag–Kastler–Araki quantum field $B(x)$ by the action of the translations of spacetime, namely $B(x) \doteq U(x)BU(x)^{-1}$. Here $U(x)$ denotes the unitary representation of the group of spacetime translations in the Hilbert space of states: $B(x)$ is affiliated to the translated region $\mathcal{O}(x) = \{y; y - x \in \mathcal{O}\}$. Then again one can consider N-point structure functions of the theory of the form $<\Psi, B(x_1)\cdots B(x_N)\Psi'>$ and $<\Psi, \tilde{B}(p_1)\cdots\tilde{B}(p_N)\Psi'>$.

To summarize the situation as it occurs in both cases, one can say the following:

1. A certain postulate of relativistic causality implies the analyticity of structure functions of a certain class, often called "Green functions," in the complex energy–momentum variables $k_j = p_j + \mathrm{i}q_j$, in particular for purely imaginary energies.
2. "Stability properties" of the states Ψ, Ψ' such as a "bounded energy content" of these states imply

the analyticity of the previous structure functions in the complex spacetime variables, in particular for purely imaginary times.

In both cases, analyticity is obtained as a basic property of the Fourier–Laplace transformation in several variables. Let V^+ denote the forward cone of the Minkowskian space ($V^+ \doteq -V^- \doteq \{x;\, x^2 \doteq x \cdot x > 0,\, x_0 > 0\}$) and let

$$\tilde{f}(p + \mathrm{i}q) = \int_{V_a^+} \mathrm{e}^{\mathrm{i}(p+iq)\cdot x}\, f(x)\mathrm{d}x \qquad [1]$$

$$g(x + \mathrm{i}y) = (2\pi)^{-4} \int_{V_p^+} \mathrm{e}^{-\mathrm{i}p(x+iy)} \tilde{g}(p)\mathrm{d}p \qquad [2]$$

be the associated reciprocal Fourier formulas, applied, respectively, to functions $f(x)$ with support contained in the translated forward cone $V_a^+ = -a + V^+$, $a \in V^+$ (or in its closure), and to functions $\tilde{g}(p)$ with support contained in the translated forward cone $V_P^+ = -P + V^+$, $P \in V^+$ of energy–momentum space (or in its closure). Then in view of the convergence properties of the previous integrals, one easily checks that $\tilde{f}(k)$ is holomorphic with possible exponential increase in the imaginary directions controlled by the bound $\mathrm{e}^{q\cdot a}$ in the tube domain $\mathcal{T}^+ = \boldsymbol{R}^4 + \mathrm{i}V^+$; similarly, $g(z)$ is holomorphic with an increase controlled by the exponential bound $\mathrm{e}^{-y\cdot P}$ in the tube domain $\mathcal{T}^- = \boldsymbol{R}^4 + \mathrm{i}V^-$.

On the one hand, for each N the structure functions $<\Psi, \tilde{\Phi}(p_1)\cdots\tilde{\Phi}(p_N)\Psi'>$ (or $<\Psi, \tilde{B}(p_1)\cdots\tilde{B}(p_N)\Psi'>$) have conical support properties of the previous type in the variables p_j, as a consequence of the relativistic shape of the energy–momentum spectrum. In both axiomatic frameworks, in fact, one postulates that there is a state of zero energy–momentum Ω, called the vacuum, and that the energy–momentum spectrum Σ, namely the joint spectrum of the generators P_μ of the Lie algebra of the group $U(x)$, is contained in the closure of V^+: this is the so-called spectral condition. A more refined assumption introduced for the requirements in particle physics is that Σ contains discrete parts localized on sheets of (mass-shell) hyperboloids inside V^+. These support properties in p-space imply that the corresponding inverse Fourier transforms $<\Psi, \Phi(x_1)\cdots\Phi(x_N)\Psi'>$ are boundary values of holomorphic functions in appropriate tube domains of the complex space variables $(z_1, \ldots, z_n)$.

On the other hand, in order to exhibit structure functions with conical support properties in x-space, one needs to build appropriate algebraic combinations of functions $<\Psi, \Phi(x_{j_1})\cdots\Phi(x_{j_N})\Psi'>$ with permuted arguments in order to take the benefit of the causality postulate, which is always formulated in terms of the commutator of two field operators. There are two versions of this postulate. In the Wightman framework, causality is expressed by the condition of local commutativity or microcausality,

$$[\Phi(x_1), \Phi(x_2)] = 0 \text{ for } (x_1 - x_2)^2 < 0 \qquad [3]$$

In the algebraic QFT framework, causality is expressed by a similar property in terms of any field $B(x)$ generated by a local observable $B \doteq B(0)$ affiliated to a region of spacetime enclosed in a given "double cone" $\mathcal{O}_b = V_b^+ \cap (-V_b^+)$. The corresponding expression of causality is

$$[B(x_1), B(x_2)] = 0 \quad \text{for } (x_1 - x_2) \notin (V_a^+ \cup (-V_a^+)) \qquad [4]$$

for all a such that $a > 2b$.

So, we see that basically, causality and spectral condition generate analyticity respectively in complexified p-space and x-space. However, the situation is more intricate, since for each N there are always several holomorphic branches (two in the case $N = 2$) in the variables $(z_1, \ldots, z_n)$ and also in the variables $(k_1, \ldots, k_n)$: each of these two sets is obtained essentially by permutations of the N vector variables. The important point is that these various branches can be seen to "communicate together," thanks to the existence of "coincidence regions" of their boundary values on the reals. Here again the roles played by causality and stability are symmetric (but inverted): while causality produces coincidence regions for the holomorphic functions in complex spacetime, spectral conditions produce coincidence regions for the holomorphic functions in complex energy–momentum space.

In view of a basic theorem of several complex variable analysis, called the edge-of-the-wedge theorem (see below in (4)), the two sets of communicating holomorphic branches actually define by mutual analytic continuation two holomorphic function $H_N^{\Psi,\Psi'}(k_1, \ldots, k_N)$ and $\mathcal{W}_N^{\Psi,\Psi'}(z_1, \ldots, z_N)$ in respective domains $D_N^{\Psi,\Psi'}$ and $\Delta_N^{\Psi,\Psi'}$. However, these two primitive domains are not natural holomorphy domains (a phenomenon which is particular to complex geometry in several variables). The problem of finding their holomorphy envelopes, namely the smallest domains $\hat{D}_N^{\Psi,\Psi'}$ and $\hat{\Delta}_N^{\Psi,\Psi'}$ in which any functions holomorphic in the primitive domains can be analytically continued, is the idealistic purpose of what has been called the analytic program of axiomatic QFT. So, we see that there is an analytic program in x-space and there is an analytic program in p-space. In practice, except for the case $N = 2$, where the complete answer is known, only a partial knowledge of the holomorphy envelopes has been obtained.

The analytic program in p-space, which is the only one to be described in the rest of this article, was often considered as physically more interesting, in view of the fact that it aims to establish analyticity properties of the scattering kernels on the complex mass shell. As a matter of fact, an important part of it concerns the derivation of the analyticity domains of dispersion relations for two-particle scattering amplitudes. This part is important from the historical viewpoint as well as from conceptual, physical, and pedagogical viewpoints (the reader may find it useful to first check the article Dispersion Relations, which illustrates how a structure function of the form $H_2^{\Psi,\Psi'}(k_1,k_2)$ can be used for that purpose with a suitable choice of the states Ψ and Ψ'). In the general development of the analytic program (in x-space as well as in p-space), it is recommended to consider the infinite set of structure functions $H_N \doteq H_N^{\Omega,\Omega}(k_1,\ldots,k_N)$ and $\mathcal{W}_N \doteq \mathcal{W}_N^{\Omega,\Omega}(z_1,\ldots,z_N)$ where Ω is the privileged vacuum state of the theory, in view of the fact that each of these sets characterizes entirely the field theory considered.

Before shifting to the analytic program in p-space, we would like to mention various points of interest of the analytic program in x-space:

1. Various results of this program have been extensively used for proving fundamental properties of QFT, such as the PCT-invariance theorem, the spin–statistics connection, etc. A good part of these can be found in the books by Streater and Wightman (1980) and by Jost (1965).
2. The functions H_N and $\mathcal{W}_N$ are holomorphic in their respective p-space and x-space "Euclidean subspaces." To make this clear, let us assume that a Lorentz frame has been chosen once for all; the linear subspace of complex spacetime (resp. energy–momentum) vectors of the form $z=(\mathrm{i}y_0,\boldsymbol{x})$ (resp. $k=(\mathrm{i}q_0,\boldsymbol{p})$) is called the "Euclidean subspace" of the corresponding complex Minkowskian space, in view of the fact that the quadratic form $z^2 \doteq z\cdot z=-(y_0^2+\boldsymbol{x}^2)$ (resp. $k^2 \doteq k\cdot k=-(q_0^2+\boldsymbol{p}^2)$) has a definite (negative) sign on that subspace. Then it has been established that (for each N) the restrictions of H_N and $\mathcal{W}_N$ to the corresponding N-vector Euclidean subspaces are the Fourier transforms of each other. This fact participates in the foundation of the Euclidean formulation of QFT or "QFT at imaginary times"; the latter has provided many important results in QFT, in particular for the rigorous study of field models (initiated by Glimm and Jaffe in the 1970s).
3. A more recent extension of QFT called thermal QFT (TQFT), which aims to study the behavior of quantum fields in a thermal bath, can be described in terms of a modified analytic program. In the latter, the spectral condition is replaced by the so-called KMS condition, which prescribes x-space analyticity properties of a particular type for the structure functions $\mathcal{W}_N$: it requires analyticity together with periodicity conditions with respect to imaginary times, the period being the inverse of the temperature (*see* Thermal Quantum Field Theory). The usual analytic structure for the theories with vacuum and spectral conditions is recovered in the zero-temperature limit.
4. In more recent investigations concerning quantum fields on (holomorphic) curved spacetimes, analyticity properties of the structure functions similar to those of thermal QFT can be established. This is the case in particular with de Sitter spacetime, for which a notion of "temperature of geometrical origin" is most simply exhibited.

In this article, an account of the general analytic program of axiomatic QFT in complex energy–momentum space will be presented; it will describe some of the methods which have been used for establishing analyticity properties of the N-point structure functions of QFT and corresponding properties of the $(n\to n')$-particle collision processes, for all n,n' such that $n\geq 2, n'\geq 2, n+n'=N$. (For a more detailed study, in particular concerning the microlocal methods, see the book by Iagolnitzer (1992)).

Concerning the important case $N=4$, this article gives complements to the results described in the article Dispersion Relations. In fact, the program allows one to justify other important analytic structures of the four-point functions and of two-particle scattering functions. They concern

- the field-theoretical basis of analyticity in the complexified variable of angular momentum, first introduced and developed in potential theory (Regge 1959);
- the Bethe–Salpeter (BS-) type structure (based on the additional postulate of asymptotic completeness), which is a relativistic field-theoretical generalization of the Lippmann–Schwinger structure of nonrelativistic scattering theory (for Schrödinger equations with Yukawa-type potentials).

The latter allows one to introduce the concept of composite particle in the field-theoretical framework (including bound states and unstable particles or "resonances") and also the concept of "Regge particle," thanks to complex angular momentum analysis.

Various Aspects of the General Analytic Program of QFT in Complex Energy–Momentum Space

The *N*-Point Structure Functions of QFT

It is proved in the Wightman QFT axiomatic framework that any QFT is completely characterized by the (infinite) sequence of its "*N*-point functions" or "vacuum expectation values" (also called "Wightman functions")

$$W_N(x_1,\ldots,x_N) \doteq < \Omega, \Phi(x_1)\cdots\Phi(x_N)\Omega >$$

which are tempered distributions on $\boldsymbol{R}^{4N}$ satisfying a set of general properties that can be split up into linear and nonlinear conditions. (This is known as the Wightman reconstruction theorem).

Linear conditions Each individual *N*-point function satisfies three sets of linear conditions which result, respectively, from:

1. *Poincaré invariance*: typically, for every Poincaré transformation g of Minkowski spacetime
$$W_N(x_1,\ldots,x_N) = W_N(gx_1,\ldots,gx_N)$$
in particular, the W_N are invariant under spacetime translations and therefore defined on the quotient subspace $\boldsymbol{R}^{4(N-1)} \doteq \boldsymbol{R}^{4N}/\boldsymbol{R}^4$ of the differences $x_j - x_k$.
2. *Microcausality*: support conditions on commutator functions of the following form:
$$C^{(j,j+1)}(x_1,\ldots,x_n) \doteq W_N(x_1,\ldots,x_j,x_{j+1},\ldots,x_N) - W_N(x_1,\ldots,x_{j+1},x_j,\ldots,x_N) = 0$$
in the region of $\boldsymbol{R}^{4N}$ defined by $(x_j - x_{j+1})^2 < 0$.
3. *Spectral condition*: support conditions on the Fourier transform $\tilde{W}_N(p_1,\ldots,p_N) = \delta(p_1+\cdots+p_N)\times\hat{w}_N(p_1,\ldots,p_{N-1})$ of W_N, which assert that $\hat{w}_N(p_1,\ldots,p_{N-1}) = 0$ if either one of the following conditions is fulfilled: $p_1+\cdots+p_j \notin \Sigma$, for $j = 1,\ldots,N-1$.

For each *N*, one can then construct a set of distributions $R_N^{(\alpha)}(x_1,\ldots,x_N)$, called "generalized retarded functions" (Araki, Ruelle, Steinmann, 1960 (see Iagolnitzer (1992, ref. [EGS])) which are appropriate linear combinations of multiple commutator functions built from W_N and multiplied by products of Heaviside step-functions $\theta(x_{j,0} - x_{k,0})$ of the differences of time coordinates. Each of these distributions $R_N^{(\alpha)}(x_1,\ldots,x_N)$ has its support contained in a convex salient cone C_α. This construction can be seen as a generalization of the decomposition [23] of the commutator $C_{\Psi,\Psi'}$ in the article Dispersion Relations. Then in view of the Laplace-transform theorem in several variables, the Fourier transform $\tilde{R}_N^{(\alpha)}(p_1,\ldots,p_N) = \delta(p_1+\cdots+p_N)\times\tilde{r}_N^{(\alpha)}([p]_N)$ is such that $\tilde{r}_N^{(\alpha)}([p]_N)$ is the boundary value of a holomorphic function $\tilde{r}_N^{(\alpha),(c)}([k]_N)$ defined in a tube $\mathcal{T}_\alpha = \boldsymbol{R}^{4(N-1)} + \mathrm{i}\tilde{C}_\alpha$. Here $[k]_N = [p]_N + \mathrm{i}[q]_N$ belongs to a $4(N-1)$-dimensional complex linear space $M_N^{(c)}$: this is the set of complex vectors $[k]_N \doteq (k_1,\ldots,k_N)$ such that $k_1+\cdots+k_N = 0$. $\tilde{C}_\alpha$ is the dual cone of C_α in the real $(4(N-1)$-dimensional) $[q]_N$-space. Geometrically, each cone $\tilde{C}_\alpha$ is defined in terms of a certain "cell" of $[q]_N$-space which is defined by prescribing consistent conditions of the form $\varepsilon_J q_J \in V^+$ with $q_J = \sum_{j\in J} q_j$ and $\varepsilon_J = \pm 1$ for all proper subsets J of the set $\{1,2,\ldots,N\}$. This is the expression of the microcausality postulate (summarized in [3] or [4]) in complex energy–momentum space. Concerning the difference between the two formulations [3] and [4], one can see that there is no geometrical difference concerning the analyticity domains, but differences for the type of increase of the structure functions in their tube domains: in the case of [3], they are bounded by powers of the energy–momenta, while in the case of [4] they may have an exponential increase governed by factors of the type $e^{q\cdot a}$.

For each *N*, the linear space generated by all the distributions $\tilde{r}_N^{(\alpha)}([p]_N)$ is constrained by a set of linear relations (called Steinmann relations) which result from algebraic expressions of discontinuities of the following type, called (generalized) "absorptive parts,"

$$\tilde{r}_N^{(\alpha)}([p]_N) - \tilde{r}_N^{(\alpha')}([p]_N) = < \Omega, [\tilde{\mathrm{R}}_{J_1}^{(\alpha_1)}([p]_{(J_1)}), \tilde{\mathrm{R}}_{J_2}^{(\alpha_2)}([p]_{(J_2)})]\Omega > \qquad [5]$$

for all pairs of adjacent cells $(\alpha,\alpha')_{(J_1,J_2)}$ in the following sense: α and α' only differ by changing the value of $\varepsilon_{J_1} = -\varepsilon_{J_2}$, (J_1,J_2) denoting any given partition of the set $\{1,2,\ldots,N\}$. In [5], the symbols $\tilde{\mathrm{R}}_{J_i}^{(\alpha_i)}$ denote generalized retarded operators of lower order and the argument $[p]_{(J)}$ stands for the set of independent 4-momenta $\{p_j; j\in J\}$. Formula [5] may be seen as an *N*-point generalization of formula [26] of Dispersion Relations for the case when the state $\Psi = \Psi'$ is replaced by Ω.

Then by applying to [5] the same argument based on spectral condition as in the exploitation of eqn [26] in Dispersion Relations, one concludes that the two distributions $\tilde{r}_N^{(\alpha)}$ and $\tilde{r}_N^{(\alpha')}$ coincide on an open set $\mathcal{R}_{\alpha,\alpha'}$ of the form $p_{J_1}^2 = p_{J_2}^2 < M_{J_1}^2$, where $p_{J_1} \doteq \sum_{j\in J_1} p_j = -p_{J_2}$. It then follows from the general "oblique edge-of-the-wedge theorem" (Epstein, 1960; see below) that the two corresponding holomorphic functions $\tilde{r}_N^{(\alpha),(c)}([k]_N)$ and $\tilde{r}_N^{(\alpha'),(c)}([k]_N)$

have a common analytic continuation in the union of their tubes together with a certain complex "connecting set," bordered by $\mathcal{R}_{\alpha,\alpha'}$. Since this argument applies to all pairs $(\alpha,\alpha')_{(J_1,J_2)}$, the following important property holds (see Iagolnitzer (1992, refs. [B2], [EGS])):

Theorem 1

(i) *All the holomorphic functions* $\tilde{r}_N^{(\alpha),(c)}([k]_N)$ *admit a common analytic continuation* $H_N([k]_N)$, *called the N-point structure function (or Green function) of the given quantum field in complex energy–momentum space. It is holomorphic in a "primitive domain"* D_N *of* $M_N^{(c)}$, *which is the union of all tubes* $\mathcal{T}_\alpha$ *together with complex "connecting sets" bordered by all the coincidence regions* $\mathcal{R}_{\alpha,\alpha'}$ *defined previously.*

(ii) *For each N the complex domain* D_N *contains the whole Euclidean subspace* $\mathcal{E}_N$ *of* $M_N^{(c)}$, *which is the set of all complex vectors* $[k]_N=(k_1,\ldots,k_N)$ *such that* $k_j=(k_{j,0},\boldsymbol{k}_j); k_{j,0}=iq_{j,0}, \boldsymbol{k}_j=\boldsymbol{p}_J$ *for* $j=1,2,\ldots,N$. (*This Euclidean subspace depends on the choice of a given Lorentz frame in Minkowski spacetime.*)

Positivity Conditions The Hilbert space framework which underlies the axioms of QFT implies (an infinite set of) positivity inequalities on the N-point structure functions of the fields. As a typical example related to the previous formula [5] when $|J_1|=|J_2|=N/2$ (for N even), one can mention the positive-definiteness property of the absorptive parts for appropriate pairs of adjacent cells $(\alpha_1,\alpha_2=-\alpha_1)_{(J_1,J_2)}$, which simply expresses the positivity of the following Hilbertian squared norm:

$$\left|\int \overline{f}([p]_{(J_2)})f([p]_{(J_1)})[\tilde{r}_N^{(\alpha)}([p]_N) - \tilde{r}_N^{(\alpha')}([p]_N)]\mathrm{d}[p]_{(J_1)}\mathrm{d}[p]_{(J_2)}\right| = \left|\int f([p]_{(J_1)})\tilde{\mathrm{R}}_J^{(\alpha_1)}([p]_{(J_1)})\Omega > \mathrm{d}[p]_{(J_1)}\right|^2 \geq 0 \quad [6]$$

Scattering Kernels of General $(n\to n')$-Particle Collisions and General Reduction Formulas

The presentation of $(2\to 2)$-particle scattering kernels in the article Dispersion Relations can be generalized to arbitrary $(n\to n')$-particle collision processes, involving n incoming massive particles $(n\geq 2)$ and n' outgoing massive particles $(n'\geq 2)$. The big "scattering matrix" or "S-matrix" in the Hilbert space of states is the collection of all partial scattering matrices $S_{n,n'}$ or of the equivalent kernels $S_{n,n'}(p_{n,\mathrm{in}};p_{n',\mathrm{out}})$, defined by a straightforward generalization of formula [20] of the quoted article:

$$\mathrm{S}_{n,n'}(\hat{f}_{n,\mathrm{in}},\hat{g}_{n',\mathrm{out}}) = \int_{\mathcal{M}_{n,n'}} \hat{f}_{n,\mathrm{in}}(p_{n,\mathrm{in}})\overline{\hat{g}_{n',\mathrm{out}}}(p_{n',\mathrm{out}}) \times S_{n,n'}(p_{n,\mathrm{in}};p_{n',\mathrm{out}})\mu_m^n(p_{n,\mathrm{in}})\mu_m^{n'}(p_{n',\mathrm{out}}) \quad [7]$$

Here we have considered for simplicity the case of collisions involving a single type of particle with mass m. In the arguments of the wave packets, the kernel, and the measures $(\mu_m^n,\mu_m^{n'}), p_{n,\mathrm{in}}$ and $p_{n',\mathrm{out}}$, respectively, denote the sets of incoming and outgoing 4-momenta $(p_1,\ldots,p_n)$ and $(p'_1,\ldots,p'_{n'})$ which all belong to the physical mass shell $H_m^+=\{p;p\in V^+,p^2=m^2\}$. By supplementing these mass-shell constraints with the relativistic law of conservation of total energy–momentum $p_1+\cdots+p_n=p'_1+\cdots+p'_{n'}$, one obtains the definition of the mass-shell manifold $\mathcal{M}_{n,n'}$ of $(n\to n')$-particle collision processes.

We shall reserve the name of scattering kernel (or scattering amplitude), denoted by $T_{n,n'}(p_{n,\mathrm{in}};p_{n',\mathrm{out}})$, to the so-called "connected component" of the S-matrix kernel $S_{n,n'}(p_{n,\mathrm{in}};p_{n',\mathrm{out}})$. By analogy with the definition of T in terms of S for the two-particle collision processes (*see* Dispersion Relations) $T_{n,n'}$ is defined by a recursive algorithm, which amounts to subtract from $S_{n,n'}$ all the components of the $(n\to n')$-collision processes that are decomposable into independent collision processes involving smaller number of particles, according to all admissible partitions of the numbers n and n'.

For any given N, let us consider all the "affiliated" scattering kernels $T_{n,n'}$ such that $n+n'=N$ and whose corresponding collision processes, also called "channels," are deduced from one another by the relevant exchange of incoming particles and outgoing antiparticles (e.g., $\Pi_1+\Pi_2+\Pi_3\to\Pi_4+\Pi_5+\Pi_6, \Pi_1+\Pi_2\to\overline{\Pi_3}+\Pi_4+\Pi_5+\Pi_6$, and $\Pi_1+\Pi_3\to\overline{\Pi_2}+\Pi_4+\Pi_5+\Pi_6$). There exist general reduction formulas according to which all these scattering kernels are restrictions to the mass-shell manifold $\mathcal{M}_{(N)}$ of appropriate boundary values of the (so-called) "amputated N-point function" $\hat{H}_N(k_1,\ldots,k_N)\doteq(k_1^2-m^2)\cdots(k_N^2-m^2)\times H_N(k_1,\ldots,k_N)$. More precisely, these reduction formulas can be written as follows:

$$T_{n,n'}(-p_{n,\mathrm{in}};p_{n',\mathrm{out}})_{|\mathcal{M}_{(N)}^{(\alpha)}} = \hat{H}_N^{(\alpha)}(p_1,\ldots,p_N)_{|\mathcal{M}_{(N)}^{(\alpha)}} \quad [8]$$

In the latter, $\hat{H}_N^{(\alpha)}$ denotes a certain boundary value of $\hat{H}_N$ on the reals: it is equal to a generalized retarded function $\tilde{r}_N^{(\alpha)}([p]_N)$ which depends in a specific way on a region of the mass shell, called $\mathcal{M}_{(N)}^{(\alpha)}$, in which

the $(n \to n')$-channel is considered. The important thing to be noted in [8] is the sign convention which attributes the notation $-p_j$ to the momentum of any incoming particle and therefore implies that p_j belongs to the negative sheet of hyperboloid $H_m^- \doteq -H_m^+$. This is the price to pay for expressing symmetrically the energy–momentum conservation law as $p_1 + p_2 + \cdots + p_N = 0$ (according to the QFT formalism), but it also displays, as a nice feature, the fact that all the affiliated scattering kernels $T_{n,n'}$ such that $n + n' = N$ are located on the various connected components of the mass shell $\mathcal{M}_{(N)}(p_j \in H_m;\ j = 1, 2, \ldots, N)$: the choice of the sheet H_m^- or H_m^+ of H_m is exactly linked to the incoming or outgoing character of the particle considered.

Remark 1 The reduction formulas are more usually expressed in terms of the Fourier transforms of the (connected parts of the) N-point amputated chronological functions $\tau_N([p]_N)$ (*see* Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools). As a matter of fact, the latter coincide with the boundary values $\tilde{r}_N^{(\alpha)}([p]_N)$ of H_N in the corresponding relevant regions $\mathcal{M}_{(N)}^{(\alpha)}$.

Remark 2 Coming back to the case of two-particle scattering amplitudes (i.e., $n = n' = 2, N = 4$), one can see that the general study presented here implies the consideration of the four-point function $H_4(k_1, k_2, k_3, k_4)$, which is a holomorphic function of three independent complex 4-momenta (since $k_1 + k_2 + k_3 + k_4 = 0$). In that case, the domain D_4 contains 32 tubes $\mathcal{T}_\alpha$ which are specified by triplets of conditions such as $q_1 \in V^+$, $q_2 \in V^+$, $q_3 \in V^+$, or $-q_1 \in V^+$, $q_1 + q_2 \in V^+$, $q_1 + q_3 \in V^+$, and those obtained by permutations of the subscripts $(1, 2, 3, 4)$ and also by a global substitution of the cone V^- to V^+.

Remark 3 The logical path from the postulates of QFT to the analyticity properties of two-particle scattering amplitudes that has been followed in the article Dispersion Relations can be seen as a partial exploitation of the general analyticity properties of the four-point function: one was specially interested there in the analyticity properties of H_4 in a single 4-momentum $k_1 = -k_3$ (at fixed real values of $p_2 = -p_4$). The "partial reduction formula" [27] of Dispersion Relations corresponds to the restriction of eqn [8] (for $N = 4$) to the linear submanifold $(p_1 = -p_3,\ p_2 = -p_4)$. It may also be worthwhile to stress the fact that, in spite of the exponential bounds on H_4 implied by the postulates of algebraic QFT, it has been possible to prove that the scattering function is still bounded by a power of s in its cut-plane (or crossing) domain; the dispersion relations with two subtractions are still justified in that case (Epstein, Glaser, Martin, 1969 (see Martin (1969, preprint))).

Off-Shell Character of D_N: Nontriviality of the Analytic Structure of the Scattering Kernels

One can now see that for each value of $N (N \geq 4)$ the situation created by complex geometry in the space $\mathbf{C}^{4(N-1)}$ of $[k]_N$ is a mere generalization of the one described in a simple situation in the article Dispersion Relations.

1. There exists a fundamental $(3N - 4)$-dimensional complex submanifold, namely the complex mass shell $\mathcal{M}_{(N)}^{(c)}$ defined by the equations $k_j^2 = m^2$; $j = 1, \ldots, N$, which connects together the various real mass-shell components $\mathcal{M}_{(N)}^{(\alpha)}$ interpreted as the various physical regions of a set of affiliated $(n \to n')$-collision processes. The problem of proving the "analyticity of $(n \to n')$-scattering functions" thus amounts to constructing such holomorphic functions on the complex manifold $\mathcal{M}_{(N)}^{(c)}$, whose boundary values on the various real regions $\mathcal{M}_{(N)}^{(\alpha)}$ would reproduce the relevant scattering kernels $T_{n,n'}(-p_{n,\text{in}}; p_{n',\text{out}})$.
2. All the tubes $\mathcal{T}_\alpha$ which generate the primitive domain D_N are off-shell domains, namely their intersections with $\mathcal{M}_{(N)}^{(c)}$ are empty. This simply comes from the fact that the conditions $q_j \in V^\pm$ (included in their definition) and $k_j^2 = m^2 > 0$ are incompatible. One can also check that adding the coincidence regions $\mathcal{R}_{\alpha,\alpha'}$ between adjacent tubes does not improve the situation. However, one can state as a relevant scope the following program.
3. *Linear program* (so-called because it only relies on the linear conditions presented in the section "N-point structure functions of QFT"): find parts of the holomorphy envelope of D_N (possibly improved by the exploitation of the Steinmann relations) whose intersections with the complex mass shell $\mathcal{M}_{(N)}^{(c)}$ are nonempty. In the best case, show that such intersections can exist which connect two different regions $\mathcal{M}_{(N)}^{(\alpha)}$ together, which means "proving the crossing property between these two regions."
4. We shall see in the following that, except for the case $N = 4$, the results of this linear program have been rather disappointing as far as reaching the complex mass shell is concerned; however, other interesting analytic structures also coming from positivity conditions and from the additional postulate of asymptotic completeness have been investigated under the general name of

nonlinear program. The "synergy" created by the combination of these two programs remains, to a large extent, to be explored.

Results of Analytic Completion in the "Linear Program"

We can only outline here some of the geometrical methods which allow one to compute parts of the holomorphy envelopes of the domains D_N. One important method, which may be used after applying suitable conformal mappings, reduces to the following basic theorem.

The tube theorem *The holomorphy envelope of a "tube domain" of the form $T_B = \mathbf{R}^n + \mathrm{i}B$, where B is an arbitrary domain in $\mathbf{R}^n$ called the basis of the tube, is the convex tube $T_{\hat{B}} = \mathbf{R}^n + \mathrm{i}\hat{B}$, where $\hat{B}$ is the convex hull of B.*

The *opposite or oblique edge-of-the-wedge theorem* (Epstein 1960 (see Streater and Wightman (1980, ch. 2, ref. 18))) is a refined local version of the tube theorem, in which the basis B is of the form $B = C_1 \cup C_2$, where C_1, C_2 are two disjoint (opposite or nonopposite) cones with apex at the origin and where T_B is replaced by a pair of "local tubes" $(T^{(\mathrm{loc})}_{C_1}, T^{(\mathrm{loc})}_{C_2})$. Here the adjective "local" means that the real parts of the variables are confined in a given open set $\mathcal{U}$ (which can be arbitrarily small). The connectedness of T_B is now replaced by the consideration of any pair of functions (f_1, f_2) holomorphic in these local tubes whose boundary values on their common real set $\mathcal{U}$ coincide. The result is that f_1 and f_2 admit a common analytic continuation f in a local tube $T^{(\mathrm{loc})}_C$, where C is the convex hull of $C_1 \cup C_2$. In the case of opposite cones $(C_1 = -C_2)$, f is then analytic in the real set $\mathcal{U}$, while in the general oblique case f is only analytic in a complex connecting set bordered by $\mathcal{U}$ (namely a set which connects $T^{(\mathrm{loc})}_{C_1}$ and $T^{(\mathrm{loc})}_{C_2}$). There exists an extended version of the edge-of-the-wedge theorem in which the boundary values of f_1 and f_2 are only defined as distributions.

For simplicity, we shall just give a very rough classification of the type of results obtained. We shall distinguish:

- analyticity domains in the space of several (possibly all) variables: they can be of global type or of microlocal type, namely restricted to complex neighborhoods of real points;
- analyticity domains in special families of one-dimensional complex manifolds; and
- combinations of one-dimensional results which generate domains in several variables by a refined use of the tube theorem, called the *Malgrange–Zerner* "*flat tube theorem*," or "flat edge-of-the-wedge theorem." In the latter, the local tubes $T^{(\mathrm{loc})}_{C_1}$ and $T^{(\mathrm{loc})}_{C_2}$ of f_1 and f_2 reduce to one-variable domains of the upper half-plane in separate variables $z_1 = x_1 + \mathrm{i}y_1, z_2 = x_2 + \mathrm{i}y_2$ but with a common range of real parts $(x_1, x_2) \in \mathcal{U}$. The data $f_1(z_1, x_2)$ and $f_2(x_1, z_2)$ have coinciding boundary values $(f_1(x_1, x_2) = f_2(x_1, x_2))$ in the limit $(y_1 \to 0, y_2 \to 0)$. The result is again the existence of a common analytic continuation to f_1 and f_2, which is a function of two complex variables $f(z_1, z_2)$ in the intersection of the quadrant $(y_1 > 0,\ y_2 > 0)$ with a complex neighborhood of $\mathcal{U}$. (Note that this result of complex analysis still holds when the real boundary values of the holomorphic functions have singularities, namely are only defined in the sense of distributions).

Global analyticity properties The following property (discovered by Streater for three-point functions) looks like an extension of the tube theorem. The holomorphy envelope of the union of two tubes $\mathcal{T}_\alpha, \mathcal{T}_{\alpha'}$ corresponding to adjacent pairs of cells $(\alpha, \alpha')_{(J_1, J_2)}$ together with a complex connecting set bordered by $\mathcal{R}_{\alpha,\alpha'} = \{[p]_N; p^2_{J_1} < m^2_{J_1}\}$ is the convex hull $\mathcal{T}_{\alpha,\alpha'}$ of the union of these tubes minus the following analytic hypersurface σ_{J_1} which can be called "a cut": $\sigma_{J_1} = \{[k]_N\colon k^2_{J_1} = m^2_{J_1} + \rho, \rho \geq 0\}$. The interest of this result (although it remains by itself an off-shell result) is that it can generate larger cut-domains by additional analytic completions, which may have intersections with the complex mass shell (see below for the case $N = 4$).

Microlocal analyticity properties In the case of the four-point function $\hat{H}_4$, it is possible to consider opposite cut-domains of the previous type, for which $\sigma_{J_1} = \sigma_{\{1,2\}}$ is the energy-cut of the channel $(1, 2 \to 3, 4)$, and for which the spectral conditions prescribe an "edge-of-the-wedge situation" in the neighborhood of the corresponding mass-shell component $\mathcal{M}_{(1,2\to3,4)}$. The result is that H_4 is proved to be holomorphic in a full complex cut-neighborhood of $\mathcal{M}_{(1,2\to3,4)}$ in the ambient complex energy–momentum space. The intersection of this local domain with the complex mass shell $\mathcal{M}^{(\mathrm{c})}_{(4)}$ is of course a full complex cut-neighborhood of $\mathcal{M}_{(1,2\to3,4)}$ in $\mathcal{M}^{(\mathrm{c})}_{(4)}$, and this proves that the corresponding scattering amplitude is the boundary value of an analytic scattering function defined as the restriction $\hat{F}(s,t) \doteq \hat{H}_{4|\mathcal{M}^{(\mathrm{c})}_{(4)}}$ of $\hat{H}_4$: it is holomorphic in a domain of complex (s,t) space deprived from the s–cut.

In the general case $N > 4$, the results are less spectacular, although a more sophisticated microlocal method involving a "generalized edge-of-the-wedge

theorem" has been applied. This method, which was one of the three methods at the origin of the chapter of mathematics called microlocal analysis (the other two being Hörmander's "analytic wave-front" method and Sato's "microfunctions" method) is based on a local version of the Fourier–Laplace transformation called the FBI transformation (see, e.g., the book on "hypo-analytic structures" by Treves (1992) and in the present context the article "Causality and local analyticity" by Bros and Iagolnitzer (1973) (see Iagolnitzer (1992, ref. [BI1])).

A first positive result (obtained at first by Hepp in 1965) is the fact that the various real boundary values of $\hat{H}_N$ admit well-defined restrictions as tempered distributions on the corresponding (real) mass shell $\mathcal{M}_{(N)}$; this result is in fact crucial for the rigorous proof of general reduction formulas. However, (according to Bros, Epstein, Glaser, 1972 (see Iagolnitzer (1992, ref. [BEG2])) the local existence of an analytic scattering function in $\mathcal{M}^{(c)}_{(N)}$ is not ensured at all points of the mass shell, but only in certain regions. A rather favourable situation still occurs for $(2 \to 3)$-particle collision amplitudes (i.e., for $N = 5$), but in the general case there are large regions of the mass shell where it is only possible to prove (at least in this linear program) that the amplitude is a sum of a limited number of boundary values of analytic functions, defined in local domains of $\mathcal{M}^{(c)}_{(N)}$ (see in this connection, Iagolnitzer (1992)).

Analyticity at fixed total energy in momentum transfer variables A remarkably simple situation had already been exploited before the general analysis of H_N leading to Theorem 1 was carried out. It is the section of the domain of the N-point function in the space of the "initial relative 4-momentum" $k = (k_1 - k_2)/2$ of the s-channel with initial 4-momenta (k_1, k_2), when the total energy–momentum $P = -(k_1 + k_2)$ with $P^2 = s$ is kept fixed and real. The remaining 4-momenta $p_3, \ldots, p_N$ such that $p_3 + \cdots + p_N = P$ are also kept fixed and real. Consider the case when P is (positive) timelike and such that $s \geq 4m^2$. Then it can be seen that one obtains analyticity of (a certain "1-vector restriction" of) H_N with respect to the vector variable k in the union of the two opposite tubes $\mathcal{T}^+ = \boldsymbol{R}^4 + \mathrm{i}V^+$, $\mathcal{T}^- = \boldsymbol{R}^4 + \mathrm{i}V^-$. Moreover, an edge-of-the-wedge situation holds in view of the spectral coincidence region of the form $k_1^2 = (-P/2 + k)^2 < M_1^2$, $k_2^2 = (-P/2 - k)^2 < M_2^2$. The corresponding holomorphy envelope is given by a Jost–Lehmann–Dyson domain (*see* Dispersion Relations), whose section by the complex mass shell $k_1^2 = k_2^2 = m^2$ turns out to give a "spherical tube domain" of the form $\{k; k = p + iq; k.P = 0, k^2 = -s/4 + m^2; |q^2| < b^2\}$. The $(2 \to N-2)$-particle scattering kernel is therefore the boundary value of a scattering function holomorphic in the previous spherical domain of complex k-space. In the special case of the two-particle scattering amplitude $F(s,t)$, one checks that the previous domain yields for each $s, s \geq 4m^2$, an ellipse of analyticity for $\hat{F}(s,t)$ in the t-plane with foci at $t = 0$ and $u = 4m^2 - s - t = 0$; this ellipse is called the Lehmann ellipse. (We have considered for simplicity the case of a single type of particle with mass m and two-particle threshold at $2m$.) In fact, the squared momentum transfer t is equal to $(k - k')^2$, if $k' = (k_3 - k_4)/2$ denotes the "final relative momentum" of the s-channel, which was here taken to be fixed and real. Moreover, by a similar argument the corresponding absorptive part, namely the discontinuity across the s-cut of the scattering amplitude, can be shown to be holomorphic in a larger ellipse with the same foci called the large Lehmann ellipse.

It is interesting to compare the previous result with the one that one obtains when the fixed vector P is chosen to be spacelike, namely when s has a negative, namely "unphysical" value with respect to the distinguished channel $(1, 2 \to 3, 4)$. For that case, the exploitation of the primitive domain D_4 shows that for all negative (unphysical) values $\zeta_i = k_i^2 < 0$; $i = 1, 2, 3, 4$, of the squared mass variables, the function $\hat{H}_4$ is holomorphic in a cut-plane of the variable t, where the cuts are the t-cut $(t = 4m^2 + \rho$, $\rho \geq 0)$ and the u-cut $(u = 4m^2 - s - t = 4m^2 + \rho'$, $\rho' \geq 0)$. This cut-plane has of course to be compared with the off- shell cut-plane domain Δ_ζ at the basis of the proof of dispersion relations (*see* Dispersion Relations). Here, however, the choice of the squared momentum transfer t as the variable of analyticity allows one to shift to another interpretation in terms of the concept of angular momentum.

Analyticity in the complex angular momentum variable In all the situations previously considered for the case $N = 4$, one can see that at fixed real values of the squared energy s and of the squared masses $\zeta = \{\zeta_i; i = 1, 2, 3, 4\}$, the complex initial and final relative 4-momenta k and k' have directions which vary on the complexified sphere $S^{(c)}$. Moreover, the corresponding restriction of $\hat{H}_4$ to that sphere turns out to be always well defined and analytic on the real part of that sphere: it therefore defines a kernel on the sphere, which, in view of Poincaré invariance, is invariant under the rotations and therefore admits a convergent expansion in Legendre polynomials. Let us call $h_\ell(s; \zeta)$ the corresponding sequence of Legendre coefficients.

In the first case considered above, this sequence coincides (all ζ_i being equal to m^2) with what the physicists call the set of partial waves $f_\ell(s)$ of the scattering amplitude. The analyticity of $\hat{H}_4$ on a complex spherical tube of $S^{(c)}$, namely of $\hat{F}(s,t)$ in the Lehmann ellipse, is then equivalent to a certain exponential decrease property with respect to ℓ of the sequence of partial waves.

In the second case, where s and the ζ_i are negative, it can be seen that the sphere S describes 4-momentum configurations which all belong to a certain Euclidean subspace $\mathcal{E}_4$ of $M_4^{(c)}$. But this situation is much more favourable from the viewpoint of analyticity, since $\hat{H}_4$ can be seen to be holomorphic on the full complex submanifold $S^{(c)} \times S^{(c)}$ minus two sets σ_t and σ_u which correspond to the t- and u-cuts of the complex t-plane. Then this larger analyticity property turns out to be equivalent to the fact that the sequence $h_\ell(s;\zeta)$ admits an interpolation $\tilde{H}(\lambda;s;\zeta)$ holomorphic in a certain half-plane of the form $\mathrm{Re}\,\lambda > \ell_0$ such that for all integers $\ell > \ell_0$ one has: $\tilde{H}(\ell;s;\zeta) = h_\ell(s;\zeta)$. The value of ℓ_0 is linked to the power bound at large momenta that must be satisfied by $\hat{H}_4$ as a consequence of the temperateness property included in the Wightman axiomatic framework (Bros and Viano 2000).

Of course, this nice analytic structure in a complex angular momentum variable could extend to the set of physical partial waves $f_\ell(s)$ if one could establish the analytic continuation of $\hat{F}(s,t)$ in a cut-plane of t containing the Lehmann ellipses, but this seems out of the possibilities at least of the linear program.

The "Nonlinear Program" and Its Two Main Aspects

The extension of the analyticity domains by positivity and the derivation of bounds by unitarity Positivity conditions of the form [6] have been extensively applied to the case $N = 4$ (namely for subsets J with two elements). The main result (Martin 1969) consists in the possibility of differentiating the forward dispersion relations with respect to t and, as a consequence, to enlarge the analyticity domain in t at fixed s: the Lehmann ellipse, whose size shrinks to zero when s tends to infinity, can then be replaced by an ellipse (i.e., the Martin ellipse) whose maximal point $t = t_{\max} > 0$ is fixed when s goes to infinity. This justifies the extension of dispersion relations in s to positive values of t; then in a second step the use of unitarity relations for the partial waves allows one to obtain Froissart-type bounds on the scattering amplitudes (see Martin (1969)).

Asymptotic completeness and BS-type structural analysis The BS equations have been at first introduced as identities of formal series in the perturbative approach of QFT, and the idea of considering such identities as exact equations having a conceptual content in the general axiomatic framework of QFT has been introduced and developed by Symanzik in 1960. However, it took a long time before its integration in the analytic program of QFT (Bros 1970 (see Iagolnitzer (1992, ref. [B1])). These developments belong to the nonlinear program since they rely on quadratic integral equations between the various N-point functions, which express the postulate of asymptotic completeness via the use of appropriate reduction formulas.

For brevity, the general set of BS-type equations for the N-point functions with $N > 4$ will not be presented. The simplest BS-type equation, which concerns the four-point function, can be written as follows:

$$\hat{H}_4(K;k,k') = B(K;k,k') + (\hat{H}_4 \circ_s B)(K;k,k') \quad [9]$$

where

$$(\hat{H}_4 \circ_s B)(K;k,k') = \int_\Gamma \hat{H}_4(K;k,k'')B(K;k'',k')G\left(\frac{K}{2}+k''\right) \times G\left(\frac{K}{2}-k''\right) \mathrm{d}_4 k'' \quad [10]$$

In the latter, the s-channel is privileged, with $s = K^2, K = -(k_1 + k_2)$; $\hat{H}_4$ is seen as a K-dependent kernel (k and k' are the initial and final relative 4-momenta already defined), and the new object B to be studied is also a K-dependent kernel. The function $G(k)$ is holomorphic in k^2 in a cut-plane except for a pole at $k^2 = m^2$ which plays a crucial role. (It is essentially the "propagator" or two-point function of the field theory considered). Apart from pathologies due to the Fredholm alternative, the correspondence between $\hat{H}_4$ and B is one-to-one, but the peculiarity concerns the integration cycle Γ of [10]: it is a complex cycle of real dimension 4, which coincides with the Euclidean space of the vector variable k'' when all the 4-momenta are Euclidean, and can always be distorted inside the analyticity domain of $\hat{H}_4$ together with the external variables. The exploitation of the Fredholm equation in complex space with "floating integration cycles" then implies that B is holomorphic at least in the primitive domain of $\hat{H}_4$.

An important geometrical aspect of the integration on the cycle Γ in [10] is the fact that this cycle is "pinched" between the pair of poles of the functions G when K^2 tends to its threshold value ($s = 4m^2$).

The type of mathematical concept encountered here is closely related to those used in the study of analyticity properties and Landau singularities of the Feynman amplitudes in the perturbative approach of QFT (in this connection, see the books by Hwa and Teplitz (1966) and by F Pham (2005) and references therein).

The first basic result is that it is equivalent for $\hat{H}_4$ to satisfy an asymptotic completeness equation in the pure two-particle region $4m^2 < s < 9m^2$ and for B to satisfy the following property called two-particle irreducibility: B satisfies dispersion relations in s such that the s-cut begins at the three-particle threshold: $s = 9m^2$.

The consequence of this extended analyticity property of B is that it generates the following type of analyticity properties for $\hat{H}_4$:

1. *The existence of a two-sheeted analytic structure* for $\hat{H}_4$ over a domain of the s-plane containing the interval $4m^2 \leq s < 9m^2$, with a square-root-type branch point at the threshold $s = 4m^2$.
2. *Composite particles*. There exists a Fredholm-type expression

$$\hat{H}_4(K; k, k') = \frac{N(K; k, k')}{D(K^2)} \qquad [11]$$

where N and D are expressed in terms of B via Fredholm determinants, which shows that in its second sheet $\hat{H}_4$ may have poles in $s = K^2$, generated by the zeros of D. These poles are interpreted as resonances or unstable particles. The generation of real poles in the first sheet (i.e., bound states) is also possible under special spectral assumptions of QFT.
3. *Complex angular momentum diagonalization of BS-type equations* (Bros and Viano 2000, 2003). The operation $\circ_s$ in the BS-type equation [9] contains not only an integration over squared-mass variables, but also a convolution product on the sphere S; the latter is transformed into a product by the Legendre expansion of four-point functions described previously in the subsection "Analyticity in the complex angular momentum variable." As a result, there is a partially diagonalized transform of eqn [9] in terms of the functions $\tilde{H}(\lambda; s; \zeta)$ and $\tilde{B}(\lambda; s; \zeta)$, which allows one to write a Fredholm formula similar to [11], namely

$$\tilde{H}(\lambda;\ s; \zeta) = \frac{\tilde{N}(\lambda;\ s; \zeta)}{\tilde{D}(\lambda; s)} \qquad [12]$$

Then under suitable increase assumptions on B, there may exist a half-plane of the form $\mathrm{Re}\,\lambda > \ell_1$ (with $\ell_1 < \ell_0$) such that $\tilde{H}(\lambda; s; \zeta)$ admits poles in the joint variables λ and s, corresponding to the concept of Regge particle: the composite particles introduced in (2) might then be integrated in the Regge particle, although they manifest themselves physically only for integral values ℓ of λ with the corresponding spin interpretation. Of course, this scenario is by no means proven to hold in the general analytic program of QFT, but we have seen that the relevant "embryonary structures" are conceptually built-in, so that the phenomenon might hopefully be produced in a definite quantum field model.
4. *Byproducts of BS-type structural analysis for* $N = 5$ *and* $N = 6$. Relativistic exact structural equations for $(3 \to 3)$-particle collision amplitudes, which generalize the Faddeev structural equations of nonrelativistic potential theory, have been shown to be valid in the energy region of "elastic" collisions (i.e., with total energy bounded by $4m$); relevant Landau singularities of tree diagrams and triangular diagrams have been exhibited as a by-product in this low-energy region (Bros, and also Combescure, Dunlop in two-dimensional field models, 1981 (see Iagolnitzer (1992, refs. [B3], [B4], [CD])). Moreover, crossing domains on the complex mass shell for $(2 \to 3)$-particle collision amplitudes have been obtained (Bros 1986 (see Iagolnitzer (1992, ref. [B1])) by conjointly using $(N = 5)$ BS-type equations together with analytic completion properties (*see*, e.g., the "Crossing lemma" in Dispersion Relations).

See also: Algebraic Approach to Quantum Field Theory; Axiomatic Quantum Field Theory; Dispersion Relations; Scattering, Asymptotic Completeness and Bound States; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools; Thermal Quantum Field Theory.

Further Reading

Bros J and Viano GA (2000) Complex angular momentum in general quantum field theory. *Annales Henri Poincaré* 1: 101–172.

Bros J and Viano GA (2003) Complex angular momentum diagonalization of the Bethe–Salpeter structure in general quantum field theory. *Annales Henri Poincaré* 4: 85–126.

Haag R (1992) *Local Quantum Physics*. Berlin: Springer.

Hwa RC and Teplitz VL (1966) *Homology and Feynman integrals*. New York: Benjamin.

Iagolnitzer D (1992) *Scattering in Quantum Field Theories: The Axiomatic and Constructive Approaches*, Princeton Series in Physics. Princeton: Princeton University Press.

Jost R (1965) *The General Theory of Quantized Fields*. Providence: American Mathematical Society.

Martin A (1969) *Scattering Theory: Unitarity, Analyticity and Crossing*, Lecture Notes in Physics. Berlin: Springer.
Pham F (2005) *Intégrales singulières*. Paris: EDP Sciences/CNRS Éditions.
Streater RF and Wightman AS (1964, 1980) *PCT, Spin and Statistics, and all that*. Princeton: Princeton University Press.
Treves F (1992) *Hypoanalytic Structures*. Princeton: Princeton University Press.

Scattering, Asymptotic Completeness and Bound States

D Iagolnitzer, CEA/DSM/SPhT, CEA/Saclay, Gif-sur-Yvette, France
J Magnen, Ecole Polytechnique, France

Introduction

Relativistic quantum field theory (QFT) has been mainly developed since the 1950s in the perturbative framework. Quantities of interest then appear as infinite sums of Feynman integrals, corresponding to infinite series expansions with respect to couplings. This approach has led to basic successes for practical purposes, but suffered due to crucial defects from conceptual and mathematical viewpoints. First, individual terms were *a priori* infinite: this was solved by perturbative renormalization. However, even so, the series remain divergent. Two rigorous approaches have been developed since the 1960s. The axiomatic approach aims to establish a general framework independent of any particular model (Lagrangian interaction) and to analyze general properties that can be derived in that framework from basic principles. The "constructive" approach aims to rigorously establish the existence of nontrivial QFT models (theories) and to directly analyze their properties. Some of the fundamental bases are described in this encyclopedia in the articles by J Bros, D Buchholz and J Summers, and by G Gallavotti, respectively. This article aims to a deeper study of particle analysis and scattering of theories. In contrast to the articles by Buchholz and Summers and G Gallavotti, it is restricted to massive theories, a rather strong restriction, but for the latter goes much beyond in particle analysis.

From a purely physical viewpoint, results remain limited: the models rigorously defined so far are weakly coupled models in spacetime dimensions 2 or 3, results on bound states depend on specific kinematical factors in these dimensions, proofs of asymptotic completeness (AC) are not yet complete, ... On the positive side, we might say that the analysis and results are of interest from both conceptual and physical viewpoints; on the other hand, these works have also largely been related and have contributed to important, purely mathematical developments, for example, in the domain of analytic functions of several complex variables, microlocal analysis, ...

The general framework of QFT based on Wightman axioms is introduced in the next section. Massive theories are characterized in that framework by a condition on the mass spectrum. Haag–Ruelle asymptotic theory then allows one to define, in the Hilbert space $\mathcal{H}$ of states, two subspaces $\mathcal{H}_{\text{in}}$ and $\mathcal{H}_{\text{out}}$ corresponding to states that are asymptotically tangent, before and after interactions, respectively, to free-particle states. The AC condition $\mathcal{H} = \mathcal{H}_{\text{in}} = \mathcal{H}_{\text{out}}$ introduces a further important implicit particle content in the theory. Collision amplitudes or scattering functions are then well defined in the space of on-mass-shell initial and final energy–momenta (satisfying energy–momentum conservation). The LSZ "reduction formulas" give their link with chronological functions of the fields.

Basic properties of scattering amplitudes that follow from the Wightman axioms are then outlined. In particular, these axioms allow one to define the "N-point functions," which are analytic in a domain of complex energy–momentum space containing the Euclidean region (imaginary energy components), and from which chronological and scattering functions can be recovered. Other results at that stage include the on-shell physical sheet analyticity properties of four-point functions, as also general asymptotic causality and local analyticity properties for $N \geq 4$.

Next, we describe results derived from AC and regularity conditions on analyticity and asymptotic causality in terms of particles. In particular, the analysis of the links between analyticity properties of irreducible kernels (satisfying Bethe–Salpeter type equations) and AC in low-energy regions are included, following ideas of K Symanzik.

The final three sections are devoted to the analysis of models.

Models of QFT have been rigorously defined in Euclidean spacetime, through cluster and, more generally, phase-space expansions which are shown

to be convergent at small coupling (and replace the nonconvergent expansions, of perturbative QFT). Examples of such models are the super-renormalizable massive φ^4 models in dimensions 2 or 3 (in the 1970s) and the "just renormalizable" massive (fermionic) Gross–Neveu model – in dimension 2 – in the 1980s. The N-point functions of these models can be shown to have exponential fall-off in Euclidean spacetime. By the usual Fourier–Laplace transform theorem, one obtains in turn analyticity properties in corresponding regions away from the Euclidean energy–momentum space.

On the other hand, *à la* Osterwalder–Schrader properties can be established in Euclidean spacetime. By analytic continuation from imaginary to real times, it is in turn shown that a corresponding nontrivial theory satisfying the Wightman axioms is recovered on the Minkowskian side. This analysis is omitted here. However, no information is obtained in that way on the mass spectrum, AC, energy–momentum space analyticity, ... Such results can be obtained through the use of irreducible kernels. This was initiated by T Spencer in the 1970s and then developed along the same line (Spencer and Zirilli, Dimock and Eckmann, Koch, Combescure, and Dunlop). We outline here the more general approach of the present authors. In the latter, irreducible kernels are directly defined through "higher-order" cluster expansions which are again convergent at sufficiently small coupling. They are shown to satisfy exponential fall-off in Euclidean spacetime with rates better than those of the N-point functions, and hence corresponding analyticity in larger regions around (and away from) the Euclidean energy–momentum space. Results will then be established by analytic continuation, from the Euclidean up to the Minkowskian energy–momentum space, of structure equations that express the N-point functions in terms of irreducible kernels. These structure equations are infinite series expansions, with again convergence properties at small coupling. In the cases $N=2$ and $N=4$ (even theories), the re-summation of these structure equations give, respectively, the Lippmann–Schwinger and Bethe–Salpeter (BS) integral equations (up to some regularization).

The one-particle irreducible (1PI) two-point kernel G_1 is analytic up to $s=(2m)^2-\varepsilon$, where ε is small at small coupling (s is the squared center of mass energy of the channel). A simple argument then allows one to show analyticity of the actual two-point function in the same region up to a pole at $k^2=m_{\mathrm{ph}}^2$: this shows the existence of a first basic physical mass m_{ph} (close at small coupling to the bare mass m). In a free theory (zero coupling) with one mass m, there is only one corresponding particle. At small coupling, the existence of other (stable) particles is not *a priori* expected; nevertheless, we will see that such particles (two-particle bound states) will occur in some models in view of kinematical threshold effects.

The 2PI four-point kernel G_2 is shown to be analytic up to $s=(4m)^2-\varepsilon$ in an even theory. On the other hand, it satisfies a (regularized) BS equation. In a way analogous to the section "AC and analyticity," starting here from the analyticity of G_2, the actual four-point function F is in turn analytic or meromorphic in that region up to the cut at $s\geq 4m^2$, and the discontinuity formula associated with AC in the low-energy region is obtained.

For some models (depending on the signs of some couplings), it will be shown that F has a pole in the physical sheet, below the two-particle threshold (at a distance from it which tends to zero as the coupling itself tends to zero). This pole then corresponds to a further stable particle.

More generally, and up to some technical problems, the structure equations should allow one to derive various discontinuity formulas of N-point functions including those associated with AC in increasingly higher-energy regions. Asymptotic causality in terms of particles and related analyticity properties (Landau singularities...) should also follow. However, in this approach, results should be obtained only for very small couplings as the energy region considered increases.

Note: Notations used are different in the next two sections on the one hand, and the final three sections on the other. These notations follow the use of, respectively, axiomatic and constructive field theory; for instance, x and p are real on the Minkowskian side in the next two sections whereas they are real on the Euclidean side in the last three sections. The mass m in the next two sections is a physical mass, whereas it is a bare mass in the last three sections (where a physical mass is noted m_{ph}).

The General Framework of Massive Field Theories

We denote by $x=(x_0,\boldsymbol{x})$ a (real) point in Minkowski spacetime with respective time and space components x_0 and $\boldsymbol{x}$ (in a given Lorentz frame); $x^2=x_0^2-\boldsymbol{x}^2$. Besides the usual spacetime dimension $d=4$, possible values 2 or 3 will also be considered. In all that follows, the unit system is such that the velocity c of light is equal to 1. Energy–momentum variables, dual (by Fourier transformation) to time and space

variables, respectively, are denoted by $p=(p_0,\boldsymbol{p})$; $p^2=p_0^2-\boldsymbol{p}^2$.

We describe below the Wightman axiomatic framework, though alternative ones such as "local quantum physics" based on the Araki–Haag–Kastler axioms may be used similarly for present purposes. For simplicity, unless otherwise stated, we consider a theory with only one basic (neutral, scalar) field A; A is defined on spacetime as an operator-valued distribution: for each test function f, $A(f)$ (formally $\int A(x)f(x)\mathrm{d}x$) is an operator in a Hilbert space $\mathcal{H}$ of states. A physical state is represented by a (normalized) vector in $\mathcal{H}$ modulo scalar multiples. It has to be physically understood as "sub specie aeternitatis" (i.e., "with all its evolution," the Heisenberg picture of quantum mechanics being always adopted). It is assumed that there exists in $\mathcal{H}$ a representation of the Poincaré group (semidirect product of pure Lorentz transformations and spacetime translations).

The Wightman axioms include:

1. local commutativity: $A(x)$ and $A(y)$ commute if $x-y$ is spacelike: $(x-y)^2<0$.
2. the spectral condition ($=$ positivity of the energy in relativistic form): the spectrum of the energy–momentum operators (infinitesimal generators of spacetime translations) is contained in the cone $V_+(p^2\geq 0, p_0\geq 0)$. In a massive theory, the spectrum is more precisely assumed to be contained in the union of the origin (that will correspond to the vacuum vector introduced next), of one or more discrete mass-shell hyperboloids $H_+(m_i)(p^2=m_i^2, p_0>0)$ with strictly positive masses m_i, and of a continuum. For simplicity, and unless otherwise stated, we consider in this section a theory with only one mass m and a continuum starting at $2m$ (but this will not be so in a theory with "two-particle bound states"). This condition introduces a first (partial) particle content of the theory. In models, physical masses will not be introduced at the outset but will have to be determined.
3. existence in $\mathcal{H}$ of a vacuum vector Ω, which is the only invariant vector under Poincaré transformations up to scalar multiples; it is moreover assumed that the vector space generated by the action of field operators on the vacuum is dense in $\mathcal{H}$.
4. Poincaré covariance of the theory.

Subspaces $\mathcal{H}_{\mathrm{in}}$ and $\mathcal{H}_{\mathrm{out}}$ of $\mathcal{H}$ can be defined by limiting procedures. To that purpose, one considers test functions $f_{j,t}(x)$ with Fourier transforms of the form $\tilde{f}_j(p)\mathrm{e}^{\mathrm{i}(p_o-[\boldsymbol{p}^2+m^2]^{1/2})t}$, where the functions $\tilde{f}_j$ have their supports in a neighborhood of the mass-shell $H_+(m)$. It can then be shown that vectors of the form $\Psi_t=A(f_{1,t})A(f_{2,t})\cdots A(f_{n,t})\Omega$ converge to limits in $\mathcal{H}$ when $t\to\pm\infty$, respectively, and that these limits depend only on the mass-shell restrictions of the test functions $\tilde{f}_{j|H_+(m)}$.

$\mathcal{H}_{\mathrm{in}}$ and $\mathcal{H}_{\mathrm{out}}$ are interpreted physically as subspaces of states that are "asymptotically tangent" before, respectively, after the interactions, to free-particle states with particles of mass m. They are in fact both isomorphic to the free-particle Fock space $\mathcal{F}$, namely the direct sum of n-particle spaces of "wave functions" depending on n on mass-shell energy–momenta $p_1,p_2,\ldots,p_n$.

AC is the assertion that $\mathcal{H}=\mathcal{H}_{\mathrm{in}}=\mathcal{H}_{\mathrm{out}}$, that is, that each state in $\mathcal{H}$ is asymptotically tangent to a free-particle state, with particles of mass m, both before and after interactions (the two free-particle states are different if there are interactions). This condition cannot be expected to always hold in the general framework introduced above, even if we restrict our attention to "physically reasonable" theories in which states of $\mathcal{H}$ are asymptotically tangent to free-particle states before and after interactions: the absence of other stable particles with different masses is not guaranteed. For instance, even if A is "neutral," the action of field operators on the vacuum might generate pairs of "charged" particles with opposite charges, whatever "charge" one might imagine. Individual charged particles cannot occur in the neutral space $\mathcal{H}$ and their mass thus does not appear in the spectral condition. Hence, such states of pairs of charged particles will not belong to $\mathcal{H}_{\mathrm{in}}$ or $\mathcal{H}_{\mathrm{out}}$ although they belong to $\mathcal{H}$. However, if the set of charged particles is known, it can be shown that the above framework might be enlarged by defining charged fields, in such a way that AC might still be valid in the enlarged framework (see the article of Buchholz and Summers). For simplicity, we restrict below our attention to the simplest theories in which AC holds in the way stated above.

If AC holds, it is shown that there exists a linear operator S from $\mathcal{H}$ to $\mathcal{H}$, called "collision operator" or "S-matrix," that relates the "initial" and "final" free-particle states to which a state in $\mathcal{H}$ is tangent before and after interactions, respectively; if AC does not hold, S can also be defined as in operator in $\mathcal{F}$. Collision amplitudes or scattering functions are the energy–momentum kernels of S for given numbers m and n of initial and final particles. As easily seen, they are well-defined distributions on the space of all initial and final on-shell energy–momenta. For convenience, we will denote by p_k the physical energy–momentum of a final particle with index $k(p_k\in H_+(m))$, and by $-p_k$ the physical energy–momentum of an initial particle $(-p_k\in H_+(m))$.

Wightman Functions, Chronological Functions, and LSZ Reduction Formulas

The N-point Wightman "functions" W_N are defined as the vacuum expectation values (VEVs) of the products of N field operators, namely:

$$W_N(x_1, x_2, \ldots, x_N) = < \Omega, A(x_1)A(x_2)\cdots A(x_N)\Omega >$$

The chronological functions T_N are the VEVs of the chronological products of the fields $A(x_1), \ldots, A(x_N)$: in the latter, fields are ordered according to decreasing values of the time components of the points x_k. T_N is essentially well defined due to local commutativity with, however, problems not treated here at coinciding points.

$\tilde{T}_N(p_1, \ldots, p_N)$ will denote the Fourier transform of T_N. In view of the invariance of the theory under spacetime translations, functions above are invariant under global spacetime translation of all points x_k together. Hence, their Fourier transforms contain an energy–momentum conservation (e.m.c.) delta function $\delta(p_1 + p_2 + \cdots + p_N)$. Connected N-point functions are defined by induction (over N) via a formula expressing each (nonconnected) function as the sum of the corresponding connected function and of products of connected functions depending on subsets of points. In contrast to nonconnected functions, the analysis shows that connected functions in energy–momentum space do not contain in general e.m.c. delta functions involving subsets of energy–momenta.

It can be shown that the two-point function $\tilde{T}_2(p_1, p_2) = \delta(p_1 + p_2)\tilde{T}_2(p_1)$ has a pole of the form $1/(p_1^2 - m^2)$ and that $\tilde{T}_N$ has similar poles for each energy–momentum variable p_k on the mass-shell. The connected, amputated chronological function $\tilde{T}_N^{\text{amp,c}}$ is defined by multiplying $(\tilde{T}_N)_{\text{connected}} = \tilde{T}_N^{\text{c}}$ (for $N \geq 2$) by the product of all factors $p_k^2 - m^2$ that cancel these poles. It is then shown that it can be restricted as a distribution to the mass-shell of any physical process with m initial and n final particles, with $m + n = N$, and that this restriction coincides with the collision amplitude of the process. A process is here characterized by fixing the initial and final indices.

The analyticity properties of interest (described below) will apply to the connected functions after factoring out their global e.m.c. delta functions.

The Analytic *N*-point Functions

The Wightman axioms (without so far AC) yield general analyticity, as also asymptotic causality, properties that we now describe. The analysis is essentially based on the interplay of support properties in x-space arising from local commutativity and the definition of chronological operators, and support properties in p-space due to the spectral condition. Support properties in x-space apply to cell and more general "paracell" functions which are VEVs of adequate combinations of products of "partial" chronological operators. It is shown that each such function has support in x-space in a closed cone C_S (with apex at the origin). Moreover, for cell functions, the cone C_S is convex and salient. Hence, in view of the usual Laplace transform theorem, the cell function in p-space (after Fourier transformation) is the boundary value of a function analytic in complex space in the tube $\text{Re}\, p$ arbitrary, $\text{Im}\, p$ in the open dual cone $\tilde{C}_S$ of C_S. It is also shown that, near any real point $P = (P_1, \ldots, P_N)$, the chronological function in p-space coincides with one or more cell functions.

Together with support properties in p-space arising from the spectral condition and the use of coincidence relations between some cell functions (in adequate real regions in p-space), one then shows the existence, for each N, of a well-defined, unique analytic function F_N, called the "analytic N-point function," whose domain of analyticity, the "primitive domain of analyticity," in complex p-space contains all the tubes $\mathcal{T}_S$ associated with the cell functions. It also contains in particular a complex neighborhood of the Euclidean energy–momentum space which consists of energy momenta P_k with real $\boldsymbol{P}_k$ and imaginary energies $(P_k)_0$. Moreover, the chronological function $\tilde{T}_N^{\text{amp,c}}$ is the boundary value of F_N at all real points P, from imaginary directions which include those of the convex envelope of the cones $\tilde{C}_S$ associated with cell functions that coincide locally with $\tilde{T}_N^{\text{amp,c}}$.

However, the primitive domain has an empty intersection with the complex mass-shell, and thus gives no result on analyticity properties of collision amplitudes on the (real or complex) mass-shell. For $N = 4$, it has been possible to largely extend the primitive domain (which is not a "natural domain of holomorphy") by computing (parts of) its holomorphy envelope, which now has a nonempty intersection with the complex mass shell. It is shown in turn that the four-point function F_4 can be restricted to the complex mass-shell in a one-sheeted domain, called the "physical sheet," that admits each (real) physical region on its boundary (there is here one physical region for each choice of the two initial and the two final indices, the corresponding physical regions being disconnected from each other). In each physical region, the collision amplitude is the boundary value of the mass-shell restriction of F_4, from the corresponding half-space of "$+i\varepsilon$" directions $\text{Im}\ s > 0$, where s is the (squared) energy of the process.

The analyticity domain on the complex mass-shell contains paths of analytic continuation between the various physical regions ("crossing property") and admits cuts s_{ij} real $\geq (2m)^2$ covering the various physical regions. From these analyticity properties in the physical sheet, one can also derive "dispersion relations" (*see* Dispersion Relations).

Asymptotic causality and analyticity properties for $N \geq 4$

No similar result has been achieved at $N > 4$, and as a matter of fact, no similar result is expected if the AC condition is not assumed. The best results achieved so far are decompositions of the collision amplitude, in various parts of its physical region, as a sum of boundary values of functions analytic in domains of the complex mass-shell. In contrast to the case $N = 4$, the sum reduces to one term only in a certain subset of the physical region. Near other points, the N-point analytic function cannot be restricted locally to the complex mass-shell, though it can be decomposed as a sum of terms which, individually, are locally analytic in a larger domain that intersects the complex mass-shell.

These analyticity properties for $N \geq 4$ are a direct consequence of (and equivalent to) an asymptotic causality property that we now outline. Let $f_{k,\tau}(p)$ be, for each index k, a test function of the form

$$f_{k,\tau}(p) = \mathrm{e}^{\mathrm{i}p.\tau u_k}\, \mathrm{e}^{-\gamma\tau|p_k - P_k|^2}$$

where each u_k is a point in spacetime, P_k is a given on-shell energy–momentum, and τ will be a spacetime dilatation parameter ($\gamma > 0$). It is well localized in p-space around the point P_k and its Fourier transform is well localized in x-space around the point τu_k up to an exponential fall-off of width $\sqrt{\gamma\tau}$ which is small compared to τ as $\tau \to \infty$.

We now consider the action of the (connected, amputated) chronological function on such test functions. A configuration $u = (u_1, \ldots, u_N)$ will be called "noncausal" at $P = (P_1, \ldots, P_N)$ if this action decays exponentially as $\tau \to \infty$. In mathematical terms, u is then outside the "essential support" or "microsupport" at P. The asymptotic causality property established, has roughly the following content: the only possible causal configurations u at P are those for which energy–momentum can be transferred from the initial to the final points in future cones. Moreover, at least two initial "extremal" points must coincide, as also two extremal final points. The simplest example is the case $N = 4$; if, for example, indices 1,2 are initial and 3, 4 final, then the only *a priori* possible causal situations are such that $u_3 = u_4$ is in the future cone of $u_1 = u_2$ (in this particular case Lorentz invariance implies that $u_3 - u_1$ must be proportional to $P_3 + P_4$). In more general cases, the possible causal configurations u depend on P.

AC and Analyticity

Asymptotic Causality in Terms of Particles and Landau Singularities

As a matter of fact, a better causality property "in terms of particles" – which is the best possible one – is expected for "physically reasonable" theories if the (stable) particles of the theory are known. (By physically reasonable, we mean the absence of "à la Martin" pathologies such as the occurrence of an infinite number of unstable particles with arbitrary long lifetime). That property expresses the idea that the only causal configurations u at P are those for which the energy–momentum can be transferred from the initial to the final points via intermediate stable particles in accordance with classical laws: there should exist a classical connected multiple scattering diagram in spacetime joining the initial and final points u_k, with physical on-shell energy–momenta for each intermediate particle and energy–momentum conservation at each (pointwise) interaction vertex.

This property, if it holds, yields in turn (and is equivalent to) improved analyticity of the analytic N-point function near real physical regions: the (on-shell) collision amplitude is the boundary value of a unique analytic function in its physical region, at least away from some "exceptional points." The boundary value (namely the collision amplitude) is moreover analytic outside Landau surfaces $L_+(\Gamma)$ of connected multiple scattering graphs Γ; and along these surfaces (which are in general smooth codimension-1 surfaces), it is in general obtained from well-specified "$+\mathrm{i}\varepsilon$" directions (that depend in general on the real point P of L_+).

Exceptional points are those that lie at the intersection of two (or several) surfaces $L_+(\Gamma_1)$, $L_+(\Gamma_2)\ldots$, with opposite causal directions, and hence having no $+\mathrm{i}\varepsilon$ directions in common (in the on-shell framework). Such points do not occur at $N = 4$ for two-body processes, in which case the surfaces L_+ are the n-particle thresholds $s = (nm)^2$, with $n \geq 2, s = (p_1 + p_2)^2$. They do occur more generally: in a $3 \to 3$ process, 1,2,3 initial, 4,5,6 final, this is the case of all points P such that $-P_1 = P_4$, $-P_2 = P_5$, $-P_3 = P_6$ which all belong to the Landau surfaces of the two graphs Γ_1, Γ_2, with only one internal line joining two interaction

vertices: in the case of Γ_1, (resp., Γ_2), the first vertex involves the external particles 1, 2, 4 (resp., 1, 3, 5), while the second one involves 3, 5, 6 (resp., 2, 4, 6). If moreover P_1, P_2, P_3 lie in a common plane, previous points P also lie on surfaces L_+ of "triangle" graphs with again opposite causal directions at P. The fact that $+i\epsilon$ directions are opposite can equally be checked for the corresponding Feynman integrals of perturbative field theory.

Remark The above points are no longer exceptional in spacetime dimension 2. In fact, all surfaces L_+ mentioned then coincide with the (on-shell) codimension-1 surface $-p_1 = p_4, -p_2 = p_5, -p_3 = p_6$, with two opposite causal directions. The previous asymptotic causality property, together with a further "causal factorization" property for causal configurations, then yields along that surface an actual factorization of the three-body (nonconnected) S-matrix into a product of two-body scattering functions modulo an analytic background. The latter vanishes outside the surface, hence is identically zero, for some special two-dimensional models.

In the absence of the AC condition, one clearly sees why the above causality in terms of particles cannot be established: as we have seen, there is *a priori* no control on the stable particles of the theory and on their masses, and pathologies such as those mentioned above cannot be excluded. Hopefully, the first problem should be solved if AC is assumed, and the second one should be removed by adequate regularity assumptions. This is the purpose of the so-called axiomatic nonlinear program, in which one also wishes to examine further problems, for example, analytic continuation into unphysical sheets, with the occurrence of possible unstable particle poles and other singularities, nature of singularities, possible multiparticle dispersion relations, . . . , to cite only a few. Results so far remain limited but provide a first insight into such problems.

The Nonlinear Axiomatic Program

Results described below are based on discontinuity formulas arising from – and essentially equivalent in adequate energy regions to – AC, together with some regularity conditions. They can be established either with or without the introduction of adequate "irreducible" kernels. The methods rely on some general preliminary results on Fredholm theory in complex space (and with complex parameters). Irreducible kernels are defined through integral (Fredholm type) equations, first in the Euclidean region (imaginary energies) and then by local distortions of integration contours allowing one to reach the Minkowskian region. From discontinuity formulas and algebraic arguments, these irreducible kernels are shown to have analyticity (or meromorphy) properties associated with the physical idea of irreducibility (see examples below).

Results obtained so far with or without irreducible kernels are comparable in the simplest cases. However, the method based on irreducible kernels gives more refined results and seems best adapted to "extricate" the analytic structure of N-point functions for $N > 4$.

$N = 4$, Two-Body Processes in the Low-Energy Region

By even theory, we mean theories in which N-point function vanishes identically for N odd.

Standard results on two-body processes with initial (resp., final) energy–momenta p_1, p_2 (resp., p'_1, p'_2) in the low-energy region $(2m)^2 \leq s < (3m)^2$ ($s = (p_1 + p_2)^2 = (p'_1 + p'_2)^2$) are based on the "off-shell unitarity equation"

$$F_+ - F_- = F_+ \star F_- \qquad [1]$$

where $F_+(p_1, p_2; p'_1, p'_2)$ and $F_-(p_1, p_2; p'_1, p'_2)$ denote, respectively, the $+i\varepsilon$ and $-i\varepsilon$ boundary values of the four-point function F_4 from above or below the cut $s \geq (2m)^2$ in the physical sheet, and $\star$ denotes on-shell convolution over two intermediate energy–momenta. This relation is a direct consequence of AC for s less than $(3m)^2$, or less than $(4m)^2$ in an even theory. When the four external energy–momentum vectors p_1, p_2, p'_1, p'_2 are put on the mass shell (on both sides of that relation), one recovers the usual elastic unitarity relation for the collision amplitude T_+ and its complex conjugate T_-:

$$T_+ - T_- = T_+ \star T_-$$

In the exploitation of these relations outlined below, a regularity condition is moreover needed, for example, the continuity of F_+ in the low-energy region.

By considering the unitarity equation as a Fredholm equation for T_+ at fixed s (in the complex mass shell), one obtains the following result: T_+ can be analytically continued as a meromorphic function of s through the cut (in the low-energy region) in a two-sheeted (d even) or multisheeted (d odd) domain around the two-particle threshold. Possible poles in the second sheet (generated by Fredholm theory) will correspond physically to unstable particles. The singularity at the two-particle threshold is of the square-root type in s for d even, or in

$1/\log s$ for d odd. The difference between the two cases is due to the power $(d-1)/2$ of s, integer or half-integer, in the kinematical factor arising from on-shell convolution. This result can also be extended to the off-shell function F_4 by applying a further argument of analytic continuation making use of the off-shell unitarity equation.

Restricting now our attention to an even theory (for simplicity), a similar result also follows from the introduction of a two 2PI BS type kernel G satisfying (and here defined from F through) a regularized BS equation of the form

$$F = G + F \circ_M G \qquad [2]$$

where $\circ_M$ denotes convolution over two intermediate energy–momenta with two-point functions on the internal lines and a regularization factor in order to avoid convergence problems at infinity (G then depends on the choice of this factor but its properties and the subsequent analysis do not). Alternatively, one may also introduce a kernel satisfying a renormalized BS equation, but this is not useful for present purposes.

Starting from the above discontinuity formula [1], one shows in turn that G is indeed "2PI" in the analytic sense:

$$G_+ = G_- \qquad [3]$$

in the low-energy region. More precisely, G is analytic or meromorphic (with poles that may arise from Fredholm theory) in a domain that includes the two-particle threshold $s=(2m)^2$, in contrast to F itself.

The proof of [3] is based on the relation independent of M (and thus leaving the M dependence implicit).

$$\circ_+ - \circ_- = \star \qquad [4]$$

(which is a nontrivial adaptation of the decomposition of a mass-shell delta function as a sum of plus and minus $i\varepsilon$ poles). A simple algebraic argument then shows essentially the equivalence between the discontinuity formulas [1] and [3].

In turn, assuming that G has no poles, this analyticity allows one to recover the two-sheetedness (d even) or multisheetedness (d odd, singularity in $1/\log$) of F, in view of the BS type equation.

$N=6$, 3–3 Process in the Low-Energy Region (Even Theory)

The result, in the neighborhood of the 3–3 physical region, is here a "structure equation" expressing the 3–3 function F in the low-energy region as a sum of "à la Feynman contributions" associated with graphs with one internal line and with triangle graphs, with two-point functions on internal lines and four-point functions at each vertex, plus a remainder R. The latter is shown to be a boundary value from $+i\varepsilon$ directions Im s positive, where $s=(p_1+p_2+p_3)^2, p_1, p_2, p_3$ denoting the energy–momentum vectors of the initial particles. Further regularity conditions are needed to recover its local physical region analyticity. The various explicit contributions that we have just mentioned yield the actual physical region Landau singularities expected in the low-energy 3–3 physical region.

A more refined result, in the approach based on irreducible kernels outlined below, applies in a larger region and then includes further à la Feynman contributions associated with 2-loop and 3-loop diagrams (the latter do not contribute to "effective" singularities in the neighborhood of the physical region).

The first result can be established from discontinuity formulas for the three-point function around two-particle thresholds, arising from AC, and "microsupport" analysis of all terms involved. In the approach based on irreducible kernels, it is useful to introduce in particular a 3PI kernel G_3 that, in contrast to the 3–3 function, will be analytic or meromorphic in a domain including the three-particle threshold. To that purpose, an adequate set of integral equations is introduced and the three-particle irreducibility of G_3 in "the analytic sense" is then established. In turn it provides the complete structure equation mentioned above.

More General Analysis

There are so far only preliminary steps in more general situations, in view of (difficult) technical problems involved and the need of *ad hoc* regularity assumption at each stage. As already mentioned, the approach based on irreducible kernels seems best adapted. The analysis should clearly involve more general irreducible kernels with various irreducibility properties with respect to various channels (and not only with respect to the basic channel considered such as the 3–3 channel in the case above). From a heuristic viewpoint, one may first consider to that purpose adequate formal expansions into (infinite) sums of "à la Feynman contributions" adapted to the energy regions under investigation. These à la Feynman contributions will involve adequate irreducible kernels in the graphical sense at each vertex, and the above expansions correspond formally to the best possible regroupings of Feynman integrals with respect to the energy region considered. From such expansions, one might

determine adequate sets of integral equations allowing one, together with regularity assumptions, to carry out an analysis similar to above.

The Models

A Euclidean field-theoretical model can be defined by a probability measure $d\mu(\varphi)$ on the space of tempered distributions φ in Euclidean spacetime, whose moments verify the Osterwalder–Schrader (or similar) axioms. The moments of $d\mu$ are, for each N, the Euclidean (Schwinger) N-point functions:

$$S(x_1, \ldots, x_N) = \int \varphi(x_1) \cdots \varphi(x_N)\, d\mu(\varphi) \qquad [5]$$

In what follows, the measure $d\mu$ will be a perturbed Gaussian measure which, for the massive φ^4 model with a volume cutoff Λ and an ultraviolet cutoff ρ, is given in d dimensions by

$$d\mu_{\Lambda,\rho} = e^{-\lambda(\rho)\int_\Lambda \varphi^4(z)\,dz + a(\rho)\int_\Lambda \varphi^2(z)dz}\, d\nu_\rho(\varphi)/Z_{\Lambda,\rho} \qquad [6]$$

where $Z_{\Lambda,\rho}$ is the normalization factor and where $d\nu_\rho(\varphi)$ is the Gaussian measure of mean zero ($\int \varphi\, d\nu = 0$) and covariance

$$C(x-y;\rho) = \int d^d p\, e^{ip(x-y)}\, e^{-p^2/\rho^2}/(\zeta(\rho)p^2 + m^2)$$

where by convention m is called the bare mass.

For $d = 2$ or 3 one can show that, for $\lambda(\rho) = \lambda$ small enough (depending on m) and $\zeta(\rho) = 1$, there exists a function $a(\rho)$ ($a(\rho) = O(\lambda)$ as $\lambda \to 0$) such that, for any set of N distinct points, the function $S(x_1, \ldots, x_N) = \lim_{\Lambda,\rho\to\infty} S_{\Lambda,\rho}(x_1, \ldots, x_N)$ exists, is not Gaussian (hence does not correspond to a trivial, free theory), and satisfies the Osterwalder–Schrader axioms. The connected part $S(x_1, \ldots, x_N)_{\text{connected}}$ has the following perturbative series:

$$\lim_{\Lambda,\rho\to\infty} \sum_n \frac{(-1)^n}{n!} \int \varphi(x_1) \ldots \varphi(x_N) \times \left[\int_\Lambda [\lambda\varphi^4 - a(\rho)\varphi^2](z)\, dz\right]^n d\nu_{\Lambda,\rho}(\varphi)\big|_{\text{connected}} \qquad [7]$$

which is the (divergent) sum of the connected renormalized (Euclidean) Feynman graphs.

The study of the perturbative series leads to the distinction of:

1. the super-renormalizable theories, where it is possible to take $\lambda(\rho), \zeta(\rho)$ not depending on ρ. In dimension 2, all the models where $\lambda\varphi^4$ is replaced by

$$c_{2p}\varphi^{2p} + c_{2p-1}\varphi^{2p-1} + \cdots + c_5\varphi^5 + \lambda\varphi^4 + c_3\varphi^3 \qquad [8]$$

also exist provided that $c_{2p} > 0$ is small enough depending on m and on the other coefficient c's and λ, and

2. the just renormalizable theories where $\lambda(\rho)$ (and possibly $\zeta(\rho)$) depend in general on ρ. In models mentioned below $\lambda(\rho) \to 0$ as $\rho \to \infty$; this characterizes "asymptotic freedom."

The proof of the existence of the N-point functions makes use of Taylor type expansions with remainder. The first orders are used to compute $\lambda(\rho), \zeta(\rho), a(\rho)$. The idea is to consider the functional integral [5] – at Λ, ρ finite – as an integral over roughly $\Lambda\rho^d$ "degrees of freedom" which are weakly coupled. This corresponds to a decomposition of the phase space (with cutoff both in x-space (the box Λ) and in p-space (roughly $|p| < \rho$)). The coupling between different regions in x-space comes from the propagators C_ρ; the coupling between different frequencies in p-space comes from the φ^4 term (the interaction vertex). The expansion is then, for each degree of freedom, a finite expansion in the coupling between this degree and the others so that, even if the expansion is perturbative up to the order $\Lambda\rho^d$, the bound on each term is qualitatively the one on a product of $\Lambda\rho^d$ finite order-independent expansions, the order of which can be fixed uniformly in ρ (and depending only on λ). To achieve this program, the propagator linking two points of distance of order L must have a decrease of order $e^{-L^{-1}|x-y|}$, that is, have momentum larger than L^{-1}, so that one must localize both in x-space and p-space ; for example, the smallest cells of phase space correspond to fields φ localized in x, p-spaces, the x-boxes being of side ρ^{-1} and the p-localization consisting of values such that roughly $(\rho/2) \le |p| \le \rho$. More generally, a generic cell (of index i) corresponds to fields φ at point x and momentum p, with x in a box of side $2^i\rho^{-1}$ and $2^{-i-1}\rho < |p| < 2^{-i}\rho$.

These expansions are mimicking the à la Wilson renormalization group. For just renormalizable theories (where $\lambda(\rho)$ depends on ρ), one is led to introduce the effective coupling constant $\lambda(2^{-i}\rho)$ whose perturbative expansion is the value at momentum zero of the sum of all the (connected, amputated) four-point functions containing only propagators of momentum (roughly) bigger than $2^{-i}\rho$ (plus $\lambda(\rho)$ which in fact tends to zero as $\rho \to \infty$).

Then by small coupling we mean a theory where $\lambda(2^{-i}\rho)/\zeta(2^{-i}\rho)^2$ is small for all i.

By convention we write $\lambda_{\text{ren}}, \zeta_{\text{ren}}, a_{\text{ren}}$ for the effective parameters of the theory at zero momentum.

The expansion obtained expresses $S_{\text{connected}}$ as a sum of terms each of them being associated to a

given set of phase-space cells which are "connected" together by "links" that are either propagators or vertices. Each term decreases exponentially with the difference $i_{\max} - i_{\min}$ of the upper and lower indices of the phase-space cells involved. Moreover, each set must contain the cells associated to the fields $\varphi(x_1) \cdots \varphi(x_N)$ whose indices are fixed by the order of magnitude of the distances between the points. On the other hand, the difference between the theory of cutoff ρ and the one of cutoff 2ρ are terms containing at least one cell of momentum of order ρ; these terms are thus small like $\mathrm{cst}(x_1, \ldots, x_N)\mathrm{e}^{-(\mathrm{cst})\rho}$, so that the limit as $\rho \to \infty$ exists.

So far, the "construction" of models is possible only at small coupling, apart from special cases. The φ^4 theory in dimension 4 is just renormalizable (from the perturbative viewpoint) but the above condition of small coupling cannot be achieved (and it is generally believed that this model cannot be defined as a nontrivial theory). A just renormalizable model has been shown to exist, namely the Gross–Neveu model which is a fermionic theory in dimension 2. The elementary particle physics models are just renormalizable but their construction has not been completed so far (in particular in view of the confinement problem). *See* Constructive Quantum Field Theory for details.

To state the result in a form convenient for our purposes here, we introduce a splitting of the covariance in two parts:

$$C(x-y;\rho) = C_M(x-y;\rho) + C_{>M}(x-y;\rho), \quad M > m$$

$$\tilde{C}_M(p;\rho) = (\mathrm{e}^{-p^2/\rho^2}/p^2 + m^2) - (\mathrm{e}^{-p^2/\rho^2}/p^2 + M^2)$$

so that $C_M(x-y)$ behaves like C at large distances but has an ultraviolet cutoff of size M, and $|C_{>M}(x-y)| \leq \mathrm{e}^{-M|x-y|}$ decreases exponentially depending on the (technical) choice of M. Let $\mathrm{d}\nu_M(\varphi)$ be the Gaussian measure of covariance C_M.

One divides also Λ in unit cubes and obtains for the connected N-point function an expansion as a sum over connected trees; a tree T is composed of lines ℓ and vertices v; each line joins two vertices or one of the external points $x_1, \ldots, x_N$ and a vertex; moreover, there are no loops.

To each line ℓ is associated a propagator $C_M(z_\ell, z'_\ell) = C_M(\ell)$.

To each vertex v are associated:

1. two subsets I_v, I'_v of $\{\ell\}$,
2. a connected set X_v of unit cubes such that all the $z_\ell, \ell \in I_v$ and all the $z'_\ell, \ell \in I'_v$ are contained in X_v; $|X_v|$ is the volume of X_v, and
3. a kernel $K_{X_v}(\{z, z'\}_v; \varphi)$

Finally, the external points are by convention z_ℓ points; then:

$$\begin{aligned} &S_{\Lambda,\rho}(x_1, \ldots, x_N)_{\mathrm{connected}} \\ &\quad = \int \mathrm{d}\nu_M(\varphi) \sum_T \frac{1}{|T|!} \sum_{\substack{\{X_v\} \\ \text{nonoverlapping}}} \\ &\qquad \times \int \left[\prod_{\substack{\ell \\ z_\ell \text{ not external}}} \mathrm{d}z_\ell \right] \left[\prod_{\ell \in T} \mathrm{d}z'_\ell C_M(\ell) \right] \\ &\qquad \times \prod_{v \in T} K_{X_v}(\{z, z'\}_v; \varphi) \end{aligned} \quad [9]$$

where for coupling small enough:

$$\int \mathrm{d}\nu_M(\varphi) \prod_{v \in T} |K_{X_v}(\{z, z'\}_v; \varphi)| \leq \prod_{v \in T} \mathrm{e}^{-M(1-\epsilon)|X_v|} \quad [10]$$

The X's are 2×2 nonoverlapping; however, it will suffice to sum over all X's (without restriction) to get a bound showing the convergence of the expansion as $\Lambda \to \infty$. In this formula the $K(., \varphi)$'s are still coupled by the measure $\mathrm{d}\nu_M(\varphi)$; all the nonperturbativity is hidden in the K's (in particular the contribution of momentum bigger than M).

As a consequence of [9] and if $a(\rho, \lambda)$ has been chosen such that $a_{\mathrm{ren}} = 0$, for M large enough and at small coupling (depending on M, m):

$$\begin{aligned} &|S(x,y)_{\mathrm{connected}}| \\ &\quad \leq |C_M(x-y)| + \int \mathrm{d}z'_1 \mathrm{d}z'_2 |C_M(x-z'_1) \\ &\qquad \times \mathrm{e}^{-M(1-\epsilon)|z'_1 - z'_2|} C_M(z'_2 - y)| + \cdots \\ &\quad \leq (\mathrm{cst})\mathrm{e}^{-m(1-\epsilon)|x-y|} \end{aligned} \quad [11]$$

More generally, the connected N-point function satisfies

$$|S(x_1, \ldots, x_N)_{\mathrm{connected}}| \leq \mathrm{cst}\ \mathrm{e}^{-m(1-\epsilon)d(x_1, \ldots, x_N)} \quad [12]$$

where $d(x_1, \ldots, x_N)$ is the length of the smallest tree joining $x_1, \ldots, x_N$, with possibly intermediate points.

The Irreducible Kernels

The 1PI Kernel and a Lippmann–Schwinger Equation

To then show that a theory – if the perturbation series heuristically shows it – contains only one particle of mass smaller than $2m(1-\epsilon)$, it is necessary to expand further the coupling between the K's in [9]. Each perturbative step relatively to this coupling will generate a sum of terms such that in each one there is a "new" propagator C_M between two K's.

The fact that in [9] the X's are nonoverlapping has the consequence that an expansion where for each pair of K_X the number of propagators C_M

remains bounded (say by $n+1$) is convergent (for small enough couplings depending on m, n); this is because, for a given X, the others must be farther and farther as their number increases, and in view of the exponential decrease (in x-space) of C_M.

We then consider the expansion where we have further expanded the two-point function $S(x,y)$ such that each term can be decomposed in the channel $x \to y$ in C_M propagators and 1PI contributions (in the sense that any line cutting such a 1PI contribution (and outside the $X's$) cuts at least two propagators); that means that these 1PI contributions are no longer coupled by the $d\nu_M(\varphi)$ measure. They are made of propagators and of K_X which still have nonoverlapping restrictions; the latter are straightforwardly expanded using a kind of (convergent) Mayer expansion; the result is finally a Lippmann–Schwinger type equation:

$$S(x,y)_{\text{connected}} = C_M(x-y) + \int dz_1\, dz_2\, C_M(x-z_1) \times G_1(z_1,z_2)\, C_M(z_2-y) + \cdots \quad [13]$$

or

$$S(x,y)_{\text{connected}} = \left[C_M \sum_{p\geq 0} [G_1 C_M]^p \right](x,y)$$

which is equivalent to

$$S_{\text{connected}} = C_M + C_M G_1 C_M + C_M G_1 S_{\text{connected}} \quad [14]$$

where G_1 is a 1PI kernel that satisfies the bound

$$|G_1(t,u)| \leq \lambda_{\text{ren}} e^{-2m(1-\epsilon)|t-u|} \quad [15]$$

In Fourier transform, eqn [14] becomes

$$F(p) = \tilde{C}_M(p) + \tilde{C}_M(p)\tilde{G}_1(p)\tilde{C}_M(p) + \tilde{C}_M(p)G_1(p)F(p) \quad [16]$$

Denoting by $\delta(p+q)\,F(p,q)$ the Fourier transform of $S(x,y)_{\text{connected}}$, we can then compute $F(p)$:

$$F(p) = \frac{(p^2+m^2)[\tilde{C}_M + \tilde{C}_M\tilde{G}_1](p)}{(p^2+m^2) - (p^2+m^2)\tilde{C}_M\tilde{G}_1(p)} \quad [17]$$

where $(p^2+m^2)\tilde{C}_M(p) \to (1-m^2/M^2)$ as $p \to 0$ and $|\tilde{G}_1(p)| \leq \lambda_{\text{ren}}\,\text{cst}(m)$ so that (as expected) F has no pole in the Euclidean region at small coupling; but, as will be seen in the next section, it has a pole outside the Euclidean region.

The 2PI Kernel and a BS Equation

From the previous discussion, it is clear that one can extract from [9] as many propagators as we want between kernels K_X. If one considers a splitting of the external points in incoming $x_1,\ldots,x_p$ and outgoing $x_{p+1},\ldots,x_N$ points, this defines a channel. One then obtains nPI kernels (in the given channel). In the same way as above, one obtains a relevant structure equation; this equation makes sense only if the kernels K_X have a decrease corresponding to n-particle irreducibility; to that purpose we take $M > nm$. The expansion converges for couplings small enough depending on m and n.

In the case $n=2$ this gives a kind of BS equation (the Lippmann–Schwinger equation corresponding to the case $n=1$); if we restrict, for simplicity, the analysis to even theories one is led to jump directly to the case $n=3$:

$$S(x_1,x_2;x_3,x_4)_{\text{connected}} = \int dz_1\, dt_1\, dz_2\, dt_2 (\circ_M)(x_1,x_2;z_1,t_1) \times G_2(z_1,t_1;z_2,t_2)(\circ_M)(z_2,t_2;x_3,x_4) + \cdots \quad [18]$$

$$S = \circ_M \sum_{p\geq 1} [G_2 \circ_M]^p$$

or

$$S = \circ_M G_2 \circ_M + \circ_M G_2 S \quad [19]$$

where

$$(\circ_M)(x_1,x_2;x_3,x_4) = S(x_1,x_3)S(x_2,x_4) + S(x_1,x_4)S(x_2,x_3)$$

and where

$$|G_2(t_1,t_2;u_1,u_2)| \leq \lambda_{\text{ren}} \exp\{-4m(1-\epsilon)\max_{i,j}(|t_i-u_j|)\} \quad [20]$$

Equation [19] once amputated, and after Fourier transformation, is eqn [2].

More General Irreducible Kernels and Structure Equations

Irreducible kernels with various degrees of irreducibility in various channels can be defined in a similar way. Corresponding expansions of N-point functions follow, in terms of integrals involving these kernels and two-point functions. These kernels are again convergent at small coupling ($\to 0$ as their irreducibility $\to \infty$) as well as the corresponding structure equations (which generalize eqn [18]).

Analyticity, AC, and Bound States

As explained in the introduction, we now proceed by analytic continuation away from the Euclidean region in complex energy–momentum space.

First, it is easily seen that the two-point function is analytic in the region $s < (2m)^2 - \epsilon$ apart from a pole at $s = m_{\text{ph}}^2$ which defines the physical mass m_{ph} (m_{ph}^2 is the zero in p^2 of the denominator in formula [17]). In view of the bounds of the previous two sections, m_{ph} is close to the "bare" mass m.

The 2PI kernel, for even theories, is shown, again by Laplace transform theorem, to be analytic and bounded in domains around and away from the Euclidean region up to $s = (4m)^2 - \epsilon$, and is of the order of λ_{ren}.

As we have seen in the section "AC and analyticity," the analyticity of G_2 entails the analytic structure of F (two-sheeted or multisheeted at the threshold). On the other hand, further poles of F can be generated by the BS integral equation [2] in the physical or unphysical sheets. If a pole in the physical sheet occurs at $s < (2m_{\text{ph}})^2$ real, it will correspond to a new particle in the theory, namely a two-particle bound state.

AC in the Low-Energy Region

The analysis of possible bound states, which will be presented in the following, will show that there might be at most one two-particle bound state of mass $m_{\text{B}} < 2m_{\text{ph}}$ which tends to $2m_{\text{ph}}$ as the couplings tends to zero.

On the other hand, for even theories, in view of the analyticity properties of the two-point function and of the 2PI kernel G_2, equation [1] holds in the region $(2m_{\text{ph}}^2) < s < (4m_{\text{ph}})^2 - \epsilon$, where $*$ is on-shell convolution with particles of mass m_{ph}.

If there is no two-particle bound state, this characterizes the AC of the theory for $s < (4m_{\text{ph}})^2 - \epsilon$.

If there is a bound state of mass m_{B}, AC is established only in the region $s < (3m_{\text{ph}})^2 - \epsilon$.

For non-even theories, the analysis is similar but requires the introduction of new irreducible kernels in view of the fact that the non-evenness opens new channels. AC in all cases can be established, for small couplings, up to $s < (3m_{\text{ph}})^2 - \epsilon$.

Analysis of Possible Two-Particle Bound States for Even Theories at Small Coupling

It can be checked that such poles of F, if there are, either lie far away in the unphysical sheet(s) or are close to the two-particle threshold ($s = (2m_{\text{ph}})^2$). This is due to the convergence, at small coupling, of the Neumann series $F = G_2 + G_2 \circ_M G_2 + \cdots$. Individual terms $G_2 \circ_M \cdots \circ_M G_2$ are, in fact, defined away from the Euclidean region by analytic continuation in a two-sheeted (d even) or multisheeted (d odd) domain around the threshold: to that purpose locally distorted integration contours (initially the Euclidean region) are introduced as in the section "AC and analyticity," so as to avoid the pole singularities of the two-point functions involved in $\circ_M$, the threshold singularities being due to the pinching of this contour between the two poles as $s \to (2m_{\text{ph}})^2$. If a fixed neighborhood of the threshold is excluded, one does obtain uniform bounds of the form $(\text{cst}\ \lambda_{\text{ren}})^q$ (for a term with q factors G_2) in any bounded domain, which ensures the convergence of the Neumann series.

It remains to study the neighborhood of the threshold. To that purpose, the following method is convenient. One shows that the convolution operator $\circ_M$ can be written in the form

$$\circ_M = g(s) * + \nabla \qquad [21]$$

where $*$ is, as in the section "AC and analyticity," on-shell convolution for $s > (2m_{\text{ph}})^2$ or is obtained by analytic continuation for complex value of s around the threshold; $g(s) = 1/2$ for d even and, if d is odd, $g(s) = (\mathrm{i}/2\pi)\log\sigma$, where $\sigma = 4m_{\text{ph}}^2 - s$. In view of this definition of $g(s)$, the operator ∇ is regular: it is an analytic one-sheeted operation around the threshold (this is equivalent to [4]), and it has no pole singularities. This property of ∇ can be established by geometric methods or by an explicit evaluation.

It is then useful to introduce a new kernel U linked to G_2 by the integral equation

$$U = G_2 + U\nabla G_2 \qquad [22]$$

In view of the regularity and bounds of ∇ and G_2, one sees (e.g., by a series expansion) that U, like G_2, is analytic in a neighborhood of the threshold and behaves in the same way at small λ_{ren}.

By a simple algebraic argument F and U are related by the integral equations

$$F = U + g(s)U * F = U + g(s)F * U \qquad [23]$$

Two-dimensional models We start the analysis with the case $d = 2$. The mass shell is trivial in this case; let f be the restriction of F to the mass shell; it depends only on $s = (p_3 + p_4)^2$ due to the mass shell and e.m.c. constraints (as also Lorentz invariance). On the mass shell, the operation $*$ becomes a mere multiplication and the integral equation [23] becomes

$$f(s) = u(s) + \frac{1}{a(s)} f(s)u(s) \qquad [24]$$

where u is the mass shell restriction of U and the factor $a(s)$ arising from $*$ is of the form $a(s) = \text{cst}\ s^{1/2}\sigma^{1/2}$, $\sigma = (2m_{\text{ph}})^2 - s$, which gives

$$f(s) = \frac{a(s)u(s)}{a(s) - u(s)} \qquad [25]$$

In turn one obtains

$$F = U + \frac{U_{|\ |}U}{a(s) - u(s)} \qquad [26]$$

where $U_{|}$ (resp., $_{|}U$) is U with p_3, p_4 (resp., p_1, p_2) restricted to the mass shell. Equation [26] completely characterizes the local structure of F in view of the local analyticity of U.

The analysis of the possible poles follows from the fact that U is equal to G_2 up to higher order in λ_{ren}; on the other hand, G_2 is equal to a first known term plus higher-order corrections in λ_{ren} (if we expand in λ_{ren} the expression for G_2 obtained in the previous section), so that the leading contribution of $u(s)$ is known and the results follow.

For a theory (see [8]) containing a $\lambda_{\text{ren}}\varphi^4$ term there is exactly one pole, which corresponds to the zero of $a(s) - u(s)$, lying in the region $(2m_{\text{ph}})^2 - \epsilon < s < (2m_{\text{ph}})^2$. This pole is either in the physical sheet for $\lambda_{\text{ren}} < 0$ or in the second sheet if $\lambda_{\text{ren}} > 0$. In the case $\lambda_{\text{ren}} < 0$, this pole corresponds to a two-particle bound state of physical mass m_{B} which tends to $2m_{\text{ph}}$ as $\lambda_{\text{ren}} \to 0$.

In a model without φ^4 term ($\lambda_{\text{ren}} = 0$) the lowest-order contribution to G_2, hence to U, is in general of the order of the square of the leading coupling, in which case there is always one bound state.

The treatment of the fermionic Gross–Neveu model, which involves spin and color indices, is analogous, with minor modifications. Equations now involve, in the two-particle region, 4×4 matrices; poles of F are now the zeros of $\det(a(s)I - m(s)u(s))$, where $m(s)$ is the 4×4 matrix obtained from 2×2 residue matrices (whose leading matrix elements are explicitly computable). The detailed analysis, which requires the consideration of different channels (various color and spin indices) is omitted.

Three-dimensional models The results are similar: F is decomposed as $F' + F''$, where F' is the $\ell = 0$ "partial wave component" of F, namely $F' = (1/2\pi)\int F\,\mathrm{d}\theta$, where θ is the "scattering angle" of the channel; its complement F'' is shown to be locally bounded in view of a further factor σ. The analysis is then analogous to the case $d = 2$ with $a(s)$ now behaving like $\text{cst}/\log\sigma$ as $\sigma \to 0$. There is, *a priori*, either no pole, or one pole in the physical sheet at $s = m_{\text{B}}^2 < (2m_{\text{ph}})^2$ with $m_{\text{B}} = 2m_{\text{ph}} + O(\mathrm{e}^{-\text{cst}/\lambda_{\text{ren}}})$, depending again on the signs of the couplings. For the existing even models such as the φ^4 model, there is no pole, hence no two-particle bound state.

Four-dimensional models The existence of the φ^4 model in dimension 4 is doubtful. If a four-dimensional model were defined, and if the 2PI kernel G_2 of a massive channel could be defined and shown to satisfy analyticity properties analogous to above, there would be no two-particle bound state at small coupling. In fact, the kinematical factor $\sigma^{(d-3)/2}$ (for d even) generated by the mass shell convolution is no longer equal to $\sigma^{-1/2}$ as in the $d = 2$ case but now to $\sigma^{1/2}$. As a consequence, the Neumann series giving F in terms of G_2 is convergent also in the neighborhood of the two-particle threshold.

Non-even theories The analysis for the non-even theories follows similar lines. As already mentioned, the analysis requires the introduction of new irreducible kernels. For the models $\lambda\varphi^4 + c_3\varphi^3$, which do exist at small couplings in dimensions 2 and 3, there will be either exactly one or no two-particle bound state, depending on the respective values of λ, c_3.

Structure Equations and AC in Higher-Energy Regions

The structure equations of the previous section provide, after analytical continuation away from the Euclidean region, a rigorous version of the analysis presented at the end of the section "AC and analyticity." The irreducible kernels can here be defined in a direct way following the previous section, together with their analyticity properties. One has then to derive the discontinuity formulas that in turn characterize AC. This program has been carried out in the $3 \to 3$ particle region, and partly in the general case. It seems possible to complete general proofs up to some technical (difficult) problems. As already mentioned, in this approach, the coupling should be taken smaller and smaller as the energy region considered increases.

See also: Axiomatic Quantum Field Theory; Constructive Quantum Field Theory; Dispersion Relations; Dynamical Systems in Mathematical Physics: An Illustration from Water Waves; Perturbation Theory and its Techniques; Quantum Chromodynamics; Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools; Scattering in Relativistic Quantum Field Theory: the Analytic Program; Schrödinger operators.

Further Reading

Bros J (1984) *r*-particle irreducible kernels, asymptotic completeness and analyticity properties of several particle collision amplitudes. *Physica A* 124: 145.

Bros J, Epstein H, and Glaser V (1965) A proof of the crossing property for two-particle amplitudes in general quantum field theory. *Communications in Mathematical Physics* 1: 240.

Bros J, Epstein H, and Glaser V (1972) Local analyticity properties of the *n*-particle scattering amplitude. *Helvetica Physica Acta* 43: 149.

Epstein H, Glaser V, and Iagolnitzer D (1981) Some analyticity properties arising from asymptotic completeness in quantum field theory. *Communications in Mathematical Physics* 80: 99.

Glimm J and Jaffe A (1981, 1987) *Quantum Physics: A Functionnal Integral Point of View*. Heidelberg: Springer.

Iagolnitzer D (1992) *Scattering in Quantum Field Theories: The Axiomatic and Constructive Approaches*, Princeton Series in Physics. Princeton University Press.
Iagolnitzer D and Magnen J (1987a) Asymptotic completeness and multiparticle structure in field theories. *Communications in Mathematical Physics* 110: 51.
Iagolnitzer D and Magnen J (1987b) Asymptotic completeness and multiparticle structure in field theories. II. Theories with renormalization: the Gross–Neveu model. *Communications in Mathematical Physics* 111: 89.
Martin A (1970) *Scattering Theory: Unitarity, Analyticity and Crossing*. Heidelberg: Springer.
Rivasseau V (1991) *From Perturbative to Constructive Renormalization*. Princeton: Princeton University Press.

Schrödinger Operators

V Bach, Johannes Gutenberg-Universität, Mainz, Germany

Schrödinger operators are linear partial differential operators of the form

$$H_V = -\Delta + V(x) \qquad [1]$$

acting on a suitable dense domain $\mathrm{dom}(H_V) \subseteq L^2(\Omega)$ in the Hilbert space of square-integrable functions on a spatial domain $\Omega \subseteq \mathbb{R}^d$, where $d \in \mathbb{N}$. Here, $H_0 = -\Delta = -\sum_{\nu=1}^{d} \partial^2/\partial x_\nu^2$ is (minus) the Laplacian on Ω, and the potential $V : \Omega \to \mathbb{R}$ acts as a multiplication operator, $[V\psi](x) := V(x)\psi(x)$.

Historical Origin and Relation to Theoretical Physics

In 1926, Schrödinger formulated quantum theory as wave mechanics and proved later that it is equivalent to Heisenberg's matrix mechanics. He proposed that the state of a physical system at time $t \in \mathbb{R}$ is given by a normalized wave function $\psi_t \in L^2(\Omega)$ whose dynamics is determined by a linear Cauchy problem: ψ_0 is the state at time $t = 0$, and for $t > 0$, it evolves according to

$$\mathrm{i}\frac{\partial \psi_t}{\partial t} = H\psi_t \qquad [2]$$

the Schrödinger equation. More generally, ψ_0 is a normalized element of a Hilbert space $\mathcal{H}$, and the Hamiltonian H_V is a self-adjoint operator, that is, $\mathrm{dom}(H_V) = \mathrm{dom}(H_V^*) \subseteq \mathcal{H}$ and $H_V = H_V^*$ on $\mathrm{dom}(H_V)$. Formally, eqn [2] is solved by the evolution operator or propagator $\exp(-\mathrm{i}tH_V)$ in the form $\psi_t = \exp(-\mathrm{i}tH_V)\psi_0$. The self-adjointness of H_V insures the existence and unitarity of the propagator $\exp(-\mathrm{i}tH_V)$, for all $t \in \mathbb{R}$, so $\|\psi_t\| = \|\psi_0\| = 1$. For physics, this unitarity is crucial, because $\|\psi_t\|^2$ is interpreted as the total probability of the system to be at time t in some state in $\mathcal{H}$. The general validity of eqn [2] as the fundamental dynamical law of all physical theories, including, for example, nonrelativistic and (special) relativistic quantum mechanics, quantum field theory, and string theory, deserves appreciation.

If the physical system under consideration is a nonrelativistic point particle of mass $m > 0$ in a potential $\widetilde{V} : \mathbb{R}^d \to \mathbb{R}$, then, according to the principles of classical (Newtonian) mechanics, its state is determined by its momentum $p \in \mathbb{R}^d$ and its position $x \in \mathbb{R}^d$, its kinetic energy is $(1/2m)p^2$, its potential energy is $\widetilde{V}(x)$, and the dynamics is given by the Hamiltonian flow generated by the Hamiltonian function $H_{\mathrm{class}}(p, x) = (1/2m)p^2 + \widetilde{V}(x)$. Schrödinger derived the Hamiltonian (operator) $H = -(\hbar^2/2m)\Delta + \widetilde{V}(x)$ in [2] from the replacement of the momentum $p \in \mathbb{R}^d$ by the momentum operator $-\mathrm{i}\hbar\nabla_x$. This prescription is called quantization and is further discussed in the section "Quantization and semiclassical limit." The Schrödinger operator H_V in [1] is then obtained after an additional unitary rescaling, $\psi(x) \mapsto \mu^{d/2}\psi(\mu x)$, by $\mu := \hbar(2m)^{-1/2}$, and a redefinition $V(x) := \widetilde{V}(x/\mu)$ of the potential.

For more details, we refer the reader to Schrödinger (1926) and Messiah (1962).

Self-Adjointness

Led by the requirement of unitarity of the propagator, the domain $\mathrm{dom}(H_V)$ in [1] is usually chosen such that H_V is self-adjoint, which, in turn, is most often established by means of the Kato–Rellich perturbation theory, briefly described below. If $V \equiv 0$, then H_0 equals the Laplacian $-\Delta$, which is a positive self-adjoint operator, provided $\mathrm{dom}(H_0) = W^2_{\mathrm{b.c.}}(\Omega)$ is the second Sobolev space with suitable conditions on the boundary $\partial\Omega$ of Ω. Typical examples are $\mathrm{dom}(H_0) = W^2(\mathbb{R}^d)$, for $\Omega = \mathbb{R}^d$, and $W^2_{\mathrm{Dir}}(\Omega)$ and $W^2_{\mathrm{Neu}}(\Omega)$ with Dirichlet or Neumann boundary conditions on $\partial\Omega$, respectively, in case that Ω is a bounded, open domain in

$\mathbb{R}^d$ with smooth boundary $\partial\Omega$. Starting from this situation, V is required to be relatively H_0-bounded, that is, that $M(V,r) := V(-\Delta + r\mathbb{1})^{-1}$ defines (extends to) a bounded operator on $L^2(\Omega)$, for any $r > 0$. If $\lim_{r\to\infty} \|M(V,r)\| < 1$, then H_V is self-adjoint on $\mathrm{dom}(H_0)$ and semibounded, that is, the infimum $\inf \sigma(H_V)$ of its spectrum $\sigma(H_V)$ is finite; in other words, $H_V \geq c\mathbb{1}$, for some $c \in \mathbb{R}$, as a quadratic form. (The semiboundedness corresponds to quasidissipativity, as a generator of the semigroup $\exp(-\beta H_V)$.)

A fairly large class of potentials fulfilling these requirements is defined by

$$\lim_{\alpha \searrow 0} \left\{ \sup_{x\in\Omega} \int_{|x-y|\leq\alpha} |x-y|^{4-d} V(y)^2 \, \mathrm{d}^d y \right\} = 0 \qquad [3]$$

for $d \neq 4$, and with $|x-y|^{4-d}$ replaced by $(\ln|x-y|)^{-1}$, for $d = 4$. For $d \leq 3$, [3] is equivalent to the uniform local square integrability of V, that is, $\sup_{x\in\Omega} \int_{|x-y|\leq 1} V(y)^2 \, \mathrm{d}^d y < \infty$. Note that [3] allows for local singularities of V, provided they are not too severe; in this respect, quantum mechanics is more general than classical mechanics. Equation [3] is a sufficient condition for $H_V = -\Delta + V$ to be self-adjoint on $\mathrm{dom}(-\Delta)$ because $\lim_{r\to\infty} \|M(V,r)\| = 0$. Moreover, as eqn [3] only misses some borderline cases, it is also almost necessary for the self-adjointness of H_V. By means of Kato's inequality, the conditions on V, especially on its positive part $V_+ := \max\{V, 0\}$, can be further relaxed. Also, if one realizes H_V as the Friedrichs extension of a semi-bounded quadratic form, the conditions to impose on V are milder. One possibly loses, however, control over the operator domain $\mathrm{dom}(H_V)$, and typically $\mathrm{dom}(-\Delta)$ is only a core for H_V.

For further details on self-adjointness, we refer the reader to Reed and Simon (1980a, b), Kato (1976), and Cycon *et al.* (1987).

Spectral Analysis

The self-adjointness of H_V establishes a functional calculus, generalizing the notion of diagonalizability of finite-dimensional self-adjoint matrices: there exists a unitary transformation $W : L^2(\Omega) \to L^2(\sigma(H_V), \mathrm{d}\mu)$ such that H_V acts on elements φ of $L^2(\sigma(H_V), \mathrm{d}\mu_{H_V})$ as a multiplication operator, $[H_V\varphi](\omega) = \omega\varphi(\omega)$. The spectral measure μ_{H_V} decomposes into an absolutely continuous (ac) part $\mu_{H_V,\mathrm{ac}}$, a pure point (pp) part $\mu_{H_V,\mathrm{pp}}$, and a singular continuous (sc) part $\mu_{H_V,\mathrm{sc}}$, mutual disjointly supported on the ac spectrum $\sigma_{\mathrm{ac}}(H_V)$, the pp spectrum $\sigma_{\mathrm{pp}}(H_V)$, and the sc spectrum $\sigma_{\mathrm{sc}}(H_V) \subseteq \mathbb{R}$, respectively, whose union is the spectrum $\sigma(H_V)$ of H_V. There is an additional decomposition of the spectrum of H_V into the discrete spectrum $\sigma_{\mathrm{disc}}(H_V)$, which consists of all isolated eigenvalues of H_V of finite multiplicity, and its complement $\sigma_{\mathrm{ess}}(H_V) = \mathbb{R} \setminus \sigma_{\mathrm{disc}}(H_V)$, the essential spectrum of H_V, as its residual spectrum is void. One of the main goals of the spectral analysis is to determine the spectral measure for a given potential V as precisely as possible.

In many applications, $\Omega = \mathbb{R}^d$ and the potential V in H_V is not only relatively H_0-bounded, but even relatively H_0-compact, that is, $M(V,1)$ is compact. In this case, $\lim_{r\to\infty} \|M(V,r)\| = 0$, insuring self-adjointness on $\mathrm{dom}(H_0)$ and semiboundedness of H_V. Moreover, a theorem of Weyl implies that its essential spectrum agrees with the one of H_0, that is, with the positive half-axis $\mathbb{R}_0^+$, and the discrete spectrum is contained in the negative half-axis $\mathbb{R}^-$. If, furthermore, $(H_0+1)^{-1}\,[x\cdot\nabla V(x)](H_0+1)^{-1}$ is compact, then the essential spectrum on the positive half-axis is purely absolutely continuous, $\sigma_{\mathrm{ess}}(H_V) \cap \mathbb{R}^+ = \sigma_{\mathrm{ac}}(H_V) \cap \mathbb{R}^+$, and hence $\sigma_{\mathrm{disc}}(H_V) \subseteq \sigma_{\mathrm{pp}}(H_V) \subseteq \sigma_{\mathrm{disc}}(H_V) \cup \{0\}$; the singular continuous spectrum is void.

We remark that the absence of singular continuous spectrum is not understood. Indeed, it is possible to explicitly construct potentials V such that $H(V)$ has singular continuous spectrum. In terms of the Baire category, singular continuous spectrum is even typical. The appearance of singular continuous spectrum can, perhaps, be easier understood in terms of the dynamical properties of $\exp[-itH_V]$, rather than the spectral analysis of its generator H_V: Singular continuous spectrum occurs when initially localized states are not bound states, but move out to infinity very slowly.

The reader is referred to Simon (2000), Reed and Simon (1980a, b) and Cycon *et al.* (1987) for further detail.

Properties of Eigenfunctions

Let us assume $\Omega = \mathbb{R}^d$, that $V \leq 0$ is nonpositive, fulfills [3], and that $\lim_{|x|\to\infty} V(x) = 0$. From the statements in the last section we conclude that $H_V = -\Delta + V(x)$ is semibounded, that the essential spectrum is the positive half-axis and that all eigenvalues are negative and of finite multiplicity, possibly accumulating only at 0. We collect some properties of the eigenfunctions $\psi_j \in L^2(\mathbb{R}^d)$ with corresponding eigenvalue $e_j < 0$, that is, $H_V\psi_j = e_j\psi_j$. The smallest eigenvalue $e_0 := \inf \sigma(H_V)$ (coinciding with the bottom of the spectrum) is simple, and the corresponding eigenfunction $\psi_0(x) > 0$ is strictly positive a.e. Elliptic regularity implies that at a given point $x \in \mathbb{R}^d$, the eigenfunction ψ_j is almost $2 - d/2$ degrees more regular than V. For example,

if $V \in C^k[B_{2\epsilon}(x)]$, for some $\epsilon > 0$, then $\psi_j \in C^{k+\ell}[B_\varepsilon(x)]$, for all $\ell < 2 - d/2$. Agmon estimates (originally obtained by S'nol and also known in mathematical physics as Combes–Thomas argument) furthermore show that, for unbounded Ω, the eigenfunction ψ_j decays exponentially: $|\psi_j(x)| \leq C_\alpha e^{-\alpha|x|}$, for any $0 < \alpha < e_j$.

For more details, see Reed and Simon (1978, 1980a, b) and Cycon *et al.* (1987).

One Dimension and Sturm–Liouville Theory

For $d = 1$, the stationary Schrödinger equation reduces to a second-order ordinary differential equation known as a Sturm–Liouville problem,

$$-\psi''(x) + V(x)\psi(x) = E\psi(x) \qquad [4]$$

on $L^2([a,b])$, with $V \in L^1([a,b])$ and independent boundary conditions at $-\infty \leq a < b \leq \infty$, say. Equation [4] admits an almost explicit solution by means of the Prüfer transformation defined by $\varphi(x) := \arctan[\psi(x)/\psi'(x)]$ and $R(x) := \ln\left(\sqrt{\psi(x)^2 + \psi'(x)^2}\right)$. The key point about the Prüfer transformation is that it effectively reduces the second-order differential equation [4] into a (nonlinear) first-order equation for φ,

$$\varphi'(x) = (E - V(x))\sin^2[\varphi(x)] + \cos^2[\varphi(x)] \qquad [5]$$

Note that [5] does not involve R and that the boundary conditions on ψ and ψ' at a and b can be easily expressed in terms of $\varphi(a)$ and $\varphi(b)$. Moreover, having determined φ on $[a,b]$ from [5], the function R is immediately obtained by integrating $R'(x) = [1 + V(x) - E]\sin[\varphi(x)]\cos[\varphi(x)]$. In case of a bounded interval, $-\infty < a < b < \infty$, or a confining potential, $\lim_{x\to\pm\infty} V(x) = \infty$, it is not difficult to derive from [5] the following basic facts: the spectrum of $H(V)$ consists only of simple eigenvalues $E_0 < E_1 < E_2 < \cdots$ with $\lim_{n\to\infty} E_n = \infty$. Moreover, the corresponding eigenfunction $\psi_n \neq 0$, $n \in \mathbb{N}_0$, with $H(V)\psi_n = E_n\psi_n$, has precisely n zeros, and Sturm's oscillation theorem holds.

See Amrein *et al.* (2005) for more details.

Quantization and Semiclassical Limit

The quantization procedure postulated by Schrödinger is the replacement of the classical momentum $p \in \mathbb{R}^d$ by the quantum-mechanical momentum operator $-i\hbar\nabla_x$. It is known (and, in fact, easy to see, cf. Messiah (1962)) that the classical Hamiltonian equation of motions is invariant under symplectic transformations, but Schrödinger's quantization procedure does not commute with symplectic changes of the classical variables. The question of the geometrically sound definition of quantization, with a general d-dimensional manifold replacing the spatial domain Ω, has attracted many mathematicians and has led to the mathematical fields of geometric quantization and deformation quantization.

It is remarkable, however, that Schrödinger himself discovered already in his early paper the fact that classical dynamics derives as the scaling limit $\hbar \to 0$ from quantum mechanics. The systematic study of the convergence of wave functions and of operators and their spectral properties is known as semiclassical analysis, which is nowadays considered to be part of microlocal analysis. We illustrate the type of results one obtains by the following example on $\Omega = \mathbb{R}^d$.

Let $F \in C_0^\infty(\mathbb{R};\mathbb{R})$ be a smooth characteristic function, compactly supported in an interval $I \subset \mathbb{R}^-$ away from the essential spectrum of the semiclassical Schrödinger operator $H_\hbar = -\hbar^2\Delta + V$ with a smooth potential $V \in C_0^\infty(\mathbb{R}^d)$ of compact support. We define the operator $F[H_\hbar]$ by functional calculus (note that $I \subset \sigma_d(H_V)$ and $F[H_\hbar]$ is of trace class).

Let, furthermore, $A_\hbar = \sum_{|\alpha|\leq M} a_\alpha(x)\partial_x^\alpha$ be a differential operator representing an observable. Then $\mathrm{tr}\{A_\hbar F[H_\hbar]\}$, which exists because the eigenfunctions of $H_\hbar$ are smooth and decay exponentially, is, up to normalization, interpreted to be the expectation of the observable $A_\hbar$ in the state represented by the spectral projection of $H_\hbar$ in I, approximated by $F[H_\hbar]$.

Semiclassical analysis then yields an asymptotic expansion of the form

$$\mathrm{tr}\{A_\hbar F[H_\hbar]\} = \hbar^{-d}(c_0 + c_1\hbar + \cdots + c_n\hbar^n + o(\hbar^n))$$

for arbitrarily large integers $n \in \mathbb{N}$. The leading-order coefficient c_0 is determined by Bohr's correspondence principle,

$$\mathrm{tr}\{A_\hbar F[H_\hbar]\} = \int_{\mathbb{R}^{2d}} a[x,p]F[p^2 + V(x)]\frac{\mathrm{d}p\,\mathrm{d}y}{(2\pi\hbar)^d} + o\left((2\pi\hbar)^{-d}\right) \qquad [6]$$

Semiclassical analysis thus provides the mathematical link between quantum and classical mechanics. The proof of [6] usually involves pseudodifferential and/or Fourier integral operators, depending on the method. Advanced topics in semiclassical analysis studied more recently are the construction of quasimodes, that is, wave functions $\psi_{E,\hbar,n}$ which solve the eigenvalue problem $(H_\hbar - E)\psi_{E,\hbar,n} = O(\hbar^n)$ up to errors of order $\hbar^n$, for arbitrarily large $n \in \mathbb{N}$, and the relation between semiclassical asymptotics

and the KAM (Kolmogorov–Arnold–Moser) theory from classical mechanics.

For more details, see Dimassi and Sjöstrand (1999), and Robert (1987). *See also* Stability Theory and KAM, KAM Theory and Celestial Mechanics in this encyclopedia.

Lieb–Thirring Inequalities

Lieb–Thirring inequalities are estimates on eigenvalue sums of $H_{-V} = -\Delta - V(x)$, where $V \geq 0$ is assumed to be non-negative (note that we changed the sign of V) and vanishing at ∞; the most important examples for these sums are the number of eigenvalues below a given $-E \leq 0$ and the sum of its negative eigenvalues, counting multiplicities. More generally, denoting by $[\lambda]_+ := \max\{\lambda, 0\}$ the positive part of $\lambda \in \mathbb{R}$, Lieb–Thirring inequalities are estimates on $\mathrm{tr}\{[-E - H_{-V}]_+^\gamma\}$, for $\gamma \geq 0$. The number of eigenvalues below $-E$ is then obtained in the limit $\gamma \to 0$, and the sum of the negative eigenvalues corresponds to $E = 0$ and $\gamma = 1$. We henceforth assume $E = 0$, for simplicity. A guess inspired by [6] with $F[\lambda] := [-\lambda]_+^\gamma$, $A = 1$, and $\hbar = 1$ then is that $\mathrm{tr}\{[-H_{-V}]_+^\gamma\}$ is approximately given by

$$\int_{\mathbb{R}^{2d}} \left[V(x) - p^2\right]_+^\gamma \frac{\mathrm{d}^d x\, \mathrm{d}^d p}{(2\pi)^d} = C_{\mathrm{SC}}(\gamma, d) \int_{\mathbb{R}^d} V(x)^{(d/2)+\gamma}\, \mathrm{d}^d x \qquad [7]$$

for a suitable constant $C_{\mathrm{SC}}(\gamma, d) > 0$ depending only on γ and d (but not on V). While this guess is wrong, it is nevertheless a useful guiding principle. Namely, in a rather large range of γ and d, there exist constants $C_{\mathrm{LT}}(\gamma, d) > 0$ such that

$$\mathrm{tr}\{[-H_{-V}]_+^\gamma\} \leq C_{\mathrm{LT}}(\gamma, d) \int_{\mathbb{R}^d} V(x)^{(d/2)+\gamma}\, \mathrm{d}^d x \qquad [8]$$

for all $V \geq 0$, for which the right-hand side is finite (with the understanding that this finiteness also insure that $[-H_{-V}]_+^\gamma$ is trace class, in the first place).

Of course, $C_{\mathrm{LT}}(\gamma, d) \geq C_{\mathrm{SC}}(\gamma, d)$, by [6]. The Lieb–Thirring conjecture, which is still open today, says that the best possible choice of $C_{\mathrm{LT}}(1, 3)$ equals $C_{\mathrm{SC}}(1, 3)$ in the physically most relevant case $\gamma = 1$ and $d = 3$. It is known that $C_{\mathrm{LT}}(\gamma, d) > C_{\mathrm{SC}}(\gamma, d)$, for $\gamma < 1$ or $d < 3$.

Lieb–Thirring estimates have been derived for various modifications of the original model, depending on the application. One of these are pseudorelativistic Hamiltonians of the form $H = T(p) - V$, where $T(p) = \sqrt{p^2 + m^2}$, with $m \geq 0$, another one includes an external magnetic field, for example, $H = (p - A)^2 - V$ (see the next and the last section).

The reader is referred to Thirring (1997), Reed and Simon (1978), and Simon (1979) for further details.

Magnetic Schrödinger Operators

Magnetic Schrödinger operators are Hamiltonians of the form

$$H_{\mathrm{mc}}(\boldsymbol{A}, V) = (\boldsymbol{p} - \boldsymbol{A}(\boldsymbol{x}))^2 - V(x) \quad \text{on } L^2(\mathbb{R}^3) \qquad [9]$$

or

$$H_{\mathrm{Pauli}}(\boldsymbol{A}, V) = [\boldsymbol{\sigma} \cdot (\boldsymbol{p} - \boldsymbol{A}(\boldsymbol{x}))]^2 - V(\boldsymbol{x}) \quad \text{on } L^2(\mathbb{R}^3) \otimes \mathbb{C}^2 \qquad [10]$$

where V is the (electrostatic) potential; as before, $\boldsymbol{A} : \mathbb{R}^3 \to \mathbb{R}^3$ is the vector potential of the magnetic field $B = \boldsymbol{\nabla} \wedge \boldsymbol{A}$, and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$ are the Pauli matrices. $H_{\mathrm{mc}}(\boldsymbol{A}, V)$ and $H_{\mathrm{pauli}}(\boldsymbol{A}, V)$ generate the dynamics of a particle moving in an external electromagnetic field of spin $s = 0$ and spin $s = 1/2$, respectively. The operator $H_{\mathrm{Pauli}}(\boldsymbol{A}, V)$ is usually called Pauli Hamiltonian, and we refer to $H_{\mathrm{mc}}(\boldsymbol{A}, V)$ as the magnetic Hamiltonian. To keep the exposition simple, we assume henceforth that A_ν and $\partial_\mu A_\nu$ are uniformly bounded, which suffices to prove the self-adjointness of both Hamiltonians.

At a first glance, the magnetic and the Pauli Hamiltonians may seem to differ only marginally, but in fact, some of their spectral properties are fundamentally different.

1. The magnetic Hamiltonian fulfills the diamagnetic inequality, $|\mathrm{e}^{-\beta H_{\mathrm{mc}}(\boldsymbol{A}, V)}(x, y)| \leq \mathrm{e}^{-\beta H_{\mathrm{mc}}(0, V)}(x, y)$, for almost all $x, y \in \mathbb{R}^3$, where $m(x, y)$ denotes the integral kernel of an operator m. As a consequence, $\inf \sigma[H_{\mathrm{mc}}(\boldsymbol{A}, V)] \leq \inf \sigma[H_{\mathrm{mc}}(0, V)] = \inf \sigma[H(V)]$, and the quadratic form of the magnetic Hamiltonian is semibounded, for all choices of $\boldsymbol{A}$, provided $H(V)$ is.
2. If $\inf \sigma[H_{\mathrm{mc}}(\boldsymbol{A}, V)]$ is an eigenvalue, the diamagnetic inequality reflects the fact that the corresponding eigenvector is not positive or of constant phase. The determination of the nodal set of eigenfunctions is a difficult task on its own.
3. For $V = 0$, the diamagnetic inequality and the minimax principle imply that $\boldsymbol{p} - \boldsymbol{A}$ has no zero eigenvalue.
4. The diamagnetic inequality fails to hold for the Pauli Hamiltonian. On the contrary, if $\boldsymbol{A}$ is carefully adjusted in $H_{\mathrm{mc}}(\boldsymbol{A}, -Z|x|^{-1})$, and Z is sufficiently large, then the corresponding

quadratic form may assume arbitrarily small values (even if the corresponding field energy is added).

5. For many choices of $\boldsymbol{A}$, the (Dirac) operator $\boldsymbol{\sigma}\cdot(\boldsymbol{p}-\boldsymbol{A})$ has a nontrivial kernel.

From (1)–(4) it is clear that the proof of stability of matter (see the next section) in presence of a magnetic field is more difficult than in absence of it. This can be illustrated by the fact that magnetic Lieb–Thirring inequalities, being the natural analog of eqn [8], are more involved to derive than the original estimate [8]. The currently best bound is of the form

$$\begin{aligned}&\mathrm{tr}\{[-H_{-V}]_+^{\gamma}\}\\&\quad\le C_{\mathrm{mLT}}\int_{\mathbb{R}^d}\Big\{[V(x)]_+^{5/2}+|B(x)|\,[V(x)]_+^{3/2}\\&\qquad+\Big(|B(x)|+L_c(x)^{-2}\Big)L_c(x)^{-1}[V(x)]_+\Big\}\,\mathrm{d}^dx\end{aligned}\quad[11]$$

for some universal $C_{\mathrm{mLT}}<\infty$, where $L_c(x)$ is a local length scale associated with B. It is nonlocal in x and somewhat reminiscent of a maximal function.

We further remark that if restricted to two dimensions, $d=2$, both the magnetic and the Pauli Hamiltonians play an important role in the theory of the (integer) quantum Hall effect.

For more details, see Simon (1979), Cycon *et al.* (1987), Rauch and Simon (1997), and Erdös and Solovej (2004). See also the article Quantum Hall Effect in this encyclopedia.

N-Body Schrödinger Operators

The origin of quantum mechanics is atomic ($K=1$ below) or molecular ($K\ge 2$) physics. If we regard the nuclei of the molecule as fixed point charges $\underline{Z}:=(Z_1,\ldots,Z_K)>0$ at respective positions $\underline{R}:=(R_1,\ldots,R_K)\in\mathbb{R}^3$, then the Hamiltonian (in convenient units) of this molecule with $N\in\mathbb{N}$ electrons is the following Schrödinger operator:

$$\begin{aligned}H_N(\underline{Z},\underline{R})=&\sum_{n=1}^{N}\left\{-\Delta_n-\sum_{k=1}^{K}\frac{Z_k}{|x_n-R_k|}\right\}\\&+\sum_{1\le m<n\le N}\frac{1}{|x_m-x_n|}\end{aligned}\quad[12]$$

defined on $\mathcal{H}^{(N)}:=\bigwedge_{n=1}^{N}L^2[\mathbb{R}^3\times\mathbb{Z}_2]\subseteq L^2[(\mathbb{R}^3\times\mathbb{Z}_2)^N]$, the space of totally antisymmetric, square-integrable wave functions in N space–spin variables $(x_1,\sigma_1),\ldots,(x_N,\sigma_N)\in\mathbb{R}^3\times\mathbb{Z}_2$. The antisymmetry of the wave function accounts for the fact that electrons are fermions and is of crucial importance. Note that the number N of electrons is possibly very large. It is clear that we cannot expect to carry out the spectral analysis of this Schrödinger operator directly, but rather only suitable approximations.

In spite of the fact that $H_N(\underline{Z},\underline{R})$ was one of the basic operators of quantum mechanics from its very beginning in the late 1920s, $H_N(\underline{Z},\underline{R})$ was, strictly speaking, not known to be self-adjoint before Kato developed the perturbation theory (described in the section "Self-adjointness") some 20 years later, which then also yielded the semiboundedness of $H_N(\underline{Z},\underline{R})$. So, the ground-state energy $E_N(\underline{Z},\underline{R}):=\inf\sigma[H_N(\underline{Z},\underline{R})]>-\infty$ is finite. From the HVZ (Hunziker–van Winter–Zishlin) theorem follows that $\inf\sigma_{\mathrm{ess}}[H_N(\underline{Z},\underline{R})]=E_{N-1}(\underline{Z},\underline{R})$, which particularly implies that $E_N(\underline{Z},\underline{R})$ is monotonically decreasing in N and negative (because $E_1(\underline{Z},\underline{R})<0$).

It is known that $E_N(\underline{Z},\underline{R})=E_{N+1}(\underline{Z},\underline{R})$ and that $H_N(\underline{Z},\underline{R})$ has no eigenvalue, for $N\ge 2Z_{\mathrm{tot}}+1$, where $Z_{\mathrm{tot}}:=\sum_{k=1}^{K}Z_k$ is the total nuclear charge of the atom. On the other hand, it is known that $E_N(\underline{Z},\underline{R})$ is an eigenvalue, provided $N<Z_{\mathrm{tot}}$. Thus, defining N_{crit} to be the smallest number such that $E_N(\underline{Z},\underline{R})$ is not an eigenvalue, for all $N\ge N_{\mathrm{crit}}$, that is, N_{crit} is the maximal number of electrons the molecule can bind, we have that $Z_{\mathrm{tot}}\le N_{\mathrm{crit}}\le 2Z_{\mathrm{tot}}+1$. In increasing precision, asymptotic neutrality, $N_{\mathrm{crit}}=Z_{\mathrm{tot}}+R(Z_{\mathrm{tot}})$, with $R(Z_{\mathrm{tot}})=o(Z_{\mathrm{tot}})$ and $R(Z)=o(Z^{5/7})$, was shown for atoms and for molecules, respectively. The ionization conjecture states that $N_{\mathrm{crit}}\le Z_{\mathrm{tot}}+C$, for some universal constant C. It is still open for the full model represented by $H_N(\underline{Z},\underline{R})$, but has been proved in the Hartree–Fock approximation. It has been proved in the Hartree–Fock approximation by Solovej.

The semiboundedness of $H_N(\underline{Z},\underline{R})$, for fixed $\underline{Z},\underline{R}$, and N, alone does not rule out a physical collapse of the matter described by $H_N(\underline{Z},\underline{R})$, but the stronger property of stability of matter does. It holds if there exists a constant C, possibly depending on $\underline{Z}$, such that

$$E_N(\underline{Z},\underline{R})+\sum_{1\le k<\ell\le K}\frac{Z_k\,Z_\ell}{|R_k-R_\ell|}\ge -C(N+K)\quad[13]$$

that is, if the ground-state energy plus the repulsive electrostatic energy of the nuclei is bounded below by a constant times the total number $N+K$ of particles in the system. Equation [13] was shown to hold for $H_N(\underline{Z},\underline{R})$.

In connection with stability of matter, Thomas–Fermi theory and the question of the limit of large nuclear charge came into the focus of research. For simplicity, we restrict ourselves to atoms, $K=1$, that is, there is one nucleus of charge $Z:=Z_1$ at the origin, $R_1=0$, and we consider $E(Z):=\min_{N\in\mathbb{N}}E_N(Z,0)$ (which amounts to fixing $N:=N_{\mathrm{crit}}$). An asymptotic expansion for $E(Z)$ of increasing

precision in Z was obtained by ever-finer estimates; presently, one knows that

$$E(Z) = E_{\mathrm{TF}} Z^{7/3} + \tfrac{1}{4} Z^2 + C_{\mathrm{DS}} Z^{5/3} + o(Z^{5/3}) \qquad [14]$$

where the leading contribution $E_{\mathrm{TF}} Z^{7/3}$ is the Thomas–Fermi energy, $(1/4)Z^2$ is the Scott correction, and $C_{\mathrm{DS}} Z^{5/3}$ is the Dirac–Schwinger term. The computation of this last term requires semiclassical analysis sketched in the section "Quantization and semiclassical limit."

For more details, see Cycon *et al.* (1987), Rauch and Simon (1997), Thirring (1997), and Solovej (2003). See also the article Stability of Matter in this encyclopedia.

Scattering Theory

The study of the properties of the propagator $\exp(-itH)$ of a self-adjoint operator $H = H^*$, as $t \to \infty$, is the concern of scattering theory. To obtain a well-defined mathematical object in this limit, it is necessary to compose $\exp(-itH)$ with the inverse of some explicitly accessible comparison dynamics before passing to the limit $t \to \infty$. If V is a short-range potential, that is, V is relatively H_0-compact and $|V(x)| \leq C|x|^{-\nu}$, for some $\nu > 1$ and $C < \infty$, then the comparison dynamics appropriate for H_V is generated by H_0: the wave operators $\Omega^{\pm}$ are defined as the strong limits

$$\Omega^{\pm} := \lim_{t \to \pm\infty} \mathrm{e}^{\mp itH_V} \mathrm{e}^{\pm itH_0} \qquad [15]$$

A general technique in scattering theory to prove the existence of such limits is Cook's argument, which formally amounts to an application of the fundamental theorem of calculus. For example, for the existence of Ω^+, one writes

$$\begin{aligned} \Omega^+ - \mathbb{1} &= \int_0^\infty \mathrm{d}t \left\{ \frac{\mathrm{d}}{\mathrm{d}t} \left(\mathrm{e}^{-itH_V} \mathrm{e}^{itH_0} \right) \right\} \\ &= -\mathrm{i} \int_0^\infty \mathrm{d}t \, \{ \mathrm{e}^{-itH_V} \, V \, \mathrm{e}^{itH_0} \} \end{aligned} \qquad [16]$$

and additionally proves the absolute integrability of $t \mapsto \mathrm{e}^{-itH_V} V \mathrm{e}^{itH_0} \varphi$, for φ in a dense subset of $\mathcal{H}$, like $\mathrm{dom}(H_0) = \mathrm{dom}(H_V)$.

Research in scattering theory in the past two decades or so was focused around the question of asymptotic completeness, which is a mathematically precise formulation

$$\mathrm{Ran}\Omega^+ = \mathrm{Ran}\Omega^- = \mathcal{H}_{\mathrm{pp}}^{\perp}(H_V) \qquad [17]$$

of the physical expectation that the states in $\mathcal{H}$ are either bound states (eigenvectors) of H_V or scattering states (states in the range of $\Omega^{\pm}$) of H_V. The intertwining property $H_V \Omega^{\pm} = \Omega^{\pm} H_0$ (which easily follows from [15]) implies that the restriction of H_V to $\mathrm{Ran}\Omega^{\pm}$ is unitarily equivalent to H_0, hence $\mathrm{Ran}\Omega^{\pm} \subseteq \mathcal{H}_{\mathrm{ac}}(H_V) \subseteq \mathcal{H}_{\mathrm{pp}}^{\perp}(H_V)$. The difficult part of the proof of asymptotic completeness is to show that $\mathcal{H}_{\mathrm{pp}}^{\perp}(H_V) \subseteq \mathrm{Ran}\Omega^{\pm}$.

Much effort has been spent to prove asymptotic completeness for N-body Schrödinger operators on $\mathcal{H}^{(N)} := \bigotimes_{n=1}^{N} L^2(\mathbb{R}^3)$ of the form

$$\begin{aligned} H_N(V) &= \sum_{n=1}^{N} \frac{-\Delta_n}{2\, m_n} + V(x) \\ \text{with } V(x) &:= \sum_{1 \leq m < n \leq N} V_{mn}(x_m - x_n) \end{aligned} \qquad [18]$$

where each pair potential V_{mn} obeys $|\partial_y^{\alpha} V_{mn}(y)| \leq C(1 + |y|)^{-\mu - |\alpha|}$, with $\alpha \in \mathbb{N}_0^d$ being a multi-index. If $\mu > 1$ for all $m \neq n$ then V is called a short-range potential. Conversely, if $0 < \mu \leq 1$ then V is a long-range potential. Note that even though each V_{mn} decays at infinity, $|x|^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \to \infty$ alone does not imply that $V(x) \to \infty$. In fact, physical intuition tells us that for a cluster $\mathcal{C}$ of N particles, whose dynamics is generated by $H_N(V)$, several scenarios for the long-time asymptotic behavior of the evolution are possible:

1. The N particles stay together in their cluster $\mathcal{C}$ whose center of mass moves in space at constant velocity.
2. The cluster breaks up into two (or even more) subclusters, $\mathcal{C}_1$ and $\mathcal{C}_2$, of N_1 and $N_2 = N - N_1$ particles, respectively, whose centers of mass drift apart from each other at constant velocities (in the short-range case). For each subcluster $\mathcal{C}_1$ and $\mathcal{C}_2$, both scenarios may appear again, after waiting sufficiently longer.
3. In the limit $t \to \infty$, possibly after going through (1) and (2) several times, the initial cluster $\mathcal{C}$ is broken up into $1 \leq K \leq N$ subclusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$, whose centers of mass drift apart from each other at constant velocities according to a free and independent dynamics of their centers of mass.

In some sense, asymptotic completeness says that nothing else than (1)–(3) can possibly happen. (Strictly speaking, asymptotic completeness is a statement about the limit $t \to \infty$ and only involves (3) – the actual behavior of $\exp[-itH_V]$ at intermediate times in terms of (1)–(3) is beyond the reach of current mathematics.) It is a key insight of scattering theory that the asymptotics of the time evolution in the sense of (3) is completely

characterized by the asymptotic velocity defined by the strong limit

$$P^+ := \lim_{t\to\infty} \left(e^{-itH_N(V)} \frac{x}{t} e^{itH_N(V)} \right) \qquad [19]$$

It is a nontrivial fact that P^+ exists, commutes with $H_N(V)$, and that bound states are precisely the states with zero asymptotic velocity, while states with nonzero asymptotic velocity are scattering states in $\mathrm{Ran}\Omega^{\pm}$. This then implies asymptotic completeness for short-range potentials. The proof of this dichotomy builds essentially upon positive commutator or Mourre estimates. Given an interval J localized (in energy) away from any eigenvalue of any possible subcluster configuration $\mathcal{C}_1, \ldots, \mathcal{C}_K$ (called thresholds), the Mourre estimate asserts the existence of a positive constant $M > 0$ and a compact operator $R \in \mathcal{B}(\mathcal{H}^{(N)})$ such that

$$\mathbb{1}_{\mathcal{J}}\, \mathrm{i}[H_N(V), A]\, \mathbb{1}_{\mathcal{J}} \geq M \mathbb{1}_{\mathcal{J}} - R \qquad [20]$$

as a quadratic form, for some suitable operator A. This operator A is often chosen to be the dilation generator $A = (1/2)\{p \cdot x + x \cdot p\}$ or a variant thereof.

Again, the proof of asymptotic completeness for long-range potentials is still more difficult and has been carried out only for $\mu > \sqrt{3} - 1$. The additional problem is the comparison dynamics of the relative motion of the clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ in (2), which is not the free one; the clusters rather influence each other even at large distances.

For more details, see Reed and Simon (1980c) and Derezinski and Gérard (1997). See also the articles Scattering in Relativistic Quantum Field Theory: Fundamental Concepts and Tools, Scattering, Asymptotic Completeness and Bound States in this encyclopedia.

Random Schrödinger Operators

Schrödinger operators $H(V_\omega)$ on $L^2(\mathbb{R}^d)$ or $\ell^2(\mathbb{Z}^d)$ with a random potential V_ω are called random Schrödinger operators. (If $H(V_\omega)$ acts on $\ell^2(\mathbb{Z}^d)$, then the (continuum) Laplacian $-\Delta$ is replaced by the discrete Laplacian on $\mathbb{Z}^d$ defined by $[-\Delta_{\mathrm{disc}} f](x) = \sum_{\nu=1}^{d} \{2f(x) - f(x - e_\nu) - f(x + e_\nu).\}$ More precisely, given a probability space (Ω, P, μ) and a random variable $\Omega \ni \omega \mapsto V_\omega$, the family $\{H(V_\omega)\}_{\omega\in\Omega}$ defines an operator-valued random variable that we refer to as a random Schrödinger operator. Random quantum systems are physically relevant as models for amorphous materials, and for solids in very heterogenous external fields or coupled to quantized fields. Suitable ergodicity assumptions on $\omega \to V_\omega$ ensure that the domain of H_ω and even many spectral properties (in particular, the spectrum $\sigma(H(V_\omega)) \subseteq \mathbb{R}$ itself) are independent of ω P-almost surely. For example, assuming an independent, identical distribution (i.i.d.) of V_ω in the discrete case on $\mathbb{Z}^d$, one arrives at the Anderson model, which has been most thoroughly studied. Its counterpart for continuum models is a Poisson-distributed V_ω. A model which also has ergodic properties, although deterministic, is the Hofstadter or the Mathieu problem. Most research has been focused on localization, that is, spatial decay properties of the resolvent $\{H(\lambda V_\omega) - E\}^{-1}(x, y)$ of $H(\lambda V_\omega)$, as $|x - y| \to \infty$, and particularly the question of presence or absence of exponential decay (localization), as this is an important indicator for the transport properties of the material under consideration. Exponential localization of eigenstates has been established for $d = 1$ or strong disorder or sufficiently high energies $E \gg 1$. Localization is also intimately related to bounds on moments of the form $\|x^{\mu/2}\psi_t\| \leq C_\mu t^\beta$. The study of the asymptotic distribution of eigenvalues close to the lowest threshold leads to the so-called Lifshitz tails.

The reader is referred to Figotin and Pastur (1992), Cycon *et al.* (1987), and Stollmann (2001).

(Pseudo)relativistic Schrödinger Operators

Schrödinger operators of the form $H(V) = p^2 + V(x)$ do not observe the invariance principles of (special) relativity, as their derivation is based in classical (Newtonian) mechanics. The free Dirac operator $D := \boldsymbol{\alpha} \cdot \boldsymbol{p} + m\beta$ (here, α_ν and β are self-adjoint 4×4 matrices) possesses the desired relativistic invariance, but it is not semibounded, and the definition of an interacting Dirac operator is notoriously difficult (and unsolved). The replacement of the kinetic energy $(1/2m)p^2$ by the Klein–Gordon operator $\sqrt{p^2 + m^2}$ is a step towards relativistic invariance, which, at the same time, yields a positive operator. This replacement may also be viewed as the restriction of the free Dirac operator to its positive-energy subspace. The virtue of this replacement is that it immediately allows for the study of interacting N-particle operators,

$$H_N^{\mathrm{rel}}(\underline{Z}, \underline{R}) = \sum_{n=1}^{N} \left\{ \sqrt{-\Delta_n + m^2} - \sum_{k=1}^{K} \frac{Z_k}{|x_n - R_k|} \right\} + \sum_{1 \leq \ell < n \leq N} \frac{1}{|x_\ell - x_n|} \qquad [21]$$

much like in [12]. Since $\sqrt{p^2 + m^2} \sim |p|$, as $p \to \infty$, the pseudorelativistic kinetic energy $\sqrt{p^2 + m^2}$ can

balance only less severe local singularities of the potential V than the nonrelativistic kinetic energy $(1/2m)p^2$. Indeed, already the quadratic form $\sqrt{p^2+m^2}-g|x|^{-1}$ on $C_0^\infty(\mathbb{R}^3)$ associated to a hydrogen-like atom is unbounded from below if $g > 2/\pi$. Hence, the stability of matter becomes a more subtle property of pseudorelativistic matter. The relaxation of the restriction onto the positive subspace of the free Dirac operator also got into the focus of research.

For more details, we refer the reader to Thirring (1997).

See also: Deformation Quantization; Elliptic Differential Equations: Linear Theory; *h*-Pseudodifferential Operators and Applications; Localization for Quasiperiodic Potentials; Nonlinear Schrödinger Equations; Normal Forms and Semiclassical Approximation; *N*-Particle Quantum Scattering; Quantum Hall Effect; Quantum Mechanical Scattering Theory; Scattering, Asymptotic Completeness and Bound States; Stability of Matter; Stationary Phase Approximation.

Further Reading

Amrein W, Hinz A, and Pearson D (2005) *Sturm–Liouville Theory – Past and Present*. Boston: Birkhäuser.

Cycon H, Froese R, Kirsch W, and Simon B (1987) *Schrödinger Operators*, 1st edn. Berlin: Springer.

Derezinski J and Gérard C (1997) *Scattering Theory of Classical and Quantum N-Particle Systems*, Text and Monographs in Physics. Berlin: Springer-Verlag.

Dimassi M and Sjöstrand J (1999) *Spectral Asymptotics in the Semi-Classical Limit*. London Mathematical Society Lecture Notes Series, vol. 268. Cambridge: Cambridge University Press.

Erdös L and Solovej JP (2004) Uniform Lieb–Thirring inequality for the three-dimensional Pauli operator with a strong non-homogeneous magnetic field. *Annales Henri Poincaré* 5: 671–741.

Figotin A and Pastur L (1992) *Spectra of Random and Almost-Periodic Operators*. Grundlehren der Mathematischen Wissenschaften, vol. 297. Berlin: Springer-Verlag.

Kato T (1976) *Perturbation Theory of Linear Operators*, 2 edn., Grundlehren der mathematischen Wissenschaften, vol. 132. Berlin: Springer-Verlag.

Messiah A (1962) *Quantum Mechanics*, 1st edn., vol. 2. Amsterdam: North-Holland.

Rauch J and Simon B (eds.) (1997) *Quasiclassical Methods*. IMA Volumes in Mathematics and Its Applications, vol. 95. Berlin: Springer-Verlag.

Reed M and Simon B (1978) *Methods of Modern Mathematical Physics IV. Analysis of Operators*, 1st edn., vol. 4. San Diego: Academic Press.

Reed M and Simon B (1980a) *Methods of Modern Mathematical Physics: I. Functional Analysis*, 2nd edn., vol. 1. San Diego: Academic Press.

Reed M and Simon B (1980b) *Methods of Modern Mathematical Physics: II. Fourier Analysis and Self-Adjointness*, 2nd edn., vol. 2. San Diego: Academic Press.

Reed M and Simon B (1980c) *Methods of Modern Mathematical Physics: III. Scattering Theory*, 2nd edn., vol. 3. San Diego: Academic Press.

Robert D (1987) *Autour de l'Approximation Semi-Classique*, 1st edn. Boston: Birkhäuser.

Schrödinger E (1926) Quantisierung als Eigenwertproblem. *Annalen der Physik* 79: 489.

Simon B (1979) *Functional Integration and Quantum Physics*, Pure and Applied Mathematics. New York: Academic Press.

Simon B (2000) Schrödinger operators in the twentieth century. *Journal of Mathematical Physics* 41: 3523–3555.

Solovej JP (2003) The ionization conjecture in Hartree–Fock theory. *Annals of Mathematics* 158: 509–576.

Stollmann P (2001) *Caught by Disorder*. Progress in Mathematical Physics, vol. 20. Boston: Birkhäuser.

Thirring W (ed.) (1997) *The Stability of Matter: From Atoms to Stars – Selecta of Elliott H. Lieb*, 2 edn. Berlin: Springer-Verlag.

Schwarz-Type Topological Quantum Field Theory

R K Kaul and T R Govindarajan, The Institute of Mathematical Sciences, Chennai, India
P Ramadevi, Indian Institute of Technology Bombay, Mumbai, India

Introduction

Topological quantum field theories (TQFTs) provide powerful tools to probe topology of manifolds, specifically in low dimensions. This is achieved by incorporating very large gauge symmetries in the theory which lead to gauge-invariant sectors with only topological degrees of freedom. These theories are of two kinds: (1) Schwarz type and (2) Witten type.

In a Witten-type topological field theory, action is a BRST exact form, so is the stress energy tensor $T_{\mu\nu}$ so that their functional averages are zero (Witten 1988). The BRST charge is associated with a certain shift symmetry. The topological observables form cohomological classes and semiclassical approximation turns out to be exact. In four dimensions, such theories involving Yang–Mills gauge fields provide a field-theoretic representation for Donaldson invariants.

On the other hand, Schwarz-type TQFTs are described by local action functionals which are not total derivatives but are explicitly independent of metric (Schwarz 1978, 1979, 1987, Witten 1989).

The examples of such theories are topological Chern–Simons (CS) theories and BF theories.

Metric independence of the action S of a Schwarz-type gauge theory implies that stress–energy tensor is zero:

$$\frac{\delta S}{\delta g_{\mu\nu}} \equiv T_{\mu\nu} = 0$$

More generally, in the gauge-fixed version of such theories, stress–energy can be BRST exact, where BRST charge corresponds to gauge fixing in contrast to Witten-type theories where corresponding BRST charge corresponds to a combination of shift symmetry and gauge symmetry. There are no local propagating degrees of freedom; the only degrees of freedom are topological. Expectation values of metric-independent operators W are also independent of the metric:

$$\frac{\delta\langle W\rangle}{\delta g_{\mu\nu}} = 0$$

Three-dimensional CS theories are of particular interest, for these provide a framework for the study of knots and links in any 3-manifold. Pioneering indications of the fact that topological invariants can be found in such a setting came in very early when A S Schwarz demonstrated that a particular topological invariant, Ray–Singer analytic torsion (which is equivalent to combinatorial Reidemeister–Franz torsion) can be interpreted in terms of the partition function of a quantum gauge field theory (Schwarz 1978, 1979). In particular, in the weak-coupling limit of CS theory of gauge group $\mathcal{G}$ on a manifold $\mathcal{M}$, contribution from each topologically distinct flat connection (characterized by the equivalence classes of homomorphisms: $\pi_1(\mathcal{M}) \to \mathcal{G}$) to the partition function is given by metric-independent Ray–Singer torsion of the flat connection up to a phase. This phase factor is also a topological invariant of framed 3-manifold $\mathcal{M}$ (Witten 1989). It was Schwarz who first discussed CS theory as a topological field theory and also conjectured that the well-known Jones polynomial may be related to it (Schwarz 1987). In his famous paper Witten (1989) not only demonstrated this connection, but also set up a general field-theoretic framework to study the topological properties of knots and links in any arbitrary 3-manifold. In addition, this framework provides a method of obtaining some new manifold invariants. As discussed by A Achúcaro and P K Townsend, CS theory also describes gravity in three-dimensional spacetime (Carlip 2003).

BF theories in three dimensions provide another framework for field-theoretic description of topological properties of knots and links. These theories with bilinear action in fields can also be defined in higher dimensions. In particular in $D = 4$, BF theory, besides describing two-dimensional generalizations of knots and links, also provides a field-theoretic interpretation of Donaldson invariants. This provides a connection of these theories with Witten-type TQFTs of Yang–Mills gauge fields. We shall not discuss BF theories in the following and refer to the article BF Theories in this Encyclopedia.

Witten (1995) has also formulated CS theories in three complex dimensions described in terms of holomorphic 1-forms. Such a theory on Calabi–Yau spaces can be interpreted as a string theory in terms of a Witten-type topological field theory of a sigma model coupled to gravity. General topological sigma models in Batalin–Vilkovisky formalism have been constructed by Alexandrov *et al.* (1997). This is a Schwarz-type theory. However, in its gauge-fixed version, it can also be interpreted as a Witten-type theory. This construction provides a general formulation from which numerous topological field theories emerge. In particular, the Witten A and B models and also multidimensional CS theories are special cases of this construction.

In the following, we shall survey three-dimensional CS theory as a description of knots/links, indicate how manifold invariants can be constructed from invariants for framed links, and also discuss its application to three-dimensional gravity.

Three-Dimensional CS Theory with Gauge Group U(1)

The simplest Schwarz-type topological field theory is the U(1) CS theory described by the action:

$$S = -\frac{1}{8\pi}\int_{\mathcal{M}} A\,\mathrm{d}A \qquad [1]$$

where A is a connection 1-form $A = A_\mu \mathrm{d}x^\mu$ and $\mathcal{M}$ is the 3-manifold, which we shall take to be S^3 for the discussion below. The action has no dependence on the metric. Besides being the U(1) gauge invariant, it is also general coordinate invariant.

In quantum CS field theory, we are interested in the functional averages of gauge-invariant and metric-independent functionals $W[A]$:

$$\langle W[A]\rangle = \frac{1}{\mathcal{Z}}\int [\mathcal{D}A]\, W[A]\exp\{\mathrm{i}kS\}$$
$$\mathcal{Z} = \int [\mathcal{D}A]\exp\{\mathrm{i}kS\} \qquad [2]$$

This theory captures some of the simple, but interesting, topological properties of knots and links in three dimensions. For a knot K, we associate a knot

operator $\oint_K A$ which is gauge invariant and also does not depend on the metric of the 3-manifold. Then for a link made of two knots K_1 and K_2, we have the loop correlation function $\langle \oint_{K_1} A \oint_{K_2} A \rangle$, which can be evaluated in terms of two-point correlator $\langle A_\mu(x) A_\nu(y) \rangle$ in R^3 (with flat metric). This correlator in Lorentz gauge ($\partial_\mu A^\mu = 0$) is:

$$\langle A_\mu(x) A_\nu(y) \rangle = \frac{\mathrm{i}}{k} \epsilon_{\mu\nu\rho} \frac{(x-y)^\rho}{|x-y|^3}$$

so that for two distinct knots K_1 and K_2

$$\left\langle \oint_{K_1} A \oint_{K_2} A \right\rangle = \frac{4\pi \mathrm{i}}{k} \mathcal{L}(K_1, K_2) \qquad [3]$$

where

$$\mathcal{L}(K_1, K_2) = \frac{1}{4\pi} \oint_{K_1} \oint_{K_2} \mathrm{d}x^\mu \mathrm{d}y^\nu \epsilon_{\mu\nu\rho} \frac{(x-y)^\rho}{|x-y|^3}$$

This integral is the well-known topological invariant called "Gauss linking number" of two distinct closed curves. It is an integer measuring the number of times one knot K_1 goes through the other knot K_2. Linking number does not depend on the location, size, or shape of the knots. In electrodynamics, it has the physical interpretation of work done to move a monopole around a knot while electric current runs through the other knot.

Abelian CS theory also provides a field-theoretic representation for another topological quantity called "self-linking number," also known as "framing number," of the knot. It is related to the functional average of $\langle \oint_K A \oint_K A \rangle$ where two loop integrals are over the same knot. Coincidence singularity is avoided by a topological loop-splitting regularization. For a knot K given by $x^\mu(s)$ parametrized along the length of the knot by s, we associate another closed curve K_f given by $y^\mu(s) = x^\mu(s) + \epsilon n^\mu(s)$, where ϵ is a small parameter and $n^\mu(s)$ is a principal normal to the curve at s. The coincidence limit is then obtained at the end by taking the limit $\epsilon \to 0$. Such a limiting procedure is called framing and knot K_f is the "frame" of knot K. Linking number of the knot K and its frame K_f is the self-linking number of the knot:

$$S\mathcal{L}(K, n^\mu) = \frac{1}{4\pi} \oint \oint \mathrm{d}x^\mu \, \mathrm{d}y^\nu \frac{\epsilon_{\mu\nu\rho}(x-y)^\rho}{|x-y|^3}$$

Hence coincidence two loop correlator is

$$\left\langle \oint_K A \oint_K A \right\rangle = \frac{4\pi \mathrm{i}}{k} S\mathcal{L}(K, n^\mu) \qquad [4]$$

Notice that the self-linking number of a knot is independent of the regularization parameter ϵ, but does depend on the topological character of the normal vector field $n^\mu(s)$. It is also related to two geometric quantities called "twist" $T(K)$ and "writhe" $w(K)$ through a theorem due to Calugareanu:

$$S\mathcal{L}(K) = T(K) + \omega(K) \qquad [5]$$

where

$$T(K) = \frac{1}{2\pi} \oint_K \mathrm{d}s \, \epsilon_{\mu\nu\rho} \frac{\mathrm{d}x^\mu}{\mathrm{d}s} n^\nu \frac{\mathrm{d}x^\rho}{\mathrm{d}s}$$

$$\omega(K) = \frac{1}{4\pi} \oint_K \mathrm{d}s \oint_K \mathrm{d}t \, \epsilon_{\mu\nu\rho} \frac{\mathrm{d}e^\mu}{\mathrm{d}s} \frac{\mathrm{d}e^\nu}{\mathrm{d}t} e^\rho$$

Here

$$e^\mu(s,t) = \frac{y^\mu(t) - y^\mu(s)}{|y(t) - y(s)|}$$

is a unit map from $K \otimes K \longrightarrow S^2$ and $n^\mu(s)$ is a normal unit vector field. $T(K)$ and $\omega(K)$ are not in general integers and represent the amount of twist and coiling of the knot. These are not topological invariants but their sum, self-linking number, is indeed always an integer and a topological invariant. This result has found interesting applications in the studies of the action of enzymes on circular DNA.

Nonabelian CS Theories

Nonabelian CS theories provide far more information about the topological properties of the manifolds as well as knots and links.

Nonabelian CS theory in a 3-manifold $\mathcal{M}$ (which as in last section is taken to be S^3) is described by the action functional

$$S = \frac{1}{4\pi} \int_{\mathcal{M}} \mathrm{tr}\left(A \wedge \mathrm{d}A + \tfrac{2}{3} A \wedge A \wedge A\right) \qquad [6]$$

where A is a gauge field 1-form which takes its value in the Lie algebra $\mathcal{LG}$ of a compact semisimple Lie group $\mathcal{G}$. For example, we may take this group to be SU(N) and $A = A^a T^a$, where T^a is the fundamental N-dimensional representation with $\mathrm{tr} T^a T^b = -1/2\delta^{ab}$. Under homotopically nontrivial gauge transformations this action is not invariant, but changes by an amount $2\pi n$ where integers n are the winding numbers characterizing the gauge transformations which fall in homotopic classes given by $\Pi_3(\mathcal{G}) = \mathcal{Z}$ for a compact semisimple group $\mathcal{G}$. However, for quantum theory what is relevant is $\exp[\mathrm{i}kS]$ which is invariant even under homotopically nontrivial gauge transformations provided the coupling k takes integer values. This quantized nature of the coupling was pointed out by Deser *et al.* (1982a, b) (and also they were first to introduce the nonabelian CS term as a gauge-invariant topological

mass term in gauge theories). So for integer k, the quantum field theory we discuss here is gauge invariant.

The topological operators are Wilson loop operators for an oriented knot K:

$$W_R[K] = \mathrm{tr}\, P \exp \oint_K A_R \qquad [7]$$

where $A_R = A^a T_R^a$ with T_R^a as the representation matrices of a finite-dimensional representation R of the $\mathcal{LG}$. P stands for the path ordering of the exponential. The observable Wilson link operator for a link $L = \bigcup_1^n K_i$, carrying representations R_i on the respective component knots, is

$$W_{R_1 R_2 \cdots R_n}[L] = \prod_1^n W_{R_i}[K_i] \qquad [8]$$

Expectation values of these operators are:

$$V_{R_1, R_2 \cdots R_n}[L] = \frac{\int [\mathcal{D}A] W_{R_1 \cdots R_n}[L] e^{ikS}}{\int [\mathcal{D}A] e^{ikS}} \qquad [9]$$

The measure $[\mathcal{D}A]$ has to be metric independent. These expectation values depend not only on the isotopy of the link L but also on the set of the representations $\{R_i\}$. These can be evaluated in principle nonperturbatively. For example, when $\mathcal{LG} = \mathrm{su}(N)$ and each of the component knot of the links carries the fundamental N-dimensional representation, the Wilson link expectation values satisfy a recursion relation involving three link diagrams which are identical except for one crossing where they differ as over crossing (L_+), under crossing (L_-), and no crossing (L_0) as shown in the Figure 1.

The expectation values of these links are related as (Witten 1989):

$$q^{N/2} V_N[L_+] - q^{-N/2} V_N[L_-] = \left(q^{1/2} - q^{-1/2}\right) V_N[L_0] \qquad [10]$$

where

$$q = \exp\left(\frac{2\pi i}{k+N}\right)$$

This is precisely the well-known skein relation for the HOMFLY polynomial. The famous Jones one-variable polynomial (whose two-variable generalization is the HOMFLY polynomial) corresponds to the case of spin-1/2 representation of SU(2) CS theory: $V_2[L] =$ Jones polynomial $[L]$, up to an overall normalization. These skein relations are sufficient to recursively find all the expectation values of links with only fundamental representation on the components. To obtain invariants for any other representation, more general methods have to be developed. A complete and explicit solution of the CS field theory is thus obtained. One such method has been reviewed in Kaul (1999). The method makes use of the following important statement:

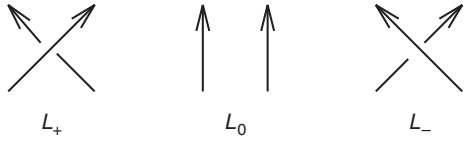


Figure 1 Skein related links.

Proposition: *CS theory on a 3-manifold $\mathcal{M}$ with boundary Σ is described by a WZNW (Wess–Zumino–Novikov–Witten) conformal field theory (CFT) on the boundary* (Figure 2).

Using the same identification, functional average for Wilson lines ending at n points on the boundary Σ is obtained from WZNW field theory on the boundary with n punctures carrying representations R_i (Figure 3):

We can represent CS functional integral as a vector (Witten 1989) in the Hilbert space $\mathcal{H}$ associated with the n-point vacuum expectation values of primary fields in WZNW conformal field theory on the boundary Σ. Next, to obtain a complete and explicit nonperturbative solution of the CS theory, the theory of knots and links and their connection to braids is invoked.

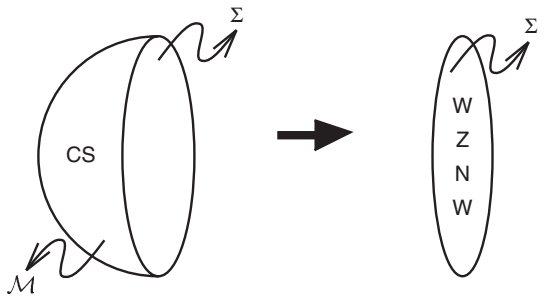


Figure 2 Relation of CS to CFT.

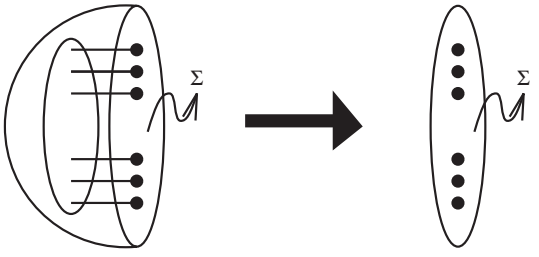


Figure 3 CS functional integrals with Wilson lines and CFT on punctured boundary.

Knots/Links and Braids

Braids have an intimate connection with knots and links which can be summarized as follows:

1. An n-braid is a collection of nonintersecting strands connecting n points on a horizontal rod to n points on another horizontal rod below strictly excluding any backward traversing of the strands. A general braid can be written as a word in terms of elementary braid generators.
2. We associate representations R_i of the group with the strands as their colors. We also put an orientation on each strand. When all the representations are identical and also all strands are unoriented, we get ordinary braids, otherwise we get colored oriented braids.
3. The colored oriented braids form a groupoid where product of the different braids is obtained by joining them with both colors and orientations matching on the joined strands. Unoriented monochromatic braids form a group.
4. A knot/link can be formed from a given braid by a process called platting. We connect adjacent strands namely the $(2i+1)$th strand to $2i$th strand carrying the same color and opposite orientations in both the rods of an even-strand braid (Figure 4a).

   There is a theorem due to Birman which states that all colored oriented knots/links can be obtained through platting. This construction is not unique.
5. There is another construction associated with braids which relates them to knots and links. We obtain a closure of a braid by connecting the ends of the first, second, third, ... strands from above to those of the respective first, second, third,... strands from below as shown in the Figure 4b. There is theorem due to Alexander which states that any knot or link can be obtained as a closure of a braid, though again not uniquely.

Link Invariants

This connection of braids to knots and links can be used to construct link invariants, say in S^3. To do so, two nonintersecting 3-balls are removed from the 3-manifold S^3 to obtain a manifold with two S^2 boundaries. Then we arrange $2n$ Wilson lines of, say SU(N) CS theory, as a $2n$-strand oriented braid carrying representations R_i in this manifold. The CS functional integral over this manifold is a state in the tensor product of the Hilbert spaces $\mathcal{H}_1 \otimes \mathcal{H}_2$ associated with conformal field theory on the two boundaries. These boundaries have $2n$ punctures carrying the set of representations $\{R_i\}$ and $\{R'_i\}$, respectively, the two sets being permutations of each other. This state can be expanded in terms of some convenient basis given by the conformal blocks for the $2n$-point correlation functions of $\mathrm{SU}(N)_k$ WZNW conformal field theory. The duality of these correlation functions represents the transformation between different bases for the Hilbert space. Their monodromy properties allow us to write down representations of the braid generators. Since an arbitrary braid is just a word in terms of these generators, this construction provides us a matrix representation $\mathcal{B}(\{R_i\},\{R'_j\})$ for the colored oriented braid in the manifold with two S^2 boundaries. Then we plat this braid by gluing two balls B_1 and B_2 with Wilson lines as shown in Figure 5.

Each of the two caps again represents a state $|\psi(\{R_j\})\rangle$ in the Hilbert space associated with the conformal field theory on punctured boundary (S^2). Platting of the braid then simply is the matrix element of braid representation $\mathcal{B}(\{R_i\},\{R'_j\})$ with respect to these states $|\psi(\{R_i\})\rangle$ and $|\psi(\{R'_j\})\rangle$ corresponding to two caps B_1, B_2. Thus, for a link in S^3 the invariant is given by the following theorem:

Theorem *The vacuum expectation value of Wilson loop operator of a link L constructed from platting of a colored oriented 2n braid with representation* $\mathcal{B}(\{R_i\},\{R'_j\})$ *is given by* (Kaul 1999):

$$V[L] = \langle\psi(\{R_i\})|\mathcal{B}(\{R_i\},\{R'_j\})|\psi(\{R'_j\})\rangle \qquad [11]$$

This theorem can be used to calculate the invariant for any arbitrary link. For an unknot U

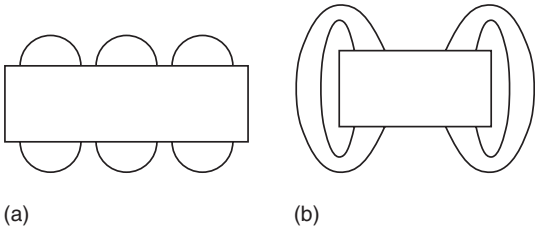


Figure 4 (a) Platting and (b) closure of braids.

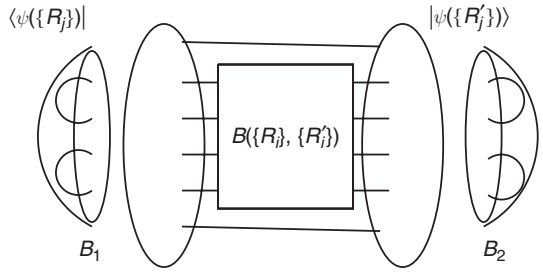


Figure 5 Construction of the link invariant.

carrying an N-dimensional representation in an SU(N) CS theory, the knot invariant is:

$$V_N[U] = [N], \quad \text{where } [N] = \frac{q^{N/2} - q^{-N/2}}{q^{1/2} - q^{-1/2}}$$

Wilson link expectation values calculated this way depend on the regularization, that is, the definition of framing used in defining coincident loop correlators. One such regularization usually used is the standard framing, where the frame for every knot is so chosen that its self-linking number is zero.

The procedure outlined here has been used for explicit computations of knot/link invariants. This has led to answers to several questions of knot theory. One such question relates to distinguishing chirality of knots (Kaul 1999). In this context, newer invariants constructed with arbitrary representations living on the knots are more powerful than the older polynomial invariants. For example, invariants with spin-3/2 representation in an SU(2) CS theory are sensitive to chirality of many knots which otherwise is not detected by Jones, HOMFLY, and Kauffman polynomials. However, invariants obtained from CS theories do not distinguish all chiral knots. There is a class of links known as "mutants" which are not distinguished by CS link invariants (Kaul 1999). A mutant link is obtained by removing a portion of weaving pattern in a link and then gluing it back after rotating it about any one of three orthogonal axes by an amount π.

The CS invariants of knots and links can also be used to construct special 3-manifold invariants. Hence, CS theory provides an important tool to study these.

Manifold Invariants from CS Theory

Different 3-manifolds can be constructed through a procedure called "surgery of framed knots and links" in S^3 (Lickorish–Wallace theorem). This construction is not unique. That is, there are many framed knots and links which give the same manifold. However, rules of this equivalence are known: these are called "Kirby moves."

Classification of 3-manifolds would involve finding a method of associating a quantity with the manifold obtained by surgery on the corresponding framed knot/link on S^3. If the Kirby moves on the framed knot/link leave this quantity unchanged, then it is a 3-manifold invariant. Knot/link invariants of nonabelian CS theories provide a method of finding such 3-manifold invariants. Equivalently, this procedure gives an algebraic meaning to the surgery construction of 3-manifolds. Details of this method for generating manifold invariants are given in Kaul (1999) and Kaul and Ramadevi (2001).

Surgery of Framed Knots/Links and Kirby Moves

As discussed earlier, frame of a knot K is an associated closed curve K_f going along the length of the knot wrapping around it certain number of times. Self-linking number (also called framing number) is equal to the linking number of the knot with its frame. There are several ways of fixing this framing. The "standard" framing is one in which the frame number of the knot, that is, the linking number of the knot and its frame is zero. On the other hand, "vertical" framing is obtained by choosing the frame vertically above the knot projected on to a plane. In such a frame, the framing number of a knot is the same as its crossing number. In constructing the 3-manifold invariants from CS theories, we need vertical framing. The framing number may be denoted by writing the integer by the side of knot. We denote a framed r-component link by $[L, \boldsymbol{f}]$ where framing $\boldsymbol{f} = (n(1), n(2), \ldots, n(r))$ is a set of integers denoting the framing number of component knots $K_1, K_2, \ldots, K_r$ in the link L.

According to the Lickorish–Wallace theorem, surgery over links with vertical framing in S^3 yields all the 3-manifolds. This surgery is performed in the following way.

Take a framed r-component link $[L, \boldsymbol{f}]$ in S^3. Thicken the component knots $K_1, K_2, \ldots, K_r$ such that the solid tubes $N_1, N_2, \ldots, N_r$ so obtained are nonintersecting. Then the compliment $S^3 - (N_1 + N_2 + \cdots + N_r)$ will have r toral boundaries. On the ith toral boundary, we imagine an appropriate curve winding $n(i)$ times around the meridian and once along the longitude. Perform a modular transformation so that this curve bounds a disk. This construction is done with each of the toral boundaries. The tubes $N_1, N_2, \ldots, N_r$ are then glued back in to the respective gaps. This surgery thus yields a new 3-manifold. This construction is not unique. The rules of equivalence for surgery on framed knots/links in S^3 are two independent Kirby moves.

Kirby move I Take an arbitrary r-component framed link $[L, \boldsymbol{f}]$ in S^3 and consider a curve C with framing number +1 going around the unlinked strands of L as in Figure 6a. We refer to this $(r+1)$-component link as $H[X]$, where X represents a weaving pattern of the strands. Kirby move I consists of twisting the disk enclosed by C in the clockwise direction from below by an amount 2π. This twisting thereby introduces new crossings

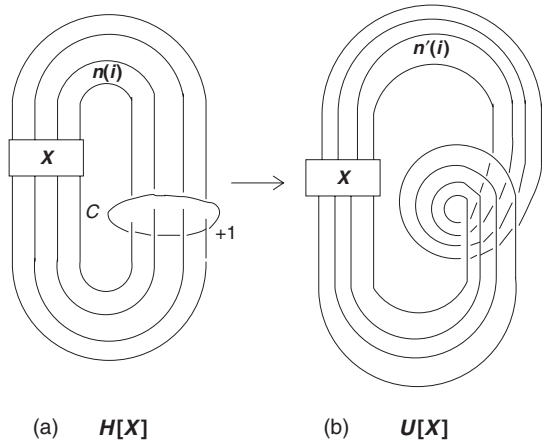


Figure 6 Kirby move I.

between the curve C and the strands enclosed by it. Then the curve C is removed giving us a new r-component link $U[X]$ of Figure 6b. Framing numbers $n'(i)$ of the component knots in link $U[X]$ are related to the framing number $n(i)$ of framed link $[L, f]$ as $n'(i) = n(i) - (\mathcal{L}(K_i, C))^2$, where $\mathcal{L}(K_i, C)$ is the linking number of knot K_i and closed curve C. The surgery of the framed links in Figures 6a and 6b will give the same 3-manifold.

Inverse Kirby move I involves removal of a curve C with framing number -1 (instead of $+1$) after making one complete anticlockwise twist from below on the disk enclosed by C. In the process the unlinked strands get twisted in the anticlockwise direction leading to changed framing numbers $n'(i) = n(i) + (\mathcal{L}(K_i, C))^2$ of the component knots K_i.

Kirby move II This move consists of removing a disjoint unknot C with framing -1 from framed link $[L, f]$ without changing the rest of the link as in Figure 7. Surgery of the two links in Figure 7 will give the same manifold.

Inverse Kirby move II involves removal of a disjoint unknot with framing $+1$ (instead of -1) from a framed link.

3-Manifold Invariants

Now a 3-manifold invariant can be constructed by an appropriate combination of the invariants of framed links in such a way that this algebraic expression is unchanged under the Kirby moves. We need for this purpose invariants for links in S^3 with vertical framing.

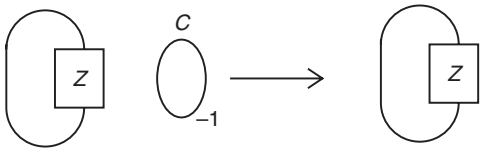


Figure 7 Kirby move II.

Let $\mathcal{M}$ be the manifold obtained from surgery of an r-component framed link $[L, f]$ in S^3. Then a manifold invariant $\hat{F}^{(\mathcal{G})}[\mathcal{M}]$ is given as a linear combination of the framed link invariants $V^{(\mathcal{G})}_{R_1,\ldots,R_r}[L, f]$, with representations $R_1, R_2, \ldots, R_r$ living on component knots, obtained from CS theory based on a compact semisimple group $\mathcal{G}$:

$$\hat{F}^{(\mathcal{G})}[\mathcal{M}] = \alpha^{-\sigma[L,f]} \sum_{R_1,\ldots R_r} \left(\prod_{i=1}^{r} \mu_{R_i} \right) \times V^{(\mathcal{G})}_{R_1,R_2,\ldots,R_r}[L, f] \qquad [12]$$

Here $\sigma[L, f]$ is the signature of the linking matrix and $\mu_{R_i} = S_{0R_i}$, $\alpha = e^{i\pi c/4}$, where c is the central charge of the associated WZNW conformal field theory and S_{0R_i} denotes the matrix element of the modular matrix S. General S-matrix elements for any compact group are given by

$$S_{R_1R_2} = (-i)^{(d-r)/2} |L_\omega/L|^{-1/2} (k + C_v)^{-1/2} \times \sum_{\omega \in W} \epsilon(\omega) \exp\left(\frac{-2\pi i}{k + C_v} (\omega(\Lambda_{R_1} + \rho), \Lambda_{R_2} + \rho) \right)$$

where W denotes the Weyl group and its elements ω are words constructed using the generator s_{α_i} – that is, $\omega = \prod_i s_{\alpha_i}$ and $\epsilon(\omega) = (-1)^{\ell(\omega)}$ with $\ell(\omega)$ as length of the word. Here Λ_{R_i}'s denotes the highest weights of the representations R_i's and ρ is the Weyl vector. The action of the Weyl generator s_α on a weight Λ_R is

$$s_\alpha(\Lambda_R) = \Lambda_R - 2\alpha \frac{(\Lambda_R, \alpha)}{(\alpha, \alpha)}$$

and $|L_\omega/L|$ is the ratio of weight and coroot lattices (equal to the determinant of the Cartan matrix for simply laced algebras). Also C_v is quadratic Casimir invariant for the adjoint representation.

It is important to stress that the expression $\hat{F}^{(\mathcal{G})}[\mathcal{M}]$ is unchanged under both Kirby moves I and II (for detailed proof, see Kaul (1999) and Kaul and Ramadevi (2001)). Notice that for every compact gauge group, we have a new 3-manifold invariant.

Few examples of 3-manifolds Table 1 lists the algebraic expressions of this invariant calculated explicitly from the formula in eqn [12] for a few 3-manifolds. All these examples can be constructed by surgery on an unknot $U(f)$ with different frame numbers f.

In Table 1 $\mathcal{L}[p, q]$ stands for Lens spaces of the type (p, q) and C_R is the quadratic Casimir invariant

Table 1 Invariants for some simple manifolds

$U(f)$	$\mathcal{M}$	$\hat{F}^{(\mathcal{G})}[\mathcal{M}]$
$U(0)$	$S^2 \times S^1$	$1/S_{00}$
$U(\pm 1)$	S^3	1
$U(+2)$	RP^3	$\alpha^{-1} \sum_R \frac{S_{0R} q^{2C_R} S_{0R}}{S_{00}}$
$U(+p)$	$\mathcal{L}[p,1]$	$\alpha^{-1} \sum_R \frac{S_{0R} q^{pC_R} S_{0R}}{S_{00}}$

for representation R of the Lie algebra of the gauge group $\mathcal{G}$.

Partition function of a CS theory on $\mathcal{M}$ is also an invariant characterizing the 3-manifold. This has been calculated for several manifolds by different methods. Invariant $\hat{F}^{(\mathcal{G})}[\mathcal{M}]$ listed above for various manifolds is related to the CS partition function $Z^{(\mathcal{G})}[\mathcal{M}]$: $\hat{F}^{(\mathcal{G})}[\mathcal{M}] = S_{00}^{-1} Z^{(\mathcal{G})}[\mathcal{M}]$. So the method of constructing 3-manifold invariants above can also be used to calculate the partition function of CS theories.

3D Gravity and CS Theory

Three-dimensional CS theory also provides a description of gravity. The 3D gravity including cosmological constant has been first discussed by Deser and Jackiw (1984). The action with cosmological constant $\Lambda = \pm 1/\ell^2$ is:

$$S = \frac{1}{16\pi G} \int_{\mathcal{M}} \mathrm{d}^3x \sqrt{-g}(R - 2\Lambda) \qquad [13]$$

G is the Newton's constant, $g_{\mu\nu}$ is the metric on the 3-manifold $\mathcal{M}$, and R is scalar curvature. Solutions of Einstein equations of motion have a constant positive (negative) curvature if Λ is positive (negative). It is also well known that there are no dynamical degrees of freedom for gravity in dimensions $D \leq 3$; it is indeed described by topological field theories. The gravity action above can be rewritten as a CS gauge theory in first-order formulation (Carlip 2003). For triads e_μ^a and spin connection ω_μ^a of Euclidean gravity, we define 1-forms $e = e_\mu^a T^a \mathrm{d}x^\mu, \omega = \omega_\mu^a T^a \mathrm{d}x^\mu$, which have values in the Lie algebra of SU(2) whose generators are $T^a = \mathrm{i}\sigma^a/2$ with σ^a as three Pauli matrices. In terms of these we define two gauge field 1-forms A and $\bar{A}$ as:

$$A = \left(\frac{\mathrm{i}e}{\ell} + \omega\right), \qquad \bar{A} = \left(\frac{\mathrm{i}e}{\ell} - \omega\right)$$

Then the Euclidean gravity action can be written in terms of two CS actions, $S_{\mathrm{CS}}[A]$ and $S_{\mathrm{CS}}[\bar{A}]$, as

$$S = kS_{\mathrm{CS}}[A] - kS_{\mathrm{CS}}[\bar{A}] \qquad [14]$$

where the coupling constant $k = \ell/(4G)$ for negative cosmological constant $\Lambda = -1/\ell^2$. The gauge group for this theory is $\mathrm{SL}(2, C)$. Infinitesimal diffeomorphisms are described by field-dependent gauge transformations. The corresponding gauge group for Minkowski gravity with negative cosmological constant Λ is $\mathrm{SO}(2, R) \otimes \mathrm{SO}(2, R)$. For positive Λ, one gets SO(3, 1) and SO(4) for Minkowski and Euclidean metrics, respectively. For $\Lambda = 0$, we have ISO(2, 1) (ISO(3)) as the gauge group for Minkowski (Euclidean) gravity. Hence, the sign of cosmological constant determines the gauge group of the CS theory.

Identification of 3D gravity with CS theory can be used with some advantage to find the partition function for a black hole in 3D gravity with negative cosmological constant. This in turn yields an expression for entropy of the black hole.

BTZ Black Hole and Its Partition Function

Only for negative Λ we have a black hole solution of the Einstein's equations. This solution, known as the BTZ black hole (Carlip 2003), in Euclidean gravity is given by the metric

$$\mathrm{d}s_{\mathrm{E}}^2 = \left(-M + \frac{r^2}{l^2} - \frac{J^2}{4r^2}\right)\mathrm{d}\tau^2 + \left(-M + \frac{r^2}{l^2} - \frac{J^2}{4r^2}\right)^{-1} \mathrm{d}r^2 + r^2\left(\mathrm{d}\phi - \frac{J}{2r}\mathrm{d}\tau\right)^2$$

It is specified by two parameters M and J (the mass and angular momentum). By a coordinate transformation, this metric can be rewritten as $\mathrm{d}s_{\mathrm{E}}^2 = (l^2/z^2)(\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2)$, with $z > 0$. This is the 3D upper-half hyperbolic space and can be rewritten using spherical polar coordinates as

$$\mathrm{d}s_{\mathrm{E}}^2 = \frac{l^2}{R^2 \sin^2 \chi}\left(\mathrm{d}R^2 + R^2\mathrm{d}\theta^2 + R^2 \sin^2\theta \mathrm{d}\chi^2\right)$$

We have the identifications $(R, \theta, \chi) \sim (R \exp\{2\pi r_+/l\}, \theta + \{2\pi r_-/l\}, \chi)$ where r_+ and r_- are the outer and inner horizon radii, respectively. It is clear from this identification that topologically the metric corresponds to a solid torus. Functional integral over this manifold represents a state in the Hilbert space specified by the mass and angular momentum. It is the microcanonical ensemble partition function and its logarithm is the entropy of the black hole.

To evaluate this partition function, the connection 1-form is kept at a constant value on the toroidal boundary through a gauge transformation. We

define local coordinates on the torus boundary $z = x + \tau y$ such that $\int_a dz = 1$, $\int_b dz = \tau$, where a (b) stands for the contractible (noncontractible) cycle of solid torus and $\tau = \tau_1 + i\tau_2$ is the modular parameter of the boundary torus. Then connection describing the black hole is

$$A = \left(\frac{-i\pi\tilde{u}}{\tau_2}d\bar{z} + \frac{i\pi u}{\tau_2}dz\right)T^3 \qquad [15]$$

where u and $\tilde{u}$ are canonically conjugate with commutation relation: $[\tilde{u}, u] = (2/\pi)\tau_2(k+2)^{-1}$. These are related to black hole parameters through holonomies of gauge field A around the a- and b-cycles (for a classical black hole solution $\Theta = 2\pi$):

$$u = -\frac{i}{2\pi}\left(-i\Theta\tau + \frac{2\pi(r_+ + i|r_-|)}{l}\right)$$

$$\tilde{u} = -\frac{i}{2\pi}\left(-i\Theta\bar{\tau} + \frac{2\pi(r_+ + i|r_-|)}{l}\right)$$

For a fixed value of connection, namely u, the functional integral is described by a state ψ_0 with no Wilson line in the bulk. The states with Wilson line carrying spin $j/2$ are given by Labastida and Ramallo:

$$\psi_j(u,\tau) = \exp\left\{\frac{\pi k}{4\tau_2}u^2\right\}\chi_j(u,\tau)$$

where the Weyl–Kac characters for affine su(2)

$$\chi_j(u,\tau) = \frac{\Theta^{(k+2)}_{j+1}(u,\tau) - \Theta^{(k+2)}_{-j-1}(u,\tau)}{\Theta^2_1(u,\tau) - \Theta^2_{-1}(u,\tau)}$$

and Θ functions are defined by

$$\Theta^k_\mu(u,\tau) = \sum_{n\in\mathcal{Z}}\exp\left\{2\pi i k\left[\left(n+\frac{\mu}{2k}\right)^2\tau + \left(n+\frac{\mu}{2k}\right)u\right]\right\}$$

Given the collection of states ψ_j, we write the partition function by choosing an appropriate ensemble for fixed mass and angular momentum. This black hole partition function is:

$$Z_{BH} = \int d\mu(\tau,\bar{\tau})\left|\sum_{j=0}^{k}(\psi_j(0,\tau))^*\psi_j(u,\tau)\right|^2$$

where modular invariant measure is $d\mu(\tau,\bar{\tau}) = d\tau\, d\bar{\tau}/\tau_2^2$. This integral can be worked out for large black hole mass and zero angular momentum in saddle-point approximation. The computation yields (Govindarajan *et al.* 2001):

$$Z_{BH} = \frac{l^2}{r_+^2}\sqrt{\frac{8r_+G}{\pi l^2}}\exp\left(\frac{2\pi r_+}{4G}\right) + \cdots \qquad [16]$$

This gives not only the leading Bekenstein–Hawking behavior of the black hole entropy S but also a subleading logarithmic term:

$$S = \ln Z_{BH} = \frac{2\pi r_+}{4G} - \frac{3}{2}\ln\frac{2\pi r_+}{4G} + \cdots$$

This is an interesting application of CS theory to 3D gravity. In fact, three-dimensional CS theory also has applications in the study of black holes in four-dimensional gravity: the boundary degrees of freedom of a black hole in 4D are also described by an SU(2) CS theory. This allows a calculation of the degrees of freedom of, for example, Schwarzschild black hole. For large area black holes, this in turn results in an expression for the entropy which, besides a Bekenstein–Hawking area term, has a logarithmic area correction with same coefficient –3/2 as above. This suggests a universal, dimension-independent, nature of the these logarithmic corrections.

See also: BF Theories; The Jones Polynomial; Knot Theory and Physics; Large-*N* and Topological Strings; Quantum 3-Manifold Invariants; Topological Quantum Field Theory: Overview.

Further Reading

Alexandrov M, Konstsevich M, Schwarz A, and Zaboronsky O (1997) The geometry of the master equations and topological quantum field theory. *International Journal of Modern Physics A* 12: 1405–1430.

Atiyah M (1989) *The Geometry and Physics of Knots*. Cambridge: Cambridge University Press.

Carlip S (2003) *Quantum Gravity in* $2+1$ *Dimensions*, Cambridge Monographs on Mathematical Physics, Cambridge: Cambridge University Press.

Deser S and Jackiw R (1984) Three-dimensional cosmological gravity: dynamics of constant curvature. *Annals of Physics* 153: 405–416.

Deser S, Jackiw R, and Templeton S (1982a) Three-dimensional massive gauge theory. *Physical Review Letters* 48: 975–978.

Deser S, Jackiw R, and Templeton S (1982b) Topological massive gauge theories. *Annals of Physics NY* 140: 372–411.

Govindarajan TR, Kaul RK, and Suneeta V (2001) Logarithmic correction to the Bekenstein–Hawking entropy of the BTZ black hole. *Classical and Quantum Gravity* 18: 2877–2886.

Kaul RK (1999) Chern–Simons theory, knot invariants, vertex models and three-manifold invariants. In: Kaul RK, Maharana J, Mukhi S, and Kalyana Rama S (eds.) *Frontiers of Field Theory, Quantum Gravity and Strings*, Horizons in World Physics, vol. 227, pp. 45–63. New York: NOVA Science Publishers.

Kaul RK and Ramadevi P (2001) Three-manifold invariants from Chern–Simons field theory with arbitrary semi-simple gauge groups. *Communications in Mathematical Physics* 217: 295–314.

Schwarz AS (1978) The partition function of degenerate quadratic functional and Ray–Singer invariants. *Letters in Mathematical Physics* 2: 247–252.

Schwarz AS (1979) The partition function of a degenerate functional. *Communications in Mathematical Physics* 67: 1–16.

Schwarz AS (1987) *New Topological Invariants in the Theory of Quantized Fields*. Abstracts in the Proceedings of International Topological Conference, Baku, Part II.

Witten E (1988) Topological quantum field theory. *Communications in Mathematical Physics* 117: 353–386.

Witten E (1989) Quantum field theory and the Jones polynomial. *Communications in Mathematical Physics* 121: 351–399.

Witten E (1995) Chern–Simons gauge theory as a string theory. *Progress in Mathematics* 133: 637–678.

Seiberg–Witten Theory

Siye Wu, University of Colorado, Boulder, CO, USA

Introduction

Gauge theory is the cornerstone of the standard model of elementary particles. The original motivation for studying supersymmetric gauge theories was phenomenological (such as the hierarchy problem). They display a large number of interesting phenomena and become the models for the dynamics of strongly coupled field theories. They also offer valuable insights to nonsupersymmetric models. In $N=1$ gauge theory, the low-energy effective superpotential is holomorphic both in the superfields and in the coupling constants. This powerful holomorphy principle, together with symmetry and various limits, often determines the effective superpotential completely. Such theories often have quantum moduli spaces where the classical singularities are smoothed out, continuous interpolation between Higgs and confinement phases, massless composite mesons and baryons, and dual theories weakly coupled at low energy. For $N=2$ pure gauge theory, the low-energy effective theory is an abelian gauge theory in which both the kinetic term and the coupling constant are determined by a holomorphic prepotential. The electric–magnetic duality is in the ambiguity of the low-energy description. Much physical information, such as the coupling constant, the Kähler metric on the quantum moduli, the monodromy around the singularities, can be incorporated in a family of elliptic curves. This low-energy exact solution is also useful to topological field theory that can be obtained from the $N=2$ theory by twisting. Much of the above was the work of Seiberg and Witten in the mid-1990s. In this article, we review some of the fascinating aspects of $N=1$ and $N=2$ supersymmetric gauge theories.

$N=1$ Gauge Theory and Seiberg Dualities

$N=1$ Yang–Mills Theory and QCD

Let G be a compact Lie group and let P be a principal G-bundle over the Minkowski space $\mathbb{R}^{3,1}$. In pure gauge theory, the dynamical variable is a connection A in P; two connections are equivalent if they are related by a gauge transformation. Let $F \in \Omega^2(\mathbb{R}^{3,1}, \mathrm{ad}\, P)$ be the curvature of A. It decomposes into the self-dual and anti-self-dual parts, that is, $F = F^+ + F^-$, where $F^{\pm} = (1/2)(F \mp \sqrt{-1} * F)$. With a suitably normalized nondegenerate bilinear form $\langle \cdot, \cdot \rangle$ on the Lie algebra $\mathfrak{g}$, the classical action is

$$\begin{aligned} S_{\mathrm{YM}}[A] &= \int_{\mathbb{R}^{3,1}} -\frac{1}{2g^2}\langle F \wedge *F\rangle + \frac{\theta}{16\pi^2}\langle F \wedge F\rangle \\ &= \int_{\mathbb{R}^{3,1}} -\frac{\tau}{8\pi}\langle F^+ \wedge F^+\rangle - \frac{\bar{\tau}}{8\pi}\langle F^- \wedge F^-\rangle \end{aligned}$$

Here $g > 0$ is the coupling constant and $\theta \in \mathbb{R}$, the θ angle, and

$$\tau = \frac{\theta}{2\pi} + \frac{4\pi\sqrt{-1}}{g^2}$$

is a complex number in the upper-half plane that incorporates both. Classically, the theory is conformally invariant and the dynamics is independent of the θ-term. At the quantum level, $\theta (\mathrm{mod} 2\pi)$ appears in the path integral and parametrizes inequivalent vacua. The coupling constant runs as energy μ varies, satisfying the renormalization group equation

$$\mu \frac{\mathrm{d}g}{\mathrm{d}\mu} = -\frac{b_0}{(4\pi)^2} g^3 + o(g^5)$$

where the right-hand side is called the β-function $\beta(g)$. This introduces, when $b_0 \neq 0$, a mass scale Λ given by

$$(\Lambda/\mu)^{b_0} = \mathrm{e}^{-8\pi^2/g(\mu)^2}$$

up to one-loop. Consequently, the classical scale invariance is lost. It is convenient to redefine Λ as a complex quantity such that

$$(\Lambda/\mu)^{b_0} = \mathrm{e}^{2\pi\sqrt{-1}\tau(\mu)}$$

For pure gauge theory, $b_0 = (11/3)\check{h}$, where $\check{h}$ is the dual Coxeter number of $\mathfrak{g}$. At high energy ($\mu \to \infty$), the coupling becomes weak ($g \to 0$); this is known as asymptotic freedom. On the contrary, the interaction becomes strong at low energy. It is believed that the theory exhibits confinement and has a mass gap.

QCD, or quantum chromodynamics, is gauge theory coupled to matter fields. Suppose the boson ϕ and the fermion ψ are in the (complex) representations R_b and R_f of G, respectively. That is, $\phi \in \Gamma(P \times_G R_b)$, or ϕ is a section of the bundle $P \times_G R_b$, and $\psi \in \Gamma(S \otimes (P \times_G R_f))$, where S is the spinor bundle over $\mathbb{R}^{3,1}$. The classical action is

$$S_{\mathrm{QCD}}[A,\phi,\psi] = S_{\mathrm{YM}}[A] + \frac{1}{g^2}\int \mathrm{d}^4x \frac{1}{2}|\nabla\phi|^2 + \sqrt{-1}(\psi, \not{\nabla}\psi) + \cdots$$

where ∇ is the covariant derivative, $\not{\nabla}$ is the Dirac operator coupled to A, and we have omitted possible mass and potential terms. The quantum theory depends sensitively on the representations R_b and R_f. In the β-function, we have

$$b_0 = \tfrac{11}{3}\check{h} - \tfrac{1}{6}\nu(R_b) - \tfrac{2}{3}\nu(R_{\mathrm{f}})$$

where $\nu(R)$ is the Dynkin index of a representation R. If $b_0 < 0$, the theory is free in the infrared but strongly interacting in the ultraviolet. If $b_0 > 0$, the converse is true; in particular, the theory exhibits asymptotic freedom. If $b_0 = 0$, the situation depends on the sign of the two or higher-loop contributions.

Pure $N = 1$ supersymmetric gauge theory is one on the superspace $\mathbb{R}^{3,1|(2,2)}$ with a constraint that the curvature vanishes in the odd directions. The dynamical variables are in the superfield strength W, a $1|(1,0)$-form valued in ad P. In components, the theory is gauge field coupled to a Majorana or Weyl fermion in the adjoint representation. Let $S^{\pm}$ be spinor bundles of positive (negative) chiralities, respectively, and let λ be a section of $S^+ \otimes \mathrm{ad}P$. The action, written both in superspace and in ordinary spacetime, is

$$S^{N=1}_{\mathrm{SYM}}[A,\lambda] = \frac{1}{4\pi}\mathrm{Im}\left(\int \mathrm{d}^4x\, \mathrm{d}^2\theta\, \tau\langle W, W\rangle\right) = S_{\mathrm{YM}}[A] + \frac{1}{g^2}\int \mathrm{d}^4x \sqrt{-1}\langle\bar{\lambda}, \not{\nabla}^+\lambda\rangle$$

Since $b_0 = 3\check{h}$, the theory is asymptotically free but strongly coupled at low energy. Classically, the theory has a $\mathrm{U}(1)_R$ chiral symmetry. However, due to anomaly, only the subgroup $\mathbb{Z}_{2\check{h}}$ survives at the quantum level. Instanton effect yields gaugino condensation $\langle\lambda\lambda\rangle \sim \Lambda^3$. The symmetry is thus further broken to $\mathbb{Z}_2$, resulting $\check{h}$ inequivalent vacua.

The $N = 1$ QCD has additional chiral superfields Φ in a representation R, including the bosons $\phi \in \Gamma(P \times_G R)$ and the fermions $\psi \in \Gamma(S^+ \otimes (P \times_G R))$. In the absence of superpotential, the action is

$$S^{N=1}_{\mathrm{SQCD}}[A,\lambda,\phi,\psi] = S^{N=1}_{\mathrm{SYM}}[A,\lambda] + \frac{1}{g^2}\int \mathrm{d}^4x\, \mathrm{d}^2\theta\, d^2\bar{\theta}\ \tfrac{1}{2}|\Phi|^2$$

In components, the second term is

$$\frac{1}{g^2}\int \mathrm{d}^4x \left(\tfrac{1}{2}|\nabla\phi|^2 + \sqrt{-1}(\psi, \not{\nabla}^+\psi) - \tfrac{1}{2}\phi^*|D|^2 + \cdots\right)$$

where $D : R \to \mathfrak{g}^*$ is the moment map of the Hamiltonian G-action on R, and we have omitted other terms containing fermionic fields. The moduli space of classical vacua is the symplectic quotient $D^{-1}(0)/G = R/\!/G$. It is the same as the Kähler quotient $R^{\mathrm{s}}/G^{\mathbb{C}}$, where the stable subset $R^{\mathrm{s}} = \{\phi \in R | G^{\mathbb{C}} \cdot \phi \cap D^{-1}(0) \neq \emptyset\}$ is open and dense in R. Again, the quantum theory depends on the representation R. Since $b_0 = 3\check{h} - (1/2)\nu(R)$, the theory is asymptotically free, infrared free, scale invariant (to one-loop) when $\nu(R) < 6\check{h}, \nu(R) > 6\check{h}, \nu(R) = 6\check{h}$, respectively. The moduli space may be lifted by a superpotential or modified by other quantum effects.

SU(N_c) Theories at Low Energy

We now consider $N = 1$ QCD with $G = \mathrm{SU}(N_c)$; N_c is the number of colors. The matter field consists of N_f copies of quarks $Q^i (1 \le i \le N_f)$ in the fundamental representation of $\mathrm{SU}(N_c)$ and N_f copies of antiquarks $Q'_{i'} (1 \le i' \le N_f)$ in the conjugate representation. Using the isomorphism of $\mathfrak{su}(N_c)$ with its dual, the moment map is

$$D(Q,Q') = \text{traceless part of } \sqrt{-1}(QQ^\dagger - Q'Q'^\dagger)$$

So $(Q,Q') \in D^{-1}(0)$ if and only if $QQ^\dagger - Q'^\dagger = cI_{N_c}$ for some $c \in \mathbb{R}$. If $N_f < N_c$, then $c = 0$ and

$$Q, Q' \sim \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_{N_f} \end{pmatrix}$$

for some $a_k \geq 0$. Generically, these $a_k > 0$ and the gauge group $SU(N_c)$ is broken to $SU(N_c - N_f)$. If $N_f \geq N_c$, then

$$Q \sim \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_{N_c} \end{pmatrix}, \qquad Q' \sim \begin{pmatrix} a'_1 & & \\ & \ddots & \\ & & a'_{N_c} \end{pmatrix}$$

where $a_k, a'_k \geq 0$ satisfy $a_k^2 - a'^2_k = c$ for some $c \in \mathbb{R}$. The gauge group is completely broken. The low-energy superfields are the mesons $M^i_{i'} = Q^i Q'_{i'}$ and, if $N_f \geq N_c$, the baryons

$$B_{i_{N_c+1}\cdots i_{N_f}} = \frac{1}{N_c!}\epsilon_{i_1\cdots i_{N_f}} Q^{i_1}\cdots Q^{i_{N_c}}$$
$$B'^{i'_{N_{c+1}}\cdots i'_{N_f}} = \frac{1}{N_c!}\epsilon^{i'_1\cdots i'_{N_f}} Q'_{i'_1}\cdots Q'_{i'_{N_c}}$$

When $N_f < N_c$, Affleck *et al.* (1984) found a dynamically generated superpotential

$$W_{\text{eff}}(\hat{M}) = (N_c - N_f)\left(\frac{\Lambda^{3N_c - N_f}}{\det M}\right)^{1/(N_c - N_f)}$$

generated by instanton effect when $N_f = N_c - 1$ and by gaugino condensation in the unbroken $SU(N_c - N_f)$ theory when $N_f < N_c - 1$. It is also the unique superpotential (up to a multiplicative constant) that is consistent with the global and supersymmetry. The potential pushes the vacuum to infinity. Therefore, contrary to the classical picture, theories with $N_f < N_c$ do not have a vacuum at the quantum level.

When $N_f \geq 3N_c$, the theory is not strongly interacting at low energy, and perturbation methods are reliable. (When $N_f = 3N_c$, the two-loop contribution to the β-function is negative.) We now look at the range $N_c \leq N_f < 3N_c$. The cases $N_f = N_c, N_c + 1$ and $N_c + 2 \leq N_f < 3N_c$ were studied in Seiberg (1994) and Seiberg (1995), respectively.

When $N_f = N_c$, the classical moduli space is $\det M = BB'$. The quantum theory at low energy consists of the fields M, B, B' satisfying the constraint $\det M - BB' = \Lambda^{2N_c}$. The quantum moduli space is smooth everywhere, and there are no additional massless particles. So the gluons are heavy throughout the moduli space. This is due to confinement near the origin, where the interaction is strong, and due to the Higgs mechanism far out in the flat direction, where the classical picture is a good approximation. We see a smooth transition between these two effects.

When $N_f = N_c + 1$, there is a dynamically generated superpotential

$$W_{\text{eff}} = \frac{1}{\Lambda^{2N_c - 1}}(B'MB - \det M)$$

The stationary points of W_{eff} are at $BB' - \wedge^{N_c} M = 0$, $BM = 0, MB' = 0$; these are precisely the constraints that the classical configuration satisfies. However, the moduli space is interpreted differently: it is embedded into a larger space, and the constraints are satisfied only at stationary points. At the singularity $\langle M \rangle = 0$, the whole global symmetry group is unbroken, and B, B' are the new massless fields resolving the singularity. So we have a continuous transition between confinement (without chiral symmetry breaking) and the Higgs mechanism in the semiclassical regime.

When $N_c + 2 \leq N_f \leq (3/2)N_c$, the original theory, called the electric theory, is still strongly coupled in the infrared. Seiberg (1995) proposed that there is a dual, magnetic theory, which is infrared free. The two theories are different classically, but are equivalent at the quantum level. The dual theory is an $N = 1 SU(\tilde{N}_c)$ gauge theory with $\tilde{N}_c = N_f - N_c$, coupled to dual quarks $\tilde{Q}_i, \tilde{Q}^{i'}$, where $1 \leq i; i' \leq N_f$ are flavor indices. In addition, the mesons $M^i_{i'}$ become fundamental fields. They are not coupled to the $SU(\tilde{N}_c)$ gauge field but interact with the dual quarks through the superpotential

$$W = \mu^{-1} M^i_{i'} \tilde{Q}_i \tilde{Q}^{i'}$$

The two theories have the same global symmetry and the same gauge-invariant operators. The dual quarks are fundamental in the magnetic theory but are solitonic excitations in the electric theory. At high energy, the electric theory is asymptotically free, while the magnetic theory is strongly coupled. At low energies, the converse is true. Therefore, reliable perturbative calculations can be performed by choosing an appropriate weakly coupled theory.

When $(3/2)N_c < N_f < 3N_c$, the theory has a nontrivial infrared fixed point. This is because up to two-loop,

$$\beta(g) = -\frac{g^3}{16\pi^2}(3N_c - N_f) + \frac{g^5}{128\pi^4}\left(2N_c N_f - 3N_c^2 - \frac{N_f}{N_c}\right) + o(g^7)$$

There is a solution $g_* > 0$ to $\beta(g) = 0$. We have $\beta(g) < 0$ when $0 < g < g_*, \beta(g) > 0$ when $g > g_*$. In the infrared limit, the coupling constant flows to $g = g_*$, where we have a nontrivial, interacting superconformal theory in four dimensions. The conformal dimension becomes anomalous and is equal to 3/2 of the charge of the chiral $U(1)_R$; for example, that of the meson $\mu^{-1}M$ is $3(N_f - N_c)/N_f > 1$ in this range.

Other Classical Gauge Groups

We now consider $N=1$ supersymmetric gauge theory and QCD with gauge groups $\mathrm{Sp}(N_c)$ and $\mathrm{SO}(N_c)$. The $\mathrm{Sp}(N_c)$ theories, studied by Intriligator and Pouliot (1995), are the simplest examples of the $N=1$ theories. We take $2N_f$ chiral superfields $Q_i (i=1,\ldots,2N_f)$ in the fundamental representation $\mathbb{C}^{2N_c} \cong \mathbb{H}^{N_c}$ of $\mathrm{Sp}(N_c)$. The number of copies must be even so that the quantum theory is free from global gauge anomaly. The gauge-invariant quantities are the mesons $M_{ij} = Q_i^a Q_j^b \omega_{ab}$, where ω is the symplectic form on $\mathbb{C}^{2N_c}$, subject to a constraint $\epsilon^{1,\ldots,2N_c+2} M_{1,2} \cdots M_{2N_c+1,2N_c+2} = 0$. Using the decomposition $\mathfrak{u}(2N_c) = \mathfrak{sp}(N_c) \oplus \sqrt{-1}\{\mathbb{H}\text{-self-adjoint matrices}\}$, the moment map $D(Q)$ is the projection of $\sqrt{-1}QQ^\dagger$ on $\mathfrak{sp}(N_c)$. So $D(Q)=0$ implies

$$Q \sim \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_{\min\{N_c,N_f\}} \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

where $a_k \geq 0$. At a generic point of the classical moduli space, the gauge group is broken to $\mathrm{Sp}(N_c - N_f)$ if $N_c > N_f$; it is completely broken if $N_c \leq N_f$.

Since $b_0 = 3(N_c+1) - N_f$, the quantum theory is infrared free if $N_f \geq 3(N_c+1)$. (When $b_0 = 0$, the two-loop β-function is negative.) When $N_f \leq N_c$, there is a dynamically generated superpotential

$$W_{\mathrm{eff}} = (N_c + 1 - N_f) \times \left(\frac{2^{N_c-1} \Lambda^{3(N_c+1)-N_f}}{\mathrm{Pf}\, M} \right)^{1/(N_c+1-N_f)}$$

pushing the vacuum to infinity.

When $N_f = N_c$, the classical moduli space $\mathrm{Pf} M = 0$ has singularities. The quantum moduli space is $\mathrm{Pf}\, M = 2^{N_c-1} \Lambda^{2(N_c+1)}$. The singularity is smoothed out and there are no light fields other than the mesons M. When $N_f = N_c + 1$, all components of M become dynamical in the low-energy theory, and there is a superpotential

$$W_{\mathrm{eff}} = -\frac{\mathrm{Pf}\, M}{2^{N_c-1} \Lambda^{2N_c+1}}$$

At the most singular point $\langle M \rangle = 0$, the global symmetry is unbroken, and all the light fields in M become massless. In both cases, there is a transition between confinement and Higgs mechanism.

When $N_c + 3 \leq N_f \leq (3/2)(N_c+1)$, there is a dual, magnetic theory which is free in the infrared. The dual theory has $2N_f$ quarks $\tilde{Q}^i$ in the fundamental representation of $\mathrm{Sp}(\tilde{N}_c)$, where $\tilde{N}_c = N_f - N_c - 2$. In addition, the mesons M_{ij} become elementary and couple to $\tilde{Q}$ through a superpotential $W = (2\mu)^{-1} M_{ij} \tilde{Q}^{ia} \tilde{Q}^{jb} \tilde{\omega}_{ab}$, where $\tilde{\omega}$ is the symplectic form on $\mathbb{C}^{2\tilde{N}_c}$. When $(3/2)(N_c+1) < N_f < 3(N_c+1)$, the theory flows to an interacting superconformal field theory in the infrared.

Theories with the $\mathrm{SO}(N_c)$ gauge group were studied by Seiberg (1995) and by Intriligator and Seiberg (1995). Since the fundamental representation is real, there is no constraint on the number N_f of quarks $Q^i (1 \leq i \leq N_f)$. The gauge invariants are the mesons $M^{ij} = Q_a^i Q_b^j \delta^{ab}$ and, if $N_f \geq N_c$, the baryons $B_{i_{N_c+1} \cdots i_{N_f}} = \epsilon_{i_1 \cdots i_{N_f}} Q^{i_1} \cdots Q^{i_{N_c}} / N_c!$ They satisfy rank $M \leq N_c$ and $BB = \wedge^{N_c} M$. Using the decomposition $\mathfrak{u}(N_c) = \mathfrak{so}(N_c) \oplus \sqrt{-1}\{\mathbb{R}\text{-self-adjoint matrices}\}$, the moment map $D(Q)$ is the projection $\sqrt{-1}QQ^\dagger$ on $\mathfrak{so}(N_c)$. If $D(Q)=0$, then up to gauge and global symmetries, Q is of the form

$$Q \sim \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_r \end{pmatrix}$$

where $a_1, \ldots, a_r > 0$ if $r = \mathrm{rank}\, Q \leq N_c$ and $a_1, \ldots, a_{N_c-1} > 0$ and $a_{N_c} \neq 0$ if $r = N_c$. Generically, the gauge group is broken to $\mathrm{SO}(N_c - N_f)$ if $N_c \geq N_f + 2$ and is totally broken if $N_c < N_f + 2$.

We have $b_0 = 3(N_c - 2) - N_f$ if $N_c \geq 5$. For $N_c = 4$, the group is $(\mathrm{SU}(2) \times \mathrm{SU}(2))/\mathbb{Z}_2$ and $b_0 = 6 - N_f$ for each SU(2) factor. If $N_c = 3$, the group is $\mathrm{SU}(2)/\mathbb{Z}_2 b_0 = 6 - 2N_f$. The theory is asymptotically free if $N_f > 3(N_c - 2)$ and infrared free if $N_f \geq 3(N_c - 2)$.

When $N_f \leq N_c - 5$, there is a dynamically generated superpotential

$$W_{\mathrm{eff}} = \frac{1}{2}(N_c - 2 - N_f) \times \left(\frac{16 \Lambda^{3(N_c-2)-N_f}}{\det M} \right)^{1/(N_c-2-N_f)}$$

lifting the classical vacuum degeneracy. The coefficient is fixed by mass deformation and by matching the SU(4) theory when $N_c = 6$.

When $N_f = N_c - 4$, the unbroken gauge group is $\mathrm{SO}(4) = (\mathrm{SU}(2) \times \mathrm{SU}(2))/\mathbb{Z}_2$ on the generic point of the moduli space. The superpotential of the original theory is

$$W_{\mathrm{eff}} = 2(\epsilon_+ + \epsilon_-) \left(\frac{\Lambda^{2(N_c-1)}}{\det M} \right)^{1/2}$$

where the choices $\epsilon_+, \epsilon_- = \pm 1$ correspond to the fact that each of the SU(2) theory has two vacua. There are two physically inequivalent branches: $\epsilon_+ = \epsilon_-$ and $\epsilon_+ = -\epsilon_-$. For $\epsilon_+ = \epsilon_-$, the superpotential pushes the vacuum to infinity. For $\epsilon_+ = -\epsilon_-, W_{\mathrm{eff}} = 0$. In the quantum theory, the singularity is smoothed out and

all the massless fermions are in M, even at the origin of the moduli space. Hence the quarks are confined.

When $N_f = N_c - 3$, the unbroken gauge group is SO(3) and the theory has two branches with

$$W_{\text{eff}} = 4(1+\epsilon)\frac{\Lambda^{2N_c-3}}{\det M}$$

where $\epsilon = \pm 1$. For $\epsilon = 1$, the quantum theory has no vacuum. For $\epsilon = -1$, $W_{\text{eff}} = 0$, but there are additional light fields $\tilde{Q}_i$ coupling to M via the superpotential $W \sim (2\mu)^{-1} M^{ij}\tilde{Q}_i\tilde{Q}_j$ near $M = 0$.

When $N_f = N_c - 2$, the low-energy theory is related to the $N = 2$ gauge theory and will be addressed in the subsection "Seiberg–Witten's low-energy solution."

When $N_f \geq N_c - 1$, we define a dual, magnetic theory whose gauge group is $\mathrm{SO}(\tilde{N}_c)$, where $\tilde{N}_c = N_f - N_c + 4$. There are N_f dual quarks $\tilde{Q}_i (1 \leq i \leq N_f)$ in the fundamental representation. This theory is infrared free if $N_f \leq (3/2)(N_c - 2)$. In the effective theory, the mesons M^{ij} become fundamental and couple with the dual quarks through a superpotential $W = (2\mu)^{-1} M^{ij}\tilde{Q}_i\tilde{Q}_j$ if $N_f \geq N_c$; there is an additional term $\det M / 64\Lambda^{2N_c-5}$ if $N_f = N_c - 1$. When $(3/2)(N_c - 2) < N_f < 3(N_c - 2)$, the theory flows to an interacting superconformal field theory in the infrared.

$N = 2$ Gauge Theory and Seiberg–Witten Duality

$N = 2$ Yang–Mills Theory

Pure $N = 2$ supersymmetric gauge theory is a special case of $N = 1$ QCD when $R = \mathfrak{g}^{\mathbb{C}}$ is the (complexified) adjoint representation of G. The moment map is $D(\phi) = (1/2\sqrt{-1})[\phi, \bar{\phi}] \in \mathfrak{g} \cong \mathfrak{g}^*(\phi \in \mathfrak{g})$. Since the fermionic fields λ and ψ are sections of the same bundle, there is a second set of supersymmetry transformations by interchanging the roles of λ and ψ. This makes the theory $N = 2$ supersymmetric. The classical action is

$$\begin{aligned} S^{N=2}_{\text{SYM}}[A, \lambda, \psi, \phi] = {} & S_{\text{YM}}[A] \\ & + \frac{1}{g^2}\int \mathrm{d}^4 x \sqrt{-1}(\langle \bar{\lambda}, \not{\nabla}\lambda\rangle \\ & + \langle \bar{\psi}, \not{\nabla}\psi\rangle) + \frac{1}{2}|\nabla\phi|^2 \\ & + \sqrt{-1}(\langle \bar{\phi}, [\lambda, \psi]\rangle + \langle \phi, [\bar{\lambda}, \bar{\psi}]\rangle) \\ & - \frac{1}{8}|[\phi, \bar{\phi}]|^2 \end{aligned}$$

The energy reaches the minimum when ϕ takes a constant value $\phi \in \mathfrak{g}^{\mathbb{C}}$ that can be conjugated by G to the Cartan subalgebra $\mathfrak{t}^{\mathbb{C}}$. ($\mathfrak{t}$ is the Lie algebra of the maximal torus T.) The classical moduli space is $\mathfrak{g}^{\mathbb{C}}/G^{\mathbb{C}} = \mathfrak{t}^{\mathbb{C}}/W$, where W is the Weyl group. At a generic $\phi \in \mathfrak{t}^{\mathbb{C}}$, the gauge group is broken to T by the Higgs mechanism. Classically, the massless degrees of freedom are excitations of ϕ and components of the gauge field in $\mathfrak{t}$. So the low-energy physics can be described by these massless fields. However, the moduli space is singular when ϕ is on the walls of the Weyl chambers. At these values, the unbroken gauge group is larger and there are extra massless fields that resolve the singularities.

Since $b_0 = 2\check{h} > 0$, the quantum theory is asymptotically free but strongly interacting at low energy. It can be shown that $N = 1$ supersymmetry already forbids a dynamically generated superpotential on $\mathfrak{t}^{\mathbb{C}}/W$. Therefore, the vacuum degeneracy is not lifted and the quantum moduli space is still a continuum. However, there are corrections to the part of classical moduli space where strong interactions occur. The quantum theory has a dynamically generated mass scale Λ. We pick the renormalization scale μ to be $|\phi|$, the typical energy scale where spontaneous symmetry breaking occurs. Far away from the origin, that is, when $|\phi| \gg |\Lambda|$, the theory is weakly interacting and the classical description of the moduli space is a good approximation. However, when $|\phi|$ is comparable to $|\Lambda|$, the classical language and perturbation methods fail due to strong interaction. At $\phi = 0$, the full gauge symmetry is restored classically. But since the theory becomes strongly interacting at low energy, it cannot be the low-energy solution of the original theory.

The classical $\mathrm{U}(1)_R$ symmetry extends to $\mathrm{U}(2)_R$, mixing λ and ψ. The $\mathrm{U}(1)_R$ subgroup in $\mathrm{U}(2)_R$ is anomalous except for a subgroup $\mathbb{Z}_{4\check{h}}$. So we have a global $\mathrm{SU}(2)_R \times_{\mathbb{Z}_2} \mathbb{Z}_{4\check{h}}$ symmetry at the quantum level. This is consistent with a continuous moduli space of vacua, if the group $\mathrm{SU}(2)_R$ is to act nontrivially. Also, the space is not a single orbit of the global symmetry group. The generator of $\mathbb{Z}_{4\check{h}}$ acts on $\mathfrak{t}^{\mathbb{C}}$ by a phase $\mathrm{e}^{-\pi\sqrt{-1}/\check{h}}$. The group $\mathbb{Z}_{4\check{h}}$ is spontaneously broken to the subgroup which acts trivially on $\mathfrak{t}^{\mathbb{C}}/W$.

We study the general form of low-energy effective Lagrangian that is consistent with $N = 2$ supersymmetry. We assume that the quantum effect does not modify the topology of the moduli space $\mathfrak{t}^{\mathbb{C}}/W$, though it may alter the singularity and its nature. Suppose U is the quantum moduli. At a generic point in U, the residual gauge group is T. In the $N = 1$ language, the theory is a supersymmetric gauged sigma model with target space U. It contains $N = 1$ vector multiplets W^I and chiral multiplets Φ^I, where $1 \leq I \leq r$, $r = \dim T$ being the rank of G. $N = 1$ supersymmetry requires that U is Kähler, with

possible singularities where the effective theory breaks down. $N=2$ supersymmetry requires further that U is special Kähler, that is, there is a flat, torsion-free connection ∇ on TU such that the Kähler form ω is parallel and such that $d_\nabla J=0$, where the complex structure J is viewed as a 1-form valued in TU. See, for example, Freed (1999). Locally, there is a holomorphic prepotential $\mathcal{F}$ and special coordinates $\{z^I\}$. Let $\tilde{z}_I=\partial\mathcal{F}/\partial z^I$ be the dual coordinates and let $\tau_{IJ}=\partial^2\mathcal{F}/\partial z^I\partial z^J=\partial\tilde{z}_I/\partial z^J$. Then $K=\mathrm{Im}(\tilde{z}_I\bar{z}^I)$ is a Kähler potential and $\omega=(\sqrt{-1}/2)\,\mathrm{Im}(\tau_{IJ})\mathrm{d}z^I\wedge\mathrm{d}\bar{z}^J$ is the Kähler form. The effective action is

$$S_{\mathrm{eff}}^{N=2}[W,\Phi]=\frac{1}{4\pi}\mathrm{Im}\left(\int \mathrm{d}^4x\,\mathrm{d}^2\theta\tfrac{1}{2}\tau_{IJ}(\Phi)(W^I,W^J)+\int \mathrm{d}^4x\,\mathrm{d}^2\theta\mathrm{d}^2\bar{\theta}K(\Phi)\right)$$

Note that both the coupling constants τ_{IJ} and the metric $\mathrm{Im}\tau_{IJ}$ on U are determined by a holomorphic function $\mathcal{F}$, which is the hallmark of $N=2$ supersymmetry.

In the bare theory with abelian gauge group T, the action is given by choosing $\mathcal{F}_0(\Phi)=(1/2)\tau_{IJ}\langle\Phi^I,\Phi^J\rangle$, where the τ_{IJ} (and hence the metric $\mathrm{Im}\tau_{IJ}$) are constants. Due to one-loop and instanton effects, $\mathcal{F}$ is no longer quadratic in the effective theory. Since τ varies on U, it cannot be holomorphic (except at a few singular points), single valued, and having a positive-definite imaginary part. The solution to this apparent contradiction is that each set of special coordinates and the expression of $\mathcal{F}$ is valid only in part of U. Solving the $N=2$ gauge theory at low energy means understanding the singularity of U in the strong coupling regime and obtaining the explicit form of $\mathcal{F}$ or τ_{IJ} in various regions of the moduli space.

Seiberg–Witten's Low-Energy Solution

We consider $N=2$ gauge theory with $G=\mathrm{SU}(2)$. The Cartan subalgebra is $\mathfrak{t}\cong\mathbb{C}$; each $a\in\mathbb{C}$ determines an element $\phi=(1/2)\left(\begin{smallmatrix}a & 0\\ 0 & -a\end{smallmatrix}\right)$ in $\mathfrak{t}$. The Weyl group $W\cong\mathbb{Z}_2$ acts on $\mathbb{C}$ by $a\mapsto\pm a$. The moduli space of classical vacua is the u-plane $\mathbb{C}/\mathbb{Z}_2$ parametrized by $u=\mathrm{tr}\,\phi^2=(1/2)a^2$. When $u\neq 0$, the gauge group is broken to U(1). The generator of $\mathbb{Z}_{4\check{b}}=\mathbb{Z}_8\subset\mathrm{U}(1)_R$ acts as $a\mapsto\sqrt{-1}a, u\mapsto -u$. The $\mathbb{Z}_8$ symmetry is broken to $\mathbb{Z}_4$; the quotient $\mathbb{Z}_2=\mathbb{Z}_8/\mathbb{Z}_4$ acts on the u-plane by $u\mapsto\pm u$.

Abelian gauge theory and $N=4$ supersymmetric gauge theory exhibit exact electric–magnetic duality in the sense that the quantum theories are identical if the coupling constant τ undergoes an $\mathrm{SL}(2,\mathbb{Z})$ transformation. Seiberg and Witten (1994a, b) proposed that this is so for the low-energy effective theory of the $N=2$ gauge theory. An $\mathrm{SL}(2,\mathbb{Z})$ transformation maps one description of the low-energy theory to another, exchanging electricity and magnetism. It is however not an exact duality of the full SU(2) theory. Rather, duality is in the ambiguity of the choice of the low-energy description. More precisely, τ is a section of a flat $\mathrm{SL}(2,\mathbb{Z})$ bundle over U. Thus, τ is multivalued and exists as a function in local charts only. So we must use different Lagrangians in different regions of the u-plane. Around the singularities where τ is not defined, nontrivial monodromy can appear.

Away from infinity, the electric theory is strongly interacting but the magnetic theory is infrared free. The dual field is $\tilde{a}=\mathrm{d}\mathcal{F}(a)/\mathrm{d}a$, and $\tau_{\mathrm{eff}}(u)=\mathrm{d}\tilde{a}/\mathrm{d}a$. The group $\mathrm{SL}(2,\mathbb{Z})$ is generated by

$$P=\begin{pmatrix}-1 & 0\\ 0 & -1\end{pmatrix},\qquad S=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}$$
$$T=\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$$

To see its action on $\left(\begin{smallmatrix}\tilde{a}\\ a\end{smallmatrix}\right)$, we use the central extension of the $N=2$ super-Poincaré algebra. In the classical theory, the central charge is $Z=(n_{\mathrm{e}}+\tau n_{\mathrm{m}})a$ from the boundary terms at infinity. As the electric–magnetic duality transformation S interchanges n_{e} and n_{m}, we have for any $\gamma\in\mathrm{SL}(2,\mathbb{Z})$, $\gamma:(n_{\mathrm{m}},n_{\mathrm{e}})\mapsto(n_{\mathrm{m}},n_{\mathrm{e}})\gamma^{-1}$. When $n_m=0$, the classical formula $Z=n_{\mathrm{e}}a$ is valid. Invariance of Z under $\mathrm{SL}(2,\mathbb{Z})$ requires that $Z=n_{\mathrm{m}}\tilde{a}+n_{\mathrm{e}}a$ at the quantum level and that $\mathrm{SL}(2,\mathbb{Z})$ acts on $\left(\begin{smallmatrix}\tilde{a}\\ a\end{smallmatrix}\right)$ homogeneously as a column vector.

When $u=(1/2)a^2$ is large, perturbation is reliable. The classical and one-loop results are $a(u)\sim\sqrt{2u}, \tilde{a}\sim(\sqrt{-1}/\pi)a\log a^2$. As u goes around infinity, the fields transforms as $a\mapsto -a, \tilde{a}\mapsto-\tilde{a}+2a$. The monodromy is $M_\infty=PT^{-2}$. The mass M of a monopole state is bounded by $M^2=P^\mu P_\mu\geq|Z|^2$, which is precisely the Bogomol'nyi bound. Now as a consequence of the $N=2$ supersymmetry, it receives no quantum corrections as long as supersymmetry is not broken at the quantum level. The states that saturate the bound are the BPS states. The BPS spectrum at $u\in U$ is a subset of $H_1(E_u,\mathbb{Z})\cong\mathbb{Z}^2$ containing the pairs $(n_{\mathrm{m}},n_{\mathrm{e}})$ realized by the dyon charges. Near infinity, the condition is that either $n_{\mathrm{e}}=\pm1, n_{\mathrm{m}}=0$ (for $W_\pm$ particles) or $n_{\mathrm{m}}=\pm1$ (for monopoles or dyons). This spectrum is invariant under the monodromy M_∞.

The nontrivial holonomy at infinity implies the existence of at least one singularity at a finite value $u=u_0$, where extra particles become massless. Seiberg and Witten (1994a, b) propose that these

particles are collective excitations in the perturbative regime. Suppose along a path connecting u_0 and some base point near infinity, a monopole of charges $(\pm 1, n_e) = (0,1)(T^{\mp n_e} S^{\pm 1})^{-1}$ becomes massless at u_0. Then by the renormalization group analysis and duality, the monodromy at u_0 is $M_{u_0} = (T^{\mp n_e} S^{\pm 1}) T^2 (T^{\mp n_e} S^{\pm 1})^{-1}$. It turns out that there are two singularities $u = \pm\Lambda^2$ with monodromies $M_{\Lambda^2} = ST^2S^{-1}$ and $M_{-\Lambda^2} = (TS)T^2(TS)^{-1}$. The particles that become massless at $\pm\Lambda^2$ are of charges $(n_m, n_e) = (1,0)$ and $(1,-1)$, respectively. The only BPS states in the strong coupling regime are those which become massless at the singularities; the others decay as u deforms towards strong interaction.

The monodromies $M_{\pm\Lambda^2}, M_\infty$ (or any two of them) generate the subgroup $\Gamma(2)$. The family of elliptic curves with these monodromies can be identified with $y^2 = (x-\Lambda^2)(x+\Lambda^2)(x-u)$ called the Seiberg–Witten curve. The singularities are at $u = \pm\Lambda^2$ and $u = \infty$, where the curve degenerates. Let

$$\lambda = \frac{\sqrt{2}}{2\pi} \frac{y\,dx}{x^2 - \Lambda^4}$$

be the Seiberg–Witten differential (of second kind on the total space E). Then in a suitable basis (α, β) of $H_1(E_u/U, \mathbb{Z})$, we have $a = \int_\alpha \lambda, \tilde{a} = \int_\beta \lambda$. At a singularity, if $\nu = n_m\beta + n_e\alpha$ is a vanishing cycle, then the dyon of charges (n_m, n_e) becomes massless. This is because its central charge is $Z = n_m\tilde{a} + n_e a = \int_\nu \lambda$. The monodromy at a singularity where ν is a vanishing cycle is given by the Picard–Lefshetz formula $M: \gamma \mapsto \gamma - 2(\gamma \cdot \nu)\nu$. At $u = \pm\Lambda^2$, the vanishing cycles are β and $\beta - \alpha$, respectively.

We return to the $N=1$ $SO(N_c)$ gauge theory with $N_f = N_c - 2$. At a generic point in the moduli space, the gauge group is broken to SO(2), which is abelian. Much of the above discussion applies to this case. By $N=1$ supersymmetry, the effective coupling τ_{eff} is holomorphic in M but is not single valued. In fact, τ_{eff} depends on $u = \det M$, which is invariant under the (anomaly free) $SU(N_f)$ symmetry. For large u, we have $e^{2\pi\sqrt{-1}\tau_{\text{eff}}} = \Lambda^{4N_c-8}/u^2$ and the monodromy around infinity is $M_\infty = PT^{-2}$. On the other hand, a large expectation value of M of rank $N_c - 3$ breaks the gauge group to SO(3) and the theory is the $N=2$ theory discussed earlier. Using these facts, Intriligator and Seiberg (1995) identified the family of elliptic curves as $y^2 = x(x - 16\Lambda^{2N_c-4})(x-u)$. There are two singularities with inequivalent physics. At $u=0$, the monodromy is ST^2S^{-1}. A pair of monopoles $\tilde{Q}^\pm$ becomes massless. They couple with M through the superpotential $W \sim (2\mu)^{-1} M^{ij}\tilde{Q}_i\tilde{Q}_j$. At $u = 16\Lambda^{2N_c-4}$, the monodromy is $(T^2S)T^2(T^2S)^{-1}$. A pair of dyons $E^\pm$ of charges ± 1 become massless. The effective action is $W_{\text{eff}} \sim (u - 16\Lambda^{2N_c-4})E^+E^-$.

Topological gauge theory is a twisted version of $N=2$ Yang–Mills theory in which the observables at high energy are the Donaldson invariants. The work of Seiberg and Witten (1994a, b) yields new insight to it and has a tremendous impact on the geometry of 4-manifolds. See Witten (1994) for the initial steps.

After the work of Seiberg and Witten (1994a, b), there has been much progress on theories with other gauge groups. If the gauge group is a compact Lie group of rank r, the u-plane is replaced by $\mathfrak{t}^{\mathbb{C}}/W$; the singularities are modified by quantum effects. The duality group is $Sp(2r, \mathbb{Z})$ or its subgroup of finite index, acting on the coupling matrix $\tau = (\tau_{IJ})$ by fractional linear transformations. For example, for $G = SU(N_c)$, the moduli space is parametrized by gauge invariants $u_2, \ldots, u_{N_c}$ defined by $\det(xI - \phi) = x^{N_c} - \sum_{i=2}^{N_c} u_i x^{N_c - i} = P_{N_c}(x, u_i)$. Classically, the singular locus is a simple singularity of type A_{N_c-1}. At the quantum level, the singularity consists of two copies of such locus shifted by $\pm\Lambda^n$ in the u_n direction. The monodromies correspond to a family of hyperelliptic curves $y^2 = P_{N_c}(x, u_i)^2 - \Lambda^{2N_c}$ of genus $N_c - 1$. The Seiberg–Witten differential is

$$\lambda = \frac{\sqrt{2}}{\pi\sqrt{-1}} \frac{\partial P_{N_c}(x, u_i)}{\partial x} \frac{x\,dx}{y} + \partial(\cdots)$$

The $N_c - 1$ independent eigenvalues a^i of ϕ and their duals $\tilde{a}_i = \partial\mathcal{F}/\partial a^i$ are the periods of λ along the $2N_c - 2$ homology cycles in the curve. For more details, the reader is referred to Klemm *et al.* (1995) and Argyes and Faraggi (1995).

$N=2$ QCD

$N=2$ supersymmetric QCD is $N=2$ Yang–Mills theory coupled to $N=2$ matter. The latter consists of $N=1$ superfields Q that form a quarternionic representation R of the gauge group G. The space R has a G-invariant hyper-Kähler structure. The hyper-Kähler moment map $\mu_{\mathbb{H}}: R \to \mathfrak{g}^* \otimes \mathrm{Im}\,\mathbb{H}$ consists of a real moment map $\mu_{\mathbb{R}}: R \to \mathfrak{g}^*$ for the Kähler structure and a complex moment map $\mu_{\mathbb{C}}: R \to (\mathfrak{g}^*)^{\mathbb{C}}$ for the holomorphic symplectic structure. As an $N=1$ theory, the matter superfields are valued in $R \times \mathfrak{g}^{\mathbb{C}}$ with a D-term $D(Q,\Phi) = \mu_{\mathbb{R}}(Q) + (1/2\sqrt{-1})[\Phi, \bar{\Phi}]$ and a superpotential $W(Q,\Phi) = \sqrt{2}\langle\mu_{\mathbb{C}}(Q), \Phi\rangle + m(Q)$, where the mass term m is a G-invariant quadratic form on R. The classical moduli space of vacua has two branches. On the Coulomb branch where $Q = 0$ and $\Phi \neq 0$, the unbroken gauge group is abelian and the

photons are massless. If $Q \neq 0$ exists in the flat directions, the gauge group is broken according to the value of Q; these are the Higgs branches. If $m = 0$, the moduli space of classical vacua is the hyper-Kähler quotient $\mu_{\mathbb{H}}^{-1}(0)/G$. The branches of two types touch at the origin, where the full gauge group is restored, and at other subvarieties in R. The global symmetry is the subgroup of $U(R)$ that commutes with the G-action on R and preserves m; it contains $U(2)_R$.

Quantum mechanically, such a theory is free from local gauge anomalies. Consistency under large gauge transformations puts a torsion condition on R, such as $\nu(R) = 0 (\mathrm{mod}\, 2)$. Since $b_0 = 2\check{h} - (1/2)\nu(R)$, the theory is asymptotically free if $\nu(R) < 4\check{h}$. If $\nu(R) = 4\check{h}$, the quantum theory is scale invariant up to one-loop (and hence to all loops), and is expected to be so nonperturbatively. If $\nu(R) > 4\check{h}$, the quantum theory may not be defined but it can be the low-energy solution of another asymptotically free theory. Due to the axial anomaly, the $U(2)_R$ global symmetry reduces to the subgroup $SU(2)_R \times_{\mathbb{Z}_2} \mathbb{Z}_{4\check{h}-\nu(R)}$. The metric on the Coulomb branch can be corrected by quantum effects, but those on the Higgs branches do not change because of the uniqueness of the hyper-Kähler metric. In the quantum theory, the Higgs branches still touch the Coulomb branch, but the photons of the Coulomb branch are the only massless gauge bosons at the point where they meet.

When $G = SU(N_c)$ we take N_f quarks $Q^i (i = 1, \ldots, N_f)$ in the fundamental representation and N_f antiquarks $\tilde{Q}_i (i = 1, \ldots, N_f)$ in the complex-conjugate representation. The moment map is the same as in $N = 1$ QCD whereas the superpotential is $W = \sqrt{2}\tilde{Q}_i \Phi Q^i + \sum_i m_i \tilde{Q}_i Q^i$. Consider the case $G = SU(2)$ as in Seiberg and Witten (1994b). Since $b_0 = 4 - N_f$, the asymptotically free theories have $N_f \leq 3$ whereas the $N_f = 4$ theory is scale invariant. As the representations on Q^i and $\tilde{Q}_i$ are isomorphic, the classical global symmetry is $O(2N_f) \times U(2)_R$ when all $m_i = 0$. The appearance of the even number of fundamental representations is necessary for the consistency of the theory at the quantum level. The $U(1)_R$ symmetry is anomalous if $N_f \neq 4$. When $N_f > 0$, $SO(2N_f)$ is anomaly free, whereas $O(2N_f)/SO(2N_f) = \mathbb{Z}_2$ is anomalous. The anomaly free subgroup of $\mathbb{Z}_2 \times U(1)_R$ is $\mathbb{Z}_{4(4-N_f)}$. Its $\mathbb{Z}_2$ subgroup acts in the same way as $\mathbb{Z}_2 \subset Z(SO(2N_f))$. A nonzero expectation value of $u = \mathrm{tr}\, \phi^2$ further breaks the symmetry to $\mathbb{Z}_4$. The quotient group that acts effectively on the u-plane (the Coulomb branch) is $\mathbb{Z}_{4-N_f}$ if $N_f > 0$ and $\mathbb{Z}_2$ if $N_f = 0$. When $N_f = 4$, the $U(1)_R$ symmetry is anomaly free but $\mathbb{Z}_2 = O(8)/SO(8)$ is still anomalous.

The $N_f = 0$ theory is the $N = 2$ pure gauge theory. In order to compare it to the $N_f > 0$ theories, we multiply n_e by 2 so that it has integer values on Q^i and $\tilde{Q}_i$, and divide a by 2 to preserve the formula $Z = n_m \tilde{a} + n_e a$. The monodromies around the singularities become $M_{\Lambda^2} = STS^{-1}$, $M_{-\Lambda^2} = (T^2 S)T(T^2 S)^{-1}$, $M_\infty = PT^{-4}$. They generate the subgroup $\Gamma_0(4)$ of $SL(2, \mathbb{Z})$. The coupling constant is

$$\tau = \frac{\theta}{\pi} + \frac{8\pi\sqrt{-1}}{g^2}$$

The Seiberg–Witten curve is $y^2 = x^3 - ux^2 + (1/4)\Lambda_0^4 x$, related to the earlier one $y^2 = (x - u)(x^2 - \Lambda_0^4)$ by an isogeny. Here and below, Λ_{N_f} is the dynamically generated scale.

For $N_f > 0$, we consider the case with zero bare masses. The simplest BPS-saturated states are the elementary quarks with mass $\sqrt{2}|a|$, which form the vector representation of $SO(2N_f)$. In addition, the quarks have fermion zero modes in the monopole background. When $n_m = 1$, each $SU(2)$ doublet of quarks has one zero mode. With N_f hypermultiplet, there are $2N_f$ zero modes in the vector representation of $SO(2N_f)$. Upon quantization, the quantum states are in the spinor representation. So the flavor symmetry is really $Spin(2N_f)$. The spectrum may also include states with $n_m > 1$. For $N_f = 2, 3, 4$, the center $Z(Spin(2N_f))$ are $\mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_4$, $\mathbb{Z}_2 \times \mathbb{Z}_2$, whose generators act on states of charges (n_m, n_e) by $((-1)^{n_e+n_m}, (-1)^{n_e})$, $\sqrt{-1}^{n_m+2n_e}$, $((-1)^{n_m}, (-1)^{n_e})$, respectively.

Suppose at a singularity on the u-plane, the low-energy theory is QED with k hypermultiplets. Let m_i be the bare mass and S_i, the $U(1)$ charge of the ith hypermultiplet. With the expectation value of ϕ, the actual masses are $|\sqrt{2}a + m_i| (1 \leq i \leq k)$. As the states form a small representation of the $N = 2$ algebra, the central charge is modified as $Z = n_m \tilde{a} + n_e a + S \cdot m/\sqrt{2}$, where $m = (m_1, \ldots, m_k)$ and $S = (S_1, \ldots, S_k)$. Under a duality transformation $M \in SL(2, \mathbb{Z})$, the column vector $(m/\sqrt{2}, \tilde{a}, a)$ is multiplied by a matrix of the form $\hat{M} = \left(\begin{smallmatrix} I_k & 0 \\ * & M \end{smallmatrix}\right)$. (For example, if $M = T$, $\hat{M}$ can be derived by one-loop analysis.) So the row vector $W = (S, n_m, n_e)$ transforms as $W \mapsto W\hat{M}^{-1}$. The transformation on (n_m, n_e) is not homogeneous when there are hypermultiplets. This phenomenon persists even when all the bare masses m_i are zero.

When $N_f = 1$, the global symmetry of the u-plane is $\mathbb{Z}_3$. There are three singularities related by this symmetry, where monopoles with charges $(n_m, n_e) = (1, 0), (1, 1)$, and $(1, 2)$ become massless. The low-energy theory at each singularity is QED with a single light hypermultiplet. Besides the photon, no other flat directions exist. This is consistent with the absence of Higgs branch in the original theory. The monodromies at the singularities are STS^{-1},

$(TS)T(TS)^{-1}, (T^2S)T(T^2S)^{-1}$, respectively, and the corresponding Seiberg–Witten family of curves is $y^2 = x^2(x-u) - (1/64)\Lambda_1^6$. The Seiberg–Witten differential is

$$\lambda = -\frac{\sqrt{2}}{4\pi}\frac{y\,dx}{x^2}$$

When $N_f = 2$, there are two singularities related by the global symmetry $\mathbb{Z}_2$ of the u-plane. The massless states at one singularity have $(n_m, n_e) = (1, 0)$ and form a spinor representation of SO(4) while those at the other have $(n_m, n_e) = (1, 1)$ and form the other spinor representation. The low-energy theory at each singularity is QED with two light hypermultiplets. There are additional flat directions along which $SO(4) \times SU(2)_R$ is broken. They form the two Higgs branches that touch the u-plane at the two singularities rather than at the origin. The metric and pattern of symmetry breaking are the same as classically. The monodromies are $ST^2S^{-1}, (TS)T^2(TS)^{-1}$. The Seiberg–Witten curve is $y^2 = (x^2 - u) - (1/64)\Lambda_2^4)(x-u)$ and the differential is

$$\lambda = -\frac{\sqrt{2}}{4\pi}\frac{y\,dx}{x^2 - \Lambda_2^4/64}$$

When $N_f = 3$, the u-plane has no global symmetry. There are two singularities. At one of them, a single monopole bound state with $(n_m, n_e) = (2, 1)$ becomes massless and there are no other light particles. At the other singularity, the massless states have $(n_m, n_e) = (1, 0)$ and form a (four-dimensional) spinor representation of SO(6) with a definite chirality. Thus, the low-energy theory is QED with four light hypermultiplets. Along the flat directions, the $SO(6) \times SU(2)_R$ symmetry is further broken. This corresponds to a single Higgs branch touching the u-plane at the singularity. Again, the metric on the Higgs branch is not modified by quantum effects. The monodromies at the two singularities are $(ST^2S)T(ST^2S)^{-1}$ and ST^4S^{-1}, respectively. The Seiberg–Witten curve is $y^2 = x^2(x-u) - (1/64)\Lambda_3^2(x-u)^2$ and the differential is

$$\lambda = \frac{\sqrt{2}}{\pi\Lambda_3}\log\left(y + \sqrt{-1}\frac{\Lambda_3}{8}\left(x - u - \frac{32}{\Lambda_3^2}x^2\right)\right)dx$$

When $N_f = 4$, the theory is characterized by classical coupling constant τ, and there are no corrections to $a = (1/2)\sqrt{2}u, \tilde{a} = \tau a$. There is only one singularity at $u = 0$, where the monodromy is P. Seiberg and Witten (1994b) postulate that the full quantum theory is $SL(2, \mathbb{Z})$ invariant, just like the $N = 4$ pure gauge theory. The elementary hypermultiplet has $(n_m, n_e) = (0, 1)$ and form the vector representation v of SO(8). Fermion zero modes give rise to hypermultiplets with $(n_m, n_e) = (1, 0), (1, 1)$ that transform under the spinor representations s, c of Spin(8). $SL(2, \mathbb{Z})$ acts on the spectrum via a homomorphism onto the outer-automorphism group S_3 of Spin(8), which then permutes v, s, and c. So duality is mixed in an interesting way with the SO(8) triality. In v, s, and c, the center $\mathbb{Z}_2 \times \mathbb{Z}_2$ acts as $((-1)^{n_m}, (-1)^{n_e}) = (1, -1), (-1, 1), (-1, -1)$, respectively. The full $SL(2, \mathbb{Z})$ invariance predicts the existence of multimonopole bound states: for every pair of relatively prime integers (p, q), there are eight states with $(n_m, n_e) = (p, q)$ that form a representation of Spin(8) on which the center acts as $((-1)^p, (-1)^q)$.

Solutions when the bare masses are nonzero are also obtained by Seiberg and Witten (1994b). The masses can be deformed to relate theories with different values of N_f. $N = 2$ QCD with a general classical gauge group has also been studied. By adding to these theories a mass term $m \operatorname{tr} \Phi^2$ that explicitly breaks the supersymmetry to $N = 1$, the dualities of Seiberg can be recovered. For $SU(N_c), SO(N_c)$ and $Sp(2N_c)$ gauge groups, see Hanany and Oz (1995), Argyes *et al.* (1996), Argyes *et al.* (1997) and references therein.

See also: Anomalies; Brane Construction of Gauge Theories; Donaldson–Witten Theory; Duality in Topological Quantum Field Theory; Effective Field Theories; Electric–Magnetic Duality; Floer Homology; Gauge Theories from Strings; Gauge Theory: Mathematical Applications; Nonperturbative and Topological Aspects of Gauge Theory; Quantum Chromodynamics; Topological Quantum Field Theory: Overview; Supersymmetric Particle Models.

Further Reading

Affleck I, Dine M, and Seiberg N (1984) Dynamical supersymmetric breaking in supersymmetric QCD. *Nuclear Physics B* 241: 493–534.

Argyes PC and Faraggi AE (1995) Vacuum structure and spectrum of $N = 2$ supersymmetric SU(n) gauge theory. *Physical Review Letters* 74: 3931–3934.

Argyes PC, Plesser MR, and Seiberg N (1996) The moduli space of vacua of $N = 2$ SUSY QCD and duality in $N = 1$ SUSY QCD. *Nuclear Physics B* 471: 159–194.

Argyes PC, Plesser MR, and Shapere AD (1997) $N = 2$ moduli spaces and $N = 1$ dualities for SO(n_c) and USp($2n_c$) super-QCD. *Nuclear Physics B* 483: 172–186.

Freed DS (1999) Special Kähler manifolds. *Communications in Mathematical Physics* 203: 31–52.

Hanany A and Oz Y (1995) On the quantum moduli space of vacua of $N = 2$ supersymmetric SU(N_c) gauge theories. *Nuclear Physics B* 452: 283–312.

Klemm A, Lerche W, Yankielowicz S, and Theisen S (1995) Simple singularities and $N = 2$ supersymmetric Yang–Mills theory. *Physics Letters B* 344: 169–175.

Intriligator K and Pouliot P (1995) Exact superpotential, quantum vacua and duality in supersymmetric SP(N_c) gauge theories. *Physics Letters B* 353: 471–476.

Intriligator K and Seiberg N (1995) Duality, monopoles, dyons, confinement and oblique confinement in supersymmetric SO(N_c) gauge theories. *Nuclear Physics B* 444: 125–160.

Seiberg N (1994) Exact results on the space of vacua of four-dimensional SUSY gauge theories. *Physical Review D* 49: 6857–6863.

Seiberg N (1995) Electric–magnetic duality in supersymmetric non-Abelian gauge theories. *Nuclear Physics B* 435: 129–146.

Seiberg N and Witten E (1994a) Electric–magnetic duality, monopole condensation, and confinement in $N=2$ supersymmetric Yang–Mills theory. *Nuclear Physics B* 426: 19–52.

Seiberg N and Witten E (1994b) Monopoles, duality and chiral symmetry breaking in $N=2$ supersymmetric QCD. *Nuclear Physics B* 431: 484–550.

Witten E (1994) Monopoles and four-manifolds. *Mathematical Research Letters* 1: 769–796.

Semiclassical Approximation *see* Stationary Phase Approximation; Normal Forms and Semiclassical Approximation

Semiclassical Spectra and Closed Orbits

Y Colin de Verdière, Université de Grenoble 1, Saint-Martin d'Hères, France

Introduction

The purpose of this article is to describe the so-called "semiclassical trace formula" (SCTF) relating the "spectrum" of a semiclassical Hamiltonian to the "periods of closed orbits" of its classical limit. SCTF formula expresses the asymptotic behavior as $\hbar \to 0$ ($\hbar = h/2\pi$) of the regularized density of states as a sum of oscillatory contributions associated to the closed orbits of the classical limit.

We will mainly present the case of the Schrödinger operator on a Riemannian manifold which contains the purely Riemannian case.

We start with a section about the history of the subject. We then give a statement of the results and a heuristic proof using Feynman integrals. This proof can be transformed into a mathematical proof which we will not give here. After that we describe some applications of the SCTF.

About the History

SCTF has several origins: on one side, Selberg trace formula (1956) is an exact summation formula concerning the case of locally symmetric spaces; this formula was interpreted by H Huber as a formula relating eigenvalues of the Laplace operator and lengths of closed geodesics (also called the "lengths spectrum") on a closed surface of curvature –1.

On the other side, around 1970, two groups of physicists developed independently asymptotic trace formulas:

- M Gutzwiller for the Schrödinger operator, using the quasiclassical approximation of the Green function (the "van Vleck's formula"); it is interesting to note that the word "trace formula" is not written, but Gutzwiller instead speaks of a new "quantization method" (the old one being "Einstein–Brillouin–Keller (EBK)" or "Bohr–Sommerfeld rules").
- R Balian and C Bloch, for the eigenfrequencies of a cavity, use what they call a "multiple reflection expansion." They asked about a possible application to Kac's problem.

At the same time, under the influence of Mark Kac's famous paper "Can one hear the shape of a drum?," mathematicians became quite interested in inverse spectral problems, mainly using heat kernel expansions (for the state of the art around 1970, see Berger *et al.* (1971)).

The SCTF was put into its final mathematical form for the Laplace operator on closed manifolds by three groups of people around 1973–75:

- Y Colin de Verdière in his thesis was using the short-time expansion of the Schrödinger kernel and an approximate Feynman path integral. He proved that the spectrum of the Laplace operator determines generically the lengths of closed geodesics.
- J Chazarain derived the qualitative form of the trace for the wave kernel using Fourier integral operators.

- Using the full power of the symbolic calculus of Fourier integral operators, H Duistermaat and V Guillemin were able to compute the main term of the singularity from the Poincaré map of the closed orbit. Their paper became a canonical reference on the subject.

After that, people were able to extend SCTF to:

- general semiclassical Hamiltonians (Helffer–Robert, Guillemin–Uribe, Meinrenken),
- manifolds with boundary (Guillemin–Melrose),
- surfaces with conical singularities and polygonal billiards (Hillairet), and
- several commuting operators (Charbonnel–Popov).

Recently, some researchers have remarked about the nonprincipal terms in the singularities expansion which come from the semiclassical Birkhoff normal form (Zelditch, Guillemin).

Selberg Trace Formula

We consider a compact hyperbolic surface X. "Hyperbolic" means that the Riemannian metric is locally $(dx^2 + dy^2)/y^2$ or is of constant curvature –1. Such a surface is the quotient $X = \boldsymbol{H}/\Gamma$ where Γ is a discrete co-compact subgroup of the group of isometries of the Poincaré half-plane $\boldsymbol{H}$. Closed geodesics of X are in bijective correspondence with nontrivial conjugacy classes of Γ. More precisely, the set of loops $C(S^1, X)$ splits into connected components associated to conjugacy classes and each component of nontrivial loops contains exactly one periodic geodesic.

Theorem 1 (Selberg trace formula). *If ρ is a real-valued function on $\mathbb{R}$ whose Fourier transform $\hat{\rho}$ is compactly supported and $\lambda_j = 1/4 + \mu_j^2$ is the spectrum of the Laplace operator on X, we have*:

$$\sum_{j=1}^{\infty} \rho(\mu \pm \mu_j) = \frac{A}{2\pi} \int_{\mathbb{R}} \rho(\mu + s) s \tanh \pi s \, ds + \sum_{\gamma \in \mathcal{P}} \sum_{n=1}^{\infty} \frac{l_\gamma}{2\pi \sinh(n l_\gamma/2)} \times \mathrm{Re}(\hat{\rho}(n l_\gamma) e^{in\mu l_\gamma})$$

where A is the area of X, $\mathcal{P}$ the set of primitive conjugacy classes of Γ and, for $\gamma \in \mathcal{P}$, l_γ is the length of the unique closed geodesic associated to γ.

A nice recent presentation of the Selberg trace formula can be found in Marklof (2003).

Semiclassical Schrödinger Operators on Riemannian Manifolds

If (X, g) is a (possibly noncompact) Riemannian manifold and $V : X \to \mathbb{R}$ a smooth function which satisfies $\liminf_{x \to \infty} V(x) = E_\infty > -\infty$, the differential operator $\hat{H} = (1/2)\hbar^2 \Delta + V$ is semibounded from below and admits self-adjoint extensions. For all those extensions, the spectrum is discrete in the interval $]-\infty, E_\infty[$ and eigenfunctions $\hat{H}\varphi_j = E_j \varphi_j$ are localized in the domain $V \leq E_j$. If X is compact and $V = 0$, we recover the case of the Laplace operator.

We will denote this part of the spectrum by

$$\inf V < E_1(\hbar) < E_2(\hbar) \leq \cdots \leq E_j(\hbar) \leq \cdots < E_\infty$$

For the Laplace operator, we have $E_j = \hbar^2 \lambda_j$, where $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_j \leq \cdots$ is the spectrum of the Laplace operator.

The SCTF can also be derived the same way for Schrödinger operators with magnetic field. One can even extend it to Hamiltonian systems which are not obtained by Legendre transform from a regular Lagrangian. In this case, Morse indices have to be replaced by the more general Maslov indices.

Classical Dynamics

Newton Flows

Euler–Lagrange equations for the Lagrangian $\mathcal{L}(x, v) := (1/2)\|v\|_g^2 - V(x)$ admit a Hamiltonian formulation on T^*X whose energy is given by $H = (1/2)\|\xi\|_g^2 + V(x)$. We will denote by X_H the Hamiltonian vector field

$$X_H := \sum_j \frac{\partial H}{\partial \xi_j} \partial_{x_j} - \frac{\partial H}{\partial x_j} \partial_{\xi_j}$$

Preservation of H by the dynamics shows immediately that the Hamiltonian flow Φ_t restricted to $H < E_\infty$ is complete.

The Hamiltonian H is the "classical limit" of $\hat{H}$; in more technical terms, H is the semiclassical principal symbol of $\hat{H}$.

If $V = 0$, $H = (1/2) g^{ij} \xi_i \xi_j$ and the flow is the geodesic flow.

Periodic Orbits

Definition 1 *A* periodic orbit (γ, T) (also denoted p.o.) of the Hamiltonian H consists of an orbit γ of X_H which is homeomorphic to a circle and a nonzero real number T so that $\Phi_T(z) = z$ for all $z \in \gamma$. We will denote by $T_0(\gamma) > 0$ (the primitive period) the smallest $T > 0$ for which $\Phi_T(z) = z$.

If (T, E) are given, $W_{T,E}$ is the set of z's so that $H(z) = E$ and $\Phi_T(z) = z$.

- The (linear) Poincaré map Π_γ of a p.o. (γ, T) with $H(\gamma) = E$: we restrict the flow to $S_E := \{H = E\}$ and take a hypersurface Σ inside S_E transversal to γ at the point z_0. The associated return map P is a local diffeomorphism fixing z_0. Its linearization $\Pi_\gamma := P'(z_0)$ is the linear Poincaré map, an inversible (symplectic) endomorphism of the tangent space $T_{z_0}\Sigma$.
- The Morse index $\iota(\gamma)$: p.o. (γ, T) is a critical point of the action integral $\int_0^T \mathcal{L}(\gamma(s), \dot\gamma(s))\, ds$ on the manifold $C^\infty(\mathbb{R}/T\mathbb{Z}, X)$. It always has a finite Morse index (Milnor 1967) which is denoted by $\iota(\gamma)$. For general Hamiltonian systems, the Morse index is replaced by the Conley–Zehnder index.
- The nullity index $\nu(\gamma)$ is the dimension of the space of infinitesimal deformations of the p.o. γ by p.o. of the same energy and period. We always have $\nu(\gamma) \geq 1$ and $\nu(\gamma) = 1 + \dim \ker(\mathrm{Id} - \Pi_\gamma)$.

Example 1 (Geodesic flows)

- Riemannian manifold with sectional curvature < 0: in this case, we have for all periodic geodesics $\iota(\gamma) = 0, \nu(\gamma) = 1$.
- Generic metrics: for a generic metric on a closed manifold, we have $\nu(\gamma) = 1$ for all periodic geodesics.
- For flat tori of dimension d: we have $\iota(\gamma) = 0$ and $\nu(\gamma) = d$.
- For sphere of dimension 2 with constant curvature: if γ_n is the nth iterate of the great circle, we have $\iota(\gamma_n) = 2|n|$ and $\nu(\gamma_n) = 3$.

It is a beautiful result of J-P Serre that any pair of points on a closed Riemannian manifold are endpoints of infinitely many distinct geodesics. Counting geometrically distinct periodic geodesics is much harder especially for simple manifolds like the spheres. It is now known that every closed Riemannian manifold admits infinitely many geometrically distinct periodic geodesics (at least, in some cases, for generic metrics, (Berger 2000 chap. V). There exists significant knowledge concerning more general Hamiltonian systems as well.

Nondegeneracy

There are several possible nondegeneracy assumptions. They can be formulated "à la Morse–Bott" (critical point of action integrals) or purely symplectically.

Definition 2 Two submanifolds Y and Z of X intersect cleanly iff $Y \cap Z$ is a manifold whose tangent space is the intersection of the tangent spaces of Y and Z.

Fixed points of a smooth map are clean if the graph of the map intersects the diagonal cleanly.

Definition 3 We will denote by (ND) the following property of the p.o. (γ_0, T_0): the fixed points of the associated (nonlinear) Poincaré map P are clean.

The set $W_{T,E}$ is ND if all p.o.'s inside are ND. $W_{T,E}$ is then a manifold of dimension $\nu(\gamma)$.

Example 2

- *Generic case*: $\nu = 1$; (ND) is equivalent to "1 is not an eigenvalue of the linear Poincaré map." In this case, we can deform the p.o. smoothly by moving the energy. This family of p.o.'s is called a cylinder of p.o.'s. The period $T(E)$ is then a smooth function of E.
- *Completely integrable systems*: $\nu = d$; (ND) is then a consequence of the so-called "isoenergetic KAM condition": assuming the Hamiltonian is expressed as $H(I_1, \ldots, I_d)$ using action-angle coordinates, this condition is that the mapping $I \to [\nabla H(I)]$ from the energy surface $H = E$ into the projective space is a local diffeomorphism. This condition implies that Diophantine invariant tori are not destructed by a small perturbation of the Hamiltonian.
- *Maximally degenerated systems*: it is the case where all orbits are periodic ($\nu = 2d - 1$). For example, the two-body problem with Newtonian potential and the geodesic flows on compact rank-1 symmetric spaces.

Canonical Measures and Symplectic Reduction

Under the hypothesis (ND), the manifold $W_{T,E}$ admits a canonical measure μ_c, invariant by Φ_t. In the case $\nu = 1$, this measure is given by $|dt|/\sqrt{\det(\mathrm{Id} - \Pi)}$.

By using a Poincaré section, it is enough to understand the following fact: if A is a symplectic linear map, the space $\ker(\mathrm{Id} - A)$ admits a canonical Lebesgue measure.

We start with the following construction: let L_1 and L_2 be two Lagrangian subspaces of a symplectic space E and $\omega_j, j = 1, 2$, be half-densities on L_j, denoted by $\omega_j \in \Omega^{1/2}(L_j)$. If $W = L_1 \cap L_2$, we have the following canonical isomorphisms: $\Omega^{1/2}(L_j) = \Omega^{1/2}(W) \otimes \Omega^{1/2}(L_j/W)$. So $\Omega^{1/2}(L_1) \otimes \Omega^{1/2}(L_2) = \Omega^{1/2}(L_1/W) \otimes \Omega^{1/2}(L_2/W) \otimes \Omega^1(W)$. $M_j = L_j/W$ are two Lagrangian subspaces of the reduced space W^o/W whose intersection is 0. Hence, by using the Liouville measure on it, we get $\Omega^{1/2}(M_1) \otimes \Omega^{1/2}(M_2) = \mathbb{C}$. Hence, we get a density $\omega_1 \star \omega_2$ on W. It turns out that the previous calculation is one of the main algebraic pieces of the symbolic calculus of

Fourier integral operators and the density $\omega_1 \star \omega_2$ arises in stationary-phase computations.

The graph of a symplectic map is equipped with a half-density by pullback of the Liouville half-density. So we can apply the previous construction to the intersection of the graph of A and the graph of the identity map.

Actions

Definition 4 If (γ, T) is a p.o., we define the following quantity which is called action of γ:

$$A(\gamma) = \int_\gamma \xi \, \mathrm{d}x$$

In the (ND) case, $A(\gamma)$ is constant on each connected component of $W_{T,E}$.

In the generic case and if $T'(E) \neq 0$ (cylinder of p.o.), p.o.'s of the cylinder are also parametrized by T (i.e., we note by γ_E the p.o. of the cylinder of energy E and γ_T the p.o. of period T). If $a(E) = A(\gamma_E)$ and $b(T) = -\int_0^T \mathcal{L}(\gamma_T(s), \dot{\gamma}_T(s))\mathrm{d}s$, $a(E)$ and $b(T)$ are Legendre transforms of each other.

Playing with Spectral Densities

We will define the "regularized spectral densities." The general idea is as follows: we want to study an $\hbar$-dependent sequence of numbers $E_j(\hbar)$ (a spectrum) in some interval $[a, b]$. We introduce a non negative function $\rho \in \mathcal{S}(\mathbb{R})$ which satisfies $\int \rho(t)\mathrm{d}t = 1$, and also $D_{\rho,\varepsilon,\hbar}(E) = \sum \rho_\varepsilon(E - E_j)$, where $\rho_\varepsilon(E) = \varepsilon^{-1}\rho(E/\varepsilon)$. It gives the analysis of the spectrum at the scale ε. Of course, we will adapt the scaling ε to the small parameter $\hbar$. If the scaling is of the size of the mean spacing of the spectrum, we will get a very precise resolution of the spectrum.

The general philosophy is:

- If $\hbar$ is the semiclassical parameter of a semiclassical Hamiltonian, the mean spacing of the eigenvalues is of order $\hbar^d$ (Weyl's law). The trace formula gives the asymptotic behavior of $D_{\rho,\varepsilon,h}(E)$ for $\varepsilon \sim \hbar$ (and hence $\varepsilon >> \overline{\Delta}E$ except if $d = 1$). This behavior is not "universal" and thus contains a significant amount information of (in our case, on periodic trajectories).
- Better resolution of the spectrum needs the use of the long-time behavior of the classical dynamics and is conjecturally universal. It means that eigenvalues seen at very small scale behave like eigenvalues of an ensemble of random matrices, the most common one being the Wigner Gaussian orthogonal ensemble (GOE) and Gaussian unitary ensemble (GUE).

We fix some interval $[a, b]$ with $b < E_\infty$.

We define $D(E) := \sum_{a \leq E_j \leq b} \delta(E_j)$ as the sum of Dirac measures at the points E_j and its h-Fourier transform as

$$Z(t) = \mathrm{trace}'(\mathrm{e}^{-\mathrm{i}t\hat{H}/\hbar}) := \sum{}' \exp(-\mathrm{i}tE_j/\hbar) \quad [1]$$

where $\sum'$ is the sum over $E_j \in [a, b]$.

The Duistermaat–Guillemin trick relates the previous behavior to asymptotics of the regularized density of eigenvalues. Let us give a function $\rho \in \mathcal{S}(\mathbb{R})$ so that $\hat{\rho}(t) = \int \mathrm{e}^{-\mathrm{i}tE}\rho(E)\mathrm{d}E$ is compactly supported and

$$\hat{\rho}(t) = 1 + O(t^\infty), \quad t \to 0 \quad [2]$$

(all moments of ρ vanish). We introduce, for $E \in [a, b]$, $D_\rho(E) := \sum'_j \frac{1}{\hbar}\rho(E - E_j/\hbar)$. $D_\rho(E)$ is independent modulo $O(h^\infty)$ of a, b. We have

$$D_\rho(E) = \frac{1}{2\pi\hbar}\int \hat{\rho}(t)Z(t)\,\mathrm{d}t$$

The idea is now to start from a semiclassical approximation of $U(t) = \mathrm{e}^{-\mathrm{i}tH/\hbar}$ and to insert it into eqn [1]. We need only a uniform approximation of $U(t)$ for $t \in \mathrm{Support}(\hat{\rho})$. From the asymptotic expansion of $Z(t)$, we will deduce the asymptotic expansion of D_ρ, the regularized eigenvalue density.

The Smoothed Density of States

The following statement expressing the smoothed density of eigenvalues is the main result of the subject. Under the (ND) assumption, it gives the existence of an asymptotic expansion for $D_\rho(E)$:

Theorem 2 *If E is not a critical value of H and the* (ND) *condition is satisfied for all p.o.'s of energy $E \in [a, b]$ and period inside the support of $\hat{\rho}$,*

$$D_\rho(E) = D_{\mathrm{Weyl}}(E) + \sum D_{W(T,E)} + O(h^\infty) \quad [3]$$

where:

(i)
$$D_{\mathrm{Weyl}}(E) = (2\pi\hbar)^{-d}\left(\sum_{j=0}^{\infty} a_j(E)\hbar^j\right)$$
with $a_0(E) = \int_{H=E} \mathrm{d}L/\mathrm{d}H$

(ii) *The sum is over all the manifolds $W_{T,E}$ so that* $T \in \mathrm{Support}(\hat{\rho})$.

(iii)
$$D_{W(T,E)} = \frac{\varepsilon}{(2\pi\mathrm{i}\hbar)^{(\nu(\gamma)+1)/2}}\,\mathrm{e}^{-\mathrm{i}\iota(\gamma)\pi/2} \times \mathrm{e}^{\mathrm{i}A(\gamma)/\hbar}\sum_{j\geq 0} b_j(E)\hbar^j$$

with

$$b_0(E) = \hat{\rho}(T) \int_{W_{T,E}} \mathrm{d}\mu_c$$

$$\varepsilon = \begin{cases} 1 & \text{if } T'(E) > 0 \\ i & \text{if } T'(E) < 0 \end{cases}$$

If $\nu(\gamma) = 1$, *we get* $b_0 = \hat{\rho}(T_\gamma) T_0 |\det(\mathrm{Id} - \Pi_\gamma)|^{-1/2}$.

The Weyl Expansion

If Support($\hat{\rho}$) is contained in $[-T_{\min}, T_{\min}]$, where $T_{\min}$ is the smallest period of a p.o. γ with $H(\gamma) = E$, and, if E is not a critical value of H, formula [3] reduces to

$$D_\rho(E) \sim (2\pi\hbar)^{-d} \left(\sum_{j=0}^{\infty} a_j(E) \hbar^j \right)$$

From the previous formula, it is possible to deduce the following estimates:

Theorem 3 *If* a, b *are not critical values of* H:

$$\#\{j | a \le E_j(\hbar) \le b\} = (2\pi\hbar)^{-d} \mathrm{volume}(a \le H \le b)(1 + O(\hbar))$$

This remainder estimate is optimal and was first shown in rather great generality by Hörmander (1968).

Derivation from the Feynman Integral

The Feynman Integral

R Feynman (Feynman and Hibbs 1965) found a geometric representation of the propagator, that is, the kernel $p(t, x, y)$ of the unitary group $\exp(-\mathrm{i}t\hat{H}/\hbar)$ using an integral (FPI := Feynman path integral) on the manifold $\Omega_{t,x,y} := \{\gamma : [0, t] \to X | \gamma(0) = x, \gamma(t) = y\}$ of paths from x to y in the time t; if $\mathcal{L}(\gamma, \dot{\gamma})$ is the Lagrangian, we have, for $t > 0$:

$$p(t, x, y) = \int_{\Omega_{t,x,y}} \exp\left(\frac{\mathrm{i}}{\hbar} \int_0^t \mathcal{L}(\gamma(s), \dot{\gamma}(s))\, \mathrm{d}s \right) |\mathrm{d}\gamma|$$

where $|\mathrm{d}\gamma|$ is a "Riemannian measure" on the manifold $\Omega_{t,x,y}$ with the natural Riemannian structure.

There is no justification FPI as a useful mathematical tool. Nevertheless, FPI gives good heuristics and right formulas.

The Trace and Loop Manifolds

Let us try a formal calculation of the partition function and its semiclassical limit. We get

$$Z(t) = \int_X |\mathrm{d}x| \int_{\Omega_{x,x,t}} \exp\left(\frac{\mathrm{i}}{\hbar} \int_0^t \mathcal{L}(\gamma(s), \dot{\gamma}(s)) \mathrm{d}s \right) |\mathrm{d}\gamma|$$

If we denote by Ω_t the manifold of paths $\gamma : \mathbb{R}/t\mathbb{Z} \to X$, (loops) and we apply Fubini (sic !), we get

$$Z(t) = \int_{\Omega_t} \exp\left(\frac{\mathrm{i}}{\hbar} \int_0^t \mathcal{L}(\gamma(s), \dot{\gamma}(s))\, \mathrm{d}s \right) |\mathrm{d}\gamma|$$

The Semiclassical Limit

We want to apply stationary phase in order to get the asymptotic expansion of $Z(t)$; critical points of $J_t : \Omega_t \to \mathbb{R}$ are the p.o.'s of the Euler–Lagrange flow and hence of the Hamiltonian flow of period t. We require the ND assumption (Morse–Bott), the Morse index, and the determinant of the Hessian:

1. The ND assumption is the original Morse–Bott one in Morse theory: we have smooth manifolds of critical points and the Hessian is transversally ND.
2. The Morse index is the Morse index of the action functional on periodic loops: $L(\gamma) := \int_0^t \mathcal{L}(\gamma(s), \dot{\gamma}(s))\mathrm{d}s$.
3. The Hessian is associated to a periodic Sturm–Liouville operator for which many regularizations have already been proposed.

In this manner, we get a sum of contributions given by the components $W_{j,t}$ of W_t:

$$Z_j(t) = (\mathrm{i}h)^{-\nu_j/2} \mathrm{e}^{(\mathrm{i}/\hbar)L(\gamma)} c_j(\hbar)$$

with $c_j(\hbar) \sim \sum_{l=0}^{\infty} c_{j,l} \hbar^l$ and

$$c_{j,0} = \frac{\mathrm{e}^{-\mathrm{i}\mu(\pi/2)}}{|\delta|^{1/2}}$$

where μ is the Morse index and δ is a regularized determinant.

The Integrable Case

As observed by Berry–Tabor, the trace formula in this case comes from Poisson summation formula using action-angle coordinates. Asymptotic of the eigenvalues to any order can then be given in the so-called quantum integrable case by Bohr–Sommerfeld rules.

The Maximally Degenerated Case

Let us assume that (X, g) is a compact Riemannian manifold for which all geodesics have the same smallest period $T_0 = 2\pi$. Then we have the following clustering property:

Theorem 4 *There exists some constant C and some integer α so that*

(i) *the spectrum of Δ is contained in the union of the intervals*

$$I_k = \left[\left(k + \frac{\alpha}{4}\right)^2 - C, \left(k + \frac{\alpha}{4}\right)^2 + C\right], \quad k = 1, 2, \ldots$$

(ii) $N(k) = \#\text{Spectrum}(\Delta) \cap I_k$ *is a polynomial function of k for k large enough.*

The property (ii) is consequence of the trace formula.

Applications to the Inverse Spectral Problem

We will now restrict ourselves to the case of the Laplace operator on a compact Riemannian manifold (X, g). The main result is as follows:

Theorem 5 (*Colin de Verdière*). *If X is given, there exists a generic subset $\mathcal{G}_X$, in the sense of Baire category, of the set of smooth Riemannian metrics on X, so that, if $g \in \mathcal{G}_X$, the length spectrum of (X, g) can be recovered from the Laplace spectrum. The set $\mathcal{G}_X$ contains all metrics with < 0 sectional curvature and (conjecturally) all metrics with < 0 sectional curvature.*

We can take for $\mathcal{G}_X$ the set of metrics for which all periodic geodesics are nondegenerate and the length spectrum is simple.

Some cancelations may occur between the asymptotic expansions of two ND periodic trajectories with the same actions if the Morse indices differ by 2 mod 4.

The Case with Boundary

If (X, g) is a smooth compact manifold with boundary, one introduces the broken geodesic flow by extending the trajectories by reflection on the boundary. SCTFs have been extended to that case by Guillemin and Melrose. Periodic geodesics which are transversal to the boundary contribute to the density of states in the same way as for periodic manifolds. Periodic geodesics inside the boundary are in general accumulation of periodic geodesics near the boundary: their contributions is therefore very complicated analytically.

Bifurcations

Let us denote by $C_H \subset \mathbb{R}^2_{T,E}$, the set of pairs (T,E) for which $W_{T,E}$ is not empty. The previous results apply to the "smooth" part of the set C_H. Among other interesting points are points $(0,E)$ with critical value E of H (Brummelhuis–Paul–Uribe) and points corresponding to bifurcation of p.o. when moving the energy.

Detailed studies of some of these points have been done, for example, the results of suitable applications of the theory of singularities of functions of finitely many variables, their deformations (catastrophe theory), and applications to stationary-phase method, and a significant body of knowledge on these subjects now exists.

SCTF and Eigenvalue Statistics

One of the main open mathematical problems is: "can one really use appropriate forms of the SCTF as quantization rules and use it in order to derive eigenvalues statistics?"

This problem is related to the fine-scale study of the eigenvalue spacings ($\varepsilon << \hbar$). It is one of the important unsolved problems of the so-called "quantum chaos." Many people think that progress in this field will allow us to solve the Bohigas–Giannoni–Schmit conjecture: "if the geodesic flow is hyperbolic, eigenvalue distribution follows random matrix asymptotics."

See also: Billiards in Bounded Convex Domains; *h*-Pseudodifferential Operators and Applications; Quantum Ergodicity and Mixing of Eigenfunctions; Random Matrix Theory in Physics; Regularization for Dynamical Zeta Functions; Resonances.

Further Reading

Balian R and Bloch C (1970) Distribution of eigenfrequencies for the wave equation in a finite domain I. *Annals of Physics* 60: 401.

Balian R and Bloch C (1971) Distribution of eigenfrequencies for the wave equation in a finite domain II. *Annals of Physics* 64: 271.

Balian R and Bloch C (1972) Distribution of eigenfrequencies for the wave equation in a finite domain III. *Annals of Physics* 69: 76.

Berger M (2000) *Riemannian Geometry During the Second Half of the Twentieth Century*. University Lecture Series, vol. 17. Providence, RI: American Mathematical Society.

Berger M, Gauduchon P, and Mazet E (1971) *Le spectre d'une variété riemannienne compacte*. Berlin–Heidelberg–New York: Springer LNM.

Colin de Verdière Y (1973) Spectre du Laplacien et longueurs des géodésiques périodiques II. *Compositio Mathematica* 27: 159–184.

Duistermaat J (1976) On the Morse index in variational calculus. *Advances in Mathematics* 21: 173–195.

Duistermaat J and Guillemin V (1975) The spectrum of positive elliptic operators and periodic geodesics. *Inventiones Mathematicae* 29: 39–79.

Feynman R and Hibbs A (1965) *Quantum Mechanics and Path Integrals*. New-York: McGraw-Hill.
Gutzwiller M (1971) Periodic orbits and classical quantization conditions. *Journal of Mathematical Physics* 12: 343–358.
Gutzwiller M (1990) *Chaos in Classical and Quantum Mechanics*. Berlin–Heidelberg–New York: Springer.
Hörmander L (1968) The spectral function of an elliptic operator. *Acta Mathematica* 121: 193–218.
Marklof J (2003) Selberg's trace formula: an introduction. Lectures given at the International School "Quantum Chaos on Hypebolic Manifolds," Schloss Reisensburg, Gunsburg, Germany, 4–11 October 2003. To appear in Springer LNP, Berlin–Heidelberg, New York. http://fr.arxiv.org
Milnor J (1967) *Morse Theory*. Annals of Mathematics Studies no. 51. Princeton, NJ: Princeton University Press.
Selberg A (1956) Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series. *Journal of the Indian Mathematical Society* 20: 47–87.

Semilinear Wave Equations

P D'Ancona, Università di Roma "La Sapienza," Rome, Italy

Introduction

A semilinear wave equation is an equation of the form

$$\Box u = F(u, u'), \quad u : \Omega \subseteq \mathbb{R} \times \mathbb{R}^n \to \mathbb{R} \qquad [1]$$

where $F : \mathbb{R}^{n+2} \to \mathbb{R}$ is a smooth function, the d'Alembert operator $\Box$ is defined as

$$\Box = D_t^2 - D_{x_1}^2 \cdots - D_{x_n}^2, \quad D_t = \frac{\partial}{\partial t}, D_{x_j} = \frac{\partial}{\partial x_j} \qquad [2]$$

and u' denotes the vector of all first-order derivatives of u:

$$u' = (D_t u, D_{x_1} u, \ldots, D_{x_n} u) \equiv (u_t, u_{x_1}, \ldots, u_{x_n})$$

Sometimes the term "semilinear" is used in a more restrictive sense and refers to the special class of equations

$$\Box u = f(u) \qquad [3]$$

The very particular case $f(u) = -mu, m > 0$, corresponds to the Klein–Gordon equation, used to model relativistic particles. True nonlinear terms of the form $f(u) = -mu - u^3, m \geq 0$ (meson equation), or $f(u) = -\sin u$ (sine-Gordon equation) have been proposed as models of self-interacting fields with a local interaction. Notice that for the physical applications it is natural to consider complex-valued functions $u(t, x)$; in the general case of eqn [1], this actually means that we are considering a 2×2 system in $\Re u$ and $\Im u$. However, the natural physical requirement of gauge invariance restricts the possible nonlinearities to the functions satisfying the condition

$$f(e^{i\theta} u) = f(u) e^{i\theta}, \quad \forall \theta \in \mathbb{R} \qquad [4]$$

Thus, in particular $f(0) = 0$ and we see that f must be of the form $f(u) = g(|u|^2)u$ for some g. Since the gauge-invariant wave equation

$$\Box u = g(|u|^2)u \qquad [5]$$

has essentially the same properties as the real-valued equation [3], it is not too restrictive to study only real-valued functions as we shall mostly do in the following.

The more general equations of the form [1], involving the derivatives of u, are encountered in several physical theories, including the nonlinear σ-models and general relativity.

However, beyond the concrete physical applications, eqn [1] is important since it is a simplified but relevant model of much more general equations and systems of mathematical physics; despite its simple structure, the semilinear wave equation presents already all the main difficulties and phenomena of nonlinear wave interaction, and it represents an ideal laboratory for such problems.

In this article we plan to give a concise but, as far as possible, comprehensive review of the main research directions concerning eqn [1], and in particular we shall focus on the global existence of both large and small nonlinear waves, and the problem of local existence for low-regularity solutions. A large part of the theory extends to nonlinear perturbations of the form $\Box u = F(u, u', u'')$ and to the fully nonlinear case; we have no space here to give an account of these developments and we must refer the reader to the books and papers cited in the "Further reading" section.

Classical Results

Equations [1] and [3] are hyperbolic with respect to the variable t. This is a precise way of stating that the "correct" problem for it is an initial-value problem (IVP) with data at some fixed time, or

more generally on some spacelike surface: this means that we assign two functions $u_0(x), u_1(x)$, called the "initial data," and we look for a function $u(t,x)$ satisfying the IVP:

$$\Box u = F(u,u'), \quad u(0,x) = u_0(x), \quad u_t(0,x) = u_1(x) \qquad [6]$$

This setting is in agreement with the physical picture of an evolution problem: the data represent the complete state of a system at a fixed time, and they uniquely determine the evolution of the system, which is described by the differential equation.

This rough statement of the problem is sufficient when working with smooth functions, as in the classical approach. By purely classical methods, that is, energy inequalities and nonlinear estimates, it is not difficult to prove the following local existence result, where $H^k = H^k(\mathbb{R}^n)$ denotes the Sobolev space of functions with k derivatives in $L^2(\mathbb{R}^n)$:

Theorem 1 *Assume F is C^∞. Let $(u_0, u_1) \in H^k \times H^{k-1}$ for some $k > 1 + n/2$. Then there exists a time $T = T(\|u_0\|_{H^k} + \|u_1\|_{H^{k-1}}) > 0$ such that problem [6] has a unique solution belonging to $(u, u_t) \in C([-T,T]; H^k) \times C([-T,T]; H^{k-1})$.*

If $F = F(u)$ depends only on u, the result holds for all $k > n/2$.

Proof We decided to include a sketchy but complete proof of this result since it shows the basic approach to nonlinear wave equations: many results of the theory, even some of the most delicate ones, are obtained by suitable variations of the contraction method, and are similar in spirit to this classical theorem.

Assume for a moment that the equation is linear so that $F = F(t,x)$ is a given smooth function of (t,x). For the linear equation $\Box u = F$, we can construct a solution u using explicit formulas. Moreover, u satisfies the energy inequality

$$E_k(t) \le E_k(0) + \int_0^t \|F(s,\cdot)\|_{H^{k-1}}\, ds \qquad [7]$$

where the energy $E_k(t)$ is defined as

$$E_k(t) = \|u(t,\cdot)\|_{H^k} + \|u_t(t,\cdot)\|_{H^{k-1}} \qquad [8]$$

Now we introduce the space $X_T = C([-T,T]; H^k) \cap C^1([-T,T]; H^{k-1})$, the space $Y_T = C([-T,T]; H^{k-1})$, the mapping $\Phi : F \to u$ that takes the function $F(t,x)$ into the solution of $\Box u = F$ (with fixed data u_0, u_1), and the mapping $\Psi(u) = F(u,u')$ which is the original right-hand side of the equation.

The energy inequality tells us that Φ is bounded from Y_T to X_T. Actually, for M large enough with respect to $E_k(0)$ (the H^k norm of the data), Φ takes any ball $B_Y(0,N)$ of Y_T into the ball $B_X(0, M + NT)$ of X_T. Moreover, if we apply [7] to the difference of two equations $\Box u = F$ and $\Box v = G$, we also see that Φ is Lipschitz continuous from Y_T to X_T, with a Lipschitz constant CT.

On the other hand, $\Psi(u) = F(u,u')$ takes X_T to Y_T, provided $k > 1 + n/2$; we can even say that it is Lipschitz continuous from $B_X(0,M)$ to $B_Y(0, C(M))$ for some function $C(M)$, with a Lipschitz constant $C_1(M)$ also depending on M. This follows easily from Moser type estimates like

$$\|F(u,u')\|_{H^{k-1}} \le \phi(\|u\|_{L^\infty})\|u\|_{H^k}, \quad k > \frac{n}{2} + 1$$

or

$$\|F(u)\|_{H^k} \le \phi(\|u\|_{L^\infty})\|u\|_{H^k}, \quad k > \frac{n}{2}$$

Now it is easy to conclude: the composition $\Phi \circ \Psi$ maps X_T into itself, and actually is a contraction of $B_X(0,M)$ into itself provided M is large enough with respect to the data, and T is small enough with respect to M. The unique fixed point is the required solution. □

The wave operator has an additional important property called the finite speed of propagation, which can be stated as follows: given the IVP

$$\Box u = 0, \qquad u(0,x) = u_0(x), \qquad u_t(0,x) = u_1(x)$$

if we modify the data "outside" a ball $B(x_0,R) \subset \mathbb{R}^n$, the values of the solution inside the cone

$$K(x_0,R) = \{(t,x) : t \ge 0,\ |x - x_0| < R - t\}$$

do not change. Notice that $K(x_0,R)$ is the cone with basis $B(x_0,R)$ and tip (R,x_0); the slope of its mantle represents the speed of propagation of the signals, which for the wave operator $\Box$ is equal to 1. The property extends without modification to the semilinear problem [6], at least for the smooth solutions given by Theorem 1. Actually, it is not difficult to modify the proof of the theorem to work on cones instead of bands $[-T,T] \times \mathbb{R}^n$; in other words, given a ball $B = B(x_0,R)$, we can assign two data $u_0 \in H^k(B)$, $u_1 \in H^{k-1}(B)(k > n/2 + 1)$ and prove the existence of a local solution on the cone $K(x_0,R)$ for some time interval $t \in [0,T]$.

In general, the finite speed of propagation allows us to localize in space most of the results and the estimates; as a rule of thumb, we expect that what is true on a band $[0,T] \times \mathbb{R}^n$ should also be true on any truncated cone $K(x_0,R) \cap \{0 \le t \le T\}$.

Symmetries

The linear wave equation can be written as the Euler–Lagrange equation of a suitable Lagrangian. This is still true for the semilinear perturbations of the form

$$\Box u + f(u) = 0 \qquad [9]$$

Indeed, denoting with $F(s) = \int_0^s f(\sigma)\,\mathrm{d}\sigma$ the primitive of f, the Lagrangian of [9] is

$$\mathcal{L}(u) = \iint \left[-\frac{1}{2}|u_t|^2 + \frac{1}{2}|\nabla_x u|^2 + F(u) \right] \mathrm{d}t\,\mathrm{d}x \qquad [10]$$

The functional $\mathcal{L}$ is not positive definite; hence, the variational approach gives only weak results. However, this point of view allows us to apply Noether's principle: any invariance of the functional is related to a conservation law of the equation. These conserved quantities can also be obtained by taking the product of the equation by a suitable multiplier, although this method is far from obvious in many cases. We describe here this circle of ideas briefly.

The functional $\mathcal{L}$ is invariant under the Poincaré group, generated by time and space translations and the Lorentz transformations ($\lambda > 1, c \neq 0$):

$$t \mapsto \frac{\lambda t - x_j/c}{\sqrt{\lambda^2 - 1}}, \qquad x_j \mapsto \frac{\lambda x_j - ct}{\sqrt{\lambda^2 - 1}} \qquad [11]$$

The infinitesimal generators of the translations are simply the partial derivatives D_t and D_{x_j}. The Lorentz transformations can be decomposed as a rotation followed by a boost, and indeed a corresponding complete set of infinitesimal generators are the operators

$$\Gamma_{jk} = x_j D_k - x_k D_j, \qquad \Gamma_j = x_j D_t + t D_j \qquad [12]$$

All the operators in the Poincaré group commute with $\Box$ exactly.

The conservation law related to time translations (time derivative) is the fundamental "conservation of energy"

$$E(t) = \int \left[\frac{1}{2}u_t^2 + \frac{1}{2}|\nabla_x u|^2 + F(u) \right] \mathrm{d}x = E(0) \qquad [13]$$

while spatial translations (spatial derivatives) lead to the conservation of momenta

$$\int u_t u_{x_j}\,\mathrm{d}x = \text{const.}, \quad j = 1, \ldots, n$$

On the other hand, infinitesimal rotations and boost [12] are connected to the conservation of angular momenta

$$\int \left[x_k D_j u - x_j D_k u \right] \cdot D_t u\,\mathrm{d}x = \text{const.}, \quad j, k = 1, \ldots, n \qquad [14]$$

and

$$\int \left[x_k e(u) + D_k u\, D_t u \right] \mathrm{d}x = \text{const.}, \quad k = 1, \ldots, n \qquad [15]$$

where

$$e(u) = \tfrac{1}{2}u_t^2 + \tfrac{1}{2}|\nabla_x u|^2 + F(u) \qquad [16]$$

is the energy density.

The Poincaré group does not exhaust the invariance properties of the free wave equation. Among the other transformations which commute or almost commute with $\Box$, we mention the spacetime dilations and inversions (which together with translations and Lorentz transformations generate the larger conformal group), the scaling $u \mapsto \lambda u$, the spatial dilations, and, in the complex-valued case, the gauge transformation $u \mapsto \mathrm{e}^{\mathrm{i}\theta} u$. In this way several useful conservation laws can be obtained, including the conformal energy identities of K Morawetz.

Strichartz Estimates

Energy estimates are very useful tools but they have some major shortcomings. The main one is clearly the large number of derivatives necessary to estimate the nonlinear term. This is why the modern theory of semilinear wave equations relies mainly on different tools, which go under the umbrella name of Strichartz estimates and express the decay properties of solutions when measured in L^p or related norms. In this section we summarize these estimates in their most general form, and try to give a feeling of the techniques involved.

Consider the following IVP for a homogeneous linear wave equation:

$$\Box u = 0, \qquad u(0, x) = 0, \qquad u_t(0, x) = f(x) \qquad [17]$$

The conservation of energy states that

$$\|u_t(t, \cdot)\|_{L^2}^2 + \|\nabla_x u(t, \cdot)\|_{L^2}^2 \equiv \|f\|_{L^2}^2 \qquad [18]$$

for all times t. Thus, we see that L^2-type norms of the solution do not decay. The interesting fact is that if we measure the solution u in a different L^p-norm, $p > 2$, the norm decays as $t \to \infty$, and the decay is fastest for the L^∞-norm.

To appreciate the dispersive phenomena at their best, let us assume that the Fourier transform of the data is localized in an annulus of order 1:

$$\operatorname{supp} \hat{f}(\xi) \subset \{1/2 \leq |\xi| \leq 2\} \qquad [19]$$

Then the corresponding solution $u(t,x)$ has the same property, and we see that

$$\|u\|_{L^2} = \|\hat{u}\|_{L^2} \leq 2\||\xi|\hat{u}\|_{L^2} \equiv 2\|\nabla u\|_{L^2} \leq 4\|u\|_{L^2}$$

We condense the last line in the shorthand notation

$$\|u\|_{L^2} \simeq \|\nabla u\|_{L^2}$$

We shall also write

$$\|v\|_X \lesssim \|w\|_Y \iff \|v\|_X \leq C\|w\|_Y \quad \text{for some } C$$

We can now rewrite the conservation of energy [20] in a very simple form; for localized data (and hence a localized solution) as in [19], we have

$$\|u(t,\cdot)\|_{L^2} \lesssim \|f\|_{L^2} \qquad [20]$$

The basic L^∞-estimate for a solution of [17] with localized data as in [19] is simply

$$\|u(t,\cdot)\|_{L^\infty} \lesssim t^{-(n-1)/2}\|f\|_{L^1} \qquad [21]$$

This estimate is well known since the 1960s; it can be proved easily by several techniques, notably by the stationary-phase method. Property [21] measures the fact that as time increases, the total energy of the solution remains constant but spreads over a region of increasing volume, due to the propagation of waves. If we interpolate between [20] and [21], we obtain the full set of dispersive estimates

$$\|u(t,\cdot)\|_{L^q} \lesssim t^{-(n-1)(1/2-1/q)}\|f\|_{L^p} \qquad \frac{1}{q}+\frac{1}{p}=1, \quad 2 \leq q \leq \infty \qquad [22]$$

Recall that we are working with localized solutions on the annulus $|\xi| \sim 1$; it is easy to extend the above estimates to general solutions by a rescaling argument, exploiting the fact that, if $u(t,x)$ is a solution of the homogeneous wave equation, $u(\lambda t, \lambda x)$ is also a solution for any constant λ. Indeed, if $\hat{f}$ (and hence $\hat{u}$) is supported in the annulus $2^{j-1} \leq |\xi| \leq 2^{j+1}$, $j \in \mathbb{Z}$, by rescaling [21], we obtain

$$\|u(t,\cdot)\|_{L^\infty} \lesssim t^{-(n-1)/2}2^{j(n-1)/2}\|f\|_{L^1} \qquad [23]$$

If f is any smooth function, not localized in frequency, we can still write it as a series

$$f = \sum_{j\in\mathbb{Z}} f_j$$

where supp $\hat{f}_j \subset \{2^{j-1} \leq |\xi| \leq 2^{j+1}\}$. The quantity

$$\|f\|_{\dot{B}^s_{1,1}} = \sum_{j\in\mathbb{Z}} 2^{js}\|f_j\|_{L^1}$$

is by definition the $\dot{B}^s_{1,1}$ Besov norm of f. Thus, summing the estimates [23] over j, we conclude that a general solution of [17] satisfies the dispersive estimate

$$\|u(t,\cdot)|_{L^\infty} \lesssim t^{-(n-1)/2}\|f\|_{\dot{B}^{(n-1)/2}_{1,1}} \qquad [24]$$

The Strichartz estimates can be obtained as a consequence of the above dispersive estimates, plus some subtle functional analytic arguments. In the general form we give here, they were proved by J Ginibre and G Velo, and in the most difficult endpoint cases by Keel and T Tao. The solution of the homogeneous problem [17] studied above can be written as

$$u(t,x) = \frac{\sin(t|D|)}{|D|}f, \quad |D| \equiv \mathcal{F}^{-1}|\xi|\mathcal{F}$$

(here $\mathcal{F}$ denotes the Fourier transform). On the other hand, the solution of the complete nonhomogeneous problem

$$\Box u = F(t,x), \quad u(0,x) = u_0, \quad u_t(0,x) = u_1(x) \qquad [25]$$

can be written by Duhamel's formula as

$$u(t,x) = \frac{\partial}{\partial t}\frac{\sin(t|D|)}{|D|}u_0 + \frac{\sin(t|D|)}{|D|}u_1 + \int_0^t \frac{\sin((t-s)|D|)}{|D|}f\,\mathrm{d}s$$

and we see that the above estimates [22] apply to all the operators appearing here. If we consider problem [25] and we assume that the data $F(t,x), u_0, u_1$ are localized in frequency so that $\hat{F}(t,\xi), \hat{u}_0, \hat{u}_1$ have support in the annulus $|\xi| \sim 1$, the Strichartz estimate takes the following form:

$$\|u\|_{L^p_I L^q} \lesssim \|u_0\|_{L^2} + \|u_1\|_{L^2} + \|F\|_{L^{\tilde{p}'}_I L^{\tilde{q}'}} \qquad [26]$$

Here the dimension is $n \geq 2$; $L^p_I L^q$ denotes the space with norm

$$\|u\|_{L^p_I L^q} = \left(\int_I \|u(t,\cdot)\|^p_{L^q(\mathbb{R}^n)}\,\mathrm{d}t\right)^{1/p}, \quad I = [0,T] \text{ or } I = \mathbb{R}$$

the indices p, q satisfy the conditions

$$\frac{1}{p}+\frac{1}{q}\frac{n-1}{2} \leq \frac{1}{2}\frac{n-1}{2}, \quad p \in [2,\infty],\ (n,p,q) \neq (3,2,\infty) \qquad [27]$$

while $\tilde{p}, \tilde{q}$ satisfy an identical condition (and p' denotes the conjugate index to p). The constant in inequality [26] is uniform with respect to the interval I.

To get the most general form of the estimates, some additional function space trickery is required. As before, a simple rescaling argument extends estimate [26] to the case of data F, u_0, u_1, whose spatial Fourier transforms are localized in the annulus $2^{j-1} \le |\xi| \le 2^{j+1}$; we obtain

$$\begin{aligned} 2^{j(1/p+n/q)}\|u\|_{L_I^p L^q} \lesssim\ & 2^{jn/2}\|u_0\|_{L^2} \\ & + 2^{j(n/2-1)}\|u_1\|_{L^2} \\ & + 2^{j(1/\tilde{p}'+n/\tilde{q}'-2)}\|F\|_{L_I^{\tilde{p}'}L^{\tilde{q}'}} \end{aligned}$$

Finally, if the data are arbitrary, we may decompose them as series of localized functions, and summing the corresponding estimates we obtain the general Strichartz estimates for the wave equation [25]: for all (p, q) and $(\tilde{p}, \tilde{q})$ as in [27],

$$\begin{aligned} \|u\|_{L_I^p \dot{B}^{1/p+n/q}_{q,2}} \lesssim\ & \|u_0\|_{\dot{H}^{n/2}} + \|u_1\|_{\dot{H}^{n/2-1}} \\ & + \|F\|_{L_I^{\tilde{p}'}\dot{B}^{1/\tilde{p}'+n/\tilde{q}'-2}_{\tilde{q}',2}} \end{aligned} \qquad [28]$$

Here, given a decomposition $f = \sum_{j\in\mathbb{Z}} f_j$, the homogeneous Besov and Sobolev norms are defined, respectively, by the identities (obvious modification for $r = \infty$):

$$\begin{aligned} \|f\|^r_{\dot{B}^s_{q,r}} &= \sum_{j\in\mathbb{Z}} 2^{jsr}\|f_j\|^r_{L^q}, \\ \|u\|_{\dot{H}^s} &= \||\xi|^s \hat{u}\|_{L^2} \simeq \|u\|_{\dot{B}^s_{2,2}} \end{aligned}$$

It is easy to convert the estimates [28] into a form that uses only the more traditional norms

$$\|f\|_{\dot{H}^s_q} \equiv \||D|^s f\|_{L^q}, \qquad |D|^\sigma \equiv \mathcal{F}^{-1}|\xi|^\sigma \mathcal{F}$$

since by the Besov–Sobolev embedding we have

$$\begin{aligned} \dot{B}^s_{q,2} &\subseteq \dot{H}^s_q \quad \text{for } 2 \le q < \infty, \\ \dot{B}^s_{q,2} &\supseteq \dot{H}^s_q \quad \text{for } 1 < q \le 2 \end{aligned}$$

Notice that if we apply to the equation and the data the operator $|D|^\sigma = \mathcal{F}^{-1}|\xi|^\sigma\mathcal{F}$, which commutes with $\Box$, the Strichartz estimate [28] can be rewritten in an apparently more general form:

$$\begin{aligned} \|u\|_{L_I^p \dot{B}^{1/p+n/q+\sigma}_{q,2}} \lesssim\ & \|u_0\|_{\dot{H}^{n/2+\sigma}} \\ & + \|u_1\|_{\dot{H}^{n/2-1+\sigma}} + \|F\|_{L_I^{\tilde{p}'}\dot{B}^{1/\tilde{p}'+n/\tilde{q}'-2+\sigma}_{\tilde{q}',2}} \end{aligned} \qquad [29]$$

In particular, it is possible to choose the indices in such a way that no derivatives appear on u and F: this choice gives

$$\begin{aligned} &\|u\|_{L^p(\mathbb{R}^{n+1})} \lesssim \|u_0\|_{\dot{H}^{1/2}} + \|u_1\|_{\dot{H}^{-1/2}} + \|F\|_{L^{p'}(\mathbb{R}^{n+1})} \\ &p = \frac{2(n+1)}{n-1} \end{aligned}$$

which is the estimate originally proved by Strichartz.

Global Large Waves

As for ordinary differential equations (ODEs), the local solutions constructed in Theorem 1 can be extended to a maximal time interval $[0, T^*]$, and a natural question arises: are these maximal solutions global, that is, is $T^* = \infty$?

For generic nonlinearities and large data, the answer is negative; in a dramatic way, in general the norm $\|u(t, \cdot)\|_{L^\infty}$ is unbounded as $t \uparrow T^* < \infty$. The reason for this is simple: using the finite speed of propagation, we can localize the equation and work on a cone; then if we take constant functions as initial data, the solution inside the cone does not depend on x, and the equation restricted to the cone effectively reduces to an ODE:

$$\begin{aligned} &\Box u = f(u) \iff y''(t) = f(y), \\ &y(t) \equiv u(t, x) \end{aligned} \qquad [30]$$

By this remark it is elementary to construct solutions of the IVP [6] that blow up in a finite time.

This construction does not apply if the equation has some positive conserved quantity. Indeed, consider a general gauge-invariant equation

$$\begin{aligned} &\Box u + g(|u|^2)u = 0, \\ &u(0, x) = u_0(x), \quad u_t(0, x) = u_1(x) \end{aligned} \qquad [31]$$

for some smooth function $g(s)$. Writing $G(s) = \int_0^s g(\sigma)\,d\sigma$, multiplying the equation by $\overline{u}_t$, and integrating over $\mathbb{R}^n$, it is easy to check that the nonlinear energy

$$E(t) = \int \left[|u_t|^2 + |\nabla_x u|^2 + G(|u|^2)\right] dx \equiv E(0) \qquad [32]$$

is constant in time, provided the solution u is smooth enough. When $G(s)$ has no definite sign, we can proceed as above and construct solutions that blow up in finite time; this is usually called the "focusing" case. However, if we assume that $G(s) \ge 0$ ("defocusing" case), the energy $E(t)$ is non-negative. The corresponding ODE, which is $y'' + g(y^2)y = 0$, has only global solutions, and one may guess that also the solutions of [31] can be extended to global ones.

This innocent-looking guess turns out to be one of the most difficult problems of the theory of nonlinear waves, and is actually largely unsolved at present.

The only general result for eqns [31] is Segal's theorem, stating that the IVP has always a global weak solution:

Theorem 2 *Let $g(s)$ be a C^1 non-negative function on $[0, +\infty)$, write $G(s) = \int_0^s g(\sigma)\,d\sigma$ and assume that for some constant C*

$$sg(s^2) \le CG(s^2), \qquad \lim_{s\to+\infty} G(s) = +\infty \qquad [33]$$

Then for any $(u_0, u_1) \in H^1 \times L^2$ *such that* $G(|u_0|^2) \in L^1$, *the IVP* [31] *has a global solution* $u(t, x)$ *in the sense of distributions, such that* $u' \in L^\infty(\mathbb{R}, L^2(\mathbb{R}^n))$ *and* $F(u) \in L^\infty(\mathbb{R}, L^1(\mathbb{R}^n))$.

The proof (see Shatah and Struwe (1998)) is delicate but elementary in spirit: by truncating the nonlinear term, we can approximate the problem at hand with a sequence of problems with global solution; then the conservation law [32] yields some extra compactness, which allows us to extract a subsequence converging to a solution of the original equation.

Thus we see that, despite its generality, this result does not shed much light on the difficulties of the problem. Indeed, the weak solution obtained might not be unique, nor smooth, and in these questions the real obstruction to solving [31] is hidden.

Notice that in the one-dimensional case $n = 1$ the solution is always unique and smooth when the data are smooth, since in this case $E(t)$ controls the L^∞-norm of u. For higher dimensions $n \geq 2$, something more can be proved if we assume that the nonlinear term has a polynomial growth:

$$sg(s^2) = |s|^{p-1}s \quad \text{for } s \text{ large}, \ p > 1 \qquad [34]$$

In particular, the defocusing wave equation with a power nonlinearity

$$\Box u + |u|^{p-1}u = 0 \qquad [35]$$

has been studied extensively. Notice that when p is close to 1, the term $|u|^{p-1}u$ becomes singular near 0; this introduces additional difficulties in the problem; for this reason, it is better to consider a smooth term as in [34].

We can summarize the best-known results concerning [31] under [34] as follows. Let $p_0(n)$ be the number

$$p_0(1) = p_0(2) = \infty$$
$$p_0(n) = 1 + \frac{4}{n-2} \quad \text{for } n \geq 3$$

Then

- in the subcritical case $1 \leq p < p_0(n)$, for any data $(u_0, u_1) \in H^1 \times L^2$, there exists a unique solution $u \in C(\mathbb{R}; H^1)$ such that $u' \in C(\mathbb{R}; L^2)$;
- the same result holds in the critical case $p = p_0(n)$ for $n \geq 3$; and
- when $3 \leq n \leq 7, 1 \leq p \leq p_0(n)$, the solution is smoother if the data are smoother.

These results have been achieved in the course of more than 30 years through the works of several authors (it is indispensable to mention at least the names of K Jörgens, I Segal, W Strauss, W von Wahl, P Brenner, H Pecher, J Ginibre, G Velo, R Glassey and the more recent contributions of J Shatah, M Struwe, L Kapitanski, M Grillakis, omitting many others). Actually modern proofs are remarkably simple, and are based again on a variation of the fixed-point argument. Roughly speaking, the linear equation $\Box u + g(|v|^2)v = 0$ defines a mapping $v \mapsto u$; the Strichartz estimates localized on a cone imply that this mapping is Lipschitz continuous in suitable spaces, the Lipschitz constant being estimated by the nonlinear energy of the solution restricted to the cone. In order to show that this mapping is actually a contraction, it is sufficient to prove that the localized energy tends to zero near the tip of the cone, that is, it cannot concentrate at a point. Once this is known, it is easy to continue the solution beyond any maximal time of existence and prove the global existence and uniqueness of the solution.

In the supercritical case $p > p_0(n)$, very little is known at present; there is some indication that the problem is much more unstable than in the subcritical case (Kumlin, Brenner, Lebeau), and there is some numerical evidence in the same direction.

Global Small Waves

It was noted already in the 1960s (Segal, Strauss) that the equation in dimension $n \geq 2$

$$\Box u = f(u), \qquad u(0,x) = \varepsilon u_0(x), \qquad u_t(0,x) = \varepsilon u_1(x)$$
$$f(u) = O(|u|^\gamma) \quad \text{for } u \sim 0$$

with small data can be considered as a perturbation of the free wave equation and admits global solutions. The phenomenon may be regarded as follows: the wave operator tends to spread waves and reduce their size (see [21]); the nonlinear term tends to concentrate the peaks and make them higher, but at the same time it makes small waves smaller. If the rate of dispersion is fast enough, the initial data are small enough, and the power of the nonlinear term is high enough, the peaks have no time to concentrate, and the solution quickly flattens out to 0. Notice that in dimension 1 there is no dispersion, and this kind of mechanism does not occur.

It was, however, F John who initiated the modern study of this question by giving the complete picture in dimension 3: for the IVP

$$\Box u = |u|^\gamma, \qquad u(0,x) = \varepsilon u_0(x)$$
$$u_t(0,x) = \varepsilon u_1(x), \quad n = 3$$

he proved that, for fixed $u_0, u_1 \in C_0^\infty$,

- if $\gamma > 1 + \sqrt{2}$ and ε is small enough, the solution is global and
- if $1 < \gamma < 1 + \sqrt{2}$ and the data are not identically zero, the solutions blow up in a finite time for all ε (i.e., the L^∞-norm is unbounded).

Later Schaeffer proved that blow-up occurs also at the critical value $\gamma = 1 + \sqrt{2}$.

W Strauss guessed the correct critical value for all dimensions – $\gamma_0(n)$ is the positive root of the algebraic equation

$$\left(\frac{n-1}{2}\gamma - \frac{n+1}{2}\right)\gamma = 1$$

and conjectured that the same picture as in dimension 3 is valid for all dimensions $n \geq 2$.

Soon Sideris proved that, for $1 < \gamma < \gamma_0(n)$ and the quite general and small data, one always has blow-up. Also it was proved by Klainerman, Shatah, Christodoulou, and others that the positive part of the conjecture was true for $\gamma > \gamma_0(n)$, with a small gap near the critical value. The gap was closed by Georgiev, Lindblad, Sogge, who proved global existence for all $\gamma > \gamma_0(n)$. We also mention that the solution at the critical value $\gamma = \gamma_0(n)$ always seems to blow up; this is settled for low dimension (Schaeffer, Yordanov, Zhang and others), but the question is still not completely clear for large dimensions.

This problem has spurred a great deal of creativity, eventually leading to very fruitful results: the different approaches have proved useful in a variety of problems, sometimes quite different from the original semilinear equation. We mention a few:

- The weighted estimates of F John are estimates of the solution in spacetime L^p norms with weights of the form $(1 + |t| + |x|)^\alpha (1 + \|t| - |x\|)^\beta$. An extension of this method was also used in the final complete proof of the conjecture.
- The vector field approach of S Klainerman. If we regard energy estimates as norms generated by the plain derivatives, it is natural to extend them to more general norms generated by vector fields commuting, or quasicommuting, with the wave operator. The conservation of energy expressed in these generalized norms has a built-in decay that allows us to prove global existence of small waves. This circle of ideas led very far, and we might even regard Christodoulou and Klainerman's proof of the stability of Minkowski space for the Einstein equation as an extreme consequence of this approach.
- The normal forms of J Shatah. The idea is to apply a nonlinear (and nonlocal) transformation to the equation in order to increase the power γ. This method is effective for a variety of equations, including the semilinear wave, Klein–Gordon, and Schrödinger equations.
- The conformal transform method of D Christodoulou. The Penrose transform takes the wave operator on $\mathbb{R}^{1+n}$ to the wave operator on a bounded subset of $\mathbb{R} \times \mathbb{S}^n$, the so-called Einstein diamond (here $\mathbb{S}^n$ is the n-dimensional sphere). Thanks to the fact that a problem of global existence is converted into a problem of local existence, the proof reduces to showing that the lifespan of the local solution becomes large enough to cover the whole diamond when ε decreases.

A similar theory has been developed for the more general semilinear equation

$$\Box u = F(u, u'), \quad F(u, u') = O(|u, u'|^\gamma) \text{ for } u \sim 0$$

but the results are less complete. The general picture is similar: for $\gamma \geq 2$ when $n \geq 4$, and for $\gamma \geq 3$ when $n = 3$, one has global small solutions, while for γ close to 1 one in general has blow-up.

A very interesting phenomenon in this context was discovered by S Klainerman: some nonlinearities with a special structure, called "null structure," behave better than the others. This structure is clearly related to the wave operator, and in the end it can be precisely explained in terms of interaction of waves in phase space. We illustrate these ideas in the most interesting special case. Consider the equation in three dimensions

$$\Box u = F(D_t u, D_x u), \quad F = O(|u'|^\gamma),\ n = 3$$

In the "cubic" case $\gamma = 3$, one has global existence for all data small enough. On the other hand, in the "quadratic" case $\gamma = 2$, it is possible to construct examples where the solution blows up in a finite time no matter how small the data. Now, assume that the nonlinear term has the following structure:

$$F(u') = aQ_0(u') + \sum_{0 \leq j < k \leq 3} c_{jk} Q_{jk}(u') + O(|u'|^3) \quad [36]$$

which is called a "null structure". Here a, c_{jk} are constants, and the quadratic forms Q are the following:

$$Q_0(u') = |D_t u|^2 - |D_{x_1} u|^2 - |D_{x_2} u|^2 - |D_{x_3} u|^2 \quad [37]$$

$$Q_{0j}(u') = D_t u \cdot D_{x_j} u - D_{x_j} u \cdot D_t u, \quad j = 1, 2, 3 \quad [38]$$

$$Q_{jk}(u') = D_{x_j} u \cdot D_{x_k} u - D_{x_k} u \cdot D_{x_j} u \quad j, k = 1, 2, 3,\ j < k \quad [39]$$

Then the problem has a global solution for all small enough data. The extensions and applications of this idea are very wide (see the "Further reading" section for further information). Another situation where the null structure plays an important role is discussed in the next section.

Low Regularity

Theorem 1, although optimal in the classical framework, is not satisfactory for a few reasons. From a physicist's point of view, requiring $n/2+1$ derivatives of the data is not meaningful, since the measurable quantities involve only low-order derivatives, the most important one being the energy, that is, the H^1-norm of the solution. Moreover, the wave equation has a rich set of conserved quantities, symmetries and decay properties which may be useful to prove stronger results, and in particular the global existence. However, many of these structures appear only at a low-regularity level (H^1 or even L^p); in order to exploit them it is essential to work with low-regularity solutions.

As an example, if we were able to prove Theorem 1 for $k=1$, then we could deduce that the local solutions can be extended to global ones in all cases when the H^1-norm is conserved. For instance, this would allow us to solve globally the equations of the form

$$\Box u + G'(|u|^2)u = 0, \quad G(s) \geq 0$$

The problem of the lowest value of s such that a unique local solution exists in H^s is quite difficult, and still not completely solved. In order to state the results we precise the definition of solution as follows: the IVP is said to be locally well posed in H^s, if, for all (u_0, u_1) in a bounded set B of $H^s \times H^{s-1}$, there exist a $T>0$, a Banach space X_T (depending on B) continuously embedded in $C([0,T]; H^s)$, and a unique solution $u \in X_T$, such that the map $(u_0, u_1) \mapsto u$ is continuous from B to X_T.

For the wave equation with a power nonlinearity

$$\Box u = |u|^p \tag{40}$$

or more generally

$$\begin{aligned} &\Box u = F(u), \quad F(u) = 0 \\ &|F(u) - F(v)| \leq C|u-v|(|u|^{p-1} + |v|^{p-1}) \end{aligned} \tag{41}$$

the picture is almost complete. Indeed, by using the scaling

$$t \mapsto \lambda t, \qquad x \mapsto \lambda x \tag{42}$$

and the Lorentz transformation

$$t \mapsto \frac{\lambda t - x_1}{\sqrt{\lambda^2 - 1}}, \qquad x \mapsto \frac{t - \lambda x_1}{\sqrt{\lambda^2 - 1}} \tag{43}$$

it is possible to show by explicit constructions that

- the equation is not locally well posed for $p(n/2 - s) \leq (n/2 + 2 - s)$ (scaling) and
- the equation is not locally well posed for $p(n/4 + 1/4 - s) \leq n/4 + 5/4 - s$ (Lorentz).

On the positive side, local well-posedness has been almost fully proved in the complementary region of indices, with the exception of a tiny spot near the endpoint $s=0$, $p=(n+5)/(n+1)$ where the problem is still open (and the conjecture is that the equation is ill posed for indices in that region). These results are due to several authors, among the others we cite C Kenig, G Ponce, L Vega, H Lindblad, C Sogge, L Kapitanski, and T Tao.

When the nonlinearity depends also on the first-order derivatives of u, the situation becomes more complex. In the general case, the best result available is still the local existence theorem (Theorem 1); the only possible refinement is the use of fractional Sobolev spaces H^s, but in general local solvability only holds for $s > n/2+1$. If we assume that $F=F(u')$ is a quadratic form in the first-order derivatives, a clever use of Strichartz estimates allows us to prove local solvability down to $s > n/2 + 1/2$ for $n \geq 3$ and $s > 7/4$ for $n=2$ (Ponce and Sideris).

However, exactly as in the case of the small nonlinear waves examined in the previous section, if the nonlinear term has a null structure the result can be improved. Indeed, when $F(u')$ is a combination of the forms [37]–[39], then local solvability and uniqueness can be proved for all $s > n/2$, as in the case of a nonlinear term of the type $F(u)$. This result is due to Klainerman, Machedon, and Selberg. Again, the proof is based on a variation of the contraction method; the additional ingredient here is the use of suitable function spaces, which are the counterpart for the wave equation of the spaces used by Bourgain in the study of the nonlinear Schrödinger equation. The norm of these spaces is defined as follows:

$$\|u\|_{H^{s,\theta}} \equiv \|\langle\xi\rangle^s \langle |\tau| - |\xi| \rangle^\theta \widetilde{u}(\tau, \xi)\|_{L^2(\mathbb{R}^{n+1})}$$

where $\langle\xi\rangle = (1+|\xi|^2)^{1/2}$ and $\widetilde{u}$ is the spacetime Fourier transform of $u(t,x)$. The wave operator can be regarded as a spacetime Fourier multiplier of the form $\tau^2 - |\xi|^2 = (|\tau| - |\xi|)(|\tau| + |\xi|)$, and we see that "inverting" the operator $\Box$ has a regularizing effect in the scale of $H^{s,\theta}$ spaces, since it decreases both

s and θ by one unit. Substantiating this formal argument and complementing it with suitable estimates for the nonlinear term requires some hard work, which is contained in the theory of bilinear estimates developed by Klainerman and his school.

See also: Evolution Equations: Linear and Nonlinear; Symmetric Hyperbolic Systems and Shock Waves; Wave Equations and Diffraction.

Further Reading

Choquet-Bruhat Y (1988) Global existence for nonlinear σ-models, Rend. Sem. Mat. Univ. Pol. Torino, Fascicolo speciale "Hyperbolic equations," 65–86.

D'Ancona P, Georgiev V, and Kubo H (2001) Weighted decay estimates for the wave equation. *Journal of Differential Equations* 177: 146–208.

Georgiev V, Lindblad H, and Sogge CD (1997) Weighted Strichartz estimates and global existence for semilinear wave equations. *American Journal of Mathematics* 119: 1291–1319.

Ginibre J and Velo G (1982) The Cauchy problem for the o(N), cp($N-1$), and $G_C(N,p)$ models. *Annals of Physics* 142: 393–415.

Ginibre J and Velo G (1995) Generalized Strichartz inequalities for the wave equation. *Journal of Functional Analysis* 133: 50–68.

Hörmander L (1997) *Lectures on Nonlinear Hyperbolic Equations*. Berlin: Springer.

John F (1979) Blow-up of solutions of nonlinear wave equations in three space dimensions. *Manuscripta Mathematica* 28(1–3): 235–268.

Keel M and Tao T (1998) Endpoint Strichartz estimates. *American Journal of Mathematics* 120: 955–980.

Klainerman S (1986) The null condition and global existence to nonlinear wave equations. *Lecture in Applied Mathematics* 23: 293–326.

Klainerman S and Selberg S (2002) Bilinear estimates and applications to nonlinear wave equations. *Communications in Contemporary Mathematics* 4: 223–295.

Lindblad H and Sogge C (1995) Existence and scattering with minimal regularity for semilinear wave equations. *Journal of Functional Analysis* 130: 357–426.

Schiff LI (1951) Nonlinear meson theory of nuclear forces I. *Physical Review* 84: 1–9.

Segal IE (1963) The global Cauchy problem for a relativistic scalar field with power interaction. *Bulletin de la Société Mathématique de France* 91: 129–135.

Shatah J (1988) Weak solutions and the development of singularities in the SU(2) σ-model. *Communications on Pure and Applied Mathematics* 41: 459–469.

Shatah J and Struwe M (1993) Regularity results for nonlinear wave equations. *Annals of Mathematics* 138: 503–518.

Shatah J and Struwe M (1998) *Geometric Wave Equations*. Courant Lecture Notes in Mathematics, vol. 2. New York: Courant Institute, New York University.

Strauss W (1989) *Nonlinear Wave Equations*. CBMS Lecture Notes, vol. 75. Providence, RI: American Mathematical Society.

Separation of Variables for Differential Equations

S Rauch-Wojciechowski, Linköping University, Linköping, Sweden
K Marciniak, Linköping University, Norrköping, Sweden

Introduction

The method of separation of variables (SoV) is a way of finding particular and general solutions of certain types of partial differential equations (PDEs). Its main idea is to consider the additive ansatz $u(x)=\sum_i w_i(x^i,\alpha)$ or the multiplicative ansatz $u(x)=\prod_i u_i(x^i,\alpha)$ for a solution of a PDE that allows for reducing this PDE to a set of (uncoupled) ordinary differential equations (ODEs) for the unknown functions $w_i(x^i,\alpha)$ or $u_i(x^i,\alpha)$ of one variable x^i, where $x=(x^1,\ldots,x^n)$. Locally, the additive ansatz is, through the change of variables $u(x)=\exp(\sum_i w_i(x^i,\alpha))$, equivalent to the multiplicative ansatz.

Many well-known equations of mathematical physics such as the heat equation, the wave equation, the Schrödinger equation, and the Hamilton–Jacobi equation are solved by separating variables in suitably chosen systems of coordinates.

Fourier Method

The SoV method can be attributed to Fourier (1945), who solved the heat equation

$$\partial_t u=\partial_{xx}u \qquad [1]$$

for distribution of temperature $u(x,t)$ in a one-dimensional metal rod (of length L) by looking first for special solutions of the product type $u(x,t)=X(x)T(t)$. This ansatz, substituted to [1], reduces it to two ODEs: $\partial_t T=-k^2T$ and $\partial_{xx}X=-k^2X$ that can be solved by quadratures:

$$T_k(t)=Ae^{-k^2t}, \qquad X_k(x)=B\cos(kx)+C\sin(kx)$$

Due to linearity of [1], any formal linear combination $u(x,t)=\sum_k c_kX_k(x)T_k(t)$ is again a solution of the heat equation and can be used for solving an initial boundary-value problem (IBVP). For instance,

in the case of the IBVP on the interval $0 \le x \le L$ and with zero boundary conditions

$$\begin{aligned} &\partial_t u = \partial_{xx} u, && 0 < t, 0 < x < L \\ &u(0,t) = u(L,t) = 0, && 0 < t \\ &u(x,0) = f(x), && 0 < x < L \end{aligned}$$

only a countable set of values for the separation constant k is admissible: $k_n = (n\pi/L)$, $n = 1, 2, \ldots$. Then the general solution has the form of the Fourier series

$$u(x,t) = \sum_{n=1}^{\infty} c_n \exp\left(-k_n^2 t\right) \sin(k_n x)$$

where the coefficients c_n are given by the integrals

$$c_n = \frac{2}{L} \int_0^L f(x) \sin(k_n x)\, dx$$

The sequence of functions $\sin(k_n x)$ is complete on the interval $[0, L]$. That means that any regular (continuous and differentiable) initial data function $f(x)$ such that $f(0) = f(L) = 0$ can be uniquely expressed as an infinite convergent sum of the orthogonal set of functions $\sin(k_n x)$. The study of mathematical properties of the Fourier expansion gave rise to the classical theory of Fourier series and Fourier integrals.

Separability of PDEs in General Setting

A general setting for an additive separability of a single, usually nonlinear, PDE has been developed by Levi-Civita (1904) and by Kalnins and Miller (1980) (see also Miller (1983)). Let

$$\begin{aligned} &H(x^1, \ldots, x^n; u, u_i, u_{ij}, u_{ijk}, \ldots) = E \\ &\quad 1 \le i, j, k \le n \end{aligned} \qquad [2]$$

be a finite-order PDE for an unknown function $u(x)$, where $u_i(x) = \partial_{x^i} u$, $u_{ij} = \partial_{x^j} \partial_{x^i} u$, etc., and E is a constant. A separable solution $u(x) = \sum_i W_i(x^i)$ satisfies the simpler equation

$$E = H(x; u, u_i, u_{ii}, \ldots) \equiv H[x, u] \qquad [3]$$

where all mixed derivatives u_{ij}, etc., disappear. If a separable solution is admissible by eqn [2], then the function $H(x; u, u_i, u_{ii}, \ldots)$ has to satisfy a set of integrability conditions following from the total derivatives of [3]. Let

$$\begin{aligned} D_i &= \partial_{x^i} + u_{i,1} \partial_u + u_{i,2} \partial_{u_{i,1}} + \cdots + u_{i,m_i+1} \partial_{u_{i,m_i}} \\ &\equiv \widetilde{D}_i + u_{i,m_i+1} \partial_{u_{i,m_i}} \end{aligned}$$

(where $u_{i,1} = u_i, u_{i,j+1} = \partial_{x^i} u_{i,j}$, etc., and m_i is the largest number l such that $\partial_{u_{i,l}} H \ne 0$) denote the operator of total derivative with respect to (w.r.t.) x^i; then, $D_i H[x, u] = 0$ or

$$u_{i,m_i+1} = -\frac{\widetilde{D}_i H}{H_{u_i,m_i}}$$

where $H_{u_i,m_i} = \partial_{u_i,m_i} H$. The integrability conditions $D_j u_{i,m_i+1} = 0$, $j \ne i$, give rise to a large set of differential conditions to be satisfied by $H[x, u]$:

$$\begin{aligned} &H_{u_i,m_i} H_{u_j,m_j} \left(\widetilde{D}_i \widetilde{D}_j H\right) + H_{u_i,m_i\, u_j,m_j} \left(\widetilde{D}_i H\right)\left(\widetilde{D}_j H\right) \\ &\quad = H_{u_j,m_j} \left(\widetilde{D}_i H\right)\left(\widetilde{D}_j H_{u_i,m_i}\right) \\ &\qquad + H_{u_i,m_i} \left(\widetilde{D}_j H\right)\left(\widetilde{D}_i H_{u_j,m_j}\right) \end{aligned} \qquad [4]$$

In general, the conditions [4] are restrictions for both H and the form of a particular separable solution $u(x)$. If [4] is satisfied identically w.r.t. all u, $u_{k,l}$, we say that the corresponding coordinate system x^i is a regular separable coordinate system; then the PDE [3] admits a $(\sum_i m_i + 1)$-parameter family of separable solutions. Most cases considered in literature are regular; since then the separable solution is usually sufficiently general for solving various IBVPs.

A given PDE, however, usually does not satisfy [4]; since these equations are not of tensorial type, the natural question arises if there exists a suitable change of coordinates $y(x)$ such that the transformed PDE satisfies [4]. Such separation coordinates may or may not exist; it is usually very difficult to decide.

Here and in what follows, we speak about separability of a single (scalar) PDE. The theory of separability of systems of PDEs is still not developed fully, although it is of relevance in the theory of Maxwell equations and of the Dirac equation.

We present here the most classical part of SoV theory: orthogonal separability of the Hamilton–Jacobi equation for geodesic motions on Riemannian manifolds.

Configurational Separation of Hamilton–Jacobi Equation on Riemannian Manifolds

Around 1842, C G J Jacobi invented the method of generating function for solving the canonical Hamilton equations

$$\begin{aligned} &\dot{x} = \frac{\partial H(x,y)}{\partial y}, && \dot{y} = -\frac{\partial H(x,y)}{\partial x} \\ &x = (x^1, \ldots, x^n) && y = (y^1, \ldots, y^n) \end{aligned} \qquad [5]$$

where $H(x, y)$ is a Hamiltonian and dot denotes the time derivative (Landau and Lifshitz 1976). In this

method, one looks for a generating function $W(x,\alpha)$ of a canonical transformation

$$y = \frac{\partial W(x,\alpha)}{\partial x}, \qquad \beta = \frac{\partial W(x,\alpha)}{\partial \alpha}$$

that transforms Hamiltonian equations [5] into simple equations for the new variables $\beta \in R^n$, $\alpha \in R^n$. Since the transformation is canonical, the transformed equations are again Hamiltonian with the new Hamiltonian $\widetilde{H}(\beta,\alpha) = H(x(\beta,\alpha), y(\beta,\alpha))$. If we choose this transformation so that $\widetilde{H}(\beta,\alpha) = \alpha_1$, then the transformed Hamilton equations become

$$\dot{\beta} = \frac{\partial \widetilde{H}(\beta,\alpha)}{\partial \alpha} = (1,0,\ldots,0)$$
$$\dot{\alpha} = -\frac{\partial \widetilde{H}(\beta,\alpha)}{\partial \beta} = 0$$

so that $\beta(t) = (t + \beta_{10}, \beta_{20}, \ldots, \beta_{n0})$, $\alpha(t) = (\alpha_{10},\ldots,\alpha_{n0}) = \text{const.}$ and the solution $x(t), y(t)$ of the Hamilton equations [5] is then given implicitly by the equations

$$\beta(t) = \frac{\partial W(x(t),\alpha)}{\partial \alpha}, \qquad y(t) = \frac{\partial W(x(t),\alpha)}{\partial x}$$

Since

$$y = \frac{\partial W(x,\alpha)}{\partial x}$$

the generating function $W(x, \alpha)$ has to satisfy (identically w.r.t. (x, α)) the first-order nonlinear PDE

$$H\left(x, \frac{\partial W(x,\alpha)}{\partial x}\right) = \alpha_1 \qquad [6]$$

This equation is called the Hamilton–Jacobi equation for the generating function $W(x,\alpha)$. It is solved when its complete integral $W(x,\alpha)$, complete means that

$$\det\left(\frac{\partial^2 W(x,\alpha)}{\partial x^i \partial \alpha_j}\right) \neq 0$$

depending on n independent constants α is known. In general, it is very difficult to find solutions of [6]. The most important method is the method of separation of variables when one looks for a solution in the form $W(x,\alpha) = \sum_{k=1}^{n} W_k(x^k,\alpha)$ which is a sum of n functions $W_k(x^k,\alpha)$, each depending on a single variable x^k and, possibly, all constants a. If the Hamilton–Jacobi equation [6] admits such a solution, then integrating this equation is reduced to integrating n (uncoupled) first-order ODEs for functions $W_k(x^k,\alpha)$. The constants α_k acquire then the meaning of integration constants.

A separable solution $W(x,\alpha)$ of [6] exists whenever the Hamiltonian $H(x,y)$ satisfies (identically) the integrability conditions [4] which in this case acquire the (nonlinear) form

$$\begin{aligned} L_{ij}(H) \equiv{}& \partial_i H \partial_j H \partial^i \partial^j H + \partial^i H \partial^j H \partial_i \partial_j H \\ & - \partial_i H \partial^j H \partial^i \partial_j H - \partial^i H \partial_j H \partial_i \partial^j H \\ ={}& 0 \quad \text{for all } i,j = 1,\ldots,n \end{aligned} \qquad [7]$$

$(\partial_i = \partial/\partial x^i, \partial^i = \partial/\partial y_i)$ found by Levi-Civita (1904).

In classical mechanics the most important Hamiltonians are natural ones:

$$H(x,y) = \frac{1}{2}\sum_{i,j} g^{ij}(x) y_i y_j + V(x) \equiv G + V \qquad [8]$$

They are defined on the cotangent bundle T^*Q of a configurational Riemannian manifold Q with the metric tensor g. The function G is the geodesic Hamiltonian associated with the metric tensor g. For such natural Hamiltonians, the Levi-Civita condition $L_{ij}(G+V) = 0$ splits into the condition $L_{ij}(G) = 0$ and a condition for the potential $V(x)$. The condition $L_{ij}(G) = 0$, depending solely on the kinetic energy term, is thus a necessary condition for coordinates x^i on Q to be separation coordinates for [8].

In the fundamental case of orthogonal separation (i.e., when $g^{ij} = 0$ for $i \neq j$), the Levi-Civita conditions $L_{ij}(G+V) = 0$ read

$$\begin{aligned} & \partial_i \partial_j g^{kk} - (\partial_i \ln g^{jj}) \partial_j g^{kk} \\ & \quad - (\partial_j \ln g^{ii}) \partial_i g^{kk} = 0, \quad i \neq j \end{aligned} \qquad [9]$$

$$\begin{aligned} & \partial_i \partial_j V - (\partial_i \ln g^{jj}) \partial_j V \\ & \quad - (\partial_j \ln g^{ii}) \partial_i V = 0, \quad i \neq j \end{aligned} \qquad [10]$$

The main questions arising here are

1. What is the algebraic form of orthogonally separable Riemannian metrics?
2. What is the form of separable coordinates on Riemannian manifolds?

The first question is answered by the Stäckel theorem (Stäckel 1891) that provides an algebraic characterization of orthogonal separability of a natural Hamiltonian $H = G + V$.

Theorem 1 *The Hamilton–Jacobi equation for the natural Hamiltonian*

$$H = G + V = \frac{1}{2}\sum_i g^{ii}(x) y_i^2 + V(x)$$

is separable in the (orthogonal) coordinates x if and only if

(i) *There exists a matrix* $\Phi = [\varphi_{ij}(x^i)]$, $\det(\Phi) \neq 0$ *(so that the row i depends only on x^i) such that* $[g^{11}, \ldots, g^{nn}]$ *is the first row of the inverse matrix* $\Psi = \Phi^{-1}$.

(ii) *The potential V has the form* $V(x) = \sum_i g^{ii} f_i(x^i)$, *where each $f_i(x^i)$ is a function of one variable x^i only.*

Such matrix Φ is called a Stäckel matrix.

Proof If

$$[g^{11}, \ldots, g^{nn}] \begin{bmatrix} \varphi_{11}(x^1) & \cdots & \varphi_{1n}(x^1) \\ \vdots & \ddots & \vdots \\ \varphi_{n1}(x^n) & \cdots & \varphi_{nn}(x^n) \end{bmatrix} = [1, 0, \ldots, 0] \quad [11]$$

then the Hamilton–Jacobi equation for H can be written as

$$\frac{1}{2}\sum_i g^{ii}\left(\frac{\partial W}{\partial x^i}\right)^2 + \sum_i g^{ii} f_i(x^i) = \alpha_1 = \alpha_1 \sum_i g^{ii}\varphi_{i1}(x^i) + \alpha_2 \sum_i g^{ii}\varphi_{i2}(x^i) + \cdots + \alpha_n \sum_i g^{ii}\varphi_{in}(x^i) \quad [12]$$

This equation admits an additively separable solution $W = \sum_i W_i(x^i)$, where the functions W_i satisfy n ODEs (separation equations):

$$\frac{1}{2}\left(\frac{\partial W_i}{\partial x^i}\right)^2 + f_i(x^i) = \alpha_1\varphi_{i1}(x^i) + \alpha_2\varphi_{i2}(x^i) + \cdots + \alpha_n\varphi_{in}(x^i) \quad i = 1, \ldots, n \quad [13]$$

By differentiating [13] w.r.t. α_j, we get

$$\varphi_{ij}(x^i) = \frac{\partial W_i}{\partial x^i}\frac{\partial^2 W_i}{\partial x^i \partial \alpha_j}$$

and thus

$$\det[\varphi_{ij}(x^i)] = \frac{\partial W_1}{\partial x^1}\cdots\frac{\partial W_n}{\partial x^n}\det\left(\frac{\partial^2 W}{\partial x^i \partial \alpha_j}\right) \neq 0$$

so that $W = \sum_i W_i(x^i)$ is indeed a complete integral of the Hamilton–Jacobi equation [12]. Conversely, if $W = \sum_i W_i(x^i)$ is a complete integral of the Hamilton–Jacobi equation [12], then by differentiating it w.r.t. α_j we get for $j = 1$

$$\sum_i g^{ii}\frac{\partial W_i}{\partial x^i}\frac{\partial^2 W_i}{\partial x^i \partial \alpha_j} = 1$$

and

$$\sum_i g^{ii}\frac{\partial W_i}{\partial x^i}\frac{\partial^2 W_i}{\partial x^i \partial \alpha_j} = 0$$

(for $j = 2, \ldots, n$), that is, the condition [11] for the Stäckel matrix

$$\Phi = \left[\frac{\partial W_i}{\partial x^i}\frac{\partial^2 W_i}{\partial x^i \partial \alpha_j}\right]$$

Further, we see that

$$V = \alpha_1 - \frac{1}{2}\sum_i g^{ii}(\partial_{x^i} W_i)^2 = \frac{1}{2}\sum_i g^{ii}\left[\alpha_1\varphi_{i1}(x^i) - \frac{1}{2}(\partial_{x^i} W_i)^2\right] = \sum_i g^{ii} f_i(x^i)$$ □

Remark 2 The Stäckel characterization of orthogonal separability is equivalent to Levi-Civita conditions [9] and [10]. It is in fact a solution of these conditions.

Remark 3 With every Stäckel matrix, one can relate a family of n quadratic in momenta Hamiltonians defined by n rows of the inverse Stäckel matrix $\Psi = \Phi^{-1} = [\psi_{kr}]$:

$$H_k = \frac{1}{2}\sum_{r=1}^n \psi_{kr} y_r^2, \quad k = 1, \ldots, n \quad [14]$$

(so that $H_1 = G$). These Hamiltonians are linearly and functionally independent; they Poisson-commute (so that they form a Liouville integrable system) and are all diagonal so that they have common eigenvectors.

These properties are the main ingredients of an intrinsic (coordinate-independent) characterization of separable geodesic Hamiltonians G in terms of involutive Killing tensors that is due to works of Eisenhart (1934), Kalnins and Miller (1980), and Benenti (1997).

Theorem 4 *A necessary and sufficient condition for the existence of an orthogonal additive separable coordinate system x for the Hamilton–Jacobi equation of the geodesic Hamiltonian $H_1 = G$ on an n-dimensional (pseudo)-Riemannian manifold is that there exist n quadratic forms* $H_r = \sum_{i,j}^n h_r^{ij}(x) y_i y_j$ *such that*

(i) *They all Poisson-commute*: $\{H_r, H_s\} = 0, 1 \leq r, s \leq n$.

(ii) *The set* $\{H_r\}_{r=1}^n$ *is linearly independent.*

(iii) *There is a basis* $\{\omega_{(j)}\}_{j=1}^n$ *of n simultaneous eigenforms for all H_r.*

If conditions(i)–(iii) are satisfied then there exist functions $g_j(x)$ *such that* $\omega_{(j)} = g_j \mathrm{d}x^j$, $j = 1, \ldots, n$.

This theorem has been further simplified by Benenti (1997), who has shown that for separability it is sufficient that g^{ij} admits a single Killing 2-tensor with simple eigenvalues and normal eigenvectors. He has also explained the role of ignorable coordinates.

These results are key ingredients of an answer to the question (2). Eisenhart (1934), starting from the fact that every separable geodesic Hamiltonian $H = G$ admits n quadratic (w.r.t. momenta y_i) integrals of motion, derived a set of nonlinear PDEs characterizing separable Riemannian metrics. He has solved these equations for spaces of constant curvature. This solution is the basis of the Kalnins and Miller's (1986) diagrammatic classification of all orthogonal separation coordinates on $\boldsymbol{R}^n$ and the sphere $\boldsymbol{S}^n$. Separable coordinates on the Minkowski space $\boldsymbol{M}^n$ have not been classified yet.

Since the work of Robertson (1927) and Eisenhart (1934), it is known that in $\boldsymbol{R}^n, \boldsymbol{S}^n$ and, in general, in the space with diagonal Ricci tensor, the (additive) separability of Hamilton–Jacobi equation for the natural Hamiltonian $H = G + V$ is equivalent to multiplicative separability of the stationary Schrödinger equation with the same potential V:

$$(\Delta + V(x))\Theta(x) = E\Theta(x) \qquad [15]$$

where

$$\Delta = \sum_{i,j=1}^{n} \frac{1}{\sqrt{\det(g)}} \partial_i \left(\sqrt{\det(g)} g^{ij} \partial_j \right)$$

is the Laplace–Beltrami operator. Usually, multiplicative separated solutions $\Theta(x) = \prod_{i=1}^{n} \Theta_i(x)$ is considered but the change of the dependent variable $u = \ln \Theta$ transforms it into an additive separable solution. If we restrict our considerations to orthogonal separation coordinates ($g^{ij} = 0$ for $i \neq j$), eqn [15] becomes

$$\sum_{i=1}^{n} \left(g^{ii}\left(u_{ii} + u_i^2\right) + \frac{1}{\sqrt{\det(g)}} \partial_i \times \left(\sqrt{\det(g)} g^{ii} \right) u_i \right) + V(x) = E$$

where $u_i = \partial_i u$, $u_{ii} = \partial_i \partial_i u$. The integrability conditions [4] for regular separation lead to the Levi-Civita condition [9] on the components g^{ii} of the metric tensor, upon comparison of the coefficients at u_i^2. The coefficients at u_{ii} yield the Robertson condition

$$\partial_i \partial_j \ln \left(\sqrt{\det(g)} g^{ii} \right) = 0, \quad i \neq j$$

and the constant terms in [4] give the Levi-Civita equation [10] meaning that $V(x) = \sum_{i=1}^{n} g^{ii} f_i(x^i)$. Eisenhart has shown that the Robertson condition is equivalent to the requirement that the Ricci tensor is diagonal: $R_{ij} = 0$, $i \neq j$ in variables x so that the Robertson condition is satisfied automatically in the Euclidean space, in spaces of constant curvature and in Einstein spaces. Thus every orthogonal coordinate system permitting multiplicative separation of the Schrödinger equation corresponds to the Stäckel form.

Jacobi Problem of Separability

In order to apply the separability theory to physical Hamiltonians $H = (1/2)p^2 + V(q)$, $p = (p_1, \ldots, p_n)$, $q = (q^1, \ldots, q^n)$, it is essential to solve the following problem: "given a potential $V(q)$, decide if there exists a point transformation $x(q)$ to some curvilinear coordinates x such that the Hamilton–Jacobi equation associated with H is separable in coordinates x, and if such transformation exists, determine it and solve the obtained Hamilton–Jacobi equation."

This problem has been raised by Jacobi (1884) in connection with the problem of finding geodesic motions on a 3-axial ellipsoid. For solving this problem Jacobi introduced his "remarkable change of coordinates" to the generalized elliptic coordinates $x(q)$ defined through zeros of the rational function

$$1 + \sum_{i=1}^{n} \frac{(q^i)^2}{(z - \lambda_i)} \equiv \frac{\prod_j (z - x^j)}{\prod_i (z - \lambda_i)} \qquad [16]$$

where the constants $\lambda_i > 0$ are all different. From the graph of the left-hand side of [16], it is easy to see that there are exactly n simple, real zeros. For given values of elliptic coordinates x^j, the values of $(q^i)^2$ are uniquely determined as residues at λ_i while Cartesian coordinates q^i are determined uniquely only in each n-tant of $\boldsymbol{R}^n$.

The Jacobi elliptic coordinates play a pivotal role in orthogonal separability on $\boldsymbol{R}^n$ and $\boldsymbol{S}^n$ since they are the mother of all other separation coordinates that can be obtained through proper and improper degenerations of λ_i's. By using these coordinates Jacobi solved not only the geodesic motions on the ellipsoid but also the motion on the ellipsoid under the action of harmonic potential $V(q) = (1/2)q^2$. He has also found separation coordinates for a system of three interacting particles on the line known today as the Calogero system. In general, however, Jacobi considered the problem of finding separation coordinates for a given potential $V(q)$ to be very difficult. In *Vörlesungen über Dynamik*, ch. 26, he writes: "The main difficulty in integrating a given

differential equation lies in introducing convenient variables, which there is no rule for finding. Therefore, we must travel the reverse path and after finding some notable substitution, look for problems to which it can be successfully applied". This statement had a profound influence on further development of SoV theory that concentrated on characterizing separable Hamiltonians (as expressed in terms of separation coordinates) and on describing and classifying separation coordinates.

The original problem of Jacobi of finding separation variables for a given natural Hamiltonian has been taken up by Rauch-Wojciechowski (1986), who found a characterization of separable potentials $V(q)$ in terms of Cartesian coordinates q_i. Its invariant geometric form has been given by Benenti. A complete criterion of separability that allows for an effective testing and calculation of separation coordinates (if they exist) for $V(q)$ has been solved by Waksjö and Rauch-Wojciechowski (2003). This criterion is directly applicable to the problem of finding SoV for the Schrödinger equation.

Criterion of Separability for $n=2$

The criterion of separability for $n=2$ can be read from the Bertrand–Darboux theorem.

Theorem 5 (Bertrand–Darboux). *For the Hamiltonian*:

$$H=\tfrac{1}{2}(p_1^2+p_2^2)+V(q_1,q_2)$$

the following statements are equivalent:

(i) *H has a functionally independent integral of motion $\{H,K\}=0$ of the form*

$$\begin{aligned}K=&\left(aq_2^2+bq_2+c\right)p_1^2+\left(aq_1^2+\widetilde{b}q_1+\widetilde{c}\right)p_2^2\\&+\left(-2aq_1q_2-bq_1-\widetilde{b}q_2+d\right)p_1p_2\\&+k(q_1,q_2)\end{aligned}$$

(ii) *The potential $V(q_1,q_2)$ satisfies the following linear second-order PDE with quadratic coefficients*

$$\begin{aligned}0=&2\left(aq_2^2-aq_1^2+bq_2-\widetilde{b}q_1+c-\widetilde{c}\right)\partial_1\partial_2V\\&+\left(-2aq_1q_2-bq_1-\widetilde{b}q_2+d\right)\left(\partial_2^2V-\partial_1^2V\right)\\&+(6aq_2+3b)\partial_1V-(6aq_1+3\widetilde{b})\partial_2V\end{aligned}\qquad[17]$$

where $a,b,\widetilde{b},c,\widetilde{c},d$ are some constants, $\partial_1=\partial_{q_1},\partial_2=\partial_{q_2}$.

(iii) *The Hamilton–Jacobi equation for H is separable in one of the four orthogonal coordinate systems in the plane: elliptic, parabolic, polar, or Cartesian.*

Remark 6 If the potential $V(q_1,q_2)$ is separable, then it admits an integral of motion K that is quadratic w.r.t. momenta and V satisfies (identically w.r.t. q_1,q_2) eqn [17] for certain values of the undetermined constants $a,b,\widetilde{b},c,\widetilde{c},d$. Since coefficients at linearly independent expressions of q_1,q_2 have to be equal to zero, the parameters $a,b,\widetilde{b},c,\widetilde{c},d$ have to satisfy a set of linear, algebraic, homogeneous equations. If there is a nonzero solution for $a,b,\widetilde{b},c,\widetilde{c},d$, then there exists an integral of motion K and separation coordinates can be determined as characteristic variables for equation [17].

Example 7 Separable cases of the Henon–Heiles potential

$$V=\tfrac{1}{2}\left(\omega_1q_1^2+\omega_2q_2^2\right)+\alpha q_1^2q_2-\tfrac{1}{3}\beta q_2^3$$

By substituting this form of V into [17], we get two sets of admissible solutions for parameters $\alpha,\beta,\omega_1,\omega_2$: (i) $\beta=-\alpha,\omega_1=\omega_2$ with V separable in rotated (by $\pi/4$) Cartesian coordinates; (ii) $\beta=-6\alpha$, ω_1,ω_2-arbitrary with V separable in the shifted parabolic coordinates. In case (ii) eqn [17] becomes

$$\begin{aligned}&2\left(q_2-\frac{1}{4\alpha}(4\omega_1-\omega_2)\right)\partial_1\partial_2V\\&\quad+q_1\left(\partial_2^2V-\partial_1^2V\right)+3\partial_1V=0\end{aligned}$$

and in its characteristic coordinates defined as $q_1=\sqrt{\xi\eta}$, $q_2=(1/2)(\xi-\eta)+(1/4\alpha)(4\omega_1-\omega_2)$ it takes the form $(\xi-\eta)\partial_\xi\partial_\eta V+\partial_\xi V+\partial_\eta V=0$ solved by $V(\xi,\eta)=(\xi+\eta)^2[f(\xi)+g(\eta)]$ which is separable in the parabolic coordinates.

Effective Criterion of Separability for Arbitrary Dimension

For $n>2$, a similar theorem characterizing separability in generalized elliptic coordinates has been formulated by Rauch-Wojciechowski (1986).

Theorem 8 (Elliptic Bertrand–Darboux). *For a natural Hamiltonian $H=(1/2)p^2+V(q)$, the following statements are equivalent*:

(i) *H has n global, functionally independent and involutive integrals of motion $\{H,K_i\}=0$, $\{K_i,K_j\}=0$, $i,j=1,\ldots,n$, having the form*

$$\begin{aligned}K_i&=\sum_{r=1,r\neq i}^{n}(\lambda_i-\lambda_r)^{-1}l_{ir}^2+p_i^2+k_i(q)\\l_{ir}&=q_ip_r-q_rp_i\end{aligned}\qquad[18]$$

(ii) *The potential V satisfies the following system of linear second-order PDEs*

$$(\lambda_i - \lambda_j)\partial_i\partial_j V - \Im_{ij}(2 + \Re)V = 0 \quad i,j = 1,\ldots,n,\ i \neq j \qquad [19]$$

$$\lambda_i\partial_i\Im_{jk}V + \lambda_j\partial_j\Im_{ki}V + \lambda_k\partial_k\Im_{ij}V = 0 \quad \textit{all } i,j,k \textit{ different} \qquad [20]$$

where $\Im_{ij} = q_i\partial_j - q_j\partial_i$, $\Re = \sum_{r=1}^n q_r\partial_r$.

(iii) *The Hamilton–Jacobi equation for H is separable in the generalized elliptic coordinates* [16] *with parameters λ_i.*

Remark 9 Equations [19]–[20] follow from the compatibility conditions that mixed derivatives of $k_i(q)$ calculated from the conditions $\{H, K_r\} = 0$, are equal. This leads to an overdetermined system [19]–[20] of PDEs for $V(q)$. Equations [19]–[20] are not linearly independent but we keep both sets [19]–[20] in the formulation of this theorem because eqns [19] give rise to the basic Bertrand–Darboux equations [21] used in the criterion of separability while eqns [20] give rise to cyclic Bertrand–Darboux equations [22] used for testing the level of spherical symmetry in the potential.

For testing elliptic separability of any given potential $V(q)$, it is necessary to introduce into eqns [19] and [20] the freedom of choice of the Euclidean reference frame (as described by the Euclidean transformation $\widetilde{q} = A^t(q - b)$, $A \in \mathrm{SO}(n)$, $b \in \boldsymbol{R}^n$). By substituting it into [19]–[20], omitting tildes and summing over one of the indices, we obtain new equations

$$\begin{aligned} 0 = \sum_{k=1}^n \big(&(\alpha q_i q_k + \beta_i q_k + \beta_k q_i + \gamma_{ik})\partial_k\partial_j V \\ &- (\alpha q_j q_k + \beta_j q_k + \beta_k q_j + \gamma_{jk})\partial_k\partial_i V \\ &+ 3\big((\alpha q_i + \beta_i)\partial_j V - (\alpha q_j + \beta_j)\partial_i V\big)\big) \\ &i,j = 1,\ldots,n,\ i \neq j \end{aligned} \qquad [21]$$

$$\begin{aligned} 0 = \sum_{l=1}^n \big(&(\gamma_{il}q_j - \gamma_{jl}q_i)\partial_k\partial_l V + (\gamma_{jl}q_k - \gamma_{kl}q_j)\partial_i\partial_l V \\ &+ (\gamma_{kl}q_i - \gamma_{il}q_k)\partial_j\partial_l V\big) \end{aligned} \qquad [22]$$

with the new coefficients α, β_i, γ_{ij} that are unconstrained despite that the orthogonal matrix A satisfies the quadratic algebraic constraint $AA^t = I$.

Theorem 8 provides the following test of elliptic separability for a potential $V(q)$ given in Cartesian coordinates.

1. Insert $V(q)$ into the Bertrand–Darboux equations [21]. This gives a system of linear, homogeneous, algebraic equations for the unknown parameters α, β_i, γ_{ij}. If $\alpha = 0$, then $V(q)$ is not separable in elliptic coordinates.
2. If $\alpha \neq 0$, set $b = -\alpha^{-1}\beta$, $S = bb^t - \alpha^{-1}\gamma$ and diagonalize S: $S = A\,\mathrm{diag}(\lambda_1,\ldots,\lambda_n)A^t$. If some eigenvalues λ_i coincide, then $V(q)$ is not separable in elliptic coordinates. Otherwise $V(q)$ is separable in the elliptic coordinates $x = (x^1,\ldots,x^n)$ given by

$$1 + \sum_{i=1}^n \frac{(\widetilde{q}^i)^2}{(z - \lambda_i)} \equiv \frac{\prod_j (z - x^j)}{\prod_i (z - \lambda_i)}$$

(compare with [16]), where $q = A\widetilde{q} + b$, with b and A found as above.

If $\alpha = 0$, $\beta \neq 0$, then there exists a similar algorithm for separability in generalized parabolic coordinates and for $\alpha = 0$, $\beta = 0$, $\gamma \neq tI$, we have separability in Cartesian coordinates if all λ_i are different. For giving an idea of what happens when degenerations occur, consider the case $\alpha = 0$, $\beta = 0$. Then the Bertrand–Darboux equations [21] are Euclidean equivalent to the canonical form $(\lambda_i - \lambda_j)\ \partial_i\partial_j V = 0$ and if all λ_i are different, then equations $\partial_i\partial_j V = 0$ imply that $V(q)$ is a sum of functions of one variable only: $V(q) = \sum_{i=1}^n V_i(q^i)$.

The main problem is to handle all possible degenerations when certain λ's coincide. Let $\lambda_1 = \cdots = \lambda_j < \lambda_{j+1} < \cdots < \lambda_n$, where $1 < j < n$. Then $V(q) = V_j(q^1,\ldots,q^j) + V_{j+1}(q^{j+1}) + \cdots + V_n(q^n)$ which means that variables $q^{j+1},\ldots,q^n$ separate off while the potential $V_j(q^1,\ldots,q^j)$ has to be tested again on $\boldsymbol{R}^j$ with the use of eqns [21]. Degenerations for $\alpha \neq 0$ or $\beta \neq 0$ are more complicated and the cyclic Bertrand–Darboux equations [22] have to be used. They unfold the level of spherical degeneracy of spheres and embedded subspheres. A complete analysis of all possible degenerations is technical. It requires considering of all possible degenerations of the sequences $\lambda_1 < \cdots < \lambda_n$ and of the related equations [21]–[22] for the potential $V(q)$. It has been proved by Waksjö and Rauch-Wojciechowski (2003) that there is a one-to-one correspondence between all possible sets of PDEs [21]–[22] characterizing separable potentials and all possible types of Riemannian metrics (in the Kalnins and Miller (1986) classification of all separable coordinates on $\boldsymbol{R}^n$ and $\boldsymbol{S}^n$) so that no completely separable case is missed. The most important is that after maximally n steps separation coordinates are always determined (if they exist) by a sequential use of the Bertrand–Darboux and cyclic Bertrand–Darboux equations [21]–[22].

Separation of Eigenvalues Problems

Eigenvalues problems (in a given domain D) of the form

$$\Delta w(q) + \lambda\rho(q)w(q) = 0, \quad \rho > 0 \qquad [23]$$

(where Δ is the Laplace operator) arise when separating the wave equation $\rho(q)u_{tt} = \Delta u$ and the diffusion equation $\rho(q)u_t = \Delta u$ (Courant and Hilbert 1989). The multiplicative ansatz $u(q,t) = w(q)g(t)$ yields eqn [23] together with $\ddot{g} = \lambda g$ or $\dot{g} = \lambda g$. The problem [23] is also used for solving the inhomogeneous equation $\Delta u = f$ with the zero boundary condition $u|_{\partial D} = 0$. In general, the properties of the eigenvalues λ_i and of the corresponding eigenfunctions w_i of the problem [23] depend on the regularity requirements for w_i and on the boundary conditions at ∂D.

For the zero boundary conditions $w(q)|_{\partial D} = 0$, one seeks a nontrivial ($w \neq 0$) solution having in the region D continuous first- and second-order derivatives. General theorems (Courant and Hilbert 1989) state that for such problems there exists a growing sequence $\{\lambda_i\}_{i=1}^{n}$ of positive eigenvalues λ_i such that $\lambda_i \to \infty$ as i increases, and that there is a related sequence of normalized eigenfunctions $\sqrt{\rho}w_1$, $\sqrt{\rho}w_2, \ldots$ that form a complete weighted-orthogonal (in the sense that $\int_D \rho w_i w_j = \delta_{ij}$, $i, j = 1, 2, \ldots$) system of functions so that every regular initial function $\nu(q)$ with $\nu(q)|_{\partial D} = 0$ may be expanded in terms of the eigenfunctions w_m in an absolutely and uniformly convergent series $\nu(q) = \sum_{m=1}^{\infty} c_m w_m(q)$ with $c_m = \int_D \rho\nu w_m$. This makes it possible to express a solution of the IBVP for the wave or for the diffusion equation with zero boundary conditions:

$$\rho(q)u_{tt} = \Delta u \text{ respectively } \rho(q)u_t = \Delta u$$
$$u(q,t)|_{\partial D} = 0, \quad u(q, t=0) = \nu(q) \qquad [24]$$

as a convergent infinite series $u(q,t) = \sum_{m=1}^{\infty} c_m w_m(q) g_m(t)$, where $g_m(t)$ satisfy $\ddot{g} = \lambda g$ respectively $\dot{g} = \lambda g$. Further determination of properties of the eigenfunctions w_n is possible only in special domains D when the problem [23] can be reduced to one-dimensional eigenvalue problems by separating variables in some suitable coordinates.

Example 10 Consider the spherical domain $r^2 = x^2 + y^2 + z^2 \leq 1$. Equation [23] with $\rho = 1$ attains in the spherical coordinates (r, φ, θ) the form

$$\Delta w + \lambda w \equiv \frac{1}{r^2 \sin\theta}\left(\partial_r(r^2 \sin\theta\, \partial_r w) + \partial_\varphi\left(\frac{1}{\sin\theta}\partial_\varphi w\right) + \partial_\theta(\sin\theta\, \partial_\theta w)\right) + \lambda w = 0$$

The ansatz $w = f(r)Y(\theta, \varphi)$ gives the separated equation

$$\frac{(r^2 f')' + \lambda r^2 f}{f} = -\frac{1}{Y\sin\theta}\left(\partial_\varphi\left(\frac{1}{\sin\theta}\partial_\varphi Y\right) + \partial_\theta(\sin\theta\, \partial_\theta Y)\right)$$

so that its both sides must be equal to a constant κ. Continuity of Y implies that it has to be periodic in φ (with period 2π) and regular at $\theta = 0$, $\theta = \pi$. It can only be satisfied for $\kappa = n(n+1)$. The left-hand side of the above equation yields then $(r^2 f'')' - n(n+1)f + \lambda r^2 f = 0$. Solutions that are regular at $r = 0$ are the Bessel functions $(1/\sqrt{r})J_{n+(1/2)}(\sqrt{\lambda}r)$. The equation for spherical harmonics

$$\frac{1}{Y\sin\theta}\left(\partial_\varphi\left(\frac{1}{\sin\theta}\partial_\varphi Y\right) + \partial_\theta(\sin\theta\, \partial_\theta Y)\right) + n(n+1)Y = 0$$

can be further multiplicatively separated by assuming $Y = \Theta(\theta)\Phi(\varphi)$. The function $P(z = \cos\theta) = \Theta(\theta)$ satisfies then the Legendre equation

$$\left((1-z^2)P'(z)\right)' + \left(n(n+1) - \frac{\sigma}{1-z^2}\right)P(z) = 0$$

$P(z)$ is regular at $z = \pm 1$ only when $\sigma = k^2$, $k = 0, 1, 2, \ldots$. The function $\Phi(\varphi)$ satisfies then $\Phi'' = -k^2\Phi$ with solutions $\Phi_k(\varphi) = a_k \cos(k\varphi) + b_k \sin(k\varphi)$. The full solution of the eigenvalue problem $\Delta w + \lambda w = 0$ has the form of an infinite series

$$w_m(r, \varphi, \theta) = \sum_{n=0}^{\infty} \frac{1}{\sqrt{r}} J_{n+(1/2)}\left(\sqrt{\lambda_{m,n}}r\right)\left[a_{n,0}P(\cos\theta) + \sum_{k=1}^{n}\left(a_{n,k}\cos(k\theta) + b_{n,k}\sin(k\theta)\right) \times P_{n,k}(\cos\theta)\right]$$

where the constants $\lambda_{m,n}$, $m = 1, 2, \ldots$, are determined by the transcendental equation $J_{n+(1/2)}(\sqrt{\lambda}) = 0$ that follows from the boundary condition $u(q,t)|_{\partial D} = 0$.

Almost all BVPs that can be reduced to one-dimensional eigenvalue problems may be considered as a special or limiting case of the Lame problem where the boundary ∂D is given by pieces of confocal quadrics corresponding to some separation coordinates. If $D = \{q(x) \in \boldsymbol{R}^3 : x_i^0 \leq x_i \leq x_i^1,\ i = 1, 2, 3\}$ is a domain defined by parametrizing q with the elliptic coordinates x_i given by [16], then the eigenvalue problem $\Delta w + \lambda w = 0$ splits into three one-dimensional equations of the form

$$\varphi(s)Y''(s) + \tfrac{1}{2}\varphi'(s)Y'(s) + (\lambda s + \mu)Y(s) = 0$$

where $\varphi(s) = 4(s-e_1)(s-e_2)(s-e_3)$ and e_i are parameters of the elliptic coordinates. This is the Lame equation; its solutions define new transcendental functions that depend on the choice of the constants λ, μ.

The approach presented here extends to diverse modifications such as vibrations with forcing term $\Delta w(q) + \lambda w(q) = f(q)$, vibrations of a nonhomogeneousmedium $\Delta w(q) + \lambda\rho(q)w(q) = 0$, the stationary Schrödinger equation $\Delta w(q) + V(q)w(q) = \lambda w(q)$ whenever the functions $\rho(q)$, $f(q)$, $V(q)$ are compatible with the separation coordinates.

Separation equations for the second-order BVP are the source of one-dimensional eigenvalue problems of the Sturm–Liouville type

$$(p(s)u')' - q(s)u + \lambda\rho(s)u = 0$$

with singularities that may occur at the endpoints of the fundamental domain. Majority of orthogonal polynomials and special functions appearing in mathematical physics are solutions of Sturm–Liouville problems.

In the complex domain the study of singularities of Laurent series solutions of the same equations led to development of theory of linear ODEs with singular points of the Fuchs class and the Böcher class.

Constructive Approach to Separability of Liouville Integrable Systems

In the constructive approach to separability, one considers simultaneously all Hamilton–Jacobi equations following from a set of n, functionally independent, commuting integrals $H_1(x,y), \ldots, H_n(x,y)$, $\{H_i, H_j\} = 0$, that define a Liouville integrable system (Sklyanin 1995).

One starts with the separation equations, a set of n decoupled ODEs for the functions $W_i(x_i, \alpha)$ depending on one variable x_i and parametric $\alpha \in \boldsymbol{R}^n$:

$$f_i\left(x_i, y_i = \frac{\partial W_i(x_i,\alpha)}{\partial x_i}; \alpha\right) = 0 \qquad [25]$$

Assume that the dependence on α_i is essential (i.e., that $\det(\partial f_i/\partial\alpha_j) \neq 0$) so that we can resolve eqns [25] w.r.t. α_i so that $\alpha_i = H_i(x,y)$ for some functions H_i. If the functions $W_i(x_i,\alpha)$ solve [25] identically w.r.t. x and α, then the function $W(x,\alpha) = \sum_{i=1}^n W_i(x_i,\alpha)$ is simultaneously an additively separable solution of eqns [25] and of the equations

$$\alpha_i \overset{x,\alpha}{\equiv} H_i\left(x, y = \frac{\partial W(x,\alpha)}{\partial x}\right), \quad i = 1, \ldots, n \qquad [26]$$

since solving [25] w.r.t. α is a purely algebraic operation. We can treat eqns [26] as a set of simultaneously separable (in the canonical variables (x,y)) Hamilton–Jacobi equations related to the Hamiltonians H_i. Assume now that

$$\det\left(\frac{\partial^2 W}{\partial x_i\, \partial\alpha_j}\right) = \det\left(\frac{\partial^2 W_i}{\partial x_i\, \partial\alpha_j}\right) \neq 0$$

i.e. that W is a complete integral for [26]. Then the Hamiltonians $H_i(x,y) = \alpha_i$ Poisson-commute since α_i can be treated as new canonical variables obtained by the canonical transformation $(x,y) \to (\beta,\alpha)$ given by

$$y = \frac{\partial W(x,\alpha)}{\partial x}, \qquad \beta = \frac{\partial W(x,\alpha)}{\partial\alpha}$$

Thus, any solvable w.r.t. α set of separation relations [25] defines a Liouville integrable system.

If we perform a canonical transformation from (x,y) to new variables (q,p), then the new set of commuting Hamiltonians $\widetilde{H}_i(q,p) = H_i(x(q,p), y(q,p))$ is also called separable.

The main problem for any given set of commuting Hamiltonians $\widetilde{H}_i(q,p)$ is to decide if there exists a canonical transformation $(q,p) \to (x,y)$ to the separation variables (x,y) so that the related Hamilton–Jacobi equations [26] are simultaneously separable. An answer to this problem is known for integrable Hamiltonians solvable through the spectral curve method (Sklyanin 1995) and for the whole class of natural Hamiltonians discussed earlier.

This approach brings new, wider perspective to the classical separability mechanism stated in the Stäckel theorem. It contains majority of all known separable Hamiltonian systems. For example, if we specify the separation relations [25] to be affine in α_i,

$$\sum_{k=1}^n f_{ik}(x_i,y_i)\alpha_k = g_i(x_i,y_i), \quad i = 1, \ldots, n \qquad [27]$$

then [27] are called generalized Stäckel separability conditions. To recover the explicit form of Hamiltonians $H_k = \alpha_k$, it is enough to solve relations [27] w.r.t. α_k. It has been proved that the Stäckel Hamiltonians in [27] constitute a quasi-bi-Hamiltonian chain. If we specify further relations [27] by assuming that functions f_{ik} do not depend on y_i and functions g_i are quadratic in y_i, then we obtain the classical Stäckel separability conditions (see Theorem 1)

$$\sum_{k=1}^n f_{ik}(x_i)\alpha_k = \frac{1}{2}g_i(x_i)y_i^2 + h_i(x_i) \qquad [28]$$

that can be solved for α_k yielding

$$\alpha_k(x,y)=\frac{1}{2}\sum_{i=1}^{n}(\Phi^{-1})_{ik}\left(y_i^2+\frac{h_i(x_i)}{g_i(x_i)}\right)$$

that is, the Stäckel Hamiltonians [14] with the Stäckel matrix $\Phi=[\varphi_{ik}]$, where $\varphi_{ik}=f_{ik}(x_i)/g_i(x_i)$. By specifying [28] further, we obtain separation relations

$$x_i^{n-1}\alpha_1+x_i^{n-2}\alpha_2+\cdots+\alpha_n=\tfrac{1}{2}g(x_i)y_i^2+h(x_i)$$

which give the so-called Benenti systems associated with conformal Killing tensors and cofactor pair systems.

Relations [27], with $g_i(x_i,y_i)$ depending exponentially on momenta y, contain several well-known systems such as periodic Toda lattice, the KdV dressing chain, and the Ruijsenaar–Schneider system. Relations with g_i cubic in momenta y yield stationary flows of Boussinesq hierarchy and integrable systems on the loop algebra sl(3).

See also: Boundary-Value Problems for Integrable Equations; Calogero–Moser–Sutherland Systems of Nonrelativistic and Relativistic Type; Elliptic Differential Equations: Linear Theory; Evolution Equations: Linear and Nonlinear; Integrable Systems: Overview; Multi-Hamiltonian Systems; Ordinary Special Functions; Recursion Operators in Classical Mechanics; Toda Lattices.

Further Reading

Benenti S (1997) Intrinsic characterization of the variable separation in the Hamilton–Jacobi equation. *Journal of Mathematical Physics* 38(12): 6578–6602.

Courant R and Hilbert D (1989) *Methods of Mathematical Physics. vol. II. Partial Differential Equations*, *Wiley Classics Library*. A Wiley-Interscience Publication. New York: Wiley.

Eisenhart LP (1934) Separable systems of Stäckel. *Annals of Mathematics* 35(2): 284–305.

Fourier J (1945) *The Analytical Theory of Heat*. New York: G. E. Stechert and Co.

Jacobi CGJ (1884) *Vörlesungen über Dynamik. Herausgegeben A. Clebsch.*, ch. 26. Berlin: Verlag von G. Reimer.

Kalnins EG and Miller W Jr. (1980) Killing tensors and variable separation for Hamilton–Jacobi and Helmholtz equations. *SIAM Journal on Mathematical Analysis* 11(6): 1011–1026.

Kalnins EG and Miller W Jr. (1986) Separation of variables on n-dimensional Riemannian manifolds. I. The n-sphere S^n and Euclidean n-space R^n. *Journal of Mathematical Physics* 27(7): 1721–1736.

Landau LD and Lifshitz EM (1976) *Course of Theoretical Physics. vol. 1. Mechanics*, Third Edition. Translated from the Russian by JB Sykes and JS Bell. Oxford–New York–Toronto, Ont: Pergamon Press.

Levi-Civita T (1904) Sulla integrazione della equazione di Hamilton-Jacobi per separazione di variabili. *Mathematische Annalen* 59: 383–397.

Miller W Jr. (1983) The technique of variable separation for partial differential equations. In: Wolf BK (ed.) *Nonlinear Phenomena* (Oaxtepec, 1982), Lecture Notes in Physics, vol. 189, pp. 184–208. Berlin: Springer.

Sklyanin EK (1995) Separation of variables. New trends. *Progress in Theoretical Physics* 118: 35–60.

Stäckel P (1897) Über die Integration der Hamilton'schen Differentialgleichung mittelst Separation der Variabeln. *Math. Ann.* 49(1): 145–147.

Waksjö C and Rauch-Wojciechowski S (2003) How to find separation coordinates for the Hamilton–Jacobi equation: a criterion of separability for natural Hamiltonian systems. *Mathematical Physics, Analysis and Geometry* 6(4): 301–348.

Wojciechowski S (1986) Review of the recent results on integrability of natural Hamiltonian systems. In: Winternitz P (ed.) *Systèmes dynamiques non linéaires: intégrabilité et comportement qualitatif*, Sém. Math. Sup, vol. 102, pp. 294–327. Montreal, QC: Presses Univ. Montréal.

Separatrix Splitting

D Treschev, Moscow State University, Moscow, Russia

Separatrices are asymptotic manifolds in dynamical systems. However, this term is applied usually in the case of a small dimension of the phase space, where these manifolds are hypersurfaces. In the context of separatrix splitting manifolds asymptotic to hyperbolic tori are usually considered, where tori of dimension 0 and 1 are called equilibrium positions and periodic trajectories, respectively. A separatrix can be stable (asymptotic as $t\to+\infty$) and unstable (asymptotic as $t\to-\infty$).

In this article we consider the case of systems with finite-dimensional phase space. Basically we deal with nonautonomous Hamiltonian systems 2π-periodic in time. However, it is useful to keep in mind the fact that the cases of autonomous Hamiltonian systems and symplectic maps are dynamically the same. Some results for non-Hamiltonian perturbations will also be presented. Hamiltonian systems with one-and-a-half or two degrees of freedom as well as area-preserving two-dimensional maps are especially important for us because the results on the separatrix splitting in this case are more clear and complete. Dynamics in such systems is essentially the same. Below we call these systems two dimensional.

We assume that all systems are at least C^∞-smooth.

Poincaré Integral

Consider a Liouville integrable Hamiltonian system. Then any separatrix either goes to infinity or joins two hyperbolic tori. From a dynamical point of view, the latter case is more interesting. If these tori are different, the situation is called heteroclinic, otherwise homoclinic. Poincaré was the first to notice that after a generic perturbation stable and unstable separatrices become different submanifolds of the phase space. This phenomenon is called the separatrix splitting.

Poincaré (1987) considered perturbations of separatrices homoclinic to a periodic solution in a Hamiltonian system with one-and-a-half degrees of freedom. In this case the system has the form

$$\dot{x} = \frac{\partial H}{\partial y}, \qquad \dot{y} = -\frac{\partial H}{\partial x}, \quad (x, y) \in D \subset \boldsymbol{R}^2 \qquad [1]$$

where D is an open domain and

$$H(x, y, t, \varepsilon) = H_0(x, y) + \varepsilon H_1(x, y, t) + O(\varepsilon^2) \qquad [2]$$

We assume that H is 2π-periodic in t and ε is a small parameter. Let $(x_0, y_0) \in D$ be an equilibrium position for the unperturbed $(\varepsilon = 0)$ system: $\operatorname{grad} H_0(x_0, y_0) = 0$. Without loss of generality, $(x_0, y_0) = 0$. In the extended phase space $D \times \boldsymbol{T}$ ($\boldsymbol{T} = \{t \bmod 2\pi\}$ is a one-dimensional torus) instead of the equilibrium we have a 2π-periodic solution $0 \times \boldsymbol{T}$. Suppose that the equilibrium (and therefore, the periodic solution) is hyperbolic and the corresponding stable and unstable separatrices $\Lambda^{s,u}$ are doubled: $\Lambda^s = \Lambda^u = \Lambda$. Let $\gamma(t)$ be a natural parametrization of Λ, that is, $(x(t), y(t)) = \gamma(t)$ is a solution of eqns [1]. In the extended phase space, we have the asymptotic surface

$$(\gamma(t + \tau), t), \quad t \in \boldsymbol{T}, \ \tau \in \boldsymbol{R}$$

For small values of ε, the perturbed system has a hyperbolic periodic solution $(\sigma_\varepsilon(t), t), \sigma_\varepsilon(t) = O(\varepsilon) \in D$ and the separatrices

$$(\gamma_\varepsilon^{s,u}(t, \tau), t), \quad \gamma_0^{s,u}(t, \tau) = \gamma(t + \tau)$$

Since the addition to the Hamiltonian of a function, depending only on t and ε, does not change the dynamics, without loss of generality we can assume that $H_1(0, 0, t) \equiv 0$. Hence the Poincaré integral

$$\mathcal{P}(\tau) = \int_{-\infty}^{+\infty} H_1(\gamma(t + \tau), t)\, \mathrm{d}t$$

converges. The function $\mathcal{P}$ carries all information on the separatrix splitting in the first approximation in ε.

Periodicity of H_1 in t implies 2π-periodicity of $\mathcal{P}(\tau)$. There is also the following obvious identity:

$$\frac{\mathrm{d}\mathcal{P}(\tau)}{\mathrm{d}\tau} = \int_{-\infty}^{+\infty} \{H_0, H_1\}(\gamma(t + \tau), t)\, \mathrm{d}t$$

where $\{,\}$ is the Poisson bracket.

Melnikov Integral

Melnikov (1963) considered general (not necessarily Hamiltonian) 2π-periodic in t perturbations:

$$\dot{x} = \frac{\partial H_0}{\partial y} + \varepsilon v_1(x, y, t) + O(\varepsilon^2)$$
$$\dot{y} = -\frac{\partial H_0}{\partial x} + \varepsilon v_2(x, y, t) + O(\varepsilon^2)$$

In this case, information on the separatrix splitting in the first approximation is contained in the Melnikov integral

$$\mathcal{M}(\tau) = \int_{-\infty}^{+\infty} v H_0(\gamma(t + \tau), t)\, \mathrm{d}t$$

where $v H_0 = v_1 \partial H_0/\partial x + v_2 \partial H_0/\partial y$.

Note that if the vector field v is Hamiltonian and H_1 is the corresponding Hamiltonian function, we have: $v H_0 = -\{H_0, H_1\}$. Hence in Hamiltonian systems we have: $\mathcal{M}(\tau) = -\mathrm{d}\mathcal{P}(\tau)/\mathrm{d}\tau$.

A multidimensional version of the Melnikov integral is presented in Wiggins (1988).

Geometric Meaning of $\mathcal{M}(\tau)$

Let Γ_T be a compact piece of the unperturbed separatrix

$$\Gamma_T = \{(x, y) \in D\colon (x, y) = \gamma(t), \ |t| \le T\}$$

Then for any $T > 0$ there exists a neighborhood U of Γ_T and symplectic coordinates (time–energy coordinates) τ, h on U such that the section of the perturbed separatrices $\Lambda_\varepsilon^{s,u}$ by the plane $\{t = 0\}$ is as follows:

$$\Lambda_\varepsilon^{s,u}|_{t=0} = \{(\tau, h) : h = h_\varepsilon^{s,u}(\tau)\}$$

where

1. $h_\varepsilon^u(\tau) = O(\varepsilon^2)$,
2. $h_\varepsilon^s(\tau) = -\varepsilon \mathcal{M}(\tau) + O(\varepsilon^2)$.

   Moreover, let $g_\varepsilon^t : D \to D$ be the phase flow of the perturbed system. The map $g_\varepsilon^{2\pi}$ is called the Poincaré map. The following statement holds.
3. For any two points $z_0, z_1 \in U$ such that $z_1 = g_\varepsilon^{2\pi}(z_0)$, let (τ_0, h_0) and (τ_1, h_1) be their time–energy coordinates. Then

$$\tau_1 = \tau_0 + 2\pi + O(\varepsilon), \qquad h_1 = h_0 + O(\varepsilon)$$

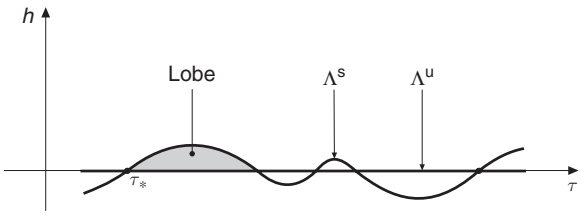


Figure 1 Perturbed separatrices in time–energy coordinates.

Existence of such coordinates has several corollaries.

- If $\mathcal{P}$ is not identically constant, the separatrices split and this splitting is of the first order in ε.
- Let τ_* be a simple zero of $\mathcal{M}$. Then the perturbed separatrices intersect transversally at a point $z_*(\varepsilon)$ with time–energy coordinates $(\tau_* + O(\varepsilon), O(\varepsilon^2), t = 0)$. Such a point $z_*(\varepsilon)$ is called a transversal homoclinic point. It generates a doubly asymptotic solution in the perturbed system.
- Consider a lobe domain $\mathcal{L}(\tau_*, \varepsilon)$ bounded by two segments of separatrices on the section $\{t = 0\}$ (see Figure 1). Let another "corner point" of the lobe $\mathcal{L}(\tau_*, \varepsilon)$ correspond to the simple zero τ_*' of $\mathcal{M}$. Then the symplectic area of $\mathcal{L}(\tau_*, \varepsilon)$ equals

$$\mathcal{A}_{\mathcal{L}}(\tau_*, \varepsilon) = -\varepsilon \int_{\tau_*}^{\tau_*'} \mathcal{M}(\tau)\, d\tau + O(\varepsilon^2)$$

A Standard Example

Consider as an example a pendulum with periodically oscillating suspension point. The Hamiltonian of the system can be presented in the form

$$H(x, y, t, \varepsilon) = \tfrac{1}{2}y^2 + \Omega^2 \cos x + \varepsilon\theta(t)\cos x \qquad [3]$$

where Ω is the "internal" frequency of the pendulum. The function θ is 2π-periodic in time. Hence the frequency of the suspension point oscillation equals 1. In this case, the unperturbed homoclinic solution $\gamma(t)$ can be computed explicitly. In particular,

$$\cos(x(t)) = 1 - 2\cosh^{-2}(\Omega t)$$

Therefore, $\mathcal{P}(\tau) = \int_{-\infty}^{+\infty} \theta(t)(\cos(x(t+\tau)) - 1)\, dt$. For example, if $\theta(t) = \cos t$, we have

$$\mathcal{P}(\tau) = -\frac{2\pi\cos\tau}{\Omega^2 \sinh(\pi/2\Omega)}$$

In this case, different lobes have the same area

$$\mathcal{A}_{\mathcal{L}} = \frac{4\varepsilon\pi}{\Omega^2 \sinh(\pi/2\Omega)} + O(\varepsilon^2)$$

Multidimensional Case

Multidimensional generalization of the Poincaré–Melnikov construction is strongly connected to the concept of a (partially) hyperbolic torus. Let (M, ω, H) be a Hamiltonian system on the $2m$-dimensional symplectic manifold (M, ω).

An invariant n-torus $N \subset M$ $(0 \le n < m)$ is called hyperbolic if there exist coordinates x, y, z on M in a neighborhood of N such that

1. $y = (y_1, \ldots, y_n)$, $x = (x_1, \ldots, x_n) \bmod 2\pi$, $z = (z^s, z^u)$, $z^{s,u} = (z_1^{s,u}, \ldots, z_l^{s,u})$, $l + n = m$;
2. $\omega = dy \wedge dx + dz^u \wedge dz^s$;
3. $N = \{(x, y, z) : y = 0,\ z = 0\}$; and
4. $H = \langle \nu, y\rangle + (1/2)\langle Ay, y\rangle + \langle z^u, \Omega(x) z^s\rangle + O_3(y, z)$,

where $\nu \in \boldsymbol{R}^n$ is a constant vector, A is a constant $n \times n$ matrix, Ω is an $l \times l$ matrix such that $\Omega(x) + \Omega^T(x)$ is positive definite for any $x \bmod 2\pi$, the symbol O_3 denotes terms of order not less than 3, and $\langle a, b\rangle = \sum a_j b_j$.

If $\det A \neq 0$, the torus is called nondegenerate. If ν is Diophantine, that is, for some $\alpha, \beta > 0$ and any $0 \neq k \in \boldsymbol{Z}^n$

$$|\langle \nu, k\rangle| \ge \alpha |k|^{-\beta}$$

the torus N is called Diophantine. The coordinates (x, y, z) are called canonical for N.

Now suppose that the Hamiltonian H depends smoothly on the parameter ε:

$$H = H_0 + \varepsilon H_1 + O(\varepsilon^2)$$

and for $\varepsilon = 0$ the system is Liouville integrable with the commuting first integrals $F_1, \ldots, F_m$:

$$\{F_j, F_k\} = 0, \quad 1 \le j, k \le m$$

Let $M_0 = \{F_1 = \cdots = F_m = 0\} \subset M$ be their zero common level and let $N \subset M$ be an n-dimensional nondegenerate Diophantine hyperbolic torus. The torus N generates the invariant Lagrangian asymptotic manifolds $\Lambda^{s,u} \subset M$. Suppose that the separatrices are doubled, that is, there is a Lagrangian manifold $\Lambda \subset \Lambda^s \cap \Lambda^u$.

Consider the perturbed Hamiltonian $H = H_0 + \varepsilon H_1 + O(\varepsilon^2)$. The torus N as well as the asymptotic manifolds $\Lambda^{s,u}$ survive the perturbation. Let N_ε be the corresponding hyperbolic torus in the perturbed system and $\Lambda_\varepsilon^{s,u}$ its asymptotic manifolds: N_ε and $\Lambda_\varepsilon^{s,u}$ depend smoothly on ε and $N_0 = N$, $\Lambda_0^{s,u} = \Lambda^{s,u}$.

Let the function $\chi(x)$ satisfy the equation

$$\langle \nu, \partial\chi(x)/\partial x\rangle + H_1(x, 0, 0) = \frac{1}{(2\pi)^n} \int_{\boldsymbol{T}^n} H_1(x, 0, 0)\, dx$$

This equation has a smooth solution unique up to an additive constant.

Consider a solution of the unperturbed Hamiltonian equations $\gamma(t) \subset \Lambda$. Let $I_j^\gamma, I_{j,l}^\gamma, 1 \le j,l \le m$ be the following quantities (Treschev 1994):

$$I_j^\gamma = \lim_{T\to+\infty}\left(-\int_{-T}^{T}\{F_j,H_1\}(\gamma(t))\,dt + \{F_j,\chi\}(\gamma(-T)) - \{F_j,\chi\}(\gamma(T))\right)$$

$$I_{j,l}^\gamma = \lim_{T\to+\infty}\left(-\int_{-T}^{T}\{F_j,\{F_l,H_1\}\}(\gamma(t))\,dt + \{F_j\{F_l,\chi\}\}(\gamma(-T)) - \{F_j\{F_l,\chi\}\}(\gamma(T))\right)$$

The numbers $I_j^\gamma, I_{j,l}^\gamma$ play the role of the first and second derivatives of the Poincaré integral at some point.

If any of the quantities $I_j^\gamma, I_{j,l}^\gamma$ does not vanish, the asymptotic manifolds $\Lambda^{s,u}$ split. Moreover, suppose that $I_1^\gamma = \cdots = I_m^\gamma = 0$ and the rank of the matrix $(I_{j,l}^\gamma)$ equals $m-1$. Then for small values of ε, the manifolds Λ^s and Λ^u intersect transversally on the energy level at points of the solution $\gamma_\varepsilon(t)$, where $\gamma_\varepsilon \to \gamma$ as $\varepsilon \to 0$.

Poincaré Integral in Multidimensional Case

Suppose that the Hamiltonian from the previous section equals

$$H(x,y,u,v,t,\varepsilon) = H_0(y,u,v) + \varepsilon H_1(x,y,u,v,t) + O(\varepsilon^2)$$

Here $x = (x_1,\ldots,x_n) \bmod 2\pi$, $y = (y_1,\ldots,y_n) \in \boldsymbol{R}^n$, and $(u,v) \in \boldsymbol{R}^2$. The symplectic structure $\omega = dy \wedge dx + dv \wedge du$.

We assume that in the unperturbed integrable system the variables separate:

$$H_0(y,u,v) = F(y) + f(u,v)$$

and the system with one degree of freedom and Hamiltonian f has a hyperbolic equilibrium $(u,v) = 0$ with a homoclinic solution $\gamma(t)$. Any torus

$$N(y^0) = \{(x,y,u,v,t)\colon y = y^0,\ u = v = 0\}$$

is a hyperbolic torus of the unperturbed system with frequency vector

$$\begin{pmatrix}\nu(y^0)\\ 1\end{pmatrix}, \quad \nu(y) = \partial F/\partial y$$

Suppose that $N = N(0)$ is Diophantine and nondegenerate. Then in the perturbed system there is smooth in ε hyperbolic torus N_ε, $N_0 = N$. Consider the Poincaré function

$$\mathcal{P}(\xi,\tau) = \int_{-\infty}^{+\infty}\Big(H_1(\xi+\nu(t+\tau),0,\gamma(t+\tau),t) - H_1(\xi+\nu(t+\tau),0,0,0,t)\Big)dt$$

Obviously, $\mathcal{P}(\xi,\tau)$ is 2π-periodic in ξ and τ.

If $\mathcal{P}$ is not identically constant, asymptotic surfaces of N_ε split in the first approximation in ε. Nondegenerate critical points of $\mathcal{P}$ correspond to transversal homoclinic solutions of the perturbed system.

Other results on the splitting of multidimensional asymptotic manifolds are presented in Arnol'd *et al.* (1988) and Lochak *et al.* (2003).

Exponentially Small Separatrix Splitting

If in the unperturbed (integrable) system there are no asymptotic manifolds, they can appear after a perturbation. Consider, for example, perturbation of a real-analytic Liouville integrable system near a simple resonance:

$$\dot{x} = \frac{\partial H}{\partial y}, \quad \dot{y} = \frac{\partial H}{\partial x}, \quad x \in \boldsymbol{T}^m, \quad y \in D \subset \boldsymbol{R}^m$$
$$H(x,y,t,\varepsilon) = H_0(y) + \varepsilon H_1(x,y,t,\varepsilon)$$

As usual, we assume 2π-periodicity in t. A simple resonance corresponds to a value of the action variable $y = y^0$ such that the frequency vector

$$\hat{\nu} = \begin{pmatrix}\nu^0\\ 1\end{pmatrix}, \quad \nu^0 = \frac{\partial H_0}{\partial y}(y^0) \in \boldsymbol{R}^m$$

(here 1 is the frequency, corresponding to the time variable) admits only one resonance. More precisely, there exists a nonzero $\hat{k} \in \boldsymbol{Z}^{m+1}$, satisfying $\langle \hat{k}, \hat{\nu}\rangle = 0$ and any $k \in \boldsymbol{Z}^{m+1}$ such that $\langle k, \hat{\nu}\rangle = 0$ is collinear with $\hat{k}$.

Without loss of generality, we can assume that $y^0 = 0$ and $\nu^0 = \binom{0}{\widetilde{\nu}}$, $\widetilde{\nu} \in \boldsymbol{R}^{m-1}$. Then the vector $\overline{\nu} = \binom{\widetilde{\nu}}{1} \in \boldsymbol{R}^m$ is nonresonant.

In a $\sqrt{\varepsilon}$-neighborhood of the resonance we have a system with fast variables $X = (x_2,\ldots,x_m,t) \bmod 2\pi$ and slow variables $Y = (x_1, \varepsilon^{-1/2}y_1, \ldots, \varepsilon^{-1/2}y_m)$ variables:

$$\dot{Y} = O(\sqrt{\varepsilon}), \qquad \dot{X} = \overline{\nu} + O(\sqrt{\varepsilon}) \qquad [4]$$

If the frequency vector $\overline{\nu}$ is Diophantine, by using the Neishtadt averaging procedure, we can reduce the dependence of the right-hand sides of eqns [4] on

the fast variables to exponentially small in ε terms. This means that there exist new symplectic variables

$$P = Y + O(\sqrt{\varepsilon}), \quad Q = X + O(\sqrt{\varepsilon})$$

(new time coincides with the old one) such that system [4] takes the form

$$\dot{P} = \sqrt{\varepsilon}F(P, \sqrt{\varepsilon}) + O\big(\exp(-a\varepsilon^{-b})\big)$$
$$\dot{Q} = \overline{\nu} + \sqrt{\varepsilon}G(P, \sqrt{\varepsilon}) + O\big(\exp(-a\varepsilon^{-b})\big)$$

with positive constants a, b.

If we neglect the exponentially small reminders, the system turns out to be integrable. Generically, it has a family of hyperbolic m-tori of the form $\{(P, Q)\colon P = \text{const.}\}$ with doubled asymptotic manifolds. However, the terms $O(\exp(-a\varepsilon^{-b}))$ generically cannot be removed completely. They produce an exponentially small splitting of the asymptotic manifolds. This splitting implies nonintegrability, chaotic behavior, Arnol'd diffusion, and other dynamical effects.

It is important to note that exponentially small splitting appears only in the analytic case. In smooth systems the splitting is much stronger.

Unfortunately, at present there are no quantitative methods for studying such splittings except obvious upper estimates and the case of two-dimensional systems.

Exponentially Small Splitting in Two-Dimensional Systems

The main results on exponentially small separatrix splitting were obtained by Lasutkin and his students (Gelfreich and others). Another effective approach was proposed by Treschev. There are no general theorems in this situation; however, many examples were studied. We discuss the splitting in the pendulum with rapidly oscillating suspension point. The Hamiltonian of the system has the form

$$H = \tfrac{1}{2}y^2 + (1 + 2b\cos(t/\varepsilon))\cos x$$

(cf. [3]). For any value of ε the circle $\{(x, y, t)\colon x = y = 0\}$ is a periodic trajectory. For small $\varepsilon > 0$ the trajectory is hyperbolic.

Poincaré integral can be formally written in this system. It predicts the area of lobes $16\pi b\varepsilon^{-1}\,\mathrm{e}^{-\pi(2\varepsilon)^{-1}}$. However, there is no reason to expect that this asymptotics of the splitting is correct. Indeed, its value is exponentially small in ε, while the error of the Poincaré–Melnikov method is in general quadratic in the perturbation. To obtain correct asymptotics of the separatrix splitting, one has to study singularities of the solutions with respect to complex time. Area of lobes in this system equals (Treschev 1997)

$$\mathcal{A}_{\mathcal{L}} = 4bf(b, \varepsilon)\varepsilon^{-1}\,\mathrm{e}^{-\pi(2\varepsilon)^{-1}}$$

Here $f(b, \varepsilon), \varepsilon \geq 0$ is a smooth function. The function $f(b, 0)$ is even and entire. It can be computed numerically as a solution of a problem which does not contain ε. The value $f(0, 0) = 4\pi$ corresponds to the Poincaré integral, but the function $f(b, 0)$ is not constant. It is possible to prove that f can be expanded in a power series in ε. Apparently, this series diverges for any $b \neq 0$.

Separatrix Splitting and Dynamics

1. Separatrix splitting can be regarded as an obstacle to the integrability of the perturbed system. However, this statement needs some comments. Doubled asymptotic surfaces in an integrable Hamiltonian system can have self-intersections. In the case of equilibrium, such intersections can even be transversal. In the literature, there is no general result saying that separatrix splitting implies nonintegrability. Some particular cases (studied by Kozlov, Ziglin, Bolotin, and others) are presented in Arnol'd *et al.* (1989). For example, in the two-dimensional case, this is seen to be true.
2. Conceptual reason for the nonintegrability, discussed in the previous item, is a complicated dynamics near the splitted separatrices. In many situations, it is possible to find in this domain a Smale horseshoe. This implies positive topological entropy, existence of nontrivial hyperbolic sets, symbolic dynamics, etc.
3. Consider a near-integrable area-preserving two-dimensional map. In the perturbed system in the vicinity of the splitted separatrices of a hyperbolic fixed point z_ε the so-called stochastic layer is formed. Here we mean the domain bounded by invariant curves, closest to the separatrices. An important quantity, describing the rate of chaos, is the area of the stochastic layer $\mathcal{A}_{\mathrm{SL}}(\varepsilon)$. It turns out (Treschev 1998b) that $\mathcal{A}_{\mathrm{SL}}(\varepsilon)$ is connected with the area of the largest lobe $\mathcal{A}_{\mathcal{L}}(\varepsilon)$ by the simple formula

$$c_1\mathcal{A}_{\mathrm{SL}}(\varepsilon) < \frac{\mathcal{A}_{\mathcal{L}}(\varepsilon)\log(\mathcal{A}_{\mathcal{L}}(\varepsilon))}{\log^2\mu} < c_2\mathcal{A}_{\mathrm{SL}}(\varepsilon)$$

   with some constants $c_1, c_2 > 0$, where μ is the largest multiplier (Lyapunov exponent) of the fixed point z_0.
4. Let $\hat{z}$ be a hyperbolic fixed point of an area-preserving two-dimensional map. The point $\hat{z}$

divides the corresponding separatrices $\Lambda^{s,u}$ in 4 branches $\Lambda^{s}_{1,2}$ and $\Lambda^{u}_{1,2}$. Suppose that the pair of branches Λ^{s}_{j} and Λ^{u}_{l} satisfies the following conditions:

- Λ^{s}_{j} and Λ^{u}_{l} lie in a compact invariant domain;
- Λ^{s}_{j} and Λ^{u}_{l} do not coincide and intersect at a homoclinic point.

Then the closures $\overline{\Lambda}^{s}_{j}, \overline{\Lambda}^{u}_{l}$ are compact invariant sets. Very little is known about these sets. For example, it is not known if their measure is positive. However, by using the Poincaré recurrence theorem, it is possible to prove (Treschev 1998a) that $\overline{\Lambda}^{s}_{j} = \overline{\Lambda}^{u}_{l}$.

See also: Averaging Methods; Billiards in Bounded Convex Domains; Hamiltonian Systems: Obstructions to Integrability; Hamiltonian Systems: Stability and Instability Theory.

Further Reading

Arnol'd V, Kozlov V, and Neishtadt A (1988) Mathematical aspects of classical and celestial mechanics. In: *Encyclopaedia of Mathematical Sciences*, vol. 3. Berlin: Springer.

Lochak P, Marco J-P, and Sauzin D (2003) On the splitting of invariant manifolds in multidimensional near-integrable Hamiltonian systems. *Memoirs of the American Mathematical Society* 163(775): viii+145.

Melnikov V (1963) On the stability of the center for time-periodic perturbations. *Trudy Moskovskogo Metern. Obschestva* 12: 3–52 (Russian). (English transl.: Trans. *Moscow Math. Soc.* 1963, pp. 1–56 (1965).)

Poincaré H (1987) *Les méthodes nouvelles de la Mécanique Céleste*, vols. 1–3. Paris: Gauthier–Villars, (Original publication: 1892, 1893, 1899). New Printing: Librairie Scientifique et Technique Albert Blanchard 9, Rue Medecin 75006, Paris.

Treschev D (1994) Hyperbolic tori and asymptotic surfaces in Hamiltonian systems. *Russian Journal of Mathematical Physics* 2(1): 93–110.

Treschev D (1997) Separatrix splitting for a pendulum with rapidly oscillating suspension point. *Russian Journal of Mathematical Physics* 5(1): 63–98.

Treschev D (1998a) Closures of asymptotic curves in a two-dimensional symplectic map. *J. Dynam. Control Systems* 4(3): 305–314.

Treschev D (1998b) Width of stochastic layers in near-integrable two-dimensional symplectic maps. *Physica D* 116(1–2): 21–43.

Wiggins S (1988) *Global Bifurcations and Chaos. Analytical Methods. Applied Mathematical Sciences*, vol. 73, 494pp New York: Springer.

Several Complex Variables: Basic Geometric Theory

A Huckleberry, Ruhr-Universität Bochum, Bochum, Germany
T Peternell, Universität Bayreuth, Bayreuth, Germany

Introduction

The rubric "several complex variables" is attached to a wide area of mathematics which involves the study of holomorphic phenomena in dimensions higher than one. In this area there are viewpoints, methods and results which range from those on the analytic side, where analytic techniques of partial differential equations (PDEs) are involved, to those of algebraic geometry which pertain to varieties defined over finite fields. Here we outline selected basic methods which are aimed at understanding global geometric phenomena. Detailed presentations of most results discussed here can be found in the basic texts (Demailly, Grauert and Fritzsche 2001, Griffiths and Harris 1978, Grauert *et al.* 1994, Grauert and Remmert 1979, 1984).

Domains in $\mathbb{C}^n$

Complex analysis begins with the study of holomorphic functions on domains D in $\mathbb{C}^n$. These are smooth complex-valued functions f which satisfy

$$\bar{\partial} f := \sum \frac{\partial f}{\partial \bar{z}_i} \mathrm{d}\bar{z}_i = 0$$

Some results from the one-dimensional theory extend to the case where $n > 1$. However, even at the early stages of development, one sees that there are many new phenomena in the higher-dimensional setting.

Extending Results from the One-Dimensional Theory

For local results one may restrict considerations to functions f which are holomorphic in a neighborhood of $0 \in \mathbb{C}^n$. The restriction of f to, for example, any complex line through 0 is holomorphic, and therefore the maximum principle can be immediately transferred to the higher-dimensional setting.

The zero-set $V(f)$ of a nonconstant holomorphic function is one-codimensional over the complex numbers (two-codimensional over the reals). Thus the identity principle must be formulated in a different way from its one-dimensional version. For example, under the usual connectivity assumptions, if f vanishes on a set E with Hausdorff dimension bigger than $2n-2$, then it vanishes identically. Here

is another useful version: if M is a real submanifold such that the real tangent space T_zM generates the full complex tangent space at one of its points, that is, $T_zM + iT_zM = T_z\mathbb{C}^n$, and $f|M \equiv 0$, then $f \equiv 0$.

In the one-dimensional theory, after choosing appropriate holomorphic coordinates, $f(z) = z^k$ for some k. This local normal form implies that nonconstant holomorphic functions are open mappings. Positive results in the mapping theory of several complex variables are discussed below. The simple example $F: \mathbb{C}^2 \to \mathbb{C}^2, (z, w) \to (z, zw)$, shows that the open mapping theorem cannot be transferred without further assumptions.

The local normal-form theorem in several complex variables is called the "Weierstraß preparation theorem." It states that after appropriate normalization of the coordinates, f is locally the product of a nonvanishing holomorphic function with a "polynomial"

$$P(z, z') = z^k + a_{k-1}(z')z^{k-1} + \cdots + a_0(z')$$

where z is a single complex variable, z' denotes the remaining $n-1$ variables, and the coefficients are holomorphic in z'. This is a strong inductive device for the local theory.

If D is a product $D = D_1 \times \cdots \times D_n$ of relatively compact domains in the complex plane $\mathbb{C}$, then repeated integration transfers the one-variable Cauchy integral formula from the D_i to D. The resulting integral is over the product $\mathrm{bd}(D_1) \times \cdots \times \mathrm{bd}(D_n)$ of the boundaries which is topologically a small set in $\mathrm{bd}(D)$. Complex analytically it is, however, large in the sense of the above identity principle.

It follows from, for example, the n-variable Cauchy integral formula that holomorphic functions agree with their convergent power series developments. As in the one-variable theory, the appropriate topology on the space $\mathcal{O}(D)$ of holomorphic functions on D is that of uniform convergence on compact subsets. In this way $\mathcal{O}(D)$ is equipped with the topology of a Fréchet space.

First Theorems on Analytic Continuation

Analytic continuation is a fundamental phenomenon in complex geometry. One type of continuation theorem which is known in the one-variable theory is of the following type: If E is a small closed set in D and $f \in \mathcal{O}(D \backslash E)$ is a holomorphic function which satisfies some growth condition near E, then it extends holomorphically to D. The notion "small" can be discussed in terms of measure, but it is more appropriate to discuss it in complex analytic terms.

An analytic subset A of D is locally the common zero set $\{a \in D; f_1(a) = \cdots = f_m(a) = 0\}$ of finitely many holomorphic functions. A function g on A is said to be holomorphic if at each $a \in A$ it is the restriction of a holomorphic function on some neighborhood of a in D. There is an appropriate notion of an irreducible component of A. If A is irreducible, it contains a dense open set A_{reg}, which is a connected k-dimensional complex manifold, that is, at each of its points a there are functions $f_1, \ldots, f_k$ which define a map $F := (f_1, \ldots, f_k)$, which is a holomorphic diffeomorphism of A_{reg} onto an open set in $\mathbb{C}^k$. The boundary A_{sing} is the set of singular points of A, which is a lower-dimensional analytic set. The dimension of an analytic set is the maximum of the dimensions of its irreducible components.

Here are typical examples of theorems on continuing holomorphic functions across small analytic sets E. If $\mathrm{codim}\, E \geq 2$, then every function which is holomorphic on $D \backslash E$ extends to a holomorphic function on D. The same is true of meromorphic functions, that is, functions which are locally defined as quotients $m = f/g$ of holomorphic functions. If f is holomorphic on D, then $g := 1/f$ is holomorphic outside the analytic set $E := V(f)$. Thus g cannot be holomorphically continued across this one-codimensional set. However, Riemann's Hebbarkeitssatz is valid in several complex variables: if f is locally bounded outside an analytic subset E of any positive codimension, then it extends holomorphically to D.

With a bit of care, continuation results of this type can be proved for (reduced) complex spaces. These are defined as paracompact Hausdorff spaces which possess charts $(U_\alpha, \varphi_\alpha)$, where the local homeomorphism φ_α identifies the open set U_α with a closed analytic subset A_α of a domain D_α in some $\mathbb{C}^{n_\alpha}$. As indicated above, a continuous function on A_α is holomorphic if at each point it can be holomorphically extended to some neighborhood of that point in D_α. Finally, just as in the case of manifolds, the compatibility between charts is guaranteed by requiring that coordinate change $\varphi_{\alpha\beta}: U_{\alpha\beta} \to U_{\beta\alpha}$ is biholomorphic, that is, it is a homeomorphism so that it and its inverse are given by holomorphic functions as $F = (f_1, \ldots, f_m)$. The discussion of irreducible components, sets of singularities, and dimension for complex spaces goes exactly in the same way as that above for analytic sets.

If E is everywhere at least two-codimensional, then the above result on continuation of meromorphic functions holds in complete generality. The Hebbarkeitssatz requires the additional condition that the complex space is normal. In many situations this causes no problem at all, because, in general, there is a canonically defined associated normal

complex space $\tilde{X}$ and a proper, surjective, finite-fibered holomorphic map $\tilde{X} \to X$ which is biholomorphic outside a nowhere-dense proper analytic subset. Difficulties can be overcome by simply lifting functions to this normalization and applying the Hebbarkeitssatz.

Continuation theorems of Hartogs-type reflect the fact that complex analysis in dimensions larger than one is really quite different from the one-variable version. The following is such a theorem. Let (z, w) be the standard coordinates in $\mathbb{C}^2$ and think of the z-axis as a parameter space for geometric figures in the w-plane. For example, let $D_z := \{(z, w) : |w| < 1\}$ be a disk and $A_z := \{(z, w) : 1 - \varepsilon < |w| < 1\}$ be an annulus. An example of a Hartogs figure H in $\mathbb{C}^2$ is the union of the family of disks D_z for $|z| < 1 - \delta$ with the family A_z of annuli for $1 - \delta \leq |z| < 1$. One should visualize the moving disks which suddenly change to moving annuli. One speaks of filling in the Hartogs figure to obtain the polydisk $\hat{H} := \{(z, w); |z|, |w| < 1\}$. Hartogs' continuation theorem states that a function which is holomorphic on H extends holomorphically to $\hat{H}$.

Cartan–Thullen Theorem

One of the major developments in complex analysis in several variables was the realization that certain convexity concepts lie behind the strong continuation properties. At the analytic level one such is defined as follows by the full algebra of holomorphic functions $\mathcal{O}(D)$. If K is a compact subset of D, then its holomorphic convex hull $\hat{K}$ is defined as the intersection of the sets $P(f) := \{p \in D : |f(p)| \leq |f|_K\}$ as f runs through $\mathcal{O}(D)$. One says that D is holomorphically convex if $\hat{K}$ is compact for every compact subset K of D.

The theorem of H. Cartan and Thullen relates this concept to analytic continuation phenomena as follows. A domain D is said to be a domain of holomorphy if, given a divergent sequence $\{z_n\} \subset D$, there exists $f \in \mathcal{O}(D)$ which is unbounded along it. In other words, the phenomenon of being able to extend all holomorphic functions on D to a truly larger domain $\hat{D}$ does not occur. The Cartan–Thullen theorem states that D is a domain of holomorphy if and only if it is holomorphically convex. In the next paragraph the relation between this type of convexity and a certain complex geometric convexity of the boundary $\mathrm{bd}(D)$ will be indicated.

Levi Theorem and the Levi Problem

Consider a smooth (local) real hypersurface Σ containing $0 \in \mathbb{C}^n$ with $n > 1$. It is the zero-set $\{\rho = 0\}$ in some neighborhood U of 0 of a smooth function with $\mathrm{d}\rho \neq 0$ on U. This is viewed as a piece of a boundary of a domain D, where $U \cap D = \{\rho < 0\}$. The real tangent space $T_0\Sigma = \mathrm{Ker}(\mathrm{d}\rho(0))$ contains a unique maximal (one-codimensional) complex subspace $T_0^{\mathbb{C}}\Sigma = \mathrm{Ker}(\partial\rho(0)) = H$. The signature of the restriction of the complex Hessian (or Levi form) $i\partial\bar{\partial}\rho$ to H is a biholomorphic invariant of Σ. In this notation the Hessian is a real alternating 2-form which is compatible with the complex structure, and its signature is defined to be the signature of the associated symmetric form.

If the restriction of this Levi form to the complex tangent space has a negative eigenvalue, that is, if the boundary $\mathrm{bd}(D)$ has a certain degree of concavity, then there is a map $F : \Delta \to U$ of the unit disk Δ which is biholomorphic onto its image with $F(0) = 0$ and otherwise $F(\mathrm{cl}(\Delta)) \subset D$. The reader can imagine pushing the image of this map into the domain to obtain a family of disks which are in the domain, and pushing it in the outward pointing direction to obtain annuli which are also in the domain. Making this precise, one builds a (higher-dimensional) Hartogs figure H at the base point 0 so that $\hat{H}$ is an open neighborhood of 0. In particular this proves the theorem of E. E. Levi: *every function holomorphic on $U \cap D$ extends to a neighborhood of* 0. This can be globally formulated as follows:

Theorem *If D is a domain of holomorphy with smooth boundary in $\mathbb{C}^n$, then* $\mathrm{bd}(D)$ *is Levi-pseudoconvex.*

Here the terminology Levi-pseudoconvex is used to denote the condition that the restriction of the Levi form to the complex tangent space of every boundary point is positive semidefinite.

One of the guiding problems of complex analysis in higher dimensions is the Levi problem. This is the converse statement to that of the Levi's theorem:

Levi Problem *Is a domain D with smooth Levi-pseudoconvex boundary in a complex manifold necessarily a domain of holomorphy?*

Stated in this form it is not true, but for domains in $\mathbb{C}^n$ it is true. As will be sketched below, under stronger assumptions on the Levi form it is almost true. However, there are still interesting open problems in complex analysis which are related to the Levi problem.

Bounded Domains and Their Automorphisms

The unit disk in the complex plane is particularly important, because, with the exceptions of projective

space $\mathbb{P}_1(\mathbb{C})$, the complex plane $\mathbb{C}$, the punctured plane $\mathbb{C}\backslash\{0\}$, and compact complex tori, it is the universal cover of every (connected) one-dimensional complex manifold.

In higher dimensions it should first be underlined that, without some further condition, there is no best bounded domain in $\mathbb{C}^n$. For example, two randomly chosen small perturbations of the unit ball $\mathbb{B}_2 := \{(z,w); |z|^2 + |w|^2 < 1\}$, with, for example, real analytic boundary, are not biholomorphically equivalent.

On the other hand, the following theorem of H. Cartan shows that bounded domains D are good candidates for covering spaces:

Theorem *Equipped with the compact open topology, the group* $\mathrm{Aut}(D)$ *of holomorphic automorphisms of* D *is a Lie group acting properly on* D.

The notion of a proper group action of a topological group on a topological space is fundamental and should be underlined. It means that if $\{x_n\}$ is a convergent sequence in the space where the group is acting, then a sequence of group elements $\{g_n\}$, with the property that $\{g_n(x_n)\}$ is convergent, itself possesses a convergent subsequence. As a consequence, isotropy groups are compact and orbits are closed.

In the context of bounded domains D this implies that if Γ is a discrete subgroup of $\mathrm{Aut}(D)$, then $X = D/\Gamma$ carries a natural structure of a complex space. If in addition Γ is acting freely, something that, with minor modifications, can be arranged, then X is a complex manifold.

Many nontrivial compact complex manifolds arise as quotients D/Γ of bounded domains. Even very concrete quotients, for example, where $D = \mathbb{B}_2$, are extremely interesting. Conversely, if $\mathrm{Aut}(D)$ contains a discrete subgroup Γ so that D/Γ is compact, then D is probably very special. For example, it is known to be holomorphically convex!

Any compact quotient $X = D/\Gamma$ of a bounded domain is projective algebraic in the sense that it can be realized as a complex (algebraic) submanifold of some complex projective space. In fact the embedding can be given by quite special Γ-invariant holomorphic tensors on D, and this in turn implies that X is of general type (see below). For further details, in particular on Cartan's theorem on the automorphism group of a bounded domain, the reader is referred to Narasimhan (1971).

Stein Manifolds

The founding fathers of the first phase of "modern complex analysis" (Cartan, Oka, and Thullen) realized that domains of holomorphy form the basic class of spaces where it would be possible to solve the important problems of the subject concerning the existence of holomorphic or meromorphic functions with reasonably prescribed properties. In fact, Oka formulated a principle which more or less states that if a complex analytic problem which is well formulated on a domain of holomorphy has a continuous solution, then it should have a holomorphic solution. Given the flexibility of continuous functions and the rigidity of holomorphic functions, this would seem impossible but in fact is true!

Beginning in the late 1930s, Stein worked on problems related to this Oka principle, in particular on those related to what we would now call the algebraic topological aspects of the subject, and he was led to formulate conditions on a general complex manifold X which should hold if problems of the above type are to be solved. First, his axiom of holomorphic convexity was simply that, given a divergent sequence $\{x_n\}$ in X, there should be a function $f \in \mathcal{O}(X)$ such that $\{f(x_n)\}$ is unbounded. Secondly, holomorphic functions should separate points in the sense that, given distinct points $x_1, x_2 \in X$, there exists $f \in \mathcal{O}(X)$ with $f(x_1) \neq f(x_2)$. Finally, globally defined holomorphic functions should give local coordinates. Assuming that X is n-dimensional, this means that, given a point $x \in X$, there exist $f_1, \ldots, f_n \in \mathcal{O}(X)$ such that $\mathrm{d}f_1(x) \wedge \cdots \wedge \mathrm{d}f_n(x) \neq 0$.

Assuming Stein's axioms, Cartan and Serre then produced a powerful theory in the context of sheaf cohomology which proved certain vanishing theorems that led to the desired existence theorems. This theory and typical applications are sketched below. Before going into this, we would like to mention that Grauert's version of the Cartan–Serre theory requires only very weak versions of Stein's axioms: (1) The connected component containing K of the holomorphic convex hull $\hat{K}$ of every compact set should be compact. (2) Given $x \in X$, there are functions $f_1, \ldots, f_m \in \mathcal{O}(X)$ so that x is an isolated point in the fiber of the map $F := (f_1, \ldots, f_m) : X \to \mathbb{C}^m$. Of course the results also hold for complex spaces.

Holomorphically convex domains in $\mathbb{C}^n$ are Stein manifolds, and since closed complex manifolds of Stein manifolds are Stein, it follows that any complex submanifold of $\mathbb{C}^n$ is Stein. In particular, affine varieties are Stein spaces. Remmert's theorem states the converse: an n-dimensional Stein manifold can be embedded as a closed complex submanifold of $\mathbb{C}^{2n+1}$. A nontrivial result of Behnke and Stein implies that every noncompact Riemann surface is also Stein.

Basic Formalism

The following first Cousin problem is typical of those which can be solved by Stein theory. Let X be a complex manifold which is covered by open sets U_i. Suppose that on each such set a meromorphic function m_i is given so that on the overlap $U_{ij} := U_i \cap U_j$ the difference $m_{ij} = m_j - m_i =: f_{ij}$ is holomorphic. This means that the distribution of polar parts of these functions is well defined. The question is whether or not there exists a globally defined meromorphic function $m \in \mathcal{M}(X)$ with these prescribed polar parts, that is, with $m - m_i =: f_i \in \mathcal{O}(U_i)$.

If one applies the Oka principle, this problem can be easily solved. For this one can assume that the covering is locally finite and take γ_i to be a partition of unity subordinate to the cover. Using standard shrinking and cut-off arguments, one extends the f_{ij} to the full space X as smooth functions so that the alternating cocycle relations $f_{ij} + f_{jk} + f_{ki} = 0$ and $f_{ij} = -f_{ji}$ still hold. Then $f_j := \sum \gamma_k f_{jk}$ is a smooth function on U_j which satisfies $f_j - f_i = f_{ij}$ on the overlap U_{ij}. It follows that $f := m_i + f_i = m_j + f_j$ is a globally well-defined "smooth" function with the prescribed polar parts. The Oka principle would then imply that there is a globally defined meromorphic function with the same property.

The basic sheaf cohomological formalism for Stein theory can be seen in the above argument. Suppose that instead of applying extension and cut-off techniques from the smooth category, we could answer positively the question "given the holomorphic functions $\{f_{ij}\}$ on the U_{ij}, do there exist holomorphic functions $\{f_i\}$ on the U_i such that $f_j - f_i = f_{ij}$ on the U_{ij}?" Then we would immediately have the desired globally defined meromorphic function $m := m_i + f_i$. This question is exactly the question of whether or not the Cech cohomology class of the alternating cocycle $\{f_{ij}\}$ vanishes.

Let us quickly summarize the language of Cech cohomology. A presheaf of abelian groups is a mapping $U \to S(U)$ which associates to every open subset of X an abelian group. Typical examples are $U \to \mathcal{O}(U)$, $U \to C^\infty(U)$, $U \to H^*(U, \mathbb{Z}), \ldots$. The last example which associates to U its topological cohomology does not localize well in terms of following the basic axioms for a sheaf: (1) Given a covering $\{U_i\}$ of an open subset U of X and elements $s_i \in S(U_i)$ with $s_j - s_i = 0$ on U_{ij}, there exists $s \in S(U)$ with $s|U_i = s_i$. (2) If $s, t \in S(U)$ are such that $s|U_i = t|U_i$ for all i, then $s = t$. For this we have assumed that the restriction mappings have been built into the definition of a presheaf.

Associated to a sheaf S on X and a covering $\mathcal{U} = \{U_i\}$ is the space of alternating q-cocycles $\mathcal{C}^q(U, S)$, which is the set of alternating maps ξ from the set of $(q+1)$-fold indices of the form $(i_0, \ldots, i_q) \mapsto s_{i_0,\ldots,i_q} \in S(U_{i_0,\ldots,i_q})$. Here $U_{i_0,\ldots,i_q} := U_{i_0} \cap \cdots \cap U_{i_q}$. The boundary mapping $\delta : \mathcal{C}^q \to \mathcal{C}^{q+1}$ is defined by $\delta(\xi)_{i_0,\ldots i_{q+1}} = \sum_k (-1)^k s_{i_0,\ldots,i_{k-1},i_{k+1},\ldots,i_{q+1}}$. It follows that $\delta^2 = 0$, and $H^*(\mathcal{U}, S)$ is defined to be the cohomology of the associated complex.

In any consideration it is necessary to refine coverings, shrink, etc., and therefore one goes to the limit $H^*(X, \mathcal{S})$ over all refinements of the coverings. The script notation $\mathcal{S}$ is used to denote that we have then localized the sheaf to the germ level. Due to a theorem of Leray one can, however, always take a suitable covering so that $H^q(\mathcal{U}, \mathcal{S}) = H^q(X, \mathcal{S})$ for all q, where now $\mathcal{S}(U)$ satisfies the above axioms.

One of the important facts in this cohomology theory is that a short exact sequence of sheaves $0 \to \mathcal{S}' \to \mathcal{S} \to \mathcal{S}'' \to 0$ yields a long exact sequence

$$0 \to H^0(X, \mathcal{S}') \to H^0(X, \mathcal{S}) \to H^0(X, \mathcal{S}'') \to H^1(X, \mathcal{S}') \to H^1(X, \mathcal{S}) \to H^1(X, \mathcal{S}'') \to \cdots$$

in cohomology.

A fundamental theorem of Stein theory, Theorem B, states that for the basic analytic sheaves $\mathcal{S}$ of complex analysis, the so-called coherent sheaves, all cohomology spaces $H^q(X, \mathcal{S})$ vanish for all $q \geq 1$. In the above example of the first Cousin problem the desired vanishing is that of $H^1(\mathcal{U}, \mathcal{O})$.

Coherent Sheaves

Numerous important sheaves in complex analysis are associated to vector bundles on complex manifolds. A holomorphic vector bundle $\pi : E \to X$ over a complex manifold is a holomorphic surjective maximal rank fibration. Every fiber $E_x := \pi^{-1}(x)$ is a complex vector space, and the vector space structure is defined holomorphically over X. For example, addition is a holomorphic map $E \times_X E \to E$. Such bundles are locally trivial, that is, there is a covering $\{U_i\}$ of the base such that $\pi^{-1}(U_i)$ is isomorphic to $U_i \times \mathbb{C}^r$ and on the overlap the gluing maps in the fibers are holomorphic maps $\varphi_{ij} : U_{ij} \to \mathrm{GL}_n(\mathbb{C})$. The number r is called the rank of the bundle. Holomorphic bundles of rank 1 are referred to as holomorphic line bundles. Of course all of these definitions make sense in other categories, for example, topological, smooth, real analytic, etc.

A holomorphic section of E over an open set U is a holomorphic map $s : U \to E$ with $\pi \circ s = \mathrm{Id}_U$. The space of these sections is denoted by $\mathcal{E}(U)$, and the map $U \to \mathcal{E}(U)$ defines a sheaf which is locally just $\mathcal{O}_X^r$. It is therefore called a locally free sheaf of

$\mathcal{O}$-modules. Conversely, by taking bases of a locally free sheaf $\mathcal{S}$ on the open sets where it is isomorphic to a direct sum $\mathcal{O}^r$, one builds an associated holomorphic vector bundle E so that $\mathcal{E} = \mathcal{S}$.

It is not possible to restrict our attention to these locally free sheaves or equivalently to holomorphic vector bundles. One important reason is that images of holomorphic vector bundle maps are not necessarily vector bundles. A related reason is that the sheaf of ideals of holomorphic functions which vanish on a given analytic set A is not always a vector bundle. This is caused by the presence of singularities in A. There are many other reasons, but these should suffice for this sketch.

The sheaves $\mathcal{S}$ that arise naturally in complex analysis are almost vector bundles. If X is the base complex manifold or complex space under consideration, then $\mathcal{S}$ will come from a vector bundle on some big open subset X_0 whose boundary is an analytic set X_1, and then on the irreducible components of X_1 it will come from vector bundles on such big open sets, etc. These sheaves are called coherent analytic sheaves of $\mathcal{O}_X$-modules. The correct algebraic definition is that locally there exists an exact sequence

$$0 \to \mathcal{O}^{p_d} \to \cdots \to \mathcal{O}^{p_1} \to \mathcal{O}^{p_0} \to \mathcal{S} \to 0 \qquad [1]$$

of sheaves of $\mathcal{O}$-modules. This implies in particular that, although $\mathcal{S}$ might not be locally free, it is locally finitely generated, and the relations among the generators are also finitely generated.

Selected Theorems

The following efficiently formulated fundamental theorem contains a great deal of information about Stein manifolds.

Theorem B *A complex space X is Stein if and only if for every coherent sheaf $\mathcal{S}$ of $\mathcal{O}_X$-modules $H^q(X, \mathcal{S}) = 0$ for all $q \geq 1$.*

Since $\mathcal{S}$ is a sheaf, it follows that $H^0(X, \mathcal{S}) = \mathcal{S}(X)$. This is referred to as the space of sections of $\mathcal{S}$ over X. As a result of Theorem B, we are able to construct sections with prescribed properties. Let us give two concrete applications (there are many more!).

Example Let A be a closed analytic subset of a Stein space X, and let $\mathcal{I}$ denote the subsheaf of $\mathcal{O}_X$ which consists of those functions which vanish on A. Note that this must be defined for every open subset U of X. Then we have the short exact sequence $0 \to \mathcal{I} \to \mathcal{O}_X \to \mathcal{O}_X/\mathcal{I} \to 0$. The restriction of $\mathcal{O}_X/\mathcal{I}$ to A is called the (reduced) structure sheaf $\mathcal{O}_A$ of A. In other words, for U open in A the space $\mathcal{O}_A(U)$ should be regarded as the space of holomorphic functions on U.

Now, $\mathcal{I}$ is a coherent sheaf on X and therefore by Theorem B the cohomology group $H^1(X, \mathcal{I})$ vanishes. Consequently, the associated long exact sequence in cohomology implies that the restriction mapping $\mathcal{O}_X(X) \to \mathcal{O}_A(A)$ is surjective. This special case of Theorem A means that every (global!) holomorphic function on A is the restriction of a holomorphic function on X. ◇

Example Let us consider the multiplicative (second) Cousin problem. In this case meromorphic functions m_i are given on the open subsets U_i of a covering $\mathcal{U}$ with the property that $m_i = f_{ij} m_j$, where f_{ij} is holomorphic and nowhere vanishing on the overlap U_{ij}. This is a distribution D of the zero and polar parts of meromorphic functions, which in complex geometry is called a divisor, and the interesting question is whether or not there exists a globally defined meromorphic function which has D as its divisor.

Now we note that $\mathrm{GL}_1(\mathbb{C}) = \mathbb{C}^*$ and thus $f_{ij} : U_{ij} \to \mathbb{C}^*$ defines a line bundle L on X and we regard it as an element of the space $H^1(X, \mathcal{O}^*)$ of equivalence classes of line bundles on X. Here $\mathcal{O}^*$ is the sheaf of nowhere-vanishing holomorphic functions on X. It is not even a sheaf of $\mathcal{O}$-modules; therefore coherence is not discussed in this case.

The long exact sequence in cohomology associated to the short exact sequence $0 \to \mathbb{Z} \longrightarrow \mathcal{O} \xrightarrow{\exp} \mathcal{O}^* \to 1$ yields an element $c_1(L) \in H^2(X, \mathbb{Z})$, which is a purely topological invariant. It is called the Chern class of L, and one knows that L is topologically trivial if and only if $c_1(L) = 0$.

Coming back to the Cousin II problem, using the same argument as in the Cousin I problem, we can solve it if and only if we can find nowhere-vanishing functions $f_i \in \mathcal{O}^*(U_i)$ with $f_i = f_{ij} f_j$. This is equivalent to finding a nowhere-vanishing section of L. But a line bundle has a nowhere-vanishing section if and only if it is isomorphic to the trivial bundle. In other words, the Cousin II problem can be solved for a given divisor D if and only if the associated line bundle $L(D)$ is trivial in $H^1(X, \mathcal{O}^*)$. For this, a necessary condition is that the Chern class $c_1(L(D))$ vanishes. But if X is Stein, this is also sufficient, because the vanishing of $H^1(X, \mathcal{O})$ together with the long exact sequence in cohomology shows that $H^1(X, \mathcal{O}^*) \xrightarrow{c_1} H^2(X, \mathbb{Z})$ is injective.

Hence, in this case we have the following precise formulation of the Oka principle: "A given divisor D on a Stein manifold is the divisor of a globally defined meromorphic function if and only if the associated line bundle is topologically trivial." ◇

A slightly refined statement from that above is the fact that on a Stein manifold the space of topological line bundles is the same as the space of holomorphic line bundles. In the case of (higher rank) vector bundles this is a deep and important theorem of Grauert. It can be formulated as follows.

Grauert's Oka principle On a Stein space the map $F: \mathrm{Vect}_{\mathrm{holo}}(X) \to \mathrm{Vect}_{\mathrm{top}}(X)$ from the space of holomorphic vector bundles to the space of topological vector bundles which forgets the complex structure is bijective.

In closing this section, a few words concerning the proofs of the major theorems, for example, Theorem B, should be mentioned. In all cases one must solve something like an additive Cousin problem and one first does this on special relatively compact subsets. For this step there are at least two different ways to proceed. One is to delicately piece together solutions which are known to exist on very special polyhedral-type domains or build up from lower-dimensional pieces of such.

Another method is to solve certain systems of PDEs on relatively compact domains where control at the boundary is given by the positivity of the Levi-form. An example of how such PDEs occur can already be seen at the level of the above Cousin I problem. At the point where we have solved it topologically, that is, the holomorphic cocycle $\{f_{ij}\}$ is a coboundary $f_{ij} = f_j - f_i$ of smooth functions, we observe that since $\bar\partial f_{ij} = 0$, it follows that $\alpha = \bar\partial f_i$ is a globally defined $(0,1)$-form. It is $\bar\partial$-closed, that is, the compatibility condition for solving the system $\bar\partial u = \alpha$ is fulfilled. If this system can be solved, then we use the solution u to adjust the topological solutions of the Cousin problem by replacing f_i by $f_i - u$. We still have $f_{ij} = f_j - f_i$, but now the f_i are holomorphic on U_i.

To obtain the global solution to a Cousin-type problem, one exhausts the Stein space by the special relatively compact subsets U_n where, by one method or another, we have solved the problem with solutions s_n. One would like to say that the s_n converge to a global solution s. However, there is no way to *a priori* guarantee this without making some sort of estimates. One main way of handling this problem is to adjust the solutions as $n \to \infty$ by an approximation procedure. For this one needs to know that holomorphic objects, for example, functions on U_n, can be approximated on U_n by objects of the same type which are defined on the bigger set U_{n+1}. This Runge-type theorem, which is a nontrivial ingredient in the whole theory, requires the introduction of an appropriate Fréchet structure on the spaces of sections of a coherent sheaf. This is in itself a point that needs some attention.

Montel's Theorem and Fredholm Mappings

If U is an open subset of a complex space X, then $\mathcal{O}(U)$ has the Fréchet topology of convergence on compact subsets K defined by the seminorms $|\cdot|_K$. Using resolutions of type (1) above, one shows that the space of sections $\mathcal{S}(U)$ of every coherent sheaf $\mathcal{S}$ also possesses a canonical Fréchet topology. This is then extended to the spaces $\mathcal{C}^q(\mathcal{U}, \mathcal{S})$, and consequently one is able to equip the cohomology spaces $H^q(X, \mathcal{S})$ with (often non-Hausdorff) quotient topology.

Elements of such cohomology groups can be regarded as obstructions to solving complex analytic problems. One often expects such obstructions, and is satisfied whenever it can be shown if there are only finitely many, that is, a finiteness theorem of the type $\dim H^q(X, \mathcal{S}) < \infty$ is desirable. Here we sketch two finiteness theorems which hold in seemingly different contexts, but their proofs are based on one principle: use the compactness guaranteed by Montel's theorem as the necessary input for the Fredholm theorem in the context of Fréchet spaces.

Recall that a continuous linear map $T: E \to F$ between topological vector spaces is said to be compact if there is an open neighborhood U of $0 \in E$ such that $T(U)$ is relatively compact in F. If Y is a relatively compact open subset of a complex space X, then Montel's theorem states that the restriction map $r_Y^X: \mathcal{O}(X) \to \mathcal{O}(Y)$ is compact. This can be extended to coherent sheaves, and using the Fredholm theorem for certain natural restriction and boundary maps, one proves the following fundamental fact.

Lemma 1 *If the restriction map $r_Y^X: H^q(X, \mathcal{S}) \to H^q(Y, \mathcal{S})$ is surjective, then $H^q(Y, \mathcal{S})$ is finite dimensional.*

Since the methods for the proof are basic in complex analysis, we outline it here. Take a covering $\tilde{\mathcal{U}}$ of X such that $H^q(\mathcal{U}, \mathcal{S}) = H^q(X, \mathcal{S})$. Then intersect its elements with Y to obtain a covering $\tilde{\mathcal{U}}$ of Y. Finally, refine that covering with refinement mapping τ to a covering $\mathcal{V}$ of Y such that $H^q(\mathcal{V}, \mathcal{S}) = H^q(X, \mathcal{S})$ and so that U_i contains $V_{\tau(i)}$ as a relatively compact subset for all i. Let $Z^q(\mathcal{U}, \mathcal{S})$ denote the kernel of the boundary map δ for the covering $\mathcal{U}$, and consider the map $Z^q(\mathcal{U}, \mathcal{S}) \oplus \mathcal{C}^{q-1}(\mathcal{V}, \mathcal{S}) \to \mathcal{C}^q(\mathcal{V}, \mathcal{S})$ which is the direct sum $\tau \oplus \delta$ of the restriction and boundary maps. By assumption it is surjective. Since δ is the difference of this map and the compact map τ, L Schwartz's version of the Fredholm theorem for Fréchet spaces implies that its image is of finite

codimension, that is, $H^q(Y,\mathcal{S}) = H^q(\mathcal{V},\mathcal{S})$ is finite dimensional.

Applying this Lemma in the case of compact spaces where $X = Y$, one has the following theorem of Cartan and Serre:

Theorem *If X is a compact complex space and $\mathcal{S}$ is a coherent sheaf on X, then* $\dim H^q(X,\mathcal{S}) < \infty$ *for all q.*

Grauert made use of this technique in solving the Levi problem for a strongly pseudoconvex relatively compact domain D with smooth boundary in a complex manifold X. Here strongly pseudoconvex means that the restriction of the Levi form to the complex tangent space of every boundary point is positive definite. To do this he sequentially made "bumps" at boundary points to obtain a finite sequence of domains $D = D_0 \subset D_1 \subset \cdots \subset D_m$ in such a way that the restriction mappings at the level of qth cohomology, $q \geq 1$, are all surjective and such that at the last step D is relatively compact in D_m. Applying the above Lemma, $\dim H^q(D,\mathcal{S}) < \infty$. Using another bumping procedure, it then follows that D is holomorphically convex and, in fact, that D is almost Stein.

This last statement means that one can guarantee that $\mathcal{O}(D)$ separates points outside of some compact subset which could contain compact subvarietes on which the global holomorphic functions are constant. In this situation one can apply Remmert's reduction theorem which implies that there is a canonically defined proper surjective holomorphic map $\pi: D \to Z$ to a Stein space which is biholomorphic outside of finitely many fibers. One says that, in order to obtain the Stein space Z, finitely many compact analytic subsets must be blown down to points.

The above mentioned reduction theorem is a general result which applies to any holomorphically convex complex space X. For this one observes that if X is holomorphically convex, then for $x \in X$ the level set $L(x) := \{y \in X;\ f(y) = f(x) \text{ for all } f \in \mathcal{O}(X)\}$ is a compact analytic subset of X. One then defines an equivalence relation: $x \sim y$ if and only if the connected component of $L(x)$ containing x and that of $L(y)$ which contains y are the same. One then equips $X/\sim$ with the quotient topology and proves that the canonical quotient $\pi: X \to X/\sim =: Z$ is proper. Finally, for U open in Z one defines $\mathcal{O}_Z(U) = \mathcal{O}_X(\pi^{-1}(U))$ and proves that, equipped with this structure, Z is a Stein space. This Remmert reduction is universal with respect to holomorphic maps to holomorphically separable complex spaces, that is, if $\varphi: X \to Y$ and $\mathcal{O}_Y(Y)$ separates the points of Y, then there exists a uniquely defined holomorphic map $\tau_\varphi: Z \to Y$ so that $\tau_\varphi \circ \pi = \varphi$. It should be noted that, even if the original space X is a complex manifold, the associated Stein space Z may be singular. This reflects the fact that it is difficult to avoid singularities in complex geometry.

Mapping Theory

Above we have attempted to make it clear that holomorphic maps play a central role in complex geometry. It is even important to regard a holomorphic function as a map. Here we outline the basic background necessary for dealing with maps and then state three basic theorems which involve proper holomorphic mappings.

Basic Facts

A holomorphic map $F: X \to Y$ between (reduced) complex spaces is a continuous map which can be represented locally as a holomorphic map between analytic subsets of the spaces in which X and Y are locally embedded. In other words, F is the restriction of a map $F = (f_1, \ldots, f_m)$ which is defined by holomorphic functions.

If X is irreducible and X and Y are one-dimensional, then a nonconstant holomorphic map $F: X \to Y$ is an open mapping. This statement is far from being true in the higher-dimensional setting. The reader need only consider the example $F: \mathbb{C}^2 \to \mathbb{C}^2, (z,w) \to (zw, z)$.

Despite the fact that holomorphic maps can be quite complicated, they have properties that in certain respects render them tenable. Let us sketch these in the case where X is irreducible. First, one notes that every fiber $F^{-1}(y)$ is a closed analytic subset of X. One defines $\operatorname{rank}_x F$ to be the codimension at x of the fiber $F^{-1}(F(x))$ at x. Then $\operatorname{rank} F := \max\{\operatorname{rank}_x F;\ x \in X\}$. It then can be shown that $\{x \in X;\ \operatorname{rank}_x F \leq k\}$ is a closed analytic subset of X for every k. Applying this for $k = \operatorname{rank} F - 1$ we see that, outside a proper closed analytic subset, F has constant maximal rank.

If $F: X \to Y$ has constant rank k in a neighborhood of some point $x \in X$, then one can choose neighborhoods U of x in X and V of $F(x)$ in Y so that $F|U$ maps U onto a closed analytic subset of Y. By restricting F to the sets where it has lower rank and applying this local-image theorem, it follows that the local images of the set where F has lower rank are at least two dimensions smaller than those of top rank. Conversely, the fiber dimension $d_F(x) := \dim_x F^{-1}(F(x))$ is semicontinuous in the sense that $d_F(x) \geq d_F(z)$ for all z near x. Finally, we note that if Y is m-dimensional, then $F: X \to Y$ is an open map if and only if it is of constant rank m.

Proper Mappings

By definition a mapping $F: X \to Y$ between topological spaces is proper if and only if the inverse image $F^{-1}(K)$ of an arbitrary compact subset in Y is compact in X. This is a more delicate condition than meets the eye. For example, if $F: X \to Y$ is a proper map and one removes one point from some fiber, then it is normally no longer proper! On the other hand, the restriction of a proper map to a closed subset is still proper.

Remmert's "Proper mapping theorem" is the first basic theorem on proper holomorphic maps:

Theorem *The image of a proper holomorphic map $F: X \to Y$ is a closed analytic subset of Y.*

Given another basic theorem of complex analysis, the reader can imagine how this might be proved. This is the continuation theorem for analytic sets due to Remmert and Stein:

> *If X is a complex space and Y is a closed analytic subset with $\dim_y Y \leq k$ for all $y \in Y$ and Z is a closed analytic subset of the complement $X \setminus Y$ with $\dim_z Z \geq k+1$ at all $z \in Z$, then the topological closure $\mathrm{cl}(Z)$ of Z in X is a closed analytic subset of X with $E = \mathrm{cl}(Z) \setminus Z = \mathrm{cl}(Z) \cap Y$ a proper analytic subset of $\mathrm{cl}(Z)$.*

Similar results hold for more general complex analytic objects. For example, closed positive currents with (locally) finite volume can be continued across any proper analytic subset (Skoda 1982). A sketch of the proof of the proper mapping theorem (for X irreducible) goes as follows. From the assumption that F is proper, the image $F(X)$ is closed. If F has constant rank k, then, by the local result stated above, its image is everywhere locally a k-dimensional analytic set. Since the image is closed, the desired result follows. If rank $F = k$ and $E := \{x \in X;\ \mathrm{rank}_x F < k\} \neq \emptyset$, then by induction $F(E)$ is a closed analytic subset of dimension at most $k-2$. Let $A := F^{-1}(E)$ and apply the previous discussion for constant rank maps to $F|(X \setminus A): X \setminus A \to Y \setminus E$. The image is a closed k-dimensional analytic subset of $Y \setminus E$ and its Remmert–Stein extension is the full image $F(X)$.

In this framework the Stein factorization theorem is an important tool. Here $F: X \to Y$ is again a proper holomorphic map which we may now assume to be surjective. Analogous to the construction of the reduction of a holomorphically convex space, one says that two points in X are equivalent if they are in the same connected component of an F-fiber. This is indeed an equivalence relation, and the quotient $Z := X/\sim$ is a complex space equipped with the direct image sheaf. Thus one decomposes F into two maps $X \to Z \to Y$, where $X \to Z$ is a canonically associated surjective map with connected fiber, and $Z \to Y$ is a finite map.

This geometric proper mapping theorem is a preview of one of the deepest results in complex analysis: Grauert's direct image theorem. This concerns the images of sheaves, not just the images of points. For this, given a sheaf $\mathcal{S}$ on X one defines the qth direct image sheaf on Y as the sheaf associated to the presheaf which attaches to an open set U in Y the cohomology space $H^q(F^{-1}(U), \mathcal{S})$. Grauert's "Bildgarbensatz" states the following: "If $F: X \to Y$ is a proper holomorphic map, then all direct image sheaves of any coherent sheaf on X are coherent on Y."

Complex Analysis and Algebraic Geometry

The interplay between these subjects has motivated research and produced deep results on both sides. Here we indicate just a few results of the type which show that objects which are *a priori* of an analytic nature are in fact algebraic geometric.

Projective Varieties

Let us begin with the algebraic geometric side of the picture where we consider algebraic subvarieties X of projective space $\mathbb{P}_n(\mathbb{C})$. If $[z_0 : z_1 : \cdots : z_n]$ are homogeneous coordinates of $\mathbb{P}_n$, such a variety is the simultaneous zero-set, $X := V(P_1, \ldots, P_m)$, of finitely many (holomorphic) homogeneous polynomials $P_i = P_i(z_0, \ldots, z_m)$. Chow's theorem states that in this context there are no further analytic phenomena:

Theorem *Closed complex analytic subsets of projective space $\mathbb{P}_n(\mathbb{C})$ are algebraic subvarieties.*

This observation has numerous consequences. For example, if $F: X \to Y$ is a holomorphic map between algebraic varieties, then, by applying Chow's theorem to its graph, it follows that F is algebraic.

Chow's theorem can be proved via an application of the Remmert–Stein theorem in a very simple situation. For this, let $\pi: \mathbb{C}^{n+1} \setminus \{0\} \to \mathbb{P}_n(\mathbb{C})$ be the standard projection, and let $Z := \pi^{-1}(X)$. Since Z is positive dimensional, by the Remmert–Stein theorem it can be extended to an analytic subset of $\mathbb{C}^{n+1}$. The resulting subvariety $K(X)$ (the cone over X) is invariant by the $\mathbb{C}^*$-action which is defined by $v \to \lambda v$ for $\lambda \in \mathbb{C}^*$. If f is a holomorphic function on $\mathbb{C}^{n+1}$ which vanishes on $K(X)$, then we develop it in homogeneous polynomials $f = \sum P_d$ and note that $\lambda(f)(z) = f(\lambda z) = \sum \lambda^d P_d$ also vanishes for all λ. Hence, all P_d vanish identically and therefore the ideal of holomorphic functions which vanish on $K(X)$

is generated by the homogeneous polynomials which vanish on $K(X)$ and consequently finitely many of these define X as a subvariety of $\mathbb{P}_n(\mathbb{C})$.

Complements of subvarieties in projective varieties occur in numerous applications and are important objects in complex geometry. Even complements $\mathbb{P}_n \backslash Y$ of subvarieties Y in the full projective space are not well understood. If Y is the intersection of a compact projective variety X with a projective hyperplane, that is, Y is a hyperplane section, then $X \backslash Y$ is affine. If Y is q-codimensional in X, then $X \backslash Y$ possesses a certain degree of Levi convexity and general theorems of Andreotti and Grauert (1962) on the finiteness and vanishing of cohomology indeed apply. However, not nearly as much is understood in this case as in the case of a hyperplane section.

Kodaira Embedding Theorem

Given that analytic subvarieties of projective space are algebraic, one would like to understand whether a given compact complex manifold or complex space can be realized as such a subvariety. Kodaira's theorem is a prototype of such an embedding theorem. Most often one formulates projective embedding theorems in the language of bundles.

For this, observe that if $L \to X$ is a holomorphic line bundle over a compact complex manifold, then its space $\Gamma(X, L)$ of holomorphic sections is a finite-dimensional vector space V. The zero-set of a section $s \in V$ is a one-codimensional subvariety of X. Let us restrict our attention to bundles which are generated by their sections which for line bundles simply means that for every $x \in X$ there is some section $s \in V$ with $s(x) \neq 0$. It then follows that for every $x \in X$ the space $H_x := \{x \in X;\, s(x) = 0\}$ is a one-codimensional vector subspace of V. Thus L defines a holomorphic map $\varphi_L : X \to \mathbb{P}(V^*), x \mapsto H_x$. Note that we must go to the projective space $\mathbb{P}(V^*)$, because a linear function defining such an H_x is only unique up to a complex multiple.

Projective embedding theorems state that under certain conditions on L the map φ_L is a holomorphic embedding, that is, it is injective and is everywhere of maximal rank in the analytic sense that its differential has maximal rank. Here we outline a complex analytic approach of Grauert for proving embedding theorems. It makes strong use of the complex geometry of bundle spaces.

Let $L \to X$ be a holomorphic line bundle over a compact complex manifold. A Hermitian bundle metric is a smoothly varying metric h in the fibers of L. This defines a norm function $v \mapsto |v|^2 := h(v, v)$ on the bundle space L. One says that L is positive if the tubular neighborhood $T := \{v \in L;\, |v|^3 < 1\}$ is strongly pseudoconcave, that is, when regarded from outside T, its boundary is strongly pseudoconvex.

To prove an embedding theorem, one must produce sections with prescribed properties. Sections of powers L^k are closely related to holomorphic functions on the dual bundle space L^*. This is due to the fact that if $\pi : L \to X$ is the bundle projection, $\pi^{-1}(U_\alpha) \cong U_\alpha \times \mathbb{C}$ is a local trivialization, and z_α is a fiber coordinate, then a holomorphic function f on L^* has a Taylor series development

$$f(v) = \sum s_\alpha(n)(\pi(v)) z_\alpha^n(v)$$

The function f is well defined on L. Hence, the transformation law for the z_α^n must be canceled out by a transformation law for the coefficient functions $s_\alpha(n)$. This implies that the $s_\alpha(n)$ are sections of L^n. Hence, proving the existence of sections in the powers of L with prescribed properties amounts to the same thing as proving the existence of holomorphic funtions on L^* with analogous properties.

The positivity assumption on L is equivalent to assuming that the tubular neighborhoods of the zero-section in L^* defined by the norm function associated to the dual metric are strongly pseudoconvex. The solution to the Levi problem, which was sketched above, then shows that L^* is holomorphically convex, and its Remmert reduction is achieved by simply blowing down its zero-section. In other words, L^* is essentially a Stein manifold, and using Stein theory, it is possible to produce enough holomorphic functions on L to show that some power L^k defines a holomorphic embedding $\varphi_{L^k} : X \to \mathbb{P}(\Gamma(X, L^k)^*)$. Bundles with this property are said to be ample, and thus we have outlined the following fact: "a line bundle which is Grauert-positive is ample."

It should be underlined that we defined the Chern class of L as the image in $H^2(X, \mathbb{Z})$ of its equivalence class in $H^1(X, \mathcal{O}^*)$, that is, in this formulation the Chern class is a Cech cohomology class. It is, however, often more useful to consider it as a deRham class where it lies in the (1, 1)-part of $H^2_{deR}(X, \mathbb{C})$. If h is a bundle metric as above, then the Levi form of the norm function is a representative $-c_1(L, h)$ of the Chern class of L^*. Thus $c_1(L, h)$ is an integral (1, 1)-form which represents $c_1(L)$. It is called the Chern form of L associated to the metric h. The following is Kodaira's formulation of his embedding theorem:

Theorem *A line bundle L is ample if and only if it possesses a metric h so that $c_1(L, h)$ is positive definite.*

Kodaira's proof of this fact follows from his vanishing theorem (*see* Several Complex Variables: Compact Manifolds) in the same way the example of Theorem A was derived from Theorem B in the

first example in the subsection "Selected theorems." That an ample bundle is positive follows immediately from the fact that if φ_{L^k} is an embedding, then its pullback of the (positive) hyperplane bundle on projective space agrees with L^k.

Finally, one asks the question "under what natural conditions can one construct a bundle L which is positive?" The following is an example of an answer which is related to geometric quantization.

Suppose that X is a compact complex manifold equipped with a symplectic structure ω, that is, ω is a d-closed, nondegenerate 2-form. One says that ω is Kählerian if it is compatible with the complex structure J in the sense that $\omega(Jv, Jw) = \omega(v, w)$ and $\omega(Jv, v) > 0$ for every v and w in every tangent space of X. Note that if L is a positive line bundle, then it possesses a Hermitian metric h such that $\omega = c_1(L, h)$ is a Kählerian structure on X.

It should be underlined that there are Kähler manifolds without positive bundles, for example, every compact complex torus $T = \mathbb{C}^n/\Gamma$ possesses the Kählerian structure which comes from the standard linear structure on $\mathbb{C}^n$. However, for $n > 1$ most such tori are not projective algebraic and therefore do not have positive bundles.

If, on the other hand, the Kählerian structure is integral, a condition that is automatic for the Chern form $c_1(L, h)$ of a bundle, then there is indeed a line bundle $L \to X$ equipped with a Hermitian metric h such that $c_1(L, h) = \omega$. The condition of integrality can be formulated in terms of the integrals of ω over homology classes being integral or that its deRham class is in the image of the deRham isomorphism from the Cech cohomology $H^2(X, \mathbb{Z}) \otimes \mathbb{C}$ to $H^2_{\text{de R}}(X, \mathbb{C})$. Coupling this with the embedding theorem for positive bundles, we have the following theorem of Kodaira:

Theorem *If (X, ω) is Kählerian and ω is integral, then X is projective algebraic.*

This result has been refined in the following important way (a conjecture of Grauert and Riemenschneider proved with different methods by Siu (1984) and by Demailly (1985)): *the same result holds if ω is only assumed to be semipositive and positive in at least one point.*

For Grauert's proof of the Kodaira embedding theorem and a number of other important and beautiful results, we recommend the original paper (Grauert 1962).

Quotients of Bounded Domains

Let D be a bounded domain in $\mathbb{C}^n$ and Γ be a discrete subgroup of $\text{Aut}(D)$ which is acting freely on D with a compact quotient $X := D/\Gamma$. For $\gamma \in \Gamma$ let $J(\gamma, z)$ be the determinant of the Jacobian $d\gamma/dz$ and, given a holomorphic function f, consider (at least formally) the Poincaré series $\sum f(\gamma(z))J(\gamma, z)^k$ of weight k. If f is bounded and $k \geq 2$, then this series converges to a holomorphic function $P(f)$ on D which satisfies the transformation rule $P(f)\,(\gamma(z)) = J(\gamma, z)^{-k} P(f)(z)$.

Now the differential volume form $\Omega := dz_1 \wedge \cdots \wedge dz_n$ transforms in the opposite way (for $k = 1$). Therefore $s(f) = P(f)(\Omega)^k$ is a Γ-invariant section of the kth power of the determinant bundle $K := \Lambda^n T^* D$ of the holomorphic cotangent bundle of D. In other words, $s(f) \in \Gamma(X, K^k)$. Since the choice of f may be varied to show that there are sufficiently many sections to separate points and to guarantee the maximal rank condition, it follows that the canonical bundle K of X is ample. Compact complex manifolds with ample canonical bundle are examples of manifolds which are said to be of general type (*see* Several Complex Variables: Compact Manifolds). Thus, this construction with Poincaré series proves the following: "Every compact quotient D/Γ of a bounded domain is of general type and is in particular projective algebraic."

See also: Gauge Theoretic Invariants of 4-Manifolds; Moduli Spaces: An Introduction; Riemann Surfaces; Several Complex Variables: Compact Manifolds; Twistor Theory: Some Applications [in Integrable Systems, Complex Geometry and String Theory].

Further Reading

Andreotti A and Grauert H (1962) Théorèmes de finitude pour la cohomologie des espaces complexes. *Bulletin de la Société Mathématique de France* 90: 193–259.

Demailly J-P (1985) Champs magnŕtique et inégalitiés de Morse pour lat d''-cohomologie. *Annales de l'Institut de Fourier* 35: 189–229.

Demailly J-P, Complex analytic and algebraic geometry, http://www-fourier.ujf-grenoble.fr/demailly.

Grauert H (1962) Uber Modifikationen und exzeptionelle analytische Mengen. *Mathematische Annalen* 129: 331–368.

Grauert H and Fritzsche K (2001) *From Holomorphic Functions to Complex Manifolds*. Heidelberg: Springer.

Griffiths PhA and Harris J (1978) *Principles of Algebraic Geometry*. New York: Wiley.

Grauert H and Remmert R (1979) *Theory of Stein Spaces*. Heidelberg: Springer.

Grauert H and Remmert R (1984) *Coherent Analytic Sheaves*. Heidelberg: Springer.

Grauert H, Peternell Th, and Remmert R (1994) Several Complex Variables VII. *Encyclopedia of Mathematical Science*, vol. 74. Heidelberg: Springer.

Narasimhan R (1971) Several Complex Variables. *Chicago Lectures in Mathematics*. Chicago, IL: University of Chicago Press.

Siu YT (1984) A vanishing theorem for semi-positive line bundles over non-Kähler manifolds. *Journal of Differntial Geometry* 19: 431–452.

Skoda H (1982) Prolongement des courants positifs fermés de masse finie. *Inventiones Mathematicae* 66: 361–376.

Several Complex Variables: Compact Manifolds

A Huckleberry, Ruhr-Universität Bochum, Bochum, Germany
T Peternell, Universität Bayreuth, Bayreuth, Germany

Introduction

The aim of this article is to give an overview of the classification theory of compact complex manifolds. Very roughly, compact manifolds can be divided into three disjoint classes:

- Projective manifolds, that is, manifolds which can be embedded into some projective space, or manifolds birational to those, usually called Moishezon manifolds. These manifolds are treated by algebraic geometric methods, but very often transcendental methods are also indispensable.
- Compact (nonalgebraic) Kähler manifolds, that is, manifolds carrying a positive closed $(1,1)$-form, or manifolds bimeromorphic to those. This class is treated mainly by transcendental methods from complex analysis and complex differential geometry. However, some algebraic methods are also of use here.
- General compact manifolds which are not bimeromorphic to Kähler manifolds. For two reasons we will essentially ignore this class in our survey. First, because of the lack of methods, not much is known, for example, there is still no complete classification of compact complex surfaces, and it is still unknown whether or not the 6-sphere carries a complex structure. And second, for the purpose of this encyclopedia, this class seems to be less important.

The main problems of classification theory can be described as follows.

- Birational classification: describe all projective (Kähler) manifolds up to birational (bimeromorphic) equivalence; find good models in every equivalence class. This includes the study of invariants.
- Biholomorphic classification: classify all projective (Kähler) manifolds with some nice property, for example, curvature, many symmetries, etc.
- Topological classification and moduli: study all complex structures on a given topological manifold – including the study of topological invariants of complex manifolds; describe complex structures up to deformations and describe moduli spaces.
- Symmetries: describe group actions and invariants – this is deeply related with the moduli problem.

In this article we will assume familiarity with basic notions and methods from several complex variables and/or algebraic geometry. In particular we refer to Several Complex Variables: Basic Geometric Theory in this encyclopedia.

We first note some standard notation used in this article. If X is a complex manifold of dimension n, then T_X will denote its holomorphic tangent bundle and Ω^p_X the sheaf of holomorphic p-forms, that is, the sheaf of sections of the bundle $\bigwedge^p T^*_X$. The bundle $\bigwedge^n T^*_X$ is usually denoted by K_X, the canonical bundle of X and its sheaf of sections is the dualizing sheaf ω_X, but frequently we will not distinguish between vector bundles and their sheaves of sections. An effective (Cartier) divisor on a normal space X is a finite linear combination $\sum n_i Y_i$, where $n_i > 0$ and $Y_i \subset X$ are irreducible reduced subvarieties of codimension, which are locally given by one equation. If L is a line bundle, then instead of $L^{\otimes m}$ we often write mL. If X is a compact variety and E a vector bundle or coherent sheaf, then the dimension of the finite-dimensional vector space $H^q(X,E)$ will be denoted by $h^q(X,E)$.

Birational Classification

Two compact manifolds X and Y are bimeromorphically equivalent, if there exist nowhere dense analytic subsets $A \subset X$ and $B \subset Y$ and a biholomorphic map $X\backslash A \to Y\backslash B$ such that the closure of the graph is an analytic set in $X \times Y$. In case X and Y are algebraic, one rather says that X and Y are birationally equivalent. This induces an isomorphism between the function fields of X and Y. If X and Y are projective or Moishezon(see below), then conversely an isomorphism of their function fields induces a birational equivalence between X and Y. Important examples are blow-ups of submanifolds; locally they can be described as follows. Suppose that locally X is an open set $U \subset \mathbb{C}^n$ with coordinates $z_1, \ldots, z_n$ and that $A \subset X$ is given by $z_1 = \cdots = z_m = 0$. Then the blow-up $\hat{X} \to X$ is the submanifold $\hat{X} \subset U \times \mathbb{P}_{n-m-1}$ given by the equations

$$y_j t_i - y_i t_j = 0$$

where t_j are homogeneous coordinates in $\mathbb{P}_{n-m-1}$.

The Chow lemma says that any birational – even rational – maps can be dominated by a sequence of blow-ups with smooth centers. Recently other factorizations ("weak factorization," using blow-ups and blow-downs) have been established.

A projective manifold is a compact manifold which is a submanifold of some projective space $\mathbb{P}_N$. Of course, a projective manifold can be embedded into projective spaces in many ways. According to Chow's theorem (*see* Several Complex Variables: Basic Geometric Theory), $X \subset \mathbb{P}_N$ is automatically given by polynomial equations and is therefore an algebraic variety. This is part of Serre's GAGA principle which roughly says that all global analytic objects on a projective manifold, for example, vector bundles or coherent sheaves and their cohomology are automatically algebraic. A compact manifold which is bimeromorphically equivalent to a projective manifold is called a Moishezon manifold. These arise naturally, for example, as quotient of group actions, compactifications, etc.

The most important birational invariant of compact manifolds is certainly the Kodaira dimension $\kappa(X)$. It is defined in three steps:

- $\kappa(X) = -\infty$ iff $h^0(mK_X) = 0$ for all $m \geq 1$.
- $\kappa(X) = 0$ iff $h^0(mK_X) \leq 1$ for all m, and $h^0(mK_X) = 1$ for some m.
- In all other cases we can consider the meromorphic map $f_m : X \to \mathbb{P}_{N(m)}$ associated to $H^0(mK_X)$ for all those m for which $h^0(mK_X) \geq 2$. Let V_m denote the (closure of the) image of f_m. Then $\kappa(X)$ is defined to be the maximal possible $\dim V_m$.

Recall that f_m is defined by $[s_0 : \cdots : s_N]$ for a given base s_i of $H^0(mK_X)$, cf. Several Complex Variables: Basic Geometric Theory.

In the same way one defines the Kodaira (or Iitaka) dimension $\kappa(L)$ of a holomorphic line bundle L (instead of $L = K_X$).

We are now going to describe geometrically the different birational equivalence classes and how to single out nice models in each class. Using methods in characteristic p, Miyaoka and Mori proved the following theorem:

Theorem 1 *Let X be a projective manifold and suppose that through a general point $x \in X$ there is a curve C such that $K_X \cdot C < 0$. Then X is uniruled, that is, there is a family of rational curves covering X.*

A rational curve is simply the image of nonconstant map $f : \mathbb{P}_1 \to X$. It is a simple matter to prove that uniruled manifolds have $\kappa(X) = -\infty$, but the converse is an important open problem. A step towards this conjecture has recently been made by Boucksom *et al.* (2004) if K_X is not pseudoeffective, that is, K_X "cannot be approximated by effective divisors," then X is uniruled. Here one also finds a discussion of the case when K_X is pseudoeffective.

Mori theory is central in birational geometry. To state the main results in this theory, we recall the notion of ampleness: a line bundle L is ample if L carries a metric of positive curvature. Alternatively some tensor power of L has enough global section to separate points and tangents and there gives an embedding into some projective space; *see* Several Complex Variables: Basic Geometric Theory for more details. The notion of nefness, which is in a certain sense the degenerate version of ampleness, plays a central role in Mori theory: a line bundle or divisor L is nef if

$$L \cdot C = \deg(L|C) \geq 0$$

for all curves $C \subset X$. Examples are those L carrying a metric of semipositive curvature, but the converse is not true. However, if L is nef, there exists for all positive $\epsilon > 0$ a metric h_ϵ with curvature $\Theta_\epsilon > -\epsilon\omega$, where ω is a fixed positive form. In this context singular metrics on L are also important. Locally they are given by $e^{-\varphi}$ with a locally integrable weight function φ and they still have a curvature current Θ. If L has a singular metric with Θ bounded from below as current by a Kähler form, then L is big, that is, $\kappa(L) = \dim X$, the birational version of ampleness. If one simply has $\Theta \geq 0$ as current, then L is pseudoeffective (and vice versa). All these positivity notions only depend on the Chern class $c_1(L)$ of L and therefore one considers the ample cone

$$K_{\text{amp}} \subset (H^{1,1}(X) \cap H^2(X, \mathbb{Z})) \otimes \mathbb{R}$$

and the cone of curves

$$\overline{NE}(X) \subset (H^{n-1,n-1}(X) \cap H^{2n-2}(X, \mathbb{Z})) \otimes \mathbb{R}$$

The ample cone is by definition the closed cone of nef divisors, the interior being the ample classes, while the cone of curves is the closed cone generated by the fundamental classes of irreducible curves.

A basic result says that these cones are dual to each other. The structure of $\overline{NE}(X)$ in the part where K_X is negative is very nice; one has the following cone theorem:

Theorem 2 *$\overline{NE}(X)$ is locally finite polyhedral in the half-space $\{K_X < 0\}$; the (geometrically) extremal rays contain classes of rational curves.*

A ray $R = \mathbb{R}_+[a]$ is said to be extremal in a closed cone K if the following holds: given $b, c \in K$ with $b + c \in R$, then $b, c \in R$. Given such an extremal ray $R \subset \overline{NE}(X)$, one can find an ample line bundle H and a rational number t such that $K_X + tH$ is nef and $K_X + tH \cdot R = 0$. Using the Kawamata–Viehweg vanishing theorem, a generalization of Kodaira's vanishing theorem, which is one of the technical corner stones of the theory, one proves the so-called

Base point free theorem *Some multiple of $K_X + tH$ is spanned by global sections and therefore defines a holomorphic map $f: X \to Y$ to some normal projective variety Y contracting exactly those curves whose classes belong to R.*

These maps are called "contractions of extremal rays" or "Mori contractions." In dimension 2 they are classical: either $X = \mathbb{P}_2$ and f is the constant map, or f is a $\mathbb{P}_1$-bundle or f is birational and the contraction of a $\mathbb{P}_1$ with normal bundle $\mathcal{O}(-1)$, that is, f contracts a (-1)-curve. In particular Y is again smooth. In the first two cases X has a very precise structure, but in the third birational case one proceeds by asking whether or not K_Y is nef. If it is not nef, we start again by choosing the contraction of an extremal ray; if K_Y is nef, then a fundamental result says that a multiple of K_Y is spanned. The class of manifolds with this property will be discussed later.

The situation in higher dimensions is much more complicated. For example, Y need no longer be smooth. However the singularities which appear are rather special.

Definition 1 A normal variety X is said to have only terminal singularities if first some multiple of the canonical (Weil) divisor K_X is a Cartier divisor, that is, a line bundle (one says that X is $\mathbb{Q}$-Gorenstein) and second if for some (hence for every) resolution of singularities $\pi: X \to X$ the following holds:

$$K_{\hat{X}} = \pi^*(K_X) + \sum a_i E_i$$

where the E_i run over the irreducible π-exceptional divisors and the a_i are strictly positive.

A brief remark concerning Weil divisors is in order: a Weil divisor is a finite linear combination $\sum a_i Y_i$ with Y_i irreducible of codimension 1, but Y_i is not necessarily locally defined by one equation. Recall that if each Y_i is given locally by one equation, then the Weil divisor is Cartier. On a smooth variety these notions coincide.

One important consequence is that $\kappa(X) = \kappa(\hat{X})$ in case of terminal singularities, which is completely false for arbitrary singularities. Also notice that terminal singularities are rational: $R^q\pi_*(\mathcal{O}_{\hat{X}}) = 0$ for $q \geq 1$. Terminal singularities occur in codimension at least 3. Thus they are not present on surfaces. In dimension 3 terminal singularities are well understood. The main point in this context is that for a birational Mori contraction the image Y often has terminal singularities.

Now the scheme of Mori theory is the following. Start with a projective manifold X. If K_X is nef, we stop; this class is discussed later. If K_X is not nef, then perform a Mori contraction $f: X \to Y$. There are two cases:

- If $\dim Y < \dim X$, then the general fiber F is a manifold with ample $-K_F$, that is, a Fano manifold (discussed in the next section). Here we stop and observe that $\kappa(X) = -\infty$. Of course one can still investigate Y and try to say more on the structure of the fibration f.
- If $\dim Y = \dim X$, then Y has terminal singularities – unless f is a small contraction which means that no divisors are contracted. Thus if f is not small, we may attempt to proceed by substituting X by Y.

As a result one must develop the entire theory for varieties with terminal singularities. The big problem arises from small contractions f. In that case K_Y cannot be $\mathbb{Q}$-Cartier and the machinery stops. So new methods are required. At this stage, other aspects of the theory lead one to attempt a certain surgery procedure which should improve the situation and allow one to continue as above. The expected surgery $Y \rightharpoonup Y'$, which takes place in codimension at least 2, is a "flip." The idea is that we should substitute a small set, namely the exceptional set of a small contraction, by some other small set (on which the canonical bundle will be positive) to improve the situation. Of course Y' should possess only terminal singularities. The existence of flips is very deep and has been proved by S Mori in dimension 3. Moreover, there cannot be an infinite sequence of flips, at least in dimension at most 4.

In summary, by performing contractions and flips one constructs from X a birational model X' with terminal singularities such that either

- $K_{X'}$ is nef in which case we call X' a minimal model for X, or
- X' admits a Fano fibration $f': X' \to Y'$ (discussed below), in which case $\kappa(X) = \kappa(X') = -\infty$.

Up to now, Mori theory (via the work of Kawamata, Kollár, Mori, Reid, Shokurov, and others) works well in dimension 3 (and possibly in the near future in dimension 4) but in higher dimensions there are big problems with the existence of flips. Of course there might be completely different and possibly less precise ways to construct a minimal model. One way is to consider the canonical ring R of a manifold of general type:

$$R = \sum H^0(mK_X)$$

If R is finitely generated as $\mathbb{C}$-algebra, then Proj(R) would be at least a canonical model which

has slightly more complicated singularities than a minimal model. However, it is known that this "finite generatedness problem" is equivalent to the existence of minimal models. On the other hand, if X is of general type with K_X nef (hence essentially ample) or more generally when some positive multiple mK_X is generated by global sections, then R is finitely generated.

We now must discuss the case of a nef canonical bundle. The behavior is predicted by the

Abundance conjecture. *If X has only terminal singularities and K_X is nef, then some multiple mK_X is spanned.*

Up to now this conjecture is known only in dimension 3 (Kawamata, Kollár, Miyaoka). In higher dimensions it is even unknown if there is a single section in some multiple mK_X. If mK_X is spanned, one considers the Stein factorization $f: X \to Y$ of the associated map, which is called the Iitaka fibration (if not birational) and we have $\dim Y = \kappa(X)$ by definition. The general fiber F is a variety with $K_F \equiv 0$, a class discussed in the next section. If f is birational, then Y will be slightly singular (so-called canonical singularities) and K_Y will be ample. Essentially we are in the case of negative Ricci curvature.

Everything that was outlined above holds for projective manifolds. In the Kähler case one would expect the same picture, but the methods completely fail, and new, analytic methods must be found. Only very few results are known in this context.

We come back to the case of a Fano fibration $f: X \to Y$. By definition the anticanoical bundle $-K_X$ is relatively ample so that the general fiber is a Fano variety. In this case there are no constraints on Y.

To see how much of the geometry of X is dictated by the rational curves, one considers the so-called rational quotient of X. Here we identify two very general points on X if they can be joined by a chain of rational curves. In that way we obtain the rational quotient

$$f: X \dashrightarrow Y$$

This map is merely meromorphic, but has the remarkable property of being "almost holomorphic," that is, the set of indeterminacies does not project onto Y. In other words, one has nice compact fibers not meeting the indeterminacy set. If Y is just a point, then all points of X can be joined by chains of rational curves and X is called rationally connected. This notion is clearly birationally invariant.

A deep theorem of Graber–Harris–Starr states that, given a Fano fibration (or a fibration with rationally connected fibers) $f: X \to Y$, then X is rationally connected if and only if Y is.

Manifolds X_n which are birational to $\mathbb{P}_n$ are called rational. If there merely exists a surjective ("dominant") rational map $\mathbb{P}_n \dashrightarrow X$, then X is said to be unirational. Of course rational (resp. unirational) manifolds are rationally connected, but to decide whether a given manifold is rational/unirational is often a very deep problem. Therefore, rational connectedness is often viewed as a practical substitute for (uni)rationality.

Often it is very important to compute the Kodaira dimension of fiber spaces. Let us fix a holomorphic surjective map $f: X \to Y$ between projective manifolds and we suppose f has connected fibers. Then the so-called conjecture C_{nm} states that

$$\kappa(X) \geq \kappa(F) + \kappa(Y)$$

where F is the general fiber of f. This conjecture is known in many cases, for example, when the general fiber is of general type, but it is wide open in general. It is deeply related to the existence of minimal models (Kawamata).

Biholomorphic Classification

In this section we discuss manifolds X with

- ample anticanonical bundles $-K_X$ (Fano manifolds),
- trivial canonical bundles, and
- ample canonical bundles K_X.

Due to the solution of the Calabi conjecture by Yau and Aubin, these classes are characterized by a Kähler metric of positive (resp. zero, resp. negative) Ricci curvature. In principle, in view of the results of Mori theory, one should rather consider varieties with terminal singularities, but we ignore this aspect completely. Philosophically, up to birational equivalence all manifolds are via fibrations somehow composed of those classes via fibrations, possibly also up to étale coverings.

Examples of Fano manifolds are hypersurfaces of degree at most $n+1$ in $\mathbb{P}_{n+1}$, Grassmannians, or more generally homogenenous varieties G/P with G semisimple and P a parabolic subgroup. Fano manifolds are simply connected. This can be seen either by classical differential geometric methods using a Kähler metric of positive curvature or via the fundamental

Theorem 3 *Fano manifolds are rationally connected.*

The only known proof of this fact uses, as in the uniruled criterion mentioned above, characteristic p methods. By just using complex methods it is not

known how to construct a single rational curve (of course, in concrete examples the rational curves are seen immediately). One still has to observe that rationally connected manifolds are simply connected, which is not so surprising, since rational curves lift to the universal cover.

At least in principle, Fano manifolds can be classified:

Theorem 4 *There are only finitely many families of Fano manifolds in every dimension.*

A family (of Fano manifolds) is a submersion $\pi: \mathcal{X} \to S$ (with S irreducible) such that all fibers are Fano manifolds. The essential step is to bound $(-K_X)^n$. An actual classification has been carried out only in dimension up to 3; in dimension 2 one finds $\mathbb{P}_2, \mathbb{P}_1 \times \mathbb{P}_1$ and the so-called del Pezzo surfaces ($\mathbb{P}_2$ blown up in at most eight points in general position). In dimension 3 there are already 17 families of Fano 3-folds with $b_2 = 1$ and 88 families with $b_2 \geq 2$.

An extremely hard question is to decide whether a given Fano manifold is rational or unirational. Even in dimension 3 this is not completely decided.

The next class to be discussed are the manifolds with trivial canonical class K_X. This means that there is a holomorphic n-form without zeros ($n = \dim X$). Important examples are tori and hypersurface in $\mathbb{P}_{n+1}$ of degree $n+2$. Simply connected manifolds with trivial canonical bundles are further divided into irreducible Calabi–Yau manifolds and irreducible symplectic manifolds. The first class is defined by requiring that there are no holomorphic p-forms for $p < \dim X$ whereas the second is characterized by the existence of a holomorphic 2-form of everywhere maximal rank. A completely different characterization is by holonomy: an irreducible Calabi–Yau manifold has SU-holonomy whereas irreducible symplectic manifolds have Sp-holonomy (with respect to a suitable Kähler metric).

The splitting theorem of Beauville–Bogomolov–Kobayashi says

Theorem 5 *Let X be a projective (or compact Kähler) manifold with trivial canonical bundle. Then there exists a finite unbranched cover $\tilde{X} \to X$ such that*

$$X = A \times \Pi X_i \times \Pi Y_j$$

with A a torus, X_i irreducible Calabi–Yau, and Y_j irreducible symplectic.

The key to the proof of this theorem is the existence of a Ricci-flat Kähler metric on X, a Kähler–Einstein metric with zero Ricci curvature. Actually one has a stronger result: instead of assuming K_X to be trivial, just assume that $c_1(X) = 0$ in $H^2(X, \mathbb{R})$. Then there exists a finite unramified cover $X \to X$ such that $K_{\tilde{X}}$ is trivial. In view of Mori theory, normal projective varieties X with at most terminal singularities and $K_X \equiv 0$ (i.e., $K_X \cdot C = 0$ for all curves) should also be investigated. It is expected that similar structure theorems hold; in particular $\pi_1(X)$ should be finite. The main difficulty is that there are no differential methods available; on the other hand an algebraic proof even for the splitting theorem in the smooth case is unknown.

Calabi–Yau manifolds play an important role in string theory and mirror symmetry (*see* Mirror Symmetry: A Geometric Survey). Here we mention two basic problems. The first is the problem of boundedness:

Are there only finitely many families of Calabi–Yau manifolds in any dimension?

This problem is wide open; in particular one might ask:

Is the Hodge number $h^{1,2}$ bounded for Calabi–Yau 3-folds?

The other problem asks for the existence of rational curves. In all known examples there are rational curves, but a general existence proof is not known. The case where $b_2(X) = 1$ seems to be particularly difficult. If $b_2(X) \geq 2$, then in may cases one can hope to find a fibration or a birational map, at least for 3-folds. Given such a map, the existence of rational curves is simple. For example, if $D \subset X$ is an irreducible hypersurface which is not nef, choose H ample and consider the *a priori* positive real number p such that $D + pH$ is on the boundary of the ample cone. Then actually p is rational and a suitable multiple $m(D + pH)$ is spanned and defines a contraction on X. This comes from "logarithmic Mori theory."

The above splitting theorem exhibits a torus factor and all holomorphic 1-forms on X come from this torus. This principle generalizes: given any projective or compact Kähler manifold X, there exists a "universal object," the Albanese torus

$$\mathrm{Alb}(X) = H^0(\Omega^1_X)^* / H_1(X, \mathbb{Z})$$

(which is algebraic if X is) together with a holomorphic map

$$\alpha : X \to \mathrm{Alb}(X)$$

the Albanese map. This Albanese map is given by integrating 1-forms and is often far from being surjective. The important property is now that, given a holomorphic 1-form ω on X, there exists a holomorphic 1-form η on the Albanese torus such that $\omega = \alpha^*(\eta)$. The universal property reads as

follows: every map $X \to T$ to a torus factors via an affine map $\mathrm{Alb}(X) \to T$.

There is a nonabelian analog, the so-called Shafarevich map, but at the moment this map is only known to be meromorphic. It is an important tool to study the fundamental group $\pi_1(X)$. We refer to Campana (1996) and Kollár (1995).

In the following, Chern classes of holomorphic vector bundles will be important. Let X be a compact complex manifold and E a holomorphic vector bundle on X. The jth Chern class of E is an element

$$c_j(E) \in H^{2j}(X, \mathbb{Q}) \cap H^{j,j}(X)$$

It can be defined, for example, by putting a Hermitian metric on E, computing the curvature of the canonical connection compatible with both the metric and the holomorphic structure and then by applying certain linear operators coming from symmetric functions such as determinant and trace. Actually Chern classes can be attached to every complex topological vector bundle on a topological manifold; then $c_j(E)$ will simply live in $H^{2j}(X, \mathbb{R})$. There is also a purely algebraic construction by Grothendieck. We refer, for example, to Fulton (1984) as well as for a discussion of the elementary functorial properties of Chern classes. Here we just recall that for a rank-r vector bundle E the first Chern class

$$c_1(E) = c_1\left(\bigwedge E\right)$$

where the Chern class of the line bundle $\bigwedge^r E$ as given in Several Complex Variables: Basic Geometric Theory actually lives in $H^2(X, \mathbb{Z})$.

Finally we discuss manifolds with ample canonical class K_X. Here moduli question often plays a central role. Moduli spaces of surfaces with fixed c_1^2 and c_2 are very intensively studied (by Catanese, Ciliberto, and others). Here, without going into details, we will concentrate on the very interesting topic of Kähler–Einstein metrics.

A Kähler metric ω is said to be Kähler–Einstein, if its Ricci curvature $\mathrm{Ric}(\omega)$ is proportional to ω. The proportionality factor λ can be taken to be $-1, 0, 1$. In case K_X is ample or trivial, Kähler–Einstein metrics always exist by Yau and Aubin (cases $\lambda = -1$, resp. $\lambda = 0$). However if X is Fano, there are obstructions, and a Kähler–Einstein metric does not always exist. An important consequence of the existence of a Kähler–Einstein metric on a manifold X_n with ample canonical class is the Miyaoka–Yau inequality:

$$c_1^2 \omega^{n-2} \leq \frac{2n+1}{n} \omega^{n-2}$$

In case of equality, X is covered by the n-dimensional unit ball.

The same inequality holds in case $K_X = 0$, and as a consequence the Chern class $c_2(X)$ is in some sense semipositive. If $c_2(X) = 0$, then some finite unramified cover of X is a torus.

There is an interesting relation to stability. Recall that a vector bundle E on a compact Kähler manifold X_n is semistable with respect to a given Kähler form ω, if for all proper coherent subsheaves $\mathcal{F} \subset E$ of rank-r the following inequality holds:

$$\frac{c_1(\mathcal{F}) \cdot \omega^{n-1}}{r} \leq \frac{c_1(E) \cdot \omega^{n-1}}{n}$$

In case of strict inequality, E is said to be stable.

The basic observation is now that the tangent bundle of a manifold with a Kähler–Einstein metric is semistable (with respect to the Kähler–Einstein metric). It is expected that Fano manifolds with $b_2 = 1$ have (semi?-)stable tangent bundles, although in certain situations they do not admit a Kähler–Einstein metric.

Again the first two Chern classes of a semistable vector bundle fulfill an inequality:

$$c_1^2(E) \cdot \omega^{n-2} \leq \frac{2r}{r-1} c_2(E) \cdot \omega^{n-2}$$

Equally important, semistable bundles with fixed numerical data form moduli spaces, this being the origin of the stability notion (Mumford). In this context, the notion of an Hermite–Einstein bundle is also important. Given a holomorphic vector bundle E with a Hermitian metric h, there is a unique connection F_h on E compatible both with h and the complex structure. F_h is a (1,1)-form with values in $\mathrm{End}(E)$. Now suppose (X, ω) is Kähler and let λF_h be the contraction of F_h with ω. Then (E, h) is said to Hermite–Einstein on (X, ω), if

$$\lambda F_h = \gamma \mathrm{id}$$

with some constant γ and id: $E \to E$ the identity. Notice that (X, ω) is Kähler–Einstein if (T_X, h) is Hermite–Einstein over (X, ω) with h the Kähler metric with Kähler form ω. It is not so difficult to see that Hermite–Einstein bundles are semistable (with respect to the underlying Kähler form) and actually are directs sum of stable Hermite–Einstein bundles. Conversely, a very deep theorem of Uhlenbeck–Yau says that *every stable vector bundle on a compact Kähler manifold is Hermite–Einstein*. This is known as the Kobayashi–Hitchin correspondence; see Lübke and Teleman (1995).

Topology, Invariants and Cohomology

Besides the Kodaira dimension there are other important invariants of compact complex manifolds. Of course there are topological invariants such as the Betti number $b_i(X) = \dim H^i(X, \mathbb{R})$ or the fundamental group $\pi_1(X)$. The fundamental group has been studied intensively in the last decade. A central question asks which groups can occur as fundamental groups of compact Kähler manifolds; another problem is the so-called Shafarevitch conjecture which says that the universal cover of a compact Kähler manifold should be holomorphically convex. We refer to Campana (1996) and Kollár (1995).

The plurigenera,

$$P_m(X) = \dim h^0(mK_X)$$

are also extremely important. Here, Siu recently proved that $P_m(X)$ is constant in families of projective manifolds. Other important invariants are $h^0(X, (\Omega^1_X)^{\otimes m})$. For example, it is conjectured that if

$$h^0(X, (\Omega^1_X)^{\otimes m}) = 0$$

for all positive m, then X is rationally connected. Tensor powers of the cotangent bundle somehow capture more of the structure of X than the Kodaria dimension but they are more difficult to treat. The relevance of the dimensions

$$h^0(X, \Omega^p_X)$$

of holomorphic forms is easier to understand. More generally one has the Hodge numbers

$$h^{p,q}(X) = \dim H^q(X, \Omega^p_X)$$

For compact Kähler manifolds, the Hodge decomposition states

$$H^r(X, \mathbb{C}) = \bigoplus_{p+q=r} H^{p,q}(X)$$

Furthermore, Hodge duality,

$$H^{p,q}(X) = H^{q,p}(X)$$

holds. These results form a cornerstone for the geometry of compact Kähler manifolds and the starting point of Hodge theory. Hodge theory is, for example, extremely important in the study of families of manifolds and moduli.

Concerning the topology of projective (Kähler) manifolds, the following two questions are very basic.

- Which invariants are topological (or diffeomorphic) invariants?
- What are the projective or Kähler structures on a given compact topological manifold?

Concerning the first, Hodge decomposition implies that the irregularity $h^0(\Omega^1_X)$ is actually a topological invariant. However it is unknown whether the number of holomorphic 2-forms is a topological invariant of Kähler 3-folds. Both questions have been intensively studied in dimension 2. However, in higher dimensions almost nothing is known. For example, it is not known whether there is projective manifold of general type of even dimension which is homeomorphic to a quadric, that is, a hypersurface of degree 2 in projective space.

Other important tools in the study of projective/Kähler manifolds are listed below.

- Cohomological methods: Riemann–Roch theorem and holomorphic Morse inequalities; vanishing theorems (Kodaira, Kawamata–Viehweg, etc.); Serre duality. References: Demailly (2000), Demailly and Lazarsfeld, Fulton (1984), Grauert *et al.* (1994), Lazarsfeld (2004).
- L^2 methods: extension theorems, singular metrics, multiplier ideals, etc. Reference: Demailly and Lazarsfeld (2001), Lazarsfeld (2004).
- Theory of currents. Reference: Demailly 2000.
- Cycle space and Douady space, resp. Chow scheme and Hilbert scheme. Reference: Fulton 1984, Grauert 1994, Kollár 1996.

We restrict our remarks on just one of these topics, vanishing theorems. The classical Kodaira–Nakano vanishing theorem says that if X is a compact manifold of dimension n with a positive (ample) line bundle L, then

$$H^q(X, L \otimes \Omega^p) = 0$$

for $p + q > n$. This is usually proved via harmonic theory, that is, by representing the cohomology space by harmonic (p, q)-forms with values in L and by computing integrals of these forms. For many purposes, for example, for Mori theory, it is important to generalize this to a line bundle which have some positivity properties but which are not ample. This works only for $p = n$, however this is the most important part of the Kodaira–Nakano vanishing. The Kawamata–Viehweg vanishing theorem in its most basic version says that given a nef and big line bundle L, then Kodaira vanishing still holds:

$$H^q(X, L \otimes K_X) = 0$$

for $q \geq 1$. But actually it is not necessary to assume L nef, in fact the following is true. Let

$$D = \sum a_i D_i$$

be an effective $\mathbb{Q}$-divisor, that is, all a_i are positive rational numbers. Let $\langle a_i \rangle$ be the fractional part of a_i and suppose that the $\mathbb{Q}$-divisor $\sum \langle a_i \rangle D_i$ has normal crossings. Let $\lceil a_i \rceil$ be the roundup of a_i and put $L = \sum \lceil a_i \rceil D_i$. If D is big and nef, then

$$H^q(X, L \otimes K_X) = 0$$

for $q \geq 1$. Of course L itself need not be nef! This generalization is technically very important and yields substantial freedom for birational manipulations. We refer to Kawamata *et al.* (1987) and Lazarsfeld (2004). Even this is not the end of the story: the Kawamata–Viehweg theorem is embedded in the broader context of the Nadel vanishing theorem where multiplier ideal sheaves come into the play. See Demailly and Lazarsfeld and Lazarsfeld (2004).

Homogeneous Manifolds

In this section we consider vector fields and holomorphic group actions on compact (Kähler) manifolds. Our main reference is Huckleberry (1990) with further literature given there.

We denote by $\mathrm{Aut}(X)$ the group of holomorphic automorphisms of the compact manifold X (well known to be a complex Lie group), and by $G := \mathrm{Aut}^0(X)$ the connected component containing the identity. The tangent space at any point of $\mathrm{Aut}^0(X)$ can naturally be identified with $H^0(X, T_X)$, the (finite-dimensional) space of holomorphic vector fields on X. In fact, by integration, a vector field determines a one-parameter group of automorphisms.

One says that X is homogeneous if G acts transitively on X. Therefore, one can write

$$X = G/H$$

where H is the isotropy subgroup of any point $x_0 \in X$, that is, the subgroup of automorphisms fixed x_0. Conversely one can take a complex Lie group G and a closed subgroup H and form the quotient G/H which is again a complex manifold and in fact homogeneous (of course not necessarily compact).

Going back to a compact manifold X, the condition to be homogenenous can be rephrased by saying that the tangent bundle is generated by global sections, that is, if $x \in X$ and $e \in T_{X,x}$, then there exists $v \in H^0(X, T_X)$ such that $v(x) = e$. The easiest case is when T_X is trivial. If X is Kähler, this is exactly the case when X is torus, $X = \mathbb{C}^n/\Gamma$ with $\Gamma \simeq \mathbb{Z}^{2n}$ a lattice, but without the Kähler assumption there are many more examples (the so-called parallelizable manifolds).

More generally, let us consider the case that the compact Kähler manifold X admits a vector field v without zeros, but X is not required to be homogeneous. Then a theorem of Lieberman says that there is a finite unramified cover $f : X \to X$ and a splitting

$$\tilde{X} \simeq F \times T$$

with T a torus, such that $f^*(v)$ is the pullback of a vector field on T. On the other hand, if v has a zero, then a classical theorem of Rosenlicht says that X is covered by rational curves, that is, X is uniruled. In particular $\kappa(X) = -\infty$. Notice also that a manifold of general type can never carry a vector field, in other words, the automorphism group is discrete, even finite.

Coming back to compact homogeneous Kähler manifolds, the first thing to study is the Albanese map. The Borel–Remmert theorem says that

$$X \simeq T \times Q$$

where T is the Albanese torus. This is proved using a maximal compact subgroup $K \subset G$ and by some averaging process over K. Moreover, Q is a rational homogeneous manifold. The structure of Q is more precisely the following. One can write $Q = S/P$ with S a semisimple Lie group and $P \subset S$ parabolic, which means that P contains a maximal connected solvable subgroup (the so-called Borel subgroup). The main ingredients of the proof are the Tits fibration, the Levi–Malcev decomposition of a Lie group into its radical and a semisimple group, and the Borel fixed point theorem:

Theorem 6 *Let $G \subset \mathrm{GL}_n(\mathbb{C})$ be a connected solvable subgroup and $X \subset \mathbb{P}_{n-1}$ be a G-stable subvariety. Then G has a fixed point on X.*

In the homogenenous Kähler case, the rationality of Q is seen by exhibiting an open subset in Q which is algebraically isomorphic to $\mathbb{C}^n$.

Now things come down to classify all rational homogenenous manifold S/P which is of course classical. Notice that all rational homogeneous manifolds are Fano. One knows that a rational homogeneous manifold with Betti number $b_2 \geq 2$ can be fibered over another rational homogenenous manifold with fibers rational homogeneous – this is actually a fiber bundle. The case that $b_2 = 1$ can be rephrased by saying that P is maximal parabolic. This fiber bundle might not be trivial as shown by the projectivized tangent bundle $\mathbb{P}(T_{\mathbb{P}_n})$.

Compact Hermitian symmetric spaces form a particularly interesting subclass of homogeneous Kähler manifolds. A manifold equipped with a

Hermitian metric is called Hermitian symmetric, if for every $x \in X$ there exists an involutive holomorphic isometry fixing x. Mok has shown the remarkable fact that the simply connected compact Hermitian symmetric spaces are exactly those simply connected compact manifolds carrying a Kähler metric with semipositive holomorphic bisectional curvature. The only manifold having a metric with positive holomorphic bisectional curvature is $\mathbb{P}_n$ (Siu-Yau, Mori).

See also: Classical Groups and Homogeneous Spaces; Einstein Manifolds; Mirror Symmetry: A Geometric Survey; Moduli Spaces: An Introduction; Riemann Surfaces; Several Complex Variables: Basic Geometric Theory; Topological Sigma Models; Twistor Theory: Some Applications [in Integrable Systems, Complex Geometry and String Theory].

Further Reading

Beltrametti M and Sommese AJ (1995) *The Adjunction Theory of Complex Projective Varieties*. Berlin: de Gruyter.

Boucksom S, Demailly JP, Paun M, and Peternell T (2004) The pseudo-effective cone of a compact Kähler manifold and varieties of negative Kodaira dimension math.AG/0405285

Campana F (1996) Kodaira dimension and fundamental group of compact Kähler manifolds. In: Andreatta M and Peternell T (eds.) *Higher Dimensional Complex Varieties*, pp. 89–162. Berlin: de Gruyter.

Demailly JP (2000) Complex analytic and algebraic geometry, http://www-fourier.ujf-grenoble.fr/ demailly.

Demailly JP and Lazarsfeld R (eds.) Vanishing Theorems and Effective Results in Algebraic Geometry, ICTP Lecture Notes, vol. 6, Trieste.

Fulton W (1984) *Intersection Theory*. Heidelberg: Springer.

Griffiths Ph and Harris J (1978) *Priniciples of Algebraic Geometry*. New York: Wiley.

Grauert H, Peternell Th, and Remmert R (1994) Several Complex Variables VII. *Encyclopedia of Mathematical Sciences*, vol. 74. Heidelberg: Springer.

Huckleberry A (1990) Actions of groups of holomorphic transformations. In: Barth W and Narasimhan R (eds.) *Several Complex Variables VI*, Encyclopedia of Mathematical Science, vol. 69, pp. 143–196. Berlin: Springer.

Kawamata Y, Matsuda K, and Matsuki K (1987) Introduction to the minimal model problem. *Advance Studies in Pure Mathematics* 10: 283–360.

Kollár J (1995) *Shafarevitch Maps and Automorphic Forms*. Princeton University Press.

Kollár J (1996) Rational Curves on Algebraic Varieties. *Ergebnisse der Mathematik und ihrer Grenzgebiete*, vol. 32. Heidelberg: Springer.

Lazarsfeld R (2004) Positivity in Algebraic Geometry I, II. *Ergebnisse der Mathematik und ihrer Grenzgebiete*, vol. 48 and 49. Heidelberg: Springer.

Lübke M and Teleman A (1995) *The Kobayashi–Hitchin Correspondence*. Singapore: World Scientific.

Matsuki K (2002) *Introduction to the Mori Program*, Universitext. Heidelberg: Springer.

Mori S (1987) Classification of higher-dimensional varieties. *Proceedings of Symposia in Pure Mathematics* 46: 269–331.

Parshin AN and Shafarevich IR (eds.) (1999) *Algebraic Geometry V – Fano Varieties*, vol. 47, *Encyclopedia of Mathematical Sciences*. Heidelberg: Springer.

Siu YT (1987) Lectures on Hermitian–Einstein Metrics for Stable Bundles and Kähler–Einstein Metrics. *DMV Seminar*, vol. 8. Basel: Birkhäuser.

Ueno K (1975) Classification Theory of Compact Complex spaces. *Lecture Notes in Math.*, vol. 439. Heidelberg: Springer.

Viehweg E (1995) Quasi-Projective Moduli for Polarozed Manifolds. *Ergebnisse der Mathematik und ihrer Grenzgebiete*, vol. 30. Heidelberg: Springer.

Shock Wave Refinement of the Friedman–Robertson–Walker Metric

B Temple, University of California at Davis, Davis, CA, USA

J A Smoller, University of Michigan, Ann Arbor, MI, USA

Introduction

In the standard model of cosmology, the expanding universe of galaxies is described by a Friedman–Robertson–Walker (FRW) metric, which in spherical coordinates has a line element given by (Blau and Guth 1987, Weinberg 1972)

$$ds^2 = -dt^2 + R^2(t)\left\{\frac{dr^2}{1-kr^2} + r^2[d\theta^2 + \sin^2\theta\, d\phi^2]\right\} \quad [1]$$

In this model, which accounts for things on the largest length scale, the universe is approximated by a space of uniform density and pressure at each fixed time, and the expansion rate is determined by the cosmological scale factor $R(t)$ that evolves according to the Einstein equations. Astronomical observations show that the galaxies are uniform on a scale of about one billion light years, and the expansion is critical – that is, $k=0$ in [1] – and so, according to [1], on the largest scale, the universe is infinite flat Euclidian space R^3 at each fixed time. Matching the Hubble constant to its observed values, and invoking the Einstein equations, the FRW model implies that the entire infinite universe R^3 emerged all at once from a singularity ($R=0$), some 14 billion years ago, and this event is referred to as the big bang.

In this article, which summarizes the work of the authors in Smoller and Temple (1995, 2003), we describe a two-parameter family of exact solutions of the Einstein equations that refine the FRW metric by a spherical shock wave cutoff. In these exact solutions, the expanding FRW metric is reduced to a region of finite extent and finite total mass at each fixed time, and this FRW region is bounded by an entropy-satisfying shock wave that emerges from the origin (the center of the explosion), at the instant of the big bang, $t = 0$. The shock wave, which marks the leading edge of the FRW expansion, propagates outward into a larger ambient spacetime from time $t = 0$ onward. Thus, in this refinement of the FRW metric, the big bang that set the galaxies in motion is an explosion of finite mass that looks more like a classical shock wave explosion than does the big bang of the standard model. (The fact that the entire infinite space R^3 emerges at the instant of the big bang, is, loosely speaking, a consequence of the Copernican principle, the principle that the Earth is not in a special place in the universe on the largest scale of things. With a shock wave present, the Copernican principle is violated, in the sense that the Earth then has a special position relative to the shock wave. But, of course, in these shock wave refinements of the FRW metric, there is a spacetime on the other side of the shock wave, beyond the galaxies, and so the scale of uniformity of the FRW metric, the scale on which the density of the galaxies is uniform, is no longer the largest length scale.)

In order to construct a mathematically simple family of shock wave refinements of the FRW metric that meet the Einstein equations exactly, we assume $k = 0$ (critical expansion), and we restrict to the case that the sound speed in the fluid on the FRW side of the shock wave is constant. That is, we assume an FRW equation of state $p = \sigma\rho$, where σ, the square of the sound speed $\sqrt{\partial p/\partial \rho}$, is constant, $0 < \sigma \leq c^2$. At $\sigma = c^2/3$, this catches the important equation of state $p = (c^2/3)\rho$ which is correct at the earliest stage of big bang physics (Weinberg 1972). Also, as σ ranges from 0 to c^2, we obtain qualitatively correct approximations to general equations of state. Taking $c = 1$ (we use the convention that $c = 1$, and Newton's constant $\mathcal{G} = 1$ when convenient), the family of solutions is then determined by two parameters, $0 < \sigma \leq 1$ and $r_* \geq 0$. The second parameter, r_*, is the FRW radial coordinate r of the shock in the limit $t \to 0$, the instant of the big bang. (Since, when $k = 0$, the FRW metric is invariant under the rescaling $r \to \alpha r$ and $R \to \alpha^{-1}R$, we fix the radial coordinate r by fixing the scale factor α with the condition that $R(t_0) = 1$ for some time t_0, say present time.) The FRW radial coordinate r is singular with respect to radial arclength $\bar{r} = rR$ at the big bang $R = 0$, so setting $r_* > 0$ does not place the shock wave away from the origin at time $t = 0$. The distance from the FRW center to the shock wave tends to zero in the limit $t \to 0$ even when $r_* > 0$. In the limit $r_* \to \infty$, we recover from the family of solutions the usual (infinite) FRW metric with equation of state $p = \sigma\rho$ – that is, we recover the standard FRW metric in the limit that the shock wave is infinitely far out. In this sense our family of exact solutions of the Einstein equations considered here represents a two-parameter refinement of the standard FRW metric.

The exact solutions for the case $r_* = 0$ were first constructed in Smoller and Temple (1995) (see also the notes by Smoller and Temple (1999)), and are qualitatively different from the solutions when $r_* > 0$, which were constructed later in Smoller and Temple (2003). The difference is that, when $r_* = 0$, the shock wave lies closer than one Hubble length from the center of the FRW spacetime throughout its motion (Smoller and Temple 2000), but when $r_* > 0$, the shock wave emerges at the big bang at a distance beyond one Hubble length. (The Hubble length depends on time, and tends to zero as $t \to 0$.) We show in Smoller and Temple (2003) that one Hubble length, equal to c/H, where $H = \dot{R}/R$, is a critical length scale in a $k = 0$ FRW metric because the total mass inside one Hubble length has a Schwarzschild radius equal exactly to one Hubble length. (Since c/H is a good estimate for the age of the universe, it follows that the Hubble length c/H is approximately the distance of light travel starting at the big bang up until the present time. In this sense, the Hubble length is a rough estimate for the distance to the further most objects visible in the universe.) That is, one Hubble length marks precisely the distance at which the Schwarzschild radius $\bar{r}_s \equiv 2M$ of the mass M inside a radial shock wave at distance $\bar{r}$ from the FRW center, crosses from inside ($\bar{r}_s < \bar{r}$) to outside ($\bar{r}_s > \bar{r}$) the shock wave. If the shock wave is at a distance closer than one Hubble length from the FRW center, then $2M < \bar{r}$ and we say that the solution lies outside the black hole, but if the shock wave is at a distance greater than one Hubble length, then $2M > \bar{r}$ at the shock, and we say that the solution lies "inside" the black hole. Since M increases like $\bar{r}^3$, it follows that $2M < \bar{r}$ for $\bar{r}$ sufficiently small, and $2M > \bar{r}$ for $\bar{r}$ sufficiently large, so there must be a critical radius at which $2M = \bar{r}$, and we show in what follows (see also Smoller and Temple (2003)) that when $k = 0$, this critical radius is exactly the Hubble length. When the parameter $r_* = 0$, the family of solutions for $0 < \sigma \leq 1$ starts at the big bang, and evolves thereafter

"outside" the black hole, satisfying $2M/\bar{r} < 1$ everywhere from $t = 0$ onward. But, when $r_* > 0$, the shock wave is further out than one Hubble length at the instant of the big bang, and the solution begins with $2M/\bar{r} > 1$ at the shock wave. From this time onward, the spacetime expands until eventually the Hubble length catches up to the shock wave at $2M/\bar{r} = 1$, and then passes the shock wave, making $2M/\bar{r} < 1$ thereafter. Thus, when $r_* > 0$, the whole spacetime begins inside the black hole (with $2M/\bar{r} > 1$ for sufficiently large $\bar{r}$), but eventually evolves to a solution outside the black hole. The time when $\bar{r} = 2M$ actually marks the event horizon of a white hole (the time reversal of a black hole) in the ambient spacetime beyond the shock wave. We show that, when $r_* > 0$, the time when the Hubble length catches up to the shock wave comes after the time when the shock wave comes into view at the FRW center, and when $2M = \bar{r}$ (assuming t is so large that we can neglect the pressure from this time onward), the whole solution emerges from the white hole as a finite ball of mass expanding into empty space, satisfying $2M/\bar{r} < 1$ everywhere thereafter. In fact, when $r_* > 0$, the zero pressure Oppenheimer–Snyder solution outside the black hole gives the large-time asymptotics of the solution (Oppenheimer and Snyder 1939, Smoller and Temple 1988, 2004 and the comments after Theorems 6–8 below).

The exact solutions in the case $r_* = 0$ give a general-relativistic version of an explosion into a static, singular, isothermal sphere of gas, qualitatively similar to the corresponding classical explosion outside the black hole (Smoller and Temple 1995). The main difference physically between the cases $r_* > 0$ and $r_* = 0$ is that, when $r_* > 0$ (the case when the shock wave emerges from the big bang at a distance beyond one Hubble length), a large region of uniform expansion is created behind the shock wave at the instant of the big bang. Thus, when $r_* > 0$, lightlike information about the shock wave propagates inward from the wave, rather than outward from the center, as is the case when $r_* = 0$ and the shock lies inside one Hubble length. (One can imagine that when $r_* > 0$, the shock wave can get out through a great deal of matter early on when everything is dense and compressed, and still not violate the speed of light bound. Thus, when $r_* > 0$, the shock wave "thermalizes," or more accurately "makes uniform," a large region at the center, early on in the explosion.) It follows that, when $r_* > 0$, an observer positioned in the FRW spacetime inside the shock wave will see exactly what the standard model of cosmology predicts, up until the time when the shock wave comes into view in the far field. In this sense, the case $r_* > 0$ gives a black hole cosmology that refines the standard FRW model of cosmology to the case of finite mass. One of the surprising differences between the case $r_* = 0$ and the case $r_* > 0$ is that, when $r_* > 0$, the important equation of state $p = \rho/3$ comes out of the analysis as special at the big bang. When $r_* > 0$, the shock wave emerges at the instant of the big bang at a finite nonzero speed (the speed of light) only for the special value $\sigma = 1/3$. In this case, the equation of state on both sides of the shock wave tends to the correct relation $p = \rho/3$ as $t \to 0$, and the shock wave decelerates to subluminous speed for all positive times thereafter (see Smoller and Temple (2003) and Theorem 8 below).

In all cases $0 < \sigma \leq 1$, $r_* \geq 0$, the spacetime metric that lies beyond the shock wave is taken to be a metric of Tolmann–Oppenheimer–Volkoff (TOV) form (Oppenheimar and Volkoff 1939):

$$ds^2 = -B(\bar{r})d\bar{t}^2 + A^{-1}(\bar{r})d\bar{r}^2 + \bar{r}^2[d\theta^2 + \sin^2\theta\, d\phi^2] \quad [2]$$

The metric [2] is in standard Schwarzschild coordinates (diagonal with radial coordinate equal to the area of the spheres of symmetry), and the metric components depend only on the radial coordinate $\bar{r}$. Barred coordinates are used to distinguish TOV coordinates from unbarred FRW coordinates for shock matching. The mass function $M(\bar{r})$ enters as a metric component through the relation

$$A = 1 - \frac{2M(\bar{r})}{\bar{r}} \quad [3]$$

The TOV metric [2] has a very different character depending on whether $A > 0$ or $A < 0$; that is, depending on whether the solution lies outside the black hole or inside the black hole. In the case $A > 0$, $\bar{r}$ is a spacelike coordinate, and the TOV metric describes a static fluid sphere in general relativity. (When $A > 0$, for example, the metric [2] is the starting point for the stability limits of Buchdahl and Chandresekhar for stars (Weinberg 1972, Smoller and Temple 1997, 1998).) When $A < 0$, $\bar{r}$ is the timelike coordinate, and [2] is a dynamical metric that evolves in time. The exact shock wave solutions are obtained by taking $\bar{r} = R(t)r$ to match the spheres of symmetry, and then matching the metrics [1] and [2] at an interface $\bar{r} = \bar{r}(t)$ across which the metrics are Lipschitz continuous. This can be done in general. In order for the interface to be a physically meaningful shock surface, we use the result in Theorem 4 below (see Smoller and Temple (1994)) that a single additional conservation constraint is sufficient to rule out δ-function sources at the shock (the Einstein equations $G = \kappa T$ are second order in the metric, and

so δ-function sources will in general be present at a Lipschitz continuous matching of metrics), and guarantee that the matched metric solves the Einstein equations in the weak sense. The Lipschitz matching of the metrics, together with the conservation constraint, leads to a system of ordinary differential equations (ODEs) that determine the shock position, together with the TOV density and pressure at the shock. Since the TOV metric depends only on $\bar{r}$, the equations thus determine the TOV spacetime beyond the shock wave. To obtain a physically meaningful outgoing shock wave, we impose the constriant $\bar{p} \leq \bar{\rho}$ to ensure that the equation of state on the TOV side of the shock is physically reasonable, and as the entropy condition we impose the condition that the shock be compressive. For an outgoing shock wave, this is the condition $\rho > \bar{\rho}$, $p > \bar{p}$, that the pressure and density be larger on the side of the shock that receives the mass flux – the FRW side when the shock wave is propagating away from the FRW center. This condition breaks the time-reversal symmetry of the equations, and is sufficient to rule out rarefaction shocks in classical gas dynamics (Smoller 1983, Smoller and Temple 2003). The ODEs, together with the equation-of-state bound and the conservation and entropy constraints, determine a unique solution of the ODEs for every $0 < \sigma \leq 1$ and $\bar{r}_* \geq 0$, and this provides the two-parameter family of solutions discussed here (Smoller and Temple 1995, 2003). The Lipschitz matching of the metrics implies that the total mass M is continuous across the interface, and so when $r_* > 0$, the total mass of the entire solution, inside and outside the shock wave, is finite at each time $t > 0$, and both the FRW and TOV spacetimes emerge at the big bang. The total mass M on the FRW side of the shock has the meaning of total mass inside the radius $\bar{r}$ at fixed time, but on the TOV side of the shock, M does not evolve according to equations that give it the interpretation as a total mass because the metric is inside the black hole. Nevertheless, after the spacetime emerges from the black hole, the total mass takes on its usual meaning outside the black hole, and time asymptotically the big bang ends with an expansion of finite total mass in the usual sense. Thus, when $r_* > 0$, our shock wave refinement of the FRW metric leads to a big bang of finite total mass.

A final comment is in order regarding our overall philosophy. The family of exact shock wave solutions described here are rough models in the sense that the equation of state on the FRW side satisfies the condition $\sigma = \text{const.}$, and the equation of state on the TOV side is determined by the equations, and therefore cannot be imposed. Nevertheless, the bounds on the equations of state imply that the equations of state are qualitatively reasonable, and we expect that this family of solutions will capture the gross dynamics of solutions when more general equations of state are imposed. For more general equations of state, other waves, such as rarefaction waves and entropy waves, would need to be present to meet the conservation constraint, and thereby mediate the transition across the shock wave. Such transitional waves would be very difficult to model in an exact solution. But, the fact that we can find global solutions that meet our physical bounds, and that are qualitatively the same for all values of $\sigma \in (0,1]$ and all initial shock positions, strongly suggests that such a shock wave would be the dominant wave in a large class of problems.

In the next section, the FRW solution is derived for the case $\sigma = \text{const.}$, and the Hubble length is discussed as a critical length scale. Subsequently, the general theorems in Smoller and Temple (1994) for matching gravitational metrics across shock waves are employed. This is followed by a discussion of the construction of the family of solutions in the case $r_* = 0$. Finally, the case $r_* > 0$ is discussed. (Details can be found in Smoller and Temple (1995, 2003, 2004).)

The FRW Metric

According to Einstein's theory of general relativity, all properties of the gravitational field are determined by a Lorentzian spacetime metric tensor g, whose line element in a given coordinate system $x = (x^0, \ldots, x^3)$ is given by

$$ds^2 = g_{ij}dx^i dx^j \qquad [4]$$

(We use the Einstein summation convention, whereby repeated up–down indices are assumed summed from 0 to 3.) The components g_{ij} of the gravitational metric g satisfy the Einstein equations

$$G^{ij} = \kappa T^{ij}, \quad T^{ij} = (\rho c^2 + p)w^i w^j + pg^{ij} \qquad [5]$$

where we assume that the stress-energy tensor T corresponds to that of a perfect fluid. Here G is the Einstein curvature tensor,

$$\kappa = \frac{8\pi\mathcal{G}}{c^4} \qquad [6]$$

is the coupling constant, $\mathcal{G}$ is Newton's gravitational constant, c is the speed of light, ρc^2 is the energy density, p is the pressure, and $\boldsymbol{w} = (w^0, \ldots, w^3)$ are the components of the 4-velocity of the fluid (cf. Weinberg 1972), and again we use the convention that $c = 1$ and $\mathcal{G} = 1$ when convenient.

Putting the metric ansatz [1] into the Einstein equations [5] gives the equations for the FRW metric (Weinberg 1972),

$$H^2 = \left(\frac{\dot{R}}{R}\right)^2 = \frac{\kappa}{3}\rho - \frac{k}{R^2} \qquad [7]$$

and

$$\dot{\rho} = -3(p+\rho)H \qquad [8]$$

The unknown quantities R, ρ, and p are assumed to be functions of the FRW coordinate time t alone, and the "dot" denotes differentiation with respect to t.

To verify that the Hubble length $\bar{r}_{\text{crit}} = 1/H$ is the limit for FRW–TOV shock matching outside a black hole, write the FRW metric [1] in standard Schwarzschild coordinates $\boldsymbol{x} = (\bar{r}, \bar{t})$, where the metric takes the form

$$\mathrm{d}s^2 = -B(\bar{r},\bar{t})\mathrm{d}\bar{t}^2 + A(\bar{r},\bar{t})^{-1}\mathrm{d}\bar{r}^2 + \bar{r}^2\mathrm{d}\Omega^2 \qquad [9]$$

and the mass function $M(\bar{r},\bar{t})$ is defined through the relation

$$A = 1 - \frac{2M}{\bar{r}} \qquad [10]$$

It is well known that a general spherically symmetric metric can be transformed to the form [9] by coordinate transformation (see Weinberg (1972) and Groah and Temple (2004)). Substituting $\bar{r} = Rr$ into [1] and diagonalizing the resulting metric, we obtain (see Smoller and Temple (2004) for details)

$$\mathrm{d}s^2 = -\frac{1}{\psi^2}\left\{\frac{1-kr^2}{1-kr^2-H^2\bar{r}^2}\right\}\mathrm{d}\bar{t}^2 + \left\{\frac{1}{1-kr^2-H^2\bar{r}^2}\right\}\mathrm{d}\bar{r}^2 + \bar{r}^2\mathrm{d}\Omega^2 \qquad [11]$$

where ψ is an integrating factor that solves the equation

$$\frac{\partial}{\partial\bar{r}}\left(\psi\frac{1-kr^2-H^2\bar{r}^2}{1-kr^2}\right) - \frac{\partial}{\partial t}\left(\psi\frac{H\bar{r}}{1-kr^2}\right) = 0 \qquad [12]$$

and the time coordinate $\bar{t} = \bar{t}(t,\bar{r})$ is defined by the exact differential

$$\mathrm{d}\bar{t} = \left(\psi\frac{1-kr^2-H^2\bar{r}^2}{1-kr^2}\right)\mathrm{d}t + \left(\psi\frac{H\bar{r}}{1-kr^2}\right)\mathrm{d}\bar{r} \qquad [13]$$

Now using [10] in [7], it follows that

$$M(t,\bar{r}) = \frac{\kappa}{2}\int_0^{\bar{r}}\rho(t)s^2\mathrm{d}s = \frac{1}{3}\frac{\kappa}{2}\rho\bar{r}^3 \qquad [14]$$

Since in the FRW metric, $\bar{r} = Rr$ measures arclength along radial geodesics at fixed time, we see from [14] that $M(t,\bar{r})$ has the physical interpretation as the total mass inside radius $\bar{r}$ at time t in the FRW metric. Restricting to the case of critical expansion $k = 0$, we see from [7], [14], and [13] that $\bar{r} = H^{-1}$ is equivalent to $2M/\bar{r} = 1$, and so at fixed time t, the following equivalences are valid:

$$\bar{r} = H^{-1} \quad \text{iff} \quad \frac{2M}{\bar{r}} = 1 \quad \text{iff} \quad A = 0 \qquad [15]$$

We conclude that $\bar{r} = H^{-1}$ is the critical length scale for the FRW metric at fixed time t in the sense that $A = 1 - 2M/\bar{r}$ changes sign at $\bar{r} = H^{-1}$, and so the universe lies inside a black hole beyond $\bar{r} = H^{-1}$, as claimed above. Now, we proved in Smoller and Temple (1998) that the standard TOV metric outside the black hole cannot be continued into $A = 0$ except in the very special case $\rho = 0$. (It takes an infinite pressure to hold up a static configuration at the event horizon of a black hole.) Thus, shock matching beyond one Hubble length requires a metric of a different character, and for this purpose, we introduce the TOV metric inside the black hole – a metric of TOV form, with $A < 0$, whose fluid is comoving with the timelike radial coordinate $\bar{r}$ (Smoller and Temple 2004).

The Hubble length $\bar{r}_{\text{crit}} = c/H$ is also the critical distance at which the outward expansion of the FRW metric exactly cancels the inward advance of a radial light ray impinging on an observer positioned at the origin of a $k = 0$ FRW metric. Indeed, by [1], a light ray traveling radially inward toward the center of an FRW coordinate system satisfies the condition

$$c^2\,\mathrm{d}t^2 = R^2\,\mathrm{d}r^2 \qquad [16]$$

so that

$$\frac{\mathrm{d}\bar{r}}{\mathrm{d}t} = \dot{R}r + R\dot{r} = H\bar{r} - c = H\left(\bar{r} - \frac{c}{H}\right) > 0 \qquad [17]$$

if and only if

$$\bar{r} > \frac{c}{H}$$

Thus, the arclength distance from the origin to an inward moving light ray at fixed time t in a $k = 0$ FRW metric will actually increase as long as the light ray lies beyond the Hubble length. An inward moving light ray will, however, eventually cross the Hubble length and reach the origin in finite proper time, due to the increase in the Hubble length with time.

We now calculate the infinite redshift limit in terms of the Hubble length. It is well known that light emitted at (t_e, r_e) at wavelength λ_e in an FRW spacetime will be observed at (t_0, r_0) at wavelength λ_0 if

$$\frac{R_0}{R_e} = \frac{\lambda_0}{\lambda_e}$$

Moreover, the redshift factor z is defined by

$$z = \frac{\lambda_0}{\lambda_e} - 1$$

Thus, infinite redshifting occurs in the limit $R_e \to 0$, where $R = 0, t = 0$ is the big bang. Consider now a light ray emitted at the instant of the big bang, and observed at the FRW origin at present time $t = t_0$. Let r_∞ denote the FRW coordinate at time $t \to 0$ of the furthest objects that can be observed at the FRW origin before time $t = t_0$. Then r_∞ marks the position of objects at time $t = 0$ whose radiation would be observed as infinitly redshifted (assuming no scattering). Note then that a shock wave emanating from $\bar{r} = 0$ at the instant of the big bang, will be observed at the FRW origin before present time $t = t_0$ only if its position r at the instant of the big bang satisfies the condition $r < r_\infty$. To estimate r_∞, note first that from [16] it follows that an incoming radial light ray in an FRW metric follows a lightlike trajectory $r = r(t)$ if

$$r - r_e = -\int_{t_e}^{t} \frac{d\tau}{R(\tau)}$$

and thus

$$r_\infty = \int_0^{t_0} \frac{d\tau}{R(\tau)} \qquad [18]$$

Using this, the following theorem can be proved (Smoller and Temple 2004).

Theorem 1 *If the pressure p satisfies the bounds*

$$0 \le p \le \tfrac{1}{3}\rho \qquad [19]$$

then, for any equation of state, the age of the universe t_0 and the infinite red shift limit r_∞ are bounded in terms of the Hubble length by

$$\frac{1}{2H_0} \le t_0 \le \frac{2}{3H_0} \qquad [20]$$

$$\frac{1}{H_0} \le r_\infty \le \frac{2}{H_0} \qquad [21]$$

(We have assumed in Theorem 1 that $R = 0$ when $t = 0$ and $R = 1$ when $t = t_0, H = H_0$.)

The next theorem gives closed-form solutions of the FRW equations [7], [8] in the case when $\sigma = \text{const}$. As a special case, we recover the bounds in [20] and [21] from the cases $\sigma = 0$ and 1/3.

Theorem 2 *Assume $k = 0$ and the equation of state*

$$p = \sigma\rho \qquad [22]$$

where σ is taken to be constant,

$$0 \le \sigma \le 1$$

then (assuming an expanding universe $\dot{R} > 0$), the solution of system [7], [8] satisfying $R = 0$ at $t = 0$ and $R = 1$ at $t = t_0$ is given by

$$\rho = \frac{4}{3\kappa(1+\sigma)^2}\frac{1}{t^2} \qquad [23]$$

$$R = \left(\frac{t}{t_0}\right)^{2/[3(1+\sigma)]} \qquad [24]$$

$$\frac{H}{H_0} = \frac{t_0}{t} \qquad [25]$$

Moreover, the age of the universe t_0 and the infinite red shift limit r_∞ are given exactly in terms of the Hubble length by

$$t_0 = \frac{2}{3(1+\sigma)}\frac{1}{H_0} \qquad [26]$$

$$r_\infty = \frac{2}{1+3\sigma}\frac{1}{H_0} \qquad [27]$$

From [27] we conclude that a shock wave will be observed at the FRW origin before present time $t = t_0$ only if its position r at the instant of the big bang satisfies the condition

$$r < \frac{2}{1+3\sigma}\frac{1}{H_0}$$

Note that r_∞ ranges from one-half to two Hubble lengths as σ ranges from 1 to 0, taking the intermediate value of one Hubble length at $\sigma = 1/3$ (cf. [21]).

Note that using [23] and [24] in [14], it follows that

$$M = \frac{\kappa}{2}\int_0^{\bar{r}} \rho(t)s^2 ds = \frac{2\bar{r}^3}{9(1+\sigma)^2 t_0^{2/(1+\sigma)}} t^{-2\sigma/(1+\sigma)} \qquad [28]$$

so $\dot{M} < 0$ if $\sigma > 0$. It follows that if $p = \sigma\rho$, $\sigma = \text{const.} > 0$, then the total mass inside radius $r = \text{const.}$ decreases in time.

The General Theory of Shock Matching

The matching of the FRW and TOV metrics in the next two sections is based on the following theorems that were derived in Smoller and Temple (1994) (Theorems 3 and 4 apply to non-lightlike shock surfaces. The lightlike case was discussed by Scott (2002).)

Theorem 3 *Let Σ denote a smooth, three-dimensional shock surface in spacetime with spacelike*

normal vector $\boldsymbol{n}$ *relative to the spacetime metric* g; *let* K *denote the second fundamental form on* Σ; *and let* G *denote the Einstein curvature tensor. Assume that the components* g_{ij} *of the gravitational metric* g *are smooth on either side of* Σ (*continuous up to the boundary on either side separately*), *and Lipschitz continuous across* Σ *in some fixed coordinate system. Then the following statements are equivalent*:

(*i*) $[K] = 0$ *at each point of* Σ.
(*ii*) *The curvature tensors* R^i_{jkl} *and* G_{ij}, *viewed as second-order operators on the metric components* g_{ij}, *produce no* δ*-function sources on* Σ.
(*iii*) *For each point* $P \in \Sigma$, *there exists a* $C^{1,1}$ *coordinate transformation defined in a neighborhood of* P, *such that*, *in the new coordinates* (*which can be taken to be the Gaussian normal coordinates for the surface*), *the metric components are* $C^{1,1}$ *functions of these coordinates.*
(*iv*) *For each* $P \in \Sigma$, *there exists a coordinate frame that is locally Lorentzian at* P, *and can be reached within the class of* $C^{1,1}$ *coordinate transformations.*

Moreover, *if any one of these equivalencies hold*, *then the Rankine–Hugoniot jump conditions*, $[G]^\sigma_i n_\sigma = 0$ (*which express the weak form of conservation of energy and momentum across* Σ *when* $G = \kappa T$), *hold at each point on* Σ.

Here $[f]$ denotes the jump in the quantity f across Σ (this being determined by the metric separately on each side of Σ because g_{ij} is only Lipschitz continuous across Σ), and by $C^{1,1}$ we mean that the first derivatives are Lipschitz continuous.

In the case of spherical symmetry, the following stronger result holds. In this case, the jump conditions $[G^{ij}]n_i = 0$, which express the weak form of conservation across a shock surface, are implied by a single condition $[G^{ij}]n_i n_j = 0$, so long as the shock is non-null, and the areas of the spheres of symmetry match smoothly at the shock and change monotonically as the shock evolves. Note that, in general, assuming that the angular variables are identified across the shock, we expect conservation to entail two conditions, one for the time and one for the radial components. The fact that the smooth matching of the spheres of symmetry reduces conservation to one condition can be interpreted as an instance of the general principle that directions of smoothness in the metric imply directions of conservation of the sources.

Theorem 4 *Assume that* g *and* $\bar{g}$ *are two spherically symmetric metrics that match Lipschitz continuously across a three-dimensional shock interface* Σ *to form the matched metric* $g \cup \bar{g}$. *That is*, *assume that* g *and* $\bar{g}$ *are Lorentzian metrics given by*

$$ds^2 = -a(t,r)dt^2 + b(t,r)dr^2 + c(t,r)d\Omega^2 \qquad [29]$$

and

$$d\bar{s}^2 = -\bar{a}(\bar{t},\bar{r})d\bar{t}^2 + \bar{b}(\bar{t},\bar{r})d\bar{r}^2 + \bar{c}(\bar{t},\bar{r})d\Omega^2 \qquad [30]$$

and that there exists a smooth coordinate transformation $\Psi : (t,r) \to (\bar{t},\bar{r})$, *defined in a neighborhood of a shock surface* Σ *given by* $r = r(t)$, *such that the metrics agree on* Σ. (*We implicitly assume that* θ *and* φ *are continuous across the surface.*) *Assume that*

$$c(t,r) = \bar{c}(\Psi(t,r)) \qquad [31]$$

in an open neighborhood of the shock surface Σ, *so that*, *in particular*, *the areas of the 2-spheres of symmetry in the barred and unbarred metrics agree on the shock surface. Assume also that the shock surface* $r = r(t)$ *in unbarred coordinates is mapped to the surface* $\bar{r} = \bar{r}(\bar{t})$ *by* $(\bar{t},\bar{r}(\bar{t})) = \Psi(t,r(t))$. *Assume*, *finally*, *that the normal* $\boldsymbol{n}$ *to* Σ *is non-null*, *and that*

$$\boldsymbol{n}(c) \neq 0 \qquad [32]$$

where $\boldsymbol{n}(c)$ *denotes the derivative of the function* c *in the direction of the vector* $\boldsymbol{n}$. *Then the following are equivalent to the statement that the components of the metric* $g \cup \bar{g}$ *in any Gaussian normal coordinate system are* $C^{1,1}$ *functions of these coordinates across the surface* Σ:

$$[G^i_j]n_i = 0 \qquad [33]$$

$$[G^{ij}]n_i n_j = 0 \qquad [34]$$

$$[K] = 0 \qquad [35]$$

Here again, $[f] = \bar{f} - f$ *denotes the jump in the quantity* f *across* Σ, *and* K *is the second fundamental form on the shock surface.*

We assume in Theorem 4 that the areas of the 2-spheres of symmetry change monotonically in the direction normal to the surface. For example, if $c = r^2$, then $\partial c/\partial t = 0$, so the assumption $\boldsymbol{n}(c) \neq 0$ is valid except when $\boldsymbol{n} = \partial/\partial t$, in which case the rays of the shock surface would be spacelike. Thus, the shock speed would be faster than the speed of light if our assumption $\boldsymbol{n}(c) \neq 0$ failed in the case $c = r^2$.

FRW–TOV Shock Matching Outside the Black Hole – The Case $r_* = 0$

To construct the family of shock wave solutions for parameter values $0 < \sigma \leq 1$ and $r_* = 0$, we match the exact solution [23]–[25] of the FRW metric [1] to the TOV metric [2] outside the black hole,

assuming $A > 0$. In this case, we can bypass the problem of deriving and solving the ODEs for the shock surface and constraints discussed above, by actually deriving the exact solution of the Einstein equations of TOV form that meets these equations. This exact solution represents the general-relativistic version of a static, singular isothermal sphere – singular because it has an inverse square density profile, and isothermal because the relationship between the density and pressure is $\bar{p} = \bar{\sigma}\bar{\rho}$, $\bar{\sigma} = \text{const.}$

Assuming the stress tensor for a perfect fluid, and assuming that the density and pressure depend only on $\bar{r}$, the Einstein equations for the TOV metric [2] outside the black hole (i.e., when $A = 1 - 2M/\bar{r} > 0$) are equivalent to the Oppenheimer–Volkoff system

$$\frac{\mathrm{d}M}{\mathrm{d}\bar{r}} = 4\pi\bar{r}^2\bar{\rho} \qquad [36]$$

$$-\bar{r}^2\frac{\mathrm{d}}{\mathrm{d}\bar{r}}\bar{p} = \mathcal{G}M\bar{\rho}\left\{1+\frac{\bar{p}}{\bar{\rho}}\right\} \times\left\{1+\frac{4\pi\bar{r}^3\bar{p}}{M}\right\}\left\{1-\frac{2\mathcal{G}M}{\bar{r}}\right\}^{-1} \qquad [37]$$

Integrating [36], we obtain the usual interpretation of M as the total mass inside radius $\bar{r}$,

$$M(\bar{r}) = \int_0^{\bar{r}} 4\pi\xi^2\bar{\rho}(\xi)\mathrm{d}\xi \qquad [38]$$

The metric component $B \equiv B(\bar{r})$ is determined from $\bar{\rho}$ and M through the equation

$$\frac{B'(\bar{r})}{B} = -2\frac{\bar{p}'(\bar{r})}{\bar{p}+\bar{\rho}} \qquad [39]$$

Assuming

$$\bar{p} = \bar{\sigma}\bar{\rho}, \qquad \bar{\rho}(\bar{r}) = \frac{\gamma}{\bar{r}^2} \qquad [40]$$

for some constants $\bar{\sigma}$ and γ, and substituting into [3], we obtain

$$M(\bar{r}) = 4\pi\gamma\bar{r} \qquad [41]$$

Putting [40] and [41] into [37] and simplifying yields the identity

$$\gamma = \frac{1}{2\pi\mathcal{G}}\left(\frac{\bar{\sigma}}{1+6\bar{\sigma}+\bar{\sigma}^2}\right) \qquad [42]$$

From [38] we obtain

$$A = 1 - 8\pi\mathcal{G}\gamma < 1 \qquad [43]$$

Applying [39] leads to

$$B = B_0\left(\frac{\bar{\rho}}{\bar{\rho}_0}\right)^{-2\bar{\sigma}/(1+\bar{\sigma})} = B_0\left(\frac{\bar{r}}{\bar{r}_0}\right)^{4\bar{\sigma}/(1+\bar{\sigma})} \qquad [44]$$

By rescaling the time coordinate, we can take $B_0 = 1$ at $\bar{r}_0 = 1$, in which case [44] reduces to

$$B = \bar{r}^{\,4\bar{\sigma}/(1+\bar{\sigma})} \qquad [45]$$

We conclude that when [42] holds, [40]–[43] and [44] provide an exact solution of the Einstein field equations of TOV type, for each $0 \le \bar{\sigma} \le 1$. (In this case, an exact solution of TOV type was first found by Tolman (1939), and rediscovered in the case $\bar{\sigma} = 1/3$ by Misner and Zapolsky (cf. Weinberg (1972 p. 320)).) By [43], these solutions are defined outside the black hole, since $2M/\bar{r} < 1$. When $\bar{\sigma} = 1/3$, [42] yields $\gamma = 3/56\pi\mathcal{G}$ (cf. Weinberg (1972, equation (11.4.13))).

To match the FRW exact solution [23]–[25] with equation of state $p = \sigma\rho$ to the TOV exact solution [40]–[45] with equation of state $\bar{p} = \bar{\sigma}\bar{\rho}$ across a shock interface, we first set $\bar{r} = Rr$ to match the spheres of symmetry, and then match the timelike and spacelike components of the corresponding metrics in standard Schwarzschild coordinates. The matching of the $\mathrm{d}\bar{r}^2$ coefficient A^{-1} yields the conservation of mass condition that implicitly gives the shock surface $\bar{r} = \bar{r}(t)$,

$$M(\bar{r}) = \frac{4\pi}{3}\rho(t)\bar{r}^3 \qquad [46]$$

Using this together with [41] gives the following two relations that hold at the shock surface:

$$\bar{r} = \sqrt{\frac{3\gamma}{\rho(t)}}$$
$$\rho = \frac{3}{4\pi}\frac{M}{\bar{r}(t)^3} = \frac{3\gamma}{\bar{r}(t)^2} = 3\bar{\rho} \qquad [47]$$

Matching the coefficient B of $\mathrm{d}\bar{t}^2$ on the shock surface determines the integrating factor ψ in a neighborhood of the shock surface by assigning initial conditions for [44]. Finally, the conservation constraint $[T_{ij}]n_in_j = 0$ leads to the single condition

$$0 = (1-A)(\rho+\bar{p})(p+\bar{\rho})^2 + \left(1-\frac{1}{A}\right)(\bar{\rho}+\bar{p})(\rho+p)^2 + (p-\bar{p})(\rho-\bar{\rho})^2 \qquad [48]$$

which upon using $p = \sigma\rho$ and $\bar{p} = \bar{\sigma}\bar{\rho}$ is satisfied assuming the condition

$$\bar{\sigma} = \tfrac{1}{2}\sqrt{9\sigma^2+54\sigma+49} - \tfrac{3}{2}\sigma - \tfrac{7}{2} \equiv H(\sigma) \qquad [49]$$

Alternatively, we can solve for σ in [49] and write this relation as

$$\sigma = \frac{\bar{\sigma}(\bar{\sigma}+7)}{3(1-\bar{\sigma})} \qquad [50]$$

This guarantees that conservation holds across the shock surface, and so it follows from Theorem 4 that all of the equivalencies in Theorem 3 hold across the shock surface. Note that $H(0)=0$, and to leading order $\bar{\sigma}=(3\sigma/7)+O(\sigma^2)$ as $\sigma\to 0$. Within the physical region $0\le\sigma,\bar{\sigma}\le 1, H'(\sigma)>0, \bar{\sigma}<\sigma$, and $H(1/3)=\sqrt{17}-4\approx 0.1231, H(1)=\sqrt{112}/2-5\approx 0.2915$.

Using the exact formulas for the FRW metric in [23]–[25], and setting $R_0=1$ at $\rho=\rho_0, t=t_0$, we obtain the following exact formulas for the shock position:

$$\bar{r}(t)=\alpha t \qquad [51]$$

$$r(t)=\bar{r}(t)R(t)^{-1}=\beta t^{(1+3\sigma)/(3+3\sigma)} \qquad [52]$$

where

$$\alpha=3(1+\sigma)\sqrt{\frac{\bar{\sigma}}{1+6\bar{\sigma}+\bar{\sigma}^2}}$$

$$\beta=\alpha^{(1+3\sigma)/(3+3\sigma)}\left(\frac{3\gamma}{\rho_0}\right)^{1/(3+3\sigma)} \qquad [53]$$

It follows from [41] that $A>0$, and from [52] that $r_*=\lim_{t\to 0} r(t)=0$. The entropy condition that the shock wave be compressive follows from the fact that $\bar{\sigma}=H(\sigma)<\sigma$. Thus, we conclude that for each $0<\sigma\le 1, r_*=0$, the solutions constructed in [40]–[53] define a one-parameter family of shock wave solutions that evolve everywhere outside the black hole, which implies that the distance from the shock wave to the FRW center is less than one Hubble length for all $t>0$.

Using [51] and [52], one can determine the shock speed, and check when the Lax characteristic condition (Smoller 1983) holds at the shock. The result is the following theorem. (Note that even when the shock speed is larger than c, only the wave, and not the sound speeds or any other physical motion, exceeds the speed of light. See Scott (2002) for the case when the shock speed is equal to the speed of light.) The reader is referred to Smoller and Temple (1995) for details.

Theorem 5 *There exist values* $0<\sigma_1<\sigma_2<1$, $(\sigma_1\approx 0.458, \sigma_2=\sqrt{5}/3\approx 0.745)$, *such that, for* $0<\sigma\le 1$, *the Lax characteristic condition holds at the shock if and only if* $0<\sigma<\sigma_1$; *and the shock speed is less than the speed of light if and only if* $0<\sigma<\sigma_2$.

The explicit solution in the case $r_*=0$ can be interpreted as a general-relativistic version of a shock wave explosion into a static, singular, isothermal sphere, known in the Newtonian case as a simple model for star formation (Smoller and Temple 2000). As the scenario goes, a star begins as a diffuse cloud of gas. The cloud slowly contracts under its own gravitational force by radiating energy out through the gas cloud as gravitational potential energy is converted into kinetic energy. This contraction continues until the gas cloud reaches the point where the mean free path for transmission of light is small enough that light is scattered, instead of being transmitted, through the cloud. The scattering of light within the gas cloud has the effect of equalizing the temperature within the cloud, and at this point the gas begins to drift toward the most compact configuration of the density that balances the pressure when the equation of state is isothermal. This configuration is a static, singular, isothermal sphere, the general-relativistic version of which is the exact TOV solution beyond the shock wave when $r_*=0$. This solution in the Newtonian case is also inverse square in the density and pressure, and so the density tends to infinity at the center of the sphere. Eventually, the high densities at the center ignite thermonuclear reactions. The result is a shock wave explosion emanating from the center of the sphere, and this signifies the birth of the star. The exact solutions when $r_*=0$ represent a general-relativistic version of such a shock wave explosion.

Shock Wave Solutions Inside the Black Hole – The Case $r_*>0$

When the shock wave is beyond one Hubble length from the FRW center, we obtain a family of shock wave solutions for each $0<\sigma\le 1$ and $r_*>0$ by shock matching the FRW metric [1] to a TOV metric of form [2] under the assumption that

$$A(\bar{r})=1-\frac{2M(\bar{r})}{\bar{r}}\equiv 1-N(\bar{r})<0 \qquad [54]$$

In this case, $\bar{r}$ is the timelike variable. Assuming that the stress tensor T is taken to be that of a perfect fluid comoving with the TOV metric, the Einstein equations $G=\kappa T$, inside the black hole, take the form (see Smoller and Temple (2004) for details)

$$\bar{p}'=\frac{\bar{p}+\bar{\rho}}{2}\frac{N'}{N-1} \qquad [55]$$

$$N'=-\left\{\frac{N}{\bar{r}}+\kappa\bar{p}\bar{r}\right\} \qquad [56]$$

$$\frac{B'}{B}=-\frac{1}{N-1}\left\{\frac{N}{\bar{r}}+\kappa\bar{\rho}\right\} \qquad [57]$$

The system [55]–[57] defines the simplest class of gravitational metrics that contain matter, evolve inside the black hole, and such that the mass function $M(\bar{r}) < \infty$ at each fixed time $\bar{r}$. System [55]–[57] for $A < 0$ differs substantially from the TOV equations for $A > 0$ because, for example, the energy density T^{00} is equated with the timelike component G^{rr} when $A < 0$, but with G^{tt} when $A > 0$. In particular, this implies that, inside the black hole, the mass function $M(\bar{r})$ does not have the interpretation as a total mass inside the radius $\bar{r}$ as it does outside the black hole.

Equations [56], [57] do not have the same character as [54], [55] and the relation $\bar{p} = \bar{\sigma}\bar{\rho}$ with $\bar{\sigma} = \text{const.}$ is inconsistent with [56], [57] together with the conservation constraint and the FRW assumption $p = \sigma\rho$ for shock matching. Thus, instead of looking for an exact solution of [56], [57] ahead of time, as in the case $r_* = 0$, we assume the FRW solution [23]–[25], and derive the ODEs that describe the TOV metrics that match this FRW metric Lipschitz-continuously across a shock surface, and then impose the conservation, entropy, and equation of state constraints at the end. Matching a given $k = 0$ FRW metric to a TOV metric inside the black hole across a shock interface leads to the system of ODEs, (see Smoller and Temple (2004) for details),

$$\frac{du}{dN} = -\left\{\frac{(1+u)}{2(1+3u)N}\right\} \times \left\{\frac{(3u-1)(\sigma-u)N + 6u(1+u)}{(\sigma-u)N + (1+u)}\right\} \quad [58]$$

$$\frac{d\bar{r}}{dN} = -\frac{1}{1+3u}\frac{\bar{r}}{N} \quad [59]$$

with conservation constraint

$$v = \frac{-\sigma(1+u) + (\sigma-u)N}{(1+u) + (\sigma-u)N} \quad [60]$$

where

$$u = \frac{\bar{p}}{\rho}, \quad v = \frac{\bar{\rho}}{\rho}, \quad \sigma = \frac{p}{\rho} \quad [61]$$

Here ρ and p denote the (known) FRW density and pressure, and all variables are evaluated at the shock. Solutions of [58]–[60] determine the (unknown) TOV metrics that match the given FRW metric Lipschitz-continuously across a shock interface, such that conservation of energy and momentum hold across the shock, and such that there are no δ-function sources at the shock (Israel 1966, Smoller and Temple 1997). Note that the dependence of [58]–[60] on the FRW metric is only through the variable σ, and so the advantage of taking $\sigma = \text{const.}$ is that the whole solution is determined by the inhomogeneous scalar equation [58] when $\sigma = \text{const.}$ We take as the entropy constraint the condition that

$$0 < \bar{p} < p, \qquad 0 < \bar{\rho} < \rho \quad [62]$$

and to insure a physically reasonable solution, we impose the equation of state constriant on the TOV side of the shock (this is equivalent to the dominant energy condition (Blau and Guth 1987))

$$0 < \bar{p} < \bar{\rho} \quad [63]$$

Condition [62] implies that outgoing shock waves are compressive. Inequalities [62] and [63] are both implied by the single condition (Smoller and Temple 2004),

$$\frac{1}{N} < \left(\frac{1-u}{1+u}\right)\left(\frac{\sigma-u}{\sigma+u}\right) \quad [64]$$

Since σ is constant, eqn [58] uncouples from [59], and thus solutions of system [58]–[60] are determined by the scalar nonautonomous equation [58]. Making the change of variable $S = 1/N$, which transforms the "big bang" $N \to \infty$ over to a rest point at $S \to 0$, we obtain

$$\frac{du}{dS} = \left\{\frac{(1+u)}{2(1+3u)S}\right\} \times \left\{\frac{(3u-1)(\sigma-u) + 6u(1+u)S}{(\sigma-u) + (1+u)S}\right\} \quad [65]$$

Note that the conditions $N > 1$ and $0 < \bar{p} < p$ restrict the domain of [65] to the region $0 < u < \sigma < 1, 0 < S < 1$. The next theorem gives the existence of solutions for $0 < \sigma \leq 1, r_* > 0$, inside the black hole (Smoller and Temple 2003).

Theorem 6 *For every σ, $0 < \sigma < 1$, there exists a unique solution $u_\sigma(S)$ of [65], such that [64] holds on the solution for all $S, 0 < S < 1$, and on this solution, $0 < u_\sigma(S) < \bar{u}$, $\lim_{S\to 0} u_\sigma(S) = \bar{u}$, where*

$$\bar{u} = \text{Min}\{1/3, \sigma\} \quad [66]$$

and

$$\lim_{S\to 1} \bar{p} = 0 = \lim_{S\to 1} \bar{\rho} \quad [67]$$

For each of these solutions $u_\sigma(S)$, the shock position is determined by the solution of [59], which in turn is determined uniquely by an initial condition which can be taken to be the FRW radial position of the shock wave at the instant of the big bang,

$$r_* = \lim_{S\to 0} r(S) > 0 \quad [68]$$

Concerning the shock speed, we have

Theorem 7 *Let $0 < \sigma < 1$. Then the shock wave is everywhere subluminous, that is, the shock speed $s_\sigma(S) \equiv s(u_\sigma(S)) < 1$ for all $0 < S \leq 1$, if and only if $\sigma \leq 1/3$.*

Concerning the shock speed near the big bang $S = 0$, the following is true:

Theorem 8 *The shock speed at the big bang $S = 0$ is given by*

$$\lim_{S\to 0} s_\sigma(S) = 0, \quad \sigma < 1/3 \qquad [69]$$

$$\lim_{S\to 0} s_\sigma(S) = \infty, \quad \sigma > 1/3 \qquad [70]$$

$$\lim_{S\to 0} s_\sigma(S) = 1, \quad \sigma = 1/3 \qquad [71]$$

Theorem 8 shows that the equation of state $p = \rho/3$ plays a special role in the analysis when $r_* > 0$, and only for this equation of state does the shock wave emerge at the big bang at a finite nonzero speed, the speed of light. Moreover, [66] implies that in this case, the correct relation $\bar{p}/\bar{\rho} = \bar{\sigma}$ is also achieved in the limit $S \to 0$. The result [67] implies that (neglecting the pressure p at this time onward), the solution continues to a $k = 0$ Oppenheimer–Snyder solution outside the black hole for $S > 1$.

It follows that the shock wave will first become visible at the FRW center $\bar{r} = 0$ at the moment $t = t_0, (R(t_0) = 1)$, when the Hubble length $H_0^{-1} = H^{-1}(t_0)$ satisfies

$$\frac{1}{H_0} = \frac{1+3\sigma}{2} r_* \qquad [72]$$

where r_* is the FRW position of the shock at the instant of the big bang. At this time, the number of Hubble lengths $\sqrt{N_0}$ from the FRW center to the shock wave at time $t = t_0$ can be estimated by

$$1 \leq \frac{2}{1+3\sigma} \leq \sqrt{N_0} \leq \frac{2}{1+3\sigma} e^{\sqrt{3\sigma}((1+3\sigma)/(1+\sigma))}$$

Thus, in particular, the shock wave will still lie beyond the Hubble length $1/H_0$ at the FRW time t_0 when it first becomes visible. Furthermore, the time $t_{\text{crit}} > t_0$ at which the shock wave will emerge from the white hole given that t_0 is the first instant at which the shock becomes visible at the FRW center, can be estimated by

$$\frac{2}{1+3\sigma} e^{\sigma/4} \leq \frac{t_{\text{crit}}}{t_0} \leq \frac{2}{1+3\sigma} e^{2\sqrt{3\sigma}/(1+\sigma)} \qquad [73]$$

for $0 < \sigma \leq 1/3$, and by the better estimate

$$e^{\sqrt{6}/4} \leq \frac{t_{\text{crit}}}{t_0} \leq e^{3/2} \qquad [74]$$

in the case $\sigma = 1/3$. Inequalities [73], [74] imply, for example, that at the Oppenheimer–Snyder limit $\sigma = 0$,

$$\sqrt{N_0} = 2, \qquad \frac{t_{\text{crit}}}{t_0} = 2$$

and in the limit $\sigma = 1/3$,

$$1.8 \leq \frac{t_{\text{crit}}}{t_0} \leq 4.5, \qquad 1 < \sqrt{N_0} \leq 4.5$$

We can conclude that at the moment t_0 when the shock wave first becomes visible at the FRW center, the shock wave must lie within 4.5 Hubble lengths of the FRW center. Throughout the expansion up until this time, the expanding universe must lie entirely within a white hole – the universe will eventually emerge from this white hole, but not until some later time t_{crit}, where t_{crit} does not exceed $4.5t_0$.

Conclusion

We believe that the existence of a wave at the leading edge of the expansion of the galaxies is the most likely possibility. The alternatives are that either the universe of expanding galaxies goes on out to infinity, or else the universe is not simply connected. Although the first possibility has been believed for most of the history of cosmology based on the Friedmann universe, we find this implausible and arbitrary in light of the shock wave refinements of the FRW metric discussed here. The second possibility, that the universe is not simply connected, has received considerable attention recently (Klarreich 2003). However, since we have not seen, and cannot create, any non-simply-connected 3-spaces on any other length scale, and since there is no observational evidence to support this, we view this as less likely than the existence of a wave at the leading edge of the expansion of the galaxies, left over from the big bang. Recent analysis of the microwave background radiation data shows a cutoff in the angular frequencies consistent with a length scale of around one Hubble length (Andy Abrecht, private communication). This certainly makes one wonder whether this cutoff is evidence of a wave at this length scale, especially given the consistency of this possibility with the case $r_* > 0$ of the family of exact solutions discussed here.

Acknowledgments

The work of JS was supported in part by NSF Applied Mathematics Grant Number DMS-010-3998, and that of BT by NSF Applied Mathematics Grant Number DMS-010-2493.

See also: Black Hole Mechanics; Cosmology: Mathematical Aspects; Newtonian Limit of General Relativity; Symmetric Hyperbolic Systems and Shock Waves.

Further Reading

Blau SK and Guth AH (1987) Inflationary cosmology. In: Hawking SW and Israel W (eds.) *Three Hundred Years of Gravitation*, pp. 524–603. Cambridge: Cambridge University Press.

Groah J, Smoller J, and Temple B (2003) Solving the Einstein equations by Lipschitz continuous metrics: shockwaves in general relativity. In: Friedlander S and Serre D (eds.) *Handbook of Mathematical Fluid Dynamics*, vol. 2, pp. 501–597. Amsterdam: North Holland.

Groah J and Temple B (2004) Shock-Wave Solutions of the Einstein Equations: Existence and Consistency by a Locally Inertial Glimm Scheme. *Memoirs of the AMS*, 84 pages, vol. 172, no. B13.

Israel W (1966) Singular hypersurfaces and thin shells in general relativity. *IL Nuovo Cimento* XLIV B(1): 1–14.

Klarreich E (2003) The shape of space. *Science News* 164–165.

Oppenheimer JR and Snyder JR (1939) On continued gravitational contraction. *Physical Review* 56: 455–459.

Scott M (2002) *General Relativistic Shock Waves Propagating at the Speed of Light*. Ph.D. thesis, UC-Davis.

Smoller J (1983) *Shock Waves and Reaction Diffusion Equations*. New York, Berlin: Springer.

Smoller J and Temple B (1994) Shock-wave solutions of the Einstein equations: the Oppenheimer–Snyder model of gravitational collapse extended to the case of non-zero pressure. *Archives for Rational and Mechanical Analysis* 128: 249–297.

Smoller J and Temple B (1995) Astrophysical shock-wave solutions of the Einstein equations. *Physical Review D* 51(6): 2733–2743.

Smoller J and Temple B (1997) Solutions of the Oppenheimer–Volkoff equations inside 9/8'ths of the Schwarzschild radius. *Communications in Mathematical Physics* 184: 597–617.

Smoller J and Temple B (1998) On the Oppenheimer–Volkov equations in general relativity. *Archives for Rational and Mechanical Analysis* 142: 177–191.

Smoller J and Temple B (2000) Cosmology with a shock wave. *Communications in Mathematical Physics* 210: 275–308.

Smoller J and Temple B (2003) Shock wave cosmology inside a black hole. *Proceedings of the National Academy of Sciences of the United States of America* 100(20): 11216–11218.

Smoller J and Temple B (2004) Cosmology, black holes, and shock waves beyond the Hubble length. *Methods and Applications of Analysis* 11(1): 77–132.

Tolman R (1939) Static solutions of Einstein's field equations for spheres of fluid. *Physical Review* 55: 364–374.

Weinberg S (1972) *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*. New York: Wiley.

Shock Waves *see* Symmetric Hyperbolic Systems and Shock Waves

Short-Range Spin Glasses: The Metastate Approach

C M Newman, New York University, New York, NY, USA
D L Stein, University of Arizona, Tucson, AZ, USA

Introduction

The nature of the low-temperature spin glass phase in short-range models remains one of the central problems in the statistical mechanics of disordered systems (Binder and Young 1986, Chowdhury 1986, Mézard *et al.* 1987, Stein 1989, Fischer and Hertz 1991, Dotsenko 2001, Newman and Stein 2003). While many of the basic questions remain unanswered, analytical and rigorous work over the past decade have greatly streamlined the number of possible scenarios for pure state structure and organization at low temperatures, and have clarified the thermodynamic behavior of these systems.

The unifying concept behind this work is that of the "metastate." It arose independently in two different constructions (Aizenman and Wehr 1990, Newman and Stein 1996b), which were later shown to be equivalent (Newman and Stein 1998a). The metastate is a probability measure on the space of all thermodynamic states. Its usefulness arises in situations where multiple "competing" pure states may be present. In such situations it may be difficult to construct individual states in a measurable and canonical way; the metastate avoids this difficulty by focusing instead on the statistical properties of the states.

An important aspect of the metastate approach is that it relates, by its very construction (Newman and Stein 1996b), the observed behavior of a system in large but finite volumes with its thermodynamic properties. It therefore serves as a (possibly indispensable) tool for analyzing and understanding both the infinite-volume and finite-volume properties of a system, particularly in cases where a straightforward interpolation between the two may be incorrect, or their relation otherwise difficult to analyze.

We will focus on the Edwards–Anderson (EA) Ising spin glass model (Edwards and Anderson

1975), although most of our discussion is relevant to a much larger class of realistic models. The EA model is described by the Hamiltonian

$$\mathcal{H}_{\mathcal{J}} = -\sum_{\langle x,y\rangle} J_{xy}\sigma_x\sigma_y \qquad [1]$$

where $\mathcal{J}$ denotes a particular realization of all of the couplings J_{xy} and the brackets indicate that the sum is over nearest-neighbor pairs only, with $x, y \in \mathbf{Z}^d$. We will take Ising spins $\sigma_x = \pm 1$; although this will affect the details of our discussion, it is unimportant for our main conclusions. The couplings J_{xy} are quenched, independent, identically distributed random variables whose common distribution μ is symmetric about zero.

States and Metastates

We are interested in both finite-volume and infinite-volume Gibbs states. For the cube of length scale L, $\Lambda_L = \{-L, -L+1, \ldots, L\}^d$, we define $\mathcal{H}_{\mathcal{J},L}$ to be the restriction of the EA Hamiltonian to Λ_L with a specified boundary condition such as free, fixed, or periodic. Then the finite-volume Gibbs distribution $\rho^{(L)}_{\mathcal{J}} = \rho^{(L)}_{\mathcal{J},\beta}$ on Λ_L (at inverse temperature $\beta = 1/T$) is

$$\rho^{(L)}_{\mathcal{J},\beta}(\sigma) = Z_L^{-1} \exp\{-\beta\mathcal{H}_{\mathcal{J},L}(\sigma)\} \qquad [2]$$

where the partition function $Z_L(\beta)$ is such that the sum of $\rho^{(L)}_{\mathcal{J},\beta}$ over all σ yields 1. (In this and all succeeding definitions, the dependence on spatial dimension d will be suppressed.)

Thermodynamic states are described by infinite-volume Gibbs measures. At fixed inverse temperature β and coupling realization $\mathcal{J}$, a thermodynamic state $\rho_{\mathcal{J},\beta}$ is the limit, as $L \to \infty$, of some sequence of such finite-volume measures (each with a specified boundary condition, which may remain the same or may change with L). A thermodynamic state $\rho_{\mathcal{J},\beta}$ can also be characterized intrinsically through the Dobrushin–Lanford–Ruelle (DLR) equations (see, e.g., Georgii 1988): for any Λ_L, the conditional distribution of $\rho_{\mathcal{J},\beta}$ (conditioned on the sigma-field generated by $\{\sigma_x : x \in \mathbf{Z}^d \backslash \Lambda_L\}$ is $\rho^{(L),\overline{\sigma}}_{\mathcal{J},\beta}$, where $\overline{\sigma}$ is given by the conditioned values of σ_x for x on the boundary of Λ_L.

Consider now the set $\mathcal{G} = \mathcal{G}(\mathcal{J},\beta)$ of all thermodynamic states at a fixed $(\mathcal{J},\beta)$. The set of extremal, or pure, Gibbs states is defined by

$$\begin{aligned}\text{ex } \mathcal{G} = \mathcal{G} \setminus \{a\rho_1 + (1-a)\rho_2 :\\ a \in (0,1);\ \rho_1, \rho_2 \in \mathcal{G};\ \rho_1 \neq \rho_2\}\end{aligned} \qquad [3]$$

and the number of pure states $\mathcal{N}(\mathcal{J},\beta)$ at $(\mathcal{J},\beta)$ is the cardinality $|\text{ex } \mathcal{G}|$ of ex $\mathcal{G}$. It is not hard to show that, in any d and for a.e. $\mathcal{J}$, the following two statements are true: (1) $\mathcal{N} = 1$ at sufficiently low $\beta > 0$; (2) at any fixed β, $\mathcal{N}$ is constant a.s. with respect to the $\mathcal{J}$'s. (The last assertion follows from the measurability and translation invariance of $\mathcal{N}$, and the translation ergodicity of the disorder distribution of $\mathcal{J}$.)

A pure state ρ_α (where α is a pure-state index) can also be intrinsically characterized by a "clustering property"; for two-point correlation functions, this reads

$$\langle\sigma_x\sigma_y\rangle_{\rho_\alpha} - \langle\sigma_x\rangle_{\rho_\alpha}\langle\sigma_y\rangle_{\rho_\alpha} \to 0 \qquad [4]$$

as $|x-y| \to \infty$. A simple observation (Newman and Stein 1992), with important consequences for spin glasses, is that if many pure states exist, a sequence of $\rho^{(L)}_{\mathcal{J},\beta}$'s, with boundary conditions and L's chosen independently of $\mathcal{J}$, will generally not have a (single) limit. We call this phenomenon "chaotic size dependence" (CSD).

We will be interested in the properties of ex $\mathcal{G}$ at low temperatures. If the spin-flip symmetry present in the EA Hamiltonian equation [1] is spontaneously broken above some dimension d_0 and below some temperature $T_c(d)$, there will be at least a pair of pure states such that their even-spin correlations are identical and their odd-spin correlations have the opposite sign. Assuming that such broken spin-flip symmetry indeed exists for $d > d_0$ and $T < T_c(d)$, the question of whether there exists more than one such pair (of spin-flip related extremal infinite-volume Gibbs distributions) is a central unresolved issue for the EA and related models. If many such pairs should exist, we can ask about the structure of their relations with one another, and how this structure would manifest itself in large but finite volumes. To do this, we use an approach, introduced by Newman and Stein (1996b), to study inhomogeneous and other systems with many competing pure states. This approach, based on an analogy with chaotic dynamical systems, requires the construction of a new thermodynamic quantity which is called the "metastate" – a probability measure $\kappa_{\mathcal{J}}$ on the thermodynamic states. The metastate allows an understanding of CSD by analyzing the way in which $\rho^{(L)}_{\mathcal{J},\beta}$ "samples" from its various possible limits as $L \to \infty$.

The analogy with chaotic dynamical systems can be understood as follows. In dynamical systems, the chaotic motion along a deterministic orbit is analyzed in terms of some appropriately selected probability measure, invariant under the dynamics. Time along the orbit is replaced, in our context, by L and the phase space of the dynamical system is replaced by the space of Gibbs states.

Newman and Stein (1996b) considered a "microcanonical ensemble" (as always, at fixed β, which will hereafter be suppressed for ease of notation) κ_N in which each of the finite-volume Gibbs states

$\rho_{\mathcal{J}}^{(L_1)}, \rho_{\mathcal{J}}^{(L_2)}, \ldots, \rho_{\mathcal{J}}^{(L_N)}$ has weight N^{-1}. The ensemble κ_N converges to a metastate $\kappa_{\mathcal{J}}$ as $N \to \infty$, in the following sense: for every (nice) function g on states (e.g., a function of finitely many correlations),

$$\lim_{N\to\infty} N^{-1} \sum_{\ell=1}^{N} g(\rho^{(L_\ell)}) = \int g(\Gamma)\, \mathrm{d}\kappa_{\mathcal{J}}(\Gamma) \qquad [5]$$

The information contained in $\kappa_{\mathcal{J}}$ effectively specifies the fraction of cube sizes L_ℓ which the system spends in different (possibly mixed) thermodynamic states Γ as $\ell \to \infty$.

A different, but in the end equivalent, approach based on $\mathcal{J}$-randomness is due to Aizenman and Wehr (1990). Here one considers the random pair $(\mathcal{J}, \rho_{\mathcal{J}}^{(L)})$, defined on the underlying probability space of $\mathcal{J}$, and takes the limit $\kappa^\dagger$ (with conditional distribution $\kappa_{\mathcal{J}}^\dagger$, given $\mathcal{J}$), via finite-dimensional distributions along some subsequence. The details are omitted here, and the reader is referred to the work by Aizenman and Wehr (1990) and Newman and Stein (1998a). We note, however, the important result that a "deterministic" subsequence of volumes can be found on which [5] is valid and also $(\mathcal{J}, \rho_{\mathcal{J}}^{(L)})$ converges, with $\kappa_{\mathcal{J}}^\dagger = \kappa_{\mathcal{J}}$ (Newman and Stein 1998a).

In what follows we use the term "metastate" as shorthand for the $\kappa_{\mathcal{J}}$ constructed using periodic boundary conditions on a sequence of volumes chosen independently of the couplings, and along which $\kappa_{\mathcal{J}} = \kappa_{\mathcal{J}}^\dagger$. We choose periodic boundary conditions for specificity; the results and claims discussed are expected to be independent of the boundary conditions used, as long as they are chosen independently of the couplings.

Low-Temperature Structure of the EA Model

There have been several scenarios proposed for the spin-glass phase of the Edwards–Anderson model at sufficiently low temperature and high dimension. These remain speculative, because it has not even been proved that a phase transition from the high-temperature phase exists at positive temperature in any finite dimension.

As noted earlier, at sufficiently high temperature in any dimension (and at all nonzero temperatures in one and presumably two dimensions, although the latter assertion has not been proved), there is a unique Gibbs state. It is conceivable that this remains the case in all dimensions and at all nonzero temperatures, in which case the metastate $\kappa_{\mathcal{J}}$ is, for a.e. $\mathcal{J}$, supported on a single, pure Gibbs state $\rho_{\mathcal{J}}$. (It is important to note, however, that in principle such a trivial metastate could occur even if $\mathcal{N} > 1$; indeed, just such a situation of "weak uniqueness" (van Enter and Fröhlich 1985, Campanino *et al.* 1987) happens in very long range spin glasses at high temperatures (Fröhlich and Zegarlinski 1987, Gandolfi *et al.* 1993).)

A phase transition has been proved to exist (Aizenman *et al.* 1987) in the Sherrington–Kirkpatrick (SK) model (Sherrington and Kirkpatrick 1975), which is the infinite-range version of the EA model. Numerical (Ogielski 1985, Ogielski and Morgenstern 1985, Binder and Young 1986, Kawashima and Young 1996) and some analytical (Fisher and Singh 1990, Thill and Hilhorst 1996) work has led to a general consensus that above some dimension (typically around three or four) there does exist a positive-temperature phase transition below which spin-flip symmetry is broken, that is, in which pure states come in pairs, as discussed below eqn [4]. Because much of the literature has focused on this possibility, we assume it in what follows, and the metastate approach turns out to be highly useful in restricting the scenarios that can occur. The simplest such scenario is a two-state picture in which, below the transition temperature T_c, there exists a single pair of global flip-related pure states $\rho_{\mathcal{J}}^{\alpha}$ and $\rho_{\mathcal{J}}^{-\alpha}$. In this case, there is no CSD for periodic boundary conditions and the metastate can be written as

$$\kappa_{\mathcal{J}} = \delta_{\frac{1}{2}\rho_{\mathcal{J}}^{\alpha} + \frac{1}{2}\rho_{\mathcal{J}}^{-\alpha}} \qquad [6]$$

That is, the metastate is supported on a single (mixed) thermodynamic state.

The two-state scenario that has received the most attention in the literature is the "droplet/scaling" picture (McMillan 1984, Fisher and Huse 1986, 1988, Bray and Moore 1985). In this picture a low-energy excitation above the ground state in Λ_L is a droplet whose surface area scales as l^{d_s}, with $l \sim O(L)$ and $d_s < d$, and whose surface energy scales as l^θ, with $\theta > 0$ (in dimensions where $T_c > 0$). More recently, an alternative picture has arisen (Krzakala and Martin 2000, Palassini and Young 2000) in which the low-energy excitations differ from those of droplet/scaling, in that their energies scale as $l^{\theta'}$, with $\theta' = 0$.

The low-temperature picture that has perhaps generated the most attention in the literature is the replica symmetry breaking (RSB) scenario (Binder and Young 1986, Marinari *et al.* 1994, 1997, Franz *et al.* 1998, Marinari *et al.* 2000, Marinari and Parisi 2000, 2001, Dotsenko 2001), which assumes a rather complicated pure-state structure, inspired by Parisi's solution of the SK model (Parisi 1979, 1983, Mézard *et al.* 1984, 1987). This is a many-state picture ($\mathcal{N} = \infty$ for a.e.

$\mathcal{J}$) in which the ordering is described in terms of the "overlaps" between states. There has been some ambiguity in how to describe such a picture for short-range models; the prevailing, or standard, view. Consider any reasonably constructed thermodynamic state $\rho_{\mathcal{J}}$ (see Newman and Stein (1998a) for more details) – e.g., the "average" over the metastate $\kappa_{\mathcal{J}}$

$$\rho_{\mathcal{J}} = \int \Gamma \, d\kappa_{\mathcal{J}}(\Gamma) \qquad [7]$$

Now choose σ and σ' from the product distribution $\rho_{\mathcal{J}}(\sigma)\rho_{\mathcal{J}}(\sigma')$. The overlap Q is defined as

$$Q = \lim_{L\to\infty} |\Lambda_L|^{-1} \sum_{x\in\Lambda_L} \sigma_x \sigma'_x \qquad [8]$$

and $P_{\mathcal{J}}(q)$ is defined to be its probability distribution.

In the standard RSB picture, $\rho_{\mathcal{J}}$ is a mixture of infinitely many pure states, each with a specific $\mathcal{J}$-dependent weight W:

$$\rho_{\mathcal{J}}(\sigma) = \sum_{\alpha} W^{\alpha}_{\mathcal{J}} \rho^{\alpha}_{\mathcal{J}}(\sigma) \qquad [9]$$

If σ is drawn from $\rho^{\alpha}_{\mathcal{J}}$ and σ' from $\rho^{\beta}_{\mathcal{J}}$, then the expression in eqn [8] equals its thermal mean,

$$q^{\alpha\beta}_{\mathcal{J}} = \lim_{L\to\infty} |\Lambda_L|^{-1} \sum_{x\in\Lambda_L} \langle\sigma_x\rangle_{\alpha}\langle\sigma_x\rangle_{\beta} \qquad [10]$$

and hence $P_{\mathcal{J}}$ is given by

$$P_{\mathcal{J}}(q) = \sum_{\alpha,\beta} W^{\alpha}_{\mathcal{J}} W^{\beta}_{\mathcal{J}} \delta(q - q^{\alpha\beta}_{\mathcal{J}}) \qquad [11]$$

The "self-overlap," or EA order parameter, is given by $q_{\mathrm{EA}} = q^{\alpha\alpha}_{\mathcal{J}}$ and (at fixed T) is thought to be independent of both α and $\mathcal{J}$ (with probability 1).

According to the standard RSB scenario, the $W^{\alpha}_{\mathcal{J}}$'s and $q^{\alpha\beta}_{\mathcal{J}}$'s are non-self-averaging (i.e., $\mathcal{J}$-dependent) quantities, except for $\alpha = \beta$ or its global flip, where $q^{\alpha\beta}_{\mathcal{J}} = \pm q_{\mathrm{EA}}$. The average $P_s(q)$ of $P_{\mathcal{J}}(q)$ over the disorder distribution of $\mathcal{J}$ is predicted to be a mixture of two delta-function components at $\pm q_{\mathrm{EA}}$ and a continuous part between them. However, it was proved by Newman and Stein (1996c) that this scenario cannot occur, because of the translation invariance of $P_{\mathcal{J}}(q)$ and the translation ergodicity of the disorder distribution. Nevertheless, the metastate approach suggests an alternative, nonstandard, RSB scenario, which is described next.

The idea behind the nonstandard RSB picture (referred to by us as the nonstandard SK picture in earlier papers) is to produce the finite-volume behavior of the SK model to the maximum extent possible. We therefore assume in this picture that in each Λ_L, the finite-volume Gibbs state $\rho^{(L)}_{\mathcal{J}}$ is well approximated deep in the interior by a mixed thermodynamic state $\Gamma^{(L)}$, decomposable into many pure states ρ_{α_L} (explicit dependence on $\mathcal{J}$ is suppressed for ease of notation). More precisely, each Γ in $\kappa_{\mathcal{J}}$ satisfies

$$\Gamma = \sum_{\alpha_\Gamma} W^{\alpha_\Gamma}_{\Gamma} \rho_{\alpha_\Gamma} \qquad [12]$$

and is presumed to have a nontrivial overlap distribution for σ, σ' from $\Gamma(\sigma)\Gamma(\sigma')$:

$$P_\Gamma(q) = \sum_{\alpha_\Gamma,\beta_\Gamma} W^{\alpha_\Gamma}_{\Gamma} W^{\beta_\Gamma}_{\Gamma} \delta(q - q_{\alpha_\Gamma\beta_\Gamma}) \qquad [13]$$

as did $\rho_{\mathcal{J}}$ in the standard RSB picture.

Because $\kappa_{\mathcal{J}}$, like its counterpart $\rho_{\mathcal{J}}$ in the standard SK picture, is translation covariant, the resulting ensemble of overlap distributions P_Γ is independent of $\mathcal{J}$. Because of the CSD present in this scenario, the overlap distribution for $\rho^{(L)}_{\mathcal{J}}$ varies with L, no matter how large L becomes. So, instead of averaging the overlap distribution over $\mathcal{J}$, the averaging must now be done over the states Γ within the metastate $\kappa_{\mathcal{J}}$, all at fixed $\mathcal{J}$:

$$P_{\mathrm{ns}}(q) = \int P_\Gamma(q)\kappa_{\mathcal{J}}(\Gamma)d\Gamma \qquad [14]$$

The $P_{\mathrm{ns}}(q)$ is the same for a.e. $\mathcal{J}$, and has a form analogous to the $P_{\mathrm{s}}(q)$ in the standard RSB picture.

However, the nonstandard RSB scenario seems rather unlikely to occur in any natural setting, because of the following result:

Theorem Newman and Stein 1998b). (*Consider two metastates constructed along (the same) deterministic sequence of Λ_L's, using two different sequences of flip-related, coupling-independent boundary conditions (such as periodic and antiperiodic). Then with probability one, these two metastates are the same.*

The proof is given by Newman and Stein (1998b), but the essential idea can be easily described here. As discussed earlier, $\kappa_{\mathcal{J}} = \kappa^{\dagger}_{\mathcal{J}}$; but $\kappa^{\dagger}_{\mathcal{J}}$ is constructed by a limit of finite-dimensional distributions, which means averaging over other couplings including the ones near the system boundary, and hence gives the same metastate for two flip-related boundary conditions.

This invariance with respect to different sequences of periodic and antiperiodic boundary conditions means essentially that the frequency of appearance of various thermodynamic states $\Gamma^{(L)}$ in finite volumes Λ_L is independent of the choice of boundary conditions. Moreover, this same

invariance property holds among any two sequences of fixed boundary conditions (and the fixed boundary condition of choice may even be allowed to vary arbitrarily along any single sequence of volumes)! It follows that, with respect to changes of boundary conditions, the metastate is extraordinarily robust.

This should rule out all but the simplest overlap structures, and in particular the nonstandard RSB and related pictures (for a full discussion, see Newman and Stein 1998b). It is therefore natural to ask whether the property of metastate invariance allows any many-state picture.

There is one such picture, namely the "chaotic pairs" picture, which is fully consistent with metastate invariance (our belief is that it is the only many-state picture that fits naturally and easily into results obtained about the metastate.)

Here the periodic boundary condition metastate is supported on infinitely many pairs of pure states, but instead of eqn [12] one has

$$\Gamma = (1/2)\rho_{\alpha_\Gamma} + (1/2)\rho_{-\alpha_\Gamma} \qquad [15]$$

with overlap

$$P_\Gamma = (1/2)\delta(q - q_{EA}) + (1/2)\delta(q + q_{EA}) \qquad [16]$$

So there is CSD in the states but not in the overlaps, which have the same form as a two-state picture in every volume. The difference is that, while in the latter case, one has the "same" pair of states in every volume, in chaotic pairs the pure-state pair varies chaotically as volume changes. If the chaotic pairs picture is to be consistent with metastate invariance in a natural way, then the number of pure-state pairs should be "uncountable." This allows for a "uniform" distribution (within the metastate) over all of the pure states, and invariance of the metastate with respect to boundary conditions could follow naturally.

Open Questions

We have discussed how the metastate approach to the EA spin glass has narrowed considerably the set of possible scenarios for low-temperature ordering in any finite dimension, should broken spin-flip symmetry occur. The remaining possibilities are either a two-state scenario, such as droplet/scaling, or the chaotic-pairs picture if there exist many pure states at some (β, d). Both have simple overlap structures. The metastate approach appears to rule out more complicated scenarios such as RSB, in which the approximate pure-state decomposition in a typical large, finite volume is a nontrivial mixture of many pure-state pairs.

Of course, this does not answer the question of which, if either, of the remaining pictures actually does occur in real spin glasses. In this section we list a number of open questions relevant to the above discussion.

Open Question 1 Determine whether a phase transition occurs in any finite dimension greater than one. If it does, find the lower critical dimension.

Existence of a phase transition does not necessarily imply two or more pure states below T_c. It could happen, for example, that in some dimension there exists a single pure state at all nonzero temperatures, with two-point spin correlations decaying exponentially above T_c and more slowly (e.g., as a power law) below T_c. This leads to:

Open Question 2 If there does exist a phase transition above some lower critical dimension, determine whether the low-temperature spin-glass phase exhibits broken spin-flip symmetry.

If broken symmetry does occur in some dimension, then of course an obvious open question is to determine the number of pure-state pairs, and hence the nature of ordering at low temperature. A (possibly) easier question (but still very difficult), and one which does not rely on knowing whether a phase transition occurs, is to determine the zero-temperature – i.e., ground state – properties of spin glasses as a function of dimensionality. A ground state is an infinite-volume spin configuration whose energy (governed by eqn [1]) cannot be lowered by flipping any finite subset of spins. That is, all ground state spin configurations must satisfy the constraint

$$\sum_{\langle x,y\rangle \in \mathcal{C}} J_{xy}\sigma_x\sigma_y \geq 0 \qquad [17]$$

along any closed loop $\mathcal{C}$ in the dual lattice.

Open Question 3 How many ground state pairs is the $T = 0$ periodic boundary condition metastate supported on, as a function of d?

The answer is known to be one for 1D, and a partial result (Newman and Stein 2000, 2001a) points towards the answer being one for 2D as well. There are no rigorous, or even heuristic (except based on underlying "ansätze") arguments in higher dimension.

An interesting – but unrealistic – spin-glass model in which the ground state structure can be exactly solved (although not yet completely rigorously) was proposed by the authors (Newman and Stein 1994, 1996a) (see also Banavar 1994). This "highly disordered" spin glass is one in which the coupling magnitudes scale nonlinearly with the volume (and so are no longer distributed independently of the volume, although they remain independent and identically distributed for each volume). The model

displays a transition in ground state multiplicity: below eight dimensions, it has only a single pair of ground states, while above eight it has uncountably many such pairs. The mechanism behind the transition arises from a mapping to invasion percolation and minimal spanning trees (Lenormand and Bories 1980, Chandler *et al*. 1982, Wilkinson and Willemsen 1983): the number of ground state pairs can be shown to equal $2^{\mathcal{N}}$, where $\mathcal{N} = \mathcal{N}(d)$ is the number of distinct global components in the "minimal spanning forest." The zero-temperature free boundary condition metastate above eight dimensions is supported on a uniform distribution (in a natural sense) on uncountably many ground state pairs.

Interestingly, the high-dimensional ground state multiplicity in this model can be shown to be unaffected by the presence of frustration, although frustration still plays an interesting role: it leads to the appearance of chaotic size dependence when free boundary conditions are used.

Returning to the more difficult problem of ground state multiplicity in the EA model, we note as a final remark that there could, in principle, exist ground state pairs that are not in the support of metastates generated through the use of coupling-independent boundary conditions. If such states exist, they may be of some interest mathematically, but are not expected to play any significant physical role. A discussion of these putative "invisible states" is given by Newman and Stein (2003).

Open Question 4 If there exists broken spin-flip symmetry at a range of positive temperatures in some dimensions, then what is the number of pure-state pairs as a function of (β, d)?

Again, the answer to this is not known above one dimension; indeed, the prerequisite existence of spontaneously broken spin-flip symmetry has not been proved in any dimension. A speculative paper by the authors (Newman and Stein 2001b), using a variant of the highly disordered model, suggests that there is at most one pair of pure states in the EA model below eight dimensions; but no rigorous arguments are known at this time.

Acknowledgments

This work was supported in part by NSF Grants DMS-01-02587 and DMS-01-02541.

See also: Glassy Disordered Systems: Dynamical Evolution; Mean Field Spin Glasses and Neural Networks; Spin Glasses.

Further Reading

Aizenman M, Lebowitz JL, and Ruelle D (1987) Some rigorous results on the Sherrington–Kirkpatrick spin glass model. *Communications in Mathematical Physics* 112: 3–20.

Aizenman M and Wehr J (1990) Rounding effects of quenched randomness on first-order phase transitions. *Communications in Mathematical Physics* 130: 489–528.

Banavar JR, Cieplak M, and Maritan A (1994) Optimal paths and domain walls in the strong disorder limit. *Physical Review Letters* 72: 2320–2323.

Binder K and Young AP (1986) Spin glasses: experimental facts, theoretical concepts, and open questions. *Review of Modern Physics* 58: 801–976.

Bray AJ and Moore MA (1985) Critical behavior of the three-dimensional Ising spin glass. *Physical Review B* 31: 631–633.

Campanino M, Olivieri E, and van Enter ACD (1987) One dimensional spin glasses with potential decay $1/r^{1+\epsilon}$. Absence of phase transitions and cluster properties. *Communications in Mathematical Physics* 108: 241–255.

Chandler R, Koplick J, Lerman K, and Willemsen JF (1982) Capillary displacement and percolation in porous media. *Journal of Fluid Mechanics* 119: 249–267.

Chowdhury D (1986) *Spin Glasses and Other Frustrated Systems*. New York: Wiley.

Dotsenko V (2001) *Introduction to the Replica Theory of Disordered Statistical Systems*. Cambridge: Cambridge University Press.

Edwards S and Anderson PW (1975) Theory of spin glasses. *Journal of Physics F* 5: 965–974.

Fischer KH and Hertz JA (1991) *Spin Glasses*. Cambridge: Cambridge University Press.

Fisher DS and Huse DA (1986) Ordered phase of short-range Ising spin-glasses. *Physical Review Letters* 56: 1601–1604.

Fisher DS and Huse DA (1988) Equilibrium behavior of the spin-glass ordered phase. *Physical Review B* 38: 386–411.

Fisher ME and Singh RRP (1990) Critical points, large-dimensionality expansions and Ising spin glasses. In: Grimmett G and Welsh DJA (eds.) *Disorder in Physical Systems*, pp. 87–111. Oxford: Clarendon Press.

Franz S, Mézard M, Parisi G, and Peliti L (1998) Measuring equilibrium properties in aging systems. *Physical Review Letters* 81: 1758–1761.

Fröhlich J and Zegarlinski B (1987) The high-temperature phase of long-range spin glasses. *Communications in Mathematical Physics* 110: 121–155.

Gandolfi A, Newman CM, and Stein DL (1993) Exotic states in long-range spin glasses. *Communications in Mathematical Physics* 157: 371–387.

Georgii HO (1988) *Gibbs Measures and Phase Transitions*. Berlin: de Gruyter.

Kawashima N and Young AP (1996) Phase transition in the three-dimensional $\pm J$ Ising spin glass. *Physical Review B* 53: R484–R487.

Krzakala F and Martin OC (2000) Spin and link overlaps in three-dimensional spin glasses. *Physical Review Letters* 85: 3013–3016.

Lenormand R and Bories S (1980) Description d'un mécanisme de connexion de liaison destiné à l'étude du drainage avec piègeage en milieu poreux. *Comptes Rendus de l'Académie des Sciences* 291: 279–282.

Marinari E and Parisi G (2000) Effects of changing the boundary conditions on the ground state of Ising spin glasses. *Physical Review B* 62: 11677–11685.

Marinari E and Parisi G (2001) Effects of a bulk perturbation on the ground state of 3*D* Ising spin glasses. *Physical Review Letters* 86: 3887–3890.
Marinari E, Parisi G, Ricci-Tersenghi F, Ruiz-Lorenzo JJ, and Zuliani F (2000) Replica symmetry breaking in spin glasses: Theoretical foundations and numerical evidences. *Journal of Statistical Physics* 98: 973–1047.
Marinari E, Parisi G, and Ritort F (1994) On the 3*D* Ising spin glass. *Journal of Physics A* 27: 2687–2708.
Marinari E, Parisi G, and Ruiz-Lorenzo J (1997) Numerical simulations of spin glass systems. In: Young AP (ed.) *Spin Glasses and Random Fields*, pp. 59–98. Singapore: World Scientific.
McMillan WL (1984) Scaling theory of Ising spin glasses. *Journal of Physics C* 17: 3179–3187.
Mézard M, Parisi G, Sourlas N, Toulouse G, and Virasoro M (1984) Nature of spin-glass phase. *Physical Review Letters* 52: 1156–1159.
Mézard M, Parisi G, and Virasoro MA (eds.) (1987) *Spin Glass Theory and Beyond*. Singapore: World Scientific.
Newman CM and Stein DL (1992) Multiple states and thermodynamic limits in short-ranged Ising spin glass models. *Physical Review B* 46: 973–982.
Newman CM and Stein DL (1994) Spin-glass model with dimension-dependent ground state multiplicity. *Physical Review Letters* 72: 2286–2289.
Newman CM and Stein DL (1996a) Ground state structure in a highly disordered spin glass model. *Journal of Statistical Physics* 82: 1113–1132.
Newman CM and Stein DL (1996b) Non-mean-field behavior of realistic spin glasses. *Physical Review Letters* 76: 515–518.
Newman CM and Stein DL (1996c) Spatial inhomogeneity and thermodynamic chaos. *Physical Review Letters* 76: 4821–4824.
Newman CM and Stein DL (1998a) Simplicity of state and overlap structure in finite volume realistic spin glasses. *Physical Review E* 57: 1356–1366.
Newman CM and Stein DL (1998b) Thermodynamic chaos and the structure of short-range spin glasses. In: Bovier A and Picco P (eds.) *Mathematics of Spin Glasses and Neural Networks*, pp. 243–287. Boston: Birkhauser.
Newman CM and Stein DL (2000) Nature of ground state incongruence in two-dimensional spin glasses. *Physical Review Letters* 84: 3966–3969.
Newman CM and Stein DL (2001a) Are there incongruent ground states in 2*D* Edwards–Anderson spin glasses? *Communications in Mathematical Physics* 224: 205–218.
Newman CM and Stein DL (2001b) Realistic spin glasses below eight dimensions: a highly disordered view. *Physical Review E* 63: 16101-1–16101-9.
Newman CM and Stein DL (2003) Topical review: Ordering and broken symmetry in short-ranged spin glasses. *Journal of Physics: Condensed Matter* 15: R1319–R1364.
Ogielski AT (1985) Dynamics of three-dimensional spin glasses in thermal equilibrium. *Physical Review B* 32: 7384–7398.
Ogielski AT and Morgenstern I (1985) Critical behavior of the three-dimensional Ising spin-glass model. *Physical Review Letters* 54: 928–931.
Palassini M and Young AP (2000) Nature of the spin glass state. *Physical Review Letters* 85: 3017–3020.
Parisi G (1979) Infinite number of order parameters for spin-glasses. *Physical Review Letters* 43: 1754–1756.
Parisi G (1983) Order parameter for spin-glasses. *Physical Review Letters* 50: 1946–1948.
Sherrington D and Kirkpatrick S (1975) Solvable model of a spin glass. *Physical Review Letters* 35: 1792–1796.
Stein DL (1989) Disordered systems: Mostly spin glasses. In: Stein DL (ed.) *Lectures in the Sciences of Complexity*, pp. 301–355. New York: Addison–Wesley.
Thill MJ and Hilhorst HJ (1996) Theory of the critical state of low-dimensional spin glass. *Journal of Physics I* 6: 67–95.
van Enter ACD and Fröhlich J (1985) Absence of symmetry breaking for *N*-vector spin glass models in two dimensions. *Communications in Mathematical Physics* 98: 425–432.
Wilkinson D and Willemsen JF (1983) Invasion percolation: A new form of percolation theory. *Journal of Physics A* 16: 3365–3376.

Sine-Gordon Equation

S N M Ruijsenaars, Centre for Mathematics and Computer Science, Amsterdam, The Netherlands

Introduction

The sine-Gordon equation

$$\left(\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial t^2}\right)\phi = \sin\,\phi \qquad [1]$$

may be viewed as a prototype for a nonlinear integrable field theory. It is manifestly invariant under spacetime translations and Lorentz boosts,

$$\begin{aligned}&(x,t) \mapsto (x-\alpha, t-\beta)\\ &(x,t) \mapsto (x\cosh\theta - t\sinh\theta, t\cosh\theta - x\sinh\theta)\end{aligned} \qquad [2]$$

It shares this relativistic invariance property with the linear Klein–Gordon equation, which is obtained upon replacing $\sin\phi$ by ϕ. (The name sine-Gordon equation is derived from this relation, and was introduced by Kruskal.) The sine-Gordon equation can also be defined and studied in the form

$$\frac{\partial^2}{\partial u\,\partial v}\tilde{\phi} = \sin\,\tilde{\phi}, \quad \tilde{\phi}(u,v) = \phi(t,x) \qquad [3]$$

where

$$u = (x+t)/2, \quad v = (x-t)/2 \qquad [4]$$

are the so-called light-cone variables.

There are two interpretations of the field $\phi(t,x)$ that are quite different, both from a physical and from a mathematical viewpoint. The first one consists in viewing it as a real-valued function, so

that [1] is simply a nonlinear PDE in two variables. In the second version, one views $\phi(t,x)$ as an operator-valued distribution on a Hilbert space. (Thus, one should smear $\phi(t,x)$ with a test function $f(t,x)$ in Schwartz space to obtain a genuine operator on the Hilbert space.) In spite of their different character, the classical and quantum field theory versions have several striking features in common, including the presence of an infinite number of conservation laws and the occurrence of solitonic excitations.

The classical sine-Gordon equation has been used as a model for various wave phenomena, including the propagation of dislocations in crystals, phase differences across Josephson junctions, torsion waves in strings and pendula, and waves along lipid membranes. It was already studied in the nineteenth century in connection with the theory of pseudospherical surfaces. The quantum version is used as a simple model for solid-state excitations.

The designation "sine-Gordon" is also used for various equations that generalize [1] or bear resemblance to it. These include the so-called homogeneous and symmetric space sine-Gordon models, discrete and supersymmetric versions, and generalizations to higher-dimensional spacetimes (i.e., in [1] the spatial derivative is replaced by the Laplace operator in several variables). In this contribution we focus on [1], however.

Our main goal is to discuss the integrability and solitonic properties, both at the classical and at the quantum level. First, we sketch the inverse-scattering transform (IST) solution to the Cauchy problem for [1]. Following Faddeev and Takhtajan, we emphasize the interpretation of the IST as an action-angle transformation for an infinite-dimensional Hamiltonian system. Next, the particle-like solutions are surveyed by using a description in terms of variables that may be viewed as relativistic action-angle coordinates. This is followed by a section on the quantum field theory version, paying special attention to the factorized scattering that is the quantum analog of the solitonic classical scattering. Finally, we sketch the intimate relation between the N-particle subspaces of the classical and quantum sine-Gordon field theory and certain integrable relativistic systems of N point particles on the line.

The Classical Version: An Integrable Hamiltonian System

In order to tie in the hyperbolic evolution equation [1] with the notion of infinite-dimensional integrable system, it is necessary to restrict attention to initial data $\phi(0,x)=\phi(x)$ and $\partial_t\phi(0,x)=\pi(x)$ with special properties. First of all, the energy functional

$$H=\int_{-\infty}^{\infty}\left(\frac{1}{2}\pi(x)^2+\frac{1}{2}\partial_x\phi(x)^2+(1-\cos\phi(x))\right)\mathrm{d}x \qquad [5]$$

and symplectic form

$$\omega=\int_{-\infty}^{\infty}\mathrm{d}\pi(x)\wedge\mathrm{d}\phi(x)\,\mathrm{d}x \qquad [6]$$

should be well defined on the phase space of initial data. Indeed, in that case [1] amounts to the Hamilton equation associated to [5] via [6].

Second, there exists a sequence of functionals

$$I_{2l+1}(\phi,\pi),\quad l\in\mathbb{Z} \qquad [7]$$

that formally Poisson-commute with H and among themselves.

In particular, H equals $2(I_1+I_{-1})$, whereas $2(I_1-I_{-1})$ equals the momentum functional

$$P=-\int_{-\infty}^{\infty}\pi(x)\partial_x\phi(x)\,\mathrm{d}x \qquad [8]$$

The functional I_{2l+1} contains x-derivatives of order up to $|2l+1|$, so one needs to require that the functions $\partial_x\phi(x)$ and $\pi(x)$ be smooth and that all of their derivatives have sufficient decrease for $x\to\pm\infty$.

A natural choice guaranteeing the latter requirements is

$$\partial_x\phi(x),\pi(x)\in S_{\mathbb{R}}(\mathbb{R}) \qquad [9]$$

where $S_{\mathbb{R}}(\mathbb{R})$ denotes the Schwartz space of real-valued functions on the line. To render the integral over $1-\cos\phi(x)$ (and similar integrals occurring for the sequence [7]) finite, one also needs to require

$$\phi(x)\to 2\pi k_{\pm},\quad x\to\pm\infty,\ k_{\pm}\in\mathbb{Z} \qquad [10]$$

On this phase space Ω of initial data, the Cauchy problem for the evolution equation [1] is not only well posed, but can be solved in explicit form by using the IST. More generally, the Hamiltonians I_{2l+1} give rise to evolution equations that are simultaneously solved via the IST, yielding an infinite sequence of commuting Hamiltonian flows on Ω.

Before sketching the overall picture resulting from the IST, it should be mentioned at this point that [1] admits explicit solutions of interest that do not belong to Ω. First, there is a class of algebro-geometric solutions that have no limits as $x\to\pm\infty$. These solutions can be obtained via finite-gap integration methods, yielding formulas involving

the Riemann theta functions associated to compact Riemann surfaces. Second, there are the tachyon solutions. They arise from the particle-like solutions that do belong to Ω by the transformation

$$\phi(t,x) \to \phi(x,t) + \pi \qquad [11]$$

(Observe that the equation of motion [1] is invariant under [11], whereas due to the finite-energy requirement [10] this is not true for solutions evolving in Ω.)

The IST via which the above Cauchy problem can be solved starts from an auxiliary system of two linear ordinary differential equations involving $\phi(0,x)$ and $\partial_t\phi(0,x)$. It is beyond our scope to describe the system in detail. The results derived from it, however, are to a large extent the same as those obtained via a simpler auxiliary linear operator that is associated to the light-cone Cauchy problem. The latter operator is of the Ablowitz–Kaup–Newell–Segur (AKNS) form. That is, the linear operator is an ordinary differential operator of Dirac type given by

$$L = \begin{pmatrix} \mathrm{i}\frac{\mathrm{d}}{\mathrm{d}x} & -\mathrm{i}q \\ \mathrm{i}r & -\mathrm{i}\frac{\mathrm{d}}{\mathrm{d}x} \end{pmatrix} \qquad [12]$$

where the external potentials $r(u)$ and $q(u)$ depend on the evolution equation at hand. For the light-cone sine-Gordon equation [3], one needs to choose

$$r = -q = (\partial_u\tilde{\phi})(u,0)/2 \qquad [13]$$

In both settings, the associated spectral features are invariant under the sine-Gordon evolution and all of the evolutions generated by the Hamiltonians I_{2l+1}, yielding the so-called isospectral flows. More specifically, if the initial data give rise to bound-state solutions of the linear problem (square-integrable wave functions), then the corresponding eigenvalues are time independent. Furthermore, due to the decay requirements on the potential in the linear system, there exist scattering solutions with plane-wave asymptotics for all initial data in Ω. A suitable normalization leads to the so-called Jost solutions $\Psi(x,\lambda)$. (Here λ is the spectral parameter, which varies over the real line for scattering solutions.) Their $x \to \pm\infty$ asymptotics is encoded in transition coefficients $a(\lambda)$ and $b(\lambda)$, with $a(\lambda)$ and $|b(\lambda)|$ being time independent, whereas $\arg b(\lambda)$ has a linear dependence on time when the potential evolves according to the sine-Gordon equation. The bound states correspond to special λ-values $\lambda_1, \ldots, \lambda_N$ with positive imaginary part (namely the zeros of the coefficient $a(\lambda)$, which is analytic in the upper-half λ-plane); their normalization coefficients $\nu_1, \ldots, \nu_N$ have an essentially linear time evolution, just as $b(\lambda)$.

The crux of the IST is now that the potentials can be reconstructed from the spectral data

$$\{b(\lambda), \lambda_1, \ldots, \lambda_N, \nu_1, \ldots, \nu_N\} \qquad [14]$$

by solving a linear integral equation of Gelfand–Levitan–Marchenko (GLM) type. (Alternatively, Riemann–Hilbert problem techniques can be used.) Hence, the nonlinear Cauchy problem can be replaced by the far simpler linear problems of determining the spectral data [14] of a linear operator (the direct problem) and then solving the linear GLM equation for the time-evolved scattering data (the inverse problem).

From the Hamiltonian perspective, the IST may be reinterpreted as a transformation to action-angle variables. The action variables are defined in terms of $|b(\lambda)|$ and $\lambda_1, \ldots, \lambda_N$. They are time independent under the sine-Gordon and higher Hamiltonian flows. The angle variables are $\arg b(\lambda)$ and suitable functions of the normalization coefficients. They depend linearly on the evolution times of the flows. The Hamiltonians can be explicitly expressed in action variables.

Next, we point out that there is a large subspace of Cauchy data $(\phi(x), \pi(x))$ that do not give rise to bound states in the auxiliary linear problem. The associated solutions are the so-called radiation solutions: they decrease to 0 for large times. These solutions can be obtained from the inverse transform involving the GLM equation by only taking $b(\lambda)$ into account.

The other extreme is to choose $b(\lambda) = 0$ and arbitrary bound states and normalization coefficients in the GLM equation. This special case of vanishing reflection leads to the particle-like solutions that are studied in the next section. For general Cauchy data, one has both $b(\lambda) \neq 0$ and a finite number of bound states. These so-called mixed solutions have a radiation component (encoded in $b(\lambda)$) which decays for asymptotic times, whereas the bound states show up for $t \to \pm\infty$ as isolated solitons, antisolitons, and breathers.

Classical Solitons, Antisolitons, and Breathers

Just as for other classical soliton equations, the case of reflectionless data can be handled in complete detail, since the GLM equation reduces to an $N \times N$ system of linear equations. The case $N = 1$ yields the 1-soliton and 1-antisoliton solutions. Resting at the origin, these one-particle solutions are given by

$$\pm 4 \arctan(\mathrm{e}^{-x}) \qquad [15]$$

and have energy 8 (cf. [5]). (We normalize all solutions by requiring

$$\lim_{x\to\infty}\phi(t,x)=0 \qquad [16]$$

Note that one can add arbitrary multiples of 2π without changing the energy H [5].) A spatial translation and Lorentz boost then yields the general solutions

$$\phi_{\pm}(t,x) = \pm 4\arctan(\exp(q-x\cosh\theta+t\sinh\theta)) \qquad [17]$$

with energy $8\cosh\theta$ and momentum $8\sinh\theta$ (cf. [8]).

Defining the topological charge of a solution (with normalization [16]) by

$$Q=\frac{1}{2\pi}\lim_{x\to-\infty}\phi(t,x) \qquad [18]$$

the different charges $Q=1$ and $Q=-1$ of the soliton and antisoliton reflect a signature associated to the special value of the spectral parameter on the imaginary axis for which a bound state in the linear problem occurs. More generally, for bound-state eigenvalues on the imaginary axis these signatures must be specified in the IST setting, a point glossed over in the previous section.

Bound states in the linear problem can also arise from λ-values off the imaginary axis, which come in pairs $ia\pm b$, with $a,b>0$. Such pairs give rise to solutions containing breathers, which can be viewed as bound states of a soliton and an antisoliton. The one-breather solution breathing at the origin is given by

$$4\arctan\left(\cot\eta\frac{\sin(t\sin\eta)}{\cosh(x\cos\eta)}\right),\quad \eta\in(0,\pi/2) \qquad [19]$$

and has energy $16\cos\eta$. A spacetime translation and Lorentz boost then yields the general solution

$$\phi_b(t,x) = 4\arctan\left(\cot\eta\frac{\sin[-\gamma/2+\sin\eta(t\cosh\theta-x\sinh\theta)]}{\cosh[y/2-\cos\eta(x\cosh\theta-t\sinh\theta)]}\right) \qquad [20]$$

which has energy $16\cosh\theta\cos\eta$ and momentum $16\sinh\theta\cos\eta$. It may be obtained by analytic continuation from the solution describing a collision between a soliton with velocity $\tanh\theta_1$ and an antisoliton with velocity $\tanh\theta_2$, taking $\theta_2<\theta_1$. The latter is given by

$$\phi_{+-}(t,x) = 4\arctan\left(\coth((\theta_1-\theta_2)/2)\frac{\sinh((\mu_1-\mu_2)/2)}{\cosh((\mu_1+\mu_2)/2)}\right) \quad \theta_2<\theta_1 \qquad [21]$$

where

$$\mu_j=q_j-x\cosh\theta_j+t\sinh\theta_j,\quad q_j,\theta_j\in\mathbb{R} \qquad [22]$$

and ϕ_b results from ϕ_{+-} by substituting

$$\theta_{\frac{1}{2}}\to\theta\pm i\eta,\qquad q_{\frac{1}{2}}\to(y\mp i\gamma)/2 \qquad [23]$$

(For the case $\theta_1<\theta_2$, one needs an extra minus sign on the right-hand side of [21].)

There is yet another possibility for an eigenvalue on the imaginary axis we have not mentioned thus far: it may have an arbitrary multiplicity, giving rise to the so-called multipole solutions. This is illustrated by the breather solution ϕ_b: when one sets $\gamma=-2q_0\eta$ and lets η tend to 0, one obtains a solution

$$\phi_{\text{sep}}(t,x) = 4\arctan\left(\frac{q_0+t\cosh\theta-x\sinh\theta}{\cosh[y/2-x\cosh\theta+t\sinh\theta]}\right) \qquad [24]$$

From a physical viewpoint, the soliton and antisoliton have just enough energy to prevent a bound state from being formed. Notice that in this case the distance between soliton and antisoliton diverges logarithmically in $|t|$ as $t\to\pm\infty$, whereas for ϕ_{+-} one obtains linear increase.

The 2-soliton and 2-antisoliton solutions can also be obtained by analytic continuation of ϕ_{+-}. They read

$$\phi_{\pm\pm}=\mp 4\arctan\left(\coth((\theta_1-\theta_2)/2)\times\frac{\cosh((\mu_1-\mu_2)/2)}{\sinh((\mu_1+\mu_2)/2)}\right),\quad \theta_2<\theta_1 \qquad [25]$$

where μ_j is given by [22]. Thus, they arise by taking $q_2\to q_2+i\pi$ and $q_1\to q_1+i\pi$ in [21], resp. The equal-signature eigenvalues corresponding to these two solutions cannot collide and move off the imaginary axis; physically speaking, equal-charge particles repel each other. The energy and momentum of the solutions [25] and [21] are given by $8\cosh\theta_1+8\cosh\theta_2$ and $8\sinh\theta_1+8\sinh\theta_2$, respectively.

Up to scale factors, the above variables θ_1,θ_2 and θ,η are the action variables resulting from the IST, whereas q_1,q_2 and y,γ are the canonically conjugated angle variables. Accordingly, the time and space translation flows (generated by H [5] and P [8], resp.) shift the angles linearly in the evolution parameters t and x.

We conclude this section with a description of the N-soliton solution and its large time asymptotics. It can be expressed in terms of the $N\times N$ matrix

$$\mathcal{L}_{jk}=\exp(\mu_j)\frac{\prod_{l\neq j}|\coth((\theta_j-\theta_l)/2)|}{\cosh((\theta_j-\theta_k)/2)} \qquad [26]$$

where μ_j is given by [22] with

$$q_1, \ldots, q_N \in \mathbb{R}, \qquad \theta_N < \cdots < \theta_1 \qquad [27]$$

Specifically, one has

$$\begin{aligned}\phi_{N+}(t,x) &= 4 \operatorname{tr} \arctan(\mathcal{L}) \\ &= -2\mathrm{i} \ln(|\mathbf{1}_N + \mathrm{i}\mathcal{L}|/|\mathbf{1}_N - \mathrm{i}\mathcal{L}|) \\ &= -2\mathrm{i} \ln\left(\left(1 + \sum_{l=1}^{N} \mathrm{i}^l S_l(\mathcal{L})\right)/\text{c.c.}\right) \end{aligned} \qquad [28]$$

where S_l is the lth symmetric function of $\mathcal{L}$. Using Cauchy's identity, one obtains the explicit formula

$$S_l = \sum_{\substack{I \subset \{1,\ldots,N\} \\ |I|=l}} \exp\left(\sum_{j \in I} \mu_j\right) \prod_{\substack{j \in I \\ k \notin I}} |\coth((\theta_j - \theta_k)/2)| \qquad [29]$$

In order to specify the $t \to \pm\infty$ asymptotics of ϕ_{N+}, we introduce the 1-soliton solutions

$$\phi_j^{\pm}(t,x) = 4 \arctan(\exp(\mu_j \mp \Delta_j/2)) \qquad [30]$$

where

$$\Delta_j = \left(\sum_{k<j} - \sum_{k>j}\right) \delta(\theta_j - \theta_k) \qquad [31]$$

$$\delta(\theta) = \ln\left(\coth(\theta/2)^2\right) \qquad [32]$$

Then, one has

$$\sup_{x \in \mathbb{R}} \left|\phi_{N+}(t,x) - \sum_{j=1}^{N} \phi_j^{\pm}(t,x)\right| = O(\exp(-|t|r)) \quad t \to \pm\infty \qquad [33]$$

where the decay rate is given by

$$r = \min_{j \neq k}(\cosh(\theta_j)|\tanh\theta_j - \tanh\theta_k|) \qquad [34]$$

Thus, the soliton profile with velocity $\tanh\theta_j$ incurs a shift $\Delta_j/\cosh\theta_j$ as a result of the collision. The factor $1/\cosh\theta_j$ may be viewed as a Lorentz contraction factor.

The Quantum Version: A Soliton Quantum Field Theory

From a perturbation-theoretic viewpoint, the quantum sine-Gordon Hamiltonian is given by

$$H = \int_{-\infty}^{\infty} : \left(\frac{1}{2}(\partial_t\phi)^2 + \frac{1}{2}(\partial_x\phi)^2 + \frac{m^2}{\beta^2}(1 - \cos\beta\phi)\right) : \mathrm{d}x, \quad m, \beta > 0 \qquad [35]$$

Here, $\phi(0,x)$ is a neutral Klein–Gordon field with mass m and the double dots denote a suitable ordering prescription. The associated equation of motion

$$\phi_{xx} - \phi_{tt} = \frac{m^2}{\beta} \sin\beta\phi \qquad [36]$$

is equivalent to [1] on the classical level, but not on the quantum level. (If $\phi(t,x)$ is a classical solution to [36], then $\beta\phi(t/m, x/m)$ solves [1].) This difference is due to the extremely singular character of interacting relativistic quantum field theory, a context in which "solving" the field theory has slowly acquired a meaning that is vastly different from the classical notion. Indeed, one can at best hope to verify [36] in the sense of expectation values in suitable quantum states, and this is precisely what has been achieved within the form-factor program sketched later on.

From the perspective of functional analysis, the existence of a well-defined Wightman field theory with all of the features mentioned below is wide open. More precisely, beginning with pioneering work by Fröhlich some 30 years ago, various authors have contributed to a mathematically rigorous construction of a sine-Gordon quantum field theory version, but to date it seems not feasible to verify that the resulting Wightman field theory has any of the explicit features we are going to sketch. (For example, not even the free character of the field theory for $\beta^2 = 4\pi$ has been established; cf. below.)

That said, we proceed to sketch some highlights of the impressive, but partly heuristic lore that has been assembled in a great many theoretical physics papers. A key result we begin with is the equivalence to a field theory that looks very different at face value. This is the massive Thirring model, formally given by the Hamiltonian

$$H_{\mathrm{T}} = \int_{-\infty}^{\infty} : \left(\Psi^*(-\mathrm{i}\gamma^5\partial_x + \gamma^0 M)\Psi + \frac{g}{2}(J_0^2 - J_1^2)\right) : \mathrm{d}x \quad M \in (0,\infty), \; g \in \mathbb{R} \qquad [37]$$

Here, $\Psi(0,x)$ is the charged Dirac field with mass M and the double dots stand for normal ordering. For the γ-algebra, one may choose

$$\gamma^0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \gamma^1 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$
$$\gamma^5 = \gamma^0\gamma^1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad [38]$$

and J_μ is the Dirac current,

$$J_0 = \Psi^*\Psi, \qquad J_1 = \Psi^*\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\Psi \qquad [39]$$

The equivalence argument (due to Coleman) consists in showing that the quantities

$$-\frac{\beta}{2\pi}\epsilon_{\mu\nu}\partial^{\nu}\phi, \qquad \frac{m^2}{\beta^2} : \cos\beta\phi : \qquad [40]$$

in the sine-Gordon theory have the same vacuum expectation values (in perturbation theory) as the massive Thirring quantities

$$: J_{\mu} :, \qquad -M : \Psi^{*}\gamma^{0}\Psi : \qquad [41]$$

resp., provided the parameters are related by

$$\frac{4\pi}{\beta^2} = 1 + \frac{g}{\pi} \qquad [42]$$

This yields an equivalence between the charge-0 sector of the massive Thirring model and the sector of the sine-Gordon theory obtained by the action of the fields [40] on the vacuum vector. But the charged sectors of the Thirring model can also be viewed as new sectors in the sine-Gordon theory, obtained by a solitonic field construction (first performed by Mandelstam).

In this picture, the fermions and antifermions in the massive Thirring model correspond to new excitations in the sine-Gordon theory, the quantum solitons and antisolitons. The latter are viewed as coherent states of the sine-Gordon "mesons" in the vacuum sector, the rest masses being related by

$$M = \frac{8m}{\beta^2}\left(1 - \frac{\beta^2}{8\pi}\right) \qquad [43]$$

in the semiclassical limit $\beta^2 \to 0$.

Even at the formal level involved in the correspondence, the theories are not believed to exist for $\beta^2 > 8\pi$ and $g < -\pi/2$, since there is positivity breakdown for this range of couplings. The free Dirac case $g = 0$ corresponds to $\beta^2 = 4\pi$. In particular, there is no interaction between the sine-Gordon solitons and antisolitons for this β-value. In the range $\beta^2 \in (4\pi, 8\pi)$ there is interaction, but bound soliton–antisoliton pairs (quantum breathers, alias sine-Gordon mesons) do not occur.

By contrast, for $\beta^2 < 4\pi$ there exist breathers with rest masses

$$m_n = 2M\sin(n+1)\alpha, \quad \alpha \equiv m/2M,$$
$$n + 1 = 1, 2, \ldots, L < \pi/2\alpha \qquad [44]$$

Thus, the "particle spectrum" consists of solitons and antisolitons with mass M and mesons $C_1, \ldots, C_L$ with masses $m_1, \ldots, m_L$ given by [44]. The latter formula was first established by semiclassical quantization of the classical breathers (Dashen–Hasslacher–Neveu), and ever since is usually called the DHN formula. Notice that for α near zero m_1 and m are nearly equal, and that for $\beta^2 \geq 4\pi$ there are no longer any sine-Gordon mesons present in the theory.

A priori, the existence of infinitely many classical conserved Hamiltonians does not even formally imply the same feature for the quantum field theory, as anomalies may occur. For the sine-Gordon and massive Thirring cases, anomalies have been shown to be absent, however. This entails not only that the number of solitons, antisolitons, and breathers in a scattering process is conserved, but also that the set of incoming rapidities equals the set of outgoing rapidities.

The latter stability features and the DHN formula [44] are corroborated by the S-matrix, which is known in complete detail. The two-body amplitudes involving solitons and antisolitons can be written in terms of the function

$$u(z) = \exp\left(\mathrm{i}\int_0^{\infty}\frac{\mathrm{d}x}{x}\frac{\sinh(\alpha - \pi/2)x}{\sinh\alpha x\cosh\pi x/2}\sin 2xz\right) \qquad [45]$$

They are given by

$$(u_{++}, t_{+-}, r_{+-}, u_{--})(\theta) = u(\theta/2)\left(1, \frac{\sinh(\pi\theta/2\alpha)}{\sinh(\pi(\mathrm{i}\pi - \theta)/2\alpha)}, \frac{\mathrm{i}\sin(\pi^2/2\alpha)}{\sinh(\pi(\mathrm{i}\pi - \theta)/2\alpha)}, 1\right) \qquad [46]$$

where θ denotes the rapidity difference. (Due to fermion statistics, one gets only one amplitude for a soliton or antisoliton pair. But a soliton and an antisoliton have opposite charge, so they can be distinguished. In that case, therefore, the notion of reflection and transmission coefficients makes sense.)

The S-matrix involving an arbitrary number of solitons, antisolitons, and their bound states is also explicitly known. The amplitudes involving no breathers are readily described in terms of the above two-body amplitudes. Indeed, the S-matrix factorizes as a sum of products of the amplitudes [46], yielding a picture of particles scattering independently in pairs, just as at the classical level. The factorization can be performed irrespective of the temporal ordering assumed for the pair scattering processes, since the four functions occurring inside the parentheses of [46] satisfy the Yang–Baxter equations.

Roughly speaking, the S-matrix for processes involving breathers can be calculated by analytic continuation from the soliton–antisoliton S-matrix. The details are however quite substantial. We only add that scattering amplitudes involving solely breathers can be expressed using only hyperbolic functions.

Since the 1980s, a lot of information has also been gathered concerning matrix elements of suitable sine-Gordon field quantities between special quantum states (form factors). Unfortunately, the correlation functions involve infinite sums of form factors that are quite difficult to control analytically. Hence, it is not known whether the correlation functions associated with the form factors give rise to a Wightman field theory with the usual axiomatic properties.

The Relation to Relativistic Calogero–Moser Systems

The behavior of the special classical solutions discussed earlier is very similar to that of classical point particles. Furthermore, the picture of classical solitons, antisolitons, and their bound states scattering independently in pairs is essentially preserved on the quantum level, just as one would expect for the quantization of an integrable particle system.

Next, we note that from the quantum viewpoint there is no physical distinction between wave functions and point particles, whereas a classical wave is a physical entity that is clearly very different from a point particle. Even so, it is a natural question whether there exist classical Hamiltonian systems of N point particles on the line whose physical characteristics (charges, bound states, scattering, etc.) are the same as those of the particle-like sine-Gordon solutions. If so, a second question is equally obvious: does the quantum version of the N-particle systems still have the same features as that of the quantum sine-Gordon excitations?

As we now sketch, the first question has been answered in the affirmative, whereas the second one has not been completely answered yet. However, all of the information on the pertinent quantum N-particle systems collected thus far points to an affirmative answer. The systems at issue are relativistic versions of the well-known nonrelativistic Calogero–Moser N-particle systems.

To begin with the classical two-particle system, its Hamiltonian is given by

$$H = (\cosh p_1 + \cosh p_2)\coth((x_1 - x_2)/2) \qquad [47]$$

on the phase space

$$\Omega = \{(x,p) \in \mathbb{R}^4 | x_2 < x_1\}, \quad \omega = \mathrm{d}x \wedge \mathrm{d}p \qquad [48]$$

Taking $x_2 \to x_2 + \mathrm{i}\pi$ yields the particle–antiparticle Hamiltonian

$$\tilde{H} = (\cosh p_1 + \cosh p_2)|\tanh((x_1 - x_2)/2)| \qquad [49]$$

on the phase space

$$\tilde{\Omega} = \{(x,p) \in \mathbb{R}^4\}, \quad \omega = \mathrm{d}x \wedge \mathrm{d}p \qquad [50]$$

The two-antiparticle Hamiltonian is again given by [47] and [48]. The interaction potential in [47] is repulsive, whereas it is attractive in [49]. Hence, any initial point in Ω gives rise to a scattering state, whereas points in $\tilde{\Omega}$ yield scattering states if and only if the reduced Hamiltonian

$$\tilde{H}_\mathrm{r} = \cosh p|\tanh(x/2)|, \quad p = (p_1 - p_2)/2$$
$$x = x_1 - x_2 \qquad [51]$$

satisfies $\tilde{H}_\mathrm{r} > 1$. More specifically, in both cases the distance $|x_1(t) - x_2(t)|$ increases linearly as $t \to \pm\infty$, the scattering (position shift) being encoded by the same function [32] as for the sine-Gordon solitons. The phase-space points on the separatrix $\{\tilde{H}_\mathrm{r} = 1\}$ have the same temporal asymptotics as the multipole solution [24], whereas the bound-state oscillations for $\tilde{H}_\mathrm{r} < 1$ match those of the breathers [20].

More generally, the Hamiltonian for N_+ particles and N_- antiparticles is given by the function

$$\sum_{j=1}^{N_+} \cosh(p_j^+) \prod_{\substack{k=1 \\ k\neq j}}^{N_+} |\coth((x_j^+ - x_k^+)/2)|$$
$$\times \prod_{l=1}^{N_-} |\tanh((x_j^+ - x_l^-)/2)| + \sum_{l=1}^{N_-} \cosh(p_l^-)$$
$$\times \prod_{\substack{m=1 \\ m\neq l}}^{N_-} |\coth((x_l^- - x_m^-)/2)|$$
$$\times \prod_{j=1}^{N_+} |\tanh((x_l^- - x_j^+)/2)| \qquad [52]$$

on the phase space

$$\Omega_{N_+,N_-} = \Big\{(x^+,p^+) \in \mathbb{R}^{2N_+}, (x^-,p^-) \in \mathbb{R}^{2N_-} | x_{N_+}^+ < \cdots < x_1^+, x_{N_-}^- < \cdots < x_1^-\Big\} \qquad [53]$$

$$\omega_{N_+,N_-} = \mathrm{d}x^+ \wedge \mathrm{d}p^+ + \mathrm{d}x^- \wedge \mathrm{d}p^- \qquad [54]$$

This defining Hamiltonian can be supplemented by $(N_+ + N_- - 1)$ independent Hamiltonians that pairwise commute. The action-angle map of this integrable system can be used to relate the scattering and bound-state behavior to that of the sine-Gordon solutions from an earlier section, yielding an exact correspondence. Indeed, the variables we used to describe the particle-like sine-Gordon solutions amount to the action-angle variables associated to [52]. Moreover, the matrix $\mathcal{L}$ [26] with $t = x = 0$

equals the Lax matrix for the N-particle system, which is the manifestation of a remarkable self-duality property of the equal-charge case. There is an equally close relation between the general particle-like solutions and the general systems encoded in [52].

As a matter of fact, the connection can be further strengthened by introducing spacetime trajectories for the solitons, antisolitons, and breathers, which are defined in terms of the evolution of an initial point in Ω_{N_+,N_-} under the time translation generator [52] and the space translation generator, obtained from [52] by the replacement $\cosh \to \sinh$. These point particle and antiparticle trajectories make it possible to follow the motion of the solitons, antisolitons, and breathers during the temporal interval in which the nonlinear interaction takes place, whereas for large times the trajectories are located at the (then) clearly discernible positions of the individual solitons, antisolitons, and breathers.

Before sketching the soliton-particle correspondence at the quantum level, we add a remark on the finite-gap solutions of the classical sine-Gordon equation, already mentioned in the paragraph containing [11]. These solutions may be viewed as generalizations of the particle-like solutions discussed earlier, and they can also be obtained via relativistic N-particle Calogero–Moser systems. The pertinent systems are generalizations of the hyperbolic systems just described to the elliptic level.

Turning now to the quantum level, we begin by mentioning that the Poisson-commuting Hamiltonians admit a quantization in terms of commuting analytic difference operators. This involves a special ordering choice of the p-dependent and x-dependent factors in the classical Hamiltonians, which is required to preserve commutativity. The resulting quantum two-body problem can be explicitly solved in terms of a generalization of the Gauss hypergeometric function. For the case of equal charges, the scattering is encoded in the sine-Gordon amplitudes $u_{\pm\pm}(\theta)$ (cf. [45] and [46]). For the unequal-charge case, one should distinguish an even and odd channel. The scattering on these channels is encoded in the sine-Gordon amplitudes $t_{+-}(\theta) \pm r_{+-}(\theta)$. Moreover, the bound-state spectrum agrees with the DHN formula [44] and the bound-state wave functions are given by hyperbolic functions.

As a consequence of these results, the physics encoded in the two-body subspace of the sine-Gordon quantum field theory is indistinguishable from that of the corresponding two-body relativistic Calogero–Moser systems. To extend this equivalence to the arbitrary-N case, one needs first of all sufficiently explicit solutions to the N-body Schrödinger equation. To date, this has only been achieved for the case of N equal charges and the special couplings for which the reflection amplitude r_{+-} vanishes. The asymptotics of the pertinent solutions is factorized in terms of $u_{\pm\pm}(\theta)$, in agreement with the sine-Gordon picture.

See also: Bäcklund Transformations; Boundary-Value Problems for Integrable Equations; Calogero–Moser–Sutherland Systems of Nonrelativistic and Relativistic Type; Infinite-dimensional Hamiltonian Systems; Integrability and Quantum Field Theory; Integrable Systems and Discrete Geometry; Integrable Systems and Inverse Scattering Method; Integrable Systems: Overview; Ljusternik–Schnirelman Theory; Solitons and Other Extended Field Configurations; Solitons and Kac–Moody Lie Algebras; Symmetries and Conservation Laws; Two-Dimensional Models; Yang–Baxter Equations.

Further Reading

Ablowitz MJ, Kaup DJ, Newell AC, and Segur H (1974) The inverse scattering transform – Fourier analysis for nonlinear problems. *Studies in Applied Mathematics* 53: 249–315.

Coleman S (1977) Classical lumps and their quantum descendants. In: Zichichi A (ed.) *New Phenomena in Subnuclear Physics, Proceedings Erice 1975*, pp. 297–421. New York: Plenum.

Faddeev LD and Takhtajan LA (1987) *Hamiltonian Methods in the Theory of Solitons*. Berlin: Springer.

Flaschka HF and Newell AC (1975) Integrable systems of nonlinear evolution equations. In: Moser J (ed.) *Dynamical Systems, Theory and Applications*, Lecture Notes in Physics, vol. 38, pp. 355–440. Berlin: Springer.

Karowski M (1979) Exact S-matrices and form factors in 1+1 dimensional field theoretic models with soliton behaviour. *Physics Reports* 49: 229–237.

Ruijsenaars SNM (2001) Sine-Gordon solitons vs. Calogero–Moser particles. In: Pakuliak S and von Gehlen G (eds.) *Proceedings of the Kiev NATO Advanced Study Institute Integrable Structures of Exactly Solvable Two-Dimensional Models of Quantum Field Theory*, NATO Science Series, vol. 35, pp. 273–292. Dordrecht: Kluwer.

Scott AC, Chu FYF, and McLaughlin DW (1973) The soliton: a new concept in applied science. *Proceedings of the Institute of Electrical and Electronics Engineers* 61: 1443–1483.

Smirnov FA (1992) Form Factors in Completely Integrable Models of Quantum Field Theory. *Advanced Series in Mathematical Physics*, vol. 14. Singapore: World Scientific.

Thacker HB (1981) Exact integrability in quantum field theory and statistical systems. *Reviews of Modern Physics* 53: 253–285.

Zamolodchikov AB and Zamolodchikov AlB (1979) Factorized S-matrices in two dimensions as the exact solutions of certain relativistic quantum field theory models. *Annals of Physics (NY)* 120: 253–291.

Singularities of the Ricci Flow

M Anderson, State University of New York at Stony Brook, Stony Brook, NY, USA

Introduction

Fix a closed n-dimensional manifold M, and let $\mathbb{M}$ be the space of Riemannian metrics on M. As in the reasoning leading to the Einstein equations in general relativity, there is basically a unique simple and natural vector field on the space $\mathbb{M}$. Namely, the tangent space $T_g\mathbb{M}$ consists of symmetric bilinear forms; besides multiples of the metric itself, the Ricci curvature Ric_g of g is the only symmetric form that depends on at most the second derivatives of the metric, and is invariant under coordinate changes, that is, a (0, 2)-tensor formed from the metric. Thus, consider

$$X_g = \mu \mathrm{Ric}_g + \lambda g$$

where μ, λ are scalars. Setting $\mu = -2$, the corresponding equation for the flow of X is

$$\frac{\mathrm{d}}{\mathrm{d}t} g(t) = -2\mathrm{Ric}_{g(t)} + \lambda g(t) \qquad [1]$$

The Ricci flow, introduced by Hamilton (1982), is obtained by setting $\lambda = 0$:

$$\frac{\mathrm{d}}{\mathrm{d}t} g(t) = -2\mathrm{Ric}_{g(t)} \qquad [2]$$

Rescaling the metric and time variable t transforms [2] into [1], with $\lambda = \lambda(t)$. For example, rescaling the Ricci flow [2] so that the volume of $(M, g(t))$ is preserved leads to the flow equation [1] with $\lambda = 2 \fint R$, twice the mean value of the scalar curvature R.

The Ricci flow [2] bears some relation with the metric part of the β-function or renormalization group (RG) flow equation

$$\frac{\mathrm{d}}{\mathrm{d}t} g(t) = \beta(g(t))$$

for the two-dimensional sigma model of maps $\Sigma^2 \to M$. The β-function is a vector field on $\mathbb{M}$, invariant under diffeomorphisms, which has an expansion of the form

$$-\beta(g) = \mathrm{Ric}_g + \varepsilon \mathrm{Riem}^2 + \cdots$$

where Riem^2 is quadratic in the Riemann curvature tensor. The Ricci flow corresponds to the one-loop term or semiclassical limit in the RG flow (cf. D'Hoker (1999) and Friedan (1985)).

Recently, G Perelman (2002, 2003a, b) has developed new insights into the geometry of the Ricci flow which has led to a solution of long-standing mathematical conjectures on the structure of 3-manifolds, namely the Thurston geometrization conjecture (Thurston 1982), and hence the Poincaré conjecture.

Basic Properties of the Ricci Flow

In charts where the coordinate functions are locally defined harmonic functions in the metric $g(t)$, [2] takes the form

$$\frac{\mathrm{d}}{\mathrm{d}t} g_{ij} = \Delta g_{ij} + Q_{ij}(g, \partial g)$$

where Δ is the Laplace operator on functions with respect to the metric $g = g(t)$ and Q is a lower-order term quadratic in g and its first-order partial derivatives. This is a nonlinear heat-type equation for g_{ij} and leads to the existence and uniqueness of solutions to the Ricci flow on some time interval starting at any smooth initial metric. This is the reason for the minus sign in [2]; a plus sign gives a backwards heat-type equation, which has no solutions in general.

The flow [2] gives a natural method to try to construct canonical metrics on the manifold M. Stationary points of the flow [2] are Ricci-flat metrics, while stationary points of the flow [1] are (Riemannian) Einstein metrics, where $\mathrm{Ric}_g = (R/n)g$, with R the scalar curvature of g. One of Hamilton's motivations for studying the Ricci flow were results on an analogous question for nonlinear sigma models. Consider maps f between Riemannian manifolds M, N with Lagrangian given by the Dirichlet energy. Eells–Sampson studied the heat equation for this action and proved that when the target N has nonpositive curvature, the flow exists for all time and converges to a stationary point of the action, that is, a harmonic map $f_\infty : M \to N$. The idea is to see if an analogous program can be developed on the space of metrics $\mathbb{M}$.

There are a number of well-known obstructions to the existence of Einstein metrics on manifolds, in particular, in dimensions 3 and 4. Thus, the Ricci flow will not exist for all time on a general manifold. Hence, it must develop singularities. A fundamental issue is to try to relate the structure of the singularities of the flow with the topology of the underlying manifold M.

A few simple qualitative features of the Ricci flow [2] are as follows: if $\mathrm{Ric}(x, t) > 0$, then the flow contracts the metric $g(t)$ near x, to the future, while

if $\mathrm{Ric}(x,t) < 0$, then the flow expands $g(t)$ near x. At a general point, there will be directions of positive and negative Ricci curvature, along which the metric locally contracts or expands. The flow preserves product structures of metrics, and preserves the isometry group of the initial metric.

The form of [2] shows that the Ricci flow continues as long as Ricci curvature remains bounded. On a bounded time interval where $\mathrm{Ric}_{g(t)}$ is bounded, the metrics $g(t)$ are quasi-isometric, that is, they have bounded distortion compared with the initial metric $g(0)$. Thus, one needs to consider evolution equations for the curvature, induced by the flow for the metric. The simplest of these is the evolution equation for the scalar curvature R:

$$\frac{\mathrm{d}}{\mathrm{d}t}R = \Delta R + 2|\mathrm{Ric}|^2 \qquad [3]$$

Evaluating [3] at a point realizing the minimum $R_{\min}$ of R on M shows that $R_{\min}$ is monotone nondecreasing along the flow. In particular, the Ricci flow preserves positive scalar curvature. Moreover, if $R_{\min}(0) > 0$, then

$$t \leq \frac{n}{2R_{\min}(0)} \qquad [4]$$

Thus, the Ricci flow exists only up to a maximal time $T \leq n/2R_{\min}(0)$ when $R_{\min}(0) > 0$. In contrast, in regions where the Ricci curvature stays negative definite, the flow exists for infinite time.

The evolution of the Ricci curvature has the same general form as [3]:

$$\frac{\mathrm{d}}{\mathrm{d}t}R_{ij} = \Delta R_{ij} + \widetilde{Q}_{ij} \qquad [5]$$

The expression for $\widetilde{Q}$ is much more complicated than the Ricci curvature term in [3] but involves only quadratic expressions in the curvature. However, $\widetilde{Q}$ involves the full Riemann curvature tensor Riem of g, and not just the Ricci curvature (as [3] involves Ricci and not just scalar curvature). An important feature of dimension 3 is that the full Riemann curvature Riem is determined algebraically by the Ricci curvature. So the Ricci flow has a much better chance of "working" in dimension 3. For example, an analysis of $\widetilde{Q}$ shows that the Ricci flow preserves positive Ricci curvature in dimension 3; if $\mathrm{Ric}_{g(0)} > 0$, then $\mathrm{Ric}_{g(t)} > 0$, for $t > 0$. This is not the case in higher dimensions. On the other hand, in any dimension >2, the Ricci flow does not preserve negative Ricci curvature, or even a general lower bound $\mathrm{Ric} \geq -\lambda$, for $\lambda > 0$. For the remainder of the article, we usually assume then that $\dim M = 3$.

The first basic result on the Ricci flow is the following, due to Hamilton (1982).

Space-form theorem. If $g(0)$ is a metric of positive Ricci curvature on a 3-manifold M, then the volume normalized Ricci flow exists for all time, and converges to the round metric on S^3/Γ, where Γ is a finite subgroup of SO(4), acting freely on S^3.

Thus, the Ricci flow "geometrizes" 3-manifolds of positive Ricci curvature. There are two further important structural results on the Ricci flow.

Curvature pinching estimate (Hamilton 1982, Ivey 1993). For $g(t)$ a solution to the Ricci flow on a closed 3-manifold M, there is a nonincreasing function $\phi : (-\infty, \infty) \to \mathbb{R}$, tending to 0 at ∞, and a constant C, depending only on $g(0)$, such that,

$$\mathrm{Riem}(x,t) \geq -C - \phi(R(x,t)) \cdot |R(x,t)| \qquad [6]$$

This estimate does not imply a lower bound on $\mathrm{Riem}(x,t)$ uniform in time. However, when combined with the fact that the scalar curvature $R(x,t)$ is uniformly bounded below (cf. [3]), it implies that $|\mathrm{Riem}|(x,t) \gg 1$ only where $R(x,t) \gg 1$. To control the size of $|\mathrm{Riem}|$, it thus suffices to obtain just an upper bound on R. This is remarkable, since the scalar curvature is a much weaker invariant of the metric than the full curvature. Moreover, at points where the curvature is sufficiently large, [6] shows that $\mathrm{Riem}(x,t)/R(x,t) \geq -\delta$, for δ small. Thus, if one scales the metric to make $R(x,t) = 1$, then $\mathrm{Riem}(x,t) \geq -\delta$. In such a scale, the metric then has almost non-negative curvature near (x,t).

Harnack estimate (Hamilton 1982). Let $(N, g(t))$ be a solution to the Ricci flow with bounded and non-negative curvature $\mathrm{Riem} \geq 0$, and suppose $g(t)$ is a complete Riemannian metric on N. Then for $0 < t_1 \leq t_2$,

$$R(x_2,t_2) \geq \frac{t_1}{t_2}\exp\left(-\frac{d^2_{t_1}(x_1,x_2)}{2(t_2-t_1)}\right)R(x_1,t_1) \qquad [7]$$

where d_{t_1} is the distance function on (M, g_{t_1}). This allows one to control the geometry of the solution at different spacetime points, given control at an initial point.

Singularity Formation

The deeper analysis of the Ricci flow is concerned with the singularities that arise in finite time. Equation [3] shows that the Ricci flow will not exist for arbitrarily long time in general. In the case of initial metrics with positive Ricci curvature, this is resolved by rescaling the Ricci flow to constant volume. However, the general situation is necessarily much more complicated. For example, any manifold which is a connected sum of S^3/Γ or $S^2 \times S^1$ factors has metrics of positive scalar curvature. For obvious

topological reasons, the volume normalized Ricci flow then cannot converge nicely to a round metric; even the renormalized flow must develop singularities.

The usual method to understand the structure of singularities, particularly in geometric PDEs, is to rescale or renormalize the solution on a sequence converging to the singularity to make the solution bounded, and try to pass to a limit of the renormalization. Such a limit solution models the singularity formation, and one hopes (or expects) that the singularity models have special features making them much simpler than an arbitrary solution of the flow.

A singularity forms for the Ricci flow only where the curvature becomes unbounded. Suppose then that $\lambda_i^2 = |\mathrm{Riem}|(x_i, t_i) \to \infty$, on a sequence of points $x_i \in M$, and times $t_i \to T < \infty$. Consider the rescaled or blow-up metrics and times

$$\bar{g}_i(\bar{t}_i) = \lambda_i^2 \phi_i^* g(t), \quad \bar{t}_i = \lambda_i^2 (t - t_i) \qquad [8]$$

where ϕ_i are diffeomorphisms giving local dilations of the manifold near x_i by the factor λ_i.

The flow $\bar{g}_i$ is also a solution of the Ricci flow, and has bounded curvature at $(x_i, 0)$. For suitable choices of x_i and t_i, the curvature will be bounded near x_i, and for nearby times to the past, $\bar{t}_i \le 0$; for example, one might choose points (x_i, t_i) where the curvature is maximal on $(M, g(t)), t \le t_i$.

The rescaling [8] expands all distances by the factor λ_i, and time by the factor λ_i^2. Thus, in effect one is studying very small regions, of spatial size on the order of $r_i = \lambda_i^{-1}$ about (x_i, t_i), and "using a microscope" to examine the small-scale features in this region on a scale of size about 1.

A limit solution of the Ricci flow, defined at least locally in space and time, will exist provided that the local volumes of the rescalings are bounded below (Gromov compactness). In terms of the original unscaled flow, this requires that the metric $g(t)$ should not be locally collapsed on the scale of its curvature, that is,

$$\mathrm{vol}\, B_{x_i}(r_i, t_i) \ge \nu r_i^n \qquad [9]$$

for some fixed but arbitrary $\nu > 0$. A maximal connected limit $(N, \bar{g}(\bar{t}), x)$ containing the base point $x = \lim x_i$, is then called a "singularity model." Observe that the topology of the limit N may well be distinct from the original manifold M, most of which may have been blown off to infinity in the rescaling.

To see the potential usefulness of this process, suppose one does have local noncollapse on the scale of the curvature, and that base points of maximal curvature in space and time $t \le t_i$ have been chosen. At least in a subsequence, one then obtains a limit solution to the Ricci flow $(N, \bar{g}(\bar{t}), x)$, based at x, defined at least for times $(-\infty, 0]$, with $\bar{g}(\bar{t})$ a complete Riemannian metric on N. Such solutions are called ancient solutions of the Ricci flow. The estimate [6] shows that the limit has non-negative curvature in dimension 3, and so [7] holds on N. Thus, the limit is indeed quite special. The topology of complete manifolds N of non-negative curvature is completely understood in dimension 3. If N is noncompact, then N is diffeomorphic to $\mathbb{R}^3, S^2 \times \mathbb{R}$, or a quotient of these spaces. If N is compact, then a slightly stronger form of the space-form theorem implies N is diffeomorphic to $S^3/\Gamma, S^2 \times S^1$, or $S^2 \times_{\mathbb{Z}_2} S^1$.

The study of the formation of singularities in the Ricci flow was initiated by Hamilton (1995). Recently, Perelman has obtained an essentially complete understanding of the singularity behavior of the Ricci flow, at least in dimension 3.

Perelman's Work

Noncollapse

Consider the Einstein–Hilbert action

$$\mathcal{R}(g) = \int_M R(g)\, \mathrm{d}V_g \qquad [10]$$

as a functional on $\mathbb{M}$. Critical points of $\mathcal{R}$ are Ricci-flat metrics. It is natural and tempting to try to relate the Ricci flow with the gradient flow of $\mathcal{R}$ (with respect to a natural L^2 metric on the space $\mathbb{M}$). However, it has long been recognized that this cannot be done directly. In fact, the gradient flow of $\mathcal{R}$ does not even exist, since it implies a backwards heat-type equation for the scalar curvature R (similar to [3] but with a minus sign before Δ).

Consider however the following functional extending $\mathcal{R}$:

$$\mathcal{F}(g, f) = \int_M (R + |\nabla f|^2) \mathrm{e}^{-f}\, \mathrm{d}V_g \qquad [11]$$

as a functional on the larger space $\mathbb{M} \times C^\infty(M, \mathbb{R})$, or equivalently a family of functionals on $\mathbb{M}$, parametrized by $C^\infty(M, \mathbb{R})$. The functional [11] also arises in string theory as the low-energy effective action; the scalar field f is called the dilaton. Fix any smooth measure $\mathrm{d}m$ on M and define the Perelman coupling by requiring that (g, f) satisfy

$$\mathrm{e}^{-f}\, \mathrm{d}V_g = \mathrm{d}m \qquad [12]$$

The resulting functional

$$\mathcal{F}^m(g,f)=\int_M (R+|\nabla f|^2)\,\mathrm{d}m \qquad [13]$$

becomes a functional on $\mathbb{M}$. (This coupling does not appear to have been considered in string theory.) The L^2 gradient flow of $\mathcal{F}^m$ is given simply by

$$\frac{\mathrm{d}\widetilde{g}}{\mathrm{d}t}=-2(\mathrm{Ric}_{\widetilde{g}}+\widetilde{D}^2 f) \qquad [14]$$

where $\widetilde{D}^2 f$ is the Hessian of f with respect to $\widetilde{g}$. The evolution equation [14] for $\widetilde{g}$ is just the Ricci flow [2] modified by an infinitesimal diffeomorphism: $\widetilde{D}^2 f=(\mathrm{d}/\mathrm{d}t)(\phi_t^*\widetilde{g})$, where $(\mathrm{d}/\mathrm{d}t)\phi_t=\widetilde{\nabla} f$. Thus, the gradient flow of $\mathcal{F}^m$ is the Ricci flow, up to diffeomorphisms. The evolution equation for the scalar field f,

$$f_t=-\widetilde{\Delta} f-\widetilde{R} \qquad [15]$$

is a backward heat equation (balancing the forward evolution of the volume form of $\widetilde{g}(t)$). Thus, this flow will not exist for general f, going forward in t. However, one of the basic points of view is to let the (pure) Ricci flow [2] flow for a time $t_0>0$. At t_0, one may then take an arbitrary $f=f(t_0)$ and flow this f backward in time ($\tau=t_0-t$) to obtain an initial value $f(0)$ for f. The choice of $f(t_0)$ determines, together with the choice of volume form of $g(0)$), (or $g(t_0)$), the measure $\mathrm{d}m$ and so the choice of $\mathcal{F}^m$. The process of passing from $\mathcal{F}$ to $\mathcal{F}^m$ corresponds to a reduction of the symmetry group of all diffeomorphisms $\mathcal{D}$ of $\mathcal{F}$ to the group $\mathcal{D}_0$ of volume-preserving diffeomorphisms; the quotient space $\mathcal{D}/\mathcal{D}_0$ has been decoupled into a space $C^\infty(M,\mathbb{R})$ of parameters.

The functionals $\mathcal{F}^m$ are not scale invariant. To achieve scale invariance, Perelman includes an explicit insertion of the scale parameter, related to time, by setting

$$\mathcal{W}(g,f,\tau)=\int\Big(\tau(|\nabla f|^2+R)+f-n\Big)\times(4\pi\tau)^{-n/2}\,\mathrm{e}^{-f}\,\mathrm{d}V \qquad [16]$$

with coupling so that $\mathrm{d}m=(4\pi\tau)^{-n/2}\,\mathrm{e}^{-f}\,\mathrm{d}V$ is fixed. The entropy functional $\mathcal{W}$ is invariant under simultaneous rescaling of τ and g, and $\tau_t=-1$. Again, the gradient flow of $\mathcal{W}$ is the Ricci flow modulo diffeomorphisms and rescalings and the stationary points of the gradient flow are the gradient Ricci solitons,

$$\mathrm{Ric}_g+D^2 f-\frac{1}{2\tau}g=0$$

for which the metrics evolve by diffeomorphisms and rescalings. Gradient solitons arise naturally as singularity models, due to the rescalings and diffeomorphisms in the blow-up procedure [8]. An important example is the cigar soliton on $\mathbb{R}^2\times\mathbb{R}$, (or $\mathbb{R}^2\times S^1$),

$$g=(1+r^2)^{-1}g_{\mathrm{Eucl}}+\mathrm{d}s^2 \qquad [17]$$

Perelman then uses the scalar field f to probe the geometry of $g(t)$. For instance, the collapse or noncollapse of the metric $g(t)$ near a point $x\in M$ can be detected from the size of $\mathcal{W}(g(t))$ by choosing e^{-f} to be an approximation to a delta function centered at (x,t). The more collapsed $g(t)$ is near x, the more negative the value of $\mathcal{W}(g(t))$. The collapse of the metric $g(t)$ on any scale in finite time is then ruled out by combining this with the fact that the entropy functional $\mathcal{W}$ is increasing along the Ricci flow.

Much more detailed information can be obtained by studying the path integral associated to the evolution equation [15] for f, given by

$$\mathcal{L}(\gamma)=\int_\gamma \sqrt{\tau}[|\dot{\gamma}(\tau)|^2+R(\gamma(\tau))]\,\mathrm{d}\tau$$

where R and $|\dot{\gamma}(\tau)|$ are computed with respect to the evolving metrics $g(\tau)$. In particular, the study of the geodesics and the associated variational theory of the length functional $\mathcal{L}$ are important in understanding the geometry of the Ricci flow near the singularities.

Singularity Models

A major accomplishment of Perelman is essentially a classification of all complete singularity models $(N,g(t))$ that arise in finite time. In the simple case where N is compact, then as noted above, N is diffeomorphic to S^3/Γ, $S^2\times S^1$, or $S^2\times_{\mathbb{Z}_2} S^1$.

In the much more important case where N is complete and noncompact, Perelman proves that the geometry of N near infinity is that of a union of ε-necks. Thus, at time 0, and at points x with $r(x)=\mathrm{dist}(x,x_0)\gg 1$, for a fixed base point x_0, a region of radius ε^{-1} about x, in the scale where $R(x)=1$, is ε-close to such a region in the standard round product metric on $S^2\times\mathbb{R}$; ε may be made arbitrarily small by choosing $r(x)$ sufficiently large. For example, this shows that the cigar soliton [17] cannot arise as a singularity model. Moreover, this structure also holds on a time interval on the order of ε^{-1} to the past, so that on such regions the solution is close to the (backwards) evolving Ricci flow on $S^2\times\mathbb{R}$.

Perelman shows that this structural result for the singularity models themselves also holds for the solution $g(t)$ very near any singularity time T. Thus, at any base point (x, t) where the curvature is sufficiently large, the rescaling as in [8] of the spacetime by the curvature is smoothly close, on large compact domains, to corresponding large domains in a complete singularity model. The "ideal" complete singularity models do actually describe the geometry and topology near any singularity. Consequently, one has a detailed understanding of the small-scale geometry and topology in a neighborhood of every point where the curvature is large on $(M, g(t))$, for t near T.

The main consequence of this analysis is the existence of canonical, almost round 2-spheres S^2 in any region of $(M, g(t))$ where the curvature is sufficiently large; the radius of the S^2's is on the order of the curvature radius. One then disconnects the manifold M into pieces, by cutting M along a judicious choice of such 2-spheres, and gluing in round 3-balls in a natural way. This surgery process allows one to excise out the regions of $(M, g(t))$ where the Ricci flow is almost singular, and thus leads to a naturally defined Ricci flow with surgery, valid for all times $t \in [0, \infty)$.

The surgery process disconnects the original connected 3-manifold M into a collection of disjoint (connected) 3-manifolds M_i, with the Ricci flow running on each. However, topologically, there is a canonical relation between M and the components M_i; M is the connected sum of $\{M_i\}$. An analysis of the long-time behavior of the volume-normalized Ricci flow confirms the expectation that the flow approaches a fixed point, that is, an Einstein metric, or collapses along 3-manifolds admitting an S^1 fibration. This then leads to the proof of Thurston's geometrization conjecture for 3-manifolds and consequently the proof of the Poincaré conjecture. It gives a full classification of all closed 3-manifolds, much like the classification of surfaces given by the classical uniformization theorem.

See also: Einstein Manifolds; Evolution Equations: Linear and Nonlinear; Minimal Submanifolds; Renormalization: General Theory; Topological Sigma Models.

Further Reading

Anderson M (1997) Scalar curvature and geometrization conjectures for 3-manifolds. In: Grove K and Petersen P (eds.) *Comparison Geometry*, MSRI Publications, vol. 30, pp. 49–82. Cambridge: Cambridge University Press.

Cao HT, Chow B, Cheng SC, and Yau ST (eds.) (2003) *Collected Papers on Ricci Flow*. Boston: International Press.

D'Hoker E (1999) String theory. In: Deligne P *et al.* (eds.) *Quantum Fields and Strings: A Course for Mathematicians*, vol. 2, Providence, RI: American Mathematical Society.

Friedan D (1985) Nonlinear models in $2 + \varepsilon$ dimensions. *Annals of Physics* 163: 318–419.

Hamilton R (1982) Three manifolds of positive Ricci curvature. *Journal of Differential Geometry* 17: 255–306.

Hamilton R (1993) The Harnack estimate for the Ricci flow. *Journal of Differential Geometry* 37: 225–243.

Hamilton R (1995) Formation of singularities in the Ricci flow. In: Yau ST (ed.) *Surveys in Differential Geometry* vol. 2, pp. 7–136. Boston, MA: International Press.

Ivey T (1993) Ricci solitons on compact three-manifolds. *Differential Geometry and Its Applications* 3: 301–307.

Perelman G (2002) *The entropy formula for the Ricci flow and its geometric applications*, math.DG/0211159.

Perelman G (2003a) *Ricci flow with surgery on three-manifolds*, math.DG/0303109.

Perelman G (2003b) *Finite extinction time for the solutions to the Ricci flow on certain three-manifolds*, math.DG/0307245.

Thurston W (1982) Three dimensional manifolds, Kleinian groups and hyperbolic geometry. *Bulletin of American Mathematical Society* 6: 357–381.

Singularity and Bifurcation Theory

J-P Françoise and C Piquet, Université P.-M. Curie, Paris VI, Paris, France

Introduction

Dynamical systems first developed from the geometry of Newton's equations (see Goodstein and Goodstein (1997)) and the question of the stability of the solar system motivated further researches inspired by celestial mechanics (cf. Siegel and Moser (1971)). Then dynamical systems developed intensively from stability theory (Lyapunov's theory) to generic properties (based on functional analysis techniques,) hyperbolic structures (Anosov's flows, Smale axiom A) and to perturbation theory (Pugh's closing lemma, KAM theorem). There are many links with ergodic theory dating back to Birkhoff's ergodic theorem (motivated by Boltzmann–Gibbs contributions to thermodynamics). These aspects have been developed in several articles of the encyclopedia (*see* Generic Properties of

Dynamical Systems; Ergodic Theory; Hyperbolic Dynamical Systems). This article develops another aspect of dynamical systems, namely bifurcation theory. In contrast, the mathematics involved relates more to local analytic geometry in the broad sense and provides local models like normal forms, uses blow-up techniques and asymptotic developments. This contains the singularity theory of functions (related to singularities of gradient flows). A recent development of the whole subject deals with bifurcation theory of fast-slow systems.

Singularity Theory of Functions

A singular point of a gradient dynamics

$$\frac{dx}{dt} = \text{grad } V(x)$$

is a critical point of the function V. Assume that the function $V: U \to R$ is defined and infinitely differentiable on an open set U. Let $x_0 \in U$ be a critical point of V.

Definition 1 The critical point x_0 is said to be of Morse type if the Hessian of V at $x_0 : D_x^2 V(x_0)$ is of maximal rank n. The corank of a singular point x_0 is the corank of the matrix $D_x^2 V(x_0)$.

Denote by O the local ring of germs of C^∞ functions at point x_0.

Definition 2 The Jacobian ideal of the function V at x_0, denoted as Jac(V), is the ideal generated in the ring O by the partial derivatives of $V: \partial V/\partial x_i, i = 1, \ldots, n$, considered as elements of the local ring O.

The singularity (or the singular point) is isolated if

$$\dim_R O/\text{Jac}(V) < \infty$$

In that case, the Milnor number is defined as the dimension

$$\mu = \dim_R O/\text{Jac}(V)$$

Local models of singularities at a point are simple expressions that germs of functions singular at this point have in local coordinates.

R Thom proposed to focus more particularly on the singularities whose Milnor number is less than or equal to 4 and whose corank is less than or equal to 2.

The list of local models $V_\lambda(x)$ of functions whose singularities at 0 display a Milnor number less than or equal to 4 and a corank less than or equal to 2 is the following:

$V_\lambda(x) = \frac{1}{3}x^3 + \lambda_1 x$, the *fold*,

$V_\lambda(x) = \frac{1}{4}x^4 + \frac{1}{2}\lambda_1 x^2 + \lambda_2 x$, the *cusp*,

$V_\lambda(x) = \frac{1}{5}x^5 + \frac{1}{3}\lambda_1 x^3 + \frac{1}{2}\lambda_2 x^2 + \lambda_3 x$, the *swallow tail*,

$V_\lambda(x) = \frac{1}{6}x^6 + \frac{1}{4}\lambda_1 x^4 + \frac{1}{3}\lambda_2 x^3 + \frac{1}{2}\lambda_3 x^2 + \lambda_4 x$, the *butterfly*,

$V_\lambda(x) = x^3 - 3xy^2 + \lambda_1(x^2 + y^2) + \lambda_2 x + \lambda_3 y$, the *elliptic umbilic*,

$V_\lambda(x) = x^3 + y^3 + \lambda_1 xy + \lambda_2 x + \lambda_3 y$, the *hyperbolic umbilic*, and

$V_\lambda(x) = y^4 + x^2 y + \lambda_1 x^2 + \lambda_2 y^2 + \lambda_3 x + \lambda_4 y$, the *parabolic umbilic*.

Consider more particularly the first four cases. The "state equation" defines the critical points of V_λ:

$$\frac{\partial V_\lambda}{\partial x} = 0$$

which contains the subset of the stable equilibrium points of the associated gradient dynamics. The nature of these equilibrium states changes at points contained in the set defined by the equation

$$\frac{\partial^2 V_\lambda}{\partial x^2} = 0$$

The projection of this set on the space of parameters contains the set of values of the parameters for which the equilibrium position is susceptible to change of topological type (in other terms to undergo a bifurcation). This set is called the catastrophe set (see **Figure 1**).

Consider now the case of umbilics where there are two state equations:

$$\frac{\partial V}{\partial x} = \frac{\partial V}{\partial y} = 0$$

The catastrophe set S is determined by one further equation:

$$\text{Hess } V = \frac{\partial^2 V}{\partial x^2}\frac{\partial^2 V}{\partial y^2} - \left(\frac{\partial^2 V}{\partial x \partial y}\right)^2 = 0$$

In both cases of hyperbolic and elliptic umbilics, the set S is a singular surface. For the last case of the parabolic umbilic, the set S is of dimension 3 and again it is only possible to represent it by a family of its sections by a variable hyperplane (see **Figure 2**).

All possible deformations (in the space of functions) of a function with an isolated singularity can be induced by a single μ-dimensional family of deformations named the "universal deformation." In general, the "codimension" of a bifurcation is the minimal number of parameters needed to display all possible phase diagrams of all possible unfoldings. Several deep mathematical techniques, like the Malgrange division theorem and preparation theorem, allowed J Mather to prove the theorem (local, then global) of existence of the universal unfolding.

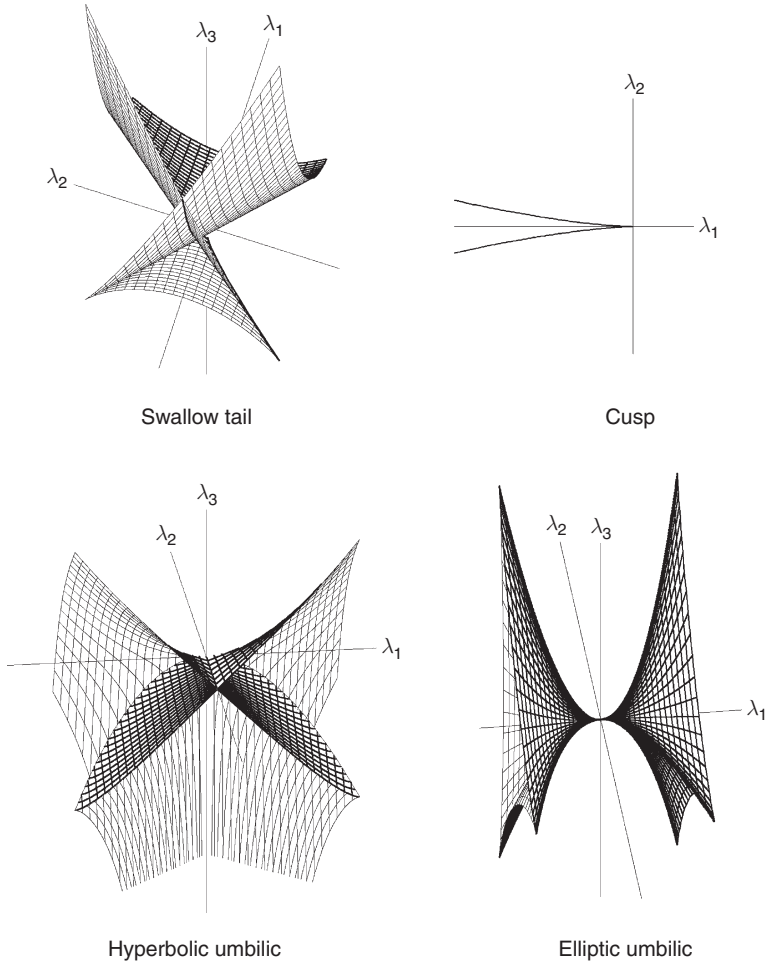


Figure 1 Examples of catastrophe sets. Adapted with permission from Françoise J-P (2005) *Oscillations en Biologie: Analyse Qualitative et Modèles* (Mathématiques et Applications, vol. 46). Heidelberg: Springer.

The theory of unfoldings of singularities can be used, for instance, to provide asymptotic expression of stationary phase integrals when critical points of the phase are not of Morse type. This relates to monodromy, Bernstein polynomials, Milnor fibration near a singular point, and simultaneous local models of forms and functions (cf. Malgrange (1974)) and *see* Feynman Path Integrals).

Singularity Theory of Vector Fields

Transcritical Bifurcation

The transcritical bifurcation is the standard mechanism for changes in stability. The local model is given by

$$\dot{x} = rx - x^2$$

For $r < 0$, there is an unstable fixed point at $x^* = r$ and a stable fixed point at $x^* = 0$. As r increases, the unstable and the stable fixed points coalesce when $r = 0$ and when $r > 0$, they exchange their stability. A simplified model of the essential physics of a laser is due to Haken (1983). It is given by

$$\dot{n} = GnN - kn$$

were n is the number of photons in the laser field, N is the number of excited atoms, and the gain term comes from the process of stimulated emission which occurs at a rate proportional to the product $n.N$. Furthermore, the number of excited atoms drops down by the emission of photons $N = N_0 - \alpha n$. Then we obtain

$$\dot{n} = (GN_0 - k)n - \alpha G n^2$$

This model displays a transcritical bifurcation, which explains in elementary terms the laser threshold.

Pitchfork Bifurcation

The local model for supercritical pitchfork bifurcation is

$$\dot{x} = rx - x^3$$

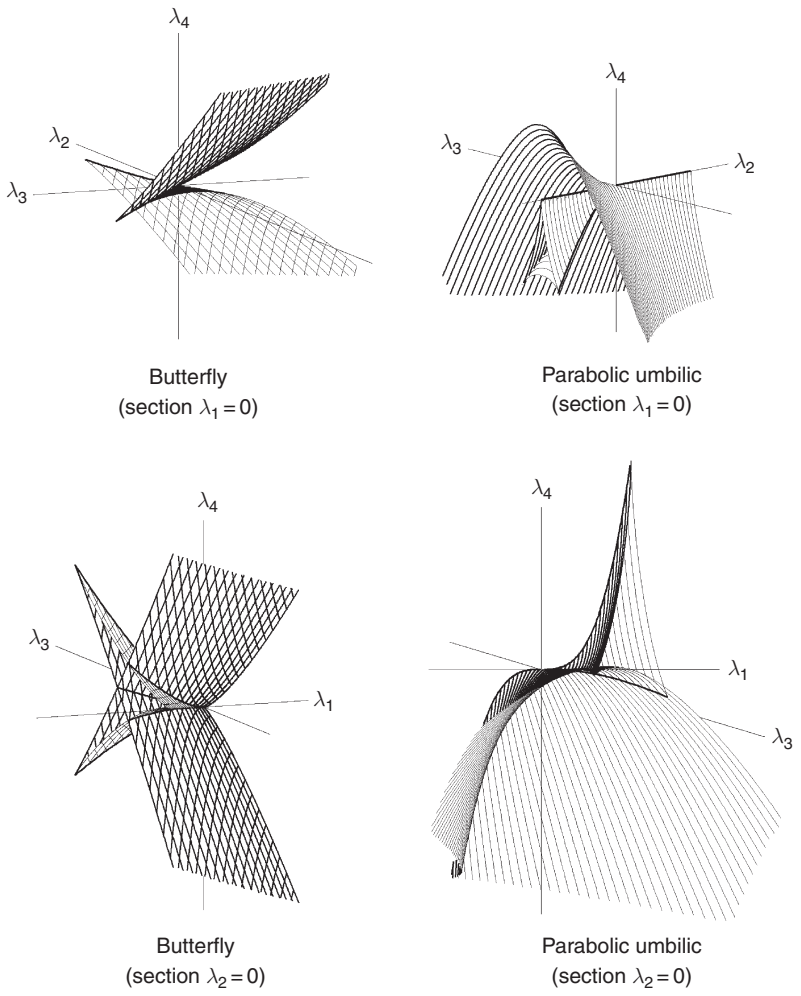


Figure 2 Sections of catastrophe sets. Adapted with permission from Françoise J-P (2005) *Oscillations en Biologie: Analyse Qualitative et Modèles* (Mathématiques et Applications, vol. 46). Heidelberg: Springer.

When the parameter $r < 0$, it displays one stable equilibrium position. As r increases, this equilibrium bifurcates (for $r > 0$) into two stable equilibria and an unstable equilibrium. Its drawing suggests "the pitchfork." In case of subcritical pitchfork bifurcation

$$\dot{x} = rx + x^3$$

there is a single stable state for $r < 0$ that bifurcates into two stable states and one unstable as $r > 0$.

Normal Forms

Local analysis of vector fields proceeds with local models called normal forms. A local vector field near a singular point (zero) is seen as a derivation of the local ring of functions which preserves the unique maximal ideal (of the functions which vanish at the singular point). It yields a linear operator of the finite-dimensional vector spaces of truncated Taylor expansions of functions. This leads to decomposition of the vector fields into semisimple and nilpotent parts (at the level of formal series). A normal form is a formal coordinate system in which the semisimple part is linear. If the vector field preserves a structure (like volume form or symplectic form) the change of coordinates which brings it to its normal form is also (volume-preserving, symplectic). The simplicity of the normal form depends on the number of allowed resonances for the eigenvalues of the first-order jet of the vector field at the singular point. The best-known example is the Birkhoff normal form of Hamiltonian vector fields that we recall now, but we should also mention the Sternberg normal form of volume-preserving vector fields.

Local analysis of a Hamiltonian vector field under symplectic changes of coordinates is the same as the local analysis of functions (namely its associated Hamiltonian). Birkhoff normal form deals with the

case of a Hamiltonian that is a perturbation at the origin:

$$H_0(p) = \sum_{j=1}^{m} \lambda_j p_j$$

$$p_j = x_j^2 + y_j^2, \quad j = 1, \ldots, m$$

where the symplectic form is

$$\omega = \sum_{j=1}^{m} \mathrm{d}x_j \wedge \mathrm{d}y_j$$

If the eigenvalues λ_j are assumed to be independent over the integers (no resonances), then there is a formal system of symplectic coordinates $\hat{p}_j, \hat{q}_j$, $j = 1, \ldots, m$, called action-angle variables, in which the Hamiltonian only depends of the action variables $\hat{p}_j$. Such a coordinate system is generically divergent because, under generic assumptions on the 3-jet of the Hamiltonian, the system displays isolated periodic orbits in any neighborhood of the origin (see Moser, Vey, Francoise). Normal forms are normally used in applications (e.g., Nekhoroshev theorem, Hopf bifurcation theorem) in their truncated versions. Birkhoff normal form was conjectured (A Weinstein) to enter in the asymptotic expansion of the fundamental solution of the wave equation on a Riemannian manifold near elliptic geodesics. This conjecture was recently proved by V Guillemin.

Stability Theory of Hamiltonian Systems, Nekhoroshev Theorem, Arnol'd Diffusion

The generic divergence of the Birkhoff normal form does not allow one to conclude about the stability of the elliptic singular point. In the case where it is convergent, the motion is trapped inside invariant tori (conservation of the actions). The KAM theorem (see Gallavotti (1983)) provides the existence of many invariant tori but, except in low dimensions, this does not exclude the existence of trajectories that would escape to infinity. Arnol'd indeed provided a mechanism and examples of such situations (this is now called Arnol'd diffusion) (*see* Introductory Articles: Classical Mechanics). This diffusion process needs some time, which is estimated below by a theorem of Nekhoroshev.

Consider the Hamiltonian

$$H_\epsilon(p, q) = h(p) + \epsilon f(p, q)$$

where $h(p)$ is strictly convex, analytic, anisochronous on the closure $\overline{U}$ of an open bounded region U of R^m and the perturbation $f(p, q)$ is analytic on $\overline{U} \times R^m$. Nekhoroshev's theorem tells that there are positive constants a, b, d, g, τ such that for any initial data p_0, q_0, the actions p do not change by more than $a\epsilon^g$ before a time bounded below by $\tau e^{b/\epsilon^d}$ (see Gallavotti (1983)).

Bifurcations of Periodic Orbits

Consider a one-parameter family of vector fields X_λ of class $C^k, k \geq 3$,

$$\dot{x} = F(x, \lambda)$$

Assume that $X_\lambda(0) = 0$ and that the linear part of the vector field at 0 has two complex-conjugated eigenvalues $\mu(\lambda)$ and $\overline{\mu(\lambda)}$ such that $\mathrm{Re}(\mu(\lambda)) > 0$ for $\lambda > 0, \mathrm{Re}(\mu(0)) = 0$ and $(\mathrm{Re}(\mu(\lambda)))/d\lambda|_{\lambda=0} \neq 0$. Then, for $\lambda > 0$ but small enough, the vector field X_λ has a periodic orbit γ_λ which tends to 0 as λ tends to 0.

This bifurcation of codimension 1 is named Hopf bifurcation and it occurs in many models.

When several oscillators (conservative or dissipative) are weakly coupled, they may display frequency locking (existence of an attractive periodic orbit) phase locking, and synchronization. The fact that we always see the same face of the Moon from the Earth can be explained by a synchronization of the rotation of the Moon onto itself with its rotation around the Earth. Synchronization also plays a fundamental role in living organisms (e.g., heart, population dynamics: see D Attenborough's movie "The Trials of Life"). It is sometimes possible to be convinced of synchronization via computer experiments, but the main theoretical approach is due to Malkin. *See* Bifurcations of Periodic Orbits, where a full mathematical proof is included.

Homoclinic Bifurcation, Newhouse's Phenomenon

Homoclinic bifurcation occurs in the family X_λ at the bifurcation value of the parameter $\lambda = 0$ if X_0 displays a singular orbit which tends to 0 both for $t \to +\infty$ and for $t \to -\infty$. In dimension 2, if λ is slightly deformed around 0, one periodic orbit may appear (or disappear). For planar systems, the Bogdanov–Takens bifurcation is the codimension-2 bifurcation, which mixes the homoclinic and the Hopf bifurcations. In dimension 3, more complicated phase diagrams may occur (such as in the Shilnikov bifurcation) with the appearance of infinitely many periodic orbits or homoclinic loops (in a stable way: Newhouse phenomenon). This eventually gives rise to strange attractors (the Roessler attractor).

The Poincaré Center-Focus Problem, Local Hilbert's 16th Problem, Abel Equations, Algebraic Moments

Hopf bifurcation theory for two-dimensional systems deals with the first case of a general situation

often referred to as degeneracies of Hopf bifurcations or alternatively Hopf–Takens bifurcations. Consider more generally a planar vector field, tangent at the origin to a linear focus:

$$\dot{x} = y + \mu x + f(x, y)$$
$$\dot{y} = -x + \mu y + g(x, y)$$

The Poincaré center-focus problem asks for necessary and sufficient conditions on the perturbation terms so that all orbits are periodic in a neighborhood of the origin. This problem is still pending in the case, for instance, where f and g are homogeneous of degrees 4 and 5. It was solved a long time ago for degrees 2 and 3. Part (b) of Hilbert's 16th Problem asks for finding a bound in terms of the degrees of polynomial perturbations for the number of limit cycles (isolated periodic orbits) in the neighborhood of the origin. In the case of homogeneous perturbations, a Cherkas transformation allows the reduction of both problems to the so-called one-dimensional periodic Abel equations:

$$\mathrm{d}y/\mathrm{d}x = p(x)y^2 + q(x)y^3$$

where p and q are trigonometric polynomials in x. A perturbative approach was developed for several years and yields a theory of algebraic moments related to Livsic's generalized problem of moments.

Fast–Slow Systems

Fast–slow systems

$$\epsilon\,\dot{x} = f(x, y), \qquad \dot{y} = g(x, y)$$

are characterized by the existence of two timescales. Variables x are called fast variables and y are called slow variables. Different approximation techniques can be used (averaging method, multiscale approach (*see* Multiscale Approaches)). The behavior of solutions is approximated as follows (when the scale ϵ is small). The orbit jumps to an attractor of the fast dynamics. This attractor may eventually lose its stability and/or bifurcate as time evolves. Then the orbit jumps to another attractor of the fast dynamics. Once again, this attractor may evolve/bifurcate/disappear, depending on the slow variables y. This explains why bifurcation theory enters in the process in a crucial way, and it has to be adapted to this special context where some new phenomena may occur (e.g., singular Hopf bifurcation theory, Canards, etc.). Fundamental tools to be used in this context are Takens theorem, Fenichel central manifold theorem, blowing-up (Dumortier–Roussarie).

Excitability is also an important feature which occurs in some fast–slow systems. Consider initial data in a neighborhood of an excitable attractive point. For some initial data, the orbit goes very quickly to the attractor. For some others instead (usually below some threshold), the orbit undergoes a long incursion in the phase diagram before turning back to the attractive point.

Singular Hopf bifurcation, hysteresis, and excitability can, for instance, occur in the electrodissolution and passivation of iron in sulfuric acid (see Alligood *et al.* (1997)).

Sometimes, the orbit leaves the neighborhood of a first attractor to jump to a second one and then this second one disappears and the orbit jumps back to the initial attractor as the slow variables have undergone a cycle. This is called a hysteresis cycle. In case one of the attractors is a point while the other is an attractive periodic orbit, it may lead to bursting oscillations. These oscillations are characterized by the periodic succession of silent phases (attractor of the fast dynamics) and active (pulsatile) phases (periodic attractor of the fast dynamics). They are ubiquitous in physiology, where they were first discovered and can be also observed in physics (laser beams) and in population dynamics.

Example

The Hindmarsh–Rose model displays bursting oscillations:

$$\epsilon\dot{x} = y - x^3 + 3x^2 + I - z$$
$$\epsilon\dot{y} = 1 - 5x^2 - y$$
$$\dot{z} = s(x - x_1) - z$$

The fast dynamics is two dimensional. For some values of the parameters, it displays an attractive node, a saddle and a repulsive focus. Under the slow variation of z, the fast dynamics displays a saddle–node bifurcation, a Hopf bifurcation from which emerges a stable limit cycle which disappears into a homoclinic bifurcation. The fast–slow system undergoes a hysteresis loop which yields to bursting oscillations.

Conclusions

Over the past three decades, mathematical techniques gathered under the names of singularity theory and bifurcation theory of dynamical systems have offered a powerful means to explore nonlinear phenomena in diverse settings. These include mechanical vibrations, lasers, superconducting circuits, and chemical oscillators. Many such instances are further developed in this encyclopedia.

See also: Bifurcation Theory; Bifurcations of Periodic Orbits; Chaos and Attractors; Entropy and Quantitative Transversality; Ergodic Theory; Feynman Path Integrals; Generic Properties of Dynamical Systems; Gravitational Lensing; Homoclinic Phenomena; Hyperbolic Dynamical Systems; Multiscale Approaches; Optical Caustics; Poisson Reduction; Stationary Phase Approximation; Symmetry and Symmetry Breaking in Dynamical Systems; Symmetry and Symplectic Reduction; Synchronization of Chaos; Weakly Coupled Oscillators.

Further Reading

Alligood KT, Sauer TD, and Yorke JA (1997) *Chaos, An Introduction to Dynamical Systems, Textbooks in Mathematical Sciences*. New York: Springer.

Alpay D and Vinikov V (eds.) (2001) *Operator Theory, System Theory and Related Topics, The Mosche Livsic Anniversary Volume, Operator Theory, Advances and Applications* vol. 123. Birkhauser.

Briskin M, Francoise JP, and Yomdin (2001) *Generalized Moments, Cener-Focus Conditions and Compositions of Polynomials*. Operator Theory, Advances and Applications 123 (in honor of M Livsic, 80th birthday).

Diener M (1994) The canard unchained, or how fast–slow dynamical systems bifurcate? *The Mathematical Intelligencer* 6: 38–49.

Francoise JP and Guillemin V (1991) On the period spectrum of a symplectic map. *Journal of Functional Analysis* 100: 317–358.

Gallavotti G (1983) *The Elements of Mechanics*. New York: Springer.

Goodstein DL and Goodstein JR (1997) *Feynmann's lost lecture*. London: Vintage.

Guckenheimer J (2004) Bifurcations of relaxation oscillations. In: Ilyashenko Y, Rousseau C, and Sabidussi G (eds.) *Normal Forms, Bifurcations and Finiteness Problems in Differential Equations. Séminaire de mathématiques supérieures de Montréal*, Nato Sciences Series, II. Mathematics, vol. 137, pp. 295–316. Kluwer.

Haken H (1983) *Synergetics*, 3rd edn. Berlin: Springer.

Keener J and Sneyd J (1998) Mathematical Physiology. *Interdisciplinary Applied Mathematics*, vol. 8. New York: Springer.

Malgrange B (1974) Intégrales asymptotiques et monodromie. *Annales de l'ENS* 7: 405–430.

May R-M (1976) Simple mathematical models with very complicated dynamics. *Nature* 261: 459–467.

Nekhoroshev V (1977) An exponential estimate of the time of stability of nearly integrable Hamiltonian systems. *Russian Mathematical Surveys* 32(6): 1–65.

Palis J and de Melo W (1982) *Geometric Theory of Dynamical Systems, An Introduction*. New York: Springer.

Perko L (2000) *Differential Equations and Dynamical Systems*, 3rd edn, Text in Applied Mathematics, vol. 7. New York: Springer.

Siegel C-L and Moser J (1971) *Lectures on Celestial Mechanics*, Die Grundleheren der mathematischen Wissenschaften, vol. 187. Berlin: Springer.

Smale S (1998) Mathematical problems for the next century. *The Mathematical Intelligencer* 20: 7–15.

Smale S. Dynamics retrospective: great problems, attempts that failed. *Physica D* 51: 267–273.

Sobolev Spaces *see* Inequalities in Sobolev Spaces

Solitons and Kac–Moody Lie Algebras

E Date, Osaka University, Osaka, Japan

Introduction

Solitons and Kac–Moody Lie algebras were born at almost the same time in the 1960s, although they did not have a connection at first. They both have roots in the history of mathematics. From the 1970s on, they became intersection points for many (previously known and new) results.

The notion of solitons has many facets and it is difficult to give a mathematically precise definition; closely related to solitons is the notion of "completely integrable systems." The latter is usually used in a much broader sense.

The terminology "soliton" was originally used for a particular phenomenon in shallow water waves. Now, in its broadest sense, it is used to represent an area of research relating to this particular phenomenon in direct or indirect ways. From the viewpoint of solitons, particular solutions of differential equations are of special interest. Although particular solutions have been studied for a long time, interest in them was overshadowed by the method of functional analysis in the 1950s. In the late nineteenth century, in parallel with the theory of algebraic functions, several studies undertook the solution of mechanical problems by elliptic or hyperelliptic integrals. Subsequently, however, there was a drop in activity in this area of work.

Originally it was hoped that this kind of phenomenon could be used for practical applications. No mention of practical application of solitons will be made in this article.

First we list several topics which constitute the main body of the notion of solitons in the early stages; we will then explain relations with Kac–Moody Lie algebras.

Birth of Solitons

The name "soliton" itself was coined by Martin D Kruskal around 1965. It was originally employed for the solitary wave solution Korteweg–de Vries (KdV) equation

$$u_t - \tfrac{1}{4}(6uu_x + u_{xxx}) = 0, \quad u = u(x,t) \qquad [1]$$

The coefficients here are not important. We can change them arbitrarily. The unknown function u, or rather $-u$, represents the height of the wave.

The solitary wave solution in question is given by

$$u(x,t) = -2c\ \mathrm{sech}^2\left(\sqrt{c}(x - ct - d)\right) \qquad [2]$$

This is a traveling-wave solution with the height of the wave proportional to the speed. This is one feature of the nonlinearity of this differential equation.

A reason for this nomenclature comes from the particle-like property of solitary wave observed via numerical computations. That is, if we have two solitons [2] with different speeds, with the faster one on the left and the slower one on the right, then after some time they collide and their shapes are distorted. After a long enough time, they are separated and recover their original shapes, the only difference being in the change of the phase shift d in [2].

Solitary waves in shallow water (like a canal) were first observed by Scott Russell in Scotland in the middle of the nineteenth century. Differential equations which possess solitary waves in shallow water as solutions were sought after Scott Russell's report. Boussinesq derived one (now called the Boussinesq equation, which contains second partial derivatives with respect to time) from the Euler equation of water wave; then in 1895 Korteweg and his student de Vries derived the KdV equation. They also showed that the KdV equation possesses solutions expressible in terms of elliptic functions.

In the 1960s Kruskal and Zabusky carried out numerical computations for the Fermi–Pasta–Ulam problem; they also came across the KdV equation and found the aforementioned phenomenon.

Inverse-Scattering Method

Kruskal and his co-workers further pursued the origin of the particle-like property of solitons and proposed the so-called inverse-scattering method.

The inverse problem of scattering theory of the one-dimensional Schrödinger operator

$$L = -\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^2 + u(x)$$

was studied by Gelfand–Dikii, Marchenko, and Krein in the 1950s, motivated by scattering theory in quantum mechanics.

It gives a one-to-one correspondence between rapidly decreasing potentials $u(x)$ and scattering data which consist of discrete eigenvalues $-\eta_j^2$ and normalization $c_j, j = 1, \ldots, n$, of the eigenfunctions corresponding to them and the reflection coefficient $r(\xi)$. The reflection coefficient represents the ratio of reflection of the unit plane wave $\mathrm{e}^{\mathrm{i}\xi x}$ by the potential field. The scattering data $\{r(\xi), \eta_j, c_i, j = 1, \ldots, n\}$ are a mathematical idealization of observable data in quantum scattering. The procedure of reconstructing a potential from given scattering data is called the inverse problem. The heart of this procedure is solving an integral equation (the Gelfand–Dikii–Marchenko equation). In the reflectionless case ($r(\xi) = 0$), this integral equation reduces to a system of linear algebraic equations.

Kruskal and co-workers found that the scattering data of these operators with solutions of [1] as potentials depend very simply on t:

$$\begin{aligned} &\eta_j(t) = \eta_j(0), \qquad c_j(t) = c_j(0)\,\mathrm{e}^{2\mathrm{i}\eta^3 t} \\ &r(\xi,t) = r(\xi,0)\,\mathrm{e}^{2\mathrm{i}\xi^3 t} \end{aligned} \qquad [3]$$

It was realized at the same time that soliton solutions correspond to a reflectionless potential ($r(\xi) = 0$) with only one discrete eigenvalue, while reflectionless potentials correspond to a nonlinear "superposition" of soliton solutions (called multisoliton solutions) and describe the interaction of solitons.

As was pointed out by Zakharov and others, the inverse-scattering method has an intimate relation with the Riemann–Hilbert problem.

Lax Representation

Looking at this invariance of the spectrum, Lax reformulated the KdV equation [1] as an evolution equation for the one-dimensional Schrödinger operator:

$$\begin{aligned} &\frac{\mathrm{d}L}{\mathrm{d}t} = [A, L], \quad L = \left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^2 + u \\ &A = \left(\frac{\partial}{\partial x}\right)^3 + \frac{3}{4}\left(u\frac{\partial}{\partial x} + \frac{\partial}{\partial x}u\right) \end{aligned} \qquad [4]$$

Here we have changed the sign of the operator for later convenience. This form of representation together with the inverse-scattering method gave a framework for finding nonlinear differential (difference) equations that have solutions with properties similar to solitons (soliton equations).

Among such are the sine-Gordon equation

$$u_{tt} - u_{xx} = \sin u$$

the nonlinear Schrödinger equation

$$iu_t + u_{xx} + |u|^2 u = 0$$

the modified KdV equation

$$u_t - \tfrac{1}{6}\left(6u^2 u_x + u_{xxx}\right) = 0$$

the Toda lattice equation

$$\begin{aligned} \frac{dQ_n}{dt} &= P_n \\ \frac{dP_n}{dt} &= -\exp(Q_n - Q_{n+1}) + \exp(Q_{n-1} - Q_n) \end{aligned} \qquad [5]$$

and so on. The first three are obtained by replacing L by a 2×2 matrix differential operator of first order. For eqn [5], the linear operator corresponding to L in the case of the KdV equation is a difference operator of order 2 and has a connection with the theory of orthogonal polynomials in one variable as well as with the theory of moment problems.

Later it was remarked that the differential operator A in eqn [4] is nothing but the differential operator part of the fractional power of L: $A = (L^{3/2})_+$. By replacing A in [4] by $(L^{(2n+1)/2})_+$ we obtain higher (nth) KdV equations.

Basic Representations of Affine Lie Algebras

In the 1960s Kac and Moody introduced independently a class of infinite-dimensional Lie algebras which are in many respects close to finite-dimensional semisimple Lie algebras. Each of them is constructed for a given generalized Cartan matrix (GCM),

$$\begin{aligned} C = (a_{ij}), \quad & a_{ii} = 2, \ a_{ij} \le 0 \text{ for } i \ne j \\ & \text{and if } a_{ij} = 0 \text{ then } a_{ji} = 0 \end{aligned} \qquad [6]$$

There is a special class of Kac–Moody Lie algebras that are now called affine Lie algebras. They correspond to positive-semidefinite GCM and are realized as central extensions of loop algebras (current algebras)

$$\mathbf{C}[\lambda, \lambda^{-1}] \otimes \mathfrak{g}$$

of finite-dimensional semisimple Lie algebras $\mathfrak{g}$. They have many applications in physics, in particular as current algebras. The Sugawara construction in current algebra plays an essential role in conformal field theory. Note that finite-dimensional semisimple Lie algebras correspond to positive-definite GCMs.

In the late 1970s, there was interest in constructing representations of these algebras after the general theory of representations was constructed. Among them was the work of Lepowsky–Wilson, who constructed basic representations of the affine Lie algebra $A_1^{(1)} (= \widehat{\mathfrak{sl}}_2)$ using differential operators of infinite order in infinitely many variables. These operators were called vertex operators by Garland, in view of the resemblance to objects in string theory. Character formulas for these new Lie algebras were intensively studied and many combinatorial identities were (re)derived.

Geometric Interpretation

How do Kac–Moody Lie algebras enter into this picture?

In the early stages of the history of solitons Kac–Moody Lie algebras appeared rather artificially. Some authors tried to understand solitons from geometric viewpoints. A typical example is the sine-Gordon equation. This equation appears as the Gauß–Codazzi equation in the theory of embeddings of two-dimensional surfaces of constant negative curvature into three-dimensional Euclidean space, while the Gauß–Weingarten equation is the linear equation that appears in the Lax representation of the sine-Gordon equation. Another approach of a geometric nature, involving the prolongation structure, was the direction initiated by Wahlquist–Estabrook. In this approach, the Lie algebra appeared in a natural way, although the nature of such Lie algebras was not so clear. This direction of research is close in spirit to the method of Cartan for treating partial differential equations.

Several authors considered generalizations of the Toda lattice equation. Bogoyavlenskii and others observed that the original Toda lattice equation [5] is related to the Cartan matrix of the affine Lie algebra of type A. Viewed in this way, it was straightforward to generalize the Toda lattice equation to Cartan matrices of another type of affine Lie algebras and also to ordinary Cartan matrices. These were typical appearances of Kac–Moody Lie algebras in the theory of solitons; they were used to produce soliton equations. The climax of this is the work of Drinfel'd–Sokolov.

It needed some time to understand another role of affine Lie algebras in the theory of solitons.

Bäcklund Transformation

In the theory of two-dimensional surfaces of constant negative curvature, a method of obtaining another surface of constant negative curvature from the given one with some parameter was known by the work of Bäcklund. If we apply this to the trivial solutions $u = 0$ of the sine-Gordon equation, we

obtain a one-soliton solution of the sine-Gordon equation. From this fact, the transformation of solutions of soliton equations to other solutions is called a Bäcklund transformation. The original Darboux transformation is a special case of a Bäcklund transformation.

Hamiltonian Formalism

Another discovery of Gardner–Greene–Kruskal–Miura was the Hamiltonian structure of the KdV equation. In the process of showing the existence of infinitely many conservation laws, they used the so-called Miura transformation, which relates the KdV and the modified KdV equation. Faddeev–Zakharov showed that the transformation to scattering data is a canonical transformation, and conserved quantities are obtained from the expansion of the reflection coefficients.

Gelfand–Dikii studied Hamiltonian structures of the KdV equation using the formal variational calculus they initiated.

M Adler was the first to try to study the KdV equation by using the orbit method known for finite-dimensional Lie algebras. It was known by the works of Kostant and Kirillov or even earlier by Lie that the co-adjoint orbits of Lie algebras admit symplectic structures (the Kostant–Kirillov bracket). Adler considered the algebra of pseudodifferential operators in one variable. This acquires the structure of Lie algebra by the commutation relation. This algebra admits a natural triangular decomposition by order. He showed that the KdV equation can be viewed as a Hamiltonian system in the co-adjoint orbit of the one-dimensional Schrödinger operator with the Kostant–Kirillov bracket. By introducing the notion of residue of pseudodifferential operators he rederived conserved quantities. The work of Drinfeld–Sokolov can be regarded as a thorough generalization of this direction. Hamiltonian structures of the KdV equation and other soliton equations are now understood in this way.

The method is also applicable to finite-dimensional Lie algebras. Symes, Kostant, and others treated the finite Toda lattice in this way.

The motion of tops, including that of Kovalevskaya, was also studied in this way.

Hirota's Method

There was another approach to soliton equations, quite different from the above. This was the method initiated by Hirota. He placed stress on the form of multisoliton solutions of the KdV equation, the sine-Gordon equation, and so on. He made a dependent-variable transformation of the KdV equation [1],

$$u = 2\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)\log f$$

This form naturally arises when we reconstruct the potential of the one-dimensional Schrödinger operator from the scattering data by solving the Gelfand–Dikii–Marchenko integral equation. In this new dependent variable, eqn [1] takes the following form:

$$\left(D_x^4 - 4D_x D_t\right) f(x,t)\cdot f(x,t) = 0$$

where the operator D_x is defined by

$$D_x(f\cdot g) = \frac{\mathrm{d}}{\mathrm{d}x'}\, f(x+x')\, g(x-x')|_{\,x'=0} \qquad [7]$$

This operator is called Hirota's bilinear differential operator. In such transformed form, he tried to solve the resulting equation in a perturbative way,

$$\begin{aligned} f = 1 &+ \sum_{j=1}^{n} \exp(2p_j x + 2p^3 t + q_j) \\ &+ \sum_{1\le j<k\le n} c_{ij}\exp(2(p_j+p_k)x \\ &+ 2(p_j^3+p_k^3)t + q_j + q_k) + \cdots \end{aligned} \qquad [8]$$

It is rather miraculous that in the soliton equation case we can truncate such a perturbative procedure at a finite point. The number of steps corresponds to the number of solitons.

Most of the soliton equations are rewritten in bilinear form with such bilinear differentiation after a suitable dependent-variable transformation. (Some equations need several new dependent variables.) Once we have a differential equation in Hirota's bilinear differential form, it always has two-soliton solutions.

Up to 1980, keywords characterizing solitons were; inverse-scattering method, Bäcklund transformation, multisolitons, Hirota's method, quasi-periodic solutions, etc. No explicit mention was made of representation theory.

Hierarchy of Soliton Equations

As was stated above, soliton equations viewed as Hamiltonian systems have infinitely many conservation laws. This implies that we can introduce infinitely many independent time variables consistently. From this viewpoint, it is natural to consider the KdV equation and its higher-order analogs simultaneously. They have many properties in common. For example, the t-dependence of the scattering data of the higher

KdV equation is given by replacing ξ^3 by ξ^{2n+1} and η_j^3 by η_j^{2n+1} in eqn [3]. The totality of soliton equations organized in this way is called a hierarchy of soliton equations; in the KdV case, it is called the KdV hierarchy. This notion of hierarchy was introduced by M Sato. He tried to understand the nature of the bilinear method of Hirota. First, he counted the number of Hirota bilinear operators of given degree for hierarchies of soliton equations. For the number of bilinear equations, M Sato and Y Sato made extensive computations and made many conjectures that involve eumeration of partitions.

Kadomtsev–Petviashvili Hierarchy

Although it was included in a family of soliton equations slightly later, the Kadomtsev–Petviashvili (KP) equation is a soliton equation in three independent variables, which first appeared in plasma physics:

$$\tfrac{3}{4}u_{yy} - \left(u_t - \tfrac{1}{4}(6uu_x + u_{xxx})\right) = 0 \qquad [9]$$

For this equation we have to replace the Lax representation by

$$\left[\left(\frac{\partial}{\partial x}\right)^2 + u - \frac{\partial}{\partial y}, \left(\frac{\partial}{\partial x}\right)^3 + \frac{3}{2}u\frac{\partial}{\partial x} + v - \frac{\partial}{\partial t}\right] = 0 \qquad [10]$$

This form of representation was introduced by Zakharov–Shabat. Sometimes it is referred to as the zero-curvature representation or the Zakharov–Shabat representation. The KP equation is universal in the sense that it contains the KdV equation [1] and the Boussinesq equation as special cases. If u does not depend on y, resp. t, this gives the KdV, resp. the Boussinesq equation.

Work of Sato

Sato stressed the importance of the study of the KP equation. He first introduced the KP hierarchy. Instead of the one-dimensional Schrödinger operator in the KdV case consider a pseudo- (micro) differential operator of first order,

$$\begin{aligned} L &= \partial + u_2(x)\partial^{-1} + u_3(x)\partial^{-3} + \cdots \\ \partial &= \frac{\partial}{\partial x_1}, \quad x = (x_1, x_2, x_3, \ldots) \end{aligned} \qquad [11]$$

Setting $B_n = (L^n)_+$, the KP hierarchy is defined by the Zakharov–Shabat representation

$$\left[\frac{\partial}{\partial x_m} - B_m, \frac{\partial}{\partial x_n} - B_n\right] = 0, \quad m, n = 2, 3, \ldots$$

If we assume that L^2 is a differential operator, we have the KdV hierarchy and the constraint that L^3 is a differential operator gives the Boussinesq hierarchy. This process is called reduction.

Sato found that character polynomials (Schur functions) solve the KP hierarchy and, based on this observation, he created the theory of the infinite-dimensional (universal) Grassmann manifold and showed that the Hirota bilinear equations are nothing but the Plücker relations for this Grassmann manifold.

Sato also gave an (infinite-dimensional) determinantal formula for Hirota's dependent variable and called the latter the τ-function. Using this τ-function, the wave function (the eigenfunction corresponding to the KP hierarchy) is expressed as

$$\begin{aligned} w(x,k) &= \exp\left(\sum_{n=1}^{\infty} x_n k^n\right) \frac{\tau(x - \epsilon(k^{-1}))}{\tau(x)} \\ \epsilon(k) &= \left(k, \frac{k^2}{2}, \frac{k^3}{3}, \ldots\right) \\ Lw &= kw \end{aligned} \qquad [12]$$

where L is given by eqn [11].

Affine Lie Algebras as Infinitesimal Transformation Groups for Soliton Equations

Date–Jimbo–Kashiwara–Miwa found another relation among soliton equations and affine Lie algebras. After noticing some similarity between the formula in the paper by Lepowsky–Wilson on the Rogers–Ramanujan identity using the vertex operators for $A_1^{(1)}$ and the formula in the computation of numbers of bilinear operators in Sato's paper, they applied the vertex operator for $A_1^{(1)}$,

$$\begin{aligned} X(p) = &\exp\left(\sum_{j=1}^{\infty} 2x_{2j-1}p^{2j-1}\right) \\ &\times \exp\left(-\sum_{j=1} \frac{2}{jp^{2j-1}} \frac{\partial}{\partial x_{2j-1}}\right) \end{aligned}$$

to 1 (which is the simplest τ-function for the KdV hierarchy), where p is a parameter. They found that the result is the τ-function corresponding to the one-soliton solution of the KP hierarchy. They also found that successive application of $X(p)$'s to 1 produced all multisoliton τ-functions. Therefore, applications of vertex operators are precisely

Bäcklund transformations. This implies that the affine Lie algebra $A_1^{(1)}$ is the infinitesimal transformation group for solutions of the KdV hierarchy.

After this discovery, it was realized that the totality of τ-functions of the KdV hierarchy is the group orbit of the highest weight vector $(=1)$ of the basic representation of $A_1^{(1)}$.

The vertex operators for the KP hierarchy were also found:

$$X(p,q) = \exp\left(\sum_{j=1}^{\infty} x_j(p^j + q^j)\right) \times \exp\left(-\sum_{j=1}\left(\frac{1}{jp^j} + \frac{1}{jq^j}\right)\frac{\partial}{\partial x_j}\right)$$

If we put $q = -p$, the vertex operator for $A_1^{(1)}$ ([12]) is recovered.

Viewed in this way the Lie algebra corresponding to the KP hierarchy is $\mathfrak{gl}_\infty (= A_\infty)$. And an embedding of $A_1^{(1)}$ into A_∞ was also found. Subsequently, the method using free fermions (Clifford algebras) was established. Frenkel–Kac had already used free fermions to construct basic representations. In this approach, the τ-functions are defined as vacuum expectation values. Based on this connection with affine Lie algebras, many conjectures of Sato on the number of bilinear equations are (re)proved by using specialized characters of affine Lie algebras.

The use of free fermions was exploited by Ishibashi–Matsuo–Ooguri to relate soliton equations with conformal field theory on Riemann surfaces. This aspect was further studied by Tsuchiya–Ueno–Yamada using D-modules.

Once such a viewpoint was established, it was easy to construct soliton equations corresponding to other affine Lie algebras. Hierarchies similar to the KP hierarchies (the simplest equation contains three variables) were also found, which correspond to Lie algebras like $\mathfrak{go}_\infty, \mathfrak{sp}_\infty$ (the BKP hierarchy, the CKP hierarchy, and so on).

Summarizing these developments, we can say that affine Lie algebras, or slightly larger ones like $\mathfrak{gl}_\infty$, appear naturally as infinitesimal transformation groups for soliton equations and the solution spaces are the (completed) group orbits of highest weight vector τ-functions of level-1 representations. The Hirota bilinear equations are the equations describing these orbits (analogs of Plücker relations).

Soon afterwards, the notion of τ-functions was introduced in the study of Painlevé equations by Okamoto, revealing Hamiltonian structures in Painlevé equations.

The Method of Drinfeld–Sokolov

The KdV or the KP hierarchies are related to scalar linear differential operators. A parallel treatment using matrix differential operators is also possible. In fact, the nonlinear Schrödinger equation, modified KdV equation, the sine-Gordon equation, etc., are treated in this way.

Drinfel'd and Sokolov gave a general framework along these lines. The first step is to choose the starting (matrix-valued) linear differential operator of order one. For that they use the language of Lie algebras.

Let us start with a matrix realization of a Lie algebra (for an affine Lie algebra, the elements are Laurent polynomials in one variable). Consider a linear differential operator of the following form:

$$L = \frac{\mathrm{d}}{\mathrm{d}x} + q(x) + \Lambda$$

where $q(x)$ is an element of the Borel subalgebra and Λ is a sum of positive Chevalley generators in the case of affine Lie algebras. By using gauge transformations (adjoint group), they consider several normal forms. One normal form is obtained by choosing a node of the corresponding Dynkin diagram. The resulting matrix system is equivalent to the one obtained by scalar Lax representation (or a slight generalization of it). In this way, the generalized KdV equations for affine Lie algebras are obtained. Another normal form is to make q $\mathfrak{h}$-valued. Soliton equations obtained in this way are called the modified KdV equations. This is a generalization of the Miura transformation. They also comment on the construction of partially modified soliton equations, which correspond to taking various parabolic subalgebras. The Hamiltonian formalism is also treated from their viewpoint.

In summary, in their approach affine algebras are used to construct soliton equations, or one can say that they consider the space of initial values of soliton equations.

They also discuss two-dimensional Toda lattices in their setting and show that modified equations in their sense are symmetries of the two-dimensional Toda lattices.

Common Features of the Roles of Affine Lie Algebras in Solitons

In τ-function approach as well as in the method of Drinfeld–Sokolov, the existence of triangular decomposition of Lie algebras was essential. In the former case, it was basic when considering highest-weight representations and, for the latter, it was used for the setup.

Special Solutions of Soliton Equations (Multisoliton and Rational Solutions)

One of the characteristic features of soliton equations is that they allow rich special solutions. Multisoliton solutions were the starting point of the whole story. They directly relate to vertex operators of affine Lie algebras.

Rational solutions (in terms of τ-function polynomial solutions) can be viewed as degenerations of multisoliton solutions. Motions of poles (or zeros) of the solutions are interesting. Airault–McKean–Moser studied the motion of poles of rational solutions of the KdV equation and found that they are identical to the motion of particles on a line (Calogero–Moser–Sutherland system). This viewpoint has now been generalized by Veselov and others.

Another discovery of Sato was that polynomial τ-functions of the KP hierarchy are precisely Schur functions (character polynomials).

In accordance with the process of reduction, polynomial τ-functions of the KdV hierarchy are Schur functions of special type.

Quasiperiodic Solutions of Soliton Equations

As mentioned above, the KdV equation admits solutions expressible in terms of elliptic functions. Dubrovin–Novikov and Its–Matveev, almost at the same time, studied solutions of the KdV equation with periodic initial condition.

To the Sturm–Liouville (i.e., one-dimensional Schrödinger) operator with periodic potential

$$L = \left(\frac{\partial}{\partial x}\right)^2 + u(x), \quad u(x+l) = u(x)$$

there corresponds the discriminant, which is an entire function of the spectral parameter. Its zeros represent the periodic and antiperiodic spectrum λ_j of the operator:

$$Lf_j(x) = \lambda_j f_j(x), \quad f_j(x+l) = \pm f_j(x)$$

It turns out that, except for a finite number of zeros, other zeros are double. Such a potential is called a finite-zone potential. These zones correspond to the spectrum of the operator in the L^2-sense. To a finite-zone potential $u(x)$ there corresponds a hyperelliptic curve

$$\mu^2 = \prod_{j=0}^{2n} (\lambda - \lambda_j)$$

with simple zeros λ_j of the discriminant as zeros of polynomials defining the curve. If we consider the Dirichlet boundary value problem for the operator L,

$$Lf = \mu, f$$
$$f(s, \mu) = 0 = f(s+l, \mu)$$

the eigenvalues are discrete and each eigenvalue μ_j is located in a zone:

$$\lambda_{2j-1} \leq \mu_j(s) \leq \lambda_{2j}$$

So, for the double zeros ($\lambda_{2j-1} = \lambda_{2j}$), the corresponding Dirichlet eigenvalue $\mu_j(s)$ does not depend on s.

Dubrovin–Novikov also showed that a finite-zone potential is a stationary solution of the higher-order KdV equation (the order being equal to the number of nontrivial zones) and the n-zonal potentials form a finite-dimensional integrable system. In other words, the linear operators L, A_n defining the nth order KdV equations commute,

$$[L, A_n] = 0$$

In passing, it was later found that such a pair of commuting linear differential operators was first studied by Burchnall–Chaundy in the 1920s. H F Baker remarked on the corresponding simultaneous eigenfunctions by relating them to multiplicative functions on algebraic curves.

The Work of Krichever

Krichever reversed the above argument, utilizing the properties of corresponding eigenfunctions as a function of the spectral parameter. In this approach, we start with a compact Riemann surface C (= nonsingular algebraic curve) of genus g. Here we apply his method to the KP hierarchy. Take a point P_0 on C together with the inverse of a local parameter k^{-1}. Also take a general divisor δ on C of degree g. Consider a function $\psi(x, P), x = (x_1, x_2, \ldots)$, with the following properties:

1. ψ is meromorphic on $C \backslash P_0$ with the pole divisor δ, and
2. near P_0, ψ behaves like

$$\psi(x, P) = \exp\left(\sum_{j=1}^{\infty} x_j k^j\right)\left(1 + \mathcal{O}(k^{-1})\right)$$

Such a ψ exists uniquely and can be constructed using the theory of abelian integrals and the Jacobi problems on algebraic curves. Such a function was called the Baker–Akhiezer function, since Akhiezer constructed it by using abelian integrals and Jacobi's

problem in his study of moment problems (orthogonal polynomials).

It was later realized that Schur had much earlier considered such functions in the study of ordinary differential equations.

It is easy to show that such a function satisfies the following linear differential equations:

$$\frac{\partial}{\partial x_n}\psi=\left(\left(\frac{\partial}{\partial x_1}\right)^n+\sum_{j=0}^{n-1}u_j(x)\left(\frac{\partial}{\partial x_1}\right)^j\right)\psi,\quad n=2,3,\ldots$$

In this way, we obtain a solution of the KP hierarchy.

If there exists a rational function $f(P)$ on C with poles only at P_0 with singular part k^n, ψ can be factorized as

$$\psi(x,P)=\exp f(P)\psi'(x',P)$$

where x' indicates the set of variables other than x_n. Consequently, we have

$$\frac{\partial}{\partial x_n}\psi(x,P)=f(P)\psi(x,P)$$

In this way, for a hyperelliptic curve C and a branch point of it, viewed as the double cover of CP^1, we recover the case of the KdV hierarchy.

Multisolitons correspond to rational algebraic curves with ordinary double points, while rational solutions correspond to further degeneration.

The study of quasiperiodic solutions of soliton equations revealed an intimate relationship with the theory of algebraic curves. One particular outcome was the characterization of Jacobian varieties among abelian varieties. This was originally posed by Schottky and subsequently reformulated by S P Novikov using soliton equations (Schottky problem, Novikov conjecture). This problem was solved through studies by Shiota, Mulase, and Arbarello–De Concini.

Another aspect was finding commutative subalgebras in the ring of linear differential operators. This problem is related to the theory of stable vector bundles on algebraic curves.

Similarity Solutions of Soliton Equations

Ablowitz and Segur have shown that the Painlevé transcendent of the second kind solves the KdV equation as a similarity solution. This was the starting point of the study of similarity solutions of soliton equations.

Flaschka and Newell tried to construct the theory of multisimilarity solutions. As a by-product, they discussed modulation of the KdV equation by using the averaging method of Whitham. This opens the way to study the quasiclassical limit of soliton equations. This aspect was further studied by Dubrovin and others in connection with topological field theory.

Quite recently, Noumi and Yamada gave a generalization of the Painlevé equation in many variables by using the idea of similarity solutions of soliton equations. In the work of Noumi–Yamada, the affine Weyl group and τ-functions play an essential role in constructing generalizations of the Painlevé equation. The shift or the unit of difference corresponds to imaginary null roots of affine Lie algebras. The idea is further applied to elliptic Painlevé equations.

Integrable Many-Body Problems

As mentioned in relation with the rational solutions of soliton equations, the theory of integrable many-body problems has an intimate relationship with the theory of solitons. Recently, Veselov and his co-workers introduced the notion of Baker–Akhiezer functions of many variables. This concerns a commutative subring of differential operators in many variables. The structure of vector bundles on algebraic varieties of higher dimensions is quite different from that of algebraic curves. For this reason, a naïve generalization of soliton equations to higher dimensions is not possible. Veselov and others have set up a class of functions which they call multidimensional Baker–Akhiezer functions. They are defined by giving a finite set of vectors in a Euclidean space. The first problem is the existence. For the existence of the multidimensional Baker–Akhiezer function the set must satisfy several constraints. This is quite different from the case of solitons. Root systems satisfy these constraints and the corresponding Baker–Akhiezer function becomes the common eigenfunction of linear differential operators appearing in the Calogero–Sutherland–Moser model corresponding to root systems.

Ball–Box Systems

Satsuma–Takahashi found a soliton-like phenomenon in cellular automata. It took much time for a mathematical explanation of this. Now it is understand that these systems are obtained by a limiting procedure from soliton equations. Sometimes this is called ultra-discretization. The system thus obtained can also be obtained from the theory of crystal bases of affine Lie algebras. They are now called ball–box systems.

Other Topics

A quantized version of the inverse-scattering method was initiated by Faddeev and his co-workers, which makes a connection with two-dimensional solvable lattice models and produced the notion of quantum groups. Through the Bethe ansatz, another relation of two-dimensional lattice models and ball–box systems has been discussed.

See also: Affine Quantum Groups; Bäcklund Transformations; Bi-Hamiltonian Methods in Soliton Theory; Coherent States; Current Algebra; Integrable Systems and Algebraic Geometry; Integrable Systems: Overview; Multi-Hamiltonian Systems; Painlevé Equations; Partial Differential Equations: Some Examples; q-Special Functions; Recursion Operators in Classical Mechanics; Sine-Gordon Equation; Toda Lattices.

Further Reading

Cherednik I (1996) *Basic Methods of Soliton Theory*. Advanced Series in Mathematical Physics, vol. 25. Singapore, New Jersey, London and Hong Kong: World Scientific.

Date E, Kashiwara M, Jimbo M, and Miwa T (1983) Transformation groups for soliton equations. In: Jimbo M and Miwa T (eds.) *Proceedings of RIMS Symposium on Non-Linear Integrable Systems – Classical Theory and Quantum theory*, pp. 39–119. Singapore: World Scientific.

Drinfel'd VG and Sokolov VV (1985) Lie algebras and equations of Korteweg–de Vries type. *Journal of Soviet Mathematics* 30: 1975–2036.

Gardner CS, Greene JM, Kruskal MD, and Miura RM (1967) Methods for solving the Korteweg–de Vries equation. *Physical Review Letter* 19: 1095–1097.

Kac VG (1990) *Infinite Dimensional Lie Algebras*, 3rd edition. Cambridge: Cambridge University Press.

Manin YuI (1979) Algebraic aspects of nonlinear differential equations. *Journal of Soviet Mathematics* 11: 1–122.

Miwa T, Jimbo M, and Date E (2000) *Solitons*, (Translated by Reid, M). Cambridge: Cambridge University Press.

Noumi M (2002) Affine Weyl group approach to Painlevé equations. In: Tatsien LI (ed.) *International Congress of Mathematicians (2002, Beijing)*, Proceedings of the International Congress of Mathematicians, August 20–28, 2002, Beijing, pp. 497–509. Beijing: Higher Education Press.

Novikov SP, Manakov SV, Pitaevskii LP, and Zakharov VE (1984) *Theory of Solitons. The Inverse Scattering Method*. New York and London: Consultants Bureau.

Sato M and Sato Y (1983) Soliton equations as dynamical systems on infinite-dimensional Grassmann manifold. In: Fujita H, Lax PD, and Strang G (eds.) *Nonlinear Partial Differential Equations in Applied Science (Tokyo, 1982)*, North-Holland Math. Stud., vol. 81, pp. 259–271. Amsterdam: North-Holland.

Solitons and Other Extended Field Configurations

R S Ward, University of Durham, Durham, UK

Introduction

A soliton is a localized lump (or string or wall, etc.) of energy, which can move without distortion, dispersion, or dissipation, and which is stable under perturbations (and collisions with other solitons). The word was coined by Zabusky and Kruskal in 1965 to describe a solitary wave with particle-like properties (as in electron, proton, etc.). Solitons are relevant to numerous areas of physics – condensed matter, cosmology, fluids/plasmas, biophysics (e.g., DNA), nuclear physics, high-energy physics, etc. Mathematically, they are modeled as solutions of appropriate partial differential equations.

Systems which admit solitons may be classified according to the mechanism by which stability is ensured. Such mechanisms include complete integrability, nontrivial topology plus dynamical balancing, and Q-balls/breathers.

Sometimes the term "soliton" is used in a restricted sense, to refer to stable localized lumps which have purely elastic interactions: solitons which collide without any radiation being emitted. This is possible only in very special systems, namely, those that are completely integrable. For these systems, soliton stability (and the elasticity of collisions) arises from a number of characteristic properties, including a precise balance between dispersion and nonlinearity, solvability by the inverse scattering transform from linear data, infinitely many conserved quantities, a Lax formulation (associated linear problem), and Bäcklund transformations. Examples of such integrable soliton systems are the sine-Gordon, Korteweg–deVries, and nonlinear Schrödinger equations.

The category of topological solitons is the most varied, and includes such examples as kinks, vortices, monopoles, skyrmions, and instantons. The requirement of dynamical balancing for these can be understood in terms of Derrick's theorem, which provides necessary conditions for a classical field theory to admit static localized solutions. The

Derrick argument involves studying what happens to the energy of a field when one changes the scale of space. If one has a scalar field (or multiplet of scalar fields) ϕ, and/or a gauge field $F_{\mu\nu}$, then the static energy E is the sum of terms such as

$$E_0 = \int V(\phi)\,\mathrm{d}^n x, \qquad E_d = \int T_d(D_j\phi)\,\mathrm{d}^n x,$$
$$E_F = \int F_{jk}F_{jk}\,\mathrm{d}^n x$$

where each integral is over (n-dimensional) space $\boldsymbol{R}^n$, $D_j\phi$ denotes the covariant spatial derivative of ϕ, and $T_d(\xi_j)$ is a real-valued polynomial of degree d. In particular, for example, we could have $T_2(D_j\phi) = (D_j\phi)(D_j\phi)$, the standard gradient term. Under the dilation $x^j \mapsto \lambda x^j$, these functionals transform as

$$E_0 \mapsto \lambda^{-n}E_0, \qquad E_d \mapsto \lambda^{d-n}E_d, \qquad E_F \mapsto \lambda^{4-n}E_F$$

In order to have a static solution (critical point of the static energy functional), one needs to have a zero exponent on λ, and/or a balance between positive and negative exponents. A negative exponent indicates a compressing force (tending to implode a localized lump), whereas a positive exponent indicates an expanding force; so to have a static lump solution, these two forces have to balance each other. For $n=1$, a system involving only a scalar field, with terms of the form E_0 and E_2, can admit static solitons (e.g., kinks); the scaling argument implies a virial theorem, which in this case says that $E_0 = E_2$. For $n=2$, one can have a scalar system with only E_2, since in this case the relevant exponent is zero (e.g., the two-dimensional sigma model). Another $n=2$ example is that of vortices in the abelian Higgs model, where the energy contains terms E_0, E_2, and E_F. For $n=3$, interesting systems have E_2 together with either E_4 (e.g., skyrmions) or E_F (e.g., monopoles). An E_0 term is optional in these cases; its presence affects, in particular, the long-range properties of the solitons. For $n=4$, one can have instantons in a pure gauge theory (term E_F only).

It should be noted that if there are no restrictions on the fields ϕ and A_j (such as those arising, e.g., from nontrivial topology), then there is a more obvious mode of instability, which will inevitably be present: $\phi \mapsto \mu\phi$ and/or $A_j \mapsto \mu A_j$, where $0 \le \mu \le 1$. In other words, the fields can simply be scaled away altogether, so that the height of the soliton (and its energy) go smoothly to zero. This can be prevented by nontrivial topology.

Another way of preventing solitons from shrinking is to allow the field to have some "internal" time dependence, so that it is stationary rather than static. For example, one could allow the complex scalar field ϕ to have the form $\phi = \psi \exp(\mathrm{i}\omega t)$, where ψ is independent of time t. This leads to something like a centrifugal force, which can have a stabilizing effect in the absence of Skyrme or magnetic terms. The corresponding solitons are Q-balls.

Kinks and Breathers

The simplest topological solitons are kinks, in systems involving a real-valued scalar field $\phi(x)$ in one spatial dimension. The dynamics is governed by the Lagrangian density

$$\mathcal{L} = \tfrac{1}{2}\left[(\phi_t)^2 - (\phi_x)^2 - W(\phi)^2\right]$$

where $W(\phi)$ is a (fixed) smooth function. The system can admit kinks if $W(\phi)$ has at least two zeros, for example, $W(A) = W(B) = 0$ with $W(\phi) > 0$ for $A < \phi < B$. Two well-known systems are: sine-Gordon (where $W(\phi) = 2\sin(\phi/2)$, $A = 0$, and $B = 2\pi$) and ϕ^4 (where $W(\phi) = 1 - \phi^2$, $A = -1$, and $B = 1$). The corresponding field equations are the Euler–Lagrange equations for $\mathcal{L}$; for example, the sine-Gordon equation is

$$\phi_{tt} - \phi_{xx} + \sin\,\phi = 0 \qquad [1]$$

Configurations satisfying the boundary conditions $\phi \to A$ as $x \to -\infty$ and $\phi \to B$ as $x \to \infty$ are called kinks (and the corresponding ones with $x = \infty$ and $x = -\infty$ interchanged are antikinks). For kink (or antikink) configurations, there is a lower bound, called the Bogomol'nyi bound, on the static energy $E[\phi]$; for kink boundary conditions, we have

$$\begin{aligned} E[\phi] &= \frac{1}{2}\int_{-\infty}^{\infty}\left[(\phi_x)^2 + W(\phi)^2\right]\mathrm{d}x \\ &= \frac{1}{2}\int_{-\infty}^{\infty}[\phi_x - W(\phi)]^2\,\mathrm{d}x + \int_{-\infty}^{\infty} W(\phi)\phi_x\,\mathrm{d}x \\ &\ge \int_A^B W(\phi)\,\mathrm{d}\phi \end{aligned}$$

with equality if and only if the Bogomol'nyi equation

$$\frac{\mathrm{d}\phi}{\mathrm{d}x} = W(\phi) \qquad [2]$$

is satisfied. A static solution of the Bogomol'nyi equation is a kink solution – it is a static minimum of the energy functional in the kink sector. For example, for the sine-Gordon system, we get $E[\phi] \ge 8$, with equality for the sine-Gordon kink

$$\phi(x) = 4\,\tan^{-1}\exp(x - x_0)$$

while for the ϕ^4 system, we get $E[\phi] \ge 4/3$, with equality for the phi-four kink

$$\phi(x) = \tanh(x - x_0)$$

These kinks are stable topological solitons; the nontrivial topology corresponds to the fact that the boundary value of $\phi(t,x)$ at $x=\infty$ is different from that at $x=-\infty$. With trivial boundary conditions (say $\phi \to A$ as $x \to \pm\infty$), stable static solitons are unlikely to exist, but solitons with periodic time dependence (which in this context are called breathers) may exist. For example, the sine-Gordon equation and the nonlinear Schrödinger equation, both, admit breathers – but these owe their existence to complete integrability. By contrast, the ϕ^4 system (which is not integrable) does not admit breathers; a collision between a ϕ^4 kink and an antikink (with suitable impact speed) produces a long-lived state which looks like a breather, but eventually decays into radiation.

In lattice systems, however, breathers are more generic. In a one-dimensional lattice system, the continuous space $\boldsymbol{R}$ is replaced by the lattice $\boldsymbol{Z}$, so $\phi(t,x)$ is replaced by $\phi_n(t)$, where $n \in \boldsymbol{Z}$. The Lagrangian is

$$L = \frac{1}{2}\sum_n \left[(\dot{\phi}_n)^2 - h^{-2}(\phi_{n+1}-\phi_n)^2 - W(\phi_n)\right]$$

where h is a positive parameter, corresponding to the dimensionless ratio between the lattice spacing and the size of a kink. The continuum limit is $h \to 0$. This system admits kink solutions as in the continuum case; and for h large enough, it admits breathers as well, but these disappear as h becomes small.

Interpreted in three dimensions, the kink becomes a domain wall separating two regions in which the order parameter ϕ takes distinct values; this has applications in such diverse areas as cosmology and condensed matter physics.

Sigma Models and Skyrmions

In a sigma model or Skyrme system, the field is a map ϕ from spacetime to a Riemannian manifold M; generally, M is taken to be a Lie group or a symmetric space. The energy density of a static field can be constructed as follows (the Lorentz-invariant extension of this gives a relativistic Lagrangian for fields on spacetime). Let ϕ^a be local coordinates on the m-dimensional manifold M, let h_{ab} denote the metric of M, and let x^j denote the spatial coordinates on space $\boldsymbol{R}^n$. An $m \times m$ matrix D is defined by

$$D_a{}^b = (\partial_j \phi^c) h_{ac} (\partial_j \phi^b)$$

where ∂_j denotes derivatives with respect to the x^j. Then the invariants $\mathcal{E}_2 = \mathrm{tr}(D) = |\partial_j \phi^a|^2$ and $\mathcal{E}_4 = (1/2)[(\mathrm{tr}\, D)^2 - \mathrm{tr}(D^2)]$ can be terms in the energy density, as well as a zeroth-order term $\mathcal{E}_0 = V(\phi^a)$ not involving derivatives of ϕ. A term of the form $\mathcal{E}_4$ is called a Skyrme term.

The boundary condition on field configurations is that ϕ tends to some constant value $\phi_0 \in M$ as $|x| \to \infty$ in $\boldsymbol{R}^n$. From the topological point of view, this compactifies $\boldsymbol{R}^n$ to S^n. In other words, ϕ extends to a map from S^n to M; and such maps are classified topologically by the homotopy group $\pi_n(M)$. For topological solitons to exist, this group has to be nontrivial.

In one spatial dimension ($n=1$) with $M=S^1$ (say), the expression $\mathcal{E}_4$ is identically zero, and we just have kink-type systems such as sine-Gordon. The simplest two-dimensional example ($n=2$) is the $O(3)$ sigma model, which has $M=S^2$ with its standard metric. In this system, the field is often expressed as a unit 3-vector field $\boldsymbol{\phi} = (\phi^1, \phi^2, \phi^3)$, with $\mathcal{E}_2 = (\partial_j \boldsymbol{\phi}) \cdot (\partial_j \boldsymbol{\phi})$. Here the configurations are classified topologically by their degree (or winding number, or topological charge) $N \in \pi_2(S^2) \cong \boldsymbol{Z}$, which equals

$$N = \frac{1}{4\pi}\int \boldsymbol{\phi} \cdot \partial_1 \boldsymbol{\phi} \times \partial_2 \boldsymbol{\phi}\, \mathrm{d}x^1\, \mathrm{d}x^2$$

Instead of $\boldsymbol{\phi}$, it is often convenient to use a single complex-valued function W related to $\boldsymbol{\phi}$ by the stereographic projection $W = (\phi^1 + \mathrm{i}\phi^2)/(1-\phi^3)$. In terms of W, the formula for the degree N is

$$N = \frac{\mathrm{i}}{2\pi}\int \frac{W_1 \overline{W}_2 - W_2 \overline{W}_1}{(1+|W|^2)^2}\, \mathrm{d}x^1\, \mathrm{d}x^2$$

and the static energy is (with $z = x^1 + \mathrm{i}x^2$)

$$\begin{aligned} E &= \int \mathcal{E}_2\, \mathrm{d}^2x \\ &= 8\int \frac{|W_z|^2 + |W_{\overline{z}}|^2}{(1+|W|^2)^2}\, \mathrm{d}^2x \\ &= 16\int \frac{|W_{\overline{z}}|^2}{(1+|W|^2)^2}\, \mathrm{d}^2x + 8\int \frac{|W_z|^2 - |W_{\overline{z}}|^2}{(1+|W|^2)^2}\, \mathrm{d}^2x \\ &= 16\int \frac{|W_{\overline{z}}|^2}{(1+|W|^2)^2}\, \mathrm{d}^2x + 8\pi N \end{aligned}$$

From this, one sees that E satisfies the Bogomol'nyi bound $E \geq 8\pi N$, and that minimal-energy solutions correspond to solutions of the Cauchy–Riemann equations $W_{\overline{z}} = 0$. To have finite energy, $W(z)$ has to be a rational function, and so solutions with winding number N correspond to rational meromorphic functions $W(z)$, of degree $|N|$. (If $N<0$, then W is a rational function of $\overline{z}$.) The energy is scale invariant (conformally invariant), and consequently these solutions are not solitons – they are not quite stable, since their size is not fixed. Adding terms $\mathcal{E}_4$ and $\mathcal{E}_0$

to the energy density fixes the soliton size, and the resulting two-dimensional Skyrme systems admit true topological solitons.

The three-dimensional case ($n=3$), with M being a simple Lie group, is the original Skyrme model of nuclear physics. If $M=\mathrm{SU}(2)$, then the integer $N \in \pi_3(\mathrm{SU}(2)) \cong \boldsymbol{Z}$ is interpreted as the baryon number. The (quantum) excitations of the ϕ-field correspond to the pions, whereas the (semiclassical) solitons correspond to the nucleons. This model emerges as an effective theory of quantum chromodynamics (QCD), in the limit where the number of colors is large. If we express the field as a function $U(x^j)$ taking values in a Lie group, then $L_j = U^{-1}\partial_j U$ takes values in the corresponding Lie algebra, and $\mathcal{E}_2$ and $\mathcal{E}_4$ take the form

$$\mathcal{E}_2 = -\tfrac{1}{2}\mathrm{tr}(L_j L_j)$$
$$\mathcal{E}_4 = -\tfrac{1}{16}\mathrm{tr}\big([L_j, L_k][L_j, L_k]\big)$$

The static energy density in the basic Skyrme system is the sum of these two terms. The static energy satisfies a Bogomol'nyi bound $E \geq 12\pi^2|N|$, and it is believed that stable solitons (skyrmions) exist for each value of N. Classical skyrmions have been investigated numerically; for values of N up to ~ 25, they turn out to resemble polyhedral shells. Comparison with nucleon phenomenology requires semiclassical quantization, and this leads to results which are at least qualitatively correct.

A variant of the Skyrme model is the Skyrme–Faddeev system, which has $n=3$ and $M=S^2$; the solitons in this case resemble loops which can be linked or knotted, and which are classified by their Hopf number $N \in \pi_3(S^2)$. In this case, the energy satisfies a lower bound of the form $E \geq cN^{3/4}$. Numerical experiments indicate that for each N, there is a minimal-energy solution with Hopf number N, and with energy close to this topological lower bound.

Abelian Higgs Vortices

Vortices live in two spatial dimensions; viewed in three dimensions, they are string-like. Two of their applications are as cosmic strings and as magnetic flux tubes in superconductors. They occur as static topological solitons in the the abelian Higgs model (or Ginzburg–Landau model), and involve a magnetic field $B=\partial_1 A_2 - \partial_2 A_1$, coupled to a complex scalar field ϕ, on the plane $\boldsymbol{R}^2$. The energy density is

$$\mathcal{E} = \tfrac{1}{2}(D_j\phi)(\overline{D_j\phi}) + \tfrac{1}{2}B^2 + \tfrac{1}{8}\lambda(1-|\phi|^2)^2 \qquad [3]$$

where $D_j\phi := \partial_j\phi - \mathrm{i}A_j\phi$, and where λ is a positive constant. The boundary conditions are

$$D_j\phi = 0, \qquad B = 0, \qquad |\phi| = 1 \qquad [4]$$

as $r \to \infty$. If we consider a very large circle C on $\boldsymbol{R}^2$, so that [4] holds on C, then $\phi|_C$ is a map from the circle C to the circle of unit radius in the complex plane, and therefore it has an integer winding number N. Thus configurations are labeled by this vortex number N.

Note that if $\mathcal{E}$ vanishes, then $B=0$ and $|\phi|=1$: the gauge symmetry is spontaneously broken, and the photon "acquires a mass": this is a standard example of spontaneous symmetry breaking.

The total magnetic flux $\int B\,\mathrm{d}^2x$ equals $2\pi N$; a proof of this is as follows. Let θ be the usual polar coordinate around C. Because $|\phi|=1$ on C, we can write $\phi = \exp[\mathrm{i}f(\theta)]$ for some function f; this f need not be single-valued, but must satisfy $f(2\pi) - f(0) = 2\pi N$ with N being an integer (in order that ϕ be single-valued). In fact, this defines the winding number. Now since $D_j\phi = \partial_j\phi - \mathrm{i}A_j\phi = 0$ on C, we have

$$A_j = -\mathrm{i}\phi^{-1}\partial_j\phi = \partial_j f$$

on C. So, using Stokes' theorem, we get

$$\begin{aligned}\int_{\boldsymbol{R}^2} B\,\mathrm{d}^2x &= \int_C A_j\,\mathrm{d}x^j \\ &= \int_0^{2\pi} \frac{\mathrm{d}f}{\mathrm{d}\theta}\,\mathrm{d}\theta \\ &= 2\pi N\end{aligned}$$

If $\lambda=1$, then the total energy $E=\int \mathcal{E}\,\mathrm{d}^2x$ satisfies the Bogomol'nyi bound $E \geq \pi N$; $E=\pi N$ if and only if a set of partial differential equations (the Bogomol'nyi equations) are satisfied. Since like charges repel, the magnetic force between vortices is repulsive. However, there is also a force from the Higgs field, and this is attractive. The balance between the two forces is determined by λ: if $\lambda > 1$, the vortices repel each other; whereas if $\lambda < 1$, the vortices attract. In the critical case $\lambda=1$, the force between vortices is exactly balanced, and there exist static multi-vortex solutions. In fact, one has the following: given N points in the plane, there exists an N-vortex solution of the Bogomol'nyi equations (and hence of the full field equations) with ϕ vanishing at the chosen points (and nowhere else). All static solutions are of this form. These solutions cannot, however, be written down explicitly in terms of elementary functions (except of course for $N=0$).

Monopoles

The abelian Higgs model does not admit three-dimensional solitons, but a nonabelian generalization does – such nonabelian Higgs solitons are called magnetic monopoles. The field content, in the simplest version, is as follows. First, there is a gauge (Yang–Mills) field $F_{\mu\nu}$, with gauge potential A_μ, and with the gauge group being a simple Lie group G. Second, there is a Higgs scalar field ϕ, transforming under the adjoint representation of G (thus ϕ takes values in the Lie algebra of G). For simplicity, G is taken to be SU(2) in what follows. So we may write $A_\mu = \mathrm{i}A^a_\mu \sigma_a, F_{\mu\nu} = \mathrm{i}F^a_{\mu\nu}\sigma_a$, and $\phi = \mathrm{i}\phi^a\sigma_a$, where σ_a are the Pauli matrices. The energy of static ($\partial_0\phi = 0 = \partial_0 A_j$), purely magnetic ($A_0 = 0$) configurations is

$$E = \int \left[\tfrac{1}{2}B^a_j B^a_j + \tfrac{1}{2}(D_j\phi)^a(D_j\phi)^a + \tfrac{1}{4}\lambda(1-\phi^a\phi^a)^2\right] \mathrm{d}^3x$$

where $B^a_j = (1/2)\epsilon_{jkl}F_{kl}$ is the magnetic field. The boundary conditions are $B^a_j \to 0$ and $\phi^a\phi^a \to 1$ as $r \to \infty$; so ϕ restricted to a large spatial 2-sphere becomes a map from S^2 to the unit 2-sphere in the Lie algebra su(2), and as such it has a degree $N \in \mathbf{Z}$. An analytic expression for N is

$$\int B^a_j (D_j\phi)^a \, \mathrm{d}^3x = 2\pi N \qquad [5]$$

At long range, the field resembles an isolated magnetic pole (a Dirac magnetic monopole), with magnetic charge $2\pi N$. Asymptotically, the SU(2) gauge symmetry is spontaneously broken to U(1), which is interpreted as the electromagnetic gauge group.

In 1974, it was observed that this system admits a smooth, finite-energy, stable, spherically symmetric $N = 1$ solution – this is the 't Hooft–Polyakov monopole. There is a Bogomol'nyi lower bound on the energy E: from $0 \le (B + D\phi)^2 = B^2 + (D\phi)^2 + 2B \cdot D\phi$, we get

$$E \ge 2\pi N + \int \tfrac{1}{4}\lambda(1-\phi^a\phi^a)^2 \, \mathrm{d}^3x \qquad [6]$$

where [5] has been used. The inequality [6] is saturated if and only if the Prasad–Sommerfield limit $\lambda = 0$ is used, and the Bogomol'nyi equations

$$(D_j\phi)^a = -B^a_j \qquad [7]$$

hold. The corresponding solitons are called Bogomol'nyi–Prasad–Sommerfield (BPS) monopoles.

The Bogomol'nyi equations [7], together with the boundary conditions described above, form a completely integrable elliptic system of partial differential equations. For any positive integer N, the space of BPS monopoles of charge N, with gauge freedom factored out, is parametrized by a $(4N-1)$-dimensional manifold $\mathcal{M}_N$. This is the moduli space of N monopoles. Roughly speaking, each monopole has a position in space (three parameters) plus a phase (one parameter), making a total of $4|N|$ parameters; an overall phase can be removed by a gauge transformation, leaving $(4|N|-1)$ parameters. In fact, it is often useful to retain the overall phase, and to work with the corresponding $4|N|$-dimensional manifold $\widetilde{\mathcal{M}}_N$. This manifold has a natural metric, which corresponds to the expression for the kinetic energy of the system. A point in $\widetilde{\mathcal{M}}_N$ represents an N-monopole configuration, and the slow-motion dynamics of N monopoles corresponds to geodesics on $\widetilde{\mathcal{M}}_N$; this is the geodesic approximation of monopole dynamics.

The $N = 1$ monopole is spherically symmetric, and the corresponding fields take a simple form; for example, the Higgs field of a 1-monopole located at $r = 0$ is

$$\phi^a = \left[\frac{\coth(2r)}{r} - \frac{1}{2r^2}\right]x^a$$

For $N > 1$, the expressions tend to be less explicit; but monopole solutions can nevertheless be characterized in a fairly complete way. The Bogomol'nyi equations [7] are a dimensional reduction of the self-dual Yang–Mills equations in $\boldsymbol{R}^4$, and BPS monopoles correspond to holomorphic vector bundles over a certain two-dimensional complex manifold ("mini-twistor space"). This leads to various other characterizations of monopole solutions, for example, in terms of certain curves ("spectral curves") on mini-twistor space, and in terms of solutions of a set of ordinary differential equations called the Nahm equations. Having all these descriptions enables one to deduce much about the monopole moduli space, and to characterize many monopole solutions. In particular, there are explicit solutions of the Nahm equations involving elliptic functions, which correspond to monopoles with certain discrete symmetries, such as a 3-monopole with tetrahedral symmetry, and a 4-monopole with the appearance and symmetries of a cube.

Yang–Mills Instantons

Consider gauge fields in four-dimensional Euclidean space $\boldsymbol{R}^4$, with gauge group G. For simplicity, in what follows, G is taken to be SU(2); one can extend much of the structure to more general groups, for example, the simple Lie groups. Let A_μ and $F_{\mu\nu}$

denote the gauge potential and gauge field. The Yang–Mills action is

$$S = -\frac{1}{4}\int \mathrm{tr}\left(F_{\mu\nu}F_{\mu\nu}\right)\mathrm{d}^4x \qquad [8]$$

where we assume a boundary condition, at infinity in $\boldsymbol{R}^4$, such that this integral converges. The Euler–Lagrange equations which describe critical points of the functional S are the Yang–Mills equations

$$D_\mu F_{\mu\nu} = 0 \qquad [9]$$

Finite-action Yang–Mills fields are called instantons. The Euclidean action [8] is used in the path-integral approach to quantum gauge field theory; therefore, instantons are crucial in understanding the path integral.

The dual of the field tensor $F_{\mu\nu}$ is

$$*F_{\mu\nu} = \tfrac{1}{2}\varepsilon_{\mu\nu\alpha\beta}F_{\alpha\beta}$$

The gauge field is self-dual if $*F_{\mu\nu} = F_{\mu\nu}$, and anti-self-dual if $*F_{\mu\nu} = -F_{\mu\nu}$. In view of the Bianchi identity $D_\mu * F_{\mu\nu} = 0$, any self-dual or anti-self-dual gauge field is automatically a solution of the Yang–Mills equations [9]. This fact also follows from the discussion below, where we see that self-dual instantons give local minima of the action.

The Yang–Mills action (and Yang–Mills equations) are conformally invariant; any finite-action solution of the Yang–Mills equations on R^4 extends smoothly to the conformal compactification S^4. Gauge fields on S^4, with gauge group SU(2), are classified topologically by an integer N, namely, the second Chern number

$$N = c_2 = -\frac{1}{8\pi^2}\int \mathrm{tr}\left(F_{\mu\nu} * F_{\mu\nu}\right)\mathrm{d}^4x \qquad [10]$$

From [8] and [10] a topological lower bound on the action is given as follows:

$$\begin{aligned} 0 &\le -\int \mathrm{tr}\left(F_{\mu\nu} - *F_{\mu\nu}\right)\left(F_{\mu\nu} - *F_{\mu\nu}\right)\mathrm{d}^4x \\ &= 8S - 16\pi^2 N \end{aligned}$$

and so $S \ge 2\pi^2 N$, with equality if and only if the field is self-dual. If $N < 0$, we get $S \ge 2\pi^2|N|$, with equality if and only if F is anti-self-dual. So the self-dual (or anti-self-dual) fields minimize the action in each topological class.

For the remainder of this section, we restrict to self-dual instantons with instanton number $N > 0$. The space (moduli space) of such instantons, with gauge equivalence factored out, is an $(8N-3)$-dimensional real manifold. In principle, all these gauge fields can be constructed using algebraic-geometry (twistor) methods: instantons correspond to holomorphic vector bundles over complex projective 3-space (twistor space). One large class of solutions which can be written out explicitly is as follows: for $N = 1$ and $N = 2$ it gives all instantons, while for $N \ge 3$ it gives a $(5N+4)$-dimensional subfamily of the full $(8N-3)$-dimensional solution space. The gauge potentials in this class have the form

$$A_\mu = \mathrm{i}\sigma_{\mu\nu}\partial_\nu \log \phi \qquad [11]$$

where the $\sigma_{\mu\nu}$ are constant matrices (antisymmetric in $\mu\nu$) defined in terms of the Pauli matrices σ_a by

$$\begin{aligned} \sigma_{10} = \sigma_{23} &= \tfrac{1}{2}\sigma_1 \\ \sigma_{20} = \sigma_{31} &= \tfrac{1}{2}\sigma_2 \\ \sigma_{30} = \sigma_{12} &= \tfrac{1}{2}\sigma_3 \end{aligned}$$

The real-valued function $\phi = \phi(x^\mu)$ is a solution of the four-dimensional Laplace equation given by

$$\phi(x^\mu) = \sum_{k=0}^{N}\frac{\lambda_k}{(x^\mu - x_k^\mu)(x^\mu - x_k^\mu)}$$

where the x_k^μ are $N+1$ distinct points in $\boldsymbol{R}^4$, and the λ_k are $N+1$ positive constants: a total of $5N+5$ parameters. It is clear from [11] that the overall scale of ϕ is irrelevant, leaving a $(5N+4)$-parameter family. For $N = 1$ and $N = 2$, symmetries reduce the parameter count further, to 5 and 13, respectively. Although ϕ has poles at the points $x = x_k$, the gauge potentials are smooth (possibly after a gauge transformation).

Finally, it is worth noting that (as one might expect) there is a gravitational analog of the gauge-theoretic structures described here. In other words, one has self-dual gravitational instantons – these are four-dimensional Riemannian spaces for which the conformal-curvature tensor (the Weyl tensor) is self-dual, and the Ricci tensor satisfies Einstein's equations $R_{\mu\nu} = \Lambda g_{\mu\nu}$. As before, such spaces can be constructed using a twistor-geometrical correspondence.

Q-Balls

A Q-ball (or nontopological soliton) is a soliton which has a periodic time dependence in a degree of freedom which corresponds to a global symmetry. The simplest class of Q-ball systems involves a complex scalar field ϕ, with an invariance under the constant phase transformation $\phi \mapsto \mathrm{e}^{\mathrm{i}\theta}\phi$; the Q-balls are soliton solutions of the form

$$\phi(t, \boldsymbol{x}) = \mathrm{e}^{\mathrm{i}\omega t}\psi(\boldsymbol{x}) \qquad [12]$$

where $\psi(\boldsymbol{x})$ is a complex scalar field depending only on the spatial variables $\boldsymbol{x}$. The best-known case is the 1-soliton solution

$$\phi(t,x) = a\sqrt{2}\exp(\mathrm{i}a^2 t)\mathrm{sech}(ax)$$

of the nonlinear Schrödinger equation $\mathrm{i}\phi_t + \phi_{xx} + \phi|\phi|^2 = 0$.

More generally, consider a system (in n spatial dimensions) with Lagrangian

$$\mathcal{L} = \tfrac{1}{2}(\partial_\mu\phi)(\partial^\mu\phi) - U(|\phi|)$$

where $\phi(x^\mu)$ is a complex-valued field. Associated with the global phase symmetry is the conserved Noether charge $Q = \int \mathrm{Im}(\bar{\phi}\phi_t)\,\mathrm{d}^n x$. Minimizing the energy of a configuration subject to Q being fixed implies that ϕ has the form [12]. Without loss of generality, we may take $\omega \geq 0$. Note that $Q = \omega I$, where $I = \int |\psi|^2\,\mathrm{d}^n x$. The energy of a configuration of the form [12] is $E = E_q + E_k + E_p$, where

$$E_q = \frac{1}{2}\int |\partial_j\psi|^2\,\mathrm{d}^n x$$

$$E_k = \tfrac{1}{2}I\omega^2 = \tfrac{1}{2}Q^2/I$$

$$E_p = \int U(|\psi|)\,\mathrm{d}^n x$$

Let us take $U(0) = 0 = U'(0)$, with the field satisfying the boundary condition $\psi \to 0$ as $r \to \infty$.

A stationary Q-lump is a critical point of the energy functional $E[\psi]$, subject to Q having some fixed value. The usual (Derrick) scaling argument shows that any stationary Q-lump must satisfy

$$(2-n)E_q - nE_p + nE_k = 0 \qquad [13]$$

For simplicity, in what follows, let us take $n \geq 3$. Define $m > 0$ by $U''(0) = m^2$; then, near spatial infinity, the Euler–Lagrange equations give $\nabla^2\psi - (m^2 - \omega^2)\psi = 0$. So, in order to satisfy the boundary condition $\psi \to 0$ as $r \to \infty$, we need $\omega < m$.

It is clear from [13] that if $U \geq (1/2)m^2|\psi|^2$ everywhere, then there can be no solution. So $K = \min[2U(|\psi|)/|\psi|^2]$ has to satisfy $K < m^2$. Also, we have

$$E_p = \int U \geq \tfrac{1}{2}KI = (K/\omega^2)E_k > (K/\omega^2)E_p \qquad [14]$$

where the final inequality comes from [13]. As a consequence, we see that ω^2 is restricted to the range

$$K < \omega^2 < m^2 \qquad [15]$$

An example which has been studied in some detail is $\mathrm{U}(f) = f^2[1 + (1-f^2)^2]$; here $m^2 = 4$ and $K = 2$, so the range of frequency for Q-balls in this system is $\sqrt{2} < \omega < 2$. The dynamics of Q-balls in systems such as these turns out to be quite complicated.

See also: Abelian Higgs Vortices; Homoclinic Phenomena; Integrable Systems: Overview; Instantons: Topological Aspects; Noncommutative Geometry from Strings; Sine-Gordon Equation; Topological Defects and Their Homotopy Classification.

Further Reading

Atiyah MF and Hitchin NJ (1988) *The Geometry and Dynamics of Magnetic Monopoles*. Princeton: Princeton University Press.

Coleman S (1988) *Aspects of Symmetry*. Cambridge: Cambridge University Press.

Drazin PG and Johnson RS (1989) *Solitons: An Introduction*. Cambridge: Cambridge University Press.

Goddard P and Mansfield P (1986) Topological structures in field theories. *Reports on Progress in Physics* 49: 725–781.

Jaffe A and Taubes C (1980) *Vortices and Monopoles*. Boston: Birkhäuser.

Lee TD and Pang Y (1992) Nontopological solitons. *Physics Reports* 221: 251–350.

Makhankov VG, Rubakov YP, and Sanyuk VI (1993) *The Skyrme Model: Fundamentals, Methods, Applications*. Berlin: Springer.

Manton NS and Sutcliffe PM (2004) *Topological Solitons*. Cambridge: Cambridge University Press.

Rajaraman R (1982) *Soliton and Instantons*. New York: North-Holland.

Rebbi C and Soliani G (1984) *Solitons and Particles*. Singapore: World Scientific.

Vilenkin A and Shellard EPS (1994) *Cosmic Strings and Other Cosmological Defects*. Cambridge: Cambridge University Press.

Ward RS and Wells RO Jr. (1990) *Twistor Geometry and Field Theory*. Cambridge: Cambridge University Press.

Zakrzewski WJ (1989) *Low Dimensional Sigma Models*. Bristol: IOP.

Source Coding in Quantum Information Theory

N Datta, University of Cambridge, Cambridge, UK
T C Dorlas, Dublin Institute for Advanced Studies, Dublin, Republic of Ireland

Introduction

Two key issues of classical and quantum information theory are storage and transmission of information. An information source produces some outputs (or signals) more frequently than others. Due to this redundancy, one can reduce the amount of space needed for its storage without compromising on its content. This data compression is done by a suitable encoding of the output of the source. In contrast, in the transmission of information through a channel, it is often advantageous to add redundancy to a message, in order to combat the effects of noise. This is done in the form of error-correcting codes. The amount of redundancy which needs to be added to the original message depends on how much noise is present in the channel (see, e.g., Nielson and Chuang (2000)). Hence, redundancy plays complementary roles in data compression and transmission of data through a noisy channel. In this review we focus only on data compression in quantum information theory.

In classical information theory, Shannon showed that there is a natural limit to the amount of compression that can be achieved. It is given by the Shannon entropy. The analogous concept in quantum information theory is the von Neumann entropy. Here, we review some of the main results of quantum data compression and the significance of the von Neumann entropy in this context.

The review is structured as follows. We first give a brief introduction to the Shannon entropy and classical data compression. This is followed by a discussion of quantum entropy and the idea behind quantum source coding. We elaborate on data compression schemes for three different classes of quantum sources, namely memoryless sources, ergodic sources, and sources modeled by Gibbs states of quantum spin systems. In the bulk of the review, we concentrate on source-dependent, fixed-length coding schemes. We conclude with a brief discussion of universal and variable-length coding. We would like to point out that this review article is by no means complete. Due to a restriction on its length, we had to leave out various important aspects and developments of quantum source coding.

Classical Data Compression

Entropy and Source Coding

A simple model of a classical information source consists of a sequence of discrete random variables $X_1, X_2, \ldots, X_n$, whose values represent the output of the source. Each random variable $X_i, 1 \leq i \leq n$, takes values x_i from a finite set, the source alphabet $\mathcal{X}$. Hence, $\underline{X}^{(n)} := (X_1, \ldots, X_n)$ takes values $\underline{x}^{(n)} := (x_1, \ldots, x_n) \in \mathcal{X}^n$. We recall the definition of entropy (or information content) of a source.

If the discrete random variables $X_1, \ldots, X_n$ which take values from a finite alphabet $\mathcal{X}$ have joint probabilities

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = p_n(x_1, \ldots, x_n)$$

then the Shannon entropy of this source is defined by

$$\begin{aligned} H(X_1, \ldots, X_n) &= -\sum_{x_1 \in \mathcal{X}} \cdots \sum_{x_n \in \mathcal{X}} p_n(x_1, \ldots, x_n) \\ &\quad \times \log p_n(x_1, \ldots, x_n) \end{aligned} \qquad [1]$$

Here and in the following, the logarithm is taken to the base 2. This is because the fundamental unit of classical information is a "bit," which takes two values 0 and 1. Notice that $H(X_1, \ldots, X_n)$ in fact only depends on the (joint) probability mass function (p.m.f.) p_n and can also be denoted as $H(p_n)$.

There are several other concepts of entropy, for example, relative entropy, conditional entropy, and mutual information. See, for example, Cover and Thomas (1991) and Nielson and Chuang (2000). It is easy to see that

1. $0 \leq H(X_1, \ldots, X_n) \leq n \log |\mathcal{X}|$, where $|\mathcal{X}|$ denotes the number of letters in the alphabet $\mathcal{X}$. Two other important properties are as follows:
2. $H(X_1, \ldots, X_n)$ is jointly concave in $X_1, \ldots, X_n$ and
3. $H(X_1, \ldots, X_n) \leq H(X_1, \ldots, X_m) + H(X_{m+1}, \ldots, X_n)$ for $m < n$.

The latter property is called subadditivity.

In the next section, analogous quantities are introduced for quantum information and the corresponding properties are stated.

Suppose that the random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.). Then the entropy of each random variable modeling the source is the same and can be denoted by $H(X)$. From the point of view of classical information theory, the Shannon entropy has an important operational definition. It quantifies the minimal

physical resources needed to store data from a classical information source and provides a limit to which data can be compressed reliably (i.e., in a manner in which the original data can be recovered later with a low probability of error). Shannon showed that the original data can be reliably obtained from the compressed version only if the rate of compression is greater than the Shannon entropy. This result is formulated in Shannon's noiseless channel coding theorem (Shannon 1918, Cover and Thomas 1991, Nielson and Chuang 2000) given later.

The Asymptotic Equipartition Property

The main idea behind Shannon's noiseless channel coding theorem is to divide the possible values $x_1, x_2, \ldots, x_n$ of random variables $X_1, \ldots, X_n$ into two classes – one consisting of sequences which have a high probability of occurrence, known as "typical sequences," and the other consisting of sequences which occur rarely, known as "atypical sequences." The idea is that there are far fewer typical sequences than the total number of possible sequences, but they occur with high probability. The existence of typical sequences follows from the so-called "asymptotic equipartition property":

Theorem 1 (AEP). *If $X_1, X_2, X_3, \ldots$ are i.i.d. random variables with p.m.f. $p(x)$, then*

$$-\frac{1}{n}\log p_n(X_1, \ldots, X_n) \xrightarrow{\mathbb{P}} H(X) \qquad [2]$$

where $H(X)$ is the Shannon entropy for a single variable, and $p_n(X_1, \ldots, X_n)$ denotes the random variable taking values $p_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$ with probabilities $p_n(x_1, \ldots, x_n)$.

This theorem has been generalized to the case of sequences of dependent variables $(X_n)_{n\in\mathbb{Z}}$ which are ergodic for the shift transformation defined below. It is easiest to formulate this for an information stream which extends from $-\infty$ to $+\infty$:

Definition A sequence $(X_n)_{n\in\mathbb{Z}}$ is called "stationary" if for any $n_1 < n_2$ and any $x_{n_1}, \ldots, x_{n_2} \in \mathcal{X}$,

$$\begin{aligned}&\mathbb{P}(X_{n_1} = x_{n_1}, \ldots, X_{n_2} = x_{n_2})\\ &\quad = \mathbb{P}(X_{n_1+1} = x_{n_1}, \ldots, X_{n_2+1} = x_{n_2})\end{aligned}$$

We define the shift transformation τ by

$$\tau\big((x_n)_{n\in\mathbb{Z}}\big) = (x'_n)_{n\in\mathbb{Z}}, \quad x'_n = x_{n-1} \qquad [3]$$

Then $(X_n)_{n\in\mathbb{Z}}$ is called "ergodic" if it is stationary and if every subset $A \subset \mathcal{X}^{\mathbb{Z}}$ such that $\tau(A) = A$ has probability 0 or 1, that is, $\mathbb{P}((X_n)_{n\in\mathbb{Z}} \in A) = 0$ or 1.

It is known that $(X_n)_{n\in\mathbb{Z}}$ is ergodic if and only if its probability distribution is extremal in the set of invariant probability measures. The generalization of Theorem 1 (McMillan 1953, Breiman 1957) now reads:

Theorem 2 (Shannon–McMillan–Breiman theorem). *Suppose that the sequence $(X_n)_{n\in\mathbb{Z}}$ is ergodic. Then*

$$\lim_{n\to\infty}\left\{-\frac{1}{n}\log p_n(X_1, \ldots, X_n)\right\} = h_{KS} \quad \text{with probability } 1 \qquad [4]$$

where h_{KS} is the Kolmogorov–Sinai entropy defined by

$$h_{\mathrm{KS}} = \lim_{n\to\infty}\frac{1}{n}H(X_1, \ldots, X_n) = \inf_n \frac{1}{n}H(X_1, \ldots, X_n) \qquad [5]$$

Remark. It follows from the subadditivity property (3) above that the sequence $(1/n)H(p_n)$ is decreasing, and it is obviously bounded below by 0.

We now define the set of typical sequences (or more precisely, ϵ-typical sequences) as follows:

Definition Let $X_1, \ldots, X_n$ be i.i.d. random variables with p.m.f. $p(x)$. Given $\epsilon > 0$, ϵ-typical set $T_\epsilon^{(n)}$ is the set of sequences $(x_1 \ldots x_n)$ for which

$$2^{-n(H(X)+\epsilon)} \le p(x_1 \ldots x_n) \le 2^{-n(H(X)-\epsilon)} \qquad [6]$$

In the case of an ergodic sequence, $H(X)$ is replaced by h_{KS} in [6].

Let $|T_\epsilon^{(n)}|$ denote the total number of typical sequences and $\mathbb{P}\{T_\epsilon^{(n)}\}$ denote the probability of the typical set. Then the following is an easy consequence of Theorem 1.

Theorem 3 (Theorem of typical sequences). *For any $\delta > 0\ \exists\, n_0(\delta) > 0$ such that $\forall n \ge n_0(\delta)$ the following hold*:

(i) $\mathbb{P}\{T_\epsilon^{(n)}\} > 1 - \delta$ *and*
(ii) $(1-\delta)2^{n(H(X)-\epsilon)} \le |T_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}$

Shannon's Noiseless Channel Coding Theorem

Shannon's noiseless channel coding theorem is a simple application of the theorem of typical sequences and says that the optimal rate at which one can reliably compress data from an i.i.d. classical information source is given by the Shannon entropy $H(X)$ of the source.

A "compression scheme" C^n of rate R maps possible sequences $\underline{x} = (x_1, \ldots, x_n)$ to a binary string of length $\lceil nR \rceil$: $C^n : \underline{x} \mapsto \underline{y} = (y_1, \ldots, y_{\lceil nR \rceil})$, where $x_i \in \mathcal{X}$; $|\mathcal{X}| = d$ and $y_i \in \{0, 1\}\, \forall\, 1 \le i \le \lceil nR \rceil$. The corresponding decompression scheme takes the $\lceil nR \rceil$

compressed bits and maps them back to a string of n letters from the alphabet $\mathcal{X}$: $D^n : \underline{y} \in \{0,1\}^{\lceil nR \rceil} \mapsto \underline{x}' = (x'_1, \ldots, x'_n)$. A compression–decompression scheme is said to be "reliable" if the probability that $\underline{x}' \neq \underline{x}$ tends to 0 as $n \to \infty$. Shannon's noiseless channel coding theorem (Shannon 1918, Cover and Thomas 1991) now states

Theorem 4 (Shannon). *Suppose that $\{X_i\}$ is an i.i.d. information source, with $X_i \sim p(x)$ and Shannon entropy $H(X)$. If $R > H(X)$ then there exists a reliable compression scheme of rate R for the source. Conversely, any compression scheme with rate $R < H(X)$ is not reliable.*

Proof (sketch). Suppose $R > H(X)$. Choose $\epsilon > 0$ such that $H(X) + \epsilon < R$. Consider the set $T_\epsilon^{(n)}$ of typical sequences. The method of compression is then to examine the output of the source, to see if it belongs to $T_\epsilon^{(n)}$. If the output is a typical sequence, then we compress the data by simply storing an index for the particular sequence using $\lceil nR \rceil$ bits in the obvious way. If the input string is not typical, then we compress the string to some fixed $\lceil nR \rceil$ bit string, for example, $(00\ldots000)$. In this case, data compression effectively fails, but, in spite of this, the compression–decompression scheme succeeds with probability tending to 1 as $n \to \infty$, since by Theorem 3 the probability of atypical sequences can be made small by choosing n large enough.

If $R < H(X)$, then any compression scheme of rate R is not reliable. This also follows from Theorem 3 by the following argument. Let $\mathcal{S}(n)$ be a collection of sequences $\underline{x}^{(n)}$ of size $|\mathcal{S}(n)| \leq 2^{\lceil nR \rceil}$. Then the subset of atypical sequences in $\mathcal{S}(n)$ is highly improbable, whereas the corresponding subset of typical sequences has probability bounded by $2^{nR} 2^{-nH(X)} \to 0$ as $n \to \infty$. □

Quantum Data Compression

Quantum Sources and Entropy

In quantum information processing systems, information is stored in quantum states of physical systems. The most general description of a quantum state is provided by a density matrix.

A "density matrix" ρ is a positive semidefinite operator on a Hilbert space $\mathcal{H}$, with $\mathrm{tr}\rho = 1$, and the expected value of an operator A on $\mathcal{H}$ is given by

$$\phi(A) = \mathrm{tr}\,(\rho A) \qquad [7]$$

The functional ϕ on $\mathcal{M} = \mathcal{B}(\mathcal{H})$, the algebra of linear operators on $\mathcal{H}$, is positive (i.e., $\phi(A) \geq 0$, if $A \geq 0$) and maps the identity $I \in \mathcal{M}$ to 1. Such a functional is also called a state. Conversely, given such a state on a finite-dimensional algebra $\mathcal{M}$, there exists a unique density matrix ρ_ϕ such that [7] holds, so the concepts can be used interchangeably. (This is not true in the infinite-dimensional case.)

The quantum analog of the Shannon entropy is called the von Neumann entropy. For any quantum state ϕ (or equivalently ρ_ϕ), it is defined by

$$S(\phi) \equiv S(\rho_\phi) := -\mathrm{tr}\big(\rho_\phi \log \rho_\phi\big) \qquad [8]$$

Here we use log to denote $\log_2$ and define $0 \log 0 \equiv 0$, as for the Shannon entropy. Let the density matrix ρ_ϕ have a spectral decomposition

$$\rho_\phi = \sum_{i=1}^{d} \lambda_i |\psi_i\rangle\langle\psi_i| \qquad [9]$$

Here $\{|\psi_i\rangle\}$ is the set of eigenvectors of ρ_ϕ. They form an orthonormal basis of the Hilbert space $\mathcal{H}$. By the fact that ρ_ϕ is positive definite and has trace 1, the eigenvalues λ_i of ρ_ϕ determine a probability distribution. When expressed in terms of the λ_i, the von Neumann entropy of ρ reduces to the Shannon entropy corresponding to this probability distribution (henceforth, the subscript ϕ of ρ_ϕ will be omitted): $S(\rho) = H(\underline{\lambda})$, where $\underline{\lambda} = \{\lambda_1, \ldots, \lambda_d\}$.

The von Neumann entropy has properties analogous to $H(X_1, \ldots, X_n)$, in particular (Ohya and Petz 1993, Nielsen and Chuang 2000)

1. $0 \leq S(\phi) \leq \log(\dim(\mathcal{H}))$;
2. $S(\phi)$ is concave in ϕ; and
3. if ϕ is a state on $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ then $S(\phi) \leq S(\phi_1) + S(\phi_2)$ if ϕ_1 and ϕ_2 are the restrictions of ϕ to $\mathcal{H}_1 \otimes I$ and $I \otimes \mathcal{H}_2$ respectively.

A "quantum information source" in general is defined by a sequence of density matrices $\rho^{(n)}$ on Hilbert spaces $\mathcal{H}_n$ of increasing dimensions N_n given by a decomposition

$$\rho^{(n)} = \sum_k p_k^{(n)} |\Psi_k^{(n)}\rangle\langle\Psi_k^{(n)}| \qquad [10]$$

where the states $|\Psi_k^{(n)}\rangle$ are interpreted as the signal states, and the numbers $p_k^{(n)} \geq 0$ with $\sum_k p_k^{(n)} = 1$, as their probabilities of occurrence. The vectors $|\Psi_k^{(n)}\rangle \in \mathcal{H}_n$ need not be mutually orthogonal.

Compression–Decompression Scheme and Fidelity

To compress data from such a source one encodes each signal state $|\Psi_k^{(n)}\rangle$ by a state $\widetilde{\rho}_k^{(n)} \in \mathcal{B}(\widetilde{\mathcal{H}}_n)$ where $\dim \widetilde{\mathcal{H}}_n = d_c(n) < N_n$. Thus, a compression scheme is a map $\mathcal{C}^{(n)} : |\Psi_k^{(n)}\rangle\langle\Psi_k^{(n)}| \mapsto \widetilde{\rho}_k^{(n)} \in \mathcal{B}(\widetilde{\mathcal{H}}_n)$. The state $\widetilde{\rho}_k^{(n)}$ is referred to as the compressed state. A corresponding decompression scheme is a map $\mathcal{D}^{(n)} : \mathcal{B}(\widetilde{\mathcal{H}}_n) \mapsto \mathcal{B}(\mathcal{H}_n)$. Both $\mathcal{C}^{(n)}$ and $\mathcal{D}^{(n)}$ must be

completely positive maps. In particular, this implies that $\mathcal{D}^{(n)}$ must be of the form

$$\mathcal{D}^{(n)}(\rho) = \sum_i D_i \rho D_i^* \qquad [11]$$

for linear operators $D_i : \widetilde{\mathcal{H}}_n \mapsto \mathcal{H}_n$ such that $\sum_i D_i^* D_i = \boldsymbol{I}$ (see Nielsen and Chuang 2000). Obviously, in order to achieve the maximum possible compression of Hilbert space dimensions per signal state, the goal must be to make the dimension $d_c(n)$ as small as possible, subject to the condition that the information carried in the signal states can be retrieved with high accuracy upon decompression.

The "rate of compression" is defined as

$$R_n := \frac{\log(\dim \widetilde{\mathcal{H}}_n)}{\log(\dim \mathcal{H}_n)} = \frac{\log d_c(n)}{\log N_n}$$

It is natural to consider the original Hilbert space $\mathcal{H}_n$ to be the n-qubit space. In this case $N_n = 2^n$ and hence $\log N_n = n$. As in the case of classical data compression, we are interested in finding the optimal limiting rate of data compression, which in this case is given by

$$R_\infty := \lim_{n\to\infty} \frac{\log d_c(n)}{n} \qquad [12]$$

Unlike classical signals, quantum signal states are not completely distinguishable. This is because they are, in general, not mutually orthogonal. As a result, perfectly reconstructing a quantum signal state from its compressed version is often an impossible task and therefore too stringent a requirement for the reliability of a compression–decompression scheme. Instead, a reasonable requirement is that a state can be reconstructed from the compressed version which is nearly indistinguishable from the original signal state. A measure of indistinguishability useful for this purpose is the average fidelity defined as follows:

$$F_n := \sum_k p_k^{(n)} \langle \Psi_k^{(n)} | \mathcal{D}^{(n)}(\widetilde{\rho}_k^{(n)}) | \Psi_k^{(n)} \rangle \qquad [13]$$

This fidelity satisfies $0 \le F_n \le 1$ and $F_n = 1$ if and only if $\mathcal{D}^{(n)}(\widetilde{\rho}_k^{(n)}) = |\Psi_k^{(n)}\rangle\langle\Psi_k^{(n)}|$ for all k. A compression–decompression scheme is said to be reliable if $F_n \to 1$ as $n \to \infty$.

The key idea behind data compression is the fact that some signal states have a higher probability of occurrence than others (these states playing a role analogous to the typical sequences of classical information theory). These signal states span a subspace of the original Hilbert space of the source and is referred to as the typical subspace.

Schumacher's Theorem for Memoryless Quantum Sources

The notion of a typical subspace was first introduced in the context of quantum information theory by Schumacher (1995) in his seminal paper. He considered the simplest class of quantum information sources, namely quantum memoryless or i.i.d sources. For such a source the density matrix $\rho^{(n)}$, defined through [10], acts on a tensor product Hilbert space $\mathcal{H}_n = \mathcal{H}^{\otimes n}$ and is itself given by a tensor product

$$\rho^{(n)} = \pi^{\otimes n} \qquad [14]$$

Here $\mathcal{H}$ is a fixed Hilbert space (representing an elementary quantum subsystem) and π is a density matrix acting on $\mathcal{H}$; for example, $\mathcal{H}$ can be a single qubit Hilbert space, in which case $\dim \mathcal{H} = 2, \mathcal{H}_n$ is the Hilbert space of n qubits and π is the density matrix of a single qubit. If the spectral decomposition of π is given by

$$\pi = \sum_{i=1}^{\dim \mathcal{H}} q_i |\phi_i\rangle\langle\phi_i| \qquad [15]$$

then the eigenvalues and eigenvectors of $\rho^{(n)}$ are given by

$$\lambda_{\underline{k}}^{(n)} = q_{k_1} q_{k_2} \dots q_{k_n} \qquad [16]$$

and

$$|\psi_{\underline{k}}^{(n)}\rangle = |\phi_{k_1}\rangle \otimes |\phi_{k_2}\rangle \otimes \dots \otimes |\phi_{k_n}\rangle \qquad [17]$$

Thus, we can write the spectral decomposition of the density matrix $\rho^{(n)}$ of an i.i.d. source as

$$\rho^{(n)} = \sum_{\underline{k}} \lambda_{\underline{k}}^{(n)} |\psi_{\underline{k}}^{(n)}\rangle\langle\psi_{\underline{k}}^{(n)}| \qquad [18]$$

where the sum is over all possible sequences $\underline{k} = (k_1 \dots k_n)$, with each k_i taking ($\dim \mathcal{H}$) values. Hence, we see that the eigenvalues $\rho^{(n)}$ are labeled by a classical sequence of indices $\underline{k} = k_1 \dots k_n$.

The von Neumann entropy of such a source is given by

$$S(\rho^{(n)}) \equiv S(\pi^{\otimes n}) = nS(\pi) = nH(X) \qquad [19]$$

where X is the classical random variable with probability distribution $\{q_i\}$.

Let $T_\epsilon^{(n)}$ be the classical typical subset of indices $(k_1 \dots k_n)$ for which

$$\left| -\frac{1}{n} \log(q_{k_1} \dots q_{k_n}) - S(\pi) \right| \le \epsilon \qquad [20]$$

as in the theorem of typical sequences. Defining $\mathcal{T}_\epsilon^{(n)}$ as the space spanned by the eigenvectors $|\psi_{\underline{k}}^{(n)}\rangle$

with $\underline{k} \in T_\epsilon^{(n)}$ then immediately yields the quantum analog of the theorem of typical sequences – Theorem 4 given below. We refer to $\mathcal{T}_\epsilon^{(n)}$ as the typical subspace (or more precisely, the ϵ–typical subspace).

Theorem 4 (Typical subspace theorem). *Fix $\epsilon > 0$. Then for any $\delta > 0 \exists\ n_0(\delta) > 0$ such that $\forall n \geq n_0(\delta)$ and $\rho^{(n)} = \pi^{\otimes n}$, the following are true*:

(i) $\mathrm{Tr}(P_\epsilon^{(n)}\rho^{(n)}) > 1 - \delta$ *and*
(ii) $(1-\delta)2^{n(S(\pi)-\epsilon)} \leq \dim(\mathcal{T}_\epsilon^{(n)}) \leq 2^{n(S(\pi)+\epsilon)}$, *where $P_\epsilon^{(n)}$ is the orthogonal projection onto the subspace $\mathcal{T}_\epsilon^{(n)}$.*

Note that $\mathrm{tr}(P_\epsilon^{(n)}\rho^{(n)})$ gives the probability of the typical subspace. As $\mathrm{tr}(P_\epsilon^{(n)}\rho^{(n)})$ approaches unity for n sufficiently large, $\mathcal{T}_\epsilon^{(n)}$ carries almost all the weight of $\rho^{(n)}$. Let $\mathcal{T}_\epsilon^{(n)\perp}$ denote the orthocomplement of the typical subspace, that is, for any pair of vectors $|\psi\rangle \in \mathcal{T}_\epsilon^{(n)}$ and $|\phi\rangle \in \mathcal{T}_\epsilon^{(n)\perp}$, $\langle\phi|\psi\rangle = 0$. It follows from the above theorem that the probability of a signal state belonging to $\mathcal{T}_\epsilon^{(n)\perp}$ can be made arbitrarily small for n sufficiently large.

Let $P_\epsilon^{(n)}$ denote the orthogonal projection onto the typical subspace $\mathcal{T}_\epsilon^{(n)}$. The encoding (compression) of the signal states $|\Psi_k^{(n)}\rangle$ of [10], is done in the following manner. $\mathcal{C}^{(n)} : |\Psi_k^{(n)}\rangle\langle\Psi_k^{(n)}| \mapsto \tilde{\rho}_k^{(n)}$, where

$$\tilde{\rho}_k^{(n)} := \alpha_k^2 |\tilde{\Psi}_k^{(n)}\rangle\langle\tilde{\Psi}_k^{(n)}| + \beta_k^2 |\Phi_0\rangle\langle\Phi_0| \qquad [21]$$

Here

$$|\tilde{\Psi}_k^{(n)}\rangle := \frac{P_\epsilon^{(n)}|\Psi_k^{(n)}\rangle}{\|P_\epsilon^{(n)}|\Psi_k^{(n)}\rangle\|} \qquad [22]$$

$$\alpha_k := \|P_\epsilon^{(n)}|\Psi_k^{(n)}\rangle\|, \quad \beta_k = \|(\boldsymbol{I} - P_\epsilon^{(n)})|\Psi_k^{(n)}\rangle\|$$

and $|\Phi_0\rangle$ is any fixed state in $\mathcal{T}_\epsilon^{(n)}$.

Obviously $\tilde{\rho}_k^{(n)} \in \mathcal{B}(\mathcal{T}_\epsilon^{(n)})$, and hence the typical subspace $\mathcal{T}_\epsilon^{(n)}$ plays the role of the compressed space. The decompression $\mathcal{D}^{(n)}(\tilde{\rho}_k^{(n)})$ is defined as the extension of $\tilde{\rho}_k^{(n)}$ on $\mathcal{T}_\epsilon^{(n)}$ to $\mathcal{H}_n$:

$$\mathcal{D}^{(n)}\left(\tilde{\rho}_{\underline{k}}^{(n)}\right) = \tilde{\rho}_{\underline{k}}^{(n)} \oplus 0$$

The fidelity of this compression–decompression scheme satisfies

$$\begin{aligned} F_n &= \sum_k p_k^{(n)} \langle\Psi_k^{(n)}|\tilde{\rho}_k^{(n)}|\Psi_k^{(n)}\rangle \\ &= \sum_k p_k^{(n)} \left[\alpha_k^2 |\langle\Psi_k^{(n)}|\tilde{\Psi}_k^{(n)}\rangle|^2 + \beta_k^2 |\langle\Psi_k^{(n)}|\Phi_0\rangle|^2\right] \\ &\geq \sum_k p_k^{(n)} \alpha_k^2 |\langle\Psi_k^{(n)}|\tilde{\Psi}_k^{(n)}\rangle|^2 = \sum_k p_k^{(n)} \alpha_k^4 \\ &\geq \sum_k p_k^{(n)} (2\alpha_k^2 - 1) = 2A_n - 1 \end{aligned} \qquad [23]$$

where $A_n = \mathrm{tr}(P_\epsilon^{(n)}\rho_n)$.

Using the typical subspace theorem, Schumacher (1995) proved the following analog of Shannon's noiseless channel coding theorem for memoryless quantum information sources:

Theorem 5 (Schumacher's quantum coding theorem). *Let $\{\rho_n, \mathcal{H}_n\}$ be an i.i.d. quantum source: $\rho_n = \pi^{\otimes n}$ and $\mathcal{H}_n = \mathcal{H}^{\otimes n}$. If $R > S(\pi)$, then there exists a reliable compression scheme of rate R. If $R < S(\pi)$, then any compression scheme of rate R is not reliable.*

Proof

(i) $R > S(\pi)$. Choose $\epsilon > 0$ such that $R > S(\pi) + \epsilon$. For a given $\delta > 0$, choose the typical subspace as above and choose n large enough so that (i) and (ii) in the typical subspace theorem hold. In particular, $A_n = \mathrm{tr}(P_\epsilon^{(n)}\rho_n) > 1 - \delta$. Thus, the fidelity tends to 1 as $n \to \infty$.

(ii) *Suppose* $R < S(\pi)$. Let the compression map be $\mathcal{C}^{(n)}$. We may assume that $\widetilde{\mathcal{H}}_n$ is a subspace of $\mathcal{H}_n$ with $\dim \widetilde{\mathcal{H}}_n = 2^{nR}$. We denote the projection onto $\widetilde{\mathcal{H}}_n$ as $\widetilde{P}_n$ and let $\tilde{\rho}_k^{(n)} = \mathcal{C}^{(n)}(|\Psi_k^{(n)}\rangle\langle\Psi_k^{(n)}|)$. Since $\tilde{\rho}_k^{(n)}$ is concentrated on $\widetilde{\mathcal{H}}_n$, we have $\tilde{\rho}_k^{(n)} \leq \widetilde{P}_n$ and hence $\mathcal{D}^{(n)}(\tilde{\rho}_k^{(n)}) \leq \mathcal{D}^{(n)}(\widetilde{P}_n)$, for any decompression map $\mathcal{D}^{(n)}$. Inserting into the definition of the fidelity, we then have

$$\begin{aligned} F &\leq \sum_k p_k^{(n)} \langle\Psi_k^{(n)}|\mathcal{D}^{(n)}(\widetilde{P}_n)|\Psi_k^{(n)}\rangle = \mathrm{tr}\left(\rho^{(n)}\mathcal{D}^{(n)}(\widetilde{P}_n)\right) \\ &\leq \sum_{\underline{k} \in T_\epsilon^{(n)}} \lambda_{\underline{k}}^{(n)} \langle\psi_{\underline{k}}^{(n)}|\mathcal{D}^{(n)}(\widetilde{P}_n)|\psi_{\underline{k}}^{(n)}\rangle + \sum_{\underline{k} \notin T_\epsilon^{(n)}} \lambda_{\underline{k}}^{(n)} \end{aligned} \qquad [24]$$

By the typical subspace theorem, the latter sum tends to 0 as $n \to \infty$, and in the sum over $\underline{k} \in T_\epsilon^{(n)}$ we have $\lambda_{\underline{k}}^{(n)} \leq 2^{-n(S(\pi)-\epsilon)}$. The first sum can therefore be bounded as follows:

$$\begin{aligned} &\sum_{\underline{k} \in T_\epsilon^{(n)}} \lambda_{\underline{k}}^{(n)} \langle\psi_{\underline{k}}^{(n)}|\mathcal{D}^{(n)}(\widetilde{P}_n)|\psi_{\underline{k}}^{(n)}\rangle \\ &\quad \leq 2^{-n(S(\pi)-\epsilon)} \sum_{\underline{k}} \langle\psi_{\underline{k}}^{(n)}|\mathcal{D}^{(n)}(\widetilde{P}_n)|\psi_{\underline{k}}^{(n)}\rangle \\ &\quad = 2^{-n(S(\pi)-\epsilon)} \mathrm{tr}\left(\mathcal{D}^{(n)}(\widetilde{P}_n)\right) \\ &\quad = 2^{-n(S(\pi)-\epsilon)} \mathrm{tr}\left(\sum_i D_i \widetilde{P}_n D_i^*\right) \\ &\quad = 2^{-n(S(\pi)-\epsilon)} 2^{nR} \end{aligned} \qquad [25]$$

by the cyclic property of the trace and the fact that $\sum_i D_i^* D_i = \boldsymbol{I}$ and $\dim \widetilde{\mathcal{H}}_n = 2^{nR}$. □

Even for a quantum source with memory, reliable data compression is achieved by looking for a typical subspace $\mathcal{T}_\epsilon^{(n)}$ of the Hilbert space $\mathcal{H}_n$ for a given $\epsilon > 0$. In the following subsections, we discuss two different classes of such sources for which one

can find typical subspaces $\mathcal{T}_\epsilon^{(n)}$ such that the fidelity F_n tends to 1 as $n \to \infty$.

Ergodic Quantum Sources

A quantum generalization of classical ergodic sources is defined as follows. First consider the analog of an infinite sequence of random variables which is a state on the infinite tensor product of a finite-dimensional $*$-algebra $\mathcal{M}$. The latter is given by the norm closure of the increasing sequence of finite tensor products

$$\mathcal{M}_\infty = \overline{\bigcup_n \otimes_{k=-n}^{n} \mathcal{M}} \qquad [26]$$

A translation-invariant state ϕ_∞ on $\mathcal{M}_\infty$ is said to be ergodic if it cannot be decomposed as a (nontrivial) convex combination of other translation-invariant states. The analog of the Kolmogorov–Sinai entropy [5] for an ergodic state ϕ_∞ is called the mean entropy and is given by

$$S_M(\phi_\infty) = \lim_{n\to\infty} \frac{1}{n} S(\phi_n) = \inf_{n\in\mathbb{N}} \frac{1}{n} S(\phi_n) \qquad [27]$$

where ϕ_n is the restriction of ϕ_∞ to $\mathcal{M}_n := \mathcal{M}^{\otimes n}$. Following Hiai and Petz (1991), we define the following quantity for any state ϕ on an arbitrary finite-dimensional $*$-algebra $\mathcal{M}$ and a given $\delta > 0$:

$$\beta_\delta(\phi) = \inf\{\log \mathrm{tr}(q) : q \in \mathcal{M}, q^* = q, \\ q^2 = q, \phi(q) \geq 1 - \delta\} \qquad [28]$$

We also define a state ϕ_∞ on $\mathcal{M}_\infty$ to be completely ergodic if it is ergodic under transformations on $\mathcal{M}_\infty$, induced by l-fold shifts on $\mathbb{Z}$, for arbitrary $l \in \mathbb{N}$. The following theorem is due to Hiai and Petz (1991), who proved it in a slightly more general setting:

Theorem 6 (Hiai and Petz). *Suppose that ϕ_∞ is a completely ergodic state on $\mathcal{M}_\infty$ and $d := \dim \mathcal{M} < \infty$, and set $\phi_n = \phi_\infty{}^{\restriction}{}_{\mathcal{M}_n}$. Then, for any $\delta > 0$, the following hold*:

(i) $$\limsup_{n\to\infty} \frac{1}{n} \beta_\delta(\phi_n) \leq S_M(\phi_\infty) \qquad [29]$$

(ii) $$\liminf_{n\to\infty} \frac{1}{n} \beta_\delta(\phi_n) \geq S_M(\phi_\infty) - \delta \log d \qquad [30]$$

Proof of (i) Choose $r > S_M(\phi_\infty)$ and let $\epsilon < r - S_M(\phi_\infty)$ and $h = r - \epsilon$. By the definition of $S_M(\phi_\infty)$, there exists $l \in \mathbb{N}$ such that $S(\phi_l) < lh$. Let $\{|e_i\rangle\}_{i=1}^{ld}$ be an orthonormal set of eigenvectors of ρ_{ϕ_l}, with corresponding eigenvalues λ_i, that is, let

$$\rho_{\phi_l} = \sum_{i=1}^{ld} \lambda_i p_i \qquad [31]$$

where $p_i = |e_i\rangle\langle e_i|$ is the projection onto $|e_i\rangle$, be the spectral decomposition for ρ_{ϕ_l}. Denote the spectrum $\mathcal{X} = \{\lambda_i\}_{i=1}^{ld}$. For $n \in \mathbb{N}$, introduce the probability measures ν_n on $\mathcal{X}^n$ by

$$\nu_n(A) = \phi_{nl}(q_A) \qquad [32]$$

where, for any $A \subset \mathcal{X}^n$, the projection q_A is defined by

$$q_A = \sum_{(\lambda_{i_1},\ldots,\lambda_{i_n})\in A} p_{i_1} \otimes \ldots \otimes p_{i_n} \qquad [33]$$

Similarly, we define ν_∞ on $\mathcal{X}^{\mathbb{Z}}$. The sequence of random variables $(X_n)_{n\in\mathbb{Z}}$ with distribution ν_∞ is then ergodic since ϕ_∞ is completely ergodic (and hence l-ergodic).

By the Shannon–McMillan–Breiman theorem (Theorem 2),

$$-\frac{1}{n} \log \nu_n(\{(x_1,\ldots,x_n)\}) \to h_{\mathrm{KS}} \qquad [34]$$

almost surely w.r.t. ν_∞, where h_{KS} is the Kolmogorov–Sinai entropy. The latter is given by $h_{\mathrm{KS}} = \lim_{n\to\infty} (1/n) H_n = \inf_{n\in\mathbb{N}} (1/n) H_n$, where

$$H_n = -\sum_{(x_1,\ldots,x_n)\in\mathcal{X}^n} \nu_n(\{(x_1,\ldots,x_n)\}) \\ \times \log \nu_n(\{(x_1,\ldots,x_n)\}) \qquad [35]$$

Notice in particular that

$$h_{\mathrm{KS}} \leq H_1 = S(\phi_l) < lh \qquad [36]$$

If let $T_\epsilon^{(n)}$ be the (typical) subset of $\mathcal{X}^n$ such that

$$-\frac{1}{n} \log \nu_n(\{(x_1,\ldots,x_n)\}) \in (h_{\mathrm{KS}} - \epsilon, h_{\mathrm{KS}} + \epsilon) \qquad [37]$$

for $(x_1,\ldots,x_n) \in T_\epsilon^{(n)}$ then we have $\nu_\infty(T_\epsilon^{(n)}) \geq 1 - \delta$ for n large enough. Moreover, since $\nu_n(\{(x_1,\ldots,x_n)\}) \geq \mathrm{e}^{-n(h_{\mathrm{KS}}+\epsilon)}$ for all $(x_1,\ldots,x_n) \in T_\epsilon^{(n)}$, and the total measure is 1,

$$|T_\epsilon^{(n)}| \leq \mathrm{e}^{n(h_{\mathrm{KS}}+\epsilon)} \leq \mathrm{e}^{n(lh+\epsilon)} \qquad [38]$$

It follows that $\mathrm{tr}(q_{T_\epsilon^{(n)}}) \leq \mathrm{e}^{n(lh+\epsilon)}$ whereas $\phi_{nl}(q_{T_\epsilon^{(n)}}) = \nu_n(T_\epsilon^{(n)}) \geq 1 - \delta$ and we conclude that

$$\frac{1}{nl} \beta_\delta(\phi_{nl}) \leq \frac{n(lh+\epsilon)}{nl} < r \qquad [39]$$

from which [29] follows upon taking $n \to \infty$, since $r > S_M(\phi_\infty)$ was arbitrary. (Notice that $\beta_\delta(\phi_n)$ is decreasing in n since $\mathcal{M}_n \subset \mathcal{M}_{n+1}$.) □

Proof of (ii) Given $\epsilon, \delta > 0$ and $n \in \mathbb{N}$, choose a projection q_n with $\phi_n(q_n) \geq 1 - \delta$ and $\log \mathrm{tr}(q_n) < \beta_\delta(\phi_n) + \epsilon$. Since $S_M(\phi_\infty) = \inf (1/n) S(\phi_n)$ we have

$S_M(\phi_\infty) \le (1/n)S(\phi_n)$. We now use the following lemma:

Lemma 7 *If ϕ is a state on a finite-dimensional $*$-algebra $\mathcal{M}$, and $q \in \mathcal{M}$ is a projection, then*

$$S(\phi) \le H(p) + \phi(q)\log \mathrm{tr}(q) + (1-\phi(q))\log \mathrm{tr}(1-q) \qquad [40]$$

where $H(p) = -p\log p - (1-p)\log(1-p)$ (the binary entropy) with $p = \phi(q)$.

Proof First notice that if $[\rho_\phi, q] = 0$ then the result [40] follows from the simple inequality:

$$-\sum_{i=1}^{m} \tilde{\lambda}_i \log \tilde{\lambda}_i \le \log m \quad \text{if} \quad \sum_{i=1}^{m} \tilde{\lambda}_i = 1 \qquad [41]$$

Indeed, diagonalizing ρ_ϕ, the eigenvalues λ_i divide into two subsets with corresponding eigenvectors belonging to the range of q, respectively, its complement. Considering the first set, we have, if $m = \dim(\mathrm{Ran}(q))$, and taking $\tilde{\lambda}_i = \lambda_i / (\sum_{i=1}^{m} \lambda_i)$ in [41],

$$-\sum_{i=1}^{m} \lambda_i \log \lambda_i \le -\left(\sum_{i=1}^{m} \lambda_i\right) \log\left(\frac{1}{m}\sum_{i=1}^{m} \lambda_i\right) = -\mathrm{tr}(q\rho_\phi)\left[\log \mathrm{tr}(q\rho_\phi) - \log \mathrm{tr}(q)\right]$$

Adding the analogous inequality for the part of the spectrum corresponding to $1-q$, we obtain [40].

In the general case, that is, if $[\rho_\phi, q] \neq 0$, define the unitary $u = 2q - 1$ and the state

$$\phi'(x) = \tfrac{1}{2}[\phi(x) + \phi(uxu)] \qquad [42]$$

Then $[\rho_{\phi'}, q] = 0$ and by concavity of $S(\phi)$ and the result for the previous case

$$H(X) + \phi(q)\log \mathrm{tr}(q) + (1-\phi(q))\log \mathrm{tr}(1-q) \ge S(\phi') \ge S(\phi) \qquad [43]$$

since $\phi'(q) = \phi(q)$. □

Continuing with the proof of (ii), we conclude that

$$S(\phi_n) \le H(p) + \phi_n(q_n)\log \mathrm{tr}(q_n) + (1-\phi(q_n))\log \mathrm{tr}(1-q_n) \le 1 + \beta_\delta(\phi_n) + \epsilon + \delta n \log d$$

Dividing by n and taking the limit we obtain (30). □

It follows from this theorem that we can define a typical subspace in the same way as in Schumacher's theorem. Indeed, given $\delta > 0$ and $\epsilon > 0$, we have that for n large enough, there exists a subspace $\mathcal{T}_\epsilon^{(n)}$ equal to the range of a projection q_n such that $\phi_n(q_n) > 1 - \delta$ and $e^{n(S_M(\phi_\infty) - \delta\log d - \epsilon)} < \dim(\mathcal{T}_\epsilon^{(n)}) = \mathrm{tr}(q_n) < e^{n(S_M(\phi_\infty)+\epsilon)}$. The proof of the quantum analog of the Shannon–McMillan theorem is then similar to that of Schumacher's theorem (Petz and Mosonyi 2001, Bjelaković *et al.* 2004):

Theorem 8 *Let ϕ_∞ be a completely ergodic stationary state on the infinite tensor product algebra $\mathcal{M}_\infty$. If $R > S_M(\phi_\infty)$, then for any decomposition of the form*

$$\rho^{(n)} = \sum p_k^{(n)} |\Psi_k^{(n)}\rangle \langle \Psi_k^{(n)}| \qquad [44]$$

there exists a reliable quantum code of rate R. Conversely, if $R < S_M(\phi_\infty)$ then any quantum compression–decompression scheme of rate R is not reliable.

Remarks Theorem 6 also holds for higher-dimensional information streams, with essentially the same proof. (The existence of the mean entropy is more complicated in that case.) The condition of complete ergodicity in this theorem is unnecessary. Indeed, Bjelaković *et al.* (2004) showed that the result remains valid (also in more than one dimensions) if the state ϕ_∞ of the source is simply ergodic. They achieved this by decomposing a general ergodic state into a finite number of l-ergodic states, and then applying the above strategy to each. It should also be mentioned that a weaker version of Theorem 6 was proved by King and Lesniewski (1998). They considered the entropy of an associated classical source, but did not show that this classical entropy can be optimized to approximate the von Neumann entropy. This had in fact already been proved by Hiai and Petz (1991). The relevance of the latter work for quantum information theory was finally pointed out by Mosonyi and Petz (2001).

Source Coding for Quantum Spin Systems

In this section we consider a class of quantum sources modeled by Gibbs states of a finite strongly interacting quantum spin system in $\Lambda \subset \mathbb{Z}^d$ with $d \ge 2$. Due to the interaction between spins, the density matrix of the source is not given by a tensor product of the density matrices of the individual spins and hence the quantum information source is non-i.i.d. We consider the density matrix to be written in the standard Gibbsian form:

$$\rho^{\omega,\Lambda} = \frac{e^{-\beta H_\Lambda^\omega}}{\Xi^{\omega,\Lambda}} \qquad [45]$$

where $\beta > 0$ is the inverse temperature. Here ω denotes the boundary condition, that is, the configuration of the spins in $\Lambda^c = \mathbb{Z}^d \setminus \Lambda$, and H_Λ^ω is the Hamiltonian acting on the spin system in Λ under this boundary condition. (see Datta and Suhov (2002)

for precise definitions of these quantities). The denominator on the right-hand side of [45] is the partition function.

Note that any faithful density matrix can be written in the form [45] for some self-adjoint operator H_Λ^ω with discrete spectrum, such that $e^{-\beta H_\Lambda^\omega}$ is trace class. However, we consider H_Λ^ω to be a small quantum perturbation of a classical Hamiltonian and require it to satisfy certain hypotheses (see Datta and Suhov (2002)). In particular, we assume that $H_\Lambda = H_{0\Lambda} + \lambda V_\Lambda$, where (1) $H_{0\Lambda}$ is a classical, finite-range, translation-invariant Hamiltonian with a finite number of periodic ground states, and the excitations of these ground states have an energy proportional to the size of their boundaries (Peierls condition); (2) λV_Λ is a translation-invariant, exponentially decaying, quantum perturbation, λ being the perturbation parameter. These hypotheses ensure that the quantum Pirogov–Sinai theory of phase transitions in lattice systems (see, e.g., Datta *et al.* (1996)) applies.

The power of quantum Pirogov–Sinai theory is such that, in proving reliable data compression for such sources, we do not need to invoke the concept of ergodicity.

Using the concavity of the von Neumann entropy $S(\rho^{\omega,\Lambda})$, one can prove that the von Neumann entropy rate (or mean entropy) of the source

$$h := \lim_{\Lambda \nearrow \mathbb{Z}^d} \frac{S(\rho^{\omega,\Lambda})}{|\Lambda|}$$

exists. For a general van Hove sequence, this follows from the strong subadditivity of the von Neumann entropy (see, e.g., Ohya and Petz (1993)).

Let $\rho^{\omega,\Lambda}$ have a spectral decomposition

$$\rho^{\omega,\Lambda} = \sum_j \lambda_j |\psi_j\rangle\langle\psi_j|$$

where the eigenvalues $\lambda_j, 1 \le j \le 2^{|\Lambda|}$, and the corresponding eigenstates $|\psi_j\rangle$, depend on ω and Λ. Let $\mathcal{P}^{\omega,\Lambda}$ denote the probability distribution $\{\lambda_j\}$ and consider a random variable $K^{\omega,\Lambda}$ which takes a value λ_j with probability λ_j:

$$K^{\omega,\Lambda}(\psi_j) = \lambda_j; \quad \mathcal{P}^{\omega,\Lambda}(K^{\omega,\Lambda} = \lambda_j) = \lambda_j$$

The data compression limit is related to asymptotical properties of the random variables $K^{\omega,\Lambda}$ as $\Lambda \nearrow \mathbb{Z}^d$. As in the case of i.i.d. sources, we prove the reliability of data compression by first proving the existence of a typical subspace. The latter follows from Theorem 9 below. The proof of this crucial theorem relies on results of quantum Pirogov–Sinai theory (Datta *et al.* 1996).

Theorem 9 *Under the above assumptions, for β large and λ small enough, for all $\epsilon > 0$*

$$\lim_{\Lambda \nearrow \mathbb{Z}^d} \mathcal{P}^{\omega,\Lambda}\left(\left|\frac{-1}{|\Lambda|}\log K^{\omega,\Lambda} - h\right| \le \epsilon\right) = \lim_{\Lambda \nearrow \mathbb{Z}^d} \sum_j \lambda_j \chi_{\{|-|\Lambda|^{-1}\log\lambda_j - h| \le \epsilon\}} = 1 \quad [46]$$

where $\chi_{\{\ldots\}}$ denotes an indicator function.

Theorem 9 is essentially a law of large numbers for random variables $(-\log K^{\omega,\Lambda})$. The statement of the theorem can be alternatively expressed as follows. For any $\epsilon > 0$,

$$\lim_{\Lambda \nearrow \mathbb{Z}^d} \mathcal{P}^{\omega,\Lambda}\left(2^{-|\Lambda|(h+\epsilon)} \le K^{\omega,\Lambda} \le 2^{-|\Lambda|(h-\epsilon)}\right) = 1 \quad [47]$$

Thus, we can define a typical subspace $\mathcal{T}_\epsilon^{\omega,\Lambda}$ by

$$\mathcal{T}_\epsilon^{\omega,\Lambda} := \text{span}\,\{|\psi_j\rangle : 2^{-|\Lambda|(h+\epsilon)} \le \lambda_j \le 2^{-|\Lambda|(h-\epsilon)}\} \quad [48]$$

It clearly satisfies the analogs of (i) and (ii) of the typical subspace theorem, which implies as before that a compression scheme of rate R is reliable if and only if $R > h$.

Universal and Variable Length Data Compression

Thus far we discussed source-dependent data compression for various classes of quantum sources. In each case data compression relied on the identification of the typical subspace of the source, which in turn required a knowledge of its density matrix. In classical information theory, there exists a generalization of the theorem of typical sequences due to Csiszár and Körner (1981) where the typical set is universal, in that it is typical for every possible probability distribution with a given entropy. This result was used by Jozsa *et al.* (1998) to construct a universal compression scheme for quantum i.i.d sources with a given von Neumann entropy S using a counting argument for symmetric subspaces. This was generalized to ergodic sources by Kaltchenko and Yang (2003) along the lines of Theorem 6. Hayashi and Matsumoto (2002) supplemented the work of Jozsa *et al.* (1998) with an estimation of the eigenvalues of the source (using the measurement smearing technique) to show that a reliable compression scheme exists for any quantum i.i.d source, independent of the value of its von Neumann entropy S, the limiting rate of compression being given by S. If one admits variable length coding, the Lempel–Ziv algorithm gives a completely universal compression scheme, independent of the value of the entropy, in the classical case (Cover and Thomas 1991). This algorithm was generalized to the quantum case for i.i.d sources by Jozsa and Presnell (2003), and to

sources modeled by Gibbs states of free bosons or fermions on a lattice by Johnson and Suhov (2002).

Another important question is the efficiency of the various coding schemes. The above-mentioned schemes for quantum i.i.d. sources are not efficient, in the sense that they have no polynomial time implementation. Recently, it was shown by Bennett *et al.* (2004) that an efficient, universal compression scheme for i.i.d sources can be constructed by employing quantum state tomography.

Acknowledgment

The authors would like to thank Y M Suhov for helpful discussions.

See also: Capacity for Quantum Information; Channels in Quantum Information Theory; Positive Maps on C^*-Algebras.

Further Reading

Bennett CH, Harrow AW, and Lloyd S (2004) Universal quantum data compression via gentle tomography, quant-ph/0403078.

Bjelaković I, Krüger T, Siegmund-Schultze R, and Szkoła A (2004) The Shannon–McMillan theorem for ergodic quantum lattice systems. *Inventiones Mathematicae* 155: 203–222.

Breiman L (1957) The individual ergodic theorem of information theory. *Annals of Mathematical Statistics* 28: 809–811.

Breiman L (1960) The individual ergodic theorem of information theory – correction note. *Annals of Mathematical Statistics* 31: 809–810.

Cover TM and Thomas JA (1991) *Elements of Information Theory*. New York: Wiley.

Csiszár I and Körner J (1981) *Information Theory. Coding Theorems for Discrete Memoryless Systems*. Budapest: Akadémiai Kiadó.

Datta N, Fernández R, and Fröhlich J (1996) Low-temperature phase diagrams of quantum lattice systems. I. Stability for quantum perturbations of classical systems with finitely-many ground states. *Journal of Statistical Physics* 84: 455–534.

Datta N and Suhov Y (2002) Data compression limit for an information source of interacting qubits. *Quantum Information Processing* 1(4): 257–281.

Hayashi M and Matsumoto K (2002) Quantum universal variable-length source coding. *Physical Review A* 66: 022311.

Hiai F and Petz D (1991) The proper formula for the relative entropy and its asymptotics in quantum probability. *Communications in Mathematical Physics* 143: 257–281.

Johnson O and Suhov YM (2002) The von Neumann entropy and information rate for integrable quantum Gibbs ensembles. *Quantum Computers and Computing* 3/1: 3–24.

Jozsa R, Horodecki M, Horodecki P, and Horodecki R (1998) Universal quantum information compression. *Physical Review Letters* 81: 1714–1717.

Jozsa R and Presnell S (2003) Universal quantum information compression and degrees of prior knowledge. *Proceedings of Royal Society of London Series A* 459: 3061–3077.

Kaltchenko A and Yang E-H (2003) Universal compression of ergodic quantum sources. *Quantum Information & Computation* 3: 359–375.

King C and Lesniewski A (1998) Quantum sources and a quantum coding theorem. *Journal of Mathematical Physics* 39: 88–101.

McMillan B (1953) The basic theorems of information theory. *Annals of Mathematical Statistics* 24: 196–219.

Nielson MA and Chuang IL (2000) *Quantum Computation and Quantum Information*. Cambridge: Cambridge University Press.

Ohya M and Petz D (1993) *Quantum Entropy and Its Use*. Heidelberg: Springer.

Petz D and Mosonyi M (2001) Stationary quantum source coding. *Journal of Mathematical Physics* 42: 4857–4864.

Schumacher B (1995) Quantum coding. *Physical Review A* 51: 2738–2747.

Shannon CE (1918) A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423, 623–656.

Spacetime Topology, Causal Structure and Singularities

R Penrose, University of Oxford, Oxford, UK

The Value of Topological Reasoning in General Relativity

Solving the equations of Einstein's general relativity (*see* General Relativity: Overview) can be an exceedingly complicated business; it is commonly found necessary to resort to numerical solutions involving very complex computer codes (*see* Computational Methods in General Relativity: The Theory). The essential content of the basic equations of the theory itself is, however, something that can be phrased in simple geometrical terms, using only basic concepts of differential geometry (*see* General Relativity: Overview). By virtue of this, it is sometimes the case, in general relativity, that geometrical arguments of various kinds – including purely topological ones (i.e., arguments depending only upon the properties of continuity or smoothness) – can be used to great effect to obtain results that are not readily accessible by standard procedures of differential equation theory or by direct numerical calculation.

One particularly significant family of situations where this kind of argument has a key role to play is in the important issue of the singularities that arise in many solutions of the Einstein equations, in which spacetime curvatures may be expected to diverge to infinity. These are exemplified, particularly, by two important classes of solutions of the

Einstein field equations in which singularities arise. In the first instance, we have cosmological models, which tend to exhibit the presence of an initial singularity referred to as the "Big Bang," as was first noted in the standard Friedmann models (which are solutions of the Einstein equations with simple matter sources; *see* Cosmology: Mathematical Aspects). Secondly, we find a final singularity (for local observers) at the endpoint of gravitational collapse to a black hole (where in the relevant region, outside the collapsing matter, Einstein's vacuum equations are normally taken to hold). In either case, there are canonical exact models, in which considerable symmetry is assumed, and where the models indeed become singular at places where the spacetime curvature diverges to infinity. For many years (prior to 1965), there had been much debate as to whether these singularities were an inevitable feature of the general physical situation under consideration, or whether the presence of singularities might be an artifact of the assumed high symmetry. The use of topological-type arguments has established that, in general terms, the occurrence of a singularity is not merely an artifact of symmetry, and cannot generally be removed by the introduction of small (finite) perturbations.

Let us first consider the standard picture, put forward in 1939 by Oppenheimer and Snyder (OS), of the gravitational collapse of an over-massive star to a black hole; see **Figure 1** (and *see* Stationary Black Holes). This assumes exact spherical symmetry. The region external to the matter is described by the well-known Schwarzschild solution of the Einstein vacuum equations, appropriately extended to inside the "Schwarzschild radius" $r = 2mG/c^2$ (G being Newton's gravitational constant and c, the speed of light, and where m is the total mass of the collapsing material; from now, for convenience, we choose units so that $G = c = 1$). In **Figure 1**, this internal extension is conveniently expressed using Eddington–Finkelstein coordinates (r, v, θ, ϕ) (see Eddington (1924) and Finkelstein (1958)), where $v = t + r + 2m\log(r - 2m)$, the metric form being

$$\mathrm{d}s^2 = (1 - 2m/r)\mathrm{d}v^2 - 2\mathrm{d}v\mathrm{d}r - r^2(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2)$$

(The signature convention $+---$ is being adopted here; *see* General Relativity: Overview.) We find that, in this model, there is a singularity (at $r = 0$) at the future endpoint of each world line of collapsing matter. Moreover, no future-timelike line starting inside the horizon can avoid reaching the singularity when we try to extend it, as a timelike curve,

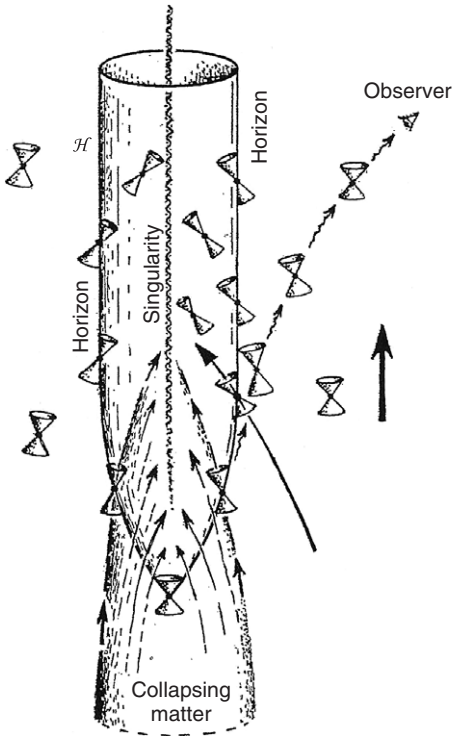


Figure 1 Spacetime diagram of collapse to a black hole. (One spatial dimension is suppressed.) Matter collapses inwards, through the 3-surface that becomes the (absolute) event horizon. No matter or information can escape the hole once it has been formed. The null cones are tangent to the horizon and allow matter or signals to pass inwards but not outwards. An external observer cannot see inside the hole, but only the matter – vastly dimmed and redshifted – just before it enters the hole. (Reproduced with permission from Penrose R. (2004) *The Road to Reality*: *a Complete Guide to the Laws of the Universe*. London: Jonathan Cape.)

indefinitely into the future, where the "horizon" is the three-dimensional region obtained by rotating, over the (θ, ϕ) 2-sphere, the null (lightlike) line which is $r = 2m$ outside the matter region and which is the extension of this line, as a null line, into the past until it meets the axis. It is easy to see that any observer's world line within this horizon is indeed trapped in this sense.

The question naturally arises: how representative is this model? Here, the singularity occurs at the center ($r = 0$), the place where all the matter is directed, and where it all reaches without rebounding. So it may be regarded as unsurprising that the density becomes infinite there. Now, let us suppose that the collapsing material is not exactly spherically symmetrical. Even if it is only slightly (though finitely) perturbed away from this symmetrical situation, having slight (but finite) transverse motions, the collapsing matter is now not all directed exactly towards the center, as it is in the OS model. One might imagine that the singularity

could now be avoided, the different portions of matter just "missing" each other and then being finally flung out again, after some complicated motions, where the density and spacetime curvatures might well become large but presumably still finite. To follow such an irregular collapse in full detail would present a very difficult task, and one would have to carry it out by numerical means. As yet, despite enormous advances in computational technique, a fully effective simulation of such a "generic" collapse is still not in hand. In any case, it is hard to make a convincing case as to whether or not a singularity arises, because as soon as metric or curvature quantities begin to diverge, the computation becomes fundamentally unreliable and simply "gives up." So we cannot really tell whether the failure is due to some genuine divergence or whether it is an artifact. It is thus fortunate that other mathematical techniques are available. Indeed, by use of a differential–topological–causal argument, we find that such perturbations do not help, at least so long as they are small enough not to alter the general character of the collapse, which we find has an "unstoppable" character, so long as a certain criterion is satisfied its early stages.

Trapped Surfaces

But how are we to characterize the collapse as "unstoppable," where no symmetries are to be assumed, and the simple picture illustrated in Figure 1 cannot be appealed to? A convenient characterization is the presence of what is called a "trapped surface." This notion generalizes a key feature of the $0 < r < 2m$ region inside the horizon of the vacuum (Eddington–Finkelstein) picture of Figure 1. To understand what this feature is, consider fixing a point s in the vacuum region of the (v, r)-plane of Figure 1. We must, of course, bear in mind that, because this plane is to be "rotated" about the central vertical axis ($r = 0$) by letting θ and ϕ vary as coordinates on a 2-sphere S^2, the point s actually describes a closed 2-surface $\mathcal{S}$ (coordinatized by θ and ϕ) with topology S^2 (so $\mathcal{S}$ is intrinsically an ordinary 2-sphere). We shall be concerned with the region $I^+(\mathcal{S})$, which is the (chronological) "future" of $\mathcal{S}$, that is, the locus of points q for which a timelike curve exists having a future endpoint at q and a past endpoint on $\mathcal{S}$. We shall also be interested, particularly, in the boundary $\partial I^+(\mathcal{S})$ of $I^+(\mathcal{S})$. This boundary is described, in Figure 1, by the pair of null curves $v = \text{const.}$ and $2r + 4m \log(r - 2m) = \text{const.}$, proceeding into the future from s (and rotated in θ and ϕ). The region $I^+(\mathcal{S})$ itself is represented by that part of Figure 1 which lies between these null curves.

We observe that, in this symmetrical case (s being chosen in the vacuum region), a characterization of s as being "trapped," in the sense that it lies in a region that is within the horizon, is that the future tangents to these null curves both point "inwards," in the sense of decreasing r. Since r is the metric radius of the S^2 of rotation, so that the element of surface area of this sphere is proportional to r^2, it follows that the surface area of the boundary $\partial I^+(\mathcal{S})$ reduces, on both branches, as we move away from $\mathcal{S}$ into the future. The three-dimensional region $\partial I^+(\mathcal{S})$ consists of two null surfaces joined along $\mathcal{S}$, in the sense that their Lorentzian normals are null 4-vectors. For each fixed value of θ and ϕ, this normal is a tangent to one or other of the two null curves of Figure 1, starting at s. For a trapped s, these normals point in the direction of decreasing r, and it follows that the divergence of these normals is negative (so $\rho > 0$ in what follows below).

In the general case, it is this property of negativity of the divergence, at $\mathcal{S}$, of both sets of Lorentzian normals (i.e., of null tangents to $\partial I^+(\mathcal{S})$), that characterizes $\mathcal{S}$ as a trapped surface, where in the general case we must also prescribe $\mathcal{S}$ to be compact and spacelike. But now there are to be no assumptions of symmetry whatever. Such a characterization is stable against small, but finite, perturbations of the location of $\mathcal{S}$, within the spacetime manifold $\mathcal{M}$, and also against small, but finite, perturbations of $\mathcal{M}$ itself.

We can think of a trapped surface in more direct physical/geometrical terms. Imagine a flash of light emitted all over some spacelike compact spherical surface such as $\mathcal{S}$, but now in ordinary flat spacetime, where for simplicity we suppose that $\mathcal{S}$ is situated in some spacelike (flat) 3-hypersurface $\mathcal{H}$, of constant time $t = 0$. There will be one component to the flash proceeding outwards and another proceeding inwards. Provided that $\mathcal{S}$ is convex, the outgoing flash will represent an initial increase of the surface area at every point of $\mathcal{S}$ and the ingoing flash, an initial decrease. In four-dimensional spacetime terms, we express this as positivity of the divergence of the outward null normal and the negativity of the divergence of the inward one. The characteristic feature of a trapped surface is that whereas the ingoing flash will still have an initially reducing surface area, the "outgoing" flash now has the curious property that its surface area is also initially decreasing, this holding at every point of $\mathcal{S}$.

Locally, this is not particularly strange. For a surface wiggling in and out, we are quite likely to find portions of ingoing flash with increasing area,

and portions of outgoing flash with decreasing area. An extreme case in Minkowski spacetime has $\mathcal{S}$ as the intersection of two past light cones. All the null normals to $\mathcal{S}$ point along the generators of these past cones, and therefore all converge into the future. Such a surface $\mathcal{S}$ (indeed spacelike) looks "trapped" everywhere locally, but fails to count as trapped, not being compact. Since there is nothing causally extreme about Minkowski space, it is appropriate not to count such surfaces as "trapped." What *is* the peculiar about a trapped surface is that both ingoing and outgoing flashes are initially decreasing in area, over the entire compact $\mathcal{S}$. (N. B. Hawking and Ellis (1973) adopt a slightly different terminology; the term "trapped," used here, refers to their "closed trapped.")

The Null Raychaudhuri Equation

What do we deduce from the existence of a trapped surface? A glance at **Figure 1** gives us some indication of the trouble. As we trace $\partial I^+(\mathcal{S})$ into the future, we find that its cross-sectional area continues to decrease, until becoming zero at the central singularity. This last feature need not reflect closely what happens in more general cases, with no spherical symmetry. But the reduction in surface area is a general property. This is the first point to appreciate in a theorem (Penrose 1965, 1968, Hawking and Ellis 1973) which indicates the profoundly disturbing physical implications of the existence of a trapped surface in physically realistic gravitational collapse, according to Einstein's general relativity. The surface-area reduction arises from a result known as "Raychaudhuri's equation," in the case of null rays – where we refer to this as the "Sachs" equations. We come to this next.

Although many different notations are used to express the needed quantities, we can here conveniently employ the spin-coefficient formalism, as described elsewhere in this Encyclopedia (*see* Spinors and Spin Coefficients).

Suppose that we have a congruence (smooth three-parameter family) of rays (null geodesics) in four-dimensional spacetime. Let ℓ^a be a real future-null vector, tangent to a null geodesic γ of the congruence, and let m^b be complex-null, also defined along γ, where its real and imginary parts are unit vectors spanning a 2-surface element orthogonal to ℓ^a at each point of γ, so we have

$$\ell_a\ell^a = 0,\ \ell_a m^a = 0,$$
$$m_a m^a = 0,\ \bar{m}_a m^a = -1,$$
$$\ell^a = \bar{\ell}^a$$

where it is assumed that each of ℓ^a, m^a is parallel-propagated along γ:

$$\ell^a \nabla_a \ell^b = 0, \qquad \ell^a \nabla_a m^b = 0$$

(∇_a denoting covariant derivative). The spin-coefficient quantities

$$\rho = m^a \bar{m}^b \nabla_a \ell_b \qquad \text{and} \qquad \sigma = m^a m^b \nabla_a \ell_b$$

are of importance. Here, the real part of ρ measures the convergence of the congruence and the imaginary part defines its rotation; σ measures its shear, where the argument of σ defines the direction (perpendicular to γ) of the axis of shear, and whose strength is defined by $|\sigma|$ (see Penrose and Rindler (1986) for a graphic description of these quantities). Defining propagation derivative along γ by

$$\mathrm{D} = \ell^a \nabla_a$$

we can write the Sachs equations as

$$\mathrm{D}\rho = \rho^2 + \bar{\sigma}\sigma + \Phi$$
$$\mathrm{D}\sigma = 2\rho\sigma + \Psi$$

where $\Phi = -(1/2)R_{ab}\ell^a\ell^b$ and $\Psi = C_{abcd}\ell^a m^b \ell^c m^d$, conventions for the Ricci tensor R_{ab} and the Weyl tensor C_{abcd} being those of General Relativity: Overview (and of Penrose and Rindler (1984)). We note that it is the real Ricci component Φ which governs the propagation of the divergence and the complex Weyl component Ψ which governs the propagation of shear, though there are some non-linear terms. The quantity Φ is normally taken non-negative, since it measures the energy flux across γ (with, in fact $\Phi = 4\pi G T_{ab}\ell^a\ell^b$, where T_{ab} is the energy tensor). The condition that $\Phi \geq 0$ at all points of spacetime and for all null directions ℓ^a, is called the "weak energy condition." (Again there is a minor discrepancy with Hawking and Ellis (1973) who adopt a somewhat stronger "weak energy condition," which is the above but where ℓ^a is also allowed to be future-timelike. Unfortunately, with this terminology, their "weak energy condition" is not strictly weaker than their "strong energy condition.")

It will now be assumed that ρ is real:

$$\rho = \bar{\rho}$$

which is always the case for propagation along the generators of a null hypersurface. The weak energy condition then has an important implication for us. We find that if A is an element of 2-surface area within the plane spanned by the real and imaginary parts of m^a, then (this area element being propagated by D along the lines γ)

$$\mathrm{D}A^{1/2} = -\rho A^{1/2}$$

As a consequence, assuming $\Phi \geq 0$,

$$\mathrm{D}^2 A^{1/2} = -(\bar{\sigma}\sigma + \Phi)A^{1/2} \leq 0$$

This tells us that once the divergence $(-\rho)$ becomes negative, then the area element must reduce to zero sometime in the future along γ, assuming that γ is future-null-complete in the sense that it extends to indefinitely large values of an affine parameter u defined along it, where an affine parameter associated with the parallel-propagated ℓ^a satisfies

$$\ell^a \nabla_a u = 1$$

Such a place where the cross-sectional area pinches down to zero is a singularity of the congruence or null hypersurface, referred to as a "caustic." (There are also terminological confusions arising from different authors defining the term "caustic" in slightly different ways. The terminology used here is slightly discrepant from that of Arnol'd (1992) (Chapter 3).)

From this property, it follows that if we have a trapped surface $\mathtt{S}$, then every generator of $\partial I^+(\mathtt{S})$, if extended indefinitely into the future, must eventually encounter a caustic. This, so far, tells us nothing about actual singularities in the spacetime $\mathtt{M}$; even Minkowski space contains many null hypersurfaces with multitudes of caustic points. However, caustics do tell us something significant about sets like $\partial I^+(\mathtt{S})$, which are the boundaries of future sets, and we come to this shortly.

Causality Properties

First, consider the basic causal relations. If a an b are two points of $\mathtt{M}$, then if there is a nontrivial future-timelike curve in $\mathtt{M}$ from a to b we say that a "chronologically" precedes b and write

$$a \ll b$$

(so it would be possible for some observer's world line to encounter first a and then b). If there is a future-null curve in $\mathtt{M}$ from a to b (trivial or otherwise), we say that a "causally" precedes b and write

$$a \prec b$$

(so it would be possible for a signal to get from a to b). We have the following elementary properties (see Penrose (1972)):

$$a \prec a$$
$$\text{if } a \ll b \text{ then } a \prec b$$
$$\text{if } a \ll b \text{ and } b \ll c \text{ then } a \ll c$$
$$\text{if } a \ll b \text{ and } b \prec c \text{ then } a \ll c$$
$$\text{if } a \prec b \text{ and } b \ll c \text{ then } a \ll c$$
$$\text{if } a \prec b \text{ and } b \prec c \text{ then } a \prec c$$

We generalize the definition of $I^+(\mathtt{S})$, above, to an arbitrary subset $\mathtt{Q}$ in $\mathtt{M}$, obtaining the *chronological future* $I^+(\mathtt{Q})$ and *past* $I^-(\mathtt{Q})$ of $\mathtt{Q}$ in $\mathtt{M}$ by

$$I^+(\mathtt{Q}) = \{q | p \ll q \text{ for some } p \in \mathtt{Q}\}$$
$$I^-(\mathtt{Q}) = \{q | q \ll p \text{ for some } p \in \mathtt{Q}\}$$

The notation $\{q|$ some property of $q\}$ denotes the set of q's with the stated property and the *causal future* $J^+(\mathtt{Q})$ and past $J^-(\mathtt{Q})$ of $\mathtt{Q}$ in $\mathtt{M}$ by

$$J^+(\mathtt{Q}) = \{q | p \prec q \text{ for some } p \in \mathtt{Q}\}$$
$$J^-(\mathtt{Q}) = \{q | q \prec p \text{ for some } p \in \mathtt{Q}\}$$

The $I^\pm(\mathtt{Q})$ are always open sets, but the $J^\pm(\mathtt{Q})$ are not always closed (though they are for any closed set $\mathtt{Q}$ in Minkowski space). Thus, the sets $I^\pm(\mathtt{Q})$ have a more uniform character than the $J^\pm(\mathtt{Q})$, and it is simpler to concentrate, here, on the $I^\pm(\mathtt{Q})$ sets.

The boundary $\partial I^+(\mathtt{Q})$ of $I^+(\mathtt{Q})$ has an elegant characterization:

$$\partial I^+(\mathtt{Q}) = \{q | I^+(q) \subseteq \partial I^+(\mathtt{Q}), \text{but } q \notin I^+(\mathtt{Q})\}$$

and the corresponding statement holds for $\partial I^-(\mathtt{Q})$. Boundaries of futures also have a relatively simple structure, as is exhibited in the following result (for which there is also a version with past and future interchanged):

Lemma *Let $\mathtt{Q} \subseteq \mathtt{M}$ be closed, and $p \in \partial I^+(\mathtt{Q}) - \mathtt{Q}$, then there exists a null geodesic on $\partial I^+(\mathtt{Q})$ with future endpoint at p and which either extends along $\partial I^+(\mathtt{Q})$ indefinitely into the past, or until it reaches a point of $\mathtt{Q}$. It can only extend into the future along $\partial I^+(\mathtt{Q})$ if p is not a caustic point of $\partial I^+(\mathtt{Q})$.*

Beyond a caustic point, the null geodesic would enter into the interior of $I^+(\mathtt{Q})$, but this also happens (more commonly) when crossing another region of null hypersurface on $\partial I^+(\mathtt{Q})$.

We wish to apply this to $\partial I^+(\mathtt{S})$, for a trapped surface $\mathtt{S}$, but we first need a further assumption that $\mathtt{S}$ lies in the interior of the (future) domain of dependence $D^+(\mathtt{H})$ of some spacelike hypersurface $\mathtt{H}$. This region is defined as the totality of points q for which every timelike curve with future endpoint q can be extended into the past until it meets $\mathtt{H}$. One can consider domains of dependence for regions $\mathtt{H}$ other than smooth spacelike surfaces, but it is usual to assume, more generally, that $\mathtt{H}$ is a closed achronal set, where "achronal" means that $\mathtt{H}$ contains no pair of points a, b for which $a \ll b$. We find that every point q in the interior $\mathrm{int}D^+(\mathtt{H})$ of $D^+(\mathtt{H})$ has the further property that all null curves into the past from q will also eventually meet $\mathtt{H}$ if extended sufficiently. The physical significance of $D^+(\mathtt{H})$ is that, for fields with locally Lorentz-invariant and deterministic evolution equations, the (appropriate) initial data on $\mathtt{H}$ will fix the fields throughout $D^+(\mathtt{H})$ (and also

throughout the similarly defined past domain of dependence $D^-(\mathrm{H})$). We find that points in the future Cauchy horizon $H^+(\mathrm{H})$, which is the future boundary of $D^+(\mathrm{H})$ defined by

$$H^+(\mathrm{H}) = D^+(\mathrm{H}) - I^-(D^+(\mathrm{H})),$$

has properties similar to the boundary of a past set, in accordance with the above lemma, and also for the past Cauchy horizon $H^-(\mathrm{H})$, defined correspondingly.

Singularity Theorems and Related Questions

Now, applying our lemma to $\partial I^+(\mathrm{S})$, for a trapped surface $\mathrm{S} \subseteq \mathrm{int}D^+(\mathrm{H})$, we find that every one of its points lies on a null-geodesic segment γ on $\partial I^+(\mathrm{S})$, with past endpoint on S (for if γ did not terminate at S it would have to reach H, which is impossible). Assuming future-null completeness and weak energy ($\Phi \geq 0$), we conclude that if extended far enough into the future, the family of such null geodesics γ must encounter a caustic, and therefore they must leave $\partial I^+(\mathrm{S})$ and enter $I^+(\mathrm{S})$. We finally conclude that $\partial I^+(\mathrm{S})$ must be a compact topological 3-manifold. Using basic theorems, we construct an everywhere timelike vector field in $\mathrm{int}D^+(\mathrm{H})$ which provides a (1–1) continuous map from the compact $\partial I^+(\mathrm{S})$ to H, yielding a contradiction if H is noncompact, thereby establishing the following (Penrose 1965, 1968):

Theorem *The requirement that there be a trapped surface which, together with its closed future, lies in the interior of the domain of dependence of a noncompact spacelike hypersurface, is incompatible with future null completeness and the weak energy condition.*

We notice that this "singularity theorem" gives no indication of the nature of the failure of future null completeness in a spatially open spacetime subject to weak positivity of energy and containing a trapped surface. The natural assumption is that in an actual physical situation of such gravitational collapse, the failure of completeness would arise at places where curvatures mount to such extreme values that classical general relativity breaks down, and must be replaced by the appropriate "quantum geometry" (*see* Quantum Geometry and its Applications, etc.). Hawking (1965) showed how this theorem (in time-reversed form) could also be applied on a cosmological scale to provide a strong argument that the Big-Bang singularity of the standard cosmologies is correspondingly stable. He subsequently introduced techniques from "Morse theory" which could be applied to timelike rather than just null geodesics and, using arguments applied to Cauchy horizons, was able to remove assumptions concerning domains of dependence (e.g., Hawking (1967)). A later theorem (Hawking and Penrose 1970) encompassed most of the earlier ones and had, as one of its implications, that virtually all spatially closed universe models, satisfying a reasonable energy condition and without closed timelike curves, would have to be singular, in this sense of "incompleteness," but again the topological-type arguments used give little indication of the nature or location of the singularities.

Another issue that is not addressed by these arguments is whether the singularities arising from gravitational collapse are inevitably "hidden," as in **Figure 1**, by the presence of a horizon – a conjecture referred to as "cosmic censorship" (see Penrose (1969, 1998)). Without this assumption, one cannot deduce that gravitational collapse, in which a trapped surface forms, will lead to a black hole, or to the alternative which would be a "naked singularity." There are many results in the literature having a bearing on this issue, but it still remains open.

A related issue is that of strong cosmic censorship which has to do with the question of whether singularities might be observable to local observers. Roughly speaking, a naked singularity would be one which is "timelike," whereas the singularities in black holes might in general be expected to be spacelike (or future-null), and in the Big Bang, spacelike (or past-null). There are ways of characterizing these distinctions purely causally, in terms of past sets or future sets (sets Q for which $\mathrm{Q} = I^-(\mathrm{Q})$ or $\mathrm{Q} = I^+(\mathrm{Q})$); see Penrose (1998). If (strong) cosmic censorship is valid, so there are no timelike singularities, the remaining singularities would be cleanly divided into past-type and future-type. In the observed universe, there appears to be a vast difference between the structure of the two, which is intimately connected with the second law of thermodynamics, there appearing to be an enormous constraint on the Weyl curvature (*see* General Relativity: Overview) in the initial singularities but not in the final ones.

Despite the likelihood of singularities arising in their time evolution, it is possible to set up initial data for the Einstein vacuum equations for a wide variety of complicated spatial topologies (*see* Einstein Equations: Initial Value Formulation). On the observational side, however, there seems to be little evidence for anything other than Euclidean spatial topology in our actual universe (which includes black holes). Speculation on the nature of spacetime at the tiniest scales, however, where quantum gravity might be relevant, often involves non-Euclidean topology, however. It may be noted that an early theorem of Geroch established that the constraints of classical Lorentzian geometry do not permit the spatial topology to change without violations of causality (closed timelike curves).

See also: Asymptotic Structure and Conformal Infinity; Boundaries for Spacetimes; Computational Methods in General Relativity: The Theory; Cosmology: Mathematical Aspects; Critical Phenomena in Gravitational Collapse; Einstein Equations: Exact Solutions; Einstein Equations: Initial Value Formulation; General Relativity: Overview; Geometric Analysis and General Relativity; Lorentzian Geometry; Quantum Cosmology; Quantum Geometry and its Applications; Spinors and Spin Coefficients; Stationary Black Holes.

Further Reading

Arno'ld, Beem JK, and Ehrlich PE (1996) *Global Lorentzian Geometry*, (2nd edn). New York: Marcel Dekker.

Eddington AS (1924) A comparison of Whitehead's and Einstein's formulas. *Nature* 113: 192.

Finkelstein D (1958) Past–future asymmetry of the gravitational field of a point particle. *Physical Review* 110: 965–967.

Hawking SW (1965) *Physical Review Letters* 15: 689.

Hawking SW (1967) The occurrence of singularities in cosmology III. Causality and singularities. *Proceedings of the Royal Society (London) A* 300: 187–.

Hawking SW and Ellis GFR (1973) *The Large-Scale Structure of Space-Time:* Cambridge: Cambridge University Press.

Hawking SW and Penrose R (1970) The singularities of gravitational collapse and cosmology. *Proceedings of the Royal Society (London) A* 314: 529–548.

Oppenheimer JR and Snyder H (1939) On continued gravitational contraction. *Physical Review* 56: 455–459.

Penrose R (1965) Gravitational collapse and space-time singularities. *Physics Review Letters* 14: 57–59.

Penrose R (1972) *Techniques of Differential Topology in Relativity*. CBMS Regional Conference Series in Applied Mathematics, no.7. Philadelphia: SIAM.

Penrose R and Rindler W (1984) *Spinors and Space-Time, Vol. 1: Two-Spinor Calculus and Relativistic Fields*. Cambridge: Cambridge University Press.

Penrose R and Rindler W (1984) *Spinors and Space-Time, Vol. 2: Spinor and Twistor Methods in Space-Time Geometry*. Cambridge: Cambridge University Press.

Penrose R (1968) Structure of space-time. In: DeWitt CM and Wheeler JA (eds.) *Battelle Rencontres, 1967 Lectures in Mathematics and Physics*. New York: Benjamin.

Penrose R (1969) Gravitational collapse: the role of general relativity. *Rivista del Nuovo Cimento Serie I, Numero Speciale* 1: 252–276.

Penrose R (1998) The question of cosmic censorship. In: Wald RM (ed.) *Black Holes and Relativistic Stars*, (reprinted in (1999) Journal of Astrophysics and Astronomy 20, 233–248.) Chicago, IL: University of Chicago Press.

Special Lagrangian Submanifolds *see* Calibrated Geometry and Special Lagrangian Submanifolds

Spectral Sequences

P Selick, University of Toronto, Toronto, ON, Canada

Introduction

Spectral sequences are a tool for collecting and distilling the information contained in an infinite number of long exact sequences. Their most common use is the calculation of homology by filtering the object under study and using a spectral sequence to pass from knowledge of the homology of the filtration quotients to that of the object itself. This article will discuss the construction of spectral sequences and the notion of convergence including conditions sufficient to guarantee convergence. Some sample applications of spectral sequences are given.

A differential on an abelian group G is a self-map $d: G \to G$ such that $d^2 = 0$. A morphism of differential groups is a map $f: G \to G'$ such that $d'f = fd$. The condition $d^2 = 0$ guarantees that $\operatorname{Im} d \subset \operatorname{Ker} d$, so to the differential group (G, d) we can associate its homology, $H(G, d) := \operatorname{Ker} d / \operatorname{Im} d$. Often G has extra structure and we require d to satisfy some compatibility condition in order that $H(G, d)$ should also have this structure. For example, a differential graded Lie algebra (L, d) requires a differential d *which satisfies the condition* $d[x, y] = [dx, y] + (-1)^{|x|}[x, dy]$. *While, for simplicity, throughout this article we will always assume that* G is an abelian group, the concepts are readily extended to the case where G is an object of some abelian category and generalizations to nonabelian situations have also been studied.

An important example of extra structure is the case where $G = \bigoplus_{n=-\infty}^{\infty} G_n$ is a graded abelian group. The appropriate compatibility condition for a differential graded group is that d should be homogeneous of degree -1. That is, $d(G_n) \subset G_{n-1}$. In many contexts it is more natural to use superscripts and regard d as having degree $+1$; the two concepts are equivalent via the reindexing convention $G^n := G_{-n}$. Another important example is that

where G forms a graded algebra, meaning that it has a multiplication $G_n \otimes G_k \to G_{n+k}$. To form a differential graded algebra, in addition to having degree -1, d is required to satisfy the Leibniz rule $d(xy) = d(x)y + -(1)^{|x|}xd(y)$ (where $|x|$ denotes the degree of x) familiar from the differentiation of differential forms.

In many cases, G itself is not the main object of interest, but is a relatively large and complicated object, $G = G(X)$, formed by applying some functor G to the object X being studied. For example, X might be some manifold and G could be the set of all differential forms on X with the exterior derivative as d. The presumption is that $H(G(X))$ carries the information we want about X in a much simpler form than the whole of $G(X)$.

A *spectral sequence* (Leray 1946) is defined simply as a sequence $((E^r, d^r))_{r=n_0, n_0+1, \ldots}$ of differential abelian groups such that $E^{r+1} = H(E^r, d^r)$. By reindexing, we could always arrange that $n_0 = 1$, but sometimes it is more natural to begin with some other integer. If all terms (E^r, d^r) of the spectral sequence have the appropriate additional structure, we might refer, for example, to a spectral sequence of Lie algebras. If there exists N such that $E^r = E^N$ for all $r \geq N$ (equivalently $d^r = 0$ for all $r \geq N$), the spectral sequence is said to "collapse" at E^N.

The definition of spectral sequence is so broad that we can say almost nothing of interest about them without putting on some additional conditions. We will begin by considering the most common type of spectral sequence, historically the one that formed the motivating example: the spectral sequence of a filtered chain complex.

Filtered Objects

To study a complicated object X, it often helps to filter X and study it one filtration at a time. A filtration $\mathcal{F}_X$ of a group X is a nested collection of subgroups

$$\mathcal{F}_X := \ldots F_n X \subset F_{n+1} X \subset \cdots \subset X \quad -\infty < n < \infty$$

A morphism $f : \mathcal{F}_X \to \mathcal{F}_Y$ of filtered groups is a homomorphism $f : X \to Y$ such that $f(F_n(X)) \subset F_n(Y)$. The groups $F_n X / F_{n-1} X$ are called the "filtration quotients" and their direct sum $\mathrm{Gr}(\mathcal{F}_X) := \bigoplus_n F_n X / F_{n-1} X$ is called the *associated graded group* of the filtered group $\mathcal{F}_X$. In cases where X has additional structure, we might define special types of filtrations satisfying some compatibility conditions so that $\mathrm{Gr}(\mathcal{F}_X)$ inherits the additional structure. For example, an *algebra filtration* of an algebra X is defined as one for which $(F_n X)(F_k X) \subset F_{n+k} X$.

Since our plan is to study X by computing $\mathrm{Gr}(\mathcal{F}_X)$, the first question we need to consider is what conditions we need to place on our filtration so that $\mathrm{Gr}(\mathcal{F}_X)$ retains enough information to recover X. Our experience from the "5-lemma" suggests that the appropriate way to phrase the requirement is to ask for conditions on the filtrations which are sufficient to conclude that $f : X \to Y$ is an isomorphism whenever $f : \mathcal{F}_X \to \mathcal{F}_Y$ is a morphism of filtered groups for which the induced $\mathrm{Gr}(f) : \mathrm{Gr}(X) \to \mathrm{Gr}(Y)$ is an isomorphism.

It is clear that $\mathrm{Gr}\mathcal{F}_X$ can tell us nothing about $X - (\cup X_n)$ so we require that $X = \cup X_n$. Similarly we need that $\cap X_n = 0$. However, the latter condition is insufficient as can be seen from the following example.

Example 1 Let $X := \bigoplus_{k=1}^{\infty} \mathbb{Z}$ and $Y := \prod_{k=1}^{\infty} \mathbb{Z}$. Set

$$F_n X := \begin{cases} X & \text{if } n \geq 0 \\ \bigoplus_{k=-n}^{\infty} \mathbb{Z} & \text{if } n < 0 \end{cases}$$

$$F_n Y := \begin{cases} Y & \text{if } n \geq 0 \\ \prod_{k=-n}^{\infty} \mathbb{Z} & \text{if } n < 0 \end{cases}$$

and let $f : X \to Y$ be the inclusion. Then $\mathrm{Gr}(f)$ is an isomorphism but f is not.

To phrase the appropriate condition we need the concept of algebraic limits. Given a sequence of objects $\{X_n\}_{n \in \mathbb{Z}}$ and morphisms $f_n : X_n \to X_{n+1}$ in some category, the "direct limit" or "colimit" of the sequence, written $\varinjlim_n F_n X$, is an object X together with morphisms $g_n : X_n \to X$ satisfying $g_{n+1} \circ f_n = g_n$, having the universal property that given any object X' together with maps $g'_n : X_n \to X'$ satisfying $g'_{n+1} \circ f_n = g'_n$, there exists a unique morphism $h : X \to X'$ such that $g'_n = h \circ g_n$ for all n. By the usual categorical argument the object X, if it exists, is unique up to isomorphism. The dual concept, "inverse limit" or simply "limit" of the sequence, written $\varprojlim_n F_n X$, is obtained by reversing the directions of the morphisms. For intuition, we note that these notions share, with the notion of limits of sequences in calculus, the properties that changing the terms X_n only for $n < N$ does not affect $\varinjlim_n F_n X$, and if the sequence stabilizes at N (i.e., the morphisms f_n are isomorphisms for all $n \geq N$), then $\varinjlim_n F_n X \cong X_N$. Similarly $\varprojlim_n F_n X$ depends only upon behavior of the sequence as $n \to -\infty$. Limits over partially ordered sets other than $\mathbb{Z}$ can also be taken but we shall not need them in this article. Although limits need not exist in general, in the category of abelian groups, both the direct and inverse limit exist for any sequence and are given explicitly by the following constructions. $\varinjlim_n F_n X = \bigoplus X_n / \sim$ where, letting $i_k : X_k \to \bigoplus X_n$ be the

canonical inclusion, the equivalence relation is generated by $i_n(x) \sim i_{n+1} f(x)$ for $x \in X_n$. $\varprojlim_n F_n X = \{(x_n) \in \Pi X_n \,|\, f_n(x_n) = x_{n+1}\ \forall n\}$.

The condition needed is that our filtrations should be *bicomplete*, defined as follows. $\mathcal{F}_X$ is called "cocomplete" if the canonical map $X \to \varinjlim_n F_n X$ is an isomorphism and $\mathcal{F}_X$ is called "complete" if $X \to \varprojlim_n X/F_n X$ is an isomorphism. $\mathcal{F}_X$ is called bicomplete if it is both complete and cocomplete. Note that $\mathcal{F}_X$ cocomplete is equivalent to $\cup F_n X = X$ but $\mathcal{F}_X$ complete is stronger than $\cap F_n X = 0$.

Theorem 1 (Comparison theorem). *Let $\mathcal{F}_X$ be bicomplete and let $\mathcal{F}_Y$ be cocomplete with $\cap F_n Y = 0$. Suppose that $f : \mathcal{F}_X \to \mathcal{F}_Y$ is a morphism such that $\mathrm{Gr}(f) : \mathrm{Gr}(X) \to \mathrm{Gr}(Y)$ is an isomorphism. Then $f : X \to Y$ is an isomorphism.*

Filtered Chain Complexes

A *chain complex* (C, d) of abelian groups consists of abelian groups C_n for $n \in \mathbb{Z}$ together with homomorphisms $d_n : C_n \to C_{n-1}$ such that $d_n \circ d_{n+1} = 0$ for all n. To the chain complex (C, d) we can associate the differential (abelian) group $(C_*, d) := \bigoplus_{n=-\infty}^{\infty} C_n$ with $d|_{C_n}$ induced by d_n. We often write simply C if the differential is understood. The dual notion in which d has degree $+1$ is called a *cochain complex* and the concepts are equivalent through our convention $C^n := C_{-n}$.

Theorem 2 (Homology commutes with direct limits). $H(\varinjlim_n C_n) = \varinjlim_n H(C_n)$.

As we shall see later, failure of homology to commute with inverse limits is a source of great complication in working with spectral sequences.

Let $\mathcal{F}_C$ be a filtered chain complex. In many applications, our goal is to compute $H_*(C)$ from a knowledge of $H_*(F_n C/F_{n-1} C)$ for all n. The overall plan, which is not guaranteed to be successful in general, would be:

1. use the given filtration on C to define a filtration on $H_*(C)$,
2. use our knowledge of $H_*(\mathrm{Gr}\, C)$ to compute $\mathrm{Gr}\, H_*(C)$,
3. reconstruct $H_*(C)$ from $\mathrm{Gr}\, H_*(C)$.

To begin, set $F_n(H_* C) := \mathrm{Im}(s_n)_*$, where $s_n : F_n(C) \to C$ is the inclusion (chain) map from the filtration. The spectral sequence which we will define for this situation can be regarded as a method of keeping track of the information contained in the infinite collection of long exact homology sequences coming from the short exact sequences $0 \to F_{n-1} C \to F_n C \to F_n C/F_{n-1} C \to 0$. When working with a long exact sequence, knowledge of two of every three terms gives a handle on computing the remaining terms but does not, in general, completely determine those terms, which explains intuitively why we have some reason to hope that a spectral sequence might be useful and also why it is not guaranteed to solve our problem.

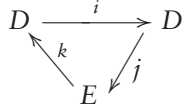
Before proceeding with our motivating example, we digress to discuss spectral sequences formed from *exact couples*.

Exact Couples

In this section, we will define exact couples, show how to associate a spectral sequence to an exact couple, and discuss some properties of spectral sequences coming from exact couples. As we shall see, a filtered chain complex gives rise to an exact couple and we will examine this spectral sequence in greater detail.

Exact couples were invented by Massey and many books use them as a convenient method of constructing spectral sequences. Other books bypass discussion of exact couples and define the spectral sequence coming from a filtered chain complex directly.

Definition 1 An "exact couple" consists of a triangle



containing abelian groups D, E, and together with homomorphisms i, j, k such that the diagram is exact at each vertex.

In the following, to avoid conflicting notation considering the many superscripts and subscripts which will be needed, we use the convention that an n-fold composition will be written $f^{\circ n}$ rather than the usual f^n.

Given an exact couple, set $d := jk : E \to E$. By exactness, $kj = 0$, so $d^{\circ 2} = jkjk = 0$ and therefore (E, d) forms a differential group. To the exact couple we can associate another exact couple, called its derived couple, as follows. Set $D' := \mathrm{Im}\ i \subset D$ and $E' := H(E, d)$. Define $i' := i|_D$ and let $j' : D' \to E'$ be given by $j'(iy) := \overline{j(y)}$, where $\overline{x}$ denotes the equivalence class of x. The map $k' : E' \to D'$ is defined by $k'(\bar{z}) := kz$. One checks that the maps j' and k' are well defined and that (D', E', i', j', k') forms an exact couple. Therefore, from our original exact couple, we can inductively form a sequence of exact couples $(D^r, E^r, i^r, j^r, k^r)_{r=1}^{\infty}$ with $D^1 := D, E^1 := E, D^r := (D^{r-1})'$

and $E^r := (E^{r-1})'$. This gives a spectral sequence $(E^r, d^r)_{r=1}^{\infty}$ with $d^r = j^r k^r$.

To the filtered chain complex $\mathcal{F}_C$, we can associate an exact couple as follows. Set $D := \bigoplus_{p,q} D_{p,q}$ where $D_{p,q} = H_{p+q}(F_pC)$ and $E := \bigoplus_{p,q} E_{p,q}$ where $E_{p,q} = H_{p+q}(F_pC/F_{p-1}C)$. The long exact homology sequences coming from the sequences $0 \to F_{p-1}C \xrightarrow{a} F_pC \xrightarrow{b} F_pC/F_{p-1}C \to 0$ give rise, for each p and q, to maps $a_* : D_{p-1,q+1} \to D_{p,q}$, $b_* : D_{p,q} \to E_{p,q}$, and $\partial : E_{p,q} \to D_{p-1,q}$. Define $i : D \to D$ to be the map whose restriction to $D_{p-1,q+1}$ is the composition of a_* with the canonical inclusion $D_{p,q} \to D$. Similarly, define $j : D \to E$ and $k : E \to D$ to be the maps whose restrictions to each summand are the compositions of b_* and ∂ with the inclusions. The indexing scheme for the bigradations is motivated by the fact that in many applications it causes all of the nonzero terms to appear in the first quadrant, so it is the most common choice, although one sometimes sees other conventions.

There is actually a second exact couple we could associate to $\mathcal{F}_C$, which yields the same spectral sequence: use the same E as above but replace D by $\bigoplus D_{p,q}$ with $D_{p,q} = H_{p+q+1}(C/F_pC)$, and define i, j, and k in a manner similar to that above.

When dealing with cohomology rather than homology, the usual starting point would be a system of inclusions of cochain complexes $\cdots F^{n+1}C \subset F^nC \subset F^{n-1} \subset \cdots \subset C$. This can be reduced to the previous case by replacing the cochain complex C by a chain complex C_* using the convention $C_p := C^{-p}$ and filtering the result by $F_nC_* := F^{-n}C$. The usual practice, equivalent to the above followed by a rotation of 180°, is to leave the original indices and instead reverse the arrows in the exact couple. In this case, it is customary to write $D_r^{p,q}$ and $E_r^{p,q}$ for the terms in the exact couple and spectral sequence.

In applications, it is often the case that E^1 is known and that our goal includes computing D^1. The example of the filtered chain complex with the assumption that we know $H_*(F_pC/F_{p-1}C)$ for all p is fairly typical.

Since each D^r is contained in D^{r-1} and each E^r is a subquotient of E^{r-1}, the terms of these exact couples get smaller as we progress. To get properties of the spectral sequence, we need to examine this process and, in particular, analyze that which remains in the spectral sequence as we let r go to infinity.

For $x \in E$, if $dx = 0$ then $\bar{x}$ belongs to E^2 and so $d^2(\bar{x})$ is defined. In the following, we shall usually simplify the notation by writing simply x in place of $\bar{x}$ and writing $d^r x = 0$ to mean "$d^r x$ is defined and equals 0."

If $dx = 0, \ldots, d^{r-1}x = 0$, then x represents an element of E^r and $d^r x$ is defined. Set $Z^r := \{x \in E \mid d^m x = 0 \; \forall m \le r\}$. Then $E^{r+1} \cong Z^r/\sim$ where $x \sim y$ if there exists $z \in E$ such that for some $t \le r$ we have $d^m z = 0$ for $m < t$ (thus $d^t z$ is defined) and $d^t z = x - y$. With this as motivation, we set $Z^\infty := \cap_r Z^r = \{x \in E \mid d^m x = 0 \; \forall m\}$ (known as the "infinite cycles") and define $E^\infty := Z^\infty/\sim$ where $x \sim y$ if there exists $z \in E$ such that for some t we have $d^m z = 0$ for $m < t$ and $d^t z = x - y$.

Notice that $D^{r+1} = \mathrm{Im} i^{\circ r} \cong D/\mathrm{Ker}\, i^{\circ r}$. There is no analog of this statement for $r = \infty$. Instead we have separate concepts so we set $D^\infty := D/\cup_r \mathrm{Ker}\, i^{\circ r}$ and ${}^\infty D := \cap_r \mathrm{Im} i^{\circ r}$. The analog of the rth-derived exact couple when $r = \infty$ is the following exact sequence.

Theorem 3 *There are maps induced by i, j, and k producing an exact sequence*

$$0 \to D^\infty \xrightarrow{i^\infty} D^\infty \xrightarrow{j^\infty} E^\infty \xrightarrow{{}^\infty k} {}^\infty D \xrightarrow{{}^\infty i} {}^\infty D$$

The fact that we were able to add the 0 term to the left of this sequence but not the right can be traced to the fact that $\varinjlim$ preserves exactness but $\varprojlim$ does not.

In our motivating example, the terms of the initial exact couple came with a bigrading $D = \bigoplus D_{p,q}$ and $E = \bigoplus E_{p,q}$ and writing $|f|$ for the bidegree of a morphism f we had: $|i| = (1, -1)$; $|j| = (0, 0)$; $|k| = (-1, 0)$; $d = (-1, 0)$. It follows that $|i^r| = (1, -1)$; $|j^r| = (-r + 1, r - 1)$; $|k^r| = (-1, 0)$; $|d^r| = (-r, r - 1)$ which is considered the standard bigrading for a bigraded exact couple. Similarly, the standard bigrading for a bigraded spectral sequence is one such that $|d^r| = (-r, r - 1)$.

We observed earlier that terms of an exact couple and its corresponding spectral sequence get smaller as $r \to \infty$ as each is a subquotient of its predecessor. Note that the bigrading is such that this applies to each pair of coordinates individually (e.g., $E_{p,q}^{r+1}$ is a subquotient of $E_{p,q}^r$) and so in particular if the p, q-position ever becomes 0 that position remains 0 forevermore.

Convergence of Graded Spectral Sequences

As noted earlier, the definition of spectral sequence is so broad that we need to put some conditions on our spectral sequences to make them useful as a computational tool. From now on, we will restrict attention to spectral sequences arising from exact couples in which $D = \bigoplus D_p$ and $E = \bigoplus E_p$ are graded with $i|_{D_p} \subset D_{p+1}, j|_{D_p} \subset E_p$, and $k|_{E_p} \subset D_{p-1}$. All the spectral sequences which have been studied

to date satisfy this condition and in fact most also have a second gradation as in the case of our motivating example. To see how to proceed, we examine that case more closely.

For a filtered chain complex $\mathcal{F}_C$ with structure maps $s_p : F_pC \to C$ we defined $F_p(H_*(C)) = \operatorname{Im} s_{p_*}$. If $x = i^{\circ(r-1)}y$ belongs to

$$D^r_{p,q} = \operatorname{Im} i^{\circ(r-1)} : H_{p+q}(F_{p-r+1}C) \to H_{p+q}(F_pC)$$

then $(s_p)_*x = (s_p)_*i^{\circ(r-1)}y = (s_{p+1})_*i^{\circ r}y = (s_{p+1})_*ix$. Therefore, we have a commutative diagram

$$\begin{array}{ccc} D^r_{p,q} & \longrightarrow & F_p(H_{p+q}(C)) \\ \downarrow i & & \downarrow \\ D^{r+1}_{p+1,q-1} & \longrightarrow & F_{p+1}(H_{p+q}(C)) \end{array}$$

yielding a map

$$\begin{aligned} D^{r+1}_{p+1,q-1}/D^r_{p,q} &\to F_{p+1}\big(H_{p+q}(C)\big)/F_p\big(H_{p+q}(C)\big) \\ &= \operatorname{Gr}_{p+1}(H_{p+q})C \end{aligned}$$

Letting r go to infinity, we get an induced map $\phi : D^\infty/i^\infty(D^\infty) \to \operatorname{Gr}(H(C))$.

Theorem 4 *If $\mathcal{F}_H(C)$ is cocomplete then*

(i) $D^\infty = F_n(H(C))$;
(ii) $\phi : D^\infty/i^\infty(D^\infty) \to \operatorname{Gr}(H(C))$ *is an isomorphism*;
(iii) *There is an exact sequence* $0 \to \operatorname{Gr}(H(C)) \xrightarrow{j^\infty} E^\infty \xrightarrow{{}^\infty k} {}^\infty D \xrightarrow{{}^\infty i} {}^\infty D$.

We say that the spectral sequence (E^r) "abuts" to $\mathcal{F}_L$ if there is an isomorphism $\operatorname{Gr}L \to E^\infty$. Here we mean an isomorphism of graded abelian groups, which makes sense since under our assumptions E^r inherits a grading from E^1 for each r. If in addition the filtration on L is cocomplete, we say that (E^r) "weakly converges" to $\mathcal{F}_L$ and if it is bicomplete we say that (E^r) "converges" (or *strongly converges*) to $\mathcal{F}_L$. The notation $(E^r) \Rightarrow \mathcal{F}_L$ (or simply $(E^r) \Rightarrow L$ when the filtration on L is either understood or unimportant) is often used in connection with convergence but there is no universal agreement as to which of the three concepts (abuts, weakly converges, or converges) it refers to! In this article, we will also use the expression (E^r) "quasiconverges" to $\mathcal{F}_L$ to mean that the spectral sequence weakly converges to $\mathcal{F}_L$ with $\cap_n F_nL = 0$. (Note: the terminology quasiconverges is nonstandard although the concept has appeared in the literature, sometimes under the name *converges*.)

While it would be overstating things to claim that convergence of the spectral sequence shows that E^∞ determines $H(C)$, it is clear that convergence is what we need in order to expect that E^∞ contains enough information to possibly reconstruct $H(C)$. The sense in which this is true is stated more precisely in the following theorem.

Theorem 5 (Spectral sequence comparison theorem). *Let $f = (f^r) : (E^r) \to \tilde{E}^r$ be a morphism of spectral sequences.*

(i) *If $f : E^N \to \tilde{E}^N$ is an isomorphism for some N, then f^r is an isomorphism for all $r \geq N$ (including $r = \infty$).*
(ii) *Suppose in addition that (E^r) converges to $\mathcal{F}_X$ and $(\tilde{E}^r)$ quasiconverges to $\mathcal{F}_{\tilde{X}}$. Let $\phi : \mathcal{F}_X \to \mathcal{F}_{\tilde{X}}$ be a morphism of filtered abelian groups which is compatible with f. (i.e., there exist isomorphisms $\eta : \operatorname{Gr} X \cong E^\infty$ and $\tilde{\eta} : \operatorname{Gr} \tilde{X} \cong \tilde{E}^\infty$ such that $f^\infty \circ \eta = \tilde{\eta} \circ \operatorname{Gr}(f)$). Then $f : X \to \tilde{X}$ is an isomorphism.*

Within the constraints provided by Theorem 5, a spectral sequence might have many limits. A typical calculation of some group Y by means of spectral sequences might proceed as an application of Theorem 5 along the lines of the following plan.

1. Subgroups F_nY forming a filtration of Y are defined, although usually not computable at this point. The subgroups are chosen in a manner that seems natural bearing in mind that to be useful it will be necessary to show convergence properties.
2. Directly or by means of an exact couple, a spectral sequence is defined in a manner that seems to be related to the filtration.
3. Some early term of the spectral sequence (usually E^1 or E^2) is calculated explicitly and the differentials d^r are calculated successively resulting in a computation of E^∞.
4. With the aid of the knowledge of E^∞, a conjecture $Y = G$ is formulated for some G.
5. A suitable filtration on G and a map of filtrations $\mathcal{F}_G \to \mathcal{F}_Y$ or $\mathcal{F}_Y \to \mathcal{F}_G$ are defined.
6. The spectral sequence arising from $\mathcal{F}_G$ is demonstrated to converge to G.
7. The original spectral sequence is demonstrated to converge to Y and Theorem 5 is applied.

The hardest steps are usually (3) and (7). For step (3), in most cases the calculations require knowledge which cannot be obtained from the spectral sequence itself, although the spectral sequence machinery plays its role in distilling the information and pointing the way to exactly what needs to be calculated. Steps (4)–(6) are frequently very easy, and often not stated explicitly, with "by construction of G" being the most common justification of (6). We now discuss the types of considerations involved in step (7).

Convergence of a spectral sequence to a desired L can be difficult to verify in general partly because

the conditions are stated in terms of some filtration (usually understood only in a theoretical sense) on an initially unknown L rather than in terms of properties of the spectral sequence itself or an exact couple from which it arose. Theorems 2 and 4(ii) give us the following extremely important special case in which we can conclude convergence to $H(C)$ of the spectral sequence for $\mathcal{F}_C$ based on conditions that are often easily checked.

Theorem 6 *If $\mathcal{F}_C$ is a filtered chain complex such that $\mathcal{F}_C$ is cocomplete and there exists M such that $H(F_nC)=0$ for $n<M$, then the spectral sequence for $\mathcal{F}_C$ converges to $H(C)$.*

Although the second hypothesis, which implies that ${}^{\infty}D=0$, is very strong it handles the large numbers of commonly used filtrations which are 0 in negative degrees.

Under the conditions of Theorem 6, inserting the bigradings into Theorem 4 gives a short exact sequence $0\to D^{\infty}_{p-1,q+1}\to D^{\infty}_{p,q}\to E^{\infty}_{p,q}\to 0$ with $D^{\infty}_{p,q}\cong F_p(H_{p+q}(X))$; equivalently

$$F_k(H_n(C))/F_{k-1}(H_n(C))\cong E^{\infty}_{k,n-k}$$

Thus, the only E^{∞}-terms relevant to the computation to $H_n(C)$ are those on the diagonal $p+q=n$. In the important case of a first quadrant spectral sequence ($E^r_{p,q}=0$ if $p<0$ or $q<0$), the number of nonzero terms on any diagonal is finite so the E^{∞}-terms on the diagonal $p+q=n$ give a finite composition series for each $H_n(C)$.

Here is an elementary example of an application of a spectral sequence.

Example 2 Let $S_*(\,)$ denote the singular chain complex, let $H_*(\,):=H_*(S_*(\,))$ denote singular homology, and let $H^{\mathrm{cell}}_*(\,)$ denote cellular homology. Let X be a CW-complex with n-skeleton $X^{(n)}$. The inclusions $S_*(X^{(n)})\to S_*(X)$ yield a filtration on $S_*(X)$. In the associated spectral sequence,

$$E^1_{p,q}=H_{p+q}\left(X^{(p)}/X^{(p-1)}\right)\cong\begin{cases}\text{free abelian group on the } p\text{-cells of } X & \text{if } q=0\\ 0 & \text{if } q\neq 0\end{cases}$$

The differential

$$d^1_{p,0}:H_p\left(X^{(p)}/X^{(p-1)}\right)\to H_{p-1}\left(X^{(p-1)}/X^{(p-2)}\right)$$

is the definition of the differential in cellular homology. Therefore,

$$E^2_{p,q}=\begin{cases}H^{\mathrm{cell}}(X) & \text{if } q=0\\ 0 & \text{if } q\neq 0\end{cases}$$

Looking at the bidegrees, the domain or range of $d^2_{p,q}$ is zero for each p and q so $d^2=0$, and similarly $d^r=0$ for all $r>2$. Therefore, the spectral sequence collapses with $E^2=E^{\infty}$. The spectral sequence converges to $H_*(X)$ so the terms on the diagonal $p+q=n$ form a composition series for $H_n(X)$. Since the $(n,0)$ term is the only nonzero term on this diagonal, $H_n(X)\cong H^{\mathrm{cell}}_n(X)$. That is, "cellular homology equals singular homology."

Returning to the general situation, set $L_{\infty}:=\varinjlim_n D_n$ and $L_{-\infty}:=\varprojlim_n D_n$. Filter L_{∞} by $F_nL_{\infty}:=\mathrm{Im}(D_n\to L_{\infty})$ and filter $L_{-\infty}$ by $F_nL_{-\infty}:=\mathrm{Ker}(L_{-\infty}\to D_n)$. It follows from the definitions that $F_nL_{\infty}=D^{\infty}_n$ and so $D^{\infty}_n/i^{\infty}(D^{\infty}_{n-1})=\mathrm{Gr}_nL_{\infty}$. At the other end, the canonical map $L_{-\infty}\to D_n$ lifts to ${}^{\infty}D_n$ yielding an injection $L_{-\infty}/F_nL_{-\infty}\to {}^{\infty}D_n$. Therefore, for each n there is an injection $\mathrm{Gr}_nL_{-\infty}\to K_n$ where $K_n=\mathrm{Ker}({}^{\infty}D_{n-1}\to {}^{\infty}D_n)$. In general, the map $L_{-\infty}\to {}^{\infty}D_n$ need not be surjective (an element could be in the image of $i^{\infty r}$ for each finite r without being part of a consistent infinite sequence), although it is surjective in the special case when ${}^{\infty}D_s\to {}^{\infty}D_{s+1}$ is surjective for each s. In the latter case we get $\mathrm{Gr}\,L_{-\infty}\cong K$. As we will see in the next section, the exact sequence of Theorem 3 extends to the right (Theorem 8) giving $\varprojlim{}^1_r Z^r=0$ as a sufficient condition that ${}^{\infty}D_s\to {}^{\infty}D_{s+1}$ be surjective for each s, where $\varprojlim{}^1$ is described in that section and (Z^r) refers to the system of inclusions $\cdots\subset Z^{r+1}\subset Z^r\subset Z^{r-1}\subset\cdots$. Thus, $\varprojlim{}^1_r Z^r=0$ is a sufficient condition for $\mathrm{Gr}\,L_{-\infty}\cong K$.

Taking into account the short exact sequence $0\to D^{\infty}/i^{\infty}(D^{\infty})\to E^{\infty}\to K\to 0$ coming from Theorem 3, the preceding discussion yields two obvious candidates for a suitable $\mathcal{F}_L$: $\mathcal{F}_{L_{\infty}}$ or $\mathcal{F}_{L_{-\infty}}$. In theory there are other possibilities, but in practice one of these two cases usually occurs. We examine them individually and see what additional conditions are required for convergence.

Case I: *Conditions for convergence to $\mathcal{F}_{L_{\infty}}$* It is easily checked from the definitions that $\varinjlim_n D^{\infty}_n=\varinjlim_n D_n$ so $\mathcal{F}_{L_{\infty}}$ is always cocomplete. Therefore, besides $\mathrm{Gr}\,L_{\infty}\cong E^{\infty}$ (equivalently, $K=0$), it is required to verify that $\mathcal{F}_{L_{\infty}}$ is complete. As we will see in the next section, the completeness condition can be restated as $\cap D_n=0$ and $\varprojlim{}^1 D_n=0$. According to the preceding discussion, under the assumption that $L_{-\infty}=\cap D^n=0$, which we need anyway as part of the requirement that $\mathcal{F}_{L_{\infty}}$ be complete, $\varprojlim{}^1_r Z_rX=0$ is sufficient to show $K=0$.

Case II: *Conditions for convergence to $\mathcal{F}_{L_{-\infty}}$* Any inverse limit is complete in its canonical filtration, so $\mathcal{F}_{L_{-\infty}}$ is always complete and the issues are whether $\mathrm{Gr}L_{-\infty}\cong E^{\infty}$ and whether $\mathcal{F}_{L_{-\infty}}$ is cocomplete. $\mathcal{F}_{L_{-\infty}}$ is cocomplete if and only if every element of

$L_{-\infty}$ lies in $\mathrm{Ker}(L_{-\infty} \to D_n)$ for some n, for which a sufficient condition is that $L_\infty = 0$ or equivalently $E^\infty \cong K$. Therefore, if the reason for the isomorphism $\mathrm{Gr}\, L_{-\infty} \cong E^\infty$ is that the maps $E^\infty \twoheadrightarrow K$ and $\mathrm{Gr}\, L_{-\infty} \rightarrowtail K$ are isomorphisms, then the rest of the convergence conditions are automatic. In particular, to deduce convergence to $\mathcal{F}_{L_{-\infty}}$ it suffices to know that $L_\infty = 0$ and $\varprojlim_r{}^1 Z_r = 0$.

Derived Functors

The left and right derived functors L_nT, R^nT of a functor T provide a measure of the amount by which the functor deviates from preserving exactness.

The category $\mathcal{I}nv$ of inverse systems indexed over $\mathbb{Z}$ (i.e., the category whose objects are diagrams of abelian groups $\cdots \to A_{n-1} \to A_n \to A_{n+1} \to \cdots$) forms an abelian category in which a sequence of morphisms $A' \to A \to A''$ is exact if and only if the sequence $A_n{}' \to A_n \to A_n{}''$ of abelian groups is exact for each n. The functor of interest to us is $\varprojlim : \mathcal{I}nv \to \mathcal{AB}$ where $\mathcal{AB}$ denotes the category of abelian groups.

Let $T: \underline{\underline{A}} \to \underline{\underline{B}}$ be an additive functor between abelian categories. Suppose that X in Obj $\underline{\underline{A}}$ has an injective resolution I_X. The definition of additive functor implies that T takes zero morphisms to zero morphisms, so TI_X forms a cochain complex in $\underline{\underline{B}}$. The *right derived functors* of T are defined by $(R^nT)(X) := H^n(TI_X)$. The result is independent of the choice of injective resolution (assuming one exists) and satisfies:

1. If T is "left exact" (meaning that T preserves monomorphisms), then $R^0T(X) = T(X)$;
2. If T preserves exactness, then $(R^nT)(X) = 0$ for $n > 0$.

Theorem 7 *Let $0 \to X' \to X \to X'' \to 0$ be a short exact sequence in $\underline{\underline{A}}$. Suppose T is left exact and that all the objects have injective resolutions. Then there is a (long) exact sequence*

$$0 \to T(X') \to T(X) \to T(X'') \to (R^1T)(X') \to \cdots$$
$$\to (R^{n-1}T)(X'') \to (R^nT)(X') \to (R^nT)(X) \to$$
$$(R^nT)(X'') \to \cdots$$

Similarly, the *left derived functors* of T are defined by using projective resolutions and have similar properties with respect to the obvious duality.

The functor $\varprojlim_n$ is left exact and in the category $\mathcal{I}nv$ every object has an injective resolution. Therefore $\varprojlim_n{}^q$ is defined and $\varprojlim_n{}^0 X_n = \varprojlim_n X_n$, where $\varprojlim_n{}^q$ denotes the derived functor $R^q(\varprojlim_n)$. It turns out that $\varprojlim_n{}^q$ is 0 for $q > 1$, but we are particularly interested in $\varprojlim_n{}^1$.

Let (X_n) be an inverse system with structure maps $i_{n-1}: X_{n-1} \to X_n$. An explicit construction for $\varprojlim^1 X_n$ is as follows. Define $\phi: \prod_n X_n \to \prod_n X_n$ by letting $\phi(x_n)$ be the sequence whose nth component is $(x_n - i_{n-1}x_{n-1})$. Then $\varprojlim_n{}^1 X_n \cong \mathrm{Coker}\, \phi$. Observe that $\mathrm{Ker}\, \phi \cong \varprojlim X_n$ according to the explicit formula for $\varprojlim X_n$ given earlier.

Recall that we defined ${}^\infty D = \cap_r \mathrm{Im}\, i^{\circ r} \cong \varprojlim_r D^r$. The exact sequence of Theorem 3 can be extended to give:

Theorem 8 *There is an exact sequence*

$$0 \to D^\infty \xrightarrow{i} D^\infty \xrightarrow{j} E^\infty \xrightarrow{k} {}^\infty D \xrightarrow{i} {}^\infty D$$
$$\xrightarrow{j} \varprojlim_r{}^1 Z^r \xrightarrow{k} \varprojlim_r{}^1 D^r \xrightarrow{i} \varprojlim_r{}^1 D^r \to 0$$

It is clear from the explicit construction that if the system (X_n) stabilizes with $X_n = G$ for all sufficiently small n, then $\varprojlim_n X = G$ and $\varprojlim_n{}^1 X = 0$. If the spectral sequence collapses at any stage then the system (Z^r) stabilizes at that point, and so for a spectral sequence which collapses, the condition $\varprojlim_r{}^1 Z^r = 0$, which arose in the discussion of convergence in the previous section, is automatic.

Let $\mathcal{F}_X$ be a filtered abelian group. Applying Theorem 7 to the short exact sequence $0 \to F_nX \to X \to X/F_nX \to 0$ of inverse systems gives an exact sequence

$$0 \to \varprojlim_n F_nX \to \varprojlim_n X \to \varprojlim_n X/F_n$$
$$\to \varprojlim_n{}^1 F_nX \to \varprojlim_n{}^1 X$$

Since $\varprojlim_n X = X$ and $\varprojlim_n{}^1 X = 0$, we get

Theorem 9 *$\mathcal{F}_X$ is complete if and only if $\varprojlim_n F_nX = 0$ and $\varprojlim_n{}^1 F_nX = 0$.*

When working with $\varprojlim_n{}^1$ the following sufficient condition for its vanishing, known as the Mittag–Leffler condition, is often useful.

Theorem 10 *Suppose A is an inverse system in which for each n there exists $k(n) \le n$ such that $\mathrm{Im}(A_i \to A_n)$ equals $\mathrm{Im}(A_{k(n)} \to A_n)$ for all $i \le k(n)$. Then $\varprojlim_n{}^1 A = 0$.*

Of course, this will not be (directly) useful in establishing $\varprojlim_n{}^1 F_nX = 0$ since the structure maps in that system are all monomorphisms.

Some Examples of Standard Spectral Sequences and Their Use

To this point we have considered the general theory of spectral sequences. The properties of the spectral sequences arising in many specific situations have

been well studied. Usually the spectral sequence would be defined either directly, through an exact couple, or by giving some filtration on a chain complex. This defines the E^1-term. Typically, a theorem would then be proved giving some formula for the resulting E^2-term. In many cases, conditions under which the spectral sequence converges may also be well known.

In this section, we shall take a brief look at the Serre spectral sequence, Atiyah–Hirzebruch spectral sequence, spectral sequence of a double complex, Grothendieck spectral sequence, change of ring spectral sequence, and Eilenberg–Moore spectral sequence, and carry out a few sample calculations.

Serre Spectral Sequence

Let $F \to X \xrightarrow{\pi} B$ be a fiber bundle (or more generally a fibration) in which the base B is a CW-complex. Define a filtration on the total space by $F_nX := \pi^{-1}B^{(n)}$. This yields a filtration on $H_*(X)$ by setting $F_nH_*(X) := \mathrm{Im}(H_*(F_nX) \to H_*(X))$. The spectral sequence coming from the exact couple in which $D^1_{p,q} := H_{p+q}(F_pX)$ and $E^1_{p,q} := H_{p+q}(F_pX, F_{p-1}X)$ is called the "Serre spectral sequence" of the fibration. Theorems from topology guarantee that this filtration is cocomplete and that $E^1_{p,q} = 0$ if either $p < 0$ or $q < 0$. Therefore, the Serre spectral sequence is always a first quadrant spectral sequence converging to $H_*(X)$.

Theorem 11 (Serre). *In the Serre spectral sequence of the fibration $F \to E \to B$ there is an isomorphism $E^2_{p,q} \cong H_p(B; {}^tH_q(F))$.*

Here ${}^tH_*(F)$ denotes a "twisted" or "local" coefficient system in which the differential is modified to take into account the action, coming from the fibration, of the fundamental groupoid of the base B on the fiber F. In the special case where B is simply connected and $\mathrm{Tor}(H_*(B), H_*(F)) = 0$, the "universal coefficient theorem" says that the E^2-term reduces to $E^2_{p,q} \cong H_p(B) \otimes H_q(F)$.

The Serre spectral sequence for cohomology, $E_2^{p,q} \cong H^p(B; {}^tH^q(F)) \Rightarrow H^{p+q}(X)$, has the advantage that it is a spectral sequence of algebras which greatly simplifies calculation of the differentials d_r which are restricted by the requirement that they satisfy the Leibniz rule with respect to the cup product on $H^*(B)$ and $H^*(F)$, and which also allows the computation of the cup product on $H^*(X)$. Since it is a first quadrant spectral sequence, convergence is not an issue.

Frequently in applications of the Serre spectral sequence, instead of using the spectral sequence to calculate $H_*(X)$ from knowledge of $H_*(F)$ and $H_*(B)$ it is instead $H_*(X)$ and one of the other two homologies which is known, and one is working backwards from the spectral sequence to find the homology of the third space.

Example 3 The universal S^1-bundle is the bundle $S^1 \to S^\infty \to \mathbb{C}P^\infty$ where S^∞ is contractible. We will calculate $H^*(\mathbb{C}P^\infty)$ from the Serre spectral sequence of this bundle, taking $H^*(S^1)$ and $H^*(S^\infty)$ as known. We also take as known that $\mathbb{C}P^\infty$ is path connected, so $H^0(\mathbb{C}P^\infty) \cong \mathbb{Z}$.

$$E_2^{p,q} \cong H^p(\mathbb{C}P^\infty) \otimes H^q(S^1) \cong \begin{cases} H^p(\mathbb{C}P^\infty) & \text{if } q = 0 \text{ or } 1 \\ 0 & \text{otherwise} \end{cases}$$

E_∞-terms on the diagonal $p + q = n$ form a composition series for $H^n(S^\infty)$ which is zero for $n \neq 0$. Therefore $E_\infty^{p,q} = 0$ unless $p = 0$ and $q = 0$, with $E_\infty^{0,0} \cong \mathbb{Z}$. Because all nonzero terms lie in the first quadrant, the bidegrees of the differentials show that $d_r(E_2^{1,0}) = 0$ for all $r \geq 2$, so $0 = E_\infty^{1,0} = E_2^{1,0} = H^1(\mathbb{C}P^\infty)$. Since $E_2^{1,q} \cong E_2^{1,0} \otimes E_2^{0,q}$, it follows that $E_2^{1,q} = 0$ for all q. Taking into the account the known zero terms, the bidegrees of the differentials show that $E_3^{0,1} \cong \mathrm{Ker}(d_2 : E_2^{0,1} \to E_2^{2,0})$ and $E_\infty^{0,1} = E_3^{0,1}$. Similarly, $E_\infty^{2,0} = E_3^{2,0} \cong \mathrm{Coker}(d_2 : E_2^{0,1} \to E_2^{2,0})$. Therefore, the vanishing of these E_∞-terms shows that $d_2 : E_2^{0,1} \cong E_2^{2,0}$ and in particular $H^2(\mathbb{C}P^\infty) \cong E_2^{0,1} = H^1(S^1) \cong \mathbb{Z}$. It follows that $E_2^{2,q} \cong \mathbb{Z} \otimes E_2^{0,q} \cong E_2^{0,q}$ for all q. With the aid of the fact that we showed $E_2^{1,1} = 0$, we can repeat the argument used to show $E_2^{1,q} = 0$ for all q to conclude that $E_2^{3,q} = 0$ for all q. Repeating the procedure, we inductively find that $E_2^{p,q} \cong E_2^{p-2,q}$ for all $p > 0$ and all q and in particular

$$H^n(\mathbb{C}P^\infty) \cong \begin{cases} \mathbb{Z} & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd} \end{cases}$$

The cup products in $H^*(\mathbb{C}P^\infty)$ can also be determined by taking advantage of the fact that the spectral sequence is a spectral sequence of algebras. Let $a \in E_2^{2,0} \cong \mathbb{Z}$ be a generator and set $x := d_2 a$. By the preceding calculation, d_2 is an isomorphism so x is a generator of $H^2(\mathbb{C}P^\infty)$. Therefore, $x \otimes a$ is a generator of $E_2^{2,2}$ and the isomorphism d_2 gives that $d_2(x \otimes a)$ is a generator of $H^4(\mathbb{C}P^\infty)$. However, $d_2(x \otimes a) = d_2(x \otimes 1)(1 \otimes a) = 0 \otimes 1 + (-1)^2(x \otimes 1)d_2 a = x^2 \otimes 1$ and thus, x^2 is a generator of $H^4(\mathbb{C}P^\infty)$. Inductively, it follows that x^n is a generator of $H^{2n}(\mathbb{C}P^\infty)$ for all n and so $H^*(\mathbb{C}P^\infty) \cong \mathbb{Z}[x]$.

When working backwards from the Serre or other first quadrant spectral sequences in which $E^2_{p,q} \cong E^2_{p,0} \otimes E^2_{0,q}$ the following analog of the comparison theorem (Theorem 5) is often useful.

Theorem 12 (Zeeman comparison theorem). *Let E and E' be first quadrant spectral sequences such that $E^2_{p,q} = E^2_{p,0} \otimes E^2_{0,q}$ and $E'^2_{p,q} = E'^2_{p,0} \otimes E'^2_{0,q}$. Let $f : E \to E'$ be a homomorphism of spectral sequences such that $f^2_{p,q} = f^2_{p,0} \otimes f^2_{0,q}$. Suppose that $f^\infty_{p,q} : E^\infty_{p,q} \to E'^\infty_{p,q}$ is an isomorphism for all p and q. Then the following are equivalent*:

(i) $f^2_{p,0} : E^2_{p,0} \to E'^2_{p,0}$ *is an isomorphism for* $p \leq n-1$;
(ii) $f^2_{0,q} : E^2_{0,q} \to E'^2_{0,q}$ *is an isomorphism for* $q \leq n$.

There is a version of the Serre spectral sequence for generalized homology theories coming from the exact couple obtained by applying the generalized homology theory to the Serre filtration of X.

Theorem 13 (Serre spectral sequence for generalized homology). *Let $F \to X \to B$ be a fibration and let Y be an (unreduced) homology theory satisfying the Milnor wedge axiom. Then there is a (right half-plane) spectral sequence with $E^2_{p,q} \cong H_p(B; {}^tY_q(F))$ converging to $Y_{p+q}(X)$.*

Cocompleteness of the filtration follows from the properties of generalized homology theories satisfying the wedge axiom (Milnor 1962), and the rest of the convergence conditions are trivial since the filtration is 0 in negative degrees. Here, unlike the Serre spectral sequence for ordinary homology, the existence of terms in the fourth quadrant opens the possibility for composition series of infinite length, although in the case where B is a finite-dimensional complex all the nonzero terms of the spectral sequence will live in the strip between $p = 0$ and $p = \dim B$ and so the filtrations will be finite.

The special case of the fibration $* \to X \to X$ yields what is known as the "Atiyah–Hirzebruch spectral sequence".

Theorem 14 (Atiyah–Hirzebruch spectral sequence). *Let X be a* CW*-complex and let Y be an (unreduced) homology theory satisfying the Milnor wedge axiom. Then there is a (right half-plane) spectral sequence with $E^2_{p,q} \cong H_p(X; Y_q(*))$ converging to $Y_{p+q}(X)$.*

In the cohomology Serre spectral sequence for generalized cohomology (including the cohomology Atiyah–Hirzebruch spectral sequence), convergence of the spectral sequence to $Y^*(X)$ is not guaranteed. Convergence to $\varprojlim_n Y^*(F_nX)$, should that occur, would be of the type discussed in case II in the section "Convergence of graded spectral sequences". Since $X_n = \emptyset$ for $n < 0$, the system defining L_∞ stabilizes to 0. Therefore, $L_\infty = 0$ and, by the discussion in that section, $\varprojlim_r{}^1 Z_r X = 0$ becomes a sufficient condition for convergence to $\varprojlim_n Y^*(F_nX)$. However since the real object of study is usually $Y^*(X)$, the spectral sequence is most useful when one is also able to show $\varprojlim^1 Y^*(F_nX) = 0$ in which case the Milnor exact sequence (Milnor 1962)

$$0 \to \varprojlim_n{}^1 Y^*(F_nX) \to Y^*(X) \to \varprojlim_n Y^*(F_nX) \to 0$$

gives $Y^*(X) \cong \varprojlim_n Y^*(F_nX)$.

If $Y^*(\,)$ has cup products then the spectral sequence has the extra structure of a spectral sequence of $Y^*(*)$-algebras. In the case where B is finite dimensional, all convergence problems disappear since the spectral sequence lives in a strip and the filtrations are finite.

Example 4 Let $K^*(\,)$ be complex K-theory. Since $K^*(*) \cong \mathbb{Z}[z, z^{-1}]$ with $|z| = 2$, in the Atiyah–Hirzebruch spectral sequence for $K^*(\mathbb{C}P^n)$ we have

$$E_2^{p,q} = \begin{cases} \mathbb{Z} & \text{if } q \text{ is even and } p \text{ is even with } 0 \leq p \leq 2n \\ 0 & \text{otherwise} \end{cases}$$

Because $\mathbb{C}P^n$ is a finite complex, the spectral sequence converges to $K^*(\mathbb{C}P^n)$. Since all the nonzero terms have even total degree and all the differentials have total degree $+1$, the spectral sequence collapses at E_2 and we conclude that $K^q(\mathbb{C}P^n) = 0$ if q is odd and that it has a composition series consisting of $(n+1)$ copies of $\mathbb{Z}$ when q is even. Since $\mathbb{Z}$ is a free abelian group, this uniquely identifies the group structure of $K^{\text{even}}(\mathbb{C}P^n)$ as $\mathbb{Z}^{n+1}$. To find the ring structure we can make use of the fact that this is a spectral sequence of $K^*(*)$-algebras. The result is $K^*(\mathbb{C}P^n) \cong K^*(*)[x]/(x^{n+1})$, where $|x| = 2$.

In the Atiyah–Hirzebruch spectral sequence for $K^*(\mathbb{C}P^\infty)$ again all the terms have even total degree so the spectral sequence collapses at E_2. We noted earlier that collapse of the spectral sequence implies that $\varprojlim_r{}^1 Z_r X = 0$ and so the spectral sequence converges to $\varprojlim_n K^*(\mathbb{C}P^n)$, where we used $F_{2n}\mathbb{C}P^\infty = \mathbb{C}P^n$. Since our preceding calculation shows that $K^*(\mathbb{C}P^n) \to K^*(\mathbb{C}P^{n-1})$ is onto, Mittag–Leffler (Theorem 10) implies that $\varprojlim^1 K^*(\mathbb{C}P^n) = 0$. Therefore, the spectral sequence converges to $K^*(\mathbb{C}P^\infty)$ and we find that $K^*(\mathbb{C}P^\infty) \cong \varprojlim_n K^*(\mathbb{C}P^n)$, which is isomorphic to the power series ring $K^*(*)[[x]]$, where $|x| = 2$.

In topology one might be interested in the Atiyah–Hirzebruch spectral sequence in the case where X is a spectrum rather than a space (a spectrum being a generalization in which cells in negative degrees are allowed including the possibility that the dimensions

of the cells are not bounded below). In such cases, the spectral sequence is no longer constrained to lie in the right half-plane and convergence criteria are not well understood for either the homology or cohomology version.

Spectral Sequence of a Double Complex

A double complex is a chain complex of chain complexes. That is, it is a bigraded abelian group $C_{p,q}$ together with two differentials $d' : C_{p,q} \to C_{p-1,q}$ and $d'' : C_{p,q} \to C_{p,q-1}$ satisfying $d' \circ d' = 0, d'' \circ d'' = 0$, and $d'd'' = d''d'$. Given a double complex C its total complex $\mathrm{Tot}\, C$ is defined by $(\mathrm{Tot}\, C)_n := \bigoplus_{p+q=n} C_{p,q}$ with differential defined by $d|_{C_{p,q}} := d' + (-1)^p d'' : C_{p,q} \to C_{p-1,q} \oplus C_{p,q-1} \subset \mathrm{Tot}_{n-1} C$.

There are two natural filtrations, $\mathcal{F}'_{\mathrm{Tot}\, C}$ and $\mathcal{F}''_{\mathrm{Tot}\, C}$, on $\mathrm{Tot}\, C$ given by

$$\left(F'_p(\mathrm{Tot}C)\right)_n = \bigoplus_{\substack{s+t=n \\ s \le p}} C_{s,t}$$

$$\left(F''_p(\mathrm{Tot}C)\right)_n = \bigoplus_{\substack{s+t=n \\ t \le p}} C_{s,t}$$

yielding two spectral sequences abutting to $H_*(\mathrm{Tot}C)$. In the first ${E'}^2_{p,q} = H_p(H_q(C_{*,*}))$ and in the other ${E''}^2_{p,q} = H_q(H_p(C_{*,*}))$. Convergence of these spectral sequences is not guaranteed, although the first will always converge if there exists N such that $C_{p,q} = 0$ for $p < N$ and the second will converge if there exists N such that $C_{p,q} = 0$ for $q < N$. From the double complex C one could instead form the product total complex $(\mathrm{Tot}^{\pi} C)_n := \prod_{p+q=n} C_{p,q}$ and proceed in a similar manner to construct the same spectral sequences with different convergence problems. In the important special case of a first quadrant double complex both spectral sequences converge and information is often obtained by playing one off against the other.

Example 5 Let M and N be R-modules. Let $\mathrm{Tor}^{'R}_*(M,N)$ and $\mathrm{Tor}^{''R}_*(M,N)$ be the derived functors of $(\_) \otimes N$ and $M \otimes (\_)$, respectively. Let P_* and Q_* be projective resolutions of M and N respectively. Define a first quadrant double complex by $C_{p,q} := P_p \otimes Q_q$. Since P_p is projective,

$$H_q(C_{p,*}) = P_p \otimes H_q(C_{p,*}) = \begin{cases} 0 & \text{if } q \neq 0 \\ N & \text{if } q = 0 \end{cases}$$

and so in the first spectral sequence of the double complex,

$${E'}^2_{p,q} = \begin{cases} 0 & \text{if } q \neq 0 \\ \mathrm{Tor}^{'R}_p(M,N) & \text{if } q = 0 \end{cases}$$

Therefore, the spectral sequence collapses to give $H_n(\mathrm{Tot}\, C) \cong \mathrm{Tor}^{'R}_n(M,N)$. Similarly, the second spectral sequence shows that $H_n(\mathrm{Tot}\, C) \cong \mathrm{Tor}^{''R}_n(M,N)$. Thus, $\mathrm{Tor}^R_*(M,N)$ can be computed equally well from a projective resolution of either variable.

The technique of using a double complex in which one spectral sequence yields the homology the total complex to which both converge can be used to prove.

Theorem 15 (Grothendieck spectral sequence). *Let $\underline{\underline{C}} \xrightarrow{F} \underline{\underline{B}} \xrightarrow{G} \underline{\underline{A}}$ be a composition of additive functors, where $\underline{\underline{C}}$, $\underline{\underline{B}}$, and $\underline{\underline{A}}$ are abelian categories. Assume that all objects in $\underline{\underline{C}}$ and $\underline{\underline{B}}$ have projective resolutions. Suppose that F takes projectives to projectives. Then for all objects C of $\underline{\underline{C}}$ there exists a (first quadrant) spectral sequence with $E^2_{p,q} = (L_pG)((L_qF)(C))$ converging to $(L_{p+q}(GF))(C)$.*

Naturally, there is a corresponding version for right derived functors.

An application of the Grothendieck spectral sequence is the following "change of rings spectral sequence." Let $f : R \to S$ be a ring homomorphism, let M be a right S-module and let N be a left R-module. Let $F(A) = S \otimes_R A$ and $G(B) = M \otimes_S B$, and note that $GF(A) = M \otimes_R A$. Applying the Grothendieck spectral sequence to the composition (left R-modules $\xrightarrow{F}$ left S-modules $\xrightarrow{G}$ abelian groups) yields a convergent spectral sequence $E^2_{p,q} \cong \mathrm{Tor}^S_p(M, \mathrm{Tor}^R_q(S,N)) \Rightarrow \mathrm{Tor}^R_{p+q}(M,N)$.

Eilenberg–Moore Spectral Sequence

For a topological group G, Milnor showed how to construct a universal G-bundle $G \to EG \to BG$ in which EG is the infinite join $G^{*\infty}$ with diagonal G-action. There is a natural filtration $F_nBG := G^{*(n+1)}/G$ on BG and therefore an induced filtration on the base of any principal G-bundle. This filtration yields a spectral sequence including as a special case a tool for calculating $H_*(BG)$ from knowledge of $H_*(G)$.

Theorem 16 *Let $G \to X \to B$ be a principal G-bundle and let $H_*(\,)$ denote homology with coefficients in a field. Then there is a first quadrant spectral sequence with $E^2_{p,q} = \mathrm{Tor}^{H_*(G)}_{pq}(H_*(X), H_*(*))$ converging to $H_{p+q}(BG)$.*

Here the group structure makes $H_*(G)$ into an algebra and $\mathrm{Tor}^A_{pq}(M,N)$ denotes degree q of the graded object formed as the pth-derived functor of the tensor product of the graded modules M and N over the graded ring A.

There is also a version (Eilenberg and Moore 1962) which, like the Serre spectral sequence, is suitable for computing $H^*(G)$ from $H^*(BG)$.

Theorem 17 *Let*

$$\begin{array}{ccc} W & \longrightarrow & Y \\ \downarrow & & \downarrow \pi \\ X & \xrightarrow{f} & B \end{array}$$

be a pullback square in which π *is a fibration and* X *and* B *are simply connected. Suppose that* $H^*(X), H^*(Y)$, *and* $H^*(B)$ *are flat* R*-modules of finite type, where* $H^*(\,)$ *denotes cohomology with coefficients in the Noetherian ring* R*. Then there is a* (*second quadrant*) *spectral sequence with* $E_2^{p,q} \cong \mathrm{Tor}_{pq}^{H^*(B)}(H^*(X), H^*(Y))$ *converging to* $H^{p+q}(W)$.

The cohomological version of the Eilenberg–Moore spectral sequence, stated above, contains the more familiar Tor for modules over an algebra. For the homological version, one must dualize these notions appropriately to define the cotensor product of comodules over a coalgebra, and its derived functors Cotor.

Provided the action of the fundamental group of B is sufficiently nice there are extensions of the Eilenberg–Moore spectral sequence to the case where B is not simply connected, although they do not always converge, and extensions to generalized (co)homology theories have also been studied.

See also: Cohomology Theories; Derived Categories; K-Theory; Spectral Theory for Linear Operators.

Further Reading

Adams JF (1974) *Stable Homotopy and Generalized Homology*. Chicago: Chicago University Press.

Atiyah M and Hirzebruch F (1969) Vector bundles and homogeneous spaces. *Proceedings of Symposia in Pure Mathematics* 3: 7–38.

Boardman JM (1999) Conditionally convergent spectral sequences. *Contemporary Mathematics* 239: 49–84.

Bousfield AK and Kan DM (1972) *Homotopy Limits, Completions and Localization*. Lecture Notes in Math, vol. 304. Berlin: Springer.

Cartan H and Eilenberg S (1956) *Homological Algebra*. Princeton: Princeton University Press.

Eilenberg S and Moore J (1962) Homology and fibrations I. Coalgebras cotensor product and its derived functors. *Commentarii Mathematici Helvetici* 40: 199–236.

Grothendieck A (1957) Sur quelques points d'algèbre homologique. *Tôhoku Mathematical Journal* 9: 119–221.

Hilton PJ and Stammbach U (1970) *A Course in Homological Algebra*. Graduate Texts in Mathematics, vol. 4. Berlin: Springer.

Koszul J-L (1947) Sur l'opérateurs de derivation dans un anneau. *Comptes Rendus de l'Académie des Sciences de Paris* 225: 217–219.

Leray J (1946) L'anneau d'une representation; Propriétés d'homologie de la projections d'un espace fibré sur sa base; Sur l'anneau d'homologie de l'espace homogène, quotient d'un groupe clos par un sous-groupe abélien, connexe, maximum. *C.R. Acad. Sci.* 222: 1366–1368; 1419–1422 and 223: 412–415.

Massey W (1952) Exact couples in algebraic topology I, II. *Annals of Mathematics* 56: 363–396.

Massey W (1953) Exact couples in algebraic topology III, IV, V. *Annals of Mathematics* 57: 248–286.

McCleary J (2001) *User's Guide to Spectral Sequences*. Cambridge Studies in Advanced Mathematics, vol. 58. Cambridge: Cambridge University Press.

Milnor J (1956) Constructions of universal bundles, II. *Annals of Mathematics* 63: 430–436.

Milnor J (1962) On axiomatic homology theory. *Pacific Journal of Mathematics* 12: 337–341.

Selick P (1997) *Introduction to Homotopy Theory*. Fields Institute Monographs, vol. 9. Providence, RI: American Mathematical Society.

Serre J-P (1951) Homologie singulière des espaces fibrés. *Annals of Mathematics* 54: 425–505.

Smith L (1969) *Lectures on the Eilenberg–Moore Spectral Sequence*. Lecture Notes in Math, vol. 134. Berlin: Springer.

Zeeman EC (1958) A proof of the comparison theorem for spectral sequences. *Proceedings of the Cambridge Philosophical Society* 54: 57–62.

Spectral Theory of Linear Operators

M Schechter, University of California at Irvine, Irvine, CA, USA

Introduction

We begin with the study of linear operators on normed vector spaces (for definitions, see, e.g., Schechter (2002) or the appendix at the end of this article). If the scalars are complex numbers, we shall call the space complex. If the scalars are real, we shall call it real.

Let X, Y be normed vector spaces. A mapping A which assigns to each element x of a set $D(A) \subset X$ a unique element $y \in Y$ is called an operator (or transformation). The set $D(A)$ on which A acts is called the domain of A. The operator A is called linear if

1. $D(A)$ is a subspace of X, and
2. $A(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 A x_1 + \alpha_2 A x_2$

for all scalars α_1, α_2 and all elements $x_1, x_2 \in D(A)$.

To begin, we shall only consider operators A with $D(A) = X$.

An operator A is called bounded if there is a constant M such that

$$\|Ax\| \leq M\|x\|, \quad x \in X \tag{1}$$

The norm of such an operator is defined by

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \tag{2}$$

It is the smallest M which works in [1]. An operator A is called continuous at a point $x \in X$ if $x_n \to x$ in X implies $Ax_n \to Ax$ in Y. A bounded linear operator is continuous at each point. For if $x_n \to x$ in X, then

$$\|Ax_n - Ax\| \leq \|A\| \cdot \|x_n - x\| \to 0$$

We also have

Theorem 1 *If a linear operator A is continuous at one point $x_0 \in X$, then it is bounded, and hence continuous at every point.*

We let $B(X, Y)$ be the set of bounded linear operators from X to Y. Under the norm [2], one easily checks that $B(X, Y)$ is a normed vector space.

The Adjoint Operator

An assignment F of a number to each element x of a vector space is called a functional and denoted by $F(x)$. If it satisfies

$$F(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 F(x_1) + \alpha_2 F(x_2) \tag{3}$$

for α_1, α_2 scalars, it is called linear. It is called bounded if

$$|F(x)| \leq M\|x\|, \quad x \in X \tag{4}$$

If F is a bounded linear functional on a normed vector space X, the norm of F is defined by

$$\|F\| = \sup_{x \in X,\, x \neq 0} \frac{|F(x)|}{\|x\|} \tag{5}$$

It is equal to the smallest number M satisfying [4].

For any normed vector space X, let X' denote the set of bounded linear functionals on X. If $f, g \in X'$, we say that $f = g$ if

$$f(x) = g(x) \text{ for all } x \in X$$

The "zero" functional is the one assigning zero to all $x \in X$. We define $h = f + g$ by

$$h(x) = f(x) + g(x), \quad x \in X$$

and $g = \alpha f$ by

$$g(x) = \alpha f(x), \quad x \in X$$

Under these definitions, X' becomes a vector space. The expression

$$\|f\| = \sup_{x \neq 0} \frac{|f(x)|}{\|x\|}, \quad f \in X' \tag{6}$$

is easily seen to be a norm. Thus, X' is a normed vector space. It is therefore natural to ask when X' will be complete. A rather surprising answer is given by

Theorem 2 *X' is a Banach space whether or not X is.*

(For the definition of a Banach space, see, e.g., Schechter (2002) or the appendix at the end of this article.)

Suppose X, Y are normed vector spaces and $A \in B(X, Y)$. For each $y' \in Y'$, the expression $y'(Ax)$ assigns a scalar to each $x \in X$. Thus, it is a functional $F(x)$. Clearly F is linear. It is also bounded since

$$|F(x)| = |y'(Ax)| \leq \|y'\| \cdot \|Ax\| \leq \|y'\| \cdot \|A\| \cdot \|x\|$$

Thus, there is an $x' \in X'$ such that

$$y'(Ax) = x'(x), \quad x \in X \tag{7}$$

This functional x' is unique. Thus, to each $y' \in Y'$ we have assigned a unique $x' \in X'$. We designate this assignment by A' and note that it is a linear operator from Y' to X'. Thus, [7] can be written in the form

$$y'(Ax) = A'y'(x) \tag{8}$$

The operator A' is called the adjoint (or conjugate) of A. We note

Theorem 3 $A' \in B(Y', X')$, *and* $\|A'\| = \|A\|$.

The adjoint has the following easily verified properties:

$$(A + B)' = A' + B' \tag{9}$$

$$(\alpha A)' = \alpha A' \tag{10}$$

$$(AB)' = B'A' \tag{11}$$

Why should we consider adjoints? One reason is as follows. Many problems in mathematics and its applications can be put in the form: given normed vector spaces X, Y and an operator $A \in B(X, Y)$, one wishes to solve

$$Ax = y \tag{12}$$

The set of all y for which one can solve [12] is called the "range" of A and is denoted by $R(A)$. The set of all x for which $Ax = 0$ is called the "null space" of A and is denoted by $N(A)$. Since A is linear, it is easily checked that $N(A)$ and $R(A)$ are subspaces of X and Y,

respectively (for definitions, see, e.g., Schechter (2002) or the appendix at the end of this article). The dimension of $N(A)$ is denoted by $\alpha(A)$.

If $y \in R(A)$, there is an $x \in X$ satisfying [12]. For any $y' \in Y'$ we have

$$y'(Ax) = y'(y)$$

Taking adjoints we get

$$A'y'(x) = y'(y)$$

If $y' \in N(A')$, this gives $y'(y) = 0$. Thus, a necessary condition that $y \in R(A)$ is that $y'(y) = 0$ for all $y' \in N(A')$. Obviously, it would be of great interest to know when this condition is also sufficient.

The Spectrum and Resolvent Sets

From this point henceforth we shall assume that $X = Y$. We can then speak of the identity operator I defined by

$$Ix = x, \quad x \in X$$

For a scalar λ, the operator λI is given by

$$\lambda Ix = \lambda x, \quad x \in X$$

We shall denote the operator λI by λ.

We shall denote the space $B(X, X)$ by $B(X)$. For any operator $A \in B(X)$, a scalar λ for which $\alpha(A - \lambda) \neq 0$ is called an eigenvalue of A. Any element $x \neq 0$ of X such that $(A - \lambda)x = 0$ is called an eigenvector (or eigenelement). The points λ for which $(A - \lambda)$ has a bounded inverse in $B(X)$ comprise the resolvent set $\rho(A)$ of A (for definitions, see, e.g., Schechter (2002) or the appendix at the end of this article). If X is a Banach space, it is the set of those λ such that $\alpha(A - \lambda) = 0$ and $R(A - \lambda) = X$. The spectrum $\sigma(A)$ of A consists of all scalars not in $\rho(A)$. The set of eigenvalues of A is sometimes called the point spectrum of A and is denoted by $P\sigma(A)$.

We note that

Theorem 4 *For A in $B(X)$, $\sigma(A') = \sigma(A)$.*

We are now going to examine the sets $\rho(A)$ and $\sigma(A)$ for arbitrary $A \in B(X)$.

Theorem 5 *$\rho(A)$ is an open set and hence $\sigma(A)$ is a closed set.*

Does every operator $A \in B(X)$ have points in its resolvent set? Yes. In fact, we have

Theorem 6 *For A in $B(X)$, set*

$$r_\sigma(A) = \inf_n \|A^n\|^{1/n} \qquad [13]$$

Then $\rho(A)$ contains all scalars λ such that $|\lambda| > r_\sigma(A)$.

Let $p(t)$ be a polynomial of the form

$$p(t) = \sum_0^n a_k t^k$$

Then for any operator $A \in B(X)$, we define the operator

$$p(A) = \sum_0^n a_k A^k$$

where we take $A^0 = I$. We have

Theorem 7 *If $\lambda \in \sigma(A)$, then $p(\lambda) \in \sigma(p(A))$ for any polynomial $p(t)$.*

Proof Since λ is a root of $p(t) - p(\lambda)$, we have

$$p(t) - p(\lambda) = (t - \lambda)q(t)$$

where $q(t)$ is a polynomial with real coefficients. Hence,

$$p(A) - p(\lambda) = (A - \lambda)q(A) = q(A)(A - \lambda) \qquad [14]$$

Now, if $p(\lambda)$ is in $\rho(p(A))$, then [14] shows that $\alpha(A - \lambda) = 0$ and $R(A - \lambda) = X$. This means that $\lambda \in \rho(A)$, and the theorem is proved. □

A symbolic way of writing Theorem 7 is

$$p(\sigma(A)) \subset \sigma(p(A)) \qquad [15]$$

Note that, in general, there may be points in $\sigma(p(A))$ which may not be of the form $p(\lambda)$ for some $\lambda \in \sigma(A)$. As an example, consider the operator on $\mathbb{R}^2$ given by

$$A(\alpha_1, \alpha_2) = (-\alpha_2, \alpha_1)$$

A has no spectrum; $A - \lambda$ is invertible for all real λ. However, A^2 has -1 as an eigenvalue. What is the reason for this? It is simply that our scalars are real. Consequently, imaginary numbers cannot be considered as eigenvalues. We shall see later that in order to obtain a more complete theory, we shall have to consider complex Banach spaces. Another question is whether every operator $A \in B(X)$ has points in its spectrum. For complex Banach spaces, the answer is yes.

The Spectral Mapping Theorem

Suppose we want to solve an equation of the form

$$p(A)x = y, \quad x, \ y \in X \qquad [16]$$

where $p(t)$ is a polynomial and $A \in B(X)$. If 0 is not in the spectrum of $p(A)$, then $p(A)$ has an inverse in $B(X)$ and, hence, [16] can be solved for all $y \in X$. So a natural question to ask is: what is the spectrum of $p(A)$? By Theorem 7 we see that it contains $p(\sigma(A))$,

but by the remark at the end of the preceding section it can contain other points. If it were true that

$$p(\sigma(A)) = \sigma(p(A)) \qquad [17]$$

then we could say that [16] can be solved uniquely for all $y \in X$ if and only if $p(\lambda) \neq 0$ for all $\lambda \in \sigma(A)$.

For a complex Banach space we have

Theorem 8 *If X is a complex Banach space, then $\mu \in \sigma(p(A))$ if and only if $\mu = p(\lambda)$ for some $\lambda \in \sigma(A)$, that is, if [17] holds.*

Proof We have proved it in one direction already (Theorem 7). To prove it in the other, let $\gamma_1, \ldots, \gamma_n$ be the (complex) roots of $p(t) - \mu$. For a complex Banach space they are all scalars. Thus,

$$p(A) - \mu = c(A - \gamma_1) \cdots (A - \gamma_n), \quad c \neq 0$$

Now suppose that all of the γ_j are in $\rho(A)$. Then each $A - \gamma_j$ has an inverse in $B(X)$. Hence, the same is true for $p(A) - \mu$. In other words, $\mu \in \rho(p(A))$. Thus, if $\mu \in \sigma(p(A))$, then at least one of the γ_j must be in $\sigma(A)$, say γ_k. Hence, $\mu = p(\gamma_k)$, where $\gamma_k \in \sigma(A)$. This completes the proof. □

Theorem 8 is called the "spectral mapping theorem" for polynomials. As mentioned before, it has the useful consequence:

Corollary 1 *If X is a complex Banach space, then eqn [16] has a unique solution for every y in X if and only if $p(\lambda) \neq 0$ for all $\lambda \in \sigma(A)$.*

Operational Calculus

Other things can be done in a complex Banach space that cannot be done in a real Banach space. For instance, we can get a formula for $p(A)^{-1}$ when it exists. To obtain this formula, we first note

Theorem 9 *If X is a complex Banach space, then $(z - A)^{-1}$ is a complex analytic function of z for $z \in \rho(A)$.*

By this, we mean that in a neighborhood of each $z_0 \in \rho(A)$, the operator $(z - A)^{-1}$ can be expanded in a "Taylor series," which converges in norm to $(z - A)^{-1}$, just like analytic functions of a complex variable.

Now, by Theorem 6, $\rho(A)$ contains the set $|z| > \|A\|$. We can expand $(z - A)^{-1}$ in powers of z^{-1} on this set. In fact, we have

Lemma 1 *If $|z| > \limsup \|A^n\|^{1/n}$, then*

$$(z - A)^{-1} = \sum_1^\infty z^{-n} A^{n-1} \qquad [18]$$

where the convergence is in the norm of $B(X)$.

Let C be any circle with center at the origin and radius greater than, say, $\|A\|$. Then, by Lemma 1,

$$\oint_C z^n (z - A)^{-1} dz = \sum_{k=1}^\infty A^{k-1} \oint_C z^{n-k} dz = 2\pi i A^n \qquad [19]$$

or

$$A^n = \frac{1}{2\pi i} \oint_C z^n (z - A)^{-1} dz \qquad [20]$$

where the line integral is taken in the right direction.

Note that the line integrals are defined in the same way as is done in the theory of functions of a complex variable. The existence of the integrals and their independence of path (so long as the integrands remain analytic) are proved in the same way. Since $(z - A)^{-1}$ is analytic on $\rho(A)$, we have

Theorem 10 *Let C be any closed curve containing $\sigma(A)$ in its interior. Then [20] holds.*

As a direct consequence of this, we have

Theorem 11 *$r_\sigma(A) = \max_{\lambda \in \sigma(A)} |\lambda|$ and $\|A^n\|^{1/n} \to r_\sigma(A)$ as $n \to \infty$.*

We can now put Lemma 1 in the following form:

Theorem 12 *If $|z| > r_\sigma(A)$, then [18] holds with convergence in $B(X)$.*

Now let b be any number greater than $r_\sigma(A)$, and let $f(z)$ be a complex-valued function that is analytic in $|z| < b$. Thus,

$$f(z) = \sum_0^\infty a_k z^k, \quad |z| < b \qquad [21]$$

We can define $f(A)$ as follows: the operators

$$\sum_0^n a_k A^k$$

converge in norm, since

$$\sum_0^\infty |a_k| \cdot \|A^k\| < \infty$$

This last statement follows from the fact that if c is any number satisfying $r_\sigma(A) < c < b$, then

$$\|A^k\|^{1/k} \leq c$$

for k sufficiently large, and the series

$$\sum_0^\infty |a_k| c^k$$

is convergent. We define $f(A)$ to be

$$\sum_0^\infty a_k A^k \qquad [22]$$

By Theorem 10, this gives

$$\begin{aligned} f(A) &= \frac{1}{2\pi i}\sum_0^\infty a_k \oint_C z^k (z-A)^{-1}\,dz \\ &= \frac{1}{2\pi i}\oint_C \sum_0^\infty a_k z^k (z-A)^{-1}\,dz \\ &= \frac{1}{2\pi i}\oint_C f(z)(z-A)^{-1}\,dz \end{aligned} \qquad [23]$$

where C is any circle about the origin with radius greater than $r_\sigma(A)$ and less than b.

We can now give the formula that we promised. Suppose $f(z)$ does not vanish for $|z| < b$. Set $g(z) = 1/f(z)$. Then $g(z)$ is analytic in $|z| < b$, and hence $g(A)$ is defined. Moreover,

$$\begin{aligned} f(A)g(A) &= \frac{1}{2\pi i}\oint_C f(z)g(z)(z-A)^{-1}\,dz \\ &= \frac{1}{2\pi i}\oint_C (z-A)^{-1}\,dz = I \end{aligned}$$

Since $f(A)$ and $g(A)$ clearly commute, we see that $f(A)^{-1}$ exists and equals $g(A)$. Hence,

$$f(A)^{-1} = \frac{1}{2\pi i}\oint_C \frac{1}{f(z)}(z-A)^{-1}\,dz \qquad [24]$$

In particular, if

$$g(z) = 1/f(z) = \sum_0^\infty c_k z^k, \quad |z| < b$$

then

$$f(A)^{-1} = \sum_0^\infty c_k A^k \qquad [25]$$

Now, suppose $f(z)$ is analytic in an open set Ω containing $\sigma(A)$, but not analytic in a disk of radius greater than $r_\sigma(A)$. In this case, we cannot say that the series [22] converges in norm to an operator in $B(X)$. However, we can still define $f(A)$ in the following way: there exists an open set ω whose closure $\bar{\omega} \subset \Omega$ and whose boundary $\partial\omega$ consists of a finite number of simple closed curves that do not intersect, and such that $\sigma(A) \subset \omega$. (That such a set always exists is left as an exercise; see, e.g., Schechter (2002).) We now define $f(A)$ by

$$f(A) = \frac{1}{2\pi i}\oint_{\partial\omega} f(z)(z-A)^{-1}\,dz \qquad [26]$$

where the line integrals are to be taken in the proper directions. It is easily checked that $f(A) \in B(X)$ and is independent of the choice of the set ω. By [23], this definition agrees with the one given above for the case when Ω contains a disk of radius greater than $r_\sigma(A)$. Note that if Ω is not connected, $f(z)$ need not be the same function on different components of Ω.

Now suppose $f(z)$ does not vanish on $\sigma(A)$. Then we can choose ω so that $f(z)$ does not vanish on $\bar{\omega}$ (this is also an exercise). Thus, $g(z) = 1/f(z)$ is analytic on an open set containing $\bar{\omega}$ so that $g(A)$ is defined. Since $f(z)g(z) = 1$, one would expect that $f(A)g(A) = g(A)f(A) = I$, in which case, it would follow that $f(A)^{-1}$ exists and is equal to $g(A)$. This follows from

Lemma 2 *If $f(z)$ and $g(z)$ are analytic in an open set Ω containing $\sigma(A)$ and*

$$h(z) = f(z)g(z)$$

then $h(A) = f(A)g(A)$.

Therefore, it follows that we have

Theorem 13 *If A is in $B(X)$ and $f(z)$ is a function analytic in an open set Ω containing $\sigma(A)$ such that $f(z) \neq 0$ on $\sigma(A)$, then $f(A)^{-1}$ exists and is given by*

$$f(A)^{-1} = \frac{1}{2\pi i}\oint_{\partial\omega} \frac{1}{f(z)}(z-A)^{-1}\,dz$$

where ω is any open set such that

(i) $\sigma(A) \subset \omega$, $\bar{\omega} \subset \Omega$,
(ii) *$\partial\omega$ consists of a finite number of simple closed curves, and*
(iii) $f(z) \neq 0$ *on* $\bar{\omega}$.

Now that we have defined $f(A)$ for functions analytic in a neighborhood of $\sigma(A)$, we can show that the spectral mapping theorem holds for such functions as well (see Theorem 8). We have

Theorem 14 *If $f(z)$ is analytic in a neighborhood of $\sigma(A)$, then*

$$\sigma(f(A)) = f(\sigma(A)) \qquad [27]$$

that is, $\mu \in \sigma(f(A))$ if and only if $\mu = f(\lambda)$ for some $\lambda \in \sigma(A)$.

Complexification

What we have just done is valid for complex Banach spaces. Suppose, however, we are dealing with a real Banach space. What can be said then?

Let X be a real Banach space. Consider the set Z of all ordered pairs $\langle x, y\rangle$ of elements of X. We set

$$\langle x_1, y_1\rangle + \langle x_2, y_2\rangle = \langle x_1 + x_2,\ y_1 + y_2\rangle$$
$$(\alpha + \mathrm{i}\beta)\langle x, y\rangle = \langle(\alpha x - \beta y), (\beta x + \alpha y)\rangle$$
$$\alpha, \beta \in \mathbb{R}$$

With these definitions, one checks easily that Z is a complex vector space. The set of elements of Z of the form $\langle x, 0\rangle$ can be identified with X. We would like to introduce a norm on Z that would make Z into a Banach space and satisfy

$$\|\langle x, 0\rangle\| = \|x\|, \quad x \in X$$

An obvious suggestion is

$$(\|x\|^2 + \|y\|^2)^{1/2}$$

However, it is soon discovered that this is not a norm on Z (why?). We have to be more careful. One that works is given by

$$\|\langle x, y\rangle\| = \max_{\alpha^2+\beta^2=1} (\|\alpha x - \beta y\|^2 + \|\beta x + \alpha y\|^2)^{1/2}$$

With this norm, Z becomes a complex Banach space having the desired properties.

Now let A be an operator in $B(X)$. We define an operator $\hat{A}$ in $B(Z)$ by

$$\hat{A}\langle x, y\rangle = \langle Ax, Ay\rangle$$

Then

$$\begin{aligned}
&\|\hat{A}\langle x, y\rangle\| \\
&\quad = \max_{\alpha^2+\beta^2=1} (\|\alpha Ax - \beta Ay\|^2 + \|\beta Ax + \alpha Ay\|^2)^{1/2} \\
&\quad = \max_{\alpha^2+\beta^2=1} (\|A(\alpha x - \beta y)\|^2 + \|A(\beta x + \alpha y)\|^2)^{1/2} \\
&\quad \le \|A\| \cdot \|\langle x, y\rangle\|
\end{aligned}$$

Thus,

$$\|\hat{A}\| \le \|A\|$$

But,

$$\|\hat{A}\| \ge \sup_{x \ne 0} \frac{\|\langle Ax, 0\rangle\|}{\|\langle x, 0\rangle\|} = \|A\|$$

Hence,

$$\|\hat{A}\| = \|A\|$$

If λ is real, then

$$(\hat{A} - \lambda)\langle x, y\rangle = \langle(A - \lambda)x, (A - \lambda)y\rangle$$

This shows that $\lambda \in \rho(\hat{A})$ if and only if $\lambda \in \rho(A)$. Similarly, if $p(t)$ is a polynomial with real coefficients, then

$$p(\hat{A})\langle x, y\rangle = \langle p(A)x, p(A)y\rangle$$

showing that $p(\hat{A})$ has an inverse in $B(Z)$ if and only if $p(A)$ has an inverse in $B(X)$. Hence, we have

Theorem 15 *Equation [16] has a unique solution for each y in X if and only if $p(\lambda) \ne 0$ for all $\lambda \in \sigma(\hat{A})$.*

In the example given earlier, the operator $\hat{A}$ has eigenvalues i and $-$i. Hence, -1 is in the spectrum of $\hat{A}^2$ and also in that of A^2. Thus, the equation

$$(A^2 + 1)x = y$$

cannot be solved uniquely for all y.

Compact Operators

Let X, Y be normed vector spaces. A linear operator K from X to Y is called compact (or completely continuous) if $D(K) = X$ and for every sequence $\{x_n\} \subset X$ such that $\|x_n\| \le C$, the sequence $\{Kx_n\}$ has a subsequence which converges in Y. The set of all compact operators from X to Y is denoted by $K(X, Y)$.

A compact operator is bounded. Otherwise, there would be a sequence $\{x_n\}$ such that $\|x_n\| \le C$, while $\|Kx_n\| \to \infty$. Then $\{Kx_n\}$ could not have a convergent subsequence. The sum of two compact operators is compact, and the same is true of the product of a scalar and a compact operator. Hence, $K(X, Y)$ is a subspace of $B(X, Y)$.

If $A \in B(X, Y)$ and $K \in K(Y, Z)$, then $KA \in K(X, Z)$. Similarly, if $L \in K(X, Y)$ and $B \in B(Y, Z)$, then $BL \in K(X, Z)$.

Suppose $K \in B(X, Y)$, and there is a sequence $\{F_n\}$ of compact operators such that

$$\|K - F_n\| \longrightarrow 0 \text{ as } n \longrightarrow \infty \qquad [28]$$

We claim that if Y is a Banach space, then K is compact.

Theorem 16 *Let X be a normed vector space and Y a Banach space. If L is in $B(X, Y)$ and there is a sequence $\{K_n\} \subset K(X, Y)$ such that*

$$\|L - K_n\| \longrightarrow 0 \text{ as } n \longrightarrow 0$$

then L is in $K(X, Y)$.

Theorem 17 *Let X be a Banach space and let K be an operator in $K(X)$. Set $A = I - K$. Then, $R(A)$ is closed in X and $\dim N(A) = \dim N(A')$ is finite.*

In particular, either $R(A) = X$ *and* $N(A) = \{0\}$, *or* $R(A) \neq X$ *and* $N(A) \neq \{0\}$.

The last statement of Theorem 17 is known as the "Fredholm alternative."

Let X, Y be Banach spaces. An operator $A \in B(X, Y)$ is said to be a Fredholm operator from X to Y if

1. $\alpha(A) = \dim N(A)$ is finite,
2. $R(A)$ is closed in Y, and
3. $\beta(A) = \dim N(A')$ is finite.

The set of Fredholm operators from X to Y is denoted by $\Phi(X, Y)$. If $X = Y$ and $K \in K(X)$, then, clearly, $I - K$ is a Fredholm operator. The index of a Fredholm operator is defined as

$$i(A) = \alpha(A) - \beta(A) \qquad [29]$$

For $K \in K(X)$, we have shown that $i(I - K) = 0$ (Theorem 17).

Theorem 18 *Let* X, Y *be normed vector spaces, and assume that* K *is in* $K(X, Y)$. *Then* K' *is in* $K(Y', X')$.

Let X be a Banach space, and suppose $K \in K(X)$. If λ is a nonzero scalar, then

$$\lambda I - K = \lambda(I - \lambda^{-1}K) \in \Phi(X) \qquad [30]$$

For an arbitrary operator $A \in B(X)$, the set of all scalars λ for which $\lambda I - A \in \Phi(X)$ is called the Φ-set of A and is denoted by Φ_A. Thus, [30] gives

Theorem 19 *If* X *is a Banach space and* K *is in* $K(X)$, *then* Φ_K *contains all scalars* $\lambda \neq 0$.

Theorem 20 *Under the hypothesis of Theorem 19,* $\alpha(K - \lambda) = 0$ *except for, at most, a denumerable set* S *of values of* λ. *The set* S *depends on* K *and has 0 as its only possible limit point. Moreover, if* $\lambda \neq 0$ *and* $\lambda \notin S$, *then* $\alpha(K - \lambda) = 0, R(K - \lambda) = X$ *and* $K - \lambda$ *has an inverse in* $B(X)$.

Unbounded Operators

In many applications, one runs into unbounded operators instead of bounded ones. This is particularly true in the case of differential equations. For instance, consider the operator d/dt on $C[0, 1]$ with domain consisting of continuously differentiable functions. It is clearly unbounded. In fact, the sequence $x_n(t) = t^n$ satisfies $\|x_n\| = 1, \|dx_n/dt\| = n \to \infty$ as $n \to \infty$. It would, therefore, be useful if some of the results that we have stated for bounded operators would also hold for unbounded ones. We shall see that, indeed, many of them do. Unless otherwise specified, X, Y, Z, and W will denote Banach spaces in this article.

Let X, Y be normed vector spaces, and let A be a linear operator from X to Y. We now officially lift our restriction that $D(A) = X$. However, if $A \in B(X, Y)$, it is still to be assumed that $D(A) = X$.

The operator A is called closed if whenever $\{x_n\} \subset D(A)$ is a sequence satisfying

$$x_n \longrightarrow x \text{ in } X, \qquad Ax_n \longrightarrow y \text{ in } Y \qquad [31]$$

then $x \in D(A)$ and $Ax = y$. Clearly, all operators in $B(X, Y)$ are closed.

To define A' for an unbounded operator, we follow the definition for bounded operators, and exercise a bit of care. We want

$$A'y'(x) = y'(Ax), \quad x \in D(A) \qquad [32]$$

Thus, we say that $y' \in D(A')$ if there is an $x' \in X'$ such that

$$x'(x) = y'(Ax), \quad x \in D(A) \qquad [33]$$

Then we define $A'y'$ to be x'. In order that this definition make sense, we need x' to be unique, that is, that $x'(x) = 0$ for all $x \in D(A)$ should imply that $x' = 0$. This is true if and only if $D(A)$ is dense in X. To summarize, we can define A' for any linear operator from X to Y provided $D(A)$ is dense in X. We take $D(A')$ to be the set of those $y' \in Y'$ for which there is an $x' \in X'$ satisfying [33]. This x' is unique, and we set $A'y' = x'$. Note that if

$$|y'(Ax)| \leq C\|x\|, \quad x \in D(A)$$

then a simple application of the Hahn–Banach theorem (see e.g., Schechter (2002) or the appendix) shows that $y' \in D(A')$.

We define unbounded Fredholm operators in the following way: let X, Y be Banach spaces. Then the set $\Phi(X, Y)$ of Fredholm operators from X to Y consists of linear operators from X to Y such that

1. $D(A)$ is dense in X,
2. A is closed,
3. $\alpha(A) = \dim N(A) < \infty$,
4. $R(A)$ is closed in Y, and
5. $\beta(A) = \dim N(A') < \infty$.

The Essential Spectrum

Let A be a linear operator on a normed vector space X. We say that $\lambda \in \rho(A)$ if $R(A - \lambda)$ is dense in X and there is a $T \in B(X)$ such that

$$\begin{aligned} T(A - \lambda) &= I \text{ on } D(A) \\ (A - \lambda)T &= I \text{ on } R(A - \lambda) \end{aligned} \qquad [34]$$

Otherwise, $\lambda \in \sigma(A)$. As before, $\rho(A)$ and $\sigma(A)$ are called the resolvent set and spectrum of A, respectively. To show the relationship of this definition to the one given before, we note the following.

Lemma 3 *If X is a Banach space and A is closed, then $\lambda \in \rho(A)$ if and only if*

$$\alpha(A-\lambda)=0, \qquad R(A-\lambda)=X \qquad [35]$$

Throughout the remainder of this section, we shall assume that X is a Banach space, and that A is a densely defined, closed linear operator on X. We ask the following question: what points of $\sigma(A)$ can be removed from the spectrum by the addition of a compact operator to A? The answer to this question is closely related to the set Φ_A. We define this to be the set of all scalars λ such that $A-\lambda \in \Phi(X)$. We have

Theorem 21 *The set Φ_A is open, and $i(A-\lambda)$ is constant on each of its components.*

We also have

Theorem 22 *$\Phi_{A+K}=\Phi_A$ for all K which are A-compact, and $i(A+K-\lambda)=i(A-\lambda)$ for all $\lambda \in \Phi_A$.*

Set

$$\sigma_e(A)=\bigcap_{K\in K(X)} \sigma(A+K)$$

We call $\sigma_e(A)$ the essential spectrum of A (there are other definitions). It consists of those points of $\sigma(A)$ which cannot be removed from the spectrum by the addition of a compact operator to A. We now characterize $\sigma_e(A)$.

Theorem 23 *$\lambda \notin \sigma_e(A)$ if and only if $\lambda \in \Phi_A$ and $i(A-\lambda)=0$.*

Normal Operators

A sequence of elements $\{\varphi_n\}$ in a Hilbert space is called orthonormal if

$$(\varphi_m,\varphi_n)=\begin{cases}0, & m\neq n\\ 1, & m=n\end{cases} \qquad [36]$$

(for definitions, see, e.g., Schechter (2002) or the appendix at the end of this article).

Let $\{\varphi_n\}$ be an orthonormal sequence (finite or infinite) in a Hilbert space H. Let $\{\lambda_k\}$ be a sequence (of the same length) of scalars satisfying

$$|\lambda_k| \le C$$

Then for each element $f \in H$, the series

$$\sum \lambda_k (f,\varphi_k)\varphi_k$$

converges in H. Define the operator A on H by

$$Af=\sum \lambda_k (f,\varphi_k)\varphi_k \qquad [37]$$

Clearly, A is a linear operator. It is also bounded, since

$$\|Af\|^2=\sum |\lambda_k|^2 |(f,\varphi_k)|^2 \le C^2\|f\|^2 \qquad [38]$$

by Bessel's inequality

$$\sum_1^\infty (f,\varphi_k)^2 \le \|f\|^2 \qquad [39]$$

For convenience, let us assume that each $\lambda_k \neq 0$ (just remove those φ_k corresponding to the λ_k that vanish). In this case, $N(A)$ consists of precisely those $f \in H$ which are orthogonal to all of the φ_k. Clearly, such f are in $N(A)$. Conversely, if $f \in N(A)$, then

$$0=(Af,\varphi_k)=\lambda_k(f,\varphi_k)$$

Hence, $(f,\varphi_k)=0$ for each k. Moreover, each λ_k is an eigenvalue of A with φ_k the corresponding eigenvector. This follows immediately from [37]. Since $\sigma(A)$ is closed, it also contains the limit points of the λ_k.

Next, we shall see that if $\lambda \neq 0$ is not a limit point of the λ_k, then $\lambda \in \rho(A)$. To show this, we solve

$$(\lambda - A)u=f \qquad [40]$$

for any $f \in H$. Any solution of [40] satisfies

$$\lambda u=f+Au=f+\sum \lambda_k (u,\varphi_k)\varphi_k \qquad [41]$$

Hence,

$$\lambda(u,\varphi_k)=(f,\varphi_k)+\lambda_k(u,\varphi_k)$$

or

$$(u,\varphi_k)=\frac{(f,\varphi_k)}{\lambda-\lambda_k} \qquad [42]$$

Substituting back in [41], we obtain

$$\lambda u=f+\sum \frac{\lambda_k (f,\varphi_k)\varphi_k}{\lambda-\lambda_k} \qquad [43]$$

Since λ is not a limit point of the λ_k, there is a $\delta>0$ such that

$$|\lambda-\lambda_k| \ge \delta, \quad k=1,2,\ldots$$

Hence, the series in [43] converges for each $f \in H$. It is an easy exercise to verify that [43] is indeed a solution of [40]. To see that $(\lambda-A)^{-1}$ is bounded, note that

$$|\lambda|\cdot\|u\| \le \|f\|+C\|f\|/\delta \qquad [44]$$

(cf. [38]). Thus, we have proved

Lemma 4 *If the operator A is given by* [37], *then $\sigma(A)$ consists of the points λ_k, their limit points and possibly 0. $N(A)$ consists of those u which are orthogonal to all of the φ_k. For $\lambda \in \rho(A)$, the solution of* [40] *is given by* [43].

We see from all this that the operator [37] has many useful properties. Therefore, it would be desirable to determine conditions under which operators are guaranteed to be of that form. For this purpose, we note another property of A. It is expressed in terms of the Hilbert space adjoint of A.

Let H_1 and H_2 be Hilbert spaces, and let A be an operator in $B(H_1, H_2)$. For fixed $y \in H_2$, the expression $Fx = (Ax, y)$ is a bounded linear functional on H_1. By the Riesz representation theorem (see, e.g., Schechter (2002) or the appendix at the end of this article), there is a $z \in H_1$ such that $Fx = (x, z)$ for all $x \in H_1$. Set $z = A^*y$. Then A^* is a linear operator from H_2 to H_1 satisfying

$$(Ax, y) = (x, A^*y) \qquad [45]$$

A^* is called the Hilbert space adjoint of A. Note the difference between A^* and the operator A' defined for a Banach space. As in the case of the operator A', we note that A^* is bounded and

$$\|A^*\| = \|A\| \qquad [46]$$

Returning to the operator A, we remove the assumption that each $\lambda_k \neq 0$ and note that

$$\begin{aligned}(Au, v) &= \sum \lambda_k (u, \varphi_k)(\varphi_k, v) \\ &= \left(u, \sum \bar{\lambda}_k (v, \varphi_k)\varphi_k\right)\end{aligned}$$

showing that

$$A^*v = \sum \bar{\lambda}_k (v, \varphi_k)\varphi_k \qquad [47]$$

(If H is a complex Hilbert space, then the complex conjugates $\bar{\lambda}_k$ of the λ_k are required. If H is a real Hilbert space, then the λ_k are real, and it does not matter.) Now, by Lemma 4, we see that each $\bar{\lambda}_k$ is an eigenvalue of A^* with φ_k a corresponding eigenvector. Note also that

$$\|A^*f\|^2 = \sum |\lambda_k|^2 |(f, \varphi_k)|^2 \qquad [48]$$

showing that

$$\|A^*f\| = \|Af\|, \quad f \in H \qquad [49]$$

An operator satisfying [49] is called normal. An important characterization is given by

Theorem 24 *An operator is normal and compact if and only if it is of the form* [37] *with $\{\varphi_k\}$ an orthonormal set and $\lambda_k \to 0$ as $k \to \infty$.*

We also have

Lemma 5 *If A is normal, then*

$$\|(A^* - \bar{\lambda})u\| = \|(A - \lambda)u\|, \quad u \in H \qquad [50]$$

Corollary 2 *If A is normal and $A\varphi = \lambda\varphi$, then $A^*\varphi = \bar{\lambda}\varphi$.*

Lemma 6 *If A is normal and compact, then it has an eigenvalue λ such that $|\lambda| = \|A\|$.*

We also have

Corollary 3 *If A is a normal compact operator, then there is an orthonormal sequence $\{\varphi_k\}$ of eigenvectors of A such that every element u in H can be written in the form*

$$u = h + \sum (u, \varphi_k)\varphi_k \qquad [51]$$

where $h \in N(A)$.

Hyponormal Operators

An operator A in $B(H)$ is called hyponormal if

$$\|A^*u\| \leq \|Au\|, \quad u \in H \qquad [52]$$

or, equivalently, if

$$([AA^* - A^*A]u, u) \leq 0, \quad u \in H \qquad [53]$$

Of course, a normal operator is hyponormal. An operator $A \in B(H)$ is called seminormal if either A or A^* is hyponormal. We have

Theorem 25 *If A is seminormal, then*

$$r_\sigma(A) = \|A\| \qquad [54]$$

We have earlier defined the essential spectrum of an operator A to be

$$\sigma_e(A) = \bigcap_{K \in K(H)} \sigma(A + K) \qquad [55]$$

It was shown that $\lambda \notin \sigma_e(A)$ if and only if $\lambda \in \Phi_A$ and $i(A - \lambda) = 0$ (Theorem 23). Let us show that we can be more specific in the case of seminormal operators.

Theorem 26 *If A is a seminormal operator, then $\lambda \in \sigma(A) \setminus \sigma_e(A)$ if and only if λ is an isolated eigenvalue with $r(A - \lambda) = \lim_{n \to \infty} \alpha[(A - \lambda)^n] < \infty$.*

Lemma 7 *If A is hyponormal, then so is $B = A - \lambda$ for any complex λ.*

Lemma 8 *If B is hyponormal with 0 an isolated point of $\sigma(B)$ and either $\alpha(B)$ or $\beta(B)$ is finite, then $B \in \Phi(H)$ and $i(B) = 0$.*

There is a simple consequence of Lemma 8.

Corollary 4 *If A is seminormal and λ is an isolated point of $\sigma(A)$, then λ is an eigenvalue of A.*

We also have the following:

Theorem 27 *Let A be a seminormal operator such that $\sigma(A)$ has no nonzero limit points. Then A is compact and normal. Thus, it is of the form [37] with the $\{\varphi_k\}$ orthonormal and $\lambda_k \to 0$.*

Corollary 5 *If A is seminormal and compact, then it is normal.*

Spectral Resolution

We saw in the section "Operational calculus" that, in a Banach space X, we can define $f(A)$ for any $A \in B(X)$ provided $f(z)$ is a function analytic in a neighborhood of $\sigma(A)$. In this section, we shall show that we can do better in the case of self-adjoint operators.

A linear operator A on a Hilbert space X is called self-adjoint if it has the property that $x \in D(A)$ and $Ax = f$ if and only if

$$(x, Ay) = (f, y), \quad y \in D(A)$$

In particular, it satisfies

$$(Ax, y) = (x, Ay), \quad x, y \in D(A)$$

A bounded self-adjoint operator is normal.

To get an idea, let A be a compact, self-adjoint operator on H. Then by Theorem 24,

$$Au = \sum \lambda_k (u, \varphi_k)\varphi_k \tag{56}$$

where $\{\varphi_k\}$ is an orthonormal sequence of eigenvectors and the λ_k are the corresponding eigenvalues of A. Now let $p(t)$ be a polynomial with real coefficients having no constant term

$$p(t) = \sum_1^m a_k t^k \tag{57}$$

Then $p(A)$ is compact and self-adjoint. Let $\mu \neq 0$ be a point in $\sigma(p(A))$. Then $\mu = p(\lambda)$ for some $\lambda \in \sigma(A)$ (Theorem 8). Now $\lambda \neq 0$ (otherwise we would have $\mu = p(0) = 0$). Hence, it is an eigenvalue of A (see the section "The spectrum and resolvent sets"). If φ is a corresponding eigenvector, then

$$\begin{aligned}[p(A) - \mu]\varphi &= \sum a_k A^k \varphi - \mu\varphi \\ &= \sum a_k \lambda^k \varphi - \mu\varphi \\ &= [p(\lambda) - \mu]\varphi = 0\end{aligned}$$

Thus μ is an eigenvalue of $p(A)$ and φ is a corresponding eigenvector. This shows that

$$p(A)u = \sum p(\lambda_k)(u, \varphi_k)\varphi_k \tag{58}$$

Now, the right-hand side of [58] makes sense if $p(t)$ is any function bounded on $\sigma(A)$ (see the section "Normal operators"). Therefore it seems plausible to define $p(A)$ by means of [58]. Of course, for such a definition to be useful, one would need certain relationships to hold. In particular, one would want $f(t)g(t) = h(t)$ to imply $f(A)g(A) = h(A)$. We shall discuss this a bit later.

If A is not compact, we cannot, in general, obtain an expansion in the form [56]. However, we can obtain something similar. In fact, we have

Theorem 28 *Let A be a self-adjoint operator in $B(H)$. Set*

$$m = \inf_{\|u\|=1} (Au, u), \qquad M = \sup_{\|u\|=1} (Au, u)$$

Then there is a family $\{E(\lambda)\}$ of orthogonal projection operators on H depending on a real parameter λ and such that:

(i) $E(\lambda_1) \leq E(\lambda_2)$ for $\lambda_1 \leq \lambda_2$;
(ii) $E(\lambda)u \to E(\lambda_0)u$ as $\lambda_0 < \lambda \to \lambda_0, u \in H$;
(iii) $E(\lambda) = 0$ for $\lambda < m, E(\lambda) = I$ for $\lambda \geq M$;
(iv) $AE(\lambda) = E(\lambda)A$; *and*
(v) *if $a < m, b \geq M$ and $p(t)$ is any polynomial, then*

$$p(A) = \int_a^b p(\lambda)\, dE(\lambda) \tag{59}$$

This means the following. Let $a = \lambda_0 < \lambda_1 < \cdots < \lambda_n = b$ be any partition of $[a, b]$, and let λ_k' be any number satisfying $\lambda_{k-1} \leq \lambda_k' \leq \lambda_k$. Then

$$\sum_1^n p(\lambda_k')[E(\lambda_k) - E(\lambda_{k-1})] \to p(A) \tag{60}$$

in $B(H)$ as $\eta = \max(\lambda_k - \lambda_{k-1}) \to 0$.

Theorem 29 *Let A be a self-adjoint operator on H. Then there is a family $\{E(\lambda)\}$ of orthogonal projection operators on H satisfying* (i) *and* (ii) *of Theorem 28 and*

(i) $$E(\lambda) \to \begin{cases} 0 & \text{as } \lambda \to -\infty \\ I & \text{as } \lambda \to +\infty \end{cases}$$

(ii) $$E(\lambda)A \subset AE(\lambda)$$

(iii) $$p(A) = \int_{-\infty}^{\infty} p(\lambda)\, dE(\lambda)$$

for any polynomial $p(t)$.

These theorems are known as the spectral theorems for self-adjoint operators.

Appendix

Here we include some background material related to the text.

Consider a collection C of elements or "vectors" with the following properties:

1. They can be added. If f and g are in C, so is $f+g$.
2. $f+(g+h)=(f+g)+h, \quad f,g,h \in C$.
3. There is an element $0 \in C$ such that $h+0=h$ for all $h \in C$.
4. For each $h \in C$ there is an element $-h \in C$ such that $h+(-h)=0$.
5. $g+h=h+g, \quad g,h \in C$.
6. For each real number $\alpha, \alpha h \in C$.
7. $\alpha(g+h)=\alpha g+\alpha h$.
8. $(\alpha+\beta)h=\alpha h+\beta h$.
9. $\alpha(\beta h)=(\alpha\beta)h$.
10. To each $h \in C$ there corresponds a real number $\|h\|$ with the following properties:
11. $\|\alpha h\|=|\alpha|\|h\|$.
12. $\|h\|=0$ if, and only if, $h=0$.
13. $\|g+h\| \leq \|g\|+\|h\|$.
14. If $\{h_n\}$ is a sequence of elements of C such that $\|h_n-h_m\| \to 0$ as $m,n \to \infty$, then there is an element $h \in C$ such that $\|h_n-h\| \to 0$ as $n \to \infty$.

A collection of objects which satisfies statements (1)–(9) and the additional statement

15. $1h=h$

is called a vector space or linear space.

A set of objects satisfying statements (1)–(13) is called a normed vector space, and the number $\|h\|$ is called the norm of h. Although statement (15) is not implied by statements (1)–(9), it is implied by statements (1)–(13). A sequence satisfying

$$\|h_n-h_m\| \to 0 \quad \text{as } m,n \to \infty$$

is called a Cauchy sequence. Property (14) states that every Cauchy sequence converges in norm to a limit (i.e., satisfies $\|h_n-h\| \to 0$ as $n \to \infty$). Property (14) is called completeness, and a normed vector space satisfying it is called a complete normed vector space or a Banach space.

We shall write

$$h_n \to h \quad \text{as } n \to \infty$$

when we mean

$$\|h_n-h\| \to 0 \quad \text{as } n \to \infty$$

A subset U of a vector space V is called a subspace of V if $\alpha_1x_1+\alpha_2x_2$ is in U whenever x_1,x_2 are in U and α_1,α_2 are scalars.

A subset U of a normed vector space X is called closed if for every sequence $\{x_n\}$ of elements in U having a limit in X, the limit is actually in U.

Consider a vector space X having a mapping (f,g) from pairs of its elements to the reals such that

1. $(\alpha f,g)=\alpha(f,g)$
2. $(f+g,h)=(f,h)+(g,h)$
3. $(f,g)=(g,f)$
4. $(f,f)>0$ unless $f=0$.

Then

$$(f,g)^2 \leq (f,f)(g,g), \quad f,g \in X \qquad [61]$$

An expression (f,g) that assigns a real number to each pair of elements of a vector space and satisfies the aforementioned properties is called a scalar (or inner) product.

If a vector space X has a scalar product (f,g), then it is a normed vector space with norm $\|f\|=(f,f)^{1/2}$. A vector space which has a scalar product and is complete with respect to the induced norm is called a Hilbert space. Every Hilbert space is a Banach space, but the converse is not true. Inequality [61] is known as the Cauchy–Schwarz inequality. $\mathbb{R}^n$ is a Hilbert space.

Let H be a Hilbert space and let (x,y) denote its scalar product. If we fix y, then the expression (x,y) assigns to each $x \in H$ a number. An assignment F of a number to each element x of a vector space is called a functional and denoted by $F(x)$. The scalar product is not the first functional we have encountered. In any normed vector space, the norm is also a functional. The functional $F(x)=(x,y)$ satisfies

$$F(\alpha_1x_1+\alpha_2x_2)=\alpha_1F(x_1)+\alpha_2F(x_2) \qquad [62]$$

for α_1,α_2 scalars. A functional satisfying [62] is called linear. Another property is

$$|F(x)| \leq M\|x\|, \quad x \in H \qquad [63]$$

which follows immediately from Schwarz's inequality (cf. [61]). A functional satisfying [63] is called bounded. The norm of such a functional is defined to be

$$\|F\|=\sup_{x\in H, x\neq 0} \frac{|F(x)|}{\|x\|}$$

Thus for y fixed, $F(x)=(x,y)$ is a bounded linear functional in the Hilbert space H. We have

Theorem 30 *For every bounded linear functional F on a Hilbert space H there is a unique element $y \in H$ such that*

$$F(x) = (x, y) \quad \text{for all } x \in H \tag{64}$$

Moreover,

$$\|y\| = \sup_{x \in H, x \neq 0} \frac{|F(x)|}{\|x\|} = \|F\| \tag{65}$$

Theorem 30 is known as the "Riesz representation theorem."

For any normed vector space X, let X' denote the set of bounded linear functionals on X. If $f, g \in X'$, we say that $f = g$ if

$$f(x) = g(x) \quad \text{for all } x \in X$$

The "zero" functional is the one assigning zero to all $x \in X$. We define $h = f + g$ by

$$h(x) = f(x) + g(x), \quad x \in X$$

and $g = \alpha f$ by

$$g(x) = \alpha f(x), \quad x \in X$$

Under these definitions, X' becomes a vector space. We have been employing the expression

$$\|f\| = \sup_{x \neq 0} \frac{|f(x)|}{\|x\|}, \quad f \in X' \tag{66}$$

This is easily seen to be a norm. Thus X' is a normed vector space.

We also have

Theorem 31 *Let M be a subspace of a normed vector space X, and suppose that $f(x)$ is a bounded linear functional on M. Set*

$$\|f\| = \sup_{x \in M, x \neq 0} \frac{|f(x)|}{\|x\|}$$

Then there is a bounded linear functional $F(x)$ on the whole of X such that

$$F(x) = f(x), \quad x \in M \tag{67}$$

and

$$\|F\| = \sup_{x \in X, x \neq 0} \frac{|F(x)|}{\|x\|} = \|f\| = \sup_{x \in M, x \neq 0} \frac{|f(x)|}{\|x\|} \tag{68}$$

Theorem 31 is known as the "Hahn–Banach theorem."

If A is a linear operator from X to Y, with $R(A) = Y$ and $N(A) = \{0\}$ (i.e., consists only of the vector 0), we can assign to each $y \in Y$ the unique solution of

$$Ax = y$$

This assignment is an operator from Y to X and is usually denoted by A^{-1} and called the inverse operator of A. It is linear because of the linearity of A. One can ask: "when is A^{-1} continuous?" or, equivalent by, "when is it bounded?" A very important answer to this question is given by

Theorem 32 *If X, Y are Banach spaces and A is a closed linear operator from X to Y with $R(A) = Y, N(A) = \{0\}$, then $A^{-1} \in B(Y, X)$.*

This theorem is sometimes referred to as the "bounded inverse theorem."

If A is self-adjoint and

$$(A - \lambda)x = 0, \qquad (A - \mu)y = 0$$

with $\lambda \neq \mu$, then

$$(x, y) = 0$$

If A has a compact inverse, its eigenvalues cannot have limit points. If A^{-1} is compact, then the eigenelements corresponding to the same eigenvalue form a finite-dimensional subspace.

See also: Ljusternik–Schnirelman Theory; Quantum Mechanical Scattering Theory; Regularization for Dynamical Zeta Functions; Spectral Sequences; Stochastic Resonance.

Further Reading

Bachman G and Narici L (1966) *Functional Analysis*. New York: Academic Press.

Banach S (1955) *Théorie des Opérations Linéaires*. New York: Chelsea.

Berberian SK (1974) *Lectures in Functional Analysis and Operator Theory*. New York: Springer.

Brown A and Pearcy C (1977) *Introduction to Operator Theory*. New York: Springer.

Day MM (1958) *Normed Linear Spaces*. Berlin: Springer.

Dunford N and Schwartz IT (1958, 1963) *Linear Operators, I, II*. New York: Wiley.

Edwards RE (1965) *Functional Analysis*. New York: Holt.

Epstein B (1970) *Linear Functional Analysis*. Philadelphia: W. B. Saunders.

Gohberg IC and Krein MS (1960) The basic propositions on defect numbers, root numbers and indices of linear operators. *Amer. Math. Soc. Transl*, Ser. 2, 13, 185–264.

Goldberg S (1966) *Unbounded Linear Operators*. New York: McGraw-Hill.

Halmos PR (1951) *Introduction to Hilbert Space*. New York: Chelsea.

Hille E and Phillips R (1957) *Functional Analysis and Sem-Groups*. Providence: American Mathematical Society.

Kato T (1966, 1976) *Perturbation Theory for Linear Operators*. Berlin: Springer.

Müller V (2003) *Spectral Theory of Linear Operators – and Spactral System in Banach Algebras*. Basel: Birkhäuser Verlag.
Reed M and Simon B (1972) *Methods of Modern Mathematical Physics, I*. Academic Press.
Riesz F and St.-Nagy B (1955) *Functional Analysis*. New York: Ungar.
Schechter M (2002) *Principles of Functional Analysis*. Providence: American Mathematical Society.
Stone MH (1932) *Linear Transformations in Hilbert Space*. Providence: American Mathematical Society.
Taylor AE (1958) *Introduction to Functional Analysis*. New York: Wiley.
Weidmann J (1980) *Linear Operators in Hilbert Space*. New York: Springer.
Yosida K (1965, 1971) *Functional Analysis*. Berlin: Springer.

Spin Foams

A Perez, Penn State University, University Park, PA, USA

Introduction

In loop quantum gravity (LQG) (*see* Loop Quantum Gravity) – a background independent formulation of quantum gravity – the full quantum dynamics is governed by the following (constraint) operator equations or quantum Einstein equations:

Gauss Law

$$\widehat{G}_i(A,E)|\Psi>:=\widehat{D_a E^a_i}|\Psi>=0$$

Vector constraint

$$\widehat{V}_a(A,E)|\Psi>:=E^a_i\widehat{F^i_{ab}(A)}|\Psi>=0$$

Scalar constraint

$$\widehat{S}(A,E)\Big|\Psi>:=\left[\sqrt{\det E}^{-1}\widehat{E^a_i E^b_j F^{ij}_{ab}(A)}+\cdots\right]\Big|\Psi>=0 \qquad [1]$$

where A^i_a is an SU(2) connection ($i=1,2,3$, $a=1,2,3$), E^a_i is its conjugate momentum (the triad field), $F^{ij}_{ab}(A)$ is the curvature of A^i_a, and D_a is the covariant derivative (*see* Canonical General Relativity). The hat means that the classical phase-space functions are promoted to operators in a kinematical Hilbert space $\mathcal{H}_{\rm kin}$; the solutions are in the so-called physical Hilbert space $\mathcal{H}_{\rm phys}$. The goal of the spin foam approach is to construct a mathematically well-defined notion of path integral for LQG as a device for computing the solutions of the previous equations.

The space of solution of the Gauss and vector constraints [1] is well understood in LQG (*see* Loop Quantum Gravity), and often also called kinematical Hilbert space $\mathcal{H}_{\rm kin}$. The solutions of the scalar constraint can be characterized by the definition of the generalized projection operator P from the kinematical Hilbert space $\mathcal{H}_{\rm kin}$ into the kernel of the scalar constraint $\mathcal{H}_{\rm phys}$. Formally, one can write P as

$$P=\text{“}\prod_{x\in}\delta(\widehat{S}(x))\text{”}=\int D[N]\exp\left[i\int_\Sigma N(x)\widehat{S(x)}\right] \qquad [2]$$

A formal argument shows that P can also be defined in a manifestly covariant manner as a regularization of the formal path integral of general relativity. In first-order variables, it becomes

$$P=\int D[e]\ D[A]\ \mu[A,e]\exp[iS_{\rm GR}(e,A)] \qquad [3]$$

where e is the tetrad field, A is the spacetime connection, and $\mu[A,e]$ denotes the appropriate measure.

In both cases, P characterizes the space of solutions of quantum Einstein equations as for any arbitrary state $|\phi>\in\mathcal{H}_{\rm kin}$ then $P|\phi>$ is a (formal) solution of [1]. Moreover, the matrix elements of P define the physical inner product $(<,>_{\rm p})$ providing the vector space of solutions of [1] with the Hilbert space structure that defines $\mathcal{H}_{\rm phys}$. Explicitly,

$$<s,s'>_{\rm p}\ :=\ <Ps,s'>$$

for $s,s'\in\mathcal{H}_{\rm kin}$.

When these matrix elements are computed in the spin network basis (see **Figure 1**) (*see* Loop Quantum Gravity), they can be expressed as a sum over amplitudes of “spin network histories”: spin foams (**Figure 2**). The latter are naturally given by foam-like combinatorial structures whose basic elements carry quantum numbers of geometry (*see* Loop Quantum Gravity). A spin foam history, from the state $|s>$ to the state $|s'>$, is denoted by a pair $(F_{s\to s'},\{j\})$, where $F_{s\to s'}$ is the 2-complex with boundary given by the graphs of the spin network states $|s'>$ and $|s>$, respectively, and $\{j\}$ is the set of spin quantum numbers labeling its edges (denoted $e\in F_{s\to s'}$) and faces (denoted $f\in F_{s\to s'}$). Vertices are denoted

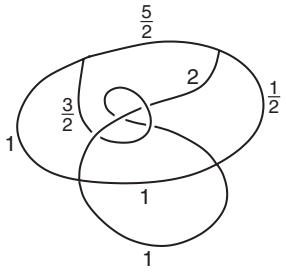


Figure 1 A spin network state is given by a graph embedded in space whose links and nodes are labeled by unitary irreducible representations of SU(2). These states form a complete basis of the kinematical Hilbert space of LQG where the operator equations [1] are defined.

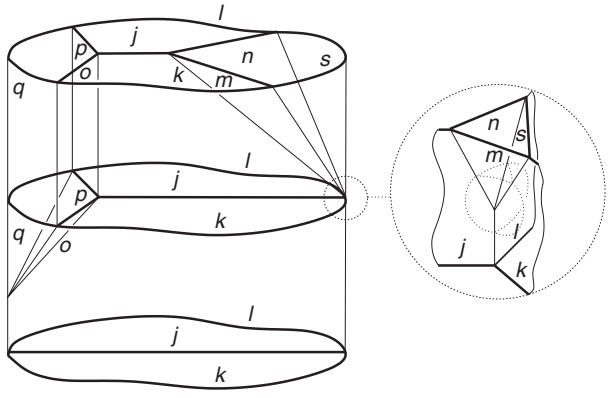


Figure 2 A spin foam as the "colored" 2-complex representing the transition between three different spin network states. A transition vertex is magnified on the right.

$v \in F_{s \to s'}$. The physical inner product can be expressed as a sum over spin foam amplitudes

$$\begin{aligned} <s', s>_{\mathrm{p}} &= <Ps', s> \\ &= \sum_{F_{s\to s'}} N(F_{s\to s'}) \sum_{\{j\}} \prod_{f\in F_{s\to s'}} A_f(j_f) \\ &\quad \times \prod_{e\in F_{s\to s'}} A_e(j_e) \prod_{v\in F_{s\to s'}} A_v(j_v) \end{aligned} \qquad [4]$$

where $N(F_{s\to s'})$ is a (possible) normalization factor, and $A_f(j_f), A_e(j_e)$, and $A_v(j_v)$ are the 2-cell or face amplitude, the edge or 1-cell amplitude, and the 0-cell or vertex amplitude, respectively. These local amplitudes depend on the spin quantum numbers labeling neighboring cells in $F_{s\to s'}$ (e.g., the vertex amplitude of the vertex magnified in Figure 2 is $A_v(j,k,l,m,n,s)$).

The underlying discreteness discovered in LQG is crucial: in the spin foam representation, the functional integral for gravity is replaced by a sum over amplitudes of combinatorial objects given by foam-like configurations (spin foams) as in [4]. A spin foam represents a possible history of the gravitational field and can be interpreted as a set of transitions through different quantum states of space. Boundary data in the path integral are given by the polymer-like excitations (spin network states, Figure 1) representing 3-geometry states in LQG.

Spin Foams in 3D Quantum Gravity

Now we introduce the concept of spin foams in a more explicit way in the context of the quantization of three-dimensional (3D) Riemannian gravity. Later in this section we will present the definition of P from the canonical and covariant viewpoint formally stated in the introduction by eqns [2] and [3], respectively.

The Classical Theory

Riemannian gravity in 3D is a theory with no local degrees of freedom, that is, a topological theory (*see* Topological Quantum Field Theory: Overview). Its action (in the first-order formalism) is given by

$$S(e,\omega) = \int_M \mathrm{tr}(e \wedge F(\omega)) \qquad [5]$$

where $M = \Sigma \times \mathbb{R}$ (for Σ an arbitrary Riemann surface), ω is an SU(2) connection, and the triad e is an su(2)-valued 1-form. The gauge symmetries of the action are the local SU(2) gauge transformations

$$\delta e = [e, \alpha], \qquad \delta\omega = d_\omega \alpha \qquad [6]$$

where α is an su(2)-valued 0-form, and the "topological" gauge transformation

$$\delta e = d_\omega \eta, \qquad \delta\omega = 0 \qquad [7]$$

where d_ω denotes the covariant exterior derivative and η is an su(2)-valued 0-form. The first invariance is manifest from the form of the action, while the second is a consequence of the Bianchi identity, $d_\omega F(\omega) = 0$. The gauge symmetries are so large that all the solutions to the equations of motion are locally pure gauge. The theory has only global or topological degrees of freedom.

Upon the standard $2+1$ decomposition (*see* Canonical General Relativity), the phase space in these variables is parametrized by the pullback to Σ of ω and e. In local coordinates, one can express them in terms of the two 2D connection A_a^i and the triad field $E_j^b = \epsilon^{bc} e_c^k \eta_{jk}$, where $a = 1, 2$ are space coordinate indices and $i, j = 1, 2, 3$ are su(2) indices. The symplectic structure is defined by

$$\{A_a^i(x), E_j^b(y)\} = \delta_a^b \delta_j^i \delta^{(2)}(x,y) \qquad [8]$$

Local symmetries of the theory are generated by the first-class constraints

$$D_b E^b_j = 0, \qquad F^i_{ab}(A) = 0 \qquad [9]$$

which are referred to as the Gauss law and the curvature constraint, respectively – the quantization of these is the analog of [1] in 4D. This simple theory has been quantized in various ways in the literature; here we will use it to introduce the spin foam quantization.

Kinematical Hilbert Space

In analogy with the 4D case, one follows Dirac's procedure finding first a representation of the basic variables in an auxiliary or kinematical Hilbert space $\mathcal{H}_{\rm kin}$. The basic states are functionals of the connection depending on the parallel transport along paths $\gamma \subset \Sigma$: the so-called holonomy. Given a connection $A^i_a(x)$ and a path γ, one defines the holonomy $h_\gamma[A]$ as the path-ordered exponential

$$h_\gamma[A] = P \exp \int_\gamma A \qquad [10]$$

The kinematical Hilbert space, $\mathcal{H}_{\rm kin}$, corresponds to the Ashtekar–Lewandowski (AL) representation of the algebra of functions of holonomies or generalized connections. This algebra is in fact a C*-algebra and is denoted Cyl (*see* Loop Quantum Gravity). Functionals of the connection act in the AL representation simply by multiplication. For example, the holonomy operator acts as follows:

$$\widehat{h_\gamma[A]}\Psi[A] = h_\gamma[A]\Psi[A] \qquad [11]$$

As in 4D, an orthonormal basis of $\mathcal{H}_{\rm kin}$ is defined by the spin network states. Each spin network is labeled by a graph $\gamma \subset \Sigma$, a set of spins $\{j_\ell\}$ labeling links $\ell \in \gamma$, and a set of intertwiners $\{\iota_n\}$ labeling nodes $n \in \gamma$ (**Figure 3**), namely:

$$s_{\gamma,\{j_\ell\},\{\iota_n\}}[A] = \bigotimes_{n\in\gamma} \iota_n \bigotimes_{\ell\in\gamma} \overset{j_\ell}{\prod} (h_\ell[A]) \qquad [12]$$

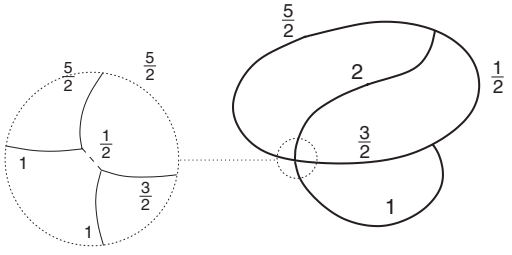


Figure 3 A spin network state in $2+1$ LQG. The decomposition of a 4-valent node in terms of basic 3-valent intertwiners is shown.

where $\overset{j}{\Pi}$ is the unitary irreducible representation matrix of spin j (for a precise definition, *see* Loop Quantum Gravity). For simplicity, we will often denote spin network states $|s>$ omitting the graph and spin labels.

Spin Foams from the Hamiltonian Formulation

The physical Hilbert space, $\mathcal{H}_{\rm phys}$, is defined by those "states" that are annihilated by the constraints. By construction, spin-network states solve the Gauss constraint – $\widehat{D_a E^a_i}|s> = 0$ – as they are manifestly SU(2) gauge invariant (*see* Loop Quantum Gravity). To complete the quantization, one needs to characterize the space of solutions of the quantum curvature constraints ($\widehat{F^i_{ab}}$), and to provide it with the physical inner product. The existence of $\mathcal{H}_{\rm phys}$ is granted by the following:

Theorem 1 *There exists a normalized positive linear form P over* Cyl, *that is*, $P(\psi^*\psi) \geq 0$ *for* $\psi \in$ Cyl *and* $P(1) = 1$, *yielding* (*through the GNS construction* (*see* Algebraic Approach to Quantum Field Theory)) *the physical Hilbert space* $\mathcal{H}_{\rm phys}$ *and the physical representation* π_p *of* Cyl.

The state P contains a very large Gelfand ideal (set of zero norm states) $J := \{\alpha \in \text{Cyl s.t. } P(\alpha^*\alpha) = 0\}$. In fact, the physical Hilbert space $\mathcal{H}_{\rm phys} := \overline{\text{Cyl}/J}$ corresponds to the quantization of finitely many degrees of freedom. This is expected in 3D gravity as the theory does not have local excitations (no "gravitons") (*see* Topological Quantum Field Theory: Overview). The representation π_p of Cyl solves the curvature constraint in the sense that for any functional $f_\gamma[A] \in$ Cyl defined on the subalgebra of functionals defined on contractible graphs $\gamma \in \Sigma$, one has that

$$\pi_p[f_\gamma]\Psi = f_\gamma[0]\Psi \qquad [13]$$

This equation expresses the fact that "$\widehat{F} = 0$" in $\mathcal{H}_{\rm phys}$ (for flat connections, parallel transport is trivial around a contractible region). For $s, s' \in \mathcal{H}_{\rm kin}$, the physical inner product is given by

$$<s, s'>_{\rm p} := P(s^* s) \qquad [14]$$

where the $*$-operation and the product are defined in Cyl.

The previous equation admits a "sum over histories" representation. We shall introduce the concept of the spin foam representation as an explicit construction of the positive linear form P which, as in [2], is formally given by

$$P = \int D[N] \exp\left(\mathrm{i} \int_\Sigma \mathrm{tr}[N\widehat{F}(A)]\right) = \prod_{x\in\Sigma} \delta[\widehat{F(A)}] \qquad [15]$$

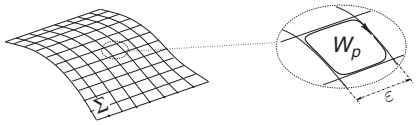


Figure 4 Cellular decomposition of the space manifold Σ (a square lattice in this example), and the infinitesimal plaquette holonomy $W_p[A]$.

where $N(x) \in \mathrm{su}(2)$. One can make the previous formal expression a rigorous definition if one introduces a regularization. Given a partition of Σ in terms of 2D plaquettes of coordinate area ϵ^2, one has that

$$\int_\Sigma \mathrm{tr}[NF(A)] = \lim_{\epsilon \to 0} \sum_{p^i} \epsilon^2 \mathrm{tr}[N_{p^i} F_{p^i}] \qquad [16]$$

where N_{p^i} and F_{p^i} are values of N^i and $\epsilon^{ab} F^i_{ab}[A]$ at some interior point of the plaquette p^i and ϵ^{ab} is the Levi-Civita tensor. Similarly, the holonomy $W_{p^i}[A]$ around the boundary of the plaquette p^i (see **Figure 4**) is given by

$$W_{p^i}[A] = \mathbb{1} + \epsilon^2 F_{p^i}(A) + \mathcal{O}(\epsilon^2) \qquad [17]$$

where $F_{p^i} = \tau_j \epsilon^{ab} F^j_{ab}(x_{p^i})$ (τ_j are the generators of su(2) in the fundamental representation). The previous two equations lead to the following definition: given $s \in \mathrm{Cyl}$ (think of spin network state based on a graph γ), the linear form $P(s)$ is defined as

$$P(s) := \lim_{\epsilon \to 0} \left\langle \Omega \prod_{p^i} \int \mathrm{d}N_{p^i} \exp(\mathrm{i}\ \mathrm{tr}[N_{p^i} W_{p^i}]), s \right\rangle \qquad [18]$$

where $<,>$ is the inner product in the AL representation and $|\Omega>$ is the "vacuum" ($1 \in \mathrm{Cyl}$) in the AL representation. The partition is chosen so that the links of the underlying graph γ border the plaquettes. One can easily perform the integration over the N_{p^i} using the identity (Peter–Weyl theorem)

$$\int \mathrm{d}N \exp(\mathrm{i}\ \mathrm{tr}[NW]) = \sum_j (2j+1) \mathrm{tr}\left[\overset{j}{\Pi}(W)\right] \qquad [19]$$

Using the previous equation

$$P(s) := \lim_{\epsilon \to 0} \prod_{p^i} \sum_{j(p^i)} (2j(p^i)+1) <\Omega\ \mathrm{tr}\left[\overset{j(p^i)}{\Pi}(W_{p^i})\right], s> \qquad [20]$$

where $j(p^i)$ is the spin labeling element of the sum [19] associated to the ith plaquette. Since the $\mathrm{tr}[\Pi^j(W)]$ commute, the ordering of plaquette operators in the previous product does not matter. It can be shown that the limit $\epsilon \to 0$ exists and one can give a closed expression of $P(s)$.

Now in the AL representation (see eqn [11]), each $\mathrm{tr}[\Pi^{j(p^i)}(W_{p^i})]$ acts by creating a closed loop in the j_{p^i} representation at the boundary of the corresponding plaquette (**Figures 5 and 6**).

One can introduce a (nonphysical) time parameter that works simply as a coordinate providing the means of organizing the series of actions of plaquette loop operators in [20]; that is, one assumes that each of the loop actions occurs at different "times." We have introduced an auxiliary time slicing (arbitrary parametrization). If one inserts the AL partition of unity

$$\mathbb{1} = \sum_{\gamma \in \Sigma} \sum_{\{j\}_\gamma} |\gamma, \{j\} >< \gamma, \{j\}| \qquad [21]$$

where the sum is over the complete basis of spin network states $\{|\gamma, \{j\}>\}$ – based on all graphs $\gamma \in \Sigma$ and with all possible spin labeling – between each time

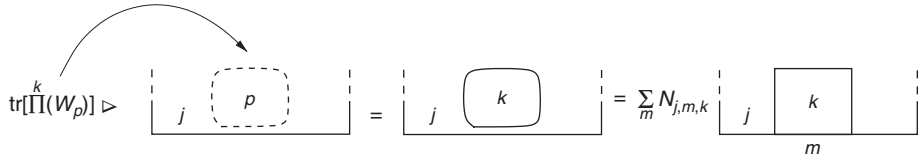


Figure 5 Graphical notation representing the action of one plaquette holonomy on a spin network state. On the right is the result written in terms of the spin network basis. The amplitude $N_{j,m,k}$ can be expressed in terms of Clebsch–Gordan coefficients.

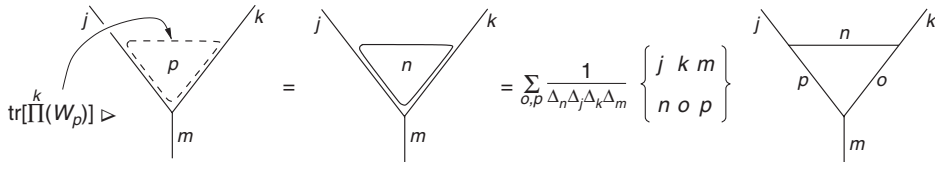


Figure 6 Graphical notation representing the action of one plaquette holonomy on a spin network vertex. The object in brackets ({}) is a 6j-symbol and $\Delta_j := 2j+1$.

slice, one arrives at a sum over spin network histories representation of $P(s)$. More precisely, $P(s)$ can be expressed as a sum over amplitudes corresponding to a series of transitions that can be viewed as the "time evolution" between the "initial" spin network s and the "final" "vacuum state" Ω. The physical inner product between spin networks s and s' is defined as

$$<s, s'>_{\mathrm{p}} := P(s^* s')$$

and can be expressed as a sum over amplitudes corresponding to transitions interpolating between the "initial" spin network s' and the "final" spin network s (e.g., Figures 7 and 8).

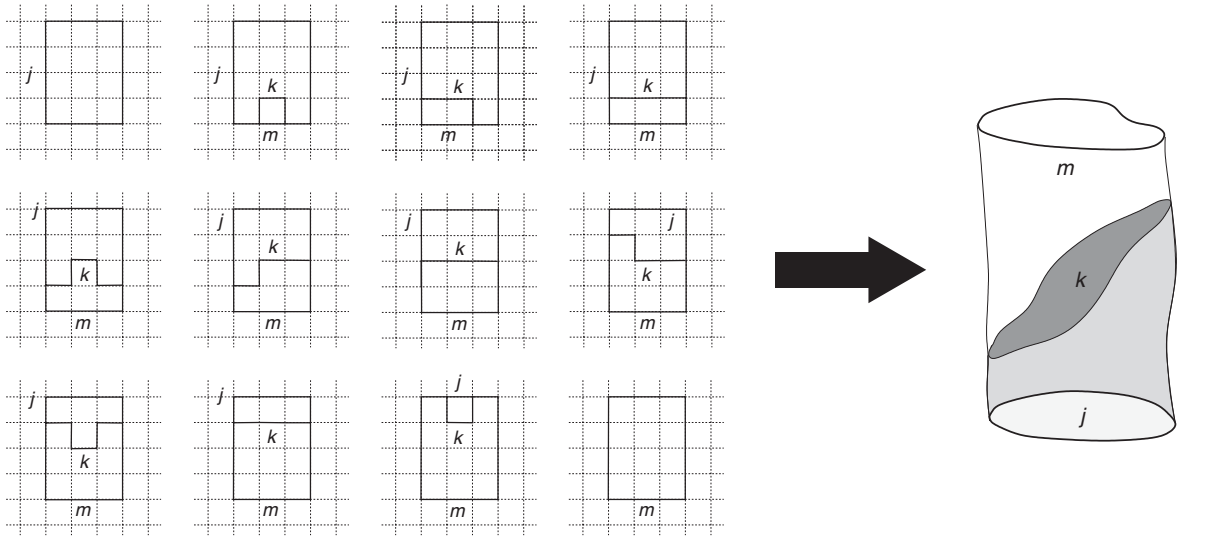


Figure 7 A set of discrete transitions in the loop-to-loop physical inner product obtained by a series of transitions as in Figure 5. On the right, the continuous spin foam representation in the limit $\epsilon \to 0$.

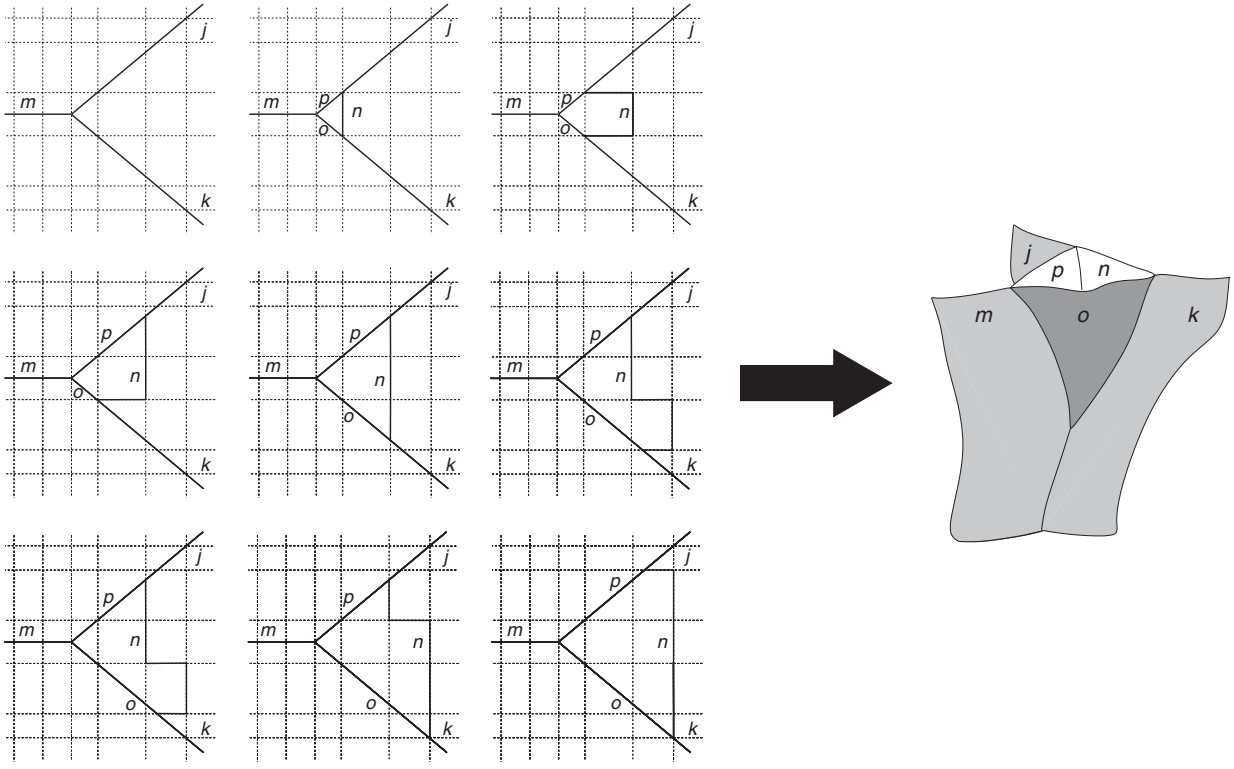
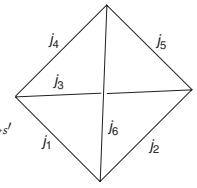


Figure 8 A set of discrete transitions representing one of the contributing histories at a fixed value of the regulator. On the right, the continuous spin foam representation when the regulator is removed.

Spin network nodes evolve into edges while spin network links evolve into 2D faces. Edges inherit the intertwiners associated to the nodes and faces inherit the spins associated to links. Therefore, the series of transitions can be represented by a 2-complex whose 1-cells are labeled by intertwiners and whose 2-cells are labeled by spins. The places where the action of the plaquette loop operators create new links (Figures 6 and 8) define 0-cells or vertices. These foam-like structures are the so-called spin foams. The spin foam amplitudes are purely combinatorial and can be explicitly computed from the simple action of the loop operator in the AL representation (*see* Loop Quantum Gravity). A particularly simple case arises when the spin network states s and s' have only 3-valent nodes. Explicitly,

$$<s, s'>_{\mathrm{p}} := P(s^* s')$$

$$= \sum_{\{j\}} \prod_{f \in F_{s \to s'}} (2j_f + 1)^{\frac{\nu_f}{2}} \prod_{v \in F_{s \to s'}} \quad [22]$$



where the notation is that of [4], and $\nu_f = 0$ if $f \cap s \neq 0 \wedge f \cap s' \neq 0, \nu_f = 1$ if $f \cap s \neq 0 \vee f \cap s' \neq 0$, and $\nu_f = 2$ if $f \cap s = 0 \wedge f \cap s' = 0$. The tetrahedral diagram denotes a $6j$-symbol: the amplitude obtained by means of the natural contraction of the four intertwiners corresponding to the 1-cells converging at a vertex. More generally, for arbitrary spin networks, the vertex amplitude corresponds to $3nj$-symbols, and $<s, s'>_{\mathrm{p}}$ takes the general form [4].

Spin Foams from the Covariant Path Integral

In this section we re-derive the spin foam representation of the physical scalar product of $2+1$ (Riemannian) quantum gravity directly as a regularization of the covariant path integral. The formal path integral for 3D gravity can be written as

$$P = \int D[e] D[A] \exp\left[\mathrm{i} \int_M \mathrm{tr}[e \wedge F(A)]\right] \quad [23]$$

Assume $M = \Sigma \times I$, where $I \subset \mathbb{R}$ is a closed (time) interval (for simplicity, we ignore boundary terms).

In order to give a meaning to the formal expression above, one replaces the 3D manifold (with boundary) M with an arbitrary cellular decomposition Δ. One also needs the notion of the associated dual 2-complex of Δ denoted by Δ^*. The dual 2-complex Δ^* is a combinatorial object defined by a set of vertices $v \in \Delta^*$ (dual to 3-cells in Δ), edges $e \in \Delta^*$ (dual to 2-cells in Δ), and faces $f \in \Delta^*$ (dual to 1-cells in Δ). The intersection of the dual 2-complex Δ^* with the boundaries defines two graphs $\gamma_1, \gamma_2 \in \Sigma$ (see Figure 9). For simplicity, we ignore the boundaries until the end of this section. The fields e and A are discretized as follows. The su(2)-valued 1-form field e is represented by the assignment of $e_f \in \mathrm{su}(2)$ to each 1-cell in Δ. We use the fact that faces in Δ^* are in one-to-one correspondence with 1-cells in Δ and label e_f with a face subindex (Figure 9). The connection field A is represented by the assignment of group elements $g_e \in \mathrm{SU}(2)$ to each edge in $e \in \Delta^*$ (see Figure 10).

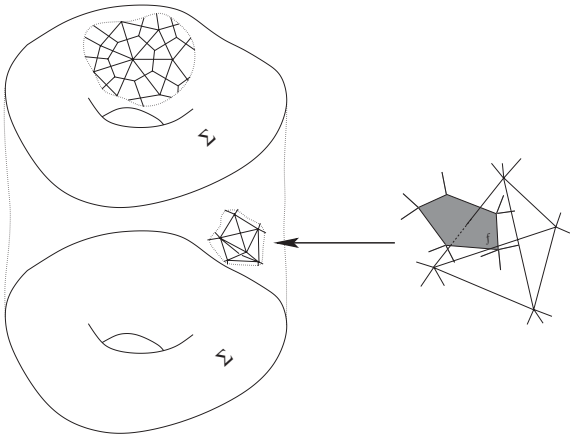


Figure 9 The cellular decomposition of $M = \Sigma \times I$ ($\Sigma = T^2$ in this example). The illustration shows part of the induced graph on the boundary and the detail of a tetrahedron in Δ and a face $f \in \Delta^*$ in the bulk.

With all this, [23] becomes the regularized version P_Δ defined as

$$P_\Delta = \int \prod_{f \in \Delta^*} \mathrm{d}e_f \prod_{e \in \Delta^*} \mathrm{d}g_e \exp\left[\mathrm{i}\ \mathrm{tr}\left[e_f W_f\right]\right] \quad [24]$$

where $\mathrm{d}e_f$ is the regular Lebesgue measure on $\mathbb{R}^3$, $\mathrm{d}g_e$ is the Haar measure on SU(2), and W_f denotes the holonomy around (spacetime) faces, that is, $W_f = g_e^1 \cdots g_e^N$ for N being the number of edges bounding the corresponding face (see Figure 10). The discretization procedure is reminiscent of the one used in standard lattice gauge theory (*see* Lattice

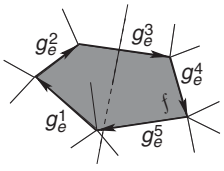


Figure 10 A (2-cell) face $f \in \Delta^*$ in a cellular decomposition of the spacetime manifold M and the corresponding dual 1-cell. The connection field is discretized by the assignment of the parallel transport group elements $g_e^i \in \mathrm{SU}(2)$ to edges $e \in \Delta^*$ ($i = 1, \ldots, 5$ in the face shown here).

Gauge Theory). The previous definition can be motivated by an analysis equivalent to the one presented in [16].

Integrating over e_f, and using [19], one obtains

$$P_\Delta = \sum_{\{j\}} \int \prod_{e\in\Delta^*} \mathrm{d}g_e \prod_{f\in\Delta^*} (2j_f+1) \times \mathrm{tr}\left[\overset{j_f}{\Pi}(g_e^1 \dots g_e^N)\right] \quad [25]$$

Now it remains to integrate over the lattice connection $\{g_e\}$. If an edge $e \in \Delta^*$ bounds n faces $f \in \Delta^*$ there will be n traces of the form $\mathrm{tr}[\Pi^{j_f}(\cdots g_e \cdots)]$ in [25] containing g_e in the argument. In order to integrate over g_e we can use the following identity:

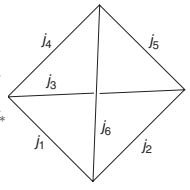
$$I^n_{\mathrm{inv}} := \int \mathrm{d}g \overset{j_1}{\Pi}(g) \otimes \overset{j_2}{\Pi}(g) \otimes \cdots \otimes \overset{j_n}{\Pi}(g) = \sum_\iota C^\iota_{j_1 j_2 \cdots j_n} \, C^{*\iota}_{j_1 j_2 \cdots j_n} \quad [26]$$

where I^n_{inv} is the projector from the tensor product of irreducible representations $\mathcal{H}_{j_1\cdots j_n} = j_1 \otimes j_2 \otimes \cdots \otimes j_n$ onto the invariant component $\mathcal{H}^0_{j_1\cdots j_n} = \mathrm{Inv}[j_1 \otimes j_2 \otimes \cdots \otimes j_n]$. On the right-hand side, we have chosen an orthonormal basis of invariant vectors (intertwiners) in $\mathcal{H}_{j_1\cdots j_n}$ to express the projector. Notice that the assignment of intertwiners to edges is a consequence of the integration over the connection. Using [26] one can write P_Δ in the general spin foam representation form [4]

$$P_\Delta = \sum_{\{f\}} \prod_{f\in\Delta^*} (2j_f+1) \prod_{v\in\Delta^*} A_v(j_v) \quad [27]$$

where $A_v(\iota_v, j_v)$ is given by the appropriate trace of the intertwiners corresponding to the edges bounded by the vertex. As in the previous section, this amplitude is given in general by an SU(2) $3Nj$-symbol. When Δ is a simplicial complex, all the edges in Δ^* are 3-valent and vertices are 4-valent. Consequently, the vertex amplitude is given by the contraction of the corresponding four 3-valent intertwiners, that is, a $6j$ symbol. In that case, the path integral takes the (Ponzano–Regge) form

$$P_\Delta = \sum_{\{j\}} \prod_{f\in\Delta^*} (2j_f+1) \prod_{v\in\Delta^*} \quad [28]$$



The labeling of faces that intersect the boundary naturally induces a labeling of the edges of the graphs γ_1 and γ_2 induced by the discretization. Thus, the boundary states are given by spin network states on γ_1 and γ_2, respectively. A careful analysis of the boundary contribution shows that only the face amplitude is modified to $(\Delta_{j_\ell})^{\nu_f/2}$, and that the spin foam amplitudes are as in eqn [22].

A crucial property of the path integral in 3D gravity (and of the transition amplitudes in general) is that it does not depend on the discretization Δ – this is due to the absence of local degrees of freedom in 3D gravity and not expected to hold in 4D. Given two different cellular decompositions Δ and Δ', one has

$$\tau^{-n_0} P_\Delta = \tau^{-n'_0} P_{\Delta'} \quad [29]$$

where n_0 is the number of 0-simplexes in Δ, and $\tau = \sum_j (2j+1)^2$. As τ is given by a divergent sum, the discretization independence statement is formal. Moreover, the sum over spins in [28] is typically divergent. Divergences occur due to infinite gauge-volume factors in the path integral corresponding to the topological gauge freedom [7]. Freidel and Louapre have shown how these divergences can be avoided by gauge-fixing unphysical degrees of freedom in [24]. In the case of 3D gravity with positive cosmological constant, the state sum generalizes to the Turaev–Viro invariant (*see* Topological Quantum Field Theory: Overview) defined in terms of the quantum group $\mathrm{SU}_q(2)$ with $q^n = 1$ where the representations are finitely many and thus $\tau < \infty$. Equation [29] is a rigorous statement in that case. No such infrared divergences appear in the canonical treatment of the previous section.

Spin Foams in 4D

Spin Foam from the Canonical Formulation

There is no rigorous construction of the physical inner product of LQG in 4D. The spin foam representation as a device for its definition has been introduced formally by Rovelli. In 4D LQG, difficulties in understanding dynamics are centered around the quantum scalar constraint $\widehat{S} = \widehat{\sqrt{\det E}^{-1} E^a_i E^b_j F^{ij}_{ab}}(A) + \cdots$ (see [1]) – the vector constraint $\widehat{V}_a(A,E)$ is solved in a simple manner (*see* Loop Quantum Gravity). The physical inner product formally becomes

$$\langle Ps, s'\rangle_{\mathrm{diff}} = \prod_x \delta[\widehat{S}(x)] = \int D[N] < \exp\left[\mathrm{i}\int_\Sigma N(x)\widehat{S}(x)\right] s, s' >_{\mathrm{diff}} = \int D[N] \sum_{n=0}^\infty \frac{i^n}{n!} < \left[\int_\Sigma N(x)\widehat{S}(x)\right]^n s, s' >_{\mathrm{diff}} \quad [30]$$

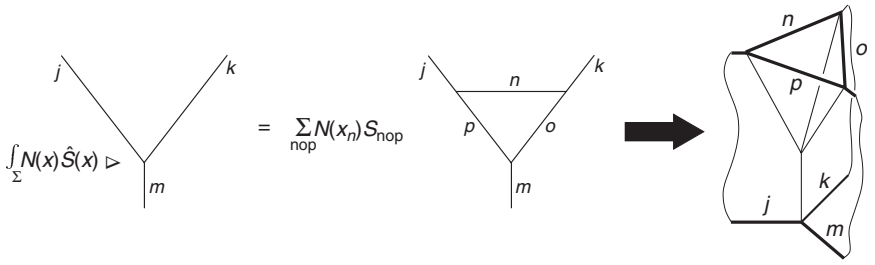


Figure 11 The action of the scalar constraint and its spin foam representation. $N(x_n)$ is the value of N at the node and S_{nop} are the matrix elements of $\widehat{S}$.

where $<,>_{\text{diff}}$ denotes the inner product in the Hilbert space of solutions of the vector constraint, and the exponential has been expanded in powers in the last expression on the right-hand side.

From early on, it was realized that smooth loop states are naturally annihilated by $\widehat{S}$ (independently of any quantization ambiguity). Consequently, $\widehat{S}$ acts only on spin network nodes. Generically, it does so by creating new links and nodes modifying the underlying graph of the spin network states (**Figure 11**).

Therefore, each term in the sum [30] represents a series of transitions – given by the local action of $\widehat{S}$ at spin network nodes – through different spin network states interpolating the boundary states s and s', respectively. The action of $\widehat{S}$ can be visualized as an "interaction vertex" in the "time" evolution of the node (**Figure 11**). As in the explicit 3D case, eqn [30] can be expressed as sum over "histories" of spin networks pictured as a system of branching surfaces described by a 2-complex whose elements inherit the representation labels on the intermediate states. The value of the "transition" amplitudes is controlled by the matrix elements of $\widehat{S}$. Therefore, although the qualitative picture is independent of quantization ambiguities, transition amplitudes are sensitive to them.

Before even considering the issue of convergence of [30], the problem with this definition is evident: every single term in the sum is a divergent integral! Therefore, this way of presenting spin foams has to be considered as formal until a well-defined regularization of [2] is provided. That is the goal of the spin foam approach.

Instead of dealing with an infinite number of constraints Thiemann recently proposed to impose one single master constraint defined as

$$M = \int_\Sigma \mathrm{d}x^3 \; \frac{S^2(x) - q^{ab} V_a(x) V_b(x)}{\sqrt{\det q(x)}} \qquad [31]$$

Using techniques developed by Thiemann, this constraint can indeed be promoted to a quantum operator acting on $\mathcal{H}_{\text{kin}}$. The physical inner product is given by

$$<s, s'>_{\text{p}} := \lim_{T\to\infty} <s, \int_{-T}^{T} \mathrm{d}t \; \mathrm{e}^{\mathrm{i}t\widehat{M}} s'> \qquad [32]$$

A spin foam representation of the previous expression could now be achieved by the standard skeletonization that leads to the path-integral representation in quantum mechanics. In this context, one splits the t-parameter in discrete steps and writes

$$\mathrm{e}^{\mathrm{i}t\widehat{M}} = \lim_{N\to\infty} [\mathrm{e}^{\mathrm{i}t\widehat{M}/N}]^N = \lim_{N\to\infty} [1 + \mathrm{i}t\widehat{M}/N]^N \qquad [33]$$

The spin foam representation follows from the fact that the action of the basic operator $1 + \mathrm{i}t\widehat{M}/N$ on a spin network can be written as a linear combination of new spin networks whose graphs and labels have been modified by the creation of new nodes (in a way qualitatively analogous to the local action shown in **Figure 11**). An explicit derivation of the physical inner product of 4D LQG along these lines is under current investigation.

Spin Foams from the Covariant Formulation

In 4D, the spin foam representation of the dynamics of LQG has been investigated more intensively in the covariant formulation. This has led to a series of constructions which are referred to as spin foam models. These treatments are related more closely to the construction based on the covariant path-integral approach of the last section. Here we illustrate the formulation which has captured much interest in the literature: the Barrett–Crane (BC) model.

Spin foam models for gravity as constrained quantum BF theory The BC model is one of the most extensively studied spin foam models for quantum gravity. To introduce the main ideas involved, we concentrate on the definition of the model in the Riemannian sector. The BC model can be formally

viewed as a spin foam quantization of SO(4) Plebanski's formulation of general relativity. Plebanski's Riemannian action depends on an SO(4) connection A, a Lie-algebra-valued 2-form B, and Lagrange multiplier fields λ and μ. Writing explicitly the Lie algebra indices, the action is given by

$$S[B,A,\lambda,\mu] = \int \left[B^{IJ}\wedge F_{IJ}(A) + \lambda_{IJKL}\, B^{IJ}\wedge B^{KL} + \mu\epsilon^{IJKL}\lambda_{IJKL}\right] \quad [34]$$

where μ is a 4-form and $\lambda_{IJKL} = -\lambda_{JIKL} = -\lambda_{IJLK} = \lambda_{KLIJ}$ is a tensor in the internal space. Variation with respect to μ imposes the constraint $\epsilon^{IJKL}\lambda_{IJKL} = 0$ on λ_{IJKL}. The Lagrange multiplier tensor λ_{IJKL} has then 20 independent components. Variation with respect to λ imposes 20 algebraic equations on the 36 components of B. The (non-degenerate) solutions to the equations obtained by varying the multipliers λ and μ are

$$B^{IJ} = \pm\epsilon^{IJKL}e_K\wedge e_L$$

and

$$B^{IJ} = \pm e^I\wedge e^J \quad [35]$$

in terms of the 16 remaining degrees of freedom of the tetrad field e^I_a. If one substitutes the first solution into the original action, one obtains Palatini's formulation of general relativity; therefore, on shell (and on the right sector), the action is that of classical gravity.

The key idea in the definition of the model is that the path integral for the theory corresponding to the action $S[B,A,0,0]$, namely

$$P_{\text{topo}} = \int D[B]D[A]\exp\left[\mathrm{i}\int\left[B^{IJ}\wedge F_{IJ}(A)\right]\right] \quad [36]$$

can be given a meaning as a spin foam sum, [4], in terms of a simple generalization of the construction of the previous section. In fact, $S[B,A,0,0]$ corresponds to a simple theory known as BF theory that is formally very similar to 3D gravity (*see* BF Theories). The result is independent of the chosen discretization because BF theory does not have local degrees of freedom (just as 3D gravity).

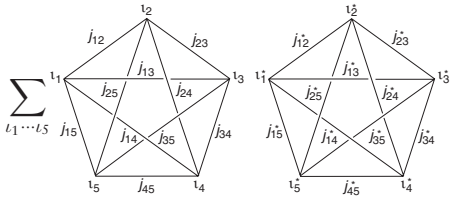
The BC model aims at providing a definition of the path integral of gravity pursuing a well-posed definition of the formal expression

$$P_{\text{GR}} = \int D[B]D[A]\delta\left[B\to\epsilon^{IJKL}e_K\wedge e_L\right] \times\exp\left[\mathrm{i}\int\left[B^{IJ}\wedge F_{IJ}(A)\right]\right] \quad [37]$$

where $D[B]D[A]\delta(B\to\epsilon^{IJKL}e_K\wedge e_L)$ means that one must restrict the sum in [36] to those configurations of the topological theory satisfying the constraints $B = *(e\wedge e)$ for some tetrad e. The remarkable fact is that this restriction can be implemented in a systematic way directly on the spin foam configurations that define P_{topo}.

In P_{topo} spin foams are labeled with spins corresponding to the unitary irreducible representations of SO(4) (given by two spin quantum numbers $(j_{\text{R}}, j_{\text{L}})$). Essentially, the factor "$\delta(B\to\epsilon^{IJKL}e_K\wedge e_L)$" restricts the set of spin foam quantum numbers to the so-called simple representations (for which $j_{\text{R}} = j_{\text{L}} = j$). This is the "quantum" version of the solution to the constraints [35]. There are various versions of this model. The simplest definition of the transition amplitudes in the BC model is given by

$$P(s^*s) = \sum_{\{j\}}\prod_{f\in F_{s\to s'}}(2j_f+1)^{\nu_f}\prod_{v\in F_{s\to s'}}$$



[38]

where we use the notation of [22], the graphs denote $15j$-symbols, and ι_i are half-integers labeling SU(2) normalized 4-intertwiners. No rigorous connection with the Hilbert space picture of LQG has yet been established. The self-dual version of Plebanski's action leads, through a similar construction, to Reisenberger's model.

The simplest amplitude in the BC model corresponds to a single 4-simplex, which can be viewed as the simplest triangulation of the 4D spacetime given by the interior of a 3-sphere (the corresponding 2-complex is shown in **Figure 12**). States of the 4-simplex are labeled by ten spins j (labeling the ten edges of the boundary spin network, see **Figure 12**) which can be shown to be related to the area in

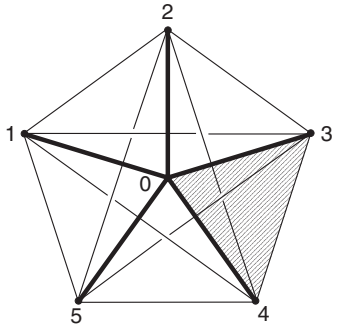


Figure 12 The dual of a 4-simplex.

Planck units of the ten triangular faces that form the 4-simplex. A first indication of the connection of the model with gravity was that the large-j asymptotics appeared to be dominated by the exponential of the Regge action (the action derived by Regge as a discretization of general relativity). This estimate was done using the stationary-phase approximation to the integral that gives the amplitude of a 4-simplex in the BC model. However, more detailed calculations showed that the amplitude is dominated by configurations corresponding to degenerate 4-simplexes. This seems to invalidate a simple connection to general relativity and is one of the main puzzles in the model.

Spin Foams as Feynman Diagrams

The main problem with the models of the previous section is that they are defined on a discretization Δ of M and that – contrary to what happens with a topological theory, for example, 3D gravity (eqn [29]) – the amplitudes depend on the discretization Δ. Various possibilities to eliminate this regulator have been discussed in the literature but no explicit results are yet known in 4D. An interesting proposal is a discretization-independent definition of spin foam models achieved by the introduction of an auxiliary field theory living on an abstract group manifold – $\mathrm{Spin}(4)^4$ and $\mathrm{SL}(2,C)^4$ for Riemannian and Lorentzian gravity, respectively. The action of the auxiliary group field theory (GFT) takes the form

$$S[\phi] = \int_{G^4} \phi^2 + \frac{\lambda}{5!}\int_{G^{10}} M^{(5)}[\phi] \qquad [39]$$

where $M^{(5)}[\phi]$ is a fifth-order monomial, and G is the corresponding group. In the simplest model, $M^{(5)}[\phi] = \phi(g_1,g_2,g_3,g_4)\phi(g_4,g_5,g_6,g_7)\times \phi(g_7,g_3,g_8,g_9)\phi(g_9,g_6,g_2,g_{10})\phi(g_{10},g_8,g_5,g_1)$. The field ϕ is required to be invariant under the (simultaneous) right action of the group on its four arguments in addition to other symmetries (not described here for simplicity). The perturbative expansion in λ of the GFT Euclidean path integral is given by

$$P = \int D[\phi]\mathrm{e}^{-S[\phi]} = \sum_{F_N} \frac{\lambda^N}{\mathrm{sym}[F_N]} A[F_N] \qquad [40]$$

where $A[F_N]$ corresponds to a sum of Feynman-diagram amplitudes for diagrams with N interaction vertices, and $\mathrm{sym}[F_N]$ denotes the standard symmetry factor. A remarkable property of this expansion is that $A[F_N]$ can be expressed as a sum over spin foam amplitudes, that is, 2-complexes labeled by unitary irreducible representations of G. Moreover, for very simple interaction $M^{(5)}[\phi]$, the spin foam amplitudes are in one-to-one correspondence to those found in the models of the previous section (e.g., the BC model). This duality is regarded as a way of providing a fully combinatorial definition of quantum gravity where no reference to any discretization or even a manifold structure is made. Transition amplitudes between spin network states correspond to n-point functions of the field theory. These models have been inspired by generalizations of matrix models applied to BF theory.

Divergent transition amplitudes can arise by the contribution of "loop" diagrams as in standard quantum field theory. In spin foams, diagrams corresponding to 2D bubbles are potentially divergent because spin labels can be arbitrarily high leading to unbounded sums in [4]. Such divergences do not occur in certain field theories dual (in the sense above) to the BC model. However, little is known about the convergence of the series in λ and the physical meaning of this constant. Nevertheless, Freidel and Louapre have shown that the series can be re-summed in certain models dual to lower-dimensional theories.

Causal Spin Foams

Let us conclude by presenting a fundamentally different construction leading to spin foams. Using the kinematical setting of LQG with the assumption of the existence of a microlocal (in the sense of Planck scale) causal structure, Markopoulou and Smolin define a general class of (causal) spin foam models for gravity. The elementary transition amplitude $A_{s_I \to s_{I+1}}$ from an initial spin network s_I to another spin network s_{I+1} is defined by a set of simple combinatorial rules based on a definition of causal propagation of the information at nodes. The rules and amplitudes have to satisfy certain causal restrictions (motivated by the standard concepts in classical Lorentzian physics). These rules generate surface-like excitations of the same kind one encounters in the previous formulations. Spin foams $F^N_{s_i \to s_f}$ are labeled by the number of times, N, these elementary transitions take place. Transition amplitudes are defined as

$$\langle s_i, s_f\rangle = \sum_N A(F^N_{s_i \to s_f}) \qquad [41]$$

which is of the generic form [4]. The models are not related to any continuum action. The only guiding principles in the construction are the restrictions imposed by causality, and the requirement of the existence of a nontrivial critical behavior that reproduces general relativity at large scales. Some indirect evidence of a possible nontrivial continuum limit has been obtained in certain versions of these models in $1+1$ dimensions.

See also: Algebraic Approach to Quantum Field Theory; BF Theories; Canonical General Relativity; Chern–Simons Models: Rigorous Results; Lattice Gauge Theory; Loop Quantum Gravity; Quantum Dynamics in Loop Quantum Gravity; Quantum Geometry and its Applications; Topological Quantum Field Theory: Overview.

Further Reading

Ashtekar A (1991) *Lectures on Nonperturbative Canonical Gravity*. Singapore: World Scientific.

Ashtekar A and Lewandowski J (2004) Background independent quantum gravity: a status report.

Baez J (1998) Spin foams. *Classical and Quantum Gravity* 15: 1827–1858.

Baez J (2000) An introduction to spin foam models of quantum gravity and BF theory. *Lecture Notes in Physics* 543: 25–94.

Baez J and Muniain JP (1995) *Gauge fields, Knots and Gravity*. Singapore: World Scientific.

Oriti D (2001) Spacetime geometry from algebra: spin foam models for nonperturbative quantum gravity. *Reports on Progress in Physics* 64: 1489–1544.

Perez A (2003) Spin foam models for quantum gravity. *Classical and Quantum Gravity* 20: R43.

Rovelli C *Quantum Gravity*. Cambridge: Cambridge University Press (to appear).

Thiemann T *Modern Canonical Quantum General Relativity*. Cambridge: Cambridge University Press (to appear).

Spin Glasses

F Guerra, Università di Roma "La Sapienza", Rome, Italy

Introduction

From a physical point of view, spin glasses, as dilute magnetic alloys, are very interesting systems. They are characterized by such features as exhibiting a new magnetic phase, where magnetic moments are frozen into disordered equilibrium orientations, without any long-range order. See, for example, Young (1987) for general reviews, and also Stein (1989) for a very readable account about the physical properties of spin glasses. The experimental laboratory study of spin glasses is a very difficult subject, because of their peculiar properties. In particular, the existence of very slowly relaxing modes, with consequent memory effects, makes it difficult to realize the very basic physical concept of a system at thermodynamical equilibrium, at a given temperature.

From a theoretical point of view some models have been proposed, which try to capture the essential physical features of spin glasses, in the frame of very simple assumptions.

The basic model has been proposed by Edwards and Anderson (1975) many years ago. It is a simple extension of the well-known nearest-neighbor Ising model. On a large region Λ of the unit lattice in d dimensions, we associate an Ising spin $\sigma(n)$ to each lattice site n, and then we introduce a lattice Hamiltonian

$$H_\Lambda(\sigma, J) = -\sum_{(n,n')} J(n,n')\sigma(n)\sigma(n') \qquad [1]$$

Here, the sum runs over all couples of nearest-neighbor sites in Λ, and J are quenched random couplings, assumed for simplicity to be independent identically distributed random variables, with centered unit Gaussian distribution. The quenched character of the J means that they do not contribute to thermodynamic equilibrium, but act as a kind of random external noise on the coupling of the σ variables. In the expression of the Hamiltonian, we have indicated with σ the set of all $\sigma(n)$, and with J the set of all $J(n,n')$. The region Λ must be taken very large, by letting it invade all lattice in the limit. The physical motivation for this choice is that for real spin glasses the interaction between the spins dissolved in the matrix of the alloy oscillates in sign according to distance. This effect is taken into account in the model through the random character of the couplings between spins.

Even though very drastic simplifications have been introduced in the formulation of this model, as compared to the extremely complicated nature of physical spin glasses, nevertheless a rigorous study of all properties emerging from the static and dynamic behavior of a thermodynamic system of this kind is far from being complete. In particular, with reference to static equilibrium properties, it is not yet possible to reach a completely substantiated description of the phases emerging in the low-temperature region. Even physical intuition gives completely different guesses for different people.

In the same way as a mean-field version can be associated to the ordinary Ising model, so it is possible for the disordered model described by [1]. Now we consider a number of sites $i = 1, 2, \ldots, N$, and let each spin $\sigma(i)$ at site i interact with all other spins, with the intervention of a quenched noise J_{ij}. The precise form of the Hamiltonian will be given in the following.

This is the mean-field model for spin glasses, introduced by Sherrington and Kirkpatrick (1975).

It is a celebrated model. Numerous articles have been devoted to its study during the years, appearing in the theoretical physics literature.

The relevance of the model stems surely from the fact that it is intended to represent some important features of the physical spin glass systems, of great interest for their peculiar properties, at least at the level of the mean-field approximation.

But another important source of interest is connected with the fact that disordered systems, of the Sherrington–Kirkpatrick type, and their generalizations, seem to play a very important role for theoretical and practical assessments about hard optimization problems, as it is shown, for example, by Mézard *et al.* (2002).

It is interesting to remark that the original paper was entitled "Solvable model of a spin-glass," while a previous draft, as told by David Sherrington, contained the even stronger designation "Exactly solvable." However, it turned out that the very natural solution devised by the authors is valid only at high temperatures, or for large external magnetic fields. At low temperatures, the proposed solution exhibits a nonphysical drawback given by a negative entropy, as properly recognized by the authors in their very first paper.

It took some years to find an acceptable solution. This was done by Giorgio Parisi in a series of papers, marking a radical departure from the previous methods. In fact, a very intense method of "spontaneous replica symmetry breaking" was developed. As a consequence, the physical content of the theory was encoded in a functional order parameter of new type, and a remarkable structure emerged for the pure states of the theory, a kind of hierarchical, ultrametric organization. These very interesting developments, due to Parisi, and his coworkers, are explained in a brilliant way in the classical book by Mézard *et al.* (1987). Part of this structure will be recalled in the following.

It is important to remark that the Parisi solution is presented in the form of an ingenious and clever "ansatz." Until few years ago, it was not known whether this ansatz would give the true solution for the model, in the so-called thermodynamic limit, when the size of the system becomes infinite, or it would be only a very good approximation for the true solution.

The general structures offered by the Parisi solution, and their possible generalizations for similar models, exhibit an extremely rich and interesting mathematical content. Very appropriately, Talagrand (2003) has used a strongly suggestive sentence in the title to his recent book: "Spin glasses: a challenge for mathematicians."

As a matter of fact, how to face this challenge is a very difficult problem. Here we would like to recall the main features of a very powerful method, yet extremely simple in its very essence, based on a comparison and interpolation argument on sets of Gaussian random variables.

The method found its first simple application in Guerra (2001), where it was shown that the Sherrington–Kirkpatrick replica symmetric approximate solution was a rigorous lower bound for the quenched free energy of the system, uniformly in the size. Then, it was possible to reach a long-awaited result (Guerra and Toninelli 2002): the convergence of the free energy density in the thermodynamic limit, by an intermediate step where the quenched free energy was shown to be subadditive in the size of the system.

Moreover, still by interpolation on families of Gaussian random variables, the first mentioned result was extended to give a rigorous proof that the expression given by the Parisi ansatz is also a lower bound for the quenched free energy of the system, uniformly in the size (Guerra 2003). The method gives not only the bound, but also the explicit form of the correction in a complex form. As a recent and very important result, along the task of facing the challenge, Michel Talagrand has been able to dominate these correction terms, showing that they vanish in the thermodynamic limit. This milestone achievement was first announced in a short note, containing only a synthetic sketch of the proof, and then presented with all details in a long paper (Talagrand 2006).

The interpolation method is also at the basis of the far-reaching generalized variational principle proved by Aizenman *et al.* (2003).

In our presentation, we will try to be as self-contained as possible. We will give all definitions, explain the basic structure of the interpolation method, and show how some of the results are obtained. We will concentrate mostly on questions connected with the free energy, its properties of subadditivity, the existence of the infinite-volume limit, and the replica bounds.

For the sake of comparison, and in order to provide a kind of warm-up, we will recall also some features of the standard elementary mean-field model of ferromagnetism, the so-called Curie–Weiss model. We will concentrate also here on the free energy, and systematically exploit elementary comparison and interpolation arguments. This will show the strict analogy between the treatment of the ferromagnetic model and the developments in the mean-field spin glass case. Basic roles will be played in the two cases, but with different expressions, by positivity and convexity properties.

Then, we will consider the problem of connecting results for the mean-field case to the short-range case. An intermediate position is occupied by the so-called diluted models. They can be studied through a generalization of the methods exploited in the mean-field case, as shown, for example, in De Sanctis (2005).

The organization of the paper is as follows. We first introduce the ferromagnetic model and discuss behavior and properties of the free energy in the thermodynamic limit, by emphasizing, in this very elementary case, the comparison and interpolation methods that will be also exploited, in a different context, in the spin glass case.

The basic features of the mean-field spin glass models are discussed next, by introducing all necessary definitions. This is followed by the introduction, for generic Gaussian interactions, of some important formulas, concerning the derivation with respect to the strength of the interaction, and the Gaussian comparison and interpolation method.

We then give simple applications to the mean-field spin glass model, in particular to the existence of the infinite-volume limit of the quenched free energy (Guerra and Toninelli 2002), and to the proof of general variational bounds, by following the useful strategy developed in Aizenman *et al.* (2003).

The main features of the Parisi representation are recalled briefly, and the main theorem concerning the free energy is stated. This is followed by a brief mention of results for diluted models.

We also attack the problem of connecting the results for the mean-field case to the more realistic short-range models.

Finally we provide conclusions and outlook for future foreseen developments.

Our treatment will be as simple as possible, by relying on the basic structural properties, and by describing methods of presumably very long lasting power. The emphasis given to the mean-field case reflects the status of research. After some years from now this review would perhaps be written according to completely different patterns.

A Warm-up. The Mean-field Ferromagnetic Model: Structure and Results

The mean-field ferromagnetic model is among the simplest models of statistical mechanics. However, it contains very interesting features, in particular a phase transition, characterized by spontaneous magnetization, at low temperatures. We refer to standard textbooks for a full treatment and a complete appreciation of the model in the frame of the theory of ferromagnetism. Here we first consider some properties of the free energy, easily obtained through comparison methods.

The generic configuration of the mean-field ferromagnetic model is defined through Ising spin variables $\sigma_i = \pm 1$, attached to each site $i = 1, 2, \ldots, N$.

The Hamiltonian of the model, in some external field of strength h, is given by the mean-field expression

$$H_N(\sigma, h) = -\frac{1}{N}\sum_{(i,j)} \sigma_i\sigma_j - h\sum_i \sigma_i \qquad [2]$$

Here, the first sum extends to all $N(N-1)/2$ site couples, and the second to all sites.

For a given inverse temperature β, let us now introduce the partition function $Z_N(\beta, h)$ and the free energy per site $f_N(\beta, h)$, according to the well-known definitions

$$Z_N(\beta, h) = \sum_{\sigma_1 \ldots \sigma_N} \exp(-\beta H_N(\sigma, h)) \qquad [3]$$

$$-\beta f_N(\beta, h) = N^{-1} E \log Z_N(\beta, h) \qquad [4]$$

It is also convenient to define the average spin magnetization

$$m = \frac{1}{N}\sum_i \sigma_i \qquad [5]$$

Then, it is immediately seen that the Hamiltonian in [2] can be equivalently written as

$$H_N(\sigma, h) = -\frac{1}{2}Nm^2 - h\sum_i \sigma_i \qquad [6]$$

where an unessential constant term has been neglected. In fact, we have

$$\sum_{(i,j)} \sigma_i\sigma_j = \frac{1}{2}\sum_{i,j;i\neq j} \sigma_i\sigma_j = \frac{1}{2}N^2m^2 - \frac{1}{2}N \qquad [7]$$

where the sum over all couples has been equivalently written as one half the sum over all i, j with $i \neq j$, and the diagonal terms with $i = j$ have been added and subtracted out. Notice that they give a constant because $\sigma_i^2 = 1$.

Therefore, the partition function in [3] can be equivalently substituted by the expression

$$Z_N(\beta, h) = \sum_{\sigma_1 \ldots \sigma_N} \exp\left(\frac{1}{2}\beta N m^2\right) \exp\left(\beta h \sum_i \sigma_i\right) \qquad [8]$$

which will be our starting point.

Our interest will be in the $\lim_{N\to\infty} N^{-1} \log Z_N(\beta, h)$. To this purpose, let us establish the important subadditivity property, holding for the splitting of the

large-N system in two smaller systems with N_1 and N_2 sites, respectively, with $N = N_1 + N_2$,

$$\log Z_N(\beta, h) \leq \log Z_{N_1}(\beta, h) + \log Z_{N_2}(\beta, h) \quad [9]$$

The proof is very simple. Let us denote, in the most natural way, by $\sigma_1, \ldots, \sigma_{N_1}$ the spin variables for the first subsystem, and by $\sigma_{N_1+1}, \ldots, \sigma_N$ the N_2 spin variables of the second subsystem. Introduce also the subsystem magnetizations m_1 and m_2, by adapting the definition [5] to the smaller systems, in such a way that

$$Nm = N_1 m_1 + N_2 m_2 \quad [10]$$

Therefore, we see that the large system magnetization m is the linear convex combination of the smaller system ones, according to the obvious

$$m = \frac{N_1}{N} m_1 + \frac{N_2}{N} m_2 \quad [11]$$

Since the mapping $m \to m^2$ is convex, we also have the general bound, holding for all values of the σ variables

$$m^2 \leq \frac{N_1}{N} m_1^2 + \frac{N_2}{N} m_2^2 \quad [12]$$

Then, it is enough to substitute the inequality in the definition [8] of $Z_N(\beta, h)$, and recognize that we achieve factorization with respect to the two subsystems, and therefore the inequality $Z_N \leq Z_{N_1} Z_{N_2}$. So we have established [9]. From subadditivity, the existence of the limit follows by standard arguments. In fact, we have

$$\lim_{N\to\infty} N^{-1} \log Z_N(\beta, h) = \inf_N N^{-1} \log Z_N(\beta, h) \quad [13]$$

Now we will calculate explicitly this limit, by introducing an order parameter M, a trial function, and an appropriate variational scheme. In order to get a lower bound, we start from the elementary inequality $m^2 \geq 2mM - M^2$, holding for any value of m and M. By inserting the inequality in the definition [8] we arrive at a factorization of the sum over σ's. The sum can be explicitly calculated, and we arrive immediately to the lower bound, uniform in the size of the system,

$$N^{-1} \log Z_N(\beta, h) \geq \log 2 + \log \cosh \beta(h + M) - \tfrac{1}{2}\beta M^2 \quad [14]$$

holding for any value of the trial order parameter M. Clearly, it is convenient to take the supremum over M. Then, we establish the optimal uniform lower bound

$$N^{-1} \log Z_N(\beta, h) \geq \sup_M \left(\log 2 + \log \cosh \beta(h + M) - \tfrac{1}{2}\beta M^2\right) \quad [15]$$

It is simple to realize that the supremum coincides with the limit as $N \to \infty$. To this purpose we follow the following simple procedure. Let us consider all possible values of the variable m. There are $N + 1$ of them, corresponding to any number K of possible spin flips, starting from a given σ configuration, $K = 0, 1, \ldots, N$. Let us consider the trivial decomposition of the identity, holding for any m,

$$1 = \sum_M \delta_{mM} \quad [16]$$

where M in the sum runs over the $N + 1$ possible values of m, and δ is Kroneker delta, being equal to 1 if $M = N$, and zero otherwise. Let us now insert [16] in the definition [8] of the partition function inside the sum over σ's, and invert the two sums. Because of the forcing $m = M$ given by the δ, we can write $m^2 = 2mM - M^2$ inside the sum. Then if we neglect the δ, by using the trivial $\delta \leq 1$, we have an upper bound, where the sum over σ's can be explicitly performed as before. Then it is enough to take the upper bound with respect to M, and consider that there are $N + 1$ terms in the now trivial sum over M, in order to arrive at the upper bound

$$N^{-1} \log Z_N(\beta, h) \leq \sup_M \left(\log 2 + \log \cosh \beta(h + M) - \tfrac{1}{2}\beta M^2\right) + N^{-1} \log(N + 1) \quad [17]$$

Therefore, by going to the limit as $N \to \infty$, we can collect all our results in the form of the following theorem giving the full characterization of the thermodynamic limit of the free energy.

Theorem 1 *For the mean-field ferromagnetic model we have*

$$\lim_{N\to\infty} N^{-1} \log Z_N(\beta, h) = \inf_N N^{-1} \log Z_N(\beta, h) \quad [18]$$

$$= \sup_M \left(\log 2 + \log \cosh \beta(h + M) - \tfrac{1}{2}\beta M^2\right) \quad [19]$$

This ends our discussion about the free energy in the ferromagnetic model.

Other properties of the model can be easily established. Introduce the Boltzmann–Gibbs state

$$\omega_N(A) = Z_N^{-1} \sum_{\sigma_1 \ldots \sigma_N} A \exp\left(\frac{1}{2}\beta N m^2\right) \exp\left(\beta h \sum_i \sigma_i\right) \quad [20]$$

where A is any function of $\sigma_1 \ldots \sigma_N$.

The observable $m(\sigma)$ becomes self-averaging under ω_N, in the infinite-volume limit, in the sense that

$$\lim_{N\to\infty} \omega_N((m - M(\beta, h))^2) = 0 \quad [21]$$

This property of m is the deep reason for the success of the strategy exploited earlier for the convergence of the free energy. Easy consequences are the following. In the infinite-volume limit, for $h \neq 0$, the Boltzmann–Gibbs state becomes a factor state

$$\lim_{N\to\infty} \omega_N(\sigma_1 \dots \sigma_s) = M(\beta,h)^s \qquad [22]$$

A phase transition appears in the form of spontaneous magnetization. In fact, while for $h=0$ and $\beta \leq 1$ we have $M(\beta,h)=0$, on the other hand, for $\beta > 1$, we have the discontinuity

$$\lim_{h\to 0^+} M(\beta,h) = -\lim_{h\to 0^-} M(\beta,h) \equiv M(\beta) > 0 \qquad [23]$$

Fluctuations can also be easily controlled. In fact, one proves that the rescaled random variable $\sqrt{N}(m - M(\beta,h))$ tends in distribution, under ω_N, to a centered Gaussian with variance given by the susceptibility

$$\chi(\beta,h) \equiv \frac{\partial}{\partial h} M(\beta,h) \equiv \frac{\beta(1-M^2)}{1-\beta(1-M^2)} \qquad [24]$$

Notice that the variance becomes infinite only at the critical point $h=0, \beta=1$, where $M=0$.

Now we are ready to attack the much more difficult spin glass model. But it will be surprising to see that, by following a simple extension of the methods described here, we will arrive at similar results.

Basic Definitions for the Mean-Field Spin Glass Model

As in the ferromagnetic case, the generic configuration of the mean-field spin glass model is defined through Ising spin variables $\sigma_i = \pm 1$, attached to each site $i = 1, 2, \dots, N$.

But now there is an external quenched disorder given by the $N(N-1)/2$ independent and identical distributed random variables J_{ij}, defined for each pair of sites. For the sake of simplicity, we assume each J_{ij} to be a centered unit Gaussian with averages $E(J_{ij}) = 0, E(J_{ij}^2) = 1$. By quenched disorder we mean that the J have a kind of stochastic external influence on the system, without contributing to the thermal equilibrium.

Now the Hamiltonian of the model, in some external field of strength h, is given by the mean-field expression

$$H_N(\sigma,h,J) = -\frac{1}{\sqrt{N}} \sum_{(i,j)} J_{ij}\sigma_i\sigma_j - h\sum_i \sigma_i \qquad [25]$$

Here, the first sum extends to all site pairs, and the second to all sites. Notice the $\sqrt{N}$, necessary to ensure a good thermodynamic behavior to the free energy.

For a given inverse temperature β, let us now introduce the disorder-dependent partition function $Z_N(\beta,h,J)$ and the quenched average of the free energy per site $f_N(\beta,h)$, according to the definitions

$$Z_N(\beta,h,J) = \sum_{\sigma_1\dots\sigma_N} \exp(-\beta H_N(\sigma,h,J)) \qquad [26]$$

$$-\beta f_N(\beta,h) = N^{-1} E \log Z_N(\beta,h,J) \qquad [27]$$

Notice that in [27] the average E with respect to the external noise is made "after" the log is taken. This procedure is called quenched averaging. It represents the physical idea that the external noise does not contribute to the thermal equilibrium. Only the σ's are thermalized.

For the sake of simplicity, it is also convenient to write the partition function in the following equivalent form. First of all let us introduce a family of centered Gaussian random variables $\mathcal{K}(\sigma)$, indexed by the configurations σ, and characterized by the covariances

$$E(\mathcal{K}(\sigma)\mathcal{K}(\sigma')) = q^2(\sigma,\sigma') \qquad [28]$$

where $q(\sigma,\sigma')$ are the overlaps between two generic configurations, defined by

$$q(\sigma,\sigma') = N^{-1}\sum_i \sigma_i\sigma_i' \qquad [29]$$

with the obvious bounds $-1 \leq q(\sigma,\sigma') \leq 1$, and the normalization $q(\sigma,\sigma)=1$. Then, starting from the definition [25], it is immediately seen that the partition function in [26] can also be written, by neglecting unessential constant terms, in the form

$$Z_N(\beta,h,J) = \sum_{\sigma_1\dots\sigma_N} \exp\left(\beta\sqrt{\frac{N}{2}}\mathcal{K}(\sigma)\right) \exp\left(\beta h \sum_i \sigma_i\right) \qquad [30]$$

which will be the starting point of our treatment.

Basic Formulas of Derivation and Interpolation

We work in the following general setting. Let U_i be a family of centered Gaussian random variables, $i = 1, \dots, K$, with covariance matrix given by $E(U_iU_j) \equiv S_{ij}$. We treat the index i now as configuration space for some statistical mechanics system, with partition function Z and quenched free energy given by

$$E \log \sum_i w_i \exp(\sqrt{t}U_i) \equiv E \log Z \qquad [31]$$

where $w_i \geq 0$ are generic weights, and t is a parameter ruling the strength of the interaction.

It would be hard to underestimate the relevance of the following derivation formula

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} E \log \sum_i w_i \exp(\sqrt{t} U_i) \\ = \frac{1}{2} E\left(Z^{-1} \sum_i w_i \exp(\sqrt{t} U_i \right) S_{ii} \\ - \frac{1}{2} E\left(Z^{-2} \sum_i \sum_j w_i w_j \exp(\sqrt{t} U_i) \cdot \right. \\ \left. \times \exp(\sqrt{t} U_j) S_{ij} \right)\end{aligned} \quad [32]$$

The proof is straightforward. First we perform directly the t-derivative. Then, we notice that the random variables appear in expressions of the form $E(U_i F)$, where F are functions of the U's. These can be easily handled through the following integration by parts formula for generic Gaussian random variables, strongly reminiscent of the Wick theorem in quantum field theory,

$$E(U_i F) = \sum_j S_{ij} E\left(\frac{\partial}{\partial U_j} F \right) \quad [33]$$

Therefore, we see that always two derivatives are involved. The two terms in [32] come from the action of the U_j derivatives, the first acting on the Boltzmann factor, and giving rise to a Kronecker δ_{ij}, the second acting on Z^{-1}, and giving rise to the minus sign and the duplication of variables.

The derivation formula can be expressed in a more compact form by introducing replicas and suitable averages. In fact, let us introduce the state ω acting on functions F of i as follows

$$\omega(F(i)) = Z^{-1} \sum_i w_i \exp(\sqrt{t} U_i) F(i) \quad [34]$$

together with the associated product state Ω acting on replicated configuration spaces $i_1, i_2, \ldots, i_s$. By performing also a global E average, finally we define the averages

$$\langle F \rangle_t \equiv E\Omega(F) \quad [35]$$

where the subscript is introduced in order to recall the t dependence of these averages.

Then, eqn [32] can be written in a more compact form

$$\frac{\mathrm{d}}{\mathrm{d}t} E \log \sum_i w_i \exp(\sqrt{t} U_i) = \tfrac{1}{2} \langle S_{i_1 i_1} \rangle - \tfrac{1}{2} \langle S_{i_1 i_2} \rangle \quad [36]$$

Our basic comparison argument will be based on the following very simple theorem.

Theorem 2 *Let U_i and $\hat{U}_i$, for $i = 1, \ldots, K$, be independent families of centered Gaussian random variables, whose covariances satisfy the inequalities for generic configurations*

$$E(U_i U_j) \equiv S_{ij} \geq E(\hat{U}_i \hat{U}_j) \equiv \hat{S}_{ij} \quad [37]$$

and the equalities along the diagonal

$$E(U_i U_i) \equiv S_{ii} = E(\hat{U}_i \hat{U}_i) \equiv \hat{S}_{ii} \quad [38]$$

then for the quenched averages we have the inequality in the opposite sense

$$E \log \sum_i w_i \exp(U_i) \leq E \log \sum_i w_i \exp(\hat{U}_i) \quad [39]$$

where the $w_i \geq 0$ are the same in the two expressions.

Considerations of this kind are present in the mathematical literature, as mentioned, for example, in Talagrand (2003).

The proof is extremely simple and amounts to a straightforward calculation. In fact, let us consider the interpolating expression

$$E \log \sum_i w_i \exp(\sqrt{t} U_i + \sqrt{1-t} \hat{U}_i) \quad [40]$$

where $0 \leq t \leq 1$. Clearly, the two expressions under comparison correspond to the values $t = 0$ and $t = 1$, respectively. By taking the derivative with respect to t, with the help of the previous derivation formula, we arrive at the evaluation of the t derivative in the form

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} E \log \sum_i w_i \exp(\sqrt{t} U_i + \sqrt{1-t} \hat{U}_i) \\ = \frac{1}{2} E\left(Z^{-1} \sum_i w_i \exp(\sqrt{t} U_i)(S_{ii} - \hat{S}_{ii} \right) \\ - \frac{1}{2} E\left(Z^{-2} \sum_i \sum_j w_i w_j \exp(\sqrt{t} U_i) \right. \\ \left. \times \exp(\sqrt{t} U_j)(S_{ij} - \hat{S}_{ij}) \right)\end{aligned} \quad [41]$$

From the conditions assumed for the covariances, we immediately see that the interpolating function is nonincreasing in t, and the theorem follows.

The derivation formula and the comparison theorem are not restricted to the Gaussian case. Generalizations in many directions are possible. For the diluted spin glass models and optimization problems we refer, for example, to Franz and Leone (2003), and to De Sanctis (2005), and references therein.

Thermodynamic Limit and the Variational Bounds

We give here some striking applications of the basic comparison theorem. Guerra and Toninelli (2002) have given a very simple proof of a long-awaited result, about the convergence of the free energy per site in the thermodynamic limit. Let us show the argument. Let us consider a system of size N and two smaller systems of sizes N_1 and N_2 respectively, with $N = N_1 + N_2$, as before in the ferromagnetic case. Let us now compare

$$E \log Z_N(\beta, h, J) = E \log \sum_{\sigma_1 \dots \sigma_N} \exp\left(\beta \sqrt{\frac{N}{2}} \mathcal{K}(\sigma)\right) \times \exp\left(\beta h \sum_i \sigma_i\right) \quad [42]$$

with

$$E \log \sum_{\sigma_1 \dots \sigma_N} \exp\left(\beta \sqrt{\frac{N_1}{2}} \mathcal{K}^{(1)}(\sigma^{(1)})\right) \times \exp\left(\beta \sqrt{\frac{N_2}{2}} \mathcal{K}^{(2)}(\sigma^{(2)})\right) \exp\left(\beta h \sum_i \sigma_i\right) \equiv E \log Z_{N_1}(\beta, h, J) + E \log Z_{N_2}(\beta, h, J) \quad [43]$$

where $\sigma^{(1)}$ stands for $\sigma_i, i = 1, \dots, N_1$, and $\sigma^{(2)}$ for $\sigma_i, i = N_1 + 1, \dots, N$. Covariances for $\mathcal{K}^{(1)}$ and $\mathcal{K}^{(2)}$ are expressed as in [28], but now the overlaps are substituted with the partial overlaps of the first and second block, q_1 and q_2, respectively. It is very simple to apply the comparison theorem. All one has to do is to observe that the obvious

$$Nq = N_1 q_1 + N_2 q_2 \quad [44]$$

analogous to [10], implies, as in [12],

$$q^2 \leq \frac{N_1}{N} q_1^2 + \frac{N_2}{N} q_2^2 \quad [45]$$

Therefore, the comparison gives the superaddivity property, to be compared with [9],

$$E \log Z_N(\beta, h, J) \geq E \log Z_{N_1}(\beta, h, J) + E \log Z_{N_2}(\beta, h, J) \quad [46]$$

From the superaddivity property the existence of the limit follows in the form

$$\lim_{N \to \infty} N^{-1} E \log Z_N(\beta, h, J) = \sup_N N^{-1} E \log Z_N(\beta, h, J) \quad [47]$$

to be compared with [13].

The second application is in the form of the Aizenman–Sims–Starr generalized variational principle. Here, we will need to introduce some auxiliary system. The denumerable configuration space is given by the values of $\alpha = 1, 2, \dots$. We introduce also weights $w_\alpha \geq 0$ for the α system, and suitably defined overlaps between two generic configurations $p(\alpha, \alpha')$, with $p(\alpha, \alpha) = 1$.

A family of centered Gaussian random variables $\hat{\mathcal{K}}(\alpha)$, now indexed by the configurations α, will be defined by the covariances

$$E(\hat{\mathcal{K}}(\alpha)\hat{\mathcal{K}}(\alpha')) = p^2(\alpha, \alpha') \quad [48]$$

We will also need a family of centered Gaussian random variables $\eta_i(\alpha)$, indexed by the sites i of our original system and the configurations α of the auxiliary system, so that

$$E(\eta_i(\alpha)\eta_{i'}(\alpha')) = \delta_{ii'} p(\alpha, \alpha') \quad [49]$$

Both the probability measure w_α, and the overlaps $p(\alpha, \alpha')$ could depend on some additional external quenched noise, which does not appear explicitly in our notation.

In the following, we will denote by E averages with respect to all random variables involved.

In order to start the comparison argument, we will consider first the case where the two σ and α systems are not coupled, so as to appear factorized in the form

$$E \log \sum_{\sigma_1 \dots \sigma_N} \sum_\alpha w_\alpha \exp\left(\beta \sqrt{\frac{N}{2}} \mathcal{K}(\sigma)\right) \times \exp\left(\beta \sqrt{\frac{N}{2}} \hat{\mathcal{K}}(\alpha)\right) \exp\left(\beta h \sum_i \sigma_i\right) \equiv E \log Z_N(\beta, h, J) + E \log \sum_\alpha w_\alpha \times \exp\left(\beta \sqrt{\frac{N}{2}} \hat{\mathcal{K}}(\alpha)\right) \quad [50]$$

In the second case, the $\mathcal{K}$ fields are suppressed and the coupling between the two systems will be taken in a very simple form, by allowing the η field to act as an external field on the σ system. In this way the σ's appear as factorized, and the sums can be explicitly performed. The chosen form for the second term in the comparison is

$$E \log \sum_{\sigma_1 \dots \sigma_N} \sum_\alpha w_\alpha \exp\left(\beta \sum_i \eta_i(\alpha)\sigma_i\right) \exp\left(\beta h \sum_i \sigma_i\right) \equiv N \log 2 + E \log \sum_\alpha w_\alpha (c_1 c_2 \dots c_N) \quad [51]$$

where we have defined

$$c_i = \cosh \beta(h + \eta_i(\alpha)) \qquad [52]$$

as arising from the sums over σ's.

Now we apply the comparison theorem. In the first case, the covariances involve the sums of squares of overlaps

$$\tfrac{1}{2}(q^2(\sigma, \sigma') + p^2(\alpha, \alpha')) \qquad [53]$$

In the second case, a very simple calculation shows that the covariances involve the overlap products

$$q(\sigma, \sigma')p(\alpha, \alpha'\prime) \qquad [54]$$

Therefore, the comparison is very easy and, by collecting all expressions, we end up with the useful estimate, as in Aizenman *et al.* (2003), holding for any auxiliary system as defined before,

$$\begin{aligned} N^{-1} E \log Z_N(\beta, h, J) \\ \leq \log 2 + N^{-1} \Bigg(E \log \sum_\alpha w_\alpha (c_1 c_2 \cdots c_N) \\ - E \log \sum_\alpha w_\alpha \exp\left(\beta \sqrt{\frac{N}{2}} \hat{\mathcal{K}}(\alpha)\right)\Bigg) \end{aligned} \qquad [55]$$

The Parisi Representation for the Free Energy

We refer to the original papers, reprinted in the extensive review given in Mézard *et al.* (2002), for the general motivations, and the derivation of the broken replica ansatz, in the frame of the ingenious replica trick. Here, we limit ourselves to a synthetic description of its general structure, independently from the replica trick.

First of all, let us introduce the convex space $\mathcal{X}$ of the functional order parameters x, as nondecreasing functions of the auxiliary variable q, both x and q taking values on the interval $[0, 1]$, that is,

$$\mathcal{X} \ni x : [0, 1] \ni q \to x(q) \in [0, 1] \qquad [56]$$

Notice that we call x the function, and $x(q)$ its values. We introduce a metric on $\mathcal{X}$ through the $L^1([0, 1], \mathrm{d}q)$-norm, where $\mathrm{d}q$ is the Lebesgue measure.

For our purposes, we will consider the case of piecewise constant functional order parameters, characterized by an integer K, and two sequences $q_0, q_1, \ldots, q_K, m_1, m_2, \ldots, m_K$ of numbers satisfying

$$\begin{aligned} 0 = q_0 \leq q_1 \leq \cdots \leq q_{K-1} \leq q_K = 1 \\ 0 \leq m_1 \leq m_2 \leq \cdots \leq m_K \leq 1 \end{aligned} \qquad [57]$$

such that

$$\begin{aligned} &x(q) = m_1 \quad \text{for } 0 = q_0 \leq q < q_1 \\ &x(q) = m_2 \quad \text{for } q_1 \leq q < q_2 \\ &\vdots \\ &x(q) = m_K \quad \text{for } q_{K-1} \leq q \leq q_K \end{aligned} \qquad [58]$$

In the following, we will find it convenient to define also $m_0 \equiv 0$, and $m_{K+1} \equiv 1$. The replica symmetric case of Sherrington and Kirkpatrick corresponds to

$$K = 2, \quad q_1 = \bar{q}, \quad m_1 = 0, \quad m_2 = 1 \qquad [59]$$

Let us now introduce the function f, with values $f(q, y; x, \beta)$, of the variables $q \in [0, 1], y \in R$, depending also on the functional order parameter x, and on the inverse temperature β, defined as the solution of the nonlinear antiparabolic equation

$$\begin{aligned} (\partial_q f)(q, y) + \tfrac{1}{2}(\partial_y^2 f)(q, y) \\ + \tfrac{1}{2} x(q)(\partial_y f)^2(q, y) = 0 \end{aligned} \qquad [60]$$

with final condition

$$f(1, y) = \log \cosh(\beta y) \qquad [61]$$

Here, we have stressed only the dependence of f on q and y.

It is very simple to integrate eqn [60] when x is piecewise constant. In fact, consider $x(q) = m_a$, for $q_{a-1} \leq q \leq q_a$, firstly with $m_a > 0$. Then, it is immediately seen that the correct solution of eqn [60] in this interval, with the right final boundary condition at $q = q_a$, is given by

$$\begin{aligned} f(q, y) \\ = \frac{1}{m_a} \log \int \exp(m_a f(q_a, y + z\sqrt{q_a - q}))\, \mathrm{d}\mu(z) \end{aligned} \qquad [62]$$

where $\mathrm{d}\mu(z)$ is the centered unit Gaussian measure on the real line. On the other hand, if $m_a = 0$, then [60] loses the nonlinear part and the solution is given by

$$f(q, y) = \int f(q_a, y + z\sqrt{q_a - q})\, \mathrm{d}\mu(z) \qquad [63]$$

which can be seen also as deriving from [62] in the limit $m_a \to 0$. Starting from the last interval K, and using [62] iteratively on each interval, we easily get the solution of [60], [61], in the case of piecewise order parameter x, as in [58], through a chain of interconnected Gaussian integrations.

Now, we introduce the following important definitions. The trial auxiliary function, associated to a given mean-field spin glass system, as described

earlier, depending on the functional order parameter x, is defined as

$$\log 2 + f(0, h; x, \beta) - \frac{\beta^2}{2} \int_0^1 q\, x(q)\, \mathrm{d}q \qquad [64]$$

Notice that in this expression the function f appears evaluated at $q = 0$, and $y = h$, where h is the value of the external magnetic field. This trial expression shoul be considered as the analog of that appearing in [14] for the ferromagnetic case.

The Parisi spontaneously broken replica symmetry expression for the free energy is given by the definition

$$-\beta f_P(\beta, h) \equiv \inf_x \left(\log 2 + f(0, h; x, \beta) - \frac{\beta^2}{2} \int_0^1 q\, x(q)\, \mathrm{d}q\right) \qquad [65]$$

where the infimum is taken with respect to all functional order parameters x. Notice that the infimum appears here, as compared to the supremum in the ferromagnetic case.

By exploiting a kind of generalized comparison argument, involving a suitably defined interpolation function, Guerra (2003) has established the following important result.

Theorem 3 *For all values of the inverse temperature β, and the external magnetic field h, and for any functional order parameter x, the following bound holds*:

$$N^{-1} E \log Z_N(\beta, h, J) \leq \log 2 + f(0, h; x, \beta) - \frac{\beta^2}{2} \int_0^1 q\, x(q)\, \mathrm{d}q$$

uniformly in N. Consequently, we have also

$$N^{-1} E \log Z_N(\beta, h, J) \leq \inf_x \left(\log 2 + f(0, h; x, \beta) - \frac{\beta^2}{2} \int_0^1 q\, x(q)\, \mathrm{d}q\right)$$

uniformly in N.

However, this result can also be understood in the framework of the generalized variational principle established by Aizenman–Sims–Starr as described earlier.

In fact, one can easily show that there exist α systems such that

$$N^{-1} E \log \sum_\alpha w_\alpha c_1 c_2 \dots c_N \equiv f(0, h; x, \beta) \qquad [66]$$

$$N^{-1} E \log \sum_\alpha w_\alpha \exp\left(\beta \sqrt{\frac{N}{2}} \hat{\mathcal{K}}(\alpha)\right) \equiv \frac{\beta^2}{2} \int_0^1 q\, x(q)\, \mathrm{d}q \qquad [67]$$

uniformly in N. This result stems from earlier work of Derrida, Ruelle, Neveu, Bolthausen, Sznitman, Aizenman, Contucci, Talagrand, Bovier, and others, and in a sense is implicit in the treatment given in Mézard *et al.* (1987). It can be reached in a very simple way. Let us sketch the argument.

First of all, let us consider the Poisson point process $y_1 \geq y_2 \geq y_3 \dots$, uniquely characterized by the following conditions. For any interval A, introduce the occupation numbers $N(A)$, defined by

$$N(A) = \sum_\alpha \chi(y_\alpha \in A) \qquad [68]$$

where $\chi(\,) = 1$, if the random variable y_α belongs to the interval A, and $\chi(\,) = 0$, otherwise. We assume that $N(A)$ and $N(B)$ are independent if the intervals A and B are disjoint, and moreover that for each A, the random variable $N(A)$ has a Poisson distribution with parameter

$$\mu(A) = \int_a^b \exp(-y)\, \mathrm{d}y \qquad [69]$$

if A is the interval (a, b), that is,

$$P(N(A) = k) = \exp(-\mu(A))\mu(A)^k / k! \qquad [70]$$

We will exploit $-y_\alpha$ as energy levels for a statistical mechanics system with configurations indexed by α. For a parameter $0 < m < 1$, playing the role of inverse temperature, we can introduce the partition function

$$v = \sum_\alpha \exp\left(\frac{y_\alpha}{m}\right) \qquad [71]$$

For m in the given interval it turns out that v is a very well defined random variable, with the sum over α extending to infinity. In fact, there is a strong inbuilt smooth cutoff in the very definition of the stochastic energy levels.

From the general properties of Poisson point processes, it is very well known that the following basic invariance property holds. Introduce a random variable b, independent of y, subject to the condition $E(\exp b) = 1$, and let b_α be independent copies. Then, the randomly biased point process $y'_\alpha = y_\alpha + b_\alpha$, $\alpha = 1, 2, \dots$, is equivalent to the original one in distribution. An immediate consequence is the following. Let f be a random variable, independent of y, such that $E(\exp f) < \infty$, and let f_α be independent copies. Then, the two random variables

$$\sum_\alpha \exp\left(\frac{y_\alpha}{m}\right) \exp(f_\alpha) \qquad [72]$$

$$\sum_\alpha \exp\left(\frac{y_\alpha}{m}\right) E(\exp(mf))^{1/m} \qquad [73]$$

have the same distribution. In particular, they can be freely substituted under averages.

The auxiliary system which gives rise to the Parisi representation according to [66] and [67], for a piecewise constant order parameter, is expressed in the following way. Now α will be a multi-index $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$, where each α_a runs on $1, 2, 3, \ldots$. Define the Poisson point process y_{α_1}, then, independently, for each value of α_1 processes $y_{\alpha_1\alpha_2}$, and so on up to $y_{\alpha_1\alpha_2\ldots\alpha_K}$. Notice that in the cascade of independent processes $y_{\alpha_1}, y_{\alpha_1\alpha_2}, \ldots, y_{\alpha_1\alpha_2\ldots\alpha_K}$, the last index refers to the numbering of the various points of the process, while the first indices denote independent copies labeled by the corresponding α's.

The weights w_α have to be chosen according to the definition

$$w_\alpha = \exp\frac{y_{\alpha_1}}{m_1}\exp\frac{y_{\alpha_1\alpha_2}}{m_2}\ldots\exp\frac{y_{\alpha_1\alpha_2\ldots\alpha_K}}{m_K} \qquad [74]$$

The cavity fields η and $\mathcal{K}$ have the following expression in terms of independent unit Gaussian random variables $J^i_{\alpha_1}, J^i_{\alpha_1\alpha_2}, \ldots, J^i_{\alpha_1\alpha_2\ldots\alpha_K}, J'_{\alpha_1}, J'_{\alpha_1\alpha_2}, \ldots, J'_{\alpha_1\alpha_2\ldots\alpha_K}$,

$$\eta_i(\alpha) = \sqrt{q_1 - q_0}\, J^i_{\alpha_1} + \sqrt{q_2 - q_1}\, J^i_{\alpha_1\alpha_2} + \cdots + \sqrt{q_K - q_{K-1}}\, J^i_{\alpha_1\alpha_2\ldots\alpha_K} \qquad [75]$$

$$\mathcal{K}(\alpha) = \sqrt{q_1^2 - q_0^2}\, J'_{\alpha_1} + \sqrt{q_2^2 - q_1^2}\, J'_{\alpha_1\alpha_2} + \cdots + \sqrt{q_K^2 - q_{K-1}^2}\, J'_{\alpha_1\alpha_2\ldots\alpha_K} \qquad [76]$$

It is immediate to verify that $E(\eta_i(\alpha)\eta_{i'}(\alpha'))$ is zero if $i \neq i'$, while

$$E(\eta_i(\alpha)\eta_i(\alpha')) = \begin{cases} 0 \quad \text{if } \alpha_1 \neq \alpha'_1 \\ q_1 \text{ if } \alpha_1 = \alpha'_1, \alpha_2 \neq \alpha'_2 \\ q_2 \text{ if } \alpha_1 = \alpha'_1, \alpha_2 = \alpha'_2, \alpha_3 \neq \alpha'_3, \\ \vdots \\ 1 \quad \text{if } \alpha_1 = \alpha'_1, \alpha_2 = \alpha'_2, \ldots, \\ \alpha_K = \alpha'_K \end{cases} \qquad [77]$$

Similarly, we have

$$E(\mathcal{K}(\alpha)\mathcal{K}(\alpha')) = \begin{cases} 0 \quad \text{if } \alpha_1 \neq \alpha'_1 \\ q_1^2 \text{ if } \alpha_1 = \alpha'_1, \alpha_2 \neq \alpha'_2 \\ q_2^2 \text{ if } \alpha_1 = \alpha'_1, \alpha_2 = \alpha'_2, \alpha_3 \neq \alpha'_3, \\ \vdots \\ 1 \quad \text{if } \alpha_1 = \alpha'_1, \alpha_2 = \alpha'_2, \ldots, \\ \alpha_K = \alpha'_K \end{cases} \qquad [78]$$

This ends the definition of the α system, associated to a given piecewise constant order parameter.

Now, it is simple to verify that [66] and [67] hold. Let us consider, for example, [66]. With the α system chosen as before, the repeated application of the stochastic equivalence of [72] and [73] will give rise to a sequence of interchained Gaussian integrations exactly equivalent to those arising from the expression for f, as solution of the eqn [60]. For [73], there are equivalent considerations.

Therefore, we see that the estimate in Theorem 3 is also a consequence of the generalized variational principle.

Up to this point we have seen how to obtain upper bounds. The problem arises whether, as in the ferromagnetic case, we can also get lower bounds, so as to shrink the thermodynamic limit to the value given by the $\inf_x$ in Theorem 3. After a short announcement, Talagrand (2005) has firmly established the complete proof of the control of the lower bound. We refer to the original paper for the complete details of this remarkable achievement. About the methods, here we only recall that in Guerra (2003) we have given also the corrections to the bounds appearing in Theorem 3, albeit in a quite complicated form. Talagrand has been able to establish that these corrections do in fact vanish in the thermodynamic limit.

In conclusion, we can establish the following extension of Theorem 1 to spin glasses.

Theorem 4 *For the mean-field spin glass model we have*

$$\lim_{N\to\infty} N^{-1} E \log Z_N(\beta, h, J) = \sup_N N^{-1} E \log Z_N(\beta, h, J) \qquad [79]$$

$$= \inf_x \left(\log 2 + f(0, h; x, \beta) - \frac{\beta^2}{2}\int_0^1 q x(q)\, dq \right) \qquad [80]$$

Diluted Models

Diluted models, in a sense, play a role intermediate between the mean-field case and the short-range case. In fact, while in the mean-field model each site is interacting with all other sites, on the other hand, in the diluted model, each site is interacting with only a fixed number of other sites. However, while for the short-range models there is a definition of distance among sites, relevant for the interaction, no such definition appears in the diluted models, where all sites are in any case equivalent. From this point of view, the diluted models are structurally similar to the mean-field models, and most of the

techniques and results explained before can be extended to them.

Let us define a typical diluted model. The quenched noise is described as follows. Let K be a Poisson random variable with parameter αN, where N is the number of sites, and α is a parameter entering the theory, together with the temperature. We consider also a sequence of independent centered random variables $J_1, J_2, \ldots$, and a sequence of discrete independent random variables $i_1, j_1, i_2, j_2, \ldots$, uniformly distributed over the set of sites $1, 2, \ldots, N$. Then we assume as Hamiltonian

$$H_N(\sigma) = -\sum_{k=0}^{K} J_k \sigma_{i_k} \sigma_{j_k} \qquad [81]$$

Only the variables σ contribute to thermodynamic equilibrium. All noise coming from K, J_k, i_k, j_k is considered quenched, and it is not explicitly indicated in our notation for H.

The role played by Gaussian integration by parts in the Sherrington–Kirckpatrick model, here is assumed by the following elementary derivation formula, holding for Poisson distributions,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} P(K = k, t\alpha N) &\equiv \frac{\mathrm{d}}{\mathrm{d}t} \exp(-t\alpha N)(t\alpha N)^k / k! \\ &= \alpha N (P(K = k-1, t\alpha N) \\ &\quad - P(K = k, t\alpha N)) \end{aligned} \qquad [82]$$

Then, all machinery of interpolation can be easily extended to the diluted models, as firstly recognized by Franz and Leone in (2003).

In this way, the superaddivity property, the thermodynamic limit, and the generalized variational principle can be easily established. We refer to Franz and Leone (2003), and De Sanctis (2005), for a complete treatment.

There is an important open problem here. While in the fully connected case, the Poisson probability cascades provide the right auxiliary α systems to be exploited in the variational principle, on the other hand in the diluted case more complicated probability cascades have been proposed, as shown, for example, in Franz and Leone (2003), and in Panchenko and Talagrand (2004). On the other hand, in De Sanctis (2005), the very interesting proposal has been made that also in the case of diluted models the Poisson probability cascades play a very important role. Of course, here the auxiliary system interacts with the original system differently, and involves a multi-overlap structure as explained in De Sanctis (2005). In this way a kind of very deep universality is emerging. Poisson probability cascades are a kind of universal class of auxiliary systems. The different models require different cavity fields ruling the interaction between the original system and the auxiliary system. But further work will be necessary in order to clarify this very important issue. For results about diluted models in the high-temperature region, we refer to Guerra and Toninelli (2004).

Short-Range Model and Its Connections with the Mean-Field Version

The investigations of the connections between the short-range version of the model and its mean-field version are at the beginning. Here, we limit ourselves to a synthetic description of what should be done, and to a short presentation of the results obtained so far.

First of all, according to the conventional wisdom, the mean-field version should be a kind of limit of the short-range model on a lattice in dimension d, when $d \to \infty$, with a proper rescaling of the strength of the Hamiltonian, of the form $d^{-1/2}$. Results of this kind are very well known in the ferromagnetic case, but the present technology of interpolation does not seem sufficient to assure a proof in the spin glass case. So, this very basic result is still missing. In analogy with the ferromagnetic case, it would be necessary to arrive at the notion of a critical dimension, beyond which the features of the mean-field case still hold, for example, in the expression of the critical exponents and in the ultrametric hierarchical structure of the pure phases, or at least for the overlap distributions. For physical dimensions less than the critical one, the short-range model would need corrections with respect to its mean-field version. Therefore, this is a completely open problem.

Moreover, always according to the conventional wisdom, the mean-field version should be a kind of limit of the short-range models, in finite fixed dimensions, as the range of the interaction goes to infinity, with proper rescaling. Important work of Franz and Toninelli shows that this is effectively the case, if a properly defined Kac limit is performed. Here, interpolation methods are effective, and we refer to Franz and Toninelli (2004), and references quoted there, for full details.

Due to the lack of efficient analytical methods, it is clear that numerical simulations play a very important role in the study of the physical properties emerging from short-range spin glass models. In particular, we refer to Marinari *et al.* (2000) for a detailed account of the evidence, coming from theoretical considerations and extensive computer simulations, that some of the more relevant features of the spontaneous replica breaking scheme of the mean field are also present in

short-range models in three dimensions. Different views are expressed, for example, in Newman and Stein (1998), where it is argued that the phase-space structure of short-range spin glass models is much simpler than that foreseen by the Parisi spontaneous replica symmetry mechanism.

Such very different views, both apparently strongly supported by reasonable theoretical considerations and powerful numerical simulations, are a natural consequence of the extraordinary difficulty of the problem.

It is clear that extensive additional work will be necessary before the clarification of the physical features exhibited by the realistic short-range spin glass models.

Conclusion and Outlook for Future Developments

As we have seen, in these last few years, there has been an impressive progress in the understanding of the mathematical structure of spin glass models, mainly due to the systematic exploration of comparison and interpolation methods. However, many important problems are still open. The most important one is to establish rigorously the full hierarchical ultrametric organization of the overlap distributions, as appears in Parisi theory, and to fully understand the decomposition in pure states of the glassy phase, at low temperatures.

Moreover, it would be important to extend these methods to other important disordered models as, for example, neural networks. Here the difficulty is that the positivity arguments, so essential in comparison methods, do not seem to emerge naturally inside the structure of the theory.

Finally, the problem of connecting properties of the short-range model, with those arising in the mean-field case, is still almost completely open.

Acknowledgment

We gratefully acknowledge useful conversations with Michael Aizenman, Pierluigi Contucci, Giorgio Parisi, and Michel Talagrand. The strategy explained in this report grew out from a systematic exploration of comparison and interpolation methods, developed in collaboration with Fabio Lucio Toninelli, and Luca De Sanctis.

This work was supported in part by MIUR (Italian Ministry of Instruction, University and Research), and by INFN (Italian National Institute for Nuclear Physics).

See also: Glassy Disordered Systems: Dynamical Evolution; Large Deviations in Equilibrium Statistical Mechanics; Mean Field Spin Glasses and Neural Networks; Short-Range Spin Glasses: The Metastate Approach; Statistical Mechanics and Combinatorial Problems.

Further Reading

Aizenman M, Sims R, and Starr S (2003) Extended variational principle for the Sherrington–Kirkpatrick spin-glass model. *Physical Review B* 68: 214403.

De Sanctis L (2005) *Structural Approachs to Spin Glasses and Optimization Problems*. Ph.D. thesis, Department of Mathematics, Princeton University.

Edwards SF and Anderson PW (1975) Theory of spin glasses. *Journal of Physics F: Metal Physics* 5: 965–974.

Franz S and Leone M (2003) Replica bounds for optimization problems and diluted spin systems. *Journal of Statistical Physics* 111: 535–564.

Franz S and Toninelli FL (2004) The Kac limit for finite-range spin glasses. *Physical Review Letters* 92: 030602.

Guerra F (2001) Sum rules for the free energy in the mean field spin glass model. *Fields Institute Communications* 30: 161.

Guerra F (2003) Broken replica symmetry bounds in the mean field spin glass model. *Communications in Mathematical Physics* 233: 1–12.

Guerra F and Toninelli FL (2002) The thermodynamic limit in mean field spin glass models. *Communications in Mathematical Physics* 230: 71–79.

Guerra F and Toninelli FL (2004) The high temperature region of the Viana–Bray diluted spin glass model. *Journal of Statistical Physics* 115: 531–555.

Marinari E, Parisi G, Ricci-Tersenghi F, Ruiz-Lorenzo JJ, and Zuliani F (2000) Replica symmetry breaking in short range spin glasses: A review of the theoretical foundations and of the numerical evidence. *Journal of Statistical Physics* 98: 973–1074.

Mézard M, Parisi G, and Virasoro MA (1987) *Spin Glass Theory and Beyond*. Singapore: World Scientific.

Mézard M, Parisi G, and Zecchina R (2002) Analytic and algorithmic solution of random satisfiability problems. *Science* 297: 812.

Newman CM and Stein DL (1998) Simplicity of state and overlap structure in finite-volume realistic spin glasses. *Physical Review E* 57: 1356–1366.

Panchenko D and Talagrand M (2004) Bounds for diluted mean-field spin glass models. *Probability Theory Related Fields* 130: 319–336.

Sherrington D and Kirkpatrick S (1975) Solvable model of a spin-glass. *Physical Review Letters* 35: 1792–1796.

Stein DL (1989) Disordered systems: Mostly spin glasses. In: Stein DL (ed.) *Lectures in the Sciences of Complexity*. New York: Addison-Wesley.

Talagrand M (2003) *Spin Glasses: A Challenge for Mathematicians. Mean Field Models and Cavity Method*. Berlin: Springer.

Talagrand M (2006) The Parisi formula. *Annals of Mathematics* 163: 221–263.

Young P (ed.) (1987) *Spin Glasses and Random Fields*. Singapore: World Scientific.

Spinors and Spin Coefficients

K P Tod, University of Oxford, Oxford, UK

Introduction

Spinors were invented by the mathematician E Cartan (see, e.g., Cartan (1981)) in the early years of the last century in the course of his study of rotation groups. The physicist Pauli reinvented what Cartan would have called the spinors of SU(2), which is the double cover of the rotation group SO(3), in order to explain the spectroscopy of alkali atoms and the anomalous Zeeman effect. For this, he needed an essential two-valuedness of the electron, an internal quantum number to contribute to the angular momentum, which is now called spin. Now the wave function becomes a two-component column vector. It is worth noting that, despite the name, Pauli resisted the picture of an electron as a spinning "thing" on the grounds that, as a representation of SU(2) which was not a representation of SO(3), it should have no classical kinematic model, which a spinning object would have.

According to the review article of van der Waerden (1960), the term "spinor" is due to Ehrenfest in 1929, and was introduced in the flurry of interest after the next important step in the evolution of spinors in the physics literature, which was the introduction of a relativistic equation for the electron by Dirac (1928).

Dirac sought a linear, first-order but Lorentz-invariant equation for the electron which was to be the square root of the linear, Lorentz-invariant but second-order Klein–Gordon equation. He assumed the equation for the wave function ψ would take the form

$$L\psi := (\mathrm{i}\gamma^a p_a + mcI)\psi = 0 \qquad [1]$$

where $p_a = -\mathrm{i}\hbar\partial/\partial x^a$ for $a = 0, 1, 2, 3$, but where γ^a are complex square matrices, of a size to be determined, and I is the corresponding identity matrix. Differentiating [1] again, one obtains the Klein–Gordon equation for ψ provided these matrices satisfy the equation

$$\gamma^a\gamma^b + \gamma^b\gamma^a = 2\eta^{ab}I \qquad [2]$$

where η^{ab} is the Minkowski metric, diag(1, −1, −1, −1).

Assuming the γ^a have been found, the usual substitution $p \to p - \mathrm{i}eA$, for a particle in a magnetic field with vector potential A, leads to the correct magnetic moment for the electron, so that this equation does describe an electron with spin in the form made familiar by Pauli.

To decide on the size of the matrices γ^a and therefore the dimension of the space of ψ's, one notices, with the aid of [2], that the following are a basis for the algebra generated by the γ^a:

$$1, \gamma^a, \gamma^{[a}\gamma^{b]}, \gamma^{[a}\gamma^b\gamma^{c]}, \gamma^{[a}\gamma^b\gamma^c\gamma^{d]} \qquad [3]$$

There are 16 elements in this basis, assuming that there are no extra identities among them, so that we might hope to find a representation as 4×4 matrices. This can be done, and Dirac gave explicit formulas in terms of Pauli matrices. The space of Dirac spinors is now a complex four-dimensional vector space, which turns out to split as the sum of a complex two-dimensional vector space S, which is referred to as a spin space, and its complex conjugate $\bar{S}$ (the relationship between a complex vector space and its complex conjugate is described in the text below and eqn [9]). Under proper, orthochronous Lorentz transformations, S transforms into itself by $\mathrm{SL}(2,\mathbb{C})$ transformation, but space and time reflections relate S to $\bar{S}$. The fact that there are two spin spaces S and $\bar{S}$ in dimension 4 is the basis of chirality: an electron is represented by a Dirac spinor, which is a pair of spinors, one in each of S and $\bar{S}$, which are related under space reflection; a particle represented just by a spinor in S cannot be invariant under space reflection.

The Clifford algebra (see Clifford Algebras and their Representations) associated with a vector space V with metric g is defined as the algebra generated by elements v, w of V with the multiplication $\cup$ satisfying

$$v \cup w + w \cup v = 2g(v, w) \qquad [4]$$

The matrices γ^a define a representation of the Clifford algebra by associating a covector v_a with a matrix $v = v_a\gamma^a$, since [2] is then equivalent to [4].

This part of the process works in any dimension n and signature s. For odd n, as, for example, with Pauli spinors, the γ^a are square matrices of size $2^N \times 2^N$, where $N = (n-1)/2$, and there is a single spin space of dimension 2^N. For even n, as with the original Dirac spinors, the γ^a are square matrices of size $2^N \times 2^N$, where $N = n/2$, but there are two spin spaces each of dimension 2^{N-1}. Reality properties of the spin spaces and the existence of other structures on them depend in an intricate way on n and s (Penrose and Rindler 1984, 1986, Benn and Tucker 1987).

The dimension of the space of spinors rises rapidly with n, which is one reason why historically spinors have been most useful in spaces of dimensions 3 and 4, where the spin space has dimension 2. In a space of dimension 11, a case considered in supergravity, the spin space already has dimension 32.

Spinors in General Relativity: Spinor Algebra

In this section, we start again with a different emphasis. Conventions follow Penrose and Rindler (1984, 1986). To introduce spinors as a calculus in a four-dimensional, Lorentzian spacetime $\mathcal{M}$, one can begin by choosing an orthonormal tetrad of vectors $(\boldsymbol{e}_0, \boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$ at a point p. The following conventions are used:

$$g(\boldsymbol{e}_a, \boldsymbol{e}_b) = \eta_{ab} = \mathrm{diag}(1, -1, -1, -1)$$

Any vector $\boldsymbol{v}$ in the tangent space $V = T_p\mathcal{M}$ at p has components v^a in this basis, which we arrange as a matrix and label in two ways:

$$\Psi(\boldsymbol{v}) = \frac{1}{\sqrt{2}}\begin{pmatrix} v^0 + v^3 & v^1 + \mathrm{i}v^2 \\ v^1 - \mathrm{i}v^2 & v^0 - v^3 \end{pmatrix} = \begin{pmatrix} v^{00'} & v^{01'} \\ v^{10'} & v^{11'} \end{pmatrix} \quad [5]$$

The reason for the factor $1/\sqrt{2}$ will be seen below, as will the rationale for the second form of the matrix. Note that $\Psi(\boldsymbol{v})$ is Hermitian and that

$$2\det\Psi(\boldsymbol{v}) = g(\boldsymbol{v}, \boldsymbol{v}) = \eta_{ab}v^a v^b \quad [6]$$

Clearly, there is a one-to-one correspondence between elements of V and Hermitian 2×2 matrices. Further, if t is any matrix in $\mathrm{SL}(2, \mathbb{C})$,then the transformation

$$\Psi(\boldsymbol{v}) \to t\Psi(\boldsymbol{v})t^\dagger \quad [7]$$

where $t^\dagger$ is the Hermitian conjugate of t, is linear in $\boldsymbol{v}$, and preserves both Hermiticity and the norm of $\boldsymbol{v}$. Thus, it must represent a Lorentz transformation. It is straightforward to check that it is a proper, orthochronous Lorentz transformation and that all such transformations arise in this way (recall that "proper" means transformations of determinant 1 so that orientation is preserved, and "orthochronous" means that future-pointing timelike or null vectors are taken to future-pointing timelike or null vectors, so that time orientation is preserved; the proper, orthochronous Lorentz group is equivalently the identity-connected component of the Lorentz group). Since both t and $-t$ give the same Lorentz transformation, this provides an explicit demonstration of the (2 – 1)-homomorphism of $\mathrm{SL}(2,\mathbb{C})$ with the proper, orthochronous Lorentz group $\mathrm{O}^{\uparrow}_{+}(1,3)$.

If the vector $\boldsymbol{v}$ in [5] is null, then the matrix has vanishing determinant, or, equivalently, it has rank 1, and so it can be written as the outer product of a two-component column vector $\alpha = (\alpha^0, \alpha^1)^{\mathrm{T}}$ and its Hermitian conjugate:

$$\Psi(\boldsymbol{v}) = \alpha\alpha^\dagger \quad [8]$$

Furthermore, under [7], α transforms as

$$\alpha \to t\alpha \quad [9]$$

The two-complex-dimensional space to which α belongs is the spin space S at p, already met in the previous section, and it follows from [8], since null vectors span V, that the tensor product $S \otimes \bar{S}$ of S with its complex conjugate vector space $\bar{S}$ is the complexification of V. Complex conjugation gives an antilinear map from S to $\bar{S}$. (One associates the complex-conjugate vector space $\bar{V}$ to any given complex vector space V as follows: scalar multiplication for V can be considered as a function $\phi : \mathbb{C} \times V \to V$ given by $\phi(z, v) = zv$, while vector addition is a map $\psi : V \times V \to V$ given by $\psi(u, v) = u + v$. Define another complex vector space by taking the same vectors and the same ψ but with scalar multiplication $\bar{\phi}$, where $\bar{\phi}(z, v) = \phi(\bar{z}, v)$. This is the complex-conjugate vector space $\bar{V}$. Given a choice of basis, we think of V as, say, n-component column vectors of complex numbers, and then $\bar{V}$ is the corresponding complex-conjugate columns.)

Conventionally, S is the space of unprimed spinors and $\bar{S}$ the space of primed spinors, and one also has the two duals S' and $\bar{S}'$ which are associated in the corresponding way to the dual V' of V. Analogously to the situation with vectors and covectors, index conventions for spinors are as follows:

$$\alpha^A \in S, \quad \beta^{A'} \in \bar{S}, \quad \gamma_A \in S', \quad \delta_{A'} \in \bar{S}'$$

where $A = 0, 1, A' = 0', 1'$.

Spinor algebra mirrors tensor algebra: a spinor $\phi^{A_1 \ldots A_p A'_1 \ldots A'_q}{}_{B_1 \ldots B_r B'_1 \ldots B'_s}$ is an element of the tensor product of p copies of S, q copies of $\bar{S}$, r copies of S', and s copies of $\bar{S}'$. The second way of writing the matrix in [5] enables the identification of a vector with a matrix to be conventionally written as

$$v^a = v^{AA'} \quad [10]$$

and then extended to any tensor $T^{a\ldots b}{}_{c\ldots d}$ by replacing each vector index, say b, with a pair BB' of spinor indices. In particular, from [8], it follows that any real null vector n^a can be written in the form

$$n^a = \nu^A \bar{\nu}^{A'}$$

for some spinor ν^A.

One must pay attention to the order of spinor indices of a given type, primed or unprimed, but by convention may permute primed and unprimed indices. A spinor with an equal number n of primed and unprimed indices corresponds to a tensor of valence n, and the tensor is real if the spinor satisfies a suitable Hermiticity relation.

Spinors may have various symmetries among their indices, much as tensors have. However, since S is two dimensional, there is only a one-dimensional space of 2-forms on S. This has two consequences: no spinor can be antisymmetric over more than two indices; and if we make a choice of canonical 2-form, all spinors can be written in terms of symmetric spinors and the canonical 2-form. This is a decomposition of spinors into irreducibles for $\mathrm{SL}(2,\mathbb{C})$.

One makes a choice of 2-form ϵ_{AB} according to

$$\epsilon_{AB} = -\epsilon_{BA}, \quad \epsilon_{01} = 1$$

There is an inverse ϵ^{AB} defined by

$$\epsilon^{AC}\epsilon_{BC} = \delta^A_B \qquad [11]$$

where δ^A_B is the Kronecker delta. The complex conjugate of ϵ_{AB} is conventionally written without an overbar as $\epsilon_{A'B'}$, and analogously $\epsilon^{A'B'}$ is the complex conjugate of ϵ^{AB}.

Because of the antisymmetry of ϵ_{AB}, order of indices is crucial in equations such as [11]. The 2-form ϵ_{AB} has a role akin to that of a metric as it provides an identification of S and its dual, according to

$$\alpha^A \to \alpha_B = \alpha^A\epsilon_{AB}$$

$$\beta_B \to \beta^A = \epsilon^{AB}\beta_B$$

with corresponding formulas for primed spinors. Note that, because of the antisymmetry of ϵ_{AB}, necessarily $\alpha_A\alpha^A = 0$ for any α^A.

With conventions made so far, it can be checked that

$$g_{ab}v^a v^b = \epsilon_{AB}\epsilon_{A'B'}v^{AA'}v^{BB'} \qquad [12]$$

for any vector v^a, where g_{ab} is the spacetime metric at p, so that

$$g_{ab} = \epsilon_{AB}\epsilon_{A'B'}$$

It is the desire to have this formula without constants that necessitates the choice of the factor $1/\sqrt{2}$ in [5].

One final piece of spinor algebra that we note is the following: given a symmetric spinor $\phi_{A_1\ldots A_n}$ there is a factorization

$$\phi_{A_1\ldots A_n} = \alpha^{(1)}_{(A_1}\cdots\alpha^{(n)}_{A_n)} \qquad [13]$$

where the round brackets indicate symmetrization over the indices $A_1,\ldots,A_n$, and the n spinors $\alpha^{(1)}_{A_1},\ldots,\alpha^{(n)}_{A_n}$, which are determined only up to reordering and rescaling, are known as the principal spinors of ϕ. To prove this, note that the principal spinors can be identified with the solutions ζ^A of the equation

$$\phi_{A_1\ldots A_n}\zeta^{A_1}\cdots\zeta^{A_n} = 0$$

and there are n of these, counting multiplicities, by the "fundamental theorem of algebra."

Spinors in General Relativity: Spinor Calculus

We now want to define spinor fields on the spacetime $\mathcal{M}$ as sections of a spinor bundle $\mathcal{S}$ whose fiber at each point is S and such that the tensor product $\mathcal{S}\otimes\bar{\mathcal{S}}$ is the complexified tangent bundle. The existence of such an $\mathcal{S}$ imposes global restrictions on $\mathcal{M}$: $\mathcal{M}$ must be orientable and time orientable, and a certain characteristic class, the second Stiefel–Whitney class, must vanish (for an explanation of these terms see, e.g., Penrose and Rindler (1984, 1986)). Assuming that $\mathcal{M}$ satisfies these conditions, spinor fields can be defined. It is convenient to retain the algebraic formulas from the previous section (e.g., [10] or [12]) but with indices now regarded as abstract (a note on the abstract index convention appears in Twistors).

By an argument analogous to that for the fundamental theorem of Riemannian geometry, there is a unique covariant derivative that satisfies the Leibniz condition, coincides with the Levi-Civita derivative on tensors and the gradient on scalars, and annihilates ϵ_{AB} and $\epsilon_{A'B'}$. Following the conventions of the previous section, the spinor covariant derivative will be denoted as $\nabla_{AA'}$. The commutator of derivatives can be written in terms of irreducible parts (for $\mathrm{SL}(2,\mathbb{C})$) according to the formula

$$\nabla_{AA'}\nabla_{BB'} - \nabla_{BB'}\nabla_{AA'} = \epsilon_{A'B'}\Delta_{AB} + \epsilon_{AB}\Delta_{A'B'}$$

where $\Delta_{AB} = \nabla_{C'(A}\nabla^{C'}_{B)}$. The definition of the Riemann curvature tensor is in terms of the Ricci identity

$$(\nabla_a\nabla_b - \nabla_b\nabla_a)v^c = R_{abd}{}^c v^d$$

and then this translates into two Ricci identities for a spinor field:

$$\Delta_{AB}\alpha_C = \chi_{ABCD}\alpha^D$$

$$\Delta_{A'B'}\alpha_C = \Phi_{A'B'CD}\alpha^D$$

The curvature spinors χ_{ABCD} and $\Phi_{A'B'CD}$ are related to the curvature tensor. The Ricci spinor $\Phi_{A'B'AB}$ is Hermitian and symmetric on both index pairs and is a multiple of the trace-free part of the Ricci tensor:

$$\Phi_{A'B'AB} = -\tfrac{1}{2}\left(R_{ab} - \tfrac{1}{4}R g_{ab}\right)$$

The spinor χ_{ABCD} is symmetric on the first and last pairs of indices and decomposes into irreducibles according to

$$\chi_{ABCD} = \Psi_{ABCD} - 2\Lambda\epsilon_{D(A}\epsilon_{B)C}$$

where $\Lambda = R/24$ in terms of the Ricci scalar or scalar curvature R, while Ψ_{ABCD}, which is totally symmetric and is known as the Weyl spinor, is related to the Weyl tensor C_{abcd} by the equation

$$C_{abcd} = \Psi_{ABCD}\epsilon_{A'B'}\epsilon_{C'D'} + \bar{\Psi}_{A'B'C'D'}\epsilon_{AB}\epsilon_{CD}$$

Thus, the ten real components of the Weyl tensor are coded into the five complex components of the Weyl spinor.

Following the last remark in the previous section, the Weyl spinor has four principal spinors, each of which defines a null direction, the principal null directions of the Weyl tensor. There is a classification of Weyl tensors, the Petrov–Pirani–Penrose classification, based on coincidences among the principal null directions (Penrose and Rindler 1984, 1986).

As a final exercise in spinor calculus, we recall the zero-rest-mass equations (see Twistors). In flat spacetime, these are the equations

$$\nabla^{A'A}\phi_{AB\ldots C} = 0$$

on a totally symmetric spinor field $\phi_{AB\ldots C}$. The field is said to have spin s if it has $2s$ indices, and the cases $s = 1/2$, 1, or 2, respectively, are the Weyl neutrino equation, the Maxwell equation, and the linearized Einstein equation. In flat spacetime, these hyperbolic equations are well understood and solvable in a variety of ways. In curved spacetime, however, if $s \geq 3/2$, then there are curvature obstructions to the existence of solutions, known as Buchdahl conditions. This can be seen at once by differentiating again, say by $\nabla^B_{A'}$, and using the spinor Ricci identity. After a little algebra, one finds

$$\Psi^{ABC}{}_{(D}\phi_{E\ldots F)ABC} = 0$$

so that, whenever the field has three or more indices, there are algebraic constraints on its components in terms of the Weyl spinor.

The Spin-Coefficient Formalism

The spin-coefficient formalism of Newman and Penrose is a formalism for spinor calculus in spacetimes (see, e.g., Penrose and Rindler (1984, 1986) and Stewart (1990)). It finds application in any calculation dealing with curvature tensors, including solving the Einstein equations. The formalism exploits the compression of terminology which the introduction of complex quantities permits.

The formalism starts with a choice of spinor dyad, a basis of spinor fields (o^A, ι^A) normalized so that $o_A\iota^A = 1$. From the dyad, one constructs a null tetrad, which is a basis of vector fields, according to the scheme

$$\ell^a = o^A\bar{o}^{A'},\ n^a = \iota^A\bar{\iota}^{A'},\ m^a = o^A\bar{\iota}^{A'},\ \bar{m}^a = \iota^A\bar{o}^{A'}$$

Given the normalization of the spinor dyad, each of the vectors in the null tetrad is null (hence the name) and all inner products are zero, except for

$$\ell^a n_a = 1 = -m^a\bar{m}_a$$

It follows that the metric can be written in the basis as

$$g_{ab} = 2\ell_{(a}n_{b)} - 2m_{(a}\bar{m}_{b)}$$

The components of the covariant derivative in the null tetrad are given separate names according to the following scheme:

$$\ell^a\nabla_a = D,\ n^a\nabla_a = \Delta,\ m^a\nabla_a = \delta,\ \bar{m}^a\nabla_a = \bar{\delta}$$

and the spin coefficients are the 12 components of the covariant derivative of the basis. Each is labeled with a Greek letter according to the following scheme:

$$\begin{aligned} Do^A &= \epsilon o^A - \kappa\iota^A, & \Delta o^A &= \gamma o^A - \tau\iota^A \\ \delta o^A &= \beta o^A - \sigma\iota^A, & \bar{\delta}o^A &= \alpha o^A - \rho\iota^A \\ D\iota^A &= \pi o^A - \epsilon\iota^A, & \Delta\iota^A &= \nu o^A - \gamma\iota^A \\ \delta\iota^A &= \mu o^A - \beta\iota^A, & \bar{\delta}\iota^A &= \lambda o^A - \alpha\iota^A \end{aligned} \qquad [14]$$

The spin coefficients code the 24 real Ricci rotation coefficients into 12 complex quantities. Some of the spin coefficients have direct geometrical interpretation. For example, the vanishing of κ is the condition for the integral curves of ℓ^a to be geodesic, while, if σ is also zero, this congruence of geodesics is shear free. The same role is played by ν and λ for the n^a-congruence. The real and imaginary parts of ρ are, respectively (minus), the expansion and the twist of the congruence of integral curves of ℓ^a.

In practice, it is often simpler to calculate the spin coefficients from the commutators of the basis vectors, now regarded as directional derivatives, as follows:

$$\begin{aligned}
\Delta D - D\Delta &= (\gamma+\bar\gamma)D + (\epsilon+\bar\epsilon)\Delta - (\bar\tau+\pi)\delta - (\tau+\bar\pi)\bar\delta \\
\delta D - D\delta &= (\bar\alpha+\beta-\bar\pi)D + \kappa\Delta - (\bar\rho+\epsilon-\bar\epsilon)\delta - \sigma\bar\delta \\
\delta\Delta - \Delta\delta &= -\bar\nu D + (\tau-\bar\alpha-\beta)\Delta + (\mu-\gamma+\bar\gamma)\delta + \bar\lambda\bar\delta \\
\bar\delta\delta - \delta\bar\delta &= (\bar\mu-\mu)D + (\bar\rho-\rho)\Delta + (\alpha-\bar\beta)\delta - (\bar\alpha-\beta)\bar\delta
\end{aligned} \qquad [15]$$

The commutator of second derivatives applied to the spinor dyad expresses the components of the curvature tensor in terms of the derivatives of the spin coefficients. Before presenting these, we adopt a convention for labeling the components of curvature. The components of the Weyl spinor are given as follows:

$$\begin{aligned}
\Psi_0 &= \Psi_{ABCD} o^A o^B o^C o^D \\
\Psi_1 &= \Psi_{ABCD} o^A o^B o^C \iota^D \\
\Psi_2 &= \Psi_{ABCD} o^A o^B \iota^C \iota^D \\
\Psi_3 &= \Psi_{ABCD} o^A \iota^B \iota^C \iota^D \\
\Psi_4 &= \Psi_{ABCD} \iota^A \iota^B \iota^C \iota^D
\end{aligned} \qquad [16]$$

so that these five complex scalars encode the ten real components of the Weyl tensor. For the Ricci spinor, set

$$\begin{aligned}
\Phi_{00} &= \Phi_{ABA'B'} o^A o^B \bar o^{A'} \bar o^{B'}, & \Phi_{01} &= \Phi_{ABA'B'} o^A o^B \bar o^{A'} \bar\iota^{B'} \\
\Phi_{02} &= \Phi_{ABA'B'} o^A o^B \bar\iota^{A'} \bar\iota^{B'}, & \Phi_{11} &= \Phi_{ABA'B'} o^A \iota^B \bar o^{A'} \bar\iota^{B'} \\
\Phi_{12} &= \Phi_{ABA'B'} o^A \iota^B \bar\iota^{A'} \bar\iota^{B'}, & \Phi_{22} &= \Phi_{ABA'B'} \iota^A \iota^B \bar\iota^{A'} \bar\iota^{B'}
\end{aligned}$$

together with $\Phi_{10} = \overline{\Phi_{01}}, \Phi_{20} = \overline{\Phi_{02}}$, and $\Phi_{21} = \overline{\Phi_{12}}$. The nine components of the trace-free Ricci tensor are encoded in these scalars of which three are real and three complex. The Ricci scalar, as before, is replaced by the real scalar $\Lambda = R/24$.

Now the commutators of covariant derivatives on the spinor dyad lead to the following system:

$$\begin{aligned}
D\rho - \bar\delta\kappa &= \rho^2 + \sigma\bar\sigma + (\epsilon+\bar\epsilon)\rho - \bar\kappa\tau \\
&\quad - (3\alpha+\bar\beta-\pi)\kappa + \Phi_{00} \\
D\sigma - \delta\kappa &= (\rho+\bar\rho+3\epsilon-\bar\epsilon)\sigma \\
&\quad - (\tau-\bar\pi+\bar\alpha+3\beta)\kappa + \Psi_0 \\
D\tau - \Delta\kappa &= (\tau+\bar\pi)\rho + (\bar\tau+\pi)\sigma + (\epsilon-\bar\epsilon)\tau \\
&\quad - (3\gamma+\bar\gamma)\kappa + \Psi_1 + \Phi_{01} \\
D\alpha - \bar\delta\epsilon &= (\rho+\bar\epsilon-2\epsilon)\alpha + \beta\bar\sigma - \bar\beta\epsilon - \kappa\lambda - \bar\kappa\gamma \\
&\quad + (\epsilon+\rho)\pi + \Phi_{10} \\
D\beta - \delta\epsilon &= (\alpha+\pi)\sigma + (\bar\rho-\bar\epsilon)\beta - (\mu+\gamma)\kappa \\
&\quad - (\bar\alpha-\bar\pi)\epsilon + \Psi_1 \\
D\gamma - \Delta\epsilon &= (\tau+\bar\pi)\alpha + (\bar\tau+\pi)\beta - (\epsilon+\bar\epsilon)\gamma - (\gamma+\bar\gamma)\epsilon \\
&\quad + \tau\kappa - \nu\kappa + \Psi_2 - \Lambda + \Phi_{11} \\
D\lambda - \bar\delta\pi &= (\rho-3\epsilon+\bar\epsilon)\lambda + \bar\sigma\mu + (\pi+\alpha-\bar\beta)\pi \\
&\quad - \nu\bar\kappa + \Phi_{20} \\
D\mu - \delta\pi &= (\bar\rho-\epsilon-\bar\epsilon)\mu + \sigma\lambda + (\bar\pi-\bar\alpha+\beta)\pi \\
&\quad - \nu\kappa + \Psi_2 + 2\Lambda \\
D\nu - \Delta\pi &= (\pi+\bar\tau)\mu + (\bar\pi+\tau)\lambda + (\gamma-\bar\gamma)\pi \\
&\quad - (3\epsilon+\bar\epsilon)\nu + \Psi_3 + \Phi_{21} \\
\Delta\lambda - \bar\delta\nu &= -(\mu+\bar\mu+3\gamma-\bar\gamma)\lambda \\
&\quad + (3\alpha+\bar\beta+\pi-\bar\tau)\nu - \Psi_4 \\
\delta\rho - \bar\delta\sigma &= (\bar\alpha+\beta)\rho - (3\alpha-\bar\beta)\sigma + (\rho-\bar\rho)\tau \\
&\quad + (\mu-\bar\mu)\kappa - \Psi_1 + \Phi_{01} \\
\delta\alpha - \bar\delta\beta &= \mu\rho - \lambda\sigma + \alpha\bar\alpha + \beta\bar\beta - 2\alpha\beta + (\rho-\bar\rho)\gamma \\
&\quad + (\mu-\bar\mu)\epsilon - \Psi_2 + \Lambda + \Phi_{11} \\
\delta\lambda - \bar\delta\mu &= (\rho-\bar\rho)\nu + (\mu-\bar\mu)\pi + (\alpha+\bar\beta)\mu \\
&\quad + (\bar\alpha-3\beta)\lambda - \Psi_3 + \Phi_{21} \\
\Delta\mu - \delta\nu &= -(\mu+\gamma+\bar\gamma)\mu - \lambda\bar\lambda + \bar\nu\pi \\
&\quad + (\bar\alpha+3\beta-\tau)\nu - \Phi_{22} \\
\Delta\beta - \delta\gamma &= (\bar\alpha+\beta-\tau)\gamma - \mu\tau + \sigma\nu + \epsilon\bar\nu \\
&\quad + (\gamma-\bar\gamma-\mu)\beta - \alpha\bar\lambda - \Phi_{12} \\
\Delta\sigma - \delta\tau &= -(\mu-3\gamma+\bar\gamma)\sigma - \bar\lambda\rho - (\tau+\beta-\bar\alpha)\tau \\
&\quad + \kappa\bar\nu - \Phi_{02} \\
\Delta\rho - \bar\delta\tau &= (\gamma+\bar\gamma-\bar\mu)\rho - \sigma\lambda + (\bar\beta-\alpha-\bar\tau)\tau \\
&\quad + \nu\kappa - \Psi_2 - 2\Lambda \\
\Delta\alpha - \bar\delta\gamma &= (\rho+\epsilon)\nu - (\tau+\beta)\lambda + (\bar\gamma-\bar\mu)\alpha \\
&\quad + (\bar\beta-\bar\tau)\gamma - \Psi_3
\end{aligned} \qquad [17]$$

Finally, it is possible to write out the Bianchi identities in this formalism. For simplicity, and with a view to an application, we do this below only for vacuum, so that the Ricci tensor is zero:

$$\begin{aligned}
D\Psi_1 - \bar\delta\Psi_0 &= (\pi-4\alpha)\Psi_0 + 2(2\rho+\epsilon)\Psi_1 - 3\kappa\Psi_2 \\
\Delta\Psi_0 - \delta\Psi_1 &= (4\gamma-\mu)\Psi_0 - 2(2\tau+\beta)\Psi_1 + 3\sigma\Psi_2 \\
D\Psi_2 - \bar\delta\Psi_1 &= -\lambda\Psi_0 + 2(\pi-\alpha)\Psi_1 + 3\rho\Psi_2 - 2\kappa\Psi_3 \\
\Delta\Psi_1 - \delta\Psi_2 &= \nu\Psi_0 + 2(\gamma-\mu)\Psi_1 - 3\tau\Psi_2 + 2\sigma\Psi_3 \\
D\Psi_3 - \bar\delta\Psi_2 &= -2\lambda\Psi_1 + 3\pi\Psi_2 + 2(\rho-\epsilon)\Psi_3 - \kappa\Psi_4 \\
\Delta\Psi_2 - \delta\Psi_3 &= 2\nu\Psi_1 - 3\mu\Psi_2 + 2(\beta-\tau)\Psi_3 + \sigma\Psi_4 \\
D\Psi_4 - \bar\delta\Psi_3 &= -3\lambda\Psi_2 + 2(\alpha+2\pi)\Psi_3 + (\rho-4\epsilon)\Psi_4 \\
\Delta\Psi_3 - \delta\Psi_4 &= 3\nu\Psi_2 - 2(\gamma+2\mu)\Psi_3 + (4\beta-\tau)\Psi_4
\end{aligned} \qquad [18]$$

The whole system is then loosely described as the spin-coefficient equations.

As a simple application, we shall prove the Goldberg–Sachs theorem: for vacuum spacetimes, a

spinor field o^A is geodesic and shear free iff it is a repeated principal spinor of the Weyl spinor.

In the spin-coefficient formalism, o^A is geodesic and shear-free iff κ and σ vanish, and, from [16], is a repeated principal spinor of the Weyl spinor provided $\Psi_0 = \Psi_1 = 0$. It will be repeated three times if also $\Psi_2 = 0$ and four times if $\Psi_3 = 0$, but one must have $\Psi_k \neq 0$ for some k if the spacetime is not to be flat.

Suppose that o^A is a (twice) repeated principal spinor of the Weyl spinor, then at once from the first two expressions in [18] both κ and σ vanish. If it is repeated three times, one gets the same result from the third and fourth expressions in [18], while if o^A is repeated four times then the fifth and sixth expressions of [18] should be used.

For the converse, suppose that $\kappa = \sigma = 0$. Then, by the first equation in [14], o^A can be rescaled to ensure that $\epsilon = 0$ and a spinor field ι^A can be chosen which is normalized against o^A and parallelly propagated along ℓ^a, so that, by the fifth equation in [14], $\pi = 0$. From the second expression in [17], one can see at once that $\Psi_0 = 0$, so that the first two equations in [18] simplify to give expressions for $D\Psi_1$ and $\delta\Psi_1$. By commuting D and δ on Ψ_1 and using the second expression of [15] with the relevant parts of [17], it can be concluded that $\Psi_1 = 0$, as required.

Another application which is easy to describe is the solution of the type-D vacuum equations. A type-D solution is one for which the Weyl spinor has two (linearly independent) repeated principal spinors. If these are taken as the normalized dyad, then from [16] only Ψ_2 is nonzero among the Ψ_k. By the Goldberg–Sachs theorem, both spinors are geodesic and shear free, so that the spin coefficients σ, κ, λ, and ν all vanish. With these conditions, the spin-coefficient equations simplify to the point that careful choices of coordinates and the remaining freedom in the dyad enable the equations to be solved explicitly. One obtains metrics that depend only on a few parameters. Analogous methods reduce the Einstein equations to simpler systems for the other vacuum algebraically special metrics, that is, the other vacuum metrics for which the Weyl spinor does not have four distinct principal null directions (Mason 1998).

The spin-coefficient formalism has also been extensively used in the study of asymptotically flat spacetimes and gravitational radiation (Penrose and Rindler 1984, 1986, Stewart 1990).

The Positive-Mass Theorem

A very important application of spinor calculus in recent years was the proof by Witten (1981) of the positive-mass (or positive-energy) theorem. The proof was motivated by ideas from supergravity and gave rise to an increased interest in spinors in general relativity.

The positive-mass theorem is the following assertion: *given an asymptotically flat spacetime $\mathcal{M}$ with a spacelike hypersurface Σ, which is topologically $\mathbb{R}^3$ and in which the dominant energy condition holds, the total (or Arnowitt–Deser–Misner (ADM)) momentum is timelike and future-pointing*. (The dominant-energy condition is the requirement that $T_{ab}U^aV^b$ is non-negative for every pair of future-pointing timelike or null vectors U^a and V^b.)

We follow the notation of Penrose and Rindler (1984, 1986), where the proof begins by considering the 2-form Ξ defined in terms of a spinor field λ^A on Σ by

$$\Xi = -\mathrm{i}\bar{\lambda}_{B'}\nabla_a\lambda_B \mathrm{d}x^a \wedge \mathrm{d}x^b$$

If λ^a tends to a constant spinor at spatial infinity on Σ, then

$$\frac{1}{4\pi G}\oint_S \Xi \to p_a\lambda^A\bar{\lambda}^{A'} \qquad [19]$$

as the spacelike spherical surface S tends to spatial infinity, where p_a is the ADM momentum. Suppose Σ has unit normal t^a, intrinsic metric $h_{ab} = g_{ab} - t_at_b$ and the dual-volume 3-form is $\mathrm{d}\Sigma^a = t^a\mathrm{d}\Sigma$. Then Stokes' theorem states that

$$\oint_S \Xi = \int_\Sigma \mathrm{d}\Xi$$

We calculate

$$\mathrm{d}\Xi = \alpha + \beta$$

where

$$\alpha = 4\pi G T_{ab}\ell^a\mathrm{d}\Sigma^b$$
$$\beta = -\mathrm{i}\epsilon_{ab}{}^{cd}\nabla_c\lambda^B\nabla_d\bar{\lambda}^{B'}\mathrm{d}\Sigma^a$$

where $\ell^a = \lambda^a\bar{\lambda}^{A'}$ and we have used the Einstein field equations to replace curvature terms in α by the energy–momentum tensor T_{ab}. Provided the matter satisfies the dominant-energy condition, α is everywhere a positive multiple of the volume form on Σ and its integral is positive (it can vanish only in vacuum). To make the integral of β positive, λ^A is required to satisfy

$$D_{AA'}\lambda^A = 0 \qquad [20]$$

where $D_a = h_a^b\nabla_b$, which is the projection of the four-dimensional covariant derivative rather than the intrinsic covariant derivative of Σ. Equation [20] is the Sen–Witten equation; it is elliptic and reduces to

the Dirac equation on a maximal surface; furthermore, given an asymptotically constant value for λ^A on an asymptotically flat 3-surface Σ with the topology of $\mathbb{R}^3$, it has a unique solution. Equation [20] removes part of the derivative of λ^A from β to leave

$$\beta = -h^{ab} D_a \lambda_C D_b \bar{\lambda}_{C'} \mathrm{d}\Sigma^c$$

Now h_{ab} is negative definite and Σ has timelike normal so that β is a positive multiple of the volume form on Σ (unless λ^A is covariantly constant, a case which is dealt with separately). Thus, the integral of $\mathrm{d}\Xi$ is non-negative and therefore, by [19], so is the inner product of the ADM momentum p_a with any null vector constructed from asymptotically constant spinors. Furthermore, this inner product is strictly positive, except in a vacuum spacetime admitting a constant spinor. Such spacetimes can be found explicitly and cannot be asymptotically flat, so that the ADM momentum is always timelike and future pointing, and vanishes only in flat spacetime.

The basic positive-energy theorem outlined above can be extended in several directions:

- to prove that the total momentum at future null infinity is also timelike and future pointing;
- to deal with surfaces Σ which have inner boundaries, for example, at black holes;
- to prove inequalities between charge and mass; and
- to deal with spacetimes which are asymptotically anti-de Sitter rather than flat.

Further Applications of Spinors

Supersymmetry is a symmetry in quantum field theory relating bosons and fermions. In the language of spinors, bosons are represented by fields with an even number of spinor indices and fermions by fields with an odd number of indices. Thus, the gauge transformations of supersymmetry are generated by spinors with a single index.

Supergravity is supersymmetry in the case that one of the fields is the graviton. A supergravity theory is labeled by an integer N for the number of independent supersymmetries and much of the numerology of these theories follows from properties of spinors. $N = 1$ supergravity contains a graviton and a spin-3/2 field coupled together, and the presence of the supersymmetry allows the Buchdahl condition to be evaded. Supergravity theory with one supersymmetry in 11 spacetime dimensions depends on one spinor, which, in 11 dimensions, has 32 components. This is as many components as eight Dirac spinors in a four-dimensional spacetime, and, by a process of dimensional reduction, $N = 1$ supergravity in 11 dimensions is related to $N = 8$ supergravity in four dimensions. For reasons related to the Buchdahl conditions, 8 is the largest N that is considered in four dimensions.

In superstring theory and in some supergravity theories, one often wishes to consider spaces with "residual supersymmetry," by which is meant that there is a spinor field satisfying a condition of covariant constancy in some connection (Candelas *et al.* 1985). The existence of such constant spinors, as a result of spinor Ricci identities analogous to those given above, typically imposes strong restrictions on the curvature. Riemannian manifolds admitting constant spinors for the Levi-Civita connection are Ricci-flat (Hitchin 1974); Lorentzian ones can often be found in terms of a few functions. Manifolds of special holomorphy, which are of interest in superstring theory, can usually be characterized as admitting special spinors (Wang 1989).

See also: Clifford Algebras and Their Representations; Dirac Operator and Dirac Field; Einstein Equations: Exact Solutions; Einstein's Equations with Matter; General Relativity: Overview; Geometric Flows and the Penrose Inequality; Index Theorems; Relativistic Wave Equations Including Higher Spin Fields; Spacetime Topology, Causal Structure and Singularities; Supergravity; Twistor Theory: Some Applications [in Integrable Systems, Complex Geometry and String Theory]; Twistors.

Further Reading

Benn IM and Tucker RW (1987) *An Introduction to Spinors and Geometry with Applications in Physics*. Bristol: Adam Hilger.

Budinich P and Trautman A (1988) *The Spinorial Chessboard*. Berlin: Springer.

Candelas P, Horowitz G, Strominger A, and Witten E (1985) Vacuum configurations for superstrings. *Nuclear Physics B* 258: 46–74.

Cartan E (1981) *The Theory of Spinors*. New York: Dover.

Chevalley CC (1954) *The Algebraic Theory of Spinors*. New York: Columbia University Press.

Dirac PAM (1928) The quantum theory of the electron. *Proceedings of the Royal Society of London* A 117: 610–624.

Harvey FR (1990) *Spinors and Calibrations*. Boston: Academic Press.

Hitchin NJ (1974) Harmonic spinors. *Advances in Mathematics*. 14: 1–55.

Mason LJ (1998) The asymptotic structure of algebraically special spacetimes. *Classical Quantum Gravity* 15: 1019–1030.

Penrose R and Rindler W (1984, 1986) *Spinors and Space–Time*. vol. 1 and 2. Cambridge: Cambridge University Press.

Stewart J (1990) *Advanced General Relativity*. Cambridge: Cambridge University Press.

van der Waerden BL (1960) Exclusion principle and spin. In: Fierz M and Weisskopf VF (eds.) *Theoretical Physics in the Twentieth Century: A Memorial Volume to Wolfgang Pauli*, pp. 199–244. New York: Interscience.

Wang MY (1989) Parallel spinors and parallel forms. *Annals of Global Analysis and Geometry* 7: 59–68.

Witten E (1981) A new proof of the positive energy theorem. *Communications in Mathematical Physics* 80: 381–402.