EMIL WOLF

EDITOR

PROGRESS IN OPTICS

VOLUME XI

CONTRIBUTORS

G. S. AGARWAL
H. YOSHINAGA
E. G. LEAN
O. BRYNGDAHL
A. V. CREWE
J. A. ARNAUD
E. W. MARCHAND

NORTH - HOLLAND PUBLISHING COMPANY - AMSTERDAM - LONDON

PROGRESS IN OPTICS

VOLUME XI

PROGRESS IN OPTICS

VOLUME XI

EDITED BY

E. WOLF

University of Rochester, N.Y., U.S.A.

Contributors

G. S. AGARWAL, H. YOSHINAGA,
E. G. LEAN, O. BRYNGDAHL,
A. V. CREWE, J. A. ARNAUD, E. W. MARCHAND

1973

NORTH-HOLLAND PUBLISHING COMPANY – AMSTERDAM · LONDON

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 61-19297
NORTH-HOLLAND ISBN: 0 7204 1511 x
AMERICAN ELSEVIER ISBN: 0 444 10497 6

PUBLISHERS:

NORTH-HOLLAND PUBLISHING COMPANY - AMSTERDAM
NORTH-HOLLAND PUBLISHING COMPANY, LTD. - LONDON

SOLE DISTRIBUTORS FOR THE U.S.A. AND CANADA:

AMERICAN ELSEVIER PUBLISHING COMPANY, INC.
52 VANDERBILT AVENUE
NEW YORK, N.Y. 10017

PRINTED IN THE NETHERLANDS

CONTENTS OF VOLUME I (1961)

CONTENTS OF VOLUME II (1963)

CONTENTS OF VOLUME III (1964)

CONTENTS OF VOLUME IV (1965)

CONTENTS OF VOLUME V (1966)

CONTENTS OF VOLUME VI (1967)

CONTENTS OF VOLUME VII (1969)

CONTENTS OF VOLUME VIII (1970)

CONTENTS OF VOLUME IX (1971)

CONTENTS OF VOLUME X (1972)

This Page Intentionally Left Blank

PREFACE

Just like its ten predecessors, the present volume of PROGRESS IN OPTICS reviews research activities in various branches of modern optics and related fields.

In the opening article G. A. Agarwal presents an account of the master equation techniques that are more and more frequently being applied to problems arising in the theory of the laser and of the interaction between laser beams and atomic systems. In the second article, H. Yoshinaga describes developments in the spectroscopic techniques in the far infrared. In the next article E. G. Lean discusses two main areas in the field of the interaction of light and acoustical waves: light diffraction by acoustic surface waves and the interaction of optical guided waves and acoustic surface waves. An article by O. Bryngdahl presents an account of the theory and novel uses of evanescent waves. These waves, whose amplitudes decay in space with position, are usually generated at sharp boundaries, or in the process of interaction of ordinary waves with a material medium. In the fifth article, A. V. Crewe discusses the production of electron probes using a field emission source. Probes of this kind, based on the quantum mechanical tunnelling process of electrons that arises close to a metal surface, are finding good uses in scanning microscopes and microanalysis. In the subsequent article J. A. Arnaud discusses novel methods for the theoretical analysis of optical beams. In the concluding article E. W. Marchand reviews the theory of lenses formed by inhomogeneous media. Until recent times lenses of this kind were largely regarded by lens designers as mathematical curiosities. However, modern technological developments promise practical realization of such optical elements.

Department of Physics and Astronomy
University of Rochester, N.Y., 14627
April 1973

EMIL WOLF

This Page Intentionally Left Blank

CONTENTS

I. MASTER EQUATION METHODS IN QUANTUM OPTICS

by G. S. AGARWAL (Rochester, N.Y.)

II. RECENT DEVELOPMENTS IN FAR INFRARED SPECTROSCOPIC TECHNIQUES

by H. YOSHINAGA (Osaka)

III. INTERACTION OF LIGHT AND ACOUSTIC SURFACE WAVES

by E. G. Lean

IV. EVANESCENT WAVES IN OPTICAL IMAGING

by O. Bryngdahl (Palo Alto, Calif.)

V. PRODUCTION OF ELECTRON PROBES USING A FIELD EMISSION SOURCE

by A. V. Crewe (Chicago, Ill.)

VI. HAMILTONIAN THEORY OF BEAM MODE PROPAGATION

by J. A. Arnaud (Holmdel, N.J.)

VII. GRADIENT INDEX LENSES

by E. W. Marchand (Rochester, N.Y.)

This Page Intentionally Left Blank

E. WOLF, PROGRESS IN OPTICS XI © NORTH-HOLLAND 1973

I

MASTER EQUATION METHODS IN QUANTUM OPTICS*

BY

G. S. AGARWAL †

Department of Physics and Astronomy,
University of Rochester, Rochester, N. Y., 14627, U.S.A.

* Work supported by the U.S. Air Force Office of Scientific Research and by the U.S. Army Research Office (Durham).

† Present address: Institut für Theoretische Physik, Universität Stuttgart, Stuttgart, Germany.

CONTENTS

§ 1. Introduction

In recent years increasing use has been made of methods of quantum statistical mechanics and stochastic processes in treatments of various problems in quantum optics. This is seen from a large number of publications dealing with the theory of lasers (GORDON [1967], HAKEN [1970], LAX [1966c, 1968a], SCULLY and LAMB [1967]), with superradiance (AGARWAL [1970; 1971b, c, e], BONIFACIO et al. [1971a, b]), with problems in nonlinear optics such as parametric oscillators (GRAHAM [1968]). Moreover some of the methods were specifically developed to treat the problems in quantum optics. These include the well known phase space methods (GLAUBER [1963, 1965], SUDARSHAN [1963], CAHILL and GLAUBER [1969a, b], AGARWAL and WOLF [1968, 1970a, b, c], LAX [1968b]). In phase space methods the c-number distribution functions for quantum systems are introduced, which in many physical situations are found to obey equations of the Fokker-Planck type. One may then use the language of stochastic processes to study various quantum systems. In quantum optics, one is usually concerned with the study of a subsystem which is a part of a large system, for example, in case of the laser one is mainly interested in the statistical properties of the emitted radiation. In this context master equation methods have played a very important role.

Master equation methods have found applications in many branches of physics such as in the theory of relaxation processes (BLOCH [1956, 1957], REDFIELD [1957, 1965], ABRAGAM [1962], ARGYRES and KELLEY [1964]), anharmonic interaction in solids (BROUT and PRIGOGINE [1956], PRIGOGINE [1962], CARRUTHERS and DY [1966]), superfluids (LANGER [1969]), transport phenomenon (CHESTER and THELLUNG [1959], VANHOVE [1959], KOHN and LUTTINGER [1957], ARGYRES [1966]), optical pumping (WILLIS [1970]), superradiance (AGARWAL [1970, 1971b, c, e], BONIFACIO et al. [1971a, b]), Brownian motion of a quantum oscillator (AGARWAL [1971d]), in the quantum theory of damping (LOUISELL and MARBURGER [1967], AGARWAL [1969]), in the kinetic theory of gases (see, for example, PRIGOGINE [1962]), in the theory of lasers (LAX [1966c], HAAKE [1969b], HAKEN [1970]), etc.

We start by giving a brief history of the subject. A master equation was first obtained by PAULI [1928]. He obtained an equation of motion for the diagonal elements ρ_{nn} of the density operator by making the statistical hypothesis of random phases at all times. Pauli's master equation has the form

$$\frac{\partial \rho_{nn}}{\partial t} = \sum_{m \neq n} \gamma_{nm} \rho_{mm} - \sum_{m \neq n} \gamma_{mn} \rho_{nn}, \tag{1.1}$$

where γ_{nm} is the transition probability per unit time for the system to make a transition from the state $|m\rangle$ to the state $|n\rangle$. It should be noted that ρ_{nn} is the probability that the system be found in the state $|n\rangle$ and, therefore, (1.1) is of the form of a rate equation. An equation of the type (1.1) for ρ_{nn} is also expected from the principle of detailed balance.

We will now outline Pauli's derivation. Our presentation follows that of VAN HOVE [1962]. We write the Hamiltonian of the system as

$$H = H_0 + gH_1, \tag{1.2}$$

where H_0 is the unperturbed Hamiltonian and gH_1 is a small perturbation. We work in a representation in which H_0 is diagonal, with eigenfunctions $|\psi_n\rangle$ and eigenvalues E_n. The state of the system at time $t+\Delta t$ is related to the state at time t by the unitary transformation

$$|\psi(t+\Delta t)\rangle = \exp\{-\mathrm{i}(H_0+gH_1)\Delta t\}|\psi(t)\rangle, \tag{1.3}$$

where $\hbar$ has been put equal to unity and this we do throughout this article. On expanding $|\psi(t+\Delta t)\rangle$ and $|\psi(t)\rangle$ in a complete set of states

$$|\psi(t+\Delta t)\rangle = \sum_n c_n(t+\Delta t)|\psi_n\rangle, \qquad |\psi(t)\rangle = \sum_n c_n(t)|\psi_n\rangle, \tag{1.4}$$

we find that the probability that the system be found in a state $|\psi_n\rangle$ at time $t+\Delta t$ is given by

$$\rho_{nn}(t+\Delta t) \equiv p_n(t+\Delta t) = |c_n(t+\Delta t)|^2$$
$$= \sum_{l,m} c_m(t)c_l^*(t)\langle\psi_n|\exp\{-\mathrm{i}(H_0+gH_1)\Delta t\}|\psi_m\rangle\langle\psi_l|\exp\{\mathrm{i}(H_0+gH_1)\Delta t\}|\psi_n\rangle. \tag{1.5}$$

Pauli made the statistical assumption of random phase at all times. This assumption enables us to ignore all the terms in (1.5) with $m \neq l$. In fact the assumption of random phases implies that the terms in (1.5) with $m \neq l$ oscillate very rapidly and so they average out to zero. Eq. (1.5) then reduces to

$$p_n(t+\Delta t) = \sum_m |\langle\psi_n|\exp\{-\mathrm{i}(H_0+gH_1)\Delta t\}|\psi_m\rangle|^2 p_m(t). \tag{1.6}$$

Since Δt is small enough the expression $|\langle\psi_n| \exp\{-i(H_0+gH_1)\Delta t\}|\psi_m\rangle|^2$ for $n \neq m$ is easily evaluated by first order perturbation theory and one finds that

$$|\langle\psi_n| \exp\{-i(H_0+gH_1)\Delta t\}|\psi_m\rangle|^2 = \gamma_{nm}\Delta t, \qquad (n \neq m), \tag{1.7}$$

where

$$\gamma_{nm} = 2\pi g^2 \delta(E_n - E_m)|\langle\psi_n|H_1|\psi_m\rangle|^2. \tag{1.8}$$

Equation (1.6) in the limit $\Delta t \to 0$ leads to the Pauli master equation (1.1) if use is also made of the following identity

$$\sum_m |\langle\psi_n| \exp\{-i(H_0+gH_1)\Delta t\}|\psi_m\rangle|^2 = 1. \tag{1.9}$$

In the derivation it has also been assumed that the spectrum is continuous. Pauli's equation is valid for times such that $\tau_c \ll t \ll t_{relax}$, where τ_c is the interaction time and t_{relax} is the relaxation time. The 'interaction time' for the problem of anharmonic interaction in solids, for example, is of the order of $(\omega_D)^{-1}$, where ω_D is the Debye frequency (see e.g. PRIGOGINE [1962]).

In the past two decades, there has been revival of interest in deriving the master equations by making far less reaching assumptions than Pauli did. The work on the derivation of the master equations can mainly be divided into three groups: (i) VAN HOVE [1955, 1957, 1962], (ii) Prigogine and his co-workers (see e.g. PRIGOGINE [1962]) and (iii) ZWANZIG [1961a, 1964]. VanHove appears to have been the first to give a rigorous derivation of the Pauli equation. He also derived master equations to all orders in perturbation. Prigogine et al. made extensive use of diagrammatic methods to derive master equations. Zwanzig developed very elegent projection operator methods for deriving the master equations to all orders in perturbation and also established the identity of various other master equations (MONTROLL [1961], PRIGOGINE and RESIBOIS [1961]).

In the present article we will not be concerned with rigorous derivation of master equations but rather in reviewing research that demonstrates the power of the master equation approach in the study of many problems in quantum optics (for rigorous derivation of master equations see for example, VAN KAMPEN [1954], VAN HOVE [1955, 1957], EMCH and SEWELL [1968]). We will show how these techniques may be employed in studies of the relaxation of oscillators and spin systems (two-level atomic systems), of lasers and of superradiance. We will make use of Zwanzig's projection operator techniques to obtain master equations for a wide variety of systems.

The density operator ρ characterizing the state of a quantum mechanical system satisfies the equation of motion

$$\partial\rho/\partial t = -\mathrm{i}\mathscr{L}\rho, \tag{1.10}$$

where the Liouville operator $\mathscr{L}$ is given by

$$\mathscr{L} \equiv [H, \quad]. \tag{1.11}$$

Zwanzig noted that the part of the density operator which is of interest can be obtained from the total density operator by suitably projecting it. To obtain an equation of motion for the diagonal elements of ρ ZWANZIG [1961a, 1964] introduced the following projection operator

$$\mathscr{P}_{mnm'n'} = \delta_{mn}\delta_{mm'}\delta_{nn'}. \tag{1.12}$$

Zwanzig regarded $\mathscr{L}$ and $\mathscr{P}$ tetradics. However, there is no need for introducing the notion of a tetradic, one may very well work instead with the projection operator

$$\mathscr{P} \ldots = \sum_m \mathscr{P}_m \operatorname{Tr}\{\mathscr{P}_m, \ldots\}, \tag{1.13}$$

where $\mathscr{P}_m$ is the projection operator onto the state $|m\rangle$, i.e.

$$\mathscr{P}_m = |m\rangle\langle m|. \tag{1.14}$$

It is seen from (1.13) that $\mathscr{P}$ projects out the diagonal elements of ρ, since the off diagonal elements of the operator $\mathscr{P}\rho$ are identically equal to zero.

We can treat both classical and quantum systems by using the projection operator methods. In classical statistical mechanics the distribution function $\Phi_N(\{q\}, \{p\}; t)$ satisfies the Liouville equation of motion

$$\frac{\partial\Phi_N}{\partial t} = \sum_i \left(\frac{\partial H}{\partial q_i}\frac{\partial}{\partial p_i} - \frac{\partial H}{\partial p_i}\frac{\partial}{\partial q_i}\right)\Phi_N, \tag{1.15}$$

which can be rewritten in the form

$$\partial\Phi_N/\partial t = -\mathrm{i}\mathscr{L}\Phi_N, \tag{1.16}$$

where the Liouville operator $\mathscr{L}$ is given by

$$\mathscr{L} = \mathrm{i}\sum_i \left(\frac{\partial H}{\partial q_i}\frac{\partial}{\partial p_i} - \frac{\partial H}{\partial p_i}\frac{\partial}{\partial q_i}\right). \tag{1.17}$$

In many problems one is only interested in the momentum distribution function $\Phi_N(\{p\}; t)$ which is obtained from $\Phi_N(\{q\}, \{p\}; t)$ by integrating over the volume V of the system:

$$\Phi_N(\{p\}; t) = \int \ldots \int \Phi_N(\{q\}, \{p\}; t)\mathrm{d}(\{q\}). \tag{1.18}$$

The relation (1.18) can be written as

$$\Phi_N(\{p\}; t) = V^N \mathscr{P} \Phi_N(\{q\}, \{p\}; t), \tag{1.19}$$

where $\mathscr{P}$ is the projection operator

$$\mathscr{P} \ldots = \frac{1}{V^N} \int \ldots \int \mathrm{d}(\{q\}) \ldots . \tag{1.20}$$

In other problems, such as anharmonic interaction in solids, one is interested in calculating the energy distribution. In this case one introduces the action-angle variables J_i and φ_i and regards the distribution as a function of $\{J_i\}$ and $\{\varphi_i\}$. Then the energy distribution is given by

$$\Phi_N(\{J_i\}; t) = (2\pi)^N \mathscr{P} \Phi_N(\{J_i\}, \{\varphi_i\}; t), \tag{1.21}$$

where

$$\mathscr{P} \ldots = \frac{1}{(2\pi)^N} \int \ldots \int \mathrm{d}(\{\varphi_i\}) \ldots . \tag{1.22}$$

In case of a system interacting with stochastic perturbations, such as a randomly modulated harmonic oscillator, the density operator of the distribution function becomes a random function. It is then preferable to work with the ensemble average of the density operator. The appropriate projection operator for this problem is

$$\mathscr{P} G = \bar{G}, \tag{1.23}$$

where the bar denotes the ensemble average with respect to the distribution of stochastic perturbations. Other examples of the projection operator will be given in subsequent sections.

We begin in § 2 with a brief account of the phase space methods. We introduce the concept of phase space distribution functions for quantum systems, a concept which was originally introduced by WIGNER [1932] and later studied in great detail and generalized by a large number of workers (MOYAL [1949], GROENEWOLD [1946], GLAUBER [1963], SUDARSHAN [1963], MEHTA and SUDARSHAN [1965], CAHILL and GLAUBER [1969a, b], AGARWAL and WOLF [1968, 1970a, b, c]). We also give in § 2 the equations of motion for these phase space distribution functions. In §§ 3–5 master equations for both classical and quantum systems are obtained. The application of these master equations to various problems in quantum optics is considered in § 4 and §§ 6–12. The problems treated in detail include relaxation of an oscillator and an atom, Brownian motion of a quantum oscillator, superradiance and relaxation of spin systems, theory of a single mode laser, systems interacting with intense external fields, parametric frequency conver-

sion, parametric oscillator and a related problem of anharmonic interaction in solids. Finally an appendix is devoted to the study of the properties of the Fokker-Planck equations.

§ 2. Phase Space Methods

In this section we present a summary of phase space methods in quantum mechanics. Consider a quantum mechanical system with one degree of freedom. Let a and a^+ be the annihilation and the creation operators* satisfying the boson commutation relations, i.e.

$$[a, a^+] = 1, \qquad [a, a] = [a^+, a^+] = 0. \tag{2.1}$$

Let ρ be the density operator characterizing the state of the system under consideration. It is possible to introduce a c-number distribution function, which we call the *phase space distribution function*, corresponding to the given density operator. The c-number distribution function is not unique. It depends on the rule of mapping that is adopted to map the operators onto c-numbers and vice versa. For a given rule of mapping, there is one and only one distribution function corresponding to a given density operator. In what follows, we will be considering only the normal, the antinormal and the Weyl rules of mapping (for a general theory see AGARWAL and WOLF [1968, 1970a, b, c]; see also LAX [1968b], CAHILL and GLAUBER [1969a, b]).

Let Ω be a linear mapping operator that transforms an arbitrary function of the c-numbers z and z^* (z^* being the complex conjugate of z) onto an operator function of a and a^+ and let Θ be the linear mapping operator, inverse to Ω, which transforms an arbitrary function of the operators a and a^+ onto a c-number function of z and z^*. Here z and z^* are the c-numbers onto which the operators a and a^+ are mapped, i.e.

$$z \rightleftarrows a, \qquad z^* \rightleftarrows a^+. \tag{2.2}$$

The mapping operators for the normal, the antinormal and the Weyl rules are now given by

$$z^{*m} z^n \xrightarrow{\Omega^{(N)}} a^{+m} a^n \xrightarrow{\Theta^{(N)}} z^{*m} z^n, \tag{2.3}$$

$$z^{*m} z^n \xrightarrow{\Omega^{(A)}} a^n a^{+m} \xrightarrow{\Theta^{(A)}} z^{*m} z^n, \tag{2.4}$$

$$z^{*m} z^n \xrightarrow{\Omega^{(W)}} (a^{+m} a^n)_W \xrightarrow{\Theta^{(W)}} z^{*m} z^n, \tag{2.5}$$

where the superscripts N, A and W on Ω and Θ stand for the normal, the antinormal and the Weyl rules of mapping. In eq. (2.5) $(a^{+m}a^n)_W$ denotes

* The superscript + denotes the Hermitean adjoint throughout this article.

the Weyl symmetrized product, i.e. $(a^{+m}a^n)_W$ is equal to the sum of all possible products involving m a^+'s and n a's divided by the total number of such terms. It is thus seen from (2.3) and (2.4) that the normal (antinormal) rule brings all the creation (annihilation) operators to the left of all the annihilation (creation) operators.

Let $\Phi^{(\Omega)}(z, z^*)$ be the phase space distribution function associated with the density operator according to a particular rule. It is obtained from ρ/π by the Θ rule of mapping, i.e.

$$\frac{\rho}{\pi} \xrightarrow{\Theta} \Phi^{(\Omega)}(z, z^*) \xrightarrow{\Omega} \frac{\rho}{\pi}. \tag{2.6}$$

The phase space distribution function so obtained is not necessarily non-negative and may even be singular, and must, in general, be regarded as a function in the sense of generalized function theory. We give below explicit expressions for the distribution functions $\Phi^{(W)}$, $\Phi^{(A)}$ and $\Phi^{(N)}$.

The distribution function $\Phi^{(W)}$, which is obtained from ρ via the Weyl rule of mapping, is known as the Wigner distribution function. In coordinate-momentum representation it was first introduced by WIGNER [1932] and studied extensively by MOYAL [1949] and GROENEWOLD [1946] (see also KUBO [1964]). For a given density operator the Wigner distribution function always exists. The following explicit expression in terms of coherent states was given by AGARWAL and WOLF [1970a]

$$\Phi^{(W)}(z, z^*) = \frac{2}{\pi^2} \exp(2|z|^2) \int \langle -\alpha|\rho|\alpha\rangle \exp\{-2(\alpha z^* - \alpha^* z)\} d^2\alpha, \tag{2.7}$$

where the integration extends over the whole complex α-plane. In eq. (2.7) $|\alpha\rangle$ is a coherent state (for properties of coherent states, see GLAUBER [1963], KLAUDER and SUDARSHAN [1968]); it is an eigenstate of the annihilation operator a with the eigenvalue α. Coherent states $|\alpha\rangle$ are known to form a complete set (KLAUDER [1960]) but are not orthogonal, more precisely

$$\frac{1}{\pi} \int |\alpha\rangle\langle\alpha| d^2\alpha = 1, \qquad \langle\alpha'|\alpha\rangle = \exp\{\alpha'^*\alpha - \tfrac{1}{2}|\alpha|^2 - \tfrac{1}{2}|\alpha'|^2\}. \tag{2.8}$$

We will refer to the distribution function $\Phi^{(A)}$, obtained from ρ via the antinormal rule of mapping, as the Sudarshan-Glauber distribution function (SUDARSHAN [1963], GLAUBER [1963]). It is often referred to as the P-representation of the density operator. In terms of $\Phi^{(A)}$ the density operator admits the following representation

$$\rho = \int \Phi^{(A)}(z, z^*)|z\rangle\langle z| d^2z. \tag{2.9}$$

The relation (2.9) has been inverted by MEHTA [1967] to obtain the following formula for $\Phi^{(A)}$

$$\Phi^{(A)}(z, z^*) = \frac{1}{\pi^2} \exp(|z|^2) \int \langle -\alpha|\rho|\alpha\rangle \exp\{|\alpha|^2 - (\alpha z^* - \alpha^* z)\} d^2\alpha. \quad (2.10)$$

The distribution function $\Phi^{(N)}$, obtained from ρ via the normal rule of mapping is given by (MEHTA and SUDARSHAN [1965], GLAUBER [1965])

$$\Phi^{(N)}(z, z^*) = \frac{1}{\pi} \langle z|\rho|z\rangle. \quad (2.11)$$

This function has the interesting properties that it exists for all class of density operators and is non-negative. These two properties make it very attractive to use in physical problems. This distribution function is thus closest to the classical distribution functions.

In calculations, we also need the c-number functions associated with any general operator G. Let $F_G^{(\Omega)}$ be the "Ω-equivalent" of G, obtained from G via the Θ rule of mapping i.e.

$$G \xrightarrow{\Theta} F_G^{(\Omega)} \xrightarrow{\Omega} G. \quad (2.12)$$

The functions $F_G^{(W)}$, $F_G^{(A)}$ and $F_G^{(N)}$ are given in terms of the operator G by the relations (2.7), (2.10) and (2.11) with ρ/π replaced by G.

The phase space distribution functions are very useful in computing the expectation values. It is easily shown that the expectation value of an operator may be expressed in the forms (AGARWAL and WOLF [1970b]; see also LAX and LOUISELL [1967])

$$\langle G\rangle = \mathrm{Tr}(\rho G) = \int \Phi^{(W)}(z, z^*) F_G^{(W)}(z, z^*) d^2z \quad (2.13)$$

$$= \int \Phi^{(A)}(z, z^*) F_G^{(N)}(z, z^*) d^2z \quad (2.14)$$

$$= \int \Phi^{(N)}(z, z^*) F_G^{(A)}(z, z^*) d^2z. \quad (2.15)$$

We have thus expressed the quantum mechanical expectation values in the same form as the averages occurring in classical statistical mechanics, i.e. averages are given as integrals over the phase space. In quantum optics, one often wishes to compute the expectation values of operators that are already in some ordered form for example the normally ordered correlation functions of the form, $\langle a^{+m} a^n\rangle$. In this case we have from (2.14)

$$\langle a^{+m} a^n\rangle = \int \Phi^{(A)}(z, z^*) z^{*m} z^n d^2z. \quad (2.16)$$

This formula brings into evidence even more clearly the formal analogy between the phase space representations and classical statistical mechanics. It should be noted that if we use (2.13) or (2.15), then we obtain instead of (2.16) (for $n = m$)

$$\langle a^{+m}a^m\rangle = (-1)^m \int \Phi^{(N)}(z, z^*)L_m(|z|^2)\mathrm{d}^2z, \tag{2.17}$$

$$= (-\tfrac{1}{2})^m \int \Phi^{(W)}(z, z^*)L_m(2|z|^2)\mathrm{d}^2z, \tag{2.18}$$

where L_m is the Laguerre polynomial of order m. We emphasize that for practical calculations, one may use any of the phase space distribution functions and in some cases we may be naturally led to use one particular distribution function.

We now consider the dynamic aspects of phase space methods (AGARWAL and WOLF [1968, 1970b, c]). The phase space distribution functions $\Phi^{(\Omega)}$ are now time dependent and an equation of motion for $\Phi^{(\Omega)}$ is obtained from eq. (1.10) by applying the mapping operator Θ to both sides of (1.10), i.e.

$$\frac{\partial}{\partial t}\Phi^{(\Omega)} = \Theta\left(\frac{\partial \rho}{\partial t}\right) = -\mathrm{i}\Theta([H, \rho]). \tag{2.19}$$

The right hand side of (2.19) may be expressed in terms of $\Phi^{(\Omega)}$ and $F_H^{(\Omega)}$ and their derivatives as follows

$$\frac{\partial}{\partial t}\Phi^{(\Omega)} = -\mathrm{i}F^{(\Omega)}\{\mathscr{D}^{(1)} - \mathscr{D}^{(2)}\}\Phi^{(\Omega)}, \tag{2.20}$$

where the operators $\mathscr{D}^{(1)}$ and $\mathscr{D}^{(2)}$ are given by (cf. AGARWAL and WOLF [1968, 1970b])

$$\mathscr{D}^{(1)} = \exp\left\{(\lambda+\tfrac{1}{2})\frac{\overleftarrow{\partial}}{\partial z}\frac{\overrightarrow{\partial}}{\partial z^*} + (\lambda-\tfrac{1}{2})\frac{\overleftarrow{\partial}}{\partial z^*}\frac{\overrightarrow{\partial}}{\partial z}\right\}, \tag{2.21}$$

$$\mathscr{D}^{(2)} = \exp\left\{(\lambda+\tfrac{1}{2})\frac{\overleftarrow{\partial}}{\partial z^*}\frac{\overrightarrow{\partial}}{\partial z} + (\lambda-\tfrac{1}{2})\frac{\overleftarrow{\partial}}{\partial z}\frac{\overrightarrow{\partial}}{\partial z^*}\right\}. \tag{2.22}$$

In eqs. (2.21) and (2.22) the arrow pointing to the left (right) indicates that the differential operator below it operates on quantities on the left (right) of the operators $\mathscr{D}^{(1)}$ and $\mathscr{D}^{(2)}$. In eq. (2.20) Ω refers to any of the three rules of mapping and λ takes values 0, $\frac{1}{2}$, and $-\frac{1}{2}$ corresponding to the Weyl, the normal and the antinormal rules of mapping. Eq. (2.20) can also be written in the form

$$\frac{\partial \Phi^{(\Omega)}}{\partial t} = -\mathrm{i}\mathscr{L}^{(\Omega)}\Phi^{(\Omega)}, \tag{2.23}$$

where the Liouville operator $\mathscr{L}^{(\Omega)}$ is now given by

$$\mathscr{L}^{(\Omega)} = F_H^{(\Omega)}\{\mathscr{D}^{(1)} - \mathscr{D}^{(2)}\}. \tag{2.24}$$

Let $K^{(\Omega)}(z, z^*, t|z_0, z_0^*, t_0)$ be the Green's function associated with eq. (2.23). The Green's function $K^{(\Omega)}$ is the solution of (2.23) with the initial condition

$$K^{(\Omega)}(z, z^*, t_0|z_0, z_0^*, t_0) = \delta^{(2)}(z - z_0). \tag{2.25}$$

The Green's function is useful in the calculation of the multitime correlation functions (LAX [1968b], GRAHAM et al. [1968], AGARWAL and WOLF [1970c]). In particular the normally-ordered time-ordered correlation functions defined by

$$\Gamma_{\mathrm{T}}^{(\mathrm{N})} \equiv \langle [a^+(t_1)]^{i_1} \ldots [a^+(t_n)]^{i_n} [a(t_n)]^{j_n} \ldots [a(t_1)]^{j_1} \rangle, \quad (t_n \geqslant t_{n-1} \geqslant \ldots \geqslant t_1), \tag{2.26}$$

where $a(t)$ and $a^+(t)$ are operators in the Heisenberg picture, may be shown to be expressible in the form

$$\Gamma_{\mathrm{T}}^{(\mathrm{N})} = \int \ldots \int \Phi^{(\mathrm{A})}(z_0, z_0^*, t_0) \times \prod_{\lambda=1}^{n} \{(z_\lambda^*)^{i_\lambda}(z_\lambda)^{j_\lambda} K^{(\mathrm{A})}(z_\lambda, z_\lambda^*, t_\lambda | z_{\lambda-1}, z_{\lambda-1}^*, t_{\lambda-1}) \mathrm{d}^2 z_\lambda\} \mathrm{d}^2 z_0 . \tag{2.27}$$

Next we give some identities which are useful in obtaining the operator form of the phase space equations of motion. Some of these identities, which are easily proved by using the general theory developed by AGARWAL and WOLF [1970b], are given by

$$\frac{\partial}{\partial z}(z\Phi^{(\mathrm{A})}) \xrightarrow{\Omega^{(\mathrm{A})}} -\frac{1}{\pi}[a^+, a\rho], \qquad \frac{\partial}{\partial z}(z^*\Phi^{(\mathrm{A})}) \xrightarrow{\Omega^{(\mathrm{A})}} -\frac{1}{\pi}[a^+, \rho a^+], \tag{2.28}$$

$$\frac{\partial}{\partial z}(z\Phi^{(\mathrm{N})}) \xrightarrow{\Omega^{(\mathrm{N})}} -\frac{1}{\pi}[a^+, \rho a], \qquad \frac{\partial}{\partial z}(z^*\Phi^{(\mathrm{N})}) \xrightarrow{\Omega^{(\mathrm{N})}} -\frac{1}{\pi}[a^+, a^+\rho]. \tag{2.29}$$

The results, which we have so far presented, are easily generalized to boson systems with many degrees of freedom (AGARWAL and WOLF [1970b]). One may develop a phase space theory for a system of fermions along similar lines. AGARWAL [1969] for example, has made use of the Schwinger's boson representation to deal with two-level atoms. The phase space theory for a system of fermions will not be discussed here.

Finally we introduce the action and angle variables J and φ defined by

$$J = |z|^2, \qquad \varphi = \frac{1}{2\mathrm{i}} \ln (z/z^*). \tag{2.30}$$

The transformation laws are given by

$$\frac{\partial}{\partial z} = J^{\frac{1}{2}} \mathrm{e}^{-\mathrm{i}\varphi} \left(\frac{\partial}{\partial J} - \frac{\mathrm{i}}{2J} \frac{\partial}{\partial \varphi} \right), \qquad \frac{\partial}{\partial z^*} = J^{\frac{1}{2}} \mathrm{e}^{\mathrm{i}\varphi} \left(\frac{\partial}{\partial J} + \frac{\mathrm{i}}{2J} \frac{\partial}{\partial \varphi} \right), \tag{2.31}$$

and

$$\int \mathrm{d}^2 z \ldots = \tfrac{1}{2} \int_0^\infty \mathrm{d}J \int_0^{2\pi} \mathrm{d}\varphi \ldots . \tag{2.32}$$

In the case when the Hamiltonian H is equal to $\omega a^+ a$ the Liouville operator $\mathscr{L}^{(\Omega)}$ is found, from (2.24) and (2.31), to be given by

$$\mathscr{L}^{(\Omega)} = \mathrm{i}\omega \frac{\partial}{\partial \varphi} . \tag{2.33}$$

For a system with many degrees of freedom and with H equal to $\sum_k \omega_k a_k^+ a_k$, the Liouville operator is

$$\mathscr{L}^{(\Omega)} = \mathrm{i} \sum_k \omega_k \frac{\partial}{\partial \varphi_k} . \tag{2.34}$$

The operator $\mathscr{L}^{(\Omega)}$ has the following eigenfunctions and eigenvalues

$$\psi(\{\nu_k\}) = (2\pi)^{-N/2} \exp\left(-\mathrm{i} \sum_k \nu_k \varphi_k\right) \qquad E(\{\nu_k\}) = \sum_k \nu_k \omega_k , \tag{2.35}$$

where ν_k vary over all positive and negative integers. The eigenfunctions $\psi(\{\nu_k\})$ are orthogonal and play a very important role in the perturbation expansions of the distribution functions in interaction problem (cf. § 12; for the use of eigenfunctions $\psi(\{\nu_k\})$ in problems in classical statistical mechanics, see PRIGOGINE [1962]).

§ 3. Master Equation for a General System

We have seen that the basic equation of motion for both classical and quantum systems can be written in the form

$$\partial f/\partial t = -\mathrm{i}\mathscr{L} f, \tag{3.1}$$

where $\mathscr{L}$ is the appropriate Liouville operator and f represents the state of the system. Table 1 gives the form of the Liouville operator $\mathscr{L}$ and f for both classical and quantum systems. The master equation is an equation

TABLE 1

The form of the Liouville operator $\mathscr{L}$, the distribution f and the commonly employed projection operators $\mathscr{P}$ for both classical and quantum systems

	The distribution f	The Liouville operator $\mathscr{L}$	Commonly employed projection operators $\mathscr{P}$
Classical systems:	$\Phi_N(\{q_i\}, \{p_i\}, t)$	$\mathrm{i}\sum_i \left\{\frac{\partial H}{\partial q_i}\frac{\partial}{\partial p_i} - \frac{\partial H}{\partial p_i}\frac{\partial}{\partial q_i}\right\}$	$\frac{1}{V^N}\int \ldots \int \mathrm{d}(\{q_i\}) \ldots$
	$\Phi_N(\{J_i\}, \{\varphi_i\}, t)$	$\mathrm{i}\sum_i \left\{\frac{\partial H}{\partial \varphi_i}\frac{\partial}{\partial J_i} - \frac{\partial H}{\partial J_i}\frac{\partial}{\partial \varphi_i}\right\}$	$\frac{1}{(2\pi)^N}\int \ldots \int \mathrm{d}(\{\varphi_i\}) \ldots$
Quantum systems:			$\sum_m \mathscr{P}_m \,\mathrm{Tr}\,\{\mathscr{P}_m \ldots\}$
Operator treatment	ρ	$[H, \quad]$	$\rho_R(0)\,\mathrm{Tr}_R \ldots$ (cf. eq. (5.7))
			$\overline{\ldots}$ (cf. eq. (1.23))
Quantum systems:		$F_H^{(\Omega)}\{\mathscr{D}^{(1)} - \mathscr{D}^{(2)}\}$	$\Phi_R^{(\Omega)}(0)\int \ldots \int \mathrm{d}^2(z_R) \ldots$
Phase space treatment	$\Phi^{(\Omega)}(\{z_i\}, \{z_i^*\}, t)$	$\mathscr{D}^{(1)} = \prod_i \exp\left\{(\lambda+\tfrac{1}{2})\frac{\overleftarrow{\partial}}{\partial z_i}\frac{\overrightarrow{\partial}}{\partial z_i^*} + (\lambda-\tfrac{1}{2})\frac{\overleftarrow{\partial}}{\partial z_i^*}\frac{\overrightarrow{\partial}}{\partial z_i}\right\}$	(cf. eq. (5.8)) $\overline{\ldots}$
		$\mathscr{D}^{(2)} = \prod_i \exp\left\{(\lambda+\tfrac{1}{2})\frac{\overleftarrow{\partial}}{\partial z_i^*}\frac{\overrightarrow{\partial}}{\partial z_i} + (\lambda-\tfrac{1}{2})\frac{\overleftarrow{\partial}}{\partial z_i}\frac{\overrightarrow{\partial}}{\partial z_i^*}\right\}$	$\left(\frac{1}{2\pi}\right)^N \int \ldots \int \mathrm{d}(\{\varphi_i\}) \ldots$

of motion for $\mathscr{P}f$, where $\mathscr{P}$ is the projection operator which projects out the relevant part of f. We will refer to f as the distribution function, with the understanding that in the operator treatment of quantum systems it represents the density operator. In Table 1, we have also listed the form of commonly used projection operators. To obtain an equation for $\mathscr{P}f$, we write f in the form

$$f = \mathscr{P}f + (1-\mathscr{P})f. \tag{3.2}$$

On taking the Laplace transform of (3.1), we obtain

$$\beta C(\beta) - f(0) = -\mathrm{i}\mathscr{L}C(\beta), \tag{3.3}$$

where $C(\beta)$ is the Laplace transform of $f(t)$, i.e.

$$C(\beta) = \int_0^\infty \mathrm{e}^{-t\beta} f(t)\mathrm{d}t, \qquad (\mathrm{Re}\,\beta \geqslant 0), \tag{3.4}$$

and where we have assumed that $\mathscr{L}$ is *explicitly time independent*. On multiplying (3.3) by $\mathscr{P}$ and $(1-\mathscr{P})$ respectively, we obtain the equations

$$\begin{aligned}\beta\mathscr{P}C(\beta) - \mathscr{P}f(0) \\ = -\mathrm{i}\mathscr{P}\mathscr{L}\mathscr{P}C(\beta) - \mathrm{i}\mathscr{P}\mathscr{L}(1-\mathscr{P})C(\beta),\end{aligned} \tag{3.5}$$

$$\begin{aligned}\beta(1-\mathscr{P})C(\beta) - (1-\mathscr{P})f(0) \\ = -\mathrm{i}(1-\mathscr{P})\mathscr{L}\mathscr{P}C(\beta) - \mathrm{i}(1-\mathscr{P})\mathscr{L}(1-\mathscr{P})C(\beta).\end{aligned} \tag{3.6}$$

The solution of (3.6) is

$$\begin{aligned}(1-\mathscr{P})C(\beta) \\ = \{\beta + \mathrm{i}(1-\mathscr{P})\mathscr{L}\}^{-1}\{-\mathrm{i}(1-\mathscr{P})\mathscr{L}\mathscr{P}C(\beta) + (1-\mathscr{P})f(0)\}.\end{aligned} \tag{3.7}$$

We substitute (3.7) in (3.5) and take the inverse Laplace transform. Then we obtain the following equation for $\mathscr{P}f$ (ZWANZIG [1961a], AGARWAL [1969])

$$\begin{aligned}\frac{\partial}{\partial t}[\mathscr{P}f(t)] + \mathrm{i}\mathscr{P}\mathscr{L}[\mathscr{P}f(t)] + \mathrm{i}\mathscr{P}\mathscr{L}\exp\{-\mathrm{i}(1-\mathscr{P})\mathscr{L}t\}(1-\mathscr{P})f(0) \\ + \int_0^t \mathscr{P}\mathscr{L}\exp\{-\mathrm{i}(1-\mathscr{P})\mathscr{L}\tau\}(1-\mathscr{P})\mathscr{L}[\mathscr{P}f(t-\tau)]\mathrm{d}\tau = 0.\end{aligned} \tag{3.8}$$

Equation (3.8) is *the master equation* for the relevant part of the distribution function. It should be noted that eq. (3.8) is an integro-differential equation and is an exact consequence of the dynamics of the system. Further simplification can be made depending on the form of the Hamiltonian and the

initial condition. In most of the physical applications the initial condition will be such that

$$(1-\mathscr{P})f(0) = 0. \tag{3.9}$$

This condition is usually referred to as *the assumption of initial random phase.*

We now obtain the explicit form of (3.8) in case when the projection operator $\mathscr{P}$ is given by eq. (1.13), i.e. we obtain the equation of motion for the diagonal elements of ρ. The condition (3.9) in this case implies that the density operator is initially diagonal. Let the Hamiltonian of the system be written as

$$H = H_0 + gH_1, \tag{3.10}$$

where gH_1 is a small perturbation. We work in a representation in which H_0 is diagonal and H_1 has no diagonal elements. Then the Liouville operator is given by

$$\mathscr{L} = \mathscr{L}_0 + g\mathscr{L}_1, \qquad \mathscr{L}_0 = [H_0, \quad], \qquad \mathscr{L}_1 = [H_1, \quad]. \tag{3.11}$$

It is easily verified from (1.13) and (3.11) that

$$\mathscr{P}\mathscr{L}_0 = \mathscr{L}_0\mathscr{P} = 0. \tag{3.12}$$

Moreover since $\mathscr{P}\rho$ has no off-diagonal elements it follows that

$$\mathscr{P}\mathscr{L}_1[\mathscr{P}\rho(t)] = 0. \tag{3.13}$$

On combining eqs. (3.9)–(3.13), eq. (3.8) reduces to

$$\frac{\partial}{\partial t}[\mathscr{P}\rho(t)] + g^2\int_0^t K(\tau)[\mathscr{P}\rho(t-\tau)]\mathrm{d}\tau = 0, \tag{3.14}$$

where the kernel $K(\tau)$ is given by

$$K(\tau) = \mathscr{P}\mathscr{L}_1 \exp\{-\mathrm{i}(1-\mathscr{P})\mathscr{L}(1-\mathscr{P})\tau\}(1-\mathscr{P})\mathscr{L}_1. \tag{3.15}$$

Equation (3.14) leads to the following equation for ρ_{nn}

$$\frac{\partial \rho_{mm}}{\partial t} + g^2\sum_n \int_0^t k_{mn}(\tau)\rho_{nn}(t-\tau)\mathrm{d}\tau = 0, \tag{3.16}$$

where

$$k_{mn}(\tau) = \langle m|[H_1, \exp\{-\mathrm{i}H_{\mathscr{P}}\tau\}(1-\mathscr{P})[H_1, \mathscr{P}_n]\exp\{\mathrm{i}H_{\mathscr{P}}\tau\}]|m\rangle, \tag{3.17}$$

and where $H_{\mathscr{P}}$ is given by

$$H_{\mathscr{P}} = H_0 + g(1-\mathscr{P})H_1(1-\mathscr{P}). \tag{3.18}$$

From the completeness of states $|n\rangle$, it follows that

$$\sum_n k_{mn}(\tau) = 0. \tag{3.19}$$

On using (3.19), we can write (3.16) in the form

$$\frac{\partial \rho_{mm}}{\partial t} = -g^2 \sum_{n \neq m} \int_0^t \{k_{mn}(\tau)\rho_{nn}(t-\tau) - k_{mn}(\tau)\rho_{mm}(t-\tau)\} \mathrm{d}\tau. \tag{3.20}$$

It should be noted that so far we have made no assumption about the strength of the perturbation, i.e. (3.20) holds to all orders in perturbation. We only assumed that ρ is initially diagonal. The master equation (3.20) is in the form of the rate equation since the first term on the right hand side represents the transitions from all other states $|n\rangle$ to the state $|m\rangle$ and the second sum represents the transitions from the state $|m\rangle$ to all other states. ZWANZIG [1964] used the expression (3.15) for the kernel $K(\tau)$ to establish the identity of the master equation (3.20) with the master equations of MONTROLL [1962] and of PRIGOGINE and RESIBOIS [1961].

The exact expression for the kernel $k_{mn}(\tau)$ is rather complicated. In the lowest order in perturbation (the Born approximation) we can replace $H_{\mathscr{I}}$ by H_0, and we then obtain the following expression for $k_{mn}(\tau)$

$$k_{mn}(\tau) = -2|\langle m|H_1|n\rangle|^2 \cos[(E_n - E_m)\tau]. \tag{3.21}$$

To obtain higher order terms, we take the Laplace transform of $K(\tau)$ and expand it in powers of $\mathscr{L}_1$. We then obtain the following series expansion (ZWANZIG [1961a])

$$\begin{aligned} K(\beta) = {} & \mathscr{P}\mathscr{L}_1 G_0(1-\mathscr{P})\mathscr{L}_1 - \mathrm{i}g\mathscr{P}\mathscr{L}_1 G_0(1-\mathscr{P})\mathscr{L}_1 G_0(1-\mathscr{P})\mathscr{L}_1 \\ & + (\mathrm{i}g)^2 \mathscr{P}\mathscr{L}_1 G_0(1-\mathscr{P})\mathscr{L}_1 G_0(1-\mathscr{P})\mathscr{L}_1 G_0(1-\mathscr{P})\mathscr{L}_1 + \ldots, \end{aligned} \tag{3.22}$$

where

$$G_0 = (\beta + \mathrm{i}\mathscr{L}_0)^{-1}. \tag{3.23}$$

The Pauli equation may be obtained from (3.20) and (3.21) in the following limits (VAN HOVE [1955, 1957], MONTROLL [1961], see also PRIGOGINE [1962]):

(i) N = number of degrees of freedom of the system $\to \infty$, and V = volume of the system $\to \infty$, such that N/V = constant,

(ii) $g \to 0$, $t \to \infty$ such that $(g^2 t)$ = constant. In these limits it is possible to replace $\rho_{nn}(t-\tau)$ in eq. (3.20) by $\rho_{nn}(t)$ and extend the upper limit of integration to ∞ (for details see MONTROLL [1961]) and we then find that

$$\frac{\partial \rho_{mm}}{\partial t} = \sum_{n \neq m} \int_0^\infty 2g^2 |\langle m|H_1|n\rangle|^2 \cos\left[(E_n - E_m)\tau\right] \mathrm{d}\tau \{\rho_{nn}(t) - \rho_{mm}(t)\}. \quad (3.24)$$

On simplification, eq. (3.24) leads to the Pauli equation (1.1).

The Pauli equation is easily generalized to open systems, i.e. a small system coupled to a large system usually called a reservoir. (Open systems will be discussed in detail in § 5.) For such systems n and m in Pauli equation refer to the states of small system and the transition probabilities are no longer symmetric, i.e. $\gamma_{nm} \neq \gamma_{mn}$. In fact if the relaxation towards thermal equilibrium at temperature T is assumed, then one finds that

$$\gamma_{nm}/\gamma_{mn} = \exp\{\beta(E_m - E_n)\}, \quad (3.25)$$

where $\beta = 1/K_B T$ and K_B is the Boltzmann constant. Eq. (3.25) is merely a statement of the principle of detailed balance*. General properties of the Pauli equation are discussed in great detail in a recent review by OPPENHEIM et al. [1967]. The Pauli's master equation can be written in the form (MONTROLL [1961])

$$\mathrm{d}\tilde{p}_n/\mathrm{d}t = \sum_m b_{nm} \tilde{p}_m, \quad (3.26)$$

where

$$\tilde{p}_n = \rho_{nn} \exp(\tfrac{1}{2}\beta E_n), \quad (3.27)$$

$$b_{nm} = \begin{cases} \exp\{\frac{1}{2}\beta(E_n - E_m)\}\gamma_{nm} & n \neq m \\ -\sum_{l \neq n} \gamma_{lm} & n = m. \end{cases} \quad (3.28)$$

It is seen from the definition of the matrix b and the fact that γ_{nm} are the transition probabilities, that the matrix b is a semi-negative definite matrix which is also symmetric (SHULER [1958]). Therefore the eigenvalues λ_k of b are real and $\lambda_k \leqq 0$. The solution of the master equation is then given as

$$p_n(t) = \sum_{m,l} \exp\{\tfrac{1}{2}\beta(E_m - E_n)\} U_{nl} U^+_{lm} p_m(0) \exp\{\lambda_l t\}, \quad (3.29)$$

where U is the orthogonal matrix which diagonalizes the matrix b. Finally it should be noted that the present analysis is not restricted to the relaxation towards thermal equilibrium and the general case is obtained by the replacement $\exp\{-\beta E_n\} \to p_n(\infty)$, where $p_n(\infty)$ is the steady state solution.

§ 4. Master Equations for Systems Interacting with Stochastic Perturbations

We consider a system which is interacting with stochastic perturbations. The stochastic perturbation may, for example, be the coupling of a spin

* For an explicit proof of the fact that (3.25) follows from microreversibility see AGARWAL [1973b].

with a fluctuating magnetic field or the dipole-dipole coupling between two spins (SLITCHER [1963], ABRAGAM [1961]). A simpler example of stochastic perturbation is a frequency modulated harmonic oscillator. We will now use the projection operator technique to obtain the master equation for such systems.

Let the Hamiltonian of the system be given by

$$H = H_0 + gH_1, \tag{4.1}$$

where H_1 represents the random perturbation. We assume that H_0 is time independent and is not stochastic in nature. Let the perturbation be of the form

$$H_1 = \sum_\alpha F_\alpha(t) G_\alpha, \tag{4.2}$$

where $F_\alpha(t)$ are the random forces and G_α are the system operators. If G_α's are not hermitian operators, then (4.2) can be made hermitian by the convention

$$F_{-\alpha}(t) = F_\alpha^*(t), \qquad G_{-\alpha} = G_\alpha^+ . \tag{4.3}$$

We further assume that the random forces $F_\alpha(t)$ are the stationary random functions with zero ensemble mean value. Let $s_{\alpha\beta}$ be the cross spectral density defined as

$$s_{\alpha\beta}(\omega) = \int_0^\infty e^{i\omega\tau} \overline{F_\alpha(t) F_\beta(t-\tau)} d\tau. \tag{4.4}$$

Since the Hamiltonian is now explicitly time dependent, it is convenient to work in the interaction picture. Various operators and c-numbers in the interaction picture will be distinguished from those in the Schrödinger picture by a subscript I. In the interaction picture we have

$$\partial \rho_I / \partial t = -ig[H_I(t), \rho_I], \tag{4.5}$$

where

$$\begin{aligned} H_I(t) &= \exp\{iH_0 t\} H_1 \exp\{-iH_0 t\}, \\ \rho_I &= \exp\{iH_0 t\} \rho \exp\{-iH_0 t\}. \end{aligned} \tag{4.6}$$

Hence the basic equation of motion in the interaction picture is of the form

$$\partial f_I / \partial t = -ig \mathscr{L}_I(t) f_I. \tag{4.7}$$

To obtain the master equation, we multiply both sides of (4.7) by $\mathscr{P}$ and $(1-\mathscr{P})$ respectively

$$\frac{\partial}{\partial t}[\mathscr{P} f_I] = -ig \mathscr{P} \mathscr{L}_I(t) \mathscr{P} f_I - ig \mathscr{P} \mathscr{L}_I(t)(1-\mathscr{P}) f_I, \tag{4.8}$$

$$\frac{\partial}{\partial t}[(1-\mathscr{P}) f_I] = -ig(1-\mathscr{P}) \mathscr{L}_I(t) \mathscr{P} f_I - ig(1-\mathscr{P}) \mathscr{L}_I(t)(1-\mathscr{P}) f_I. \tag{4.9}$$

The solution of (4.9) is given by

$$(1-\mathscr{P})f_{\rm I}(t) = \mathscr{U}(t,0)[(1-\mathscr{P})f_{\rm I}(0)] - ig\int_0^t \mathscr{U}(t,\tau)(1-\mathscr{P})\mathscr{L}_{\rm I}(\tau)[\mathscr{P}f_{\rm I}(\tau)]d\tau, \tag{4.10}$$

where

$$\mathscr{U}(t,\tau) = T\exp\left\{-ig\int_\tau^t (1-\mathscr{P})\mathscr{L}_{\rm I}(t')dt'\right\}, \tag{4.11}$$

and where T is Dyson's time ordering operator. On substituting from (4.10) into (4.8), we obtain the following master equation (AGARWAL [1969])

$$\frac{\partial}{\partial t}[\mathscr{P}f_{\rm I}(t)] + ig\mathscr{P}\mathscr{L}_{\rm I}(t)[\mathscr{P}f_{\rm I}(t)] + ig\mathscr{P}\mathscr{L}_{\rm I}(t)\mathscr{U}(t,0)[(1-\mathscr{P})f_{\rm I}(0)] + g^2\int_0^t \mathscr{P}\mathscr{L}_{\rm I}(t)\mathscr{U}(t,\tau)(1-\mathscr{P})\mathscr{L}_{\rm I}(\tau)[\mathscr{P}f_{\rm I}(\tau)]d\tau = 0. \tag{4.12}$$

For the problem under consideration the projection operator is given by eq. (1.23), viz.

$$\mathscr{P}G = \bar{G}, \tag{4.13}$$

where the bar denotes the ensemble average with respect to the distribution of the random forces $F_\alpha(t)$. On making use of the condition of initial random phase (eq. (3.9)) and of the property of the random forces $F_\alpha(t)$, assumed above, eq. (4.12) reduces to

$$\frac{\partial\overline{f_{\rm I}(t)}}{\partial t} + g^2\sum_{\alpha,\beta}\int_0^t \mathscr{L}_\alpha(t)\overline{F_\alpha(t)\mathscr{U}(t,t-\tau)F_\beta(t-\tau)}\mathscr{L}_\beta(t-\tau)\bar{f}_{\rm I}(t-\tau)d\tau = 0, \tag{4.14}$$

where $\mathscr{L}_\alpha(t)$ is the Liouville operator corresponding to the operator $G_\alpha(t)$. We have not made any assumption about the strength of the interaction and eq. (4.14) is therefore a master equation to all orders in the coupling coefficient g for systems interacting with stochastic perturbations.

We now make the commonly used approximations (ABRAGAM [1961], BLOCH [1956], REDFIELD [1957, 1965], SLITCHER [1963]): (i) Born approximation in which we replace $\mathscr{U}(t,\tau)$ by unity, (ii) short memory approximation in which we replace $\overline{f_{\rm I}(t-\tau)}$ under the integral sign in (4.14) by $\overline{f_{\rm I}(t)}$ and (iii) take the long time limit. The master equation (4.14) then reduces to

$$\frac{\partial\overline{f_{\rm I}(t)}}{\partial t} + g^2\sum_{\alpha,\beta}\int_0^\infty \overline{F_\alpha(t)F_\beta(t-\tau)}\mathscr{L}_\alpha(t)\mathscr{L}_\beta(t-\tau)\bar{f}_{\rm I}(t)d\tau = 0. \tag{4.15}$$

On taking the matrix elements with respect to the unperturbed states, this equation reduces to REDFIELD's [1957, 1965] master equation.

We now discuss two applications of the master equation (4.14). We first consider an elementary model of a laser, known as the *phase diffusion model* (GLAUBER [1965], LAMB [1965], KLAUDER and SUDARSHAN [1968, p. 228]; a more realistic model of the laser will be discussed in § 10). Let us consider a randomly modulated* harmonic oscillator for which the Hamiltonian is given by

$$H = \omega a^+ a + gF(t)a^+ a, \tag{4.16}$$

where $F(t)$ is the random force. We assume that $F(t)$ is a real delta correlated Gaussian random process (cf. STRATONOVICH [1963]) with zero mean, i.e. one for which

$$\overline{F(t)} = 0, \qquad \overline{F(t_1)F(t_2)} = \frac{2D}{g^2}\delta(t_1 - t_2). \tag{4.17}$$

For the Hamiltonian (4.16), it is easily seen from eq. (2.20) that the Sudarshan-Glauber distribution function in the interaction picture satisfies the equation

$$\frac{\partial \Phi_I}{\partial t} = -i\mathscr{L}_I(t)\Phi_I, \tag{4.18}$$

where

$$\mathscr{L}_I(t) = gF(t)\left\{z^* \frac{\partial}{\partial z^*} - z\frac{\partial}{\partial z}\right\}, \tag{4.19}$$

and where we have dropped the superscript A from $\Phi^{(A)}$. On substituting (4.19) in (4.14) and on using the properties of the random force $F(t)$, we find that $\overline{\Phi_I}$ satisfies the following (exact) master equation

$$\frac{\partial}{\partial t}\overline{\Phi_I} + D\left(z\frac{\partial}{\partial z} - z^*\frac{\partial}{\partial z^*}\right)^2 \overline{\Phi_I} = 0. \tag{4.20}$$

On introducing the polar coordinates r and θ ($z = re^{i\theta}$), eq. (4.20) becomes

$$\frac{\partial \overline{\Phi_I}}{\partial t} = D\frac{\partial^2 \overline{\Phi_I}}{\partial \theta^2}. \tag{4.21}$$

It should be noted that (4.21) is just the diffusion equation for diffusion on a circular ring. It is easily seen that the Green's function associated with (4.21) is given by the expression

$$\overline{K_I}(r, \theta, t | r_0, \theta_0, 0) = \frac{1}{2\pi r}\delta(r - r_0)\sum_{n=-\infty}^{+\infty} \exp\{in(\theta - \theta_0) - Dn^2 t\}. \tag{4.22}$$

Various normally-ordered time-ordered correlation functions can be calculated by using (4.22) and (2.27). In particular one finds that

* Generalizations of this kind of stochastic processes have been treated by FOX [1972].

$$\langle a^+(t)a(0)\rangle = \langle a^+(0)a(0)\rangle e^{i\omega t - Dt}, \tag{4.23}$$

which obviously corresponds to a Lorenztian spectrum with half width equal to D. It should also be noted that there are no amplitude fluctuations in this model and it is often used to describe the behavior of a laser beam far above threshold.

Next we use the master equation (4.15) to obtain Bloch's phenomenological equations (see, for example, ABRAGAM [1961]). In problems of relaxation Redfield's equation leads to relaxation at infinite temperature. This is so because the heat bath has been treated classically. The correct relaxation behavior is obtained by replacing $\overline{f_I(t)}$ in (4.15) by $\overline{f_I(t)} - f_{th}$, where f_{th} is the equilibrium distribution function at temperature T and is given by

$$\rho_{th} = \exp(-H_0/K_B T)/\operatorname{Tr}\{\exp(-H_0/K_B T)\}. \tag{4.24}$$

The Hamiltonian for the case of spin relaxation can be written as

$$H_0 = -\omega S^z, \qquad gH_1 = -\gamma(S^+F^+(t) + S^-F^-(t) + S^zF^z(t)). \tag{4.25}$$

We assume that the random forces are such that

$$s_{++} = s_{--} = s_{+z} = s_{-z} = 0, \qquad s_{+-} \neq 0, \qquad s_{zz} \neq 0, \tag{4.26}$$

where $s_{\alpha\beta}$ are defined by eq. (4.4). On substituting (4.25) and (4.26) in eq. (4.15), we find the following master equation

$$\begin{aligned}\frac{\partial \bar{\rho}_I}{\partial t} = &-\gamma^2 s_{+-}(\omega)[S^+, [S^-, \bar{\rho}_I - \rho_{th}]] \\ &-\gamma^2 s^*_{+-}(\omega)[S^-, [S^+, \bar{\rho}_I - \rho_{th}]] - \gamma^2 s_{zz}(0)[S^z, [S^z, \bar{\rho}_I]].\end{aligned} \tag{4.27}$$

Equation (4.27) leads to the following equations of motion for the macroscopic mean values

$$\frac{\partial}{\partial t}\langle S^+\rangle = -i\omega\langle S^+\rangle - \{2s^*_{+-}(\omega) + s_{zz}(0)\}\gamma^2\langle S^+\rangle, \tag{4.28}$$

$$\frac{\partial}{\partial t}\langle S^z\rangle = -2\gamma^2(s^*_{+-}(\omega) + s_{+-}(\omega))(\langle S^z\rangle - \langle S^z\rangle_{th}), \tag{4.29}$$

where

$$\langle S^z\rangle_{th} = \operatorname{Tr}\{S^z\rho_{th}\}. \tag{4.30}$$

In deriving (4.28) and (4.29) we have also assumed that the spin system under consideration is a spin-$\frac{1}{2}$ system. Let T_1 and T_2 be the longitudinal and transverse relaxation coefficients. Then eqs. (4.28), (4.29) are the Bloch's phenomenological equations, with

$$T_1^{-1} = 4\gamma^2 \operatorname{Re}\{s_{+-}(\omega)\}, \qquad T_2^{-1} = \gamma^2\{2 \operatorname{Re} s_{+-}(\omega) + s_{zz}(0)\}. \quad (4.31)$$

It should be noted that the imaginary part of $s_{+-}(\omega)$ causes the frequency shift. Bloch equations are extensively used in optical problems such as in connection with self-induced transparency (McCall and Hahn [1969]).

§ 5. Master Equations for Open Systems

We have already defined an open system as a system which is coupled to another large system, usually called reservoir. Most of the systems which we encounter in physics are open systems. Some of the problems, involving open systems, in quantum optics are those of lasers, relaxation of oscillators and two-level atoms, superradiance, parametric oscillators. In problems like superradiance the radiation field plays the role of the reservoir. In this section we obtain the master equation for the reduced density operator (phase space distribution function) of the sub-system of interest.

We denote the system of interest by S and the reservoir by R. We write the total Hamiltonian of the open system as

$$H = H_S + H_R + H_{RS}, \quad (5.1)$$

where H_S and H_R are the unperturbed Hamiltonians of the system S and of the reservoir R respectively and H_{RS} is the interaction Hamiltonian between the system and the reservoir. We assume, for the sake of simplicity, that the Hamiltonian is explicitly time independent. Let $\rho_{S+R}\{\Phi_{S+R}(\{z_S\}, \{z_S^*\}; \{z_R\}, \{z_R^*\}; t)\}$ be the density operator {phase space distribution function} for the combined system. The reduced density operator corresponding to the system S is given by

$$\rho_S(t) = \operatorname{Tr}_R\{\rho_{S+R}(t)\}, \quad (5.2)$$

where Tr_R indicates the trace over the reservoir variables. The reduced phase space distribution function is given by

$$\Phi_S(\{z_S\}, \{z_S^*\}; t) = \int \Phi_{S+R}(\{z_S\}, \{z_S^*\}; \{z_R\}, \{z_R^*\}; t) d^2(\{z_R\}). \quad (5.3)$$

We make the following two assumptions:

(i) The reservoir is initially in a state of thermal equilibrium at temperature T i.e. $\rho_R(0)$ is given by

$$\rho_R(0) = \exp\{-H_R/K_B T\}/\operatorname{Tr}\exp\{-H_R/K_B T\}, \quad (5.4)$$

or, more generally, $\rho_R(0)$ is a function of H_R only.

(ii) The system S and the reservoir R are initially uncorrelated, i.e.

$$\rho_{S+R}(0) = \rho_S(0)\rho_R(0). \tag{5.5}$$

The first assumption is not necessary but has been made for the sake of simplicity (for a rigorous treatment of open systems see, for example, EMCH and SEWELL [1968]).

It is clear from (5.2) that the reduced density operator can be obtained from the total density operator by suitably projecting it, i.e.

$$\mathscr{P}\rho_{S+R}(t) = \rho_R(0)\rho_S(t), \tag{5.6}$$

where the projection operator $\mathscr{P}$ is given by (ARGYRES and KELLEY [1964]; ARGYRES [1966])

$$\mathscr{P} \ldots = \rho_R(0)\,\mathrm{Tr}_R \ldots . \tag{5.7}$$

The corresponding projection operator for the phase space distribution functions is given by (AGARWAL [1969])

$$\mathscr{P} \ldots = \Phi_R(0) \int \mathrm{d}^2(\{z_R\}) \ldots , \tag{5.8}$$

where $\Phi_R(0)$ is the phase space distribution function corresponding to the density operator $\rho_R(0)$. The Liouville operator $\mathscr{L}$ may be written as

$$\mathscr{L} = \mathscr{L}_S + \mathscr{L}_R + \mathscr{L}_{RS}. \tag{5.9}$$

It is easily verified from (5.5)–(5.9) that the projection operator $\mathscr{P}$ has the properties

$$\mathscr{P}\mathscr{L}_S = \mathscr{L}_S\mathscr{P}, \qquad \mathscr{P}\mathscr{L}_R = \mathscr{L}_R\mathscr{P} = 0, \qquad (1-\mathscr{P})f_{S+R}(0) = 0. \tag{5.10}$$

On using (3.8), (5.9) and (5.10), we find that $\mathscr{P}f$ satisfies the master equation

$$\frac{\partial}{\partial t}[\mathscr{P}f(t)] + \mathrm{i}(\mathscr{L}_S + \mathscr{P}\mathscr{L}_{RS}\mathscr{P})[\mathscr{P}f(t)] + \int_0^t \mathscr{P}\mathscr{L} \exp\left[-\mathrm{i}\tau(1-\mathscr{P})\mathscr{L}\right](1-\mathscr{P})\mathscr{L}[\mathscr{P}f(t-\tau)]\mathrm{d}\tau = 0. \tag{5.11}$$

Eq. (5.11) may be transformed in the standard fashion (see for example, EMCH and SEWELL [1968]) to the following form

$$\frac{\partial}{\partial t}[\mathscr{P}f(t)] = -\mathrm{i}\mathscr{L}_S[\mathscr{P}f(t)] - \mathrm{i}\mathscr{P}\mathscr{L}_{RS}[\mathscr{P}f(t)] - \int_0^t K(t-\tau)[\mathscr{P}f(\tau)]\mathrm{d}\tau, \tag{5.12}$$

where the kernel $K(\tau)$ is given by

$$K(\tau) = \mathscr{P}\mathscr{L}_{RS}\, U_0(\tau)(1-\mathscr{P})U(\tau)\mathscr{L}_{RS}. \tag{5.13}$$

In eq. (5.13) $U_0(t)$ and $U(t)$ are the time development operators defined by

$$U_0(t) = \exp\{-i(\mathscr{L}_R + \mathscr{L}_S)t\}, \tag{5.14}$$

$$U(t) = T \exp\left\{-i\int_0^t (1-\mathscr{P})U_0(-\tau)\mathscr{L}_{RS}\, U_0(\tau)(1-\mathscr{P})d\tau\right\}, \tag{5.15}$$

where T is Dyson's time ordering operator.

Equation (5.12) is the desired master equation for the reduced density operator (or the phase space distribution function). This master equation is exact and is an integro-differential equation. The first term on the right hand side of (5.12) describes the unperturbed motion of the system S. The second term describes the effect of the reservoir R to first order in the interaction H_{RS} and its effect can generally be taken into account by a simple renormalization of the energy levels of the system S. The last term describes the effect of the reservoir to all other orders in perturbation H_{RS}. In the case when (5.12) is the master equation for the phase space distribution function, it is a c-number differential equation and as we will see later, its Markovian form is useful in establishing connection with the Fokker-Planck equations of the stochastic processes.

In applications it is preferable to employ the master equation for the distribution function f in the interaction picture. On transforming (5.12) to the interaction picture we obtain the equation

$$\frac{\partial}{\partial t}[\mathscr{P}f_I(t)] + i\mathscr{P}\mathscr{L}_{RSI}(t)[\mathscr{P}f_I(t)] + \int_0^t \mathscr{P}\mathscr{L}_{RSI}(t)\mathscr{U}(t,\tau)(1-\mathscr{P})\mathscr{L}_{RSI}(\tau)[\mathscr{P}f_I(\tau)]d\tau = 0, \tag{5.16}$$

where

$$\mathscr{U}(t,\tau) = T\exp\left\{-i\int_\tau^t dt'(1-\mathscr{P})\mathscr{L}_{RSI}(t')(1-\mathscr{P})\right\}. \tag{5.17}$$

In eq. (5.16) $\mathscr{L}_{RSI}(t)$ is the Liouville operator $\mathscr{L}_{RS}$ in the interaction picture, i.e.

$$\mathscr{L}_{RSI}(t) = \exp\{i(\mathscr{L}_R + \mathscr{L}_S)\tau\}\mathscr{L}_{RS}\exp\{-i(\mathscr{L}_R + \mathscr{L}_S)\tau\}. \tag{5.18}$$

We now consider the case of the system S interacting weakly with the reservoir R. In the lowest Born approximation we replace $\mathscr{U}(t,\tau)$ by unity. We further assume that the properties of the reservoir are such that

$$\mathscr{P}\mathscr{L}_{\mathrm{RSI}}(t)\mathscr{P}f_{\mathrm{I}}(\tau) = 0. \tag{5.19}$$

In most of the applications which we will be considering this condition will be satisfied (if it is not, its effect can be taken into account by renormalization). Then (5.16) reduces to

$$\frac{\partial}{\partial t}[\mathscr{P}f(t)] + \int_0^t \mathscr{P}\mathscr{L}_{\mathrm{RS}}(t)\mathscr{L}_{\mathrm{RS}}(\tau)[\mathscr{P}f(\tau)]\mathrm{d}\tau = 0, \tag{5.20}$$

where, for the sake of brevity, we have suppressed the subscript I. The Markovian form of the approximate master equation (5.20) will be the basis for all the applications which we consider in the rest of the article.

We write the explicit form of (5.20) when f is the phase space distribution function and when H_{RS} is of the form

$$H_{\mathrm{RS}} = \sum_k (G_{\mathrm{R}k} G_{\mathrm{S}k}^{+} + \mathrm{H.C.}), \tag{5.21}$$

where $G_{\mathrm{R}k}(G_{\mathrm{S}k})$ is an operator acting on the reservoir (system) variables alone. In the interaction picture we have

$$H_{\mathrm{RS}}(t) = \sum_k (G_{\mathrm{R}k}(t)G_{\mathrm{S}k}^{+}(t) + \mathrm{H.C.}). \tag{5.22}$$

The Liouville operator $\mathscr{L}_{\mathrm{RS}}(t)$ is then given by (cf. eq. (2.24))

$$\mathscr{L}_{\mathrm{RS}}(t) = \sum_k \{\psi_{\mathrm{R}k}(t)\mathscr{D}_{\mathrm{R}}^{(1)}\psi_{\mathrm{S}k}(t)\mathscr{D}_{\mathrm{S}}^{(1)} + \psi_{\mathrm{R}k}^{*}(t)\mathscr{D}_{\mathrm{R}}^{(1)}\psi_{\mathrm{S}k}(t)\mathscr{D}_{\mathrm{S}}^{(1)} + \mathrm{C.C.}\}, \tag{5.23}$$

where $\psi_{\mathrm{R}k}(t)$ is the Ω-equivalent of $G_{\mathrm{R}k}(t)$ and $\psi_{\mathrm{S}k}(t)$ is the Ω-equivalent of $G_{\mathrm{S}k}(t)$. On substituting from (5.23) in (5.20), we obtain the equation

$$\frac{\partial \Phi_{\mathrm{S}}^{(\Omega)}}{\partial t} + \int_0^t K(t,\tau)\mathrm{d}\tau = 0, \tag{5.24}$$

where

$$\begin{aligned} K(t,\tau) = \sum_{kl} \{ & \Gamma_{11\mathrm{R}}^{lk}(t,\tau)[\psi_{\mathrm{S}l}^{*}(t)\mathscr{D}_{\mathrm{S}}^{(1)}(\psi_{\mathrm{S}k}^{*}(\tau)\mathscr{D}_{\mathrm{S}}^{(1)}\Phi_{\mathrm{S}}^{(\Omega)}(\tau))] \\ & + \Gamma_{12\mathrm{R}}^{lk}(t,\tau)[\psi_{\mathrm{S}l}^{*}(t)\mathscr{D}_{\mathrm{S}}^{(1)}(\psi_{\mathrm{S}k}(\tau)\mathscr{D}_{\mathrm{S}}^{(1)}\Phi_{\mathrm{S}}^{(\Omega)}(\tau))] \\ & - \Gamma_{11\mathrm{R}}^{kl}(\tau,t)[\psi_{\mathrm{S}l}^{*}(t)\mathscr{D}_{\mathrm{S}}^{(1)}(\psi_{\mathrm{S}k}^{*}(\tau)\mathscr{D}_{\mathrm{S}}^{(2)}\Phi_{\mathrm{S}}^{(\Omega)}(\tau))] \\ & - \Gamma_{21\mathrm{R}}^{kl}(\tau,t)[\psi_{\mathrm{S}l}^{*}(t)\mathscr{D}_{\mathrm{S}}^{(1)}(\psi_{\mathrm{S}k}(\tau)\mathscr{D}_{\mathrm{S}}^{(2)}\Phi_{\mathrm{S}}^{(\Omega)}(\tau))] \\ & + \Gamma_{21\mathrm{R}}^{lk}(t,\tau)[\psi_{\mathrm{S}l}(t)\mathscr{D}_{\mathrm{S}}^{(1)}(\psi_{\mathrm{S}k}^{*}(\tau)\mathscr{D}_{\mathrm{S}}^{(1)}\Phi_{\mathrm{S}}^{(\Omega)}(\tau))] \\ & + \Gamma_{22\mathrm{R}}^{lk}(t,\tau)[\psi_{\mathrm{S}l}(t)\mathscr{D}_{\mathrm{S}}^{(1)}(\psi_{\mathrm{S}k}(\tau)\mathscr{D}_{\mathrm{S}}^{(1)}\Phi_{\mathrm{S}}^{(\Omega)}(\tau))] \\ & - \Gamma_{12\mathrm{R}}^{kl}(\tau,t)[\psi_{\mathrm{S}l}(t)\mathscr{D}_{\mathrm{S}}^{(1)}(\psi_{\mathrm{S}k}^{*}(\tau)\mathscr{D}_{\mathrm{S}}^{(2)}\Phi_{\mathrm{S}}^{(\Omega)}(\tau))] \\ & - \Gamma_{22\mathrm{R}}^{kl}(\tau,t)[\psi_{\mathrm{S}l}(t)\mathscr{D}_{\mathrm{S}}^{(1)}(\psi_{\mathrm{S}k}(\tau)\mathscr{D}_{\mathrm{S}}^{(2)}\Phi_{\mathrm{S}}^{(\Omega)}(\tau))] + \mathrm{C.C.}\}. \end{aligned} \tag{5.25}$$

In eq. (5.25) $\Gamma^{kl}_{ij,\mathrm{R}}(t,\tau)$ are the elements of the reservoir correlation matrix defined by

$$\Gamma^{kl}{}_{,\mathrm{R}}(t,\tau) = \begin{pmatrix} \mathrm{tr}\,(\rho_{\mathrm{R}}(0)G_{\mathrm{R}k}(t)G_{\mathrm{R}l}(\tau)) & \mathrm{tr}\,(\rho_{\mathrm{R}}(0)G_{\mathrm{R}k}(t)G^{+}_{\mathrm{R}l}(\tau)) \\ \mathrm{tr}\,(\rho_{\mathrm{R}}(0)G^{+}_{\mathrm{R}k}(t)G_{\mathrm{R}l}(\tau)) & \mathrm{tr}\,(\rho_{\mathrm{R}}(0)G^{+}_{\mathrm{R}k}(t)G^{+}_{\mathrm{R}l}(\tau)) \end{pmatrix}. \quad (5.26)$$

We now discuss the effect of the weak external field on the system S which is interacting with a reservoir. Let the effect of the external field be represented by the Hamiltonian $H_{\mathrm{ext}}(t)$. Then the master equation (5.24) is modified, having the form

$$\frac{\partial \Phi^{(\Omega)}_{\mathrm{S}}}{\partial t} + \int_0^t K(t,\tau)\mathrm{d}\tau + \mathrm{i}\mathscr{L}_{\mathrm{ext}}(t)\Phi^{(\Omega)}_{\mathrm{S}} = 0, \quad (5.27)$$

where

$$\mathscr{L}_{\mathrm{ext}}(t) = [H_{\mathrm{ext}}(t), \quad], \quad (5.28)$$

and all the quantities are in the interaction picture.

§ 6. Relaxation of a Harmonic Oscillator

We first apply the theory developed in § 5 to the problem of a harmonic oscillator interacting with a reservoir. We present a fairly detailed treatment of this problem because this elementary example illustrates very clearly the techniques developed in previous sections. Our presentation follows closely that of AGARWAL [1969] (for various other treatments see MONTROLL and SHULER [1957], SENITZKY [1960, 1961], LOUISELL and WALKER [1965], WEIDLICH and HAAKE [1965], LOUISELL and MARBURGER [1967], OPPENHEIM, SHULER and WEISS [1967], LAX [1966c]).

The Hamiltonian of the total system consisting of the oscillator S and the reservoir R is assumed to be given by

$$H = \omega a^{+}a + \sum_k \omega_k a_k^{+} a_k + \sum_k (g_k a_k^{+} a + \mathrm{H.C.}), \quad (6.1)$$

where we have also assumed, for the sake of simplicity, that the reservoir is made up of harmonic oscillators. The operators a, a^{+}, a_k and a_k^{+} satisfy the commutation relations

$$[a, a^{+}] = 1, \qquad [a_k, a_k^{+}] = 1, \quad (6.2)$$

and all other commutators vanish. We first obtain the master equation for the reduced phase-space distribution function which is obtained from the density operator by the normal rule of mapping. The functions $\psi_{k\mathrm{R}}$ and $\psi_{k\mathrm{S}}$ for the problem under consideration are given by

$$\psi_{k\mathrm{R}}(t) = g_k z_k \exp\{-\mathrm{i}\omega_k t\}, \qquad \psi_{k\mathrm{S}}(t) = z \exp\{-\mathrm{i}\omega t\}. \quad (6.3)$$

The operators $\mathscr{D}_S^{(1)}$ and $\mathscr{D}_S^{(2)}$ (defined by eqs. (2.21) and (2.22) respectively) for the normal rule of mapping are given by

$$\mathscr{D}_S^{(1)} = \exp\left(\frac{\overleftarrow{\partial}}{\partial z}\frac{\overrightarrow{\partial}}{\partial z^*}\right), \qquad \mathscr{D}_S^{(2)} = \exp\left(\frac{\overleftarrow{\partial}}{\partial z^*}\frac{\overrightarrow{\partial}}{\partial z}\right), \tag{6.4}$$

and the correlation matrix $\Gamma_R^{kl}(t, \tau)$ is easily seen to be given by

$$\Gamma_R^{kl}(t, \tau) = |g_k|^2 \begin{pmatrix} 0 & (\langle n_k\rangle + 1)\exp(-i\omega_k(t-\tau)) \\ \langle n_k\rangle \exp(i\omega_k(t-\tau)) & 0 \end{pmatrix} \delta_{kl}. \tag{6.5}$$

In eq. (6.5) $\langle n_k\rangle$ is the average occupation number for the kth mode of the oscillator of the heat bath and is equal to

$$\langle n_k\rangle = (\exp\{\beta\omega_k\} - 1)^{-1}, \qquad \beta = 1/K_B T, \tag{6.6}$$

where K_B is the Boltzmann constant and T is the temperature of the heat bath. On substituting eqs. (6.3)–(6.6) in eq. (5.24), we find that $\Phi_S^{(N)}(z, z^*)$ satisfies the following master equation

$$\frac{\partial \Phi^{(N)}(t)}{\partial t} = \int_0^t \left\{ \sum_k |g_k|^2 \exp[-i(\omega_k - \omega)\tau] \right.$$
$$\left. \times \left[\frac{\partial}{\partial z}(z\Phi^{(N)}(t-\tau)) + (1+\langle n_k\rangle)\frac{\partial^2}{\partial z \partial z^*}\Phi^{(N)}(t-\tau)\right] + \text{C.C.}\right\} d\tau, \tag{6.7}$$

where $\Phi^{(N)}(t)$ corresponds to the density operator in the interaction picture and where, for the sake of brevity, we have dropped the subscript S from $\Phi_S^{(N)}$. It is seen that (6.7) is a non-Markovian equation, i.e. the time rate change of $\Phi^{(N)}$ at time t depends on the values of $\Phi^{(N)}$ at all earlier times. The equation (6.7) can be solved exactly by taking its Laplace transform. However, in what follows we consider only the Markovian behavior of an oscillator interacting with the heat bath (for some of the non-Markovian effects associated with (6.7) see HAAKE [1969a] and for a general discussion of non-Markovian effects see ZWANZIG [1961b], AGARWAL [1973c]).

We assume that the bath oscillators are closely spaced in frequency so that $\sum_k$ can be replaced by $\int d\omega_k h(\omega_k) \ldots$, where $h(\omega_k)d\omega_k$ is the number of oscillators with frequencies lying between ω_k and $\omega_k + d\omega_k$. Moreover we assume that the reservoir correlations have a short correlation time τ_c so that for $t \gg \tau_c$ we can replace $\Phi^{(N)}(t-\tau)$ in (6.7) by $\Phi^{(N)}(t)$ and extend the upper limit of integration to infinity (short memory approximation). Then eq. (6.7) reduces to

$$\frac{\partial \Phi^{(N)}}{\partial t} = (\gamma + i\Delta\omega)\frac{\partial}{\partial z}(z\Phi^{(N)}) + \gamma(1 + \langle n(\omega)\rangle)\frac{\partial^2 \Phi^{(N)}}{\partial z \partial z^*} + \text{C.C.}, \quad (6.8)$$

where

$$\gamma = \pi h(\omega)|g(\omega)|^2, \qquad \Delta\omega = P\int d\omega_k\, h(\omega_k)|g(\omega_k)|^2(\omega - \omega_k)^{-1}, \quad (6.9)$$

and where P denotes the Cauchy principal value of the integral. The distribution function $\Phi^{(N)}$ in (6.8) corresponds to the density operator in the interaction picture and on transforming to the Schrödinger picture (cf. AGARWAL [1971d, Appendix A]) we find that $\Phi^{(N)}$ satisfies the following master equation

$$\frac{\partial \Phi^{(N)}}{\partial t} = \{i(\omega + \Delta\omega) + \gamma\}\frac{\partial}{\partial z}(z\Phi^{(N)}) + \gamma(1 + \langle n(\omega)\rangle)\frac{\partial^2 \Phi^{(N)}}{\partial z \partial z^*} + \text{C.C.}. \quad (6.10)$$

It is seen from (6.10) that $\Delta\omega$ causes the frequency shift and its effect can be taken into account by frequency renormalization*. In what follows we will ignore the effect of this term. We may similarly obtain the master equation for the Wigner distribution function and the Sudarshan-Glauber distribution function. We write collectively these master equations as follows

$$\frac{\partial \Phi^{(\Omega)}}{\partial t} = (i\omega + \gamma)\frac{\partial}{\partial z}(z\Phi^{(\Omega)}) + \gamma(\lambda + \tfrac{1}{2} + \langle n(\omega)\rangle)\frac{\partial^2 \Phi^{(\Omega)}}{\partial z \partial z^*} + \text{C.C.}, \quad (6.11)$$

where we recall that λ is $+\frac{1}{2}$ for the normal rule of mapping, $-\frac{1}{2}$ for the antinormal rule of mapping and is zero for the Weyl rule of mapping.

It should be noted that the master equation under the approximations discussed above, is of the form of a Fokker-Planck equation and we can now use the methods of the stochastic processes to study the dynamics of our system. The stochastically equivalent Langevin equations corresponding to the Fokker-Planck process (6.11) are given by (cf. eq. (A.35))

$$\dot{z} = -i\omega z - \gamma z + F(t), \qquad \dot{z}^* = i\omega z^* - \gamma z^* + F^*(t), \quad (6.12)$$

where $F(t)$ is a delta correlated complex Gaussian random process with the properties

$$\langle F(t)\rangle = \langle F^*(t)\rangle = \langle F(t)F(t')\rangle = \langle F^*(t)F^*(t')\rangle = 0,$$
$$\langle F(t)F^*(t')\rangle = 2\gamma(\lambda + \tfrac{1}{2} + \langle n(\omega)\rangle)\delta(t - t'), \quad (6.13)$$

and all the higher order linked moments (cumulants) of $F(t)$ vanish. For the calculation of the moments one can use either the Fokker-Planck equation (6.11) or the Langevin equations (6.12). Since the Langevin

* Cf. footnote on p. 45.

equations (6.12) correspond to the quantum system, it is clear that the moments calculated from (6.12) would correspond to normally ordered moments $\langle a^{+m}a^n\rangle$ for $\lambda = -\frac{1}{2}$, to anti-normally ordered moments $\langle a^n a^{+m}\rangle$ for $\lambda = \frac{1}{2}$ and to Weyl ordered moments $\langle (a^{+m}a^n)_W\rangle$ for $\lambda = 0$. The solution of (6.12) is given by

$$z(t) = z(0)\mathrm{e}^{-\mathrm{i}\omega t-\gamma t} + \int_0^t F(t-\tau)\mathrm{e}^{-\mathrm{i}\omega\tau-\gamma\tau}\mathrm{d}\tau. \tag{6.14}$$

On using (6.13) and (6.14) it follows that

$$\langle z^*(t)z(t)\rangle = \langle z^*(0)z(0)\rangle \mathrm{e}^{-2\gamma t} + (\lambda+\tfrac{1}{2}+\langle n(\omega)\rangle)(1-\mathrm{e}^{-2\gamma t}), \tag{6.15}$$

and, therefore, for $\lambda = -\frac{1}{2}$, (6.15) leads to

$$\langle a^+(t)a(t)\rangle = \langle a^+(0)a(0)\rangle \mathrm{e}^{-2\gamma t} + \langle n(\omega)\rangle(1-\mathrm{e}^{-2\gamma t}). \tag{6.16}$$

It is seen from (6.16) that the average occupation number at time t is a sum of two terms: a term which depends on the initial occupation number and a term which depends on the temperature of the reservoir and which vanishes as the temperature $\to 0$.

The equation of motion for the diagonal elements ρ_{nn} of the reduced density operator can be obtained by using the relation (which can be obtained from eq. (2.9) by taking the diagonal matrix elements with respect to the Fock state $|n\rangle$)

$$\rho_{nn} = \int \Phi^{(A)}(z, z^*)\, \frac{|z|^{2n}\mathrm{e}^{-|z|^2}}{n!}\, \mathrm{d}^2 z, \tag{6.17}$$

and the eq. (6.11) for $\lambda = -\frac{1}{2}$. A straightforward calculation shows that the diagonal elements satisfy the master equation

$$\begin{aligned}\frac{\partial \rho_{nn}}{\partial t} &= 2\gamma\langle n(\omega)\rangle n\rho_{n-1,n-1} + 2\gamma(1+\langle n(\omega)\rangle)(n+1)\rho_{n+1\,n+1} \\ &\quad -2\gamma\{n(1+2\langle n(\omega)\rangle)+\langle n(\omega)\rangle\}\rho_{nn}.\end{aligned} \tag{6.18}$$

The master equation (6.18) is of the form of the Pauli equation. The flow of the probability is schematically represented in Fig. 1. It is seen from this figure that the steady state solution is given by

$$2\gamma(n+1)\langle n(\omega)\rangle\rho_{nn} = 2\gamma(n+1)(1+\langle n(\omega)\rangle)\rho_{n+1\,n+1}, \tag{6.19}$$

which leads to the Bose-Einstein distribution for ρ_{nn}

$$\rho_{nn} = \langle n(\omega)\rangle^n/(1+\langle n(\omega)\rangle)^{n+1}. \tag{6.20}$$

The operator form of the master equation is easily obtained by using (6.11)

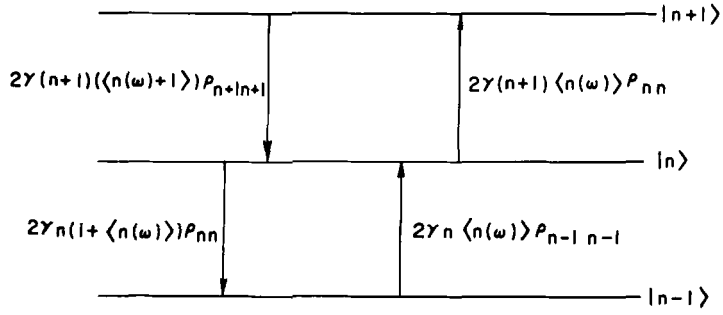


Fig. 1. Schematic representation of eq. (6.18).

for $\lambda = -\frac{1}{2}$ and by using the identities (2.28). The reduced density operator satisfies the master equation

$$\frac{\partial \rho}{\partial t} = -i\omega[a^+ a, \rho] - \gamma\{a^+ a\rho - 2a\rho a^+ + \rho a^+ a\} - 2\gamma\langle n(\omega)\rangle[a^+, [a, \rho]]. \tag{6.21}$$

We now present the solution of the relaxation eq. (6.11). It should be noted that (6.11) is of the form of a linearized Fokker-Planck equation. From the solution (A.24) of the linearized Fokker-Planck equation it can be shown that the Green's function $K^{(\Omega)}(z, z^*, t|z_0, z_0^*, 0)$ associated with eq. (6.11) is given by

$$K^{(\Omega)}(z, z^*, t|z_0, z_0^*, 0) = \{\pi(1-e^{-2\gamma t})(\lambda+\tfrac{1}{2}+\langle n(\omega)\rangle)\}^{-1} \times \exp\left[-|z - z_0 e^{-\gamma t - i\omega t}|^2\{(1-e^{-2\gamma t})(\lambda+\tfrac{1}{2}+\langle n(\omega)\rangle)\}^{-1}\right]. \tag{6.22}$$

The distribution function $\Phi^{(\Omega)}$ is then obtained from the relation

$$\Phi^{(\Omega)}(z, z^*, t) = \int K^{(\Omega)}(z, z^*, t|z_0, z_0^*, 0)\Phi^{(\Omega)}(z_0, z_0^*, 0)\mathrm{d}^2 z_0, \tag{6.23}$$

where $\Phi^{(\Omega)}(z_0, z_0^*, 0)$ is the distribution function at time $t = 0$ and $K^{(\Omega)}$ is given by (6.22). If the initial state of the oscillator is a coherent state then $\Phi^{(A)}(z, z^*, t)$ is given by (6.22) with $\lambda = -\frac{1}{2}$. In addition if the reservoir is at zero temperature $\langle n(\omega)\rangle \to 0$, then $\Phi^{(A)}$ is given by

$$\Phi^{(A)}(z, z^*, t) = \delta^{(2)}(z - z_0 e^{-i\omega t - \gamma t}). \tag{6.24}$$

The result (6.24) shows that if the oscillator is initially in a coherent state then it remains in a coherent state with exponentially decaying amplitude. A similar result was found for closed systems, characterized by the Hamiltonian of the form (6.1), by MEHTA and SUDARSHAN [1966].

The multi-time correlation functions are easily computed by using (6.22) and (2.27). In particular it can be shown that

$$\langle a^+(t)a(0)\rangle = \langle a^+(0)a(0)\rangle e^{i\omega t-\gamma t}, \tag{6.25}$$

$$\langle a^+(0)a^+(t)a(t)a(0)\rangle = \langle a^+(0)a(0)\rangle\langle n(\omega)\rangle(1-e^{-2\gamma t}) + \langle (a^+(0))^2(a(0))^2\rangle e^{-2\gamma t}. \tag{6.26}$$

One may also compute a time dependent entropy (AGARWAL [1971a]) for the oscillator system. It can be shown that the quantum entropy S for the oscillator system, if initially it was in a coherent state, is given by

$$S = K_B\{(\sigma+1)\ln(\sigma+1)-\sigma\ln\sigma\}, \tag{6.27}$$

where

$$\sigma = \langle n(\omega)\rangle(1-e^{-2\gamma t}). \tag{6.28}$$

Thus one may introduce the concept of a time dependent temperature $T(t)$ by (see also LOUISELL and WALKER [1965])

$$T(t) = \omega\{K_B \ln(1+\sigma^{-1})\}^{-1}. \tag{6.29}$$

In fact the reduced density operator $\rho(t)$ which is obtained from the solution of (6.21) under the initial condition $\rho(0) = |z_0\rangle\langle z_0|$ is given by

$$\rho(t) = \frac{\exp[-\beta(t)(a^+ - z_0^* e^{i\omega t-\gamma t})(a - z_0 e^{-i\omega t-\gamma t})]}{\mathrm{Tr}\exp[-\beta(t)(a^+ - z_0^* e^{i\omega t-\gamma t})(a - z_0 e^{-i\omega t-\gamma t})]}, \tag{6.30}$$

where

$$\beta(t) \equiv \omega\{K_B T(t)\}^{-1}. \tag{6.31}$$

Thus the oscillator system at time t is found in thermal equilibrium (in terms of displaced coordinates) at a time dependent temperature $T(t)$ given by eq. (6.29).

We now present the solution of the Pauli equation (6.18). We introduce the generating function $Q(x, t)$ defined by

$$Q(x, t) = \sum_0^\infty (1-x)^n \rho_{nn}(t). \tag{6.32}$$

It is easily shown that $Q(x, t)$ satisfies the following first order differential equation

$$\partial Q/\partial t = -2\gamma\langle n(\omega)\rangle xQ - 2\gamma[1+\langle n(\omega)\rangle x]x\,\partial Q/\partial x. \tag{6.33}$$

This equation is easily solved by the method of characteristics (SNEDDON [1957, Chap. II]) and the solution is

$$Q(x, t) = \{1+\langle n(\omega)\rangle x(1-e^{-2\gamma t})\}^{-1} Q_0[x e^{-2\gamma t}\{1+\langle n(\omega)\rangle x(1-e^{-2\gamma t})\}^{-1}], \tag{6.34}$$

TABLE 2

The time dependence of the diagonal elements $\rho_{n,n}(t)$, for the problem of harmonic oscillator relaxation, for a class of initial distributions

Initial distribution	$\rho_{n,n}(0)$	$Q_0(x)$	$\rho_{n,n}(t)$	Remarks
Bose-Einstein distribution	$\bar{n}^n/(\bar{n}+1)^{n+1}$	$(1+\bar{n}x)^{-1}$	$\delta^n/(1+\delta)^{n+1}$ $\delta \equiv \langle n(\omega)\rangle + (\bar{n} - \langle n(\omega)\rangle)\mathrm{e}^{-2\gamma t}$	The distribution remains Bose-Einstein
Poisson distribution (coherent state)	$\dfrac{\bar{n}^n \mathrm{e}^{-\bar{n}}}{n!}$	$\mathrm{e}^{-\bar{n}x}$	$\dfrac{\mu^n}{(1+\mu)^{n+1}} \exp\left(-\bar{n}\,\mathrm{e}^{-2\gamma t}/(1+\mu)\right)$ $\times L_n(-\bar{n}\,\mathrm{e}^{-2\gamma t}/\mu(1+\mu))$ $\mu \equiv \langle n(\omega)\rangle(1-\mathrm{e}^{-2\gamma t})$	This distribution corresponds to a $\Phi^{(A)}$ which is equal to the convolution of two probability distributions, one corresponding to a coherent state and the other to thermal equilibrium
Fock state $\lvert m\rangle$	$\delta_{n,m}$	$(1-x)^m$	$\dfrac{(1-\mathrm{e}^{-2\gamma t})^{n+m}(\mathrm{e}^{\beta\omega}-1)\mathrm{e}^{m\beta\omega}}{(\mathrm{e}^{\beta\omega}-\mathrm{e}^{-2\gamma t})^{n+m+1}} \times$ $\times F\left[-n, -m, 1; \left(\dfrac{\sinh\frac{1}{2}\beta\omega}{\sinh\gamma t}\right)^2\right]$ $\beta \equiv 1/K_{\mathrm{B}}T$	This distribution at zero temperature becomes the binomial distribution with parameter $\exp(-2\gamma t)$

where $Q_0(x)$ is the generating function at time $t = 0$. The time dependence of the diagonal elements of ρ is then calculated from

$$\rho_{nn}(t) = \frac{(-1)^n}{n!} \frac{\partial^n Q(x, t)}{\partial x^n}\bigg|_{x=1}. \tag{6.35}$$

In Table 2, we give the values of $\rho_{nn}(t)$ for several important initial distributions. At zero temperature ($\langle n(\omega)\rangle = 0$), eqs. (6.34) and (6.35) lead to the following result for the time dependence of $\rho_{nn}(t)$

$$\rho_{nn}(t) = \sum_{m \geq n} \binom{m}{n} e^{-2\gamma t n}(1 - e^{-2\gamma t})^{m-n} \rho_{mm}(0). \tag{6.36}$$

We next consider the effect of a weak external field on the relaxation of the oscillator. The Hamiltonian of our system in the presence of an external field is given by

$$H = \omega a^+ a + \sum_k \omega_k a_k^+ a_k + \sum_k (g_k a_k^+ a + \text{H.C.}) + g_e[\varepsilon^*(t) a + \text{H.C.}], \tag{6.37}$$

where $\varepsilon(t)$ is a c-number driving field. The reduced phase space distribution function satisfies the master equation

$$\frac{\partial}{\partial t} \Phi^{(\Omega)} = (i\omega + \gamma) \frac{\partial}{\partial z}(z\Phi^{(\Omega)}) + i\varepsilon(t) g_e \frac{\partial}{\partial z} \Phi^{(\Omega)} + \gamma(\lambda + \tfrac{1}{2} + \langle n(\omega)\rangle) \frac{\partial^2 \Phi^{(\Omega)}}{\partial z \partial z^*} + \text{C.C.} \tag{6.38}$$

The Fokker-Planck equation (6.38) is stochastically equivalent to the Langevin equations

$$\dot{z} = -i\omega z - \gamma z - i\varepsilon(t) + F(t), \qquad \dot{z}^* = i\omega z^* - \gamma z^* + i\varepsilon^*(t) + F^*(t), \tag{6.39}$$

where $F(t)$ is a delta correlated complex Gaussian random process with properties given by eq. (6.13) and where we have set $g_e = 1$. Eq. (6.38) has the form of a linearized Fokker-Planck equation and the Green's function associated with (6.38) is, therefore, given by

$$K^{(\Omega)}(z, z^*, t | z_0, z_0^*, 0) = \frac{\exp\{-|z - \langle z(t)\rangle|^2 / \{(\lambda + \tfrac{1}{2} + \langle n(\omega)\rangle)(1 - e^{-2\gamma t})\}\}}{\pi(\lambda + \tfrac{1}{2} + \langle n(\omega)\rangle)(1 - e^{-2\gamma t})}, \tag{6.40}$$

where

$$\langle z(t)\rangle = z_0 e^{-i\omega t - \gamma t} - i \int_0^t \varepsilon(t - \tau) e^{-i\omega\tau - \gamma\tau} d\tau. \tag{6.41}$$

On comparing (6.22) and (6.40), we find that the effect of the external field is to displace the center of the Gaussian distribution. The other

dynamical properties may be calculated from (6.40) and we will not discuss them here (they are considered in LOUISELL and MARBURGER [1967]).

SCHWINGER [1961] has discussed the problem of the relaxation of the oscillator from a different standpoint. He considers the augmented Hamiltonian

$$H = H_0 + K^*(t)a + K(t)a^+, \tag{6.42}$$

where H_0 is given by (6.1) and $K(t)$ is an external force. He introduced the functional (see also SCULLY and WHITNEY [1972])

$$Z = \mathrm{Tr}\,\{U_-^+(t, t_0)U_+(t, t_0)\rho(t_0)\}, \tag{6.43}$$

where

$$U_+(t, t_0) = T \exp\left\{-\mathrm{i}\int_{t_0}^{t}(K_+^*(\tau)a(\tau) + K_+(\tau)a^+(\tau))\mathrm{d}\tau\right\}, \tag{6.44}$$

$$U_-^+(t, t_0) = \tilde{T} \exp\left\{\mathrm{i}\int_{t_0}^{t}(K_-^*(\tau)a(\tau) + K_-(\tau)a^+(\tau))\mathrm{d}\tau\right\}, \tag{6.45}$$

and T and $\tilde{T}$ are the operators for the chronological and the antichronological orderings respectively. The operators $a(\tau)$, $a^+(\tau)$ are in the interaction picture with respect to H_0. The variation of the functional Z is given by

$$\begin{aligned}\delta \ln Z = -\mathrm{i}\int_{t_0}^{t}\{&\mathscr{A}_+^*(\tau)\delta K_+(\tau) + \mathscr{A}_+(\tau)\delta K_+^*(\tau) \\ &- \mathscr{A}_-^*(\tau)\delta K_-(\tau) - \mathscr{A}_-(\tau)\delta K_-^*(\tau)\}\mathrm{d}\tau,\end{aligned} \tag{6.46}$$

where

$$\mathscr{A}_+(\tau) \equiv \mathrm{i}\frac{\delta \ln Z}{\delta K_+^*(\tau)}, \qquad \mathscr{A}_+^*(\tau) \equiv \mathrm{i}\frac{\delta \ln Z}{\delta K_+(\tau)}, \tag{6.47}$$

$$\mathscr{A}_-(\tau) \equiv -\mathrm{i}\frac{\delta \ln Z}{\delta K_-^*(\tau)}, \qquad \mathscr{A}_-^*(\tau) \equiv -\mathrm{i}\frac{\delta \ln Z}{\delta K_-(\tau)}. \tag{6.48}$$

Thus the functional $\mathscr{A}_+^*(t)$ is given by

$$\begin{aligned}\mathscr{A}_+^*(t) = \mathrm{Tr}\Big\{&U_-^+(t, t_0)T \\ &\times \left[a^+(t)\exp\left\{-\mathrm{i}\int_{t_0}^{t}(K_+^*(\tau)a(\tau) + K_+(\tau)a^+(\tau))\mathrm{d}\tau\right\}\right]\rho(t_0)\Big\}.\end{aligned} \tag{6.49}$$

The functionals $\mathscr{A}_-^*(\tau)$, $\mathscr{A}_-(\tau)$, $\mathscr{A}_+(\tau)$ are given by expressions similar to (6.49). The functional Z is then calculated by using the equations of motion for the functional $\mathscr{A}_+(\tau)$, $\mathscr{A}_+^*(\tau)$. The correlation functions are obtained from Z by the functional differentiation of Z. (The details can be found in a recent review article by SCULLY and WHITNEY [1972].)

The functional Z is closely related to the phase space distribution functions. For example when $K_+(t) = K^*(t) = 0$ then (6.43) reduces to

$$Z = \left\langle \tilde{T} \exp\left\{ i \int_{t_0}^{t} K_-(\tau) a^+(\tau) d\tau \right\} T \exp\left\{ i \int_{t_0}^{t} K_+^*(\tau) a(\tau) d\tau \right\} \right\rangle, \quad (6.50)$$

which is seen to be the generating functional for the normally-ordered time-ordered correlation functions of the form (2.26). Moreover, in the special case when

$$K_-(\tau) = -i\alpha\delta(\tau - t), \quad K_+^*(\tau) = i\alpha^*\delta(\tau - t), \quad (6.51)$$

(6.50) reduces to

$$Z = \langle \exp(\alpha a^+) \exp(-\alpha^* a) \rangle, \quad (6.52)$$

and the Fourier transform of (6.52) is just the Sudarshan-Glauber distribution function.

The model which we have discussed in this section is of great value in quantum optics. It is used to describe the damping of the field mode. Consider a nonlinear process in which several field modes are interacting with each other and let this interaction be described by the Hamiltonian H. The equation of motion for the Sudarshan-Glauber distribution function in the absence of losses is of the form (2.23). The losses are taken into account by assuming that each field mode is interacting with its own reservoir characterized by the parameters γ_i and $\langle n_i \rangle$. Then the equation of motion for $\Phi^{(A)}$, when losses are taken into account, is of the form

$$\frac{\partial \Phi^{(A)}}{\partial t} = -i\mathscr{L}^{(A)}\Phi^{(A)} + \sum_i \left\{ (i\omega_i + \gamma_i) \frac{\partial}{\partial z_i} (z_i \Phi^{(A)}) + \gamma_i \langle n_i \rangle \frac{\partial^2 \Phi^{(A)}}{\partial z_i \partial z_i^*} + \text{C.C.} \right\}. \quad (6.53)$$

We apply this procedure to obtain the Fokker-Planck equation for the parametric oscillator. The Hamiltonian for this system, in the interaction picture, is given by

$$H = (i\chi a_1^+ a_2^+ a_3 + \text{H.C.}) - i(F_p^*(t) a_3 - \text{H.C.}), \quad (6.54)$$

where we have ignored the effects of detuning. The parameter χ is related to the nonlinear susceptibility of the medium (for the form of χ see GRAHAM [1968]). In (6.54) $F_p(t)$ is the external pump field. On using (6.54) and (2.24) generalized to the case of many degrees of freedom (see Table 1) $\mathscr{L}^{(A)}$ is easily calculated. On using this value of $\mathscr{L}^{(A)}$ and (6.53), we obtain the Fokker-Planck equation

$$\frac{\partial \Phi^{(A)}}{\partial t} = \left\{ \frac{\partial}{\partial z_1}(\gamma_1 z_1 - \chi z_2^* z_3) + \frac{\partial}{\partial z_2}(\gamma_2 z_2 - \chi z_1^* z_3) \right.$$
$$\left. + \frac{\partial}{\partial z_3}(\gamma_3 z_3 - F_p(t) + \chi z_1 z_2) + \chi \frac{\partial^2}{\partial z_1 \partial z_2} z_3 + \sum_{i=1}^{3} \gamma_i \langle n_i \rangle \frac{\partial^2}{\partial z_i \partial z_i^*} + \text{C.C.} \right\} \Phi^{(A)}. \tag{6.55}$$

(6.55) is the basic equation of GRAHAM [1968]. Below and above threshold, this equation can be solved by linearization and by quasilinearization procedures respectively.

§ 7. Brownian Motion of a Quantum Oscillator

In this section we treat the problem of the Brownian motion of a quantum oscillator. Our approach is based on the master equation satisfied by the distribution function corresponding to the oscillator system alone. The treatment presented here follows closely the one given by AGARWAL [1971d] (for various other approaches see SCHWINGER [1961], ULLERSMA [1966a, b]). The Brownian motion of a classical oscillator is discussed in great detail in the papers of KRAMERS [1940], CHANDRASEKHAR [1943] and WANG and UHLENBECK [1945].

The displacement $q(t)$ of the oscillator satisfies the equation of motion

$$\mathrm{d}^2 q/\mathrm{d}t^2 + 2\gamma\, \mathrm{d}q/\mathrm{d}t + \omega^2 q = F(t)/m, \tag{7.1}$$

where 2γ is the phenomenological damping coefficient, ω is the natural frequency of oscillation and m is the mass of the particle. $F(t)$ is a random force which is assumed to be a delta correlated real Gaussian process with zero mean, i.e.

$$\langle F(t) \rangle = 0, \qquad \langle F(t) F(t') \rangle = 2D\delta(t-t'), \tag{7.2}$$

where D is the diffusion coefficient. Eq. (7.1) is equivalent to the following two first order differential equations

$$\dot{q} = p/m, \qquad \dot{p} = -2\gamma p - m\omega^2 q + F(t), \tag{7.3}$$

where p is the momentum of the particle. Equations (7.3) are the Langevin equations describing the Brownian motion of a classical oscillator and the corresponding random process is a two dimensional Gaussian Markov process (cf. Appendix and also STRATONOVICH [1963]).

We now discuss the Brownian motion of a quantum oscillator. The Brownian motion is described by a model Hamiltonian which is taken to be the one describing the interaction between the oscillator and the reservoir at temperature T. This Hamiltonian is given by

$$H = \omega a^+ a + \sum_j \omega_j a_j^+ a_j + \sum_j \{g_j(a+a^+)a_j^+ + \text{H.C.}\}, \tag{7.4}$$

where contrary to the case discussed in § 6, we have retained the terms of the form aa_j and $a^+ a_j^+$. We use the Hamiltonian (7.4) and make approximations similar to the ones made in § 6. We then find that $\Phi^{(\Omega)}(z, z^*; t)$, which is the reduced phase space distribution function, satisfies the master equation (AGARWAL [1971d])

$$\frac{\partial \Phi^{(\Omega)}}{\partial t} = \left\{ i\omega \frac{\partial}{\partial z} z + \gamma \frac{\partial}{\partial z}(z - z^*) - \tfrac{1}{2}\gamma(\langle n(\omega)\rangle + \lambda + \tfrac{1}{2}) \times \left(\frac{\partial}{\partial z} - \frac{\partial}{\partial z^*}\right)^2 + \text{C.C.} \right\} \Phi^{(\Omega)}. \tag{7.5}$$

The parameters γ and $\langle n(\omega)\rangle$ are defined by (6.9) and (6.6) respectively. We also recall that the parameter λ is $-\frac{1}{2}$ for the Sudarshan-Glauber distribution function, $\frac{1}{2}$ for the distribution function obtained with the normal rule of mapping and is zero for the Wigner distribution function. We make the transformation to the real variables q and p defined by

$$z = \left(\frac{m\omega}{2}\right)^{\frac{1}{2}} q + i\left(\frac{1}{2m\omega}\right)^{\frac{1}{2}} p, \qquad z^* = \left(\frac{m\omega}{2}\right)^{\frac{1}{2}} q - i\left(\frac{1}{2m\omega}\right)^{\frac{1}{2}} p. \tag{7.6}$$

Then the Fokker-Planck equation (7.5) transforms into the equation

$$\frac{\partial \Phi^{(\Omega)}}{\partial t} = \left\{ -\frac{\partial}{\partial q}(p/m) + \frac{\partial}{\partial p}(m\omega^2 q + 2\gamma p) + 2m\omega\gamma(\langle n(\omega)\rangle + \lambda + \tfrac{1}{2}) \frac{\partial^2}{\partial p^2} \right\} \Phi^{(\Omega)}. \tag{7.7}$$

The Fokker-Planck equation (7.7) is stochastically equivalent to the Langevin equations (cf. Appendix)

$$\dot{q} = p/m, \qquad \dot{p} = -2\gamma p - m\omega^2 q + F(t), \tag{7.8}$$

where $F(t)$ is a real Gaussian random process with the properties

$$\langle F(t)\rangle = 0, \qquad \langle F(t)F(t')\rangle = 2D\delta(t-t'). \tag{7.9}$$

Here the diffusion coefficient D is given by

$$D = 2m\omega\gamma(\langle n(\omega)\rangle + \lambda + \tfrac{1}{2}). \tag{7.10}$$

It should be noted that the Langevin equations (7.8) are of the same form as the equations (7.3) describing the Brownian motion of a classical oscillator. It should be borne in mind that in the classical case q and p are the position and momentum variables whereas in the quantum case q and p

are the c-number variables onto which the operators q and p are mapped by the Ω-rule of mapping. Since we are dealing with a quantum system certain care should be exercised in the proper interpretation of moments calculated from (7.8). It is convenient to transform to the complex variables z and z^*. We then obtain the following Langevin equations (which are stochastically equivalent to the Fokker-Planck equation (7.5))

$$\dot{z} = -i\omega z - \gamma(z - z^*) + \mathscr{F}(t), \qquad \dot{z}^* = i\omega z^* - \gamma(z^* - z) + \mathscr{F}^*(t), \tag{7.11}$$

where $\mathscr{F}(t)$ is a complex Gaussian random process with the properties

$$\begin{aligned}\langle \mathscr{F}(t)\rangle = \langle \mathscr{F}^*(t)\rangle = 0, \qquad & \langle \mathscr{F}(t)\mathscr{F}^*(t')\rangle = -\langle \mathscr{F}(t)\mathscr{F}(t')\rangle \\ = -\langle \mathscr{F}^*(t)\mathscr{F}^*(t')\rangle &= +2\gamma(\langle n(\omega)\rangle + \lambda + \tfrac{1}{2})\delta(t-t').\end{aligned} \tag{7.12}$$

As remarked earlier in § 6, the moments calculated from (7.11) correspond to the normally ordered moments $\langle a^{+m}a^n\rangle$ for $\lambda = -\frac{1}{2}$, to antinormally ordered moments $\langle a^n a^{+m}\rangle$ for $\lambda = +\frac{1}{2}$ and to Weyl ordered moments $\langle (a^{+m}a^n)_W\rangle$ for $\lambda = 0$. It should also be noted that the diffusion coefficient D, given by (7.10), in the high temperature limit goes over to the classical value, viz.

$$D_{cl} = 2\gamma m K_B T. \tag{7.13}$$

The Green's function $K^{(\Omega)}(z, z^*, t|z_0, z_0^*, 0)$ corresponding to the Fokker-Planck equation (7.5), which is in the form of a linearized Fokker-Planck equation, is easily seen to be given by

$$\begin{aligned}K^{(\Omega)}(z, z^*, t|z_0, z_0^*, 0) &= (\pi^2 \Delta_0)^{-\frac{1}{2}} \\ &\times \exp\{\Delta_0^{-1}[\mu(z^* - \langle z^*(t)\rangle)^2 + \mu^*(z - \langle z(t)\rangle)^2 - \tau|(z - \langle z(t)\rangle)|^2\},\end{aligned} \tag{7.14}$$

where

$$\langle z(t)\rangle = \left(\cos\omega_0 t - \frac{i\omega}{\omega_0}\sin\omega_0 t\right) z_0 e^{-\gamma t} + \frac{\gamma}{\omega_0}\sin\omega_0 t\, z_0^* e^{-\gamma t}, \tag{7.15}$$

$$\mu = -\left(\frac{\gamma}{\omega_0}\right)(\langle n(\omega)\rangle + \lambda + \tfrac{1}{2})\sin\omega_0 t\left(\cos\omega_0 t - \frac{i\omega}{\omega_0}\sin\omega_0 t\right) e^{-2\gamma t}, \tag{7.16}$$

$$\tau = (\langle n(\omega)\rangle + \lambda + \tfrac{1}{2})\left[1 - \left(1 + \frac{2\gamma^2}{\omega_0^2}\sin^2\omega_0 t\right) e^{-2\gamma t}\right], \tag{7.17}$$

and

$$\omega_0 = (\omega^2 - \gamma^2)^{\frac{1}{2}}, \qquad \Delta_0 = (\tau^2 - 4\mu\mu^*). \tag{7.18}$$

The Green's function (7.14) is in the form of complex Gaussian distribution which is centered at $z = \langle z(t)\rangle$. The steady state solution of (7.7) is given by

$$\Phi_{\rm st}^{(\Omega)} = \{\pi(\langle n(\omega)\rangle + \lambda + \tfrac{1}{2})\}^{-1} \exp\{-|z|^2(\langle n(\omega)\rangle + \lambda + \tfrac{1}{2})^{-1}\}, \quad (7.19)$$

which corresponds to an oscillator in thermal equilibrium at temperature T.

We conclude this section by giving the operator form of the master equation (7.5). This is easily obtained from (7.5) by the substitution $\lambda = -\frac{1}{2}$ and by using the identities (2.28). The result is

$$\frac{\partial \rho}{\partial t} = -i\omega[a^+ a, \rho] - \gamma(a^+ a\rho - 2a\rho a^+ + \rho a^+ a + a^2\rho - a\rho a - a^+\rho a^+ + \rho a^{+2}) - \gamma\langle n(\omega)\rangle(2[a^+, [a, \rho]] + [a^+, [a^+, \rho]] + [a, [a, \rho]]). \quad (7.20)$$

§ 8. Relaxation of an Atom

We next consider the relaxation of an atom with unperturbed energy eigenvalues E_l and eigenfunctions $|l\rangle$. We write the Hamiltonian of the atom interacting with the reservoir in the form

$$H = \sum_l E_l A_{ll} + H_{\rm R} + \sum_{kl} v_{kl} A_{kl}, \quad (8.1)$$

where the operators A_{kl} are given by

$$A_{kl} = |k\rangle\langle l|. \quad (8.2)$$

The v_{kl}'s in eq. (8.1) are the functions of the reservoir operators. We leave v_{kl} arbitrary. The Liouville operator in the interaction picture is given by the expression

$$\mathscr{L}_{\rm RS}(t) = \sum_{kl} v_{kl}(t) A_{kl} \exp(i\omega_{kl} t), \quad (8.3)$$

where

$$\omega_{kl} = (E_k - E_l), \qquad v_{kl}(t) = \exp\{iH_{\rm R} t\} v_{kl} \exp\{-iH_{\rm R} t\}. \quad (8.4)$$

We assume that the condition of the initial random phase and the condition (5.19) are satisfied. Then on combining (8.3) and (5.20), we find that the reduced density operator corresponding to the atomic system alone satisfies the master equation

$$\frac{\partial \rho}{\partial t} = \sum_{klmn} \{(A_{mn}\rho A_{kl} - A_{kn}\rho\delta_{lm})\gamma^+_{klmn} + (A_{kl}\rho A_{mn} - \rho A_{ml}\delta_{nk})\gamma^-_{mnkl}\} \exp\{i(\omega_{kl} + \omega_{mn})t\}, \quad (8.5)$$

where we have, for the sake of brevity, suppressed the subscript S from ρ. The density operator ρ, in (8.5), is in the interaction picture. γ^+_{klmn} and γ^-_{mnkl} are given by (LOUISELL [1969], LAX [1966c])

$$\gamma^{+}_{klmn} = \int_0^\infty \langle v_{kl}(t) v_{mn}(0)\rangle \exp\{-i\omega_{mn} t\} dt, \tag{8.6}$$

$$\gamma^{-}_{mnkl} = \int_0^\infty \langle v_{mn}(0) v_{kl}(t)\rangle \exp\{-i\omega_{mn} t\} dt \tag{8.7}$$

$$= (\gamma^{+}_{lknm})^*. \tag{8.8}$$

In eqs. (8.6) and (8.7) $\langle v_{kl}(t) v_{mn}(0)\rangle$ and $\langle v_{mn}(0) v_{kl}(t)\rangle$ are the reservoir correlation functions. In deriving (8.5) we also made the Markovian approximation which is justified for times $t \gg \tau_c$ where τ_c is the reservoir correlation time which is assumed to be very short. We also make the rotating wave approximation, i.e. we drop the rapidly oscillating terms in (8.5). Eq. (8.5) then reduces to

$$\frac{\partial \rho}{\partial t} = \sum_{klmn} \{(A_{mn}\rho A_{kl} - A_{kn}\rho\delta_{lm})\gamma^{+}_{klmn} + (A_{kl}\rho A_{mn} - \rho A_{ml}\delta_{nk})\gamma^{-}_{mnkl}\}, \tag{8.9}$$

where summation in (8.9) is over those values of k, l, m, n which satisfy the relation

$$\omega_{mn} + \omega_{kl} = 0, \qquad \text{i.e.} \quad E_m - E_n + E_k - E_l = 0. \tag{8.10}$$

We assume that the energy levels of the atom are non-degenerate and are unevenly spaced (the case of evenly spaced energy levels is treated in § 9). Then (8.10) will be satisfied in the following three cases

$$\text{(i)}\ k = n, m = l; \quad \text{(ii)}\ k = l, m = n; \quad \text{(iii)}\ k = l = m = n. \tag{8.11}$$

On grouping the terms in eq. (8.9) according to (8.11), we find that (8.9) reduces to

$$\frac{\partial \rho}{\partial t} = \sum_{k \neq l} \gamma_{lk} A_{lk}\rho A_{kl} + \sum_{kl} (\gamma^{+}_{llkk} + \gamma^{-}_{llkk}) A_{kk}\rho A_{ll} - \sum_{kl} \{\gamma^{+}_{kllk} A_{kk}\rho + \gamma^{-}_{kllk}\rho A_{kk}\}, \tag{8.12}$$

where we have set

$$\gamma_{lk} = \gamma^{+}_{kllk} + \gamma^{-}_{kllk} \tag{8.13}$$

$$= \int_{-\infty}^{+\infty} \langle v_{kl}(t) v_{lk}(0)\rangle \exp\{-i\omega_{lk} t\} dt, \tag{8.14}$$

where the last relation follows on combining eqs. (8.6), (8.7) and (8.13). Eq. (8.12) is the desired master equation for the reduced density operator ρ (in the interaction picture) describing the relaxation of the atom. On taking the matrix elements of both sides of (8.12), we obtain the equation

$$\frac{\partial \rho_{ij}}{\partial t} = \delta_{ij} \sum_{k \neq i} \gamma_{ik} \rho_{kk} - \Gamma^{c}_{ji} \rho_{ij}, \tag{8.15}$$

where Γ^{c}_{ij} is defined by

$$\Gamma^{c}_{ij} = -(\gamma^{+}_{iijj} + \gamma^{-}_{iijj}) + \sum_{l} (\gamma^{+}_{jlli} + \gamma^{-}_{iill}). \tag{8.16}$$

It is easily seen from (8.14) that γ_{lk} is the transition probability per unit time that the atom makes a transition from the state $|k\rangle$ to the state $|l\rangle$. For a reservoir in thermal equilibrium at temperature T, one can further show that

$$\gamma_{kl} = \gamma_{lk} \exp(\beta \omega_{lk}). \tag{8.17}$$

Let us denote by Γ_{ij} the real part of Γ^{c}_{ij}. Then it is easy to show (LOUISELL [1969]) that Γ_{ii} is the transition probability that the reservoir causes the atom to make the transition from a state $|i\rangle$ to all other states, i.e.

$$\Gamma_{ii} = \sum_{k \neq i} \gamma_{ki}. \tag{8.18}$$

Moreover from the definition (8.16), the following relation is easily established

$$\Gamma_{ij} = \tfrac{1}{2}(\Gamma_{ii} + \Gamma_{jj}) + \Gamma^{ph}_{ij}, \tag{8.19}$$

where Γ^{ph}_{ij} is given by (LOUISELL [1969])

$$\Gamma^{ph}_{ij} = \pi \sum_{E_R} |\langle E_R | v_{ii} - v_{jj} | E_R \rangle|^2 \rho_R(E_R), \tag{8.20}$$

and where E_R and $|E_R\rangle$ are the energy eigenvalues and the energy eigenfunctions of H_R and the $\rho_R(0)$ is assumed to be a function of H_R only. It should be noted that Γ^{ph}_{ij} appears only in the equations of motion for the off-diagonal elements and this term describes the damping by virtual processes. On putting $i = j$ in (8.15), we obtain the Pauli type equation

$$\partial \rho_{ii} / \partial t = \sum_{k \neq i} (\gamma_{ik} \rho_{kk} - \gamma_{ki} \rho_{ii}), \tag{8.21}$$

where use has been made of the relation (8.18).

So far our discussion for the relaxation of an atom has been quite general. We now specialize to the case of a two-level atom. Let $|1\rangle$ and $|2\rangle$ be the ground and the excited states of the atom respectively. The operators A_{kl} are then related to the spin angular momentum operators for spin $\frac{1}{2}$-value (cf. FEYNMAN, VERNON and HELLWARTH [1957]) by the equation

$$A_{21} \equiv |2\rangle\langle 1| = S^{+}, \quad A_{12} \equiv |1\rangle\langle 2| = S^{-}, \quad \tfrac{1}{2}(A_{22} - A_{11}) = S^{z}. \tag{8.22}$$

The relaxation equation (8.12) in case of a two-level atomic system can be written in the form (see also RISKEN [1970])

$$\frac{\partial \rho}{\partial t} = \tfrac{1}{2}\gamma_{12}(2S^-\rho S^+ - \rho S^+ S^- - S^+ S^- \rho) + \tfrac{1}{2}\gamma_{21}(2S^+\rho S^- - \rho S^- S^+ - S^- S^+ \rho) + \Gamma_{12}^{\mathrm{ph}}(2S^z\rho S^z - \rho S^z S^z - S^z S^z \rho). \tag{8.23}$$

Equation (8.23) leads to the following equations for the macroscopic mean values

$$\frac{\partial}{\partial t}\langle S^+\rangle = \mathrm{i}\omega\langle S^+\rangle - \langle S^+\rangle/T_2, \tag{8.24}$$

$$\frac{\partial}{\partial t}\langle S^z\rangle = -(\langle S^z\rangle - \langle S^z\rangle_{\mathrm{st}})/T_1, \tag{8.25}$$

where

$$\omega = E_2 - E_1, \tag{8.26}$$

and T_1 and T_2 are the longitudinal and transverse relaxation times defined by

$$T_1^{-1} = (\gamma_{12}+\gamma_{21}), \qquad T_2^{-1} = (\tfrac{1}{2}\gamma_{12}+\tfrac{1}{2}\gamma_{21}+\Gamma_{12}^{\mathrm{ph}}). \tag{8.27}$$

In eq. (8.25) $\langle S^z\rangle_{\mathrm{st}}$ is the steady state value of $\langle S^z\rangle$ and is given by

$$\langle S^z\rangle_{\mathrm{st}} = -\tfrac{1}{2}(\gamma_{12}-\gamma_{21})/(\gamma_{12}+\gamma_{21}), \tag{8.28}$$

which in case of thermal reservoir reduces to

$$\langle S^z\rangle_{\mathrm{st}} = -\tfrac{1}{2}\tanh(\tfrac{1}{2}\beta\omega), \tag{8.29}$$

where use of (8.17) has also been made. Equations (8.24) and (8.25) are recognized to be the familiar Bloch equations which predict the correct steady state behavior, i.e. relaxation at a finite temperature T rather than at infinite temperature which was found in the semiclassical treatment given in § 4.

Finally we mention that if the reservoir is made up of harmonic oscillators and is initially in a state of thermal equilibrium, then one finds the following expression for the transition probabilities (AGARWAL [1969], LOUISELL [1969])

$$\gamma_{12} = 2\pi h(\omega)|g(\omega)|^2(\langle n(\omega)\rangle + 1), \tag{8.30}$$

$$\gamma_{21} = 2\pi h(\omega)|g(\omega)|^2\langle n(\omega)\rangle, \tag{8.31}$$

and $\Gamma_{12}^{\mathrm{ph}}$ is zero since the atom is assumed to have no permanent dipole moment. $\langle n(\omega)\rangle$ is given by eq. (6.6) and $g(\omega)$, $h(\omega)$ have the same meaning as in § 6.

We have considered here the relaxation of an atom whose energy levels are unevenly spaced and are non-degenerate*. For the case of a two-level atom, the problem was equivalent to the relaxation of a spin $-\frac{1}{2}$ system. In the next section we consider the problem of relaxation of N two-level atoms. The problem of relaxation of general spin systems is treated in papers of BLOCH [1956, 1957], REDFIELD [1957, 1965] and in the book of ABRAGAM [1961].

§ 9. Incoherent and Coherent (Superradiance) Spontaneous Emission

In this section we consider the spontaneous emission from a collection of N identical two-level atoms, using the master equation techniques. This system was originally studied by DICKE [1954], who found that under certain conditions the radiation rate is proportional to the square of the number of atoms. This coherent emission of radiation is known as superradiance and has recently been the subject of many investigations (DILLARD and ROBL [1969], LEHMBERG [1970a], DIALETIS [1970], AGARWAL [1970, 1971b, c, e], REHLER and EBERLY [1971], BONIFACIO, SCHWENDIMAN and HAAKE [1971a, b], WILLIS and PICARD [1973]).

The Hamiltonian for a collection of N identical two-level atoms interacting with a quantized radiation field can be shown to be given by (see e.g. HAKEN [1970])

$$H = \omega \sum_j S_j^z + \sum_{ks} \omega_{ks} a_{ks}^+ a_{ks} + \sum_{ks} \sum_j (S_j^+ a_{ks} g_{jks} + \text{H.C.}). \tag{9.1}$$

In deriving (9.1), the nonresonant terms have been ignored, i.e. we have made the rotating wave approximation. In eq. (9.1) a_{ks} and a_{ks}^+ are the annihilation and the creation operators associated with the ks mode of the radiation field and $S_j^\pm$, S_j^z are the components of the spin angular momentum operator (corresponding to spin $-\frac{1}{2}$ value) associated with the jth atom (cf. eq. (8.22); FEYNMAN, VERNON and HELLWARTH [1957]). g_{jks} is the coupling constant and is given by

$$g_{jks} = (-i\omega/c)(2\pi c/L^3)^{\frac{1}{2}} k^{-\frac{1}{2}} (\boldsymbol{\varepsilon}_{ks} \cdot \boldsymbol{d}) \exp(i\boldsymbol{k} \cdot \boldsymbol{R}_j), \tag{9.2}$$

where $\boldsymbol{R}_j$ is the position vector of the jth atom and $\boldsymbol{d}$ is the dipole moment matrix element. All other symbols have the usual meaning.

Since we are studying spontaneous emission, the initial state of the system is given by

$$\rho_{S+R}(0) = \rho_S(0) |\{0\}\rangle\langle\{0\}|, \tag{9.3}$$

* The equation (8.15) at zero temperature and its generalizations describe spontaneous emission from multi-level atoms, AGARWAL [1973a].

where $|\{0\}\rangle$ is the vacuum state of the radiation field and $\rho_S(0)$ is the initial state of the atomic system which we leave arbitrary. It is clear that the radiation field in the vacuum state interacting with atoms behaves as a reservoir at zero temperature interacting with the atomic system. The problem of spontaneous emission thus reduces to that of relaxation at zero temperature. We can now use the master equation techniques of § 5 to obtain the equation of motion for the reduced density operator corresponding to the atomic system alone. It is found that, in the Born approximation and in the Markovian approximation, the reduced density operator satisfies the master equation (AGARWAL [1970])

$$\frac{\partial \rho}{\partial t} = -\mathrm{i}\sum_{ij} \Delta_{ij}[S_i^+ S_j^-, \rho] - \sum_{ij} \gamma_{ij}\{S_i^+ S_j^- \rho - 2S_j^- \rho S_i^+ + \rho S_i^+ S_j^-\}, \quad (9.4)$$

where*

$$\gamma_{ij} = \left(\frac{2\pi^2\omega^2}{c}\right)\left(\frac{1}{2\pi}\right)^3 \int \mathrm{d}^3k \exp\{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{R}_i - \boldsymbol{R}_j)\}\left(|\boldsymbol{d}|^2 - \frac{|\boldsymbol{d}\cdot\boldsymbol{k}|^2}{|\boldsymbol{k}|^2}\right)\frac{\delta(kc-\omega)}{k}, \quad (9.5)$$

$$\Delta_{ij} = \left(\frac{2\pi\omega^2}{c}\right)\left(\frac{1}{2\pi}\right)^3 \times \mathrm{P}\int \mathrm{d}^3k \exp\{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{R}_i - \boldsymbol{R}_j)\}\left(|\boldsymbol{d}|^2 - \frac{|\boldsymbol{d}\cdot\boldsymbol{k}|^2}{k^2}\right) k^{-1}(\omega - kc)^{-1}, \quad (9.6)$$

and where P denotes the Cauchy principal part. In (9.4) ρ refers to the reduced density operator in the interaction picture and we have dropped the subscript S on ρ_S. The master equation (9.4) is valid both for small and for large systems. By a small system we mean a system whose linear dimensions are small compared to a wavelength. In what follows we consider mainly small systems. For such systems $\gamma_{ij} \to \gamma$, independent of the indices i and j. On evaluation γ is found to be

$$\gamma = (2\omega^3|\boldsymbol{d}|^2)/(3c^3), \quad (9.7)$$

and the coefficient Δ_{ii} is related to the frequency shifts; its value may be obtained by renormalization. The various results on coherent and incoherent spontaneous emission follow from the master equation (9.4). In what follows, we will also ignore the effect of the frequency shift terms involving Δ_{ij}.

For one two-level atom, eq. (9.4) is easily solved and one finds that

$$\langle S^+(t)\rangle = \langle S^+(0)\rangle \mathrm{e}^{\mathrm{i}\omega t - \gamma t}, \qquad \langle S^z(t)\rangle = -\tfrac{1}{2} + \langle S^z(0) + \tfrac{1}{2}\rangle \mathrm{e}^{-2\gamma t}. \quad (9.8)$$

* The correct value of Δ_{ij} is obtained by making rotating wave approximation on the master equation rather than the Hamiltonian itself (AGARWAL [1973a]).

It should be noted that the results (9.8) correspond to the well known 'exponential decay' (WEISSKOPF and WIGNER [1930]). 2γ is equal to the inverse life time of a single atom. The density operator ρ is given by

$$\rho = \tfrac{1}{2} + 2\langle S^z\rangle S^z + \langle S^+\rangle S^- + \langle S^-\rangle S^+. \tag{9.9}$$

It is easily deduced from (9.8) and (9.9) that $\rho^2 \neq \rho$, i.e. because of spontaneous emission the atom is left in a mixed state unless $t = \infty$.

For the case of two atoms the master equation (9.4) can be easily solved. Here we only give the time dependence of the total energy $W(t) = \sum_{i=1}^{2}\langle S_i^z\rangle$ (in units of ω)

$$W(t) = (W(0) + \tfrac{3}{4} + \langle \boldsymbol{S}_1 \cdot \boldsymbol{S}_2\rangle_0 + 4\gamma t\langle \boldsymbol{S}_1^+ S_1^- S_2^+ S_2^-\rangle_0)\mathrm{e}^{-4\gamma t} - (\tfrac{3}{4} + \langle \boldsymbol{S}_1 \cdot \boldsymbol{S}_2\rangle_0), \tag{9.10}$$

where $\langle\ \rangle_0$ refers to the mean value at time $t = 0$. It is seen from (9.10) that

$$W(t) \xrightarrow{t\to\infty} -(\tfrac{3}{4} + \langle \boldsymbol{S}_1 \cdot \boldsymbol{S}_2\rangle_0), \tag{9.11}$$

which shows that the atoms are left in a state which is determined from the initial state. If the atoms are initially excited to a permutationally symmetric state ($\langle \boldsymbol{S}_1 \cdot \boldsymbol{S}_2\rangle_0 = \tfrac{1}{4}$), then only each atom will be found in its ground state.

We now consider the general problem of N two-level atoms. It is easily seen from (9.4) that the mean energy $\langle S_i^z\rangle$ of the ith atom obeys the equation

$$\frac{\partial}{\partial t}\langle S_i^z\rangle + \gamma \sum_j (\langle S_i^+ S_j^-\rangle + \text{C.C.}) = 0. \tag{9.12}$$

The radiation rate, which we denote by $I(t)$, is defined as the time rate change of the total energy of the system. Evidently

$$I(t) = -\omega \sum_j \frac{\partial}{\partial t}\langle S_j^z\rangle = 2\gamma\omega \sum_{ij}\langle S_i^+ S_j^-\rangle, \tag{9.13}$$

where the correlation $\langle S_i^+ S_j^-\rangle$ is to be obtained from the solution of the master equation (9.4). A perturbation theoretic result is obtained by replacing $\rho(t)$ in (9.13) by $\rho(0)$ and is given by

$$I_0 = 2\gamma\omega \sum_{ij} \mathrm{tr}\,(\rho(0) S_i^+ S_j^-). \tag{9.14}$$

We now assume that the system was initially excited to a state of the form

$$\rho(0) = \prod |\theta_0, \varphi_0\rangle_{i\,i}\langle\theta_0, \varphi_0|, \tag{9.15}$$

where

$$|\theta_0, \varphi_0\rangle_i = \cos(\tfrac{1}{2}\theta_0)e^{i\varphi_0/2}|-\rangle_i + \sin(\tfrac{1}{2}\theta_0)e^{-i\varphi_0/2}|+\rangle_i, \qquad (9.16)$$

and where $|+\rangle_i$ and $|-\rangle_i$ are the excited and ground states of the ith atom. The state (9.15) is characterized by two parameters θ_0 and φ_0 and can be obtained by exciting the atomic system by an external field (see e.g. DIALETIS [1970]). On substituting (9.15) and (9.16) in (9.14), we find that I_0 is given by

$$I_0 = 2\gamma\omega N \sin^2(\tfrac{1}{2}\theta_0)\{1+(N-1)\cos^2(\tfrac{1}{2}\theta_0)\}, \qquad (9.17)$$

which in the special cases leads to

$$I_0 \propto \begin{cases} N & \text{if } \theta_0 = \pi, \\ N^2 & \text{if } \theta_0 = \tfrac{1}{2}\pi. \end{cases} \qquad (9.18)$$

A system excited to a state of the form (9.16) with $\theta_0 = \frac{1}{2}\pi$ gives, therefore, rise to superradiant emission (DICKE [1954]) even though there are no correlations among different atoms. One should note that in this case the dipole moment is maximum. The author has referred to this type of superradiance as the superradiance of first kind (AGARWAL [1971c]).

To study the collective behavior of the system, we introduce, following DICKE [1954], the operators $S^{\pm}$, S^z defined by

$$S^{\pm} = \sum_i S_i^{\pm}, \qquad S^z = \sum_i S_i^z. \qquad (9.19)$$

In terms of the collective variables, the master equation (9.4) becomes

$$\partial\rho/\partial t = -\gamma(S^+S^-\rho - 2S^-\rho S^+ + \rho S^+S^-). \qquad (9.20)$$

Starting from a different standpoint the master equation (9.20) has also been obtained by BONIFACIO, SCHWENDIMANN and HAAKE [1971a]. The perturbation theoretic result (9.14) in terms of the collective variables may be written as

$$I_0 = 2\gamma\omega \text{ tr } \{\rho(0)S^+S^-\}. \qquad (9.21)$$

If we assume that the system was initially excited to the state $|\frac{1}{2}N, m\rangle$, (Dicke state, which is the simultaneous eigenstate of the operators S^2 and S^z with eigenvalues $\frac{1}{2}N(\frac{1}{2}N+1)$ and m respectively) then (9.21) reduces to

$$I_0 = 2\gamma\omega\{\tfrac{1}{2}N(\tfrac{1}{2}N+1) - m^2 + m\}. \qquad (9.22)$$

It is seen from (9.22) that for large N

$$I_0 \propto \begin{cases} N & \text{if } m = \tfrac{1}{2}N, \\ N^2 & \text{if } m = 0. \end{cases} \qquad (9.23)$$

We conclude from (9.23) that a system excited to a state $|\frac{1}{2}N, 0\rangle$ leads

to the superradiant emission (DICKE [1954]). We refer to this type of superradiance as the superradiance of second kind (AGARWAL [1971c]). One may further show that $|\frac{1}{2}N, 0\rangle$ is a state in which the dipole moment is zero and in which there are correlations among different atoms, one has, for example,

$$\langle S_i^z S_j^z\rangle - \langle S_i^z\rangle\langle S_j^z\rangle = -\{4(N-1)\}^{-1}. \tag{9.24}$$

It is, therefore, clear that the atomic correlations play an important role in the superradiant emission of the second kind. The radiation from the state $|\frac{1}{2}N, \frac{1}{2}N\rangle$ is incoherent to start with but, as we will see, it becomes superradiant as the system develops in time. This is another example of the superradiant emission of the second kind, since the dipole moment of the system is found to be zero (cf. eq. (9.31)).

To obtain the time dependence of the radiation rate one must solve either the master equation (9.4) or (9.20). For small systems S^2 is a constant of motion. On taking the matrix elements of both sides of (9.20), we obtain the equation

$$\partial\rho_{mn}/\partial t = 2\gamma\{(v_{m+1}\, v_{n+1})^{\frac{1}{2}}\rho_{m+1,n+1} - \tfrac{1}{2}(v_m + v_n)\rho_{mn}\}, \tag{9.25}$$

where

$$\rho_{mn} \equiv \langle S, m|\rho|S, n\rangle, \tag{9.26}$$

and $|S, m\rangle$ is the eigenstate of the operators S^2 and S^z, with eigenvalues $S(S+1)$ and m respectively. In the theory of superradiance S is referred to as 'cooperation number'. In eq. (9.25) $2\gamma v_m$ is the transition probability that the atomic system makes a transition from the state $|m\rangle$ to the state $|m-1\rangle$ and is given by

$$v_m = (S-m+1)(S+m). \tag{9.27}$$

For the diagonal elements eq. (9.25) is an equation of the Pauli type. The exact solution of (9.25) is given by (AGARWAL [1970])

$$\rho_{m,n}(\beta) = \sum_{l\geqq 0}\{\prod_{k=1}^{l}(v_{m+k}\, v_{n+k})^{\frac{1}{2}}\}\{\prod_{k=0}^{l}[\beta + \tfrac{1}{2}(v_{m+k} + v_{n+k})]^{-1}\}\rho_{m+l,n+l}(0), \tag{9.28}$$

where $\rho_{m,n}(\beta)$ is the Laplace transform of $\rho_{m,n}(t)$ defined by a relation of the form (3.4). It is seen from (9.28) that the steady state solution is given by

$$\rho_{m,n}(\infty) = \delta_{m,n}\delta_{m,-S}, \tag{9.29}$$

where S is the cooperation number. Thus the steady state value of the energy is

$$W(\infty) = -S. \tag{9.30}$$

A special case of (9.30) is given by eq. (9.11). It should be noted that the cooperation number takes only the integer or half integer values.

From (9.28) one may deduce that

$$\langle S^{\pm}(t)\rangle = 0 \qquad \text{if} \quad \rho(0) = |S, m\rangle\langle S, m|, \tag{9.31}$$

i.e., a system which starts in a state with zero dipole moment remains in a state with zero dipole moment and such a system will, therefore, not show superradiance of the first kind.

The analytic solution (9.28) is too involved for practial calculations, unless N is a small number. For large values of N, we may resort to approximate methods. Before discussing the approximate methods, we will discuss another exactly soluble model. We will consider what happens if each two-level atom is replaced by a harmonic oscillator. The superradiant emission from a system of harmonic oscillators has been studied in detail by AGARWAL [1970, 1971b].

For studying the oscillator system it is convenient to work with the phase space distribution functions. The spontaneous emission from a collection of identical harmonic oscillators is described by the following master equation for Sudarshan-Glauber distribution function

$$\frac{\partial \Phi^{(A)}}{\partial t} = \gamma \sum_{ij} \left\{ \frac{\partial}{\partial z_i} (z_j \Phi^{(A)}) + \text{C.C.} \right\}. \tag{9.32}$$

The Green's function corresponding to eq. (9.32) is given by

$$K^{(A)}(\{z_i\}, \{z_i^*\}, t | \{z_i^0\}, \{z_i^{0*}\}, 0) = \prod_{i=1}^{N} \delta^{(2)}(z_i - \bar{z}_i), \tag{9.33}$$

where

$$\bar{z}_i = z_i^0 - \frac{1}{N}(1 - e^{-\gamma t N}) \sum_{j=1}^{N} z_j^0. \tag{9.34}$$

We assume that each of the oscillators was initially excited to a coherent state $|z_i^0\rangle$. Then the radiation rate from such a system is given by

$$I(t) = 2\gamma\omega e^{-2\gamma t N} |\sum_{i=1}^{N} z_i^0|^2, \tag{9.35}$$

and is proportional to N^2 if $z_i^0 = z_0$ and the decay constant is N times larger than that due to a single oscillator. The coherent state $|\{z_0\}\rangle$ is thus a superradiant state for the oscillator system and this is an example of the superradiance of the first kind. The enhanced decay rate for the case of two classical oscillators has been observed in a recent experiment by LAMA, JODOIN and MANDEL [1972]. The steady state solution is given by

$$\Phi^{(A)}(\{z_i\}, \{z_i^*\}, \infty) = \prod_{i=1}^{N} \delta^{(2)}\left[z_i - \left(z_i^0 - \frac{1}{N}\sum_{j=1}^{N} z_j^0\right)\right], \qquad (9.36)$$

which shows that, in the steady state, each oscillator is not necessarily left in its ground state.

It should also be noted that for the initial coherent state excitation, no correlations are induced among different oscillators. However, if the system is initially excited to an incoherent state, such as a Fock state, then correlations are induced among different oscillators because of spontaneous emission (AGARWAL [1971b]).

We now discuss approximate expressions for the radiation rate from a collection of two-level atoms. We have seen that the radiation rate can be calculated if the time dependence of the correlation function $\langle S_i^+ S_j^- \rangle$ is known. The equation of motion for $\langle S_i^+ S_j^- \rangle$ is coupled to the higher order correlation function of the form $\langle S_i^+ S_j^- S_k^z \rangle$ (see AGARWAL [1970]). So in order to obtain the radiation rate one has to solve the whole hierarchy of equations. The approximate result is obtained by closing the hierarchy of equations, making suitable approximations on the higher order correlation functions. Various approximate procedures are discussed in AGARWAL [1971c, e]. From the permutation symmetry of the problem it follows that $\langle \boldsymbol{S}_i \cdot \boldsymbol{S}_j \rangle = \frac{1}{4}$, if the system is initially excited to a permutationally symmetric state. This relation enables us to express $\langle S_i^+ S_j^- \rangle$ in terms of $\langle S_i^+ S_i^- S_j^+ S_j^- \rangle$, i.e.

$$(\langle S_i^+ S_j^- \rangle + \text{C.C.}) = 2\{\langle S_i^+ S_i^- \rangle - \langle S_i^+ S_i^- S_j^+ S_j^- \rangle\}, \qquad (i \neq j). \qquad (9.37)$$

On using (9.37), eq. (9.12) may be written as

$$\frac{\partial}{\partial t}\langle S_i^z \rangle + 2\gamma N \langle S_i^+ S_i^- \rangle - 2\gamma(N-1)\langle S_i^+ S_j^+ S_i^- S_j^- \rangle = 0, \qquad (i \neq j). \qquad (9.38)$$

Our approximate procedure consists of expressing $\langle S_i^+ S_i^- S_j^+ S_j^- \rangle$ in terms of the one-particle mean values. The nature of the approximation depends on the initial excitation.

We assume that the atomic system was initially excited to a state of the form (9.15) with $\theta_0 < \pi$. We make an "Hartree type" of approximation on the two-particle mean value

$$\langle S_i^+ S_i^- S_j^+ S_j^- \rangle \approx \langle S_i^+ S_i^- \rangle \langle S_j^+ S_j^- \rangle, \qquad (i \neq j). \qquad (9.39)$$

On substituting (9.39) in (9.38), we obtain a simple equation for $\langle S_i^z \rangle$ which is easily solved. We then find the following expression for the radiation rate

$$I(t) = \frac{\omega\gamma N^3}{2(N-1)} \operatorname{sech}^2\{N\gamma(t-\tau)\}, \quad \tau = (2N\gamma)^{-1} \ln \frac{N-1}{N \cot^2(\frac{1}{2}\theta_0)+1}. \tag{9.40}$$

A similar result has been obtained by REHLER and EBERLY [1971]. The result (9.40) agrees well with the numerical solution of the master equation (9.20) (BONIFACIO, SCHWENDIMAN and HAAKE [1971b]). A possible experimental observation of the superradiant decay has been made by COMPAAN and ABELLA [1971]. The approximation (9.39) was analyzed in AGARWAL [1971e], where an improved result for the radiation rate was also given.

Superradiance of the second kind is much more complicated because one must take into account the atomic correlations. The two cases corresponding to the initial excitations $|\frac{1}{2}N, \frac{1}{2}N\rangle$ and $|\frac{1}{2}N, 0\rangle$ are discussed by AGARWAL [1971c, e] and by BONIFACIO, SCHWENDIMAN and HAAKE [1971b].

We conclude this section by making some remarks about the relaxation of an atom with equidistant energy levels. It is clear that (9.20) describes the relaxation at zero temperature of a spin (spin value S) or the relaxation of an atom with $(2S+1)$ equidistant energy levels. It can be easily shown that the relaxation, at finite temperature T, of an atom with $(2S+1)$ equidistant energy levels is described by the equation*

$$\begin{aligned}\frac{\partial\rho}{\partial t} = &-\gamma(1+\langle n(\omega)\rangle)\{S^+S^-\rho - 2S^-\rho S^+ + \rho S^+S^-\} \\ &-\gamma\langle n(\omega)\rangle\{S^-S^+\rho - 2S^+\rho S^- + \rho S^-S^+\},\end{aligned} \tag{9.41}$$

where $\langle n(\omega)\rangle$ is defined by (6.6). The master equation for the diagonal elements is given by

$$\begin{aligned}\frac{\partial\rho_{m,m}}{\partial t} = &\ 2\gamma(1+\langle n(\omega)\rangle)(v_{m+1}\rho_{m+1,m+1} - v_m\rho_{m,m}) \\ &-2\gamma\langle n(\omega)\rangle(v_{m+1}\rho_{m,m} - v_m\rho_{m-1,m-1}).\end{aligned} \tag{9.42}$$

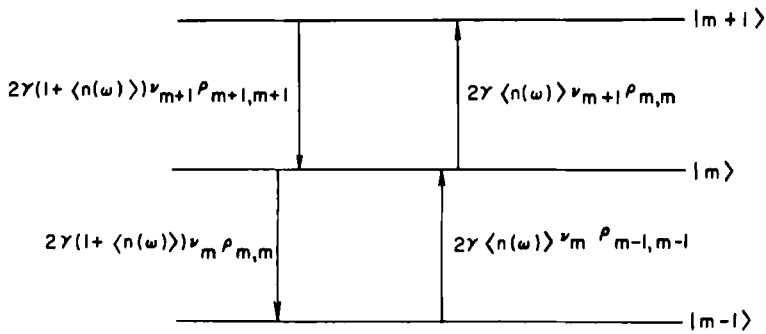


Fig. 2. Schematic representation of eq. (9.42).

* This master equation may be used to study spontaneous emission in presence of black body radiation (AGARWAL [1973a]).

This rate equation is schematically represented in Fig. 2 and has the following steady state solution

$$\rho_{m+1,m+1}(\infty) = \frac{\langle n(\omega)\rangle}{(1+\langle n(\omega)\rangle)}\rho_{m,m}(\infty) = \mathrm{e}^{-\beta\omega}\rho_{m,m}(\infty). \tag{9.43}$$

§ 10. Laser Master Equation

We use the techniques developed in previous sections to obtain the master equation for a single mode laser (for various theories of lasers see, for example, LAX [1966c, 1968a], LAX and LOUISELL [1969], LAMB [1964, 1965], LAMB and SCULLY [1967], HAKEN [1970], RISKEN [1970]). A laser system typically consists of active atoms, electromagnetic field and pump and loss mechanisms. We assume that each atom is a two-level system and that each atom and the field mode is coupled to its own reservoir. The reservoirs describe the effects of loss mechanism and the pump. We have already treated the effect of reservoirs on each atom and on an oscillator (which is the electromagnetic field mode in the present case) in §§ 6, 8 (see eqs. (6.21), (8.23)). The density operator ρ corresponding to the coupled atoms-field system satisfies the equation

$$\partial\rho/\partial t = -\mathrm{i}\mathscr{L}\rho, \tag{10.1}$$

where $\mathscr{L}$ is the effective Liouville operator and is non-Hermitian since we have already taken into account the effects of loss and pump mechanisms. $\mathscr{L}$, therefore, consists of two parts: a reversible part and an irreversible part. The reversible part of $\mathscr{L}$ may be written as

$$\mathscr{L}_{\mathrm{rev}} = \mathscr{L}_{\mathrm{A}} + \mathscr{L}_{\mathrm{F}} + \mathscr{L}_{\mathrm{AF}}, \tag{10.2}$$

where $\mathscr{L}_{\mathrm{A}} + \mathscr{L}_{\mathrm{F}}$ describe the unperturbed motion of the atoms and the field and $\mathscr{L}_{\mathrm{AF}}$ describes the interaction between the active atoms and the field. The corresponding Hamiltonians are given by

$$H_{\mathrm{A}} = \sum_{j=1}^{N}\omega_j S_j^z, \qquad H_{\mathrm{F}} = \omega a^+ a, \qquad H_{\mathrm{AF}} = \sum_{j=1}^{N} g_j(S_j a^+ + \mathrm{H.C.}). \tag{10.3}$$

The irreversible part of the Liouville operator is given by

$$\mathscr{L}_{\mathrm{irr}} = \mathrm{i}\Lambda_{\mathrm{F}} + \mathrm{i}\Lambda_{\mathrm{A}}, \tag{10.4}$$

where the operators Λ_{F} and Λ_{A} are given by (see eqs. (6.21), (8.23))

$$\Lambda_F G = \gamma\{[aG, a^+]+[a, Ga^+]+2\langle n(\omega)\rangle[[a, G], a^+]\}, \tag{10.5}$$

$$\Lambda_A G = \sum_{j=1}^{N}(\tfrac{1}{2}\gamma_{12}\{[S_j^-, GS_j^+]+[S_j^- G, S_j^+]\} + \tfrac{1}{2}\gamma_{21}\{[S_j^+, GS_j^-]+[S_j^+ G, S_j^-]\}+\Gamma_{12}^{ph}\{[S_j^z, GS_j^z]+[S_j^z G, S_j^z]\}). \tag{10.6}$$

Let $\gamma_{||}$ and $\gamma_\perp$ be the inverses of the longitudinal and transverse relaxation times T_1 and T_2 respectively, given by

$$\gamma_{||} = \gamma_{12}+\gamma_{21}, \qquad \gamma_\perp = \tfrac{1}{2}(\gamma_{12}+\gamma_{21}+2\Gamma_{12}^{ph}), \tag{10.7}$$

and let σ be the unsaturated inversion defined by

$$\sigma = (\gamma_{21}-\gamma_{12})/(\gamma_{12}+\gamma_{21}). \tag{10.8}$$

The master equation (10.1) contains both the atomic and the field variables. Many of the experiments on lasers concern the statistical properties of the laser light and it is, therefore, appropriate to obtain an equation of motion involving the field variables only. Let ρ_F be the reduced density operator corresponding to the field alone. It is defined in the usual way, viz.

$$\rho_F = \mathrm{Tr}_A\,\rho, \tag{10.9}$$

where Tr_A denotes the trace over the atomic variables. Various methods have been developed to eliminate the atomic variables from the original master equation (see LAX and LOUISELL [1967, 1969], ARZT et al. [1966], GORDON [1967], and HAAKE [1969b]). Here we discuss the method due to Haake who used the projection operator techniques to eliminate the atomic variables from the master equation (10.1).

It is easily seen that the derivation given in § 3 goes through even if $\mathscr{L}$ is non-Hermitian or non-real so that $\mathscr{P}\rho$ satisfies equations of the form given by (3.8), viz.

$$\frac{\partial}{\partial t}[\mathscr{P}\rho(t)]+\mathrm{i}\mathscr{P}\mathscr{L}[\mathscr{P}\rho(t)]+\mathrm{i}\mathscr{P}\mathscr{L}\exp[-\mathrm{i}(1-\mathscr{P})\mathscr{L}t](1-\mathscr{P})\rho(0) + \int_0^t \mathscr{P}\mathscr{L}\exp[-\mathrm{i}\tau(1-\mathscr{P})\mathscr{L}](1-\mathscr{P})\mathscr{L}\mathscr{P}\rho(t-\tau)\mathrm{d}\tau = 0, \tag{10.10}$$

where $\mathscr{L}$ is given by the sum of (10.2) and (10.4). For the purpose of eliminating the atomic variables, the projection operator is given by

$$\mathscr{P}\ldots = G_A\,\mathrm{Tr}_A\ldots, \tag{10.11}$$

where the operator G_A will be chosen such that

$$\mathscr{L}_A G_A = \Lambda_A G_A = 0; \qquad \mathrm{Tr}\,G_A = 1. \tag{10.12}$$

It is easily checked that a possible choice of G_A is

$$G_A = \prod_{i=1}^{N} (\tfrac{1}{2} + \sigma S_i^z), \tag{10.13}$$

where σ is the unsaturated inversion given by (10.8). The choice of the operator G_A of the form (10.13) is motivated by the fact that for a reservoir, in thermal equilibrium, and made of two-level atoms, the appropriate G_A is equal to $G_A = \prod_{i=1}^{N} (\frac{1}{2} - \tanh(\frac{1}{2}\beta\omega) S_i^z)$. The reduced density operator ρ_F is then given by

$$\rho_F = \mathrm{Tr}_A (\mathscr{P}\rho). \tag{10.14}$$

It is also easily verified that the projection operator $\mathscr{P}$ defined by (10.11) and (10.13) satisfies the identifies

$$\begin{aligned} &\mathscr{P}\mathscr{L}_F = \mathscr{L}_F \mathscr{P}, \qquad \mathscr{P}\Lambda_F = \Lambda_F \mathscr{P}, \qquad \mathscr{P}\mathscr{L}_A = \mathscr{P}\Lambda_A = 0, \\ &\mathscr{L}_A \mathscr{P} = \Lambda_A \mathscr{P} = 0, \qquad \mathscr{P}(\mathscr{L}_{AF})^n \mathscr{P} = 0, \end{aligned} \tag{10.15}$$

where n is an odd integer. On using (10.15), eq. (10.10) reduces to

$$\frac{\partial \rho_F}{\partial t} + \mathrm{i}(\mathscr{L}_F + \mathrm{i}\Lambda_F)\rho_F + I(t) + \int_0^t K(\tau)\rho_F(t-\tau)\mathrm{d}\tau = 0, \tag{10.16}$$

where the inhomogeneous term $I(t)$ and the kernel $K(t)$ are given by

$$I(t) = \mathrm{i}\,\mathrm{Tr}_A \mathscr{L}_{AF} \exp\{-\mathrm{i}(1-\mathscr{P})\mathscr{L}t\}(1-\mathscr{P})\rho(0), \tag{10.17}$$

$$K(t) = \mathrm{Tr}_A \mathscr{L}_{AF} \exp\{-\mathrm{i}(1-\mathscr{P})\mathscr{L}t\}\mathscr{L}_{AF} G_A. \tag{10.18}$$

The inhomogeneous term $I(t)$ in (10.16) depends on how the laser system has been initially turned on and is important only in the study of its transient behavior. In what follows we consider only the steady state properties and therefore we ignore the term $I(t)$. The kernel $K(t)$ can be simplified in the standard manner (cf. § 5) and may be written in the form

$$K(t) = \mathrm{Tr}_A \mathscr{L}_{AF} U_0(t)(1-\mathscr{P})U(t)\mathscr{L}_{AF} G_A, \tag{10.19}$$

where

$$U_0(t) = \exp\{-\mathrm{i}(\mathscr{L}_F + \mathrm{i}\Lambda_F + \mathscr{L}_A + \mathrm{i}\Lambda_A)t\}, \tag{10.20}$$

$$U(t) = T \exp\left\{-\mathrm{i}\int_0^t \mathrm{d}\tau U_0(-\tau)(1-\mathscr{P})\mathscr{L}_{AF}(1-\mathscr{P})U_0(\tau)\right\}. \tag{10.21}$$

On using the identities (10.15), the $K(t)$ may be rewritten as (cf. eq. (3.22))

$$K(t) = \mathrm{Tr}_A \mathscr{L}_{AF} \sum_{n=0}^{\infty} (-i)^{2n} \int_0^t d\tau_1 \int_0^{\tau_1} d\tau_2 \cdots \int_0^{\tau_{2n-1}} d\tau_{2n}$$
$$\times \{U_0(t-\tau_1)(1-\mathscr{P})\mathscr{L}_{AF} U_0(\tau_1-\tau_2)(1-\mathscr{P})\mathscr{L}_{AF} \cdots$$
$$\cdots \mathscr{L}_{AF} U_0(\tau_{2n})(1-\mathscr{P})\mathscr{L}_{AF} G_A\}. \tag{10.22}$$

Eq. (10.16) is still exact, i.e. it contains the effect of the coherent interaction H_{AF} to all orders. Further simplification can be made by terminating the series (10.22). It is well known in the theory of lasers that the Born approximation is inadequate (LAMB [1964]), since in the Born approximation one is taking into account the effect of the atoms on the field but not the effect of the field back on atoms. In order to describe the saturation effects in the atomic system, which are of great importance in the operation of lasers, it turns out sufficient to retain the first two terms in (10.22). As remarked by Haake, this approximation contains the hitherto known theories of the laser.

For the sake of simplicity, we consider only the case of homogeneous line at resonance i.e. $g_j = g$, $\omega_j = \omega$. We also assume that the field damping γ is much smaller than $\gamma_{||}$ and $\gamma_\perp$. This assumption is justified for most gas lasers. We apply the mapping operator $\Theta^{(A)}$ to both sides of (10.16) and use eqs. (10.3)–(10.6), and (10.22) (which is terminated at $n = 1$). We then find that the Sudarshan-Glauber distribution function satisfies the equation (HAAKE [1969b])

$$\frac{\partial \Phi^{(A)}}{\partial t} = i\omega \left(\frac{\partial}{\partial z} z - \frac{\partial}{\partial z^*} z^*\right) \Phi^{(A)} + \gamma \left\{\frac{\partial}{\partial z} z + \frac{\partial}{\partial z^*} z^* + 2\langle n(\omega)\rangle \frac{\partial^2}{\partial z\, \partial z^*}\right\} \Phi^{(A)}$$
$$+ \int_0^t k(\tau) \Phi^{(A)}(t-\tau) d\tau, \tag{10.23}$$

where the kernel $k(\tau)$ is given by

$$k(\tau) = \left(\frac{\partial}{\partial z} z + \frac{\partial}{\partial z^*} z^*\right) (-\alpha_l \psi_1(\tau) + \alpha_{nl} \psi_2(\tau) z^* z) + 4q\psi_1(\tau) \frac{\partial^2}{\partial z\, \partial z^*}. \tag{10.24}$$

In eq. (10.24) α_l, α_{nl} and q are the linear gain, nonlinear gain and fluctuation respectively, defined by

$$\alpha_l = \frac{Ng^2\sigma}{\gamma_\perp}, \qquad \alpha_{nl} = \frac{4Ng^4\sigma}{\gamma_\perp^2 \gamma_{||}}, \qquad g = \frac{Ng^2(1+\sigma)}{4\gamma_\perp}, \tag{10.25}$$

and $\psi_1(\tau)$ and $\psi_2(\tau)$ are the retardation functions defined by

$$\psi_1(\tau) = \gamma_\perp \exp(-\gamma_\perp \tau), \tag{10.26}$$

$$\psi_2(\tau) = \frac{\gamma_\perp^2 \gamma_{||}}{(\gamma_\perp - \gamma_{||})^2} [\exp(-\gamma_{||}\tau) - \{1 + \tau(\gamma_\perp - \gamma_{||})\} \exp(-\gamma_\perp \tau)]. \tag{10.27}$$

In writing (10.24) we have also ignored certain correction terms. These terms are of the order of $(\alpha_{nl}/q)^{\frac{1}{2}}$, which is quite small in the threshold region. These correction terms are of purely quantum mechanical origin and are discussed in detail in a recent review article by RISKEN [1970]. Since we ignore such quantum mechanical corrections, we will from now on regard the field variable as a c-number variable and the distribution $\Phi^{(A)}$ as a classical distribution, denoted by Φ. The master equation (10.23) is still an integro-differential equation. We now make the Markovian approximation (adiabatic approximation), which is made in most theories of the laser. Under this approximation (10.23) reduces to

$$\frac{\partial \Phi}{\partial t} = \frac{\partial}{\partial z}\{z(\gamma + i\omega - \alpha_l + \alpha_{nl} z^* z)\Phi\} + (\gamma\langle n(\omega)\rangle + 2q)\frac{\partial^2 \Phi}{\partial z\, \partial z^*} + \text{C.C.}. \tag{10.28}$$

We introduce the normalized variables, defined by

$$b = \left(\frac{\alpha_{nl}}{q}\right)^{\frac{1}{2}} z, \qquad \tau = (q\alpha_{nl})^{\frac{1}{2}} t, \qquad p = \frac{-\gamma + \alpha_l}{(q\alpha_{nl})^{\frac{1}{2}}}, \tag{10.29}$$

and we ignore the relatively small term $\langle n(\omega)\rangle\gamma$, eq. (10.28) then reduces to

$$\frac{\partial \Phi}{\partial t} = -\frac{\partial}{\partial b}[(p - b^* b)b\Phi] - \frac{\partial}{\partial b^*}[(p - b^* b)b^*\Phi] + 4\frac{\partial^2 \Phi}{\partial b\, \partial b^*}, \tag{10.30}$$

where Φ refers to the distribution function in the interaction picture. Eq. (10.30) is the well known Fokker-Planck equation describing the behavior of a single mode laser. This Fokker-Planck equation is very basic in the theory of the laser and has been studied in great detail (RISKEN [1970], LAX and LOUISELL [1967], RISKEN and VOLLMER [1967a, b], HEMPSTEAD and LAX [1967]). The Fokker-Planck equation (10.30) is stochastically equivalent to the Langevin equations

$$\dot{b} = (p - b^* b)b + F(t), \qquad \dot{b}^* = (p - b^* b)b^* + F^*(t), \tag{10.31}$$

where $F(t)$ is a delta correlated Gaussian random process with the properties

$$\langle F(t)\rangle = \langle F^*(t)\rangle = 0, \qquad \langle F(t)F(t')\rangle = \langle F^*(t)F^*(t')\rangle = 0,$$
$$\langle F(t)F^*(t')\rangle = 4\delta(t - t'). \tag{10.32}$$

It should be noted that the eq. (10.31), without fluctuations, is the same as that for a rotating wave Van der Pol oscillator (see e.g. DAVIS [1962]) and that is why this theory is also referred to as the Van der Pol oscillator model of laser light. For this model it is easily verified that the detailed balance

condition (A11) is satisfied and, therefore, the steady state solution is given by

$$\Phi_{st}(b, b^*) = \left(\frac{\mathcal{N}}{2\pi}\right) \exp\{-\tfrac{1}{4}|b|^4 + \tfrac{1}{2}|b|^2 p\}, \tag{10.33}$$

where $\mathcal{N}$ is the normalization constant. Thus the mean amplitude of the field is zero. This is contrary to the result obtained by ignoring the fluctuations (LAMB [1964]). The stationary distribution (10.33) is Gaussian centered at $|b|^2 = p$, in the variable $|b|^2$ and is truncated at $|b|^2 = 0$. From (10.33) it is easily shown that the cumulants $K_n(p)$ of the intensity distribution are given by (RISKEN [1970])

$$K_{n+1}(p) = 2\frac{\mathrm{d}}{\mathrm{d}p} K_n(p), \qquad K_1(p) = p + \frac{2}{\sqrt{\pi}} \mathrm{e}^{-\frac{1}{4}p^2}(1 + \mathrm{erf}\,(\tfrac{1}{2}p))^{-1}, \tag{10.34}$$

where erf (x) is the error function of x.

It does not seem possible to obtain an analytic expression for the Green's function associated with eq. (10.30). Numerical computations of the various statistical properties of the laser light, which follow from (10.30), have been carried out by RISKEN and VOLLMER [1967a, b] and by HEMPSTEAD and LAX [1967]. An excellent treatment of the statistical properties of the laser light is given in a recent review article by RISKEN [1970].

Recently the steady state distribution function (10.33) has attracted a great deal of attention in connection with the analogy between the laser threshold region and the second order phase transitions (GRAHAM and HAKEN [1970], SCULLY and DIGIORGIO [1970], GROSSMAN and RICHTER [1971]), for there is a close correspondence between (10.33) and the logarithm of Ginzburg-Landau energy functional (GINZBURG and LANDAU [1950]). We consider here the effect of a weak external field of the form $\varepsilon_0 \mathrm{e}^{-\mathrm{i}\nu t}$ and make the rotating wave approximation. The Langevin equations (10.31) are then modified to (g = coupling constant between the laser field and external field)

$$\dot{x}_1 = -(p - x_1^2 - x_2^2)x_1 + F_1 + \mathrm{Re}\, \mathrm{i}g\, \varepsilon_0 \mathrm{e}^{-\mathrm{i}\nu t + \mathrm{i}\omega t}, \tag{10.35}$$

$$\dot{x}_2 = -(p - x_1^2 - x_2^2)x_2 + F_2 + \mathrm{Im}\, \mathrm{i}g\, \varepsilon_0 \mathrm{e}^{-\mathrm{i}\nu t + \mathrm{i}\omega t}, \tag{10.36}$$

where we have introduced the real variables defined by $b = x_1 + \mathrm{i}x_2$, $F(t) = F_1(t) + \mathrm{i}F_2(t)$. We define the linear susceptibility $\chi_{ij}(\nu)$ by

$$\langle x_i \rangle = \mathrm{Re} \sum_j \chi_{ij}(\nu) \mathrm{e}^{\mathrm{i}\nu t} e_i, \tag{10.37}$$

where $e_1 = \mathrm{i}g\varepsilon_0$, $e_2 = -g\varepsilon_0$. Then it may be shown on using the properties of the Fokker-Planck equation (10.30) that (AGARWAL [1972])

$$\chi_{12}(\nu) = \chi_{21}(\nu) = 0, \tag{10.38}$$

$$\chi_{11}(\nu) = \chi_{22}(\nu) = \tfrac{1}{2}K_1(p) \sum_{m=0}^{\infty} V_m^g \lambda_{1m}(\lambda_{1m} + i\nu)^{-1}, \tag{10.39}$$

where V_m^g and λ_{1m} are the well known matrix elements and the eigenvalues respectively (RISKEN [1970]). From the behavior of V_m^g, λ_{1m} and K_1 as a function of the pump parameter p (see Figs. 2, 4 of RISKEN [1970]) it is found that $\chi_{11}(\nu)$ and $\chi_{22}(\nu)$ are continuous functions of the pump parameter. This result disagrees with a prediction of the theory of DIGIORGIO and SCULLY [1970], who found that the static susceptibility diverges both from above and from below as $|p|^{-1}$ at threshold. The reason for this disagreement is that Digiorgio and Scully have ignored the statistical fluctuations and it is the neglect of these fluctuations that leads to the singular behavior of the susceptibility.

§ 11. Master Equations for Strongly Interacting Quantum Systems in Contact with Heat Baths

In this section we obtain the master equations for strongly interacting quantum systems in contact with heat baths. An example of such systems is that of the relaxation of a spin system interacting with strong external fields. The theory for the relaxation of spin systems in presence of external fields has been developed by BLOCH [1957], REDFIELD [1957, 1965] and by others (see e.g. ARGYRES and KELLEY [1964]). This theory also enables to study spontaneous emission in the presence of strong external fields.

We write the total Hamiltonian of the system in the form

$$H = H_S + H_R + H_{RS} + H_{ext}(t), \tag{11.1}$$

where $H_{ext}(t)$ is the interaction Hamiltonian due to the external field acting on the system S. Let H_{01} and H_{02} be defined by

$$H_{01} = H_S + H_R, \qquad H_{02} = H_S + H_R + H_{ext}(t). \tag{11.2}$$

It is clear from (11.2) that H_{01} and H_{02} are the unperturbed parts of H with respect to the interaction Hamiltonians $H_{RS} + H_{ext}(t)$ and H_{RS} respectively. We introduce the operators $\sigma_1(t)$ and $\sigma_2(t)$ defined by

$$\sigma_i(t) = U_{0i}^{+}(t, 0)\rho(t)U_{0i}(t, 0), \qquad (i = 1, 2), \tag{11.3}$$

where the time evolution operators U_{0i} are given by

$$U_{0i}(t, \tau) = T \exp\left\{-i\int_{\tau}^{t} dt' H_{0i}(t')\right\}, \tag{11.4}$$

and where H_{0i} is given by (11.2). The unitary transformations, given by U_{01} and U_{02}, lead to two interaction pictures to which we will refer as the interaction picture [1] and the interaction picture [2] respectively. It should be noted that the interaction picture [2], as defined here, is known as the Furry picture in quantum electrodynamics (see e.g. SCHWEBER [1961; § 15g]).

It is easily seen that U_{01} and U_{02} are related by

$$U_{02}(t,\tau) = U_{01}(t,0)V(t,\tau)U_{01}^{+}(\tau,0), \tag{11.5}$$

where

$$V(t,\tau) = T\exp\left\{-\mathrm{i}\int_{\tau}^{t}\mathrm{d}t' H_{\mathrm{ext}\,1}(t')\right\}. \tag{11.6}$$

In eq. (11.6) $H_{\mathrm{ext}\,1}(t)$ is the operator $H_{\mathrm{ext}}(t)$ in the interaction picture [1] i.e.

$$H_{\mathrm{ext}\,1}(t) = U_{01}^{+}(t,0)H_{\mathrm{ext}}(t)U_{01}(t,0). \tag{11.7}$$

It should also be noted that $\sigma_1(t)$ and $\sigma_2(t)$ are related by

$$\sigma_2(t) = V^{+}(t,0)\sigma_1(t)V(t,0). \tag{11.8}$$

From the definition (11.3), it is clear that $\sigma_2(t)$ satisfies the equation

$$\partial\sigma_2/\partial t = -\mathrm{i}[H_{\mathrm{RS}2}(t),\sigma_2] \equiv -\mathrm{i}\mathscr{L}_{\mathrm{RS}2}(t)\sigma_2, \tag{11.9}$$

where

$$H_{\mathrm{RS}2}(t) = U_{02}^{+}(t,0)H_{\mathrm{RS}}U_{02}(t,0), \tag{11.10}$$

and where $\mathscr{L}_{\mathrm{RS}2}(t)$ is the Liouville operator in the interaction picture [2]. The Liouville equation (11.9) contains both the system variables and the reservoir variables. The master equation is obtained from (11.9) by the elimination of the reservoir coordinates. By applying the projection operator $\mathscr{P}$ to both sides of (11.9) and by using the procedure of § 5, we obtain the master equation (cf. eq. (5.16))

$$\frac{\partial}{\partial t}[\mathscr{P}\sigma_2(t)] + \mathrm{i}\mathscr{P}\mathscr{L}_{\mathrm{RS}2}(t)[\mathscr{P}\sigma_2(t)] + \int_0^t \mathrm{d}\tau\,\mathscr{P}\mathscr{L}_{\mathrm{RS}2}(t)\,\mathscr{U}(t,\tau)(1-\mathscr{P})\mathscr{L}_{\mathrm{RS}2}(\tau)[\mathscr{P}\sigma_2(\tau)] = 0, \tag{11.11}$$

where

$$\mathscr{U}(t,\tau) = T\exp\left\{-\mathrm{i}\int_{\tau}^{t}\mathrm{d}t'(1-\mathscr{P})\mathscr{L}_{\mathrm{RS}2}(t')(1-\mathscr{P})\right\}, \tag{11.12}$$

and where the projection operator $\mathscr{P}$ is given by (5.7). The master equation (11.11) contains the effect of the reservoir interaction to all orders. We again consider the approximate master equation obtained from (11.11) by making the Born approximation and the Markovian approximation (for some of

the non-Markovian effects in spin resonance saturation see ARGYRES and KELLEY [1964]). Under these approximations we recover the theory of Bloch and Redfield. The equation so obtained will be valid for $\tau_c \ll t \leqq \tau_r$ where τ_c and τ_r are the typical reservoir correlation time and the relaxation time respectively. The master equation (11.11) in these two approximations reduces to

$$\frac{\partial}{\partial t}[\mathscr{P}\sigma_2(t)] + \int_0^\infty \mathscr{P}\mathscr{L}_{RS2}(t)\mathscr{L}_{RS2}(t-\tau)[\mathscr{P}\sigma_2(t)]\mathrm{d}\tau = 0. \quad (11.13)$$

In deriving (11.13) we have also assumed that the properties of the reservoir are such that

$$\mathscr{P}\mathscr{L}_{RS2}(t)\mathscr{P}\sigma_2(t) = 0. \quad (11.14)$$

On using (11.8) and (11.13), we obtain the following master equation $\mathscr{P}\sigma_1(t)$

$$\frac{\partial}{\partial t}[\mathscr{P}\sigma_1(t)] + \mathrm{i}[H_{ext\,1}(t), \mathscr{P}\sigma_1] + \rho_R(0)\,\mathrm{Tr}_R \int_0^\infty [H_{RS1}(t), [V(t, t-\tau)H_{RS1}(t-\tau)V^+(t, t-\tau), \mathscr{P}\sigma_1(t)]]\mathrm{d}\tau = 0. \quad (11.15)$$

Eq. (11.15) is a basic equation for the description of strongly driven quantum systems. We now use eq. (11.15) to study spontaneous emission in presence of strong fields.

We assume that the atoms (two-level) are driven by a resonant optical field of the form

$$\varepsilon(t) = \varepsilon(\hat{x}\cos\omega t + \hat{y}\sin\omega t). \quad (11.16)$$

$H_{ext}(t)$ then is given by

$$H_{ext}(t) = -\tfrac{1}{2}|d|\varepsilon\sum_j (S_j^+ \mathrm{e}^{-\mathrm{i}\omega t} + \mathrm{H.C.}). \quad (11.17)$$

The Hamiltonian $H_S + H_R + H_{RS}$ is given by eq. (9.1). For the system under consideration we have

$$H_{ext\,1}(t) = -\tfrac{1}{2}|d|\varepsilon\sum_j (S_j^+ + \mathrm{H.C.}), \quad (11.18)$$

$$H_{RS1}(t) = \sum_{jks} (S_j^+ a_{ks} g_{ks} \exp\{-\mathrm{i}\omega_{ks} t + \mathrm{i}\omega t\} + \mathrm{H.C.}), \quad (11.19)$$

$$V(t, \tau) = \prod_{j=1}^{N} \exp\{\tfrac{1}{2}|d|\varepsilon\mathrm{i}(S_j^+ + S_j^-)(t-\tau)\}. \quad (11.20)$$

On using the properties of the spin angular momentum operators, it is easy to show that

$$V(t, t-\tau)H_{\mathrm{RSI}}(t-\tau)V^{+}(t, t-\tau) = \sum_{jks}[g_{ks}a_{ks}\exp\{-\mathrm{i}(t-\tau)\omega_{ks}\}$$
$$\times\{\cos^2(\tfrac{1}{2}\omega_1\tau)S_j^{+} + \sin^2(\tfrac{1}{2}\omega_1\tau)S_j^{-} - 2\mathrm{i}\sin(\tfrac{1}{2}\omega_1\tau)\cos(\tfrac{1}{2}\omega_1\tau)S_j^{z}\} + \mathrm{H.C.}], \tag{11.21}$$

where

$$\omega_1 = |d|\varepsilon. \tag{11.22}$$

On substituting (11.21) and (11.19) in (11.15), we find that the reduced density operator $\rho(t)$, for the atomic system, in the interaction picture [1] satisfies the master equation

$$\frac{\partial\rho}{\partial t} - \mathrm{i}\sum_j(\tfrac{1}{2}\omega_1)[S_j^{+} + S_j^{-}, \rho] + \sum_{ij}\{\tfrac{1}{4}(2\kappa+\kappa_+)[S_i^{+}, S_j^{-}\rho]$$
$$+\tfrac{1}{4}(2\kappa-\kappa_+)[S_i^{+}, S_j^{+}\rho] + \tfrac{1}{2}\kappa_-[S_i^{+}, S_j^{z}\rho] + \mathrm{H.C.}\} = 0, \tag{11.23}$$

where

$$\kappa = \gamma - \mathrm{i}\Omega, \qquad \kappa_{\pm} = \kappa(\omega+\omega_1) \pm \kappa(\omega-\omega_1), \tag{11.24}$$

and where (cf. eqs. (9.5), (9.6))

$$\gamma(\omega) = \sum_{ks}\pi|g_{ks}|^2\delta(\omega-\omega_{ks}), \qquad \Omega(\omega) = \sum_{ks}|g_{ks}|^2P(\omega_{ks}-\omega)^{-1}. \tag{11.25}$$

The master equation (11.23) should be compared with the one for weak external fields (cf. eqs. (5.27) and eq. (9.4)),

$$\frac{\partial\rho}{\partial t} - \mathrm{i}\sum_j(\tfrac{1}{2}\omega_1)[S_j^{+} + S_j^{-}, \rho] + \sum_{ij}\{\kappa[S_i^{+}, S_j^{-}\rho] + \mathrm{H.C.}\} = 0, \tag{11.26}$$

which can be obtained from (11.23) by the approximation $\kappa_+ \approx 2\kappa$, $\kappa_- \approx 0$. The master equation (11.23) leads to the following equation for the macroscopic mean value of an operator Q (in the rotating coordinate frame)

$$\frac{\partial}{\partial t}\langle Q\rangle + \mathrm{i}\sum_j(\tfrac{1}{2}\omega_1)\langle[S_j^{+} + S_j^{-}, Q]\rangle + \sum_{ij}\{-\tfrac{1}{4}(2\kappa+\kappa_+)\langle[S_i^{+}, Q]S_j^{-}\rangle$$
$$-\tfrac{1}{4}(2\kappa-\kappa_+)\langle[S_i^{+}, Q]S_j^{+}\rangle - \tfrac{1}{2}\kappa_-\langle[S_i^{+}, Q]S_j^{z}\rangle + \tfrac{1}{4}(2\kappa^{*}+\kappa_+^{*})\langle S_i^{+}[S_j^{-}, Q]\rangle$$
$$+\tfrac{1}{4}(2\kappa^{*}-\kappa_+^{*})\langle S_i^{-}[S_j^{-}, Q]\rangle + \tfrac{1}{2}\kappa_-^{*}\langle S_i^{z}[S_j^{-}, Q]\rangle\} = 0. \tag{11.27}$$

The master equation (11.23) or the mean value eq. (11.27) is too complicated to be solved for the case of N two-level atoms. We present its solution for the case of one two-level atom. For one atom we have

$$\frac{\partial}{\partial t}\langle S^-\rangle = -i\omega_1\langle S^z\rangle - \tfrac{1}{4}(2\kappa+\kappa_+)\langle S^-\rangle + \tfrac{1}{4}(2\kappa-\kappa_+)\langle S^+\rangle + \tfrac{1}{4}\gamma_-, \quad (11.28)$$

$$\frac{\partial}{\partial t}\langle S^z\rangle = \tfrac{1}{2}i\omega_1\langle S^+ - S^-\rangle + \tfrac{1}{4}\kappa_-\langle S^+\rangle + \tfrac{1}{4}\kappa_-^*\langle S^-\rangle - \tfrac{1}{2}(2\gamma+\gamma_+)\langle S^z+\tfrac{1}{2}\rangle. \quad (11.29)$$

Eqs. (11.28) and (11.29) have been recently obtained by LEHMBERG [1970b]. Similar equations have been derived by BLOCH [1957] and by others in connection with spin resonance saturation. If we ignore the frequency shifts, the steady state solution of eqs. (11.28) and (11.29) is given by

$$\mathrm{Re}\,\langle S^+\rangle = \frac{1}{2}\frac{\gamma_-}{\gamma_+} \approx \frac{\omega_1}{2}\frac{\mathrm{d}}{\mathrm{d}\omega}\ln\gamma(\omega); \quad \mathrm{Im}\,\langle S^+\rangle \approx -\left(\frac{\omega_1}{2\gamma}\right)\left\{1+2\left(\frac{\omega_1}{2\gamma}\right)^2\right\}^{-1};$$
$$\langle S^z\rangle \approx -\tfrac{1}{2}\{1+2(\omega_1/2\gamma)^2\}^{-1}, \quad (11.30)$$

where in obtaining (11.30) we also assumed that $\omega_1 \ll \omega$ so that $\kappa_+ \approx 2\kappa$, $\kappa_- \approx 2\omega_1\,\mathrm{d}\kappa/\mathrm{d}\omega$. Hence $\mathrm{Re}\,\langle S^+\rangle \approx (3\omega_1/2\omega)$ and thus the in-phase component continues to increase as the strength of the external field increases. The out of phase components saturates and then decreases. Lehmberg has noted that this effect can result in enhanced coherent scattering and fluorescence.

We now consider another example of strongly interacting quantum systems. We discuss the parametric frequency conversion in a medium with losses. In the absence of losses, the frequency conversion is described by the model Hamiltonian

$$H = \omega_a a^+ a + \omega_b b^+ b + g(ab^+ + \mathrm{H.C.}), \quad (11.31)$$

where a and b are the annihilation operators associated with two field modes. We have assumed that the pump field is very strong, so that it can be treated as a c-number. The losses in the medium can be taken into account by using the prescription given in § 6 (cf. eq. (6.53)), i.e. by assuming that each field mode is interacting with its own reservoir. From (11.31), (6.53) and (2.24), we obtain the following equation for the Sudarshan-Glauber distribution function $\Phi^{(A)}(\alpha, \alpha^*, \beta, \beta^*; t)$

$$\frac{\partial\Phi^{(A)}}{\partial t} = \left\{(i\omega_a+\gamma_a)\frac{\partial}{\partial\alpha}\alpha + (i\omega_b+\gamma_b)\frac{\partial}{\partial\beta}\beta + ig\frac{\partial}{\partial\beta}\alpha + ig\frac{\partial}{\partial\alpha}\beta + \gamma_a\langle n_a\rangle\frac{\partial^2}{\partial\alpha\,\partial\alpha^*}\right.$$
$$\left. + \gamma_b\langle n_b\rangle\frac{\partial^2}{\partial\beta\,\partial\beta^*} + \mathrm{C.C.}\right\}\Phi^{(A)}. \quad (11.32)$$

In order to keep the analysis as simple as possible, we also assume that

$\gamma_a = \gamma_b = \gamma$, $\langle n_a \rangle = \langle n_b \rangle = \langle n \rangle$, $\omega_a = \omega_b = \omega$. The Green's function associated with (11.32) (which is in the form of a linearized Fokker-Planck equation) is given by (WALLS [1970])

$$K^{(A)}(\alpha, \alpha^*, \beta, \beta^*, t|\alpha_0, \alpha_0^*, \beta_0, \beta_0^*, 0) = \{\pi\langle n\rangle(1-e^{-2\gamma t})\}^{-2} \times \exp\left\{\frac{-\frac{1}{2}(|(\alpha+\beta)-(\alpha_0+\beta_0)e^{-i\omega t-igt-\gamma t}|^2+|(\alpha-\beta)-(\alpha_0-\beta_0)e^{-i\omega t+igt-\gamma t}|^2)}{\langle n\rangle(1-e^{-2\gamma t})}\right\}. \tag{11.33}$$

The steady state solution is, therefore, given by

$$\Phi_{st}^{(A)} = (\pi\langle n\rangle)^{-2} \exp\{-(|\alpha|^2+|\beta|^2)/\langle n\rangle\}, \tag{11.34}$$

which implies that

$$\rho_{st} = \exp\left\{-\frac{(\omega a^+ a+\omega b^+ b)}{K_B T}\right\} \Big/ \operatorname{Tr} \exp\left\{-\frac{(\omega a^+ a+\omega b^+ b)}{K_B T}\right\}. \tag{11.35}$$

It is seen that (11.35) is not consistent with the Hamiltonian (11.31). This is due to the fact that the interaction between the two modes has not been treated properly when we assumed that each mode is interacting with its own reservoir. The effect of the coupling can be taken into account in the irreversible part of the master equation by using a procedure similar to that used in connection with interactions with strong external fields. We write the total Hamiltonian in the form

$$H = H_a + H_b + H_{ab} + H_{R_a} + H_{R_b} + H_{aR_a} + H_{bR_b}, \tag{11.36}$$

where H_{aR_a} and H_{bR_b} are the interaction Hamiltonian between the mode a and the reservoir R_a and the mode b and the reservoir R_b respectively. The reduced density operator, in the interaction picture, satisfies the master equation

$$\frac{\partial\rho}{\partial t} + i[H_{ab}(t), \rho] + \operatorname{Tr}_{R_a}\int_0^\infty [H_{aR_a}(t), [V(t, t-\tau)H_{aR_a}(t-\tau) \times V^+(t, t-\tau), \rho_{R_a}(0)\rho(t)]]d\tau + \operatorname{Tr}_{R_b}\int_0^\infty [H_{bR_b}(t), [V(t, t-\tau)H_{bR_b}(t-\tau) \times V^+(t, t-\tau), \rho_{R_b}(0)\rho(t)]]d\tau = 0, \tag{11.37}$$

where

$$H_{aR_a}(t) = \sum_j g_{j_a} a^+ a_j \exp\{-i\omega_j t + i\omega_a t\} + \text{H.C.}, \tag{11.38}$$

$$H_{bR_b}(t) = \sum_j g_{j_b} b^+ b_j \exp\{-i\omega_j t + i\omega_b t\} + \text{H.C.}, \tag{11.39}$$

and

$$V(t, \tau) = \exp\{-i(t-\tau)g(a^+ b + ab^+)\}. \tag{11.40}$$

In deriving (11.37) we made use of our conventional Born and short memory approximations. On simplifying (11.37) and on making the transformation to the Schrödinger picture we obtain the master equation

$$\begin{aligned}\frac{\partial\rho}{\partial t} &= -\mathrm{i}\omega_a[a^+a,\rho]-\mathrm{i}\omega_b[b^+b,\rho]-\mathrm{i}g[ab^+ + a^+b,\rho]\\ &\quad+\{\gamma_{a+}[a,\rho a^+]+\gamma_{b+}[b,\rho b^+]+\gamma_{a-}[a,\rho b^+]+\gamma_{b-}[b,\rho a^+]+\text{H.C.}\}\\ &\quad+\Gamma_{a+}\{[a^+,\rho a]+[a\rho,a^+]\}+\Gamma_{b+}\{[b^+,\rho b]+[b\rho,b^+]\}\\ &\quad+\tfrac{1}{2}\Gamma_{a-}\{[[b^+,\rho],a]+[[b,\rho],a^+]\}+\tfrac{1}{2}\Gamma_{b-}\{[[a^+,\rho],b]+[[a,\rho],b^+]\},\end{aligned} \tag{11.41}$$

where

$$\begin{aligned}\gamma_{i\pm} &= \tfrac{1}{2}(\gamma(\omega_i+g)\pm\gamma(\omega_i-g)),\\ \Gamma_{i\pm} &= \gamma(\omega_i+g)\langle n(\omega_i+g)\rangle\pm\gamma(\omega_i-g)\langle n(\omega_i-g)\rangle,\end{aligned} \tag{11.42}$$

and $\gamma(\omega)$ and $\langle n(\omega)\rangle$ are given by eqs. (6.9) and (6.6) respectively. This derivation of (11.41) differs from that of WALLS [1970], who considered the interaction of each of the normal modes of the Hamiltonian (11.31) with the heat bath. The Sudarshan-Glauber distribution function now satisfies the equation

$$\begin{aligned}\frac{\partial\Phi^{(A)}}{\partial t} &= \left\{(\gamma_+ +\mathrm{i}\omega)\frac{\partial}{\partial\alpha}\alpha+(\gamma_+ +\mathrm{i}\omega)\frac{\partial}{\partial\beta}\beta+\mathrm{i}g\left(\frac{\partial}{\partial\beta}\alpha+\frac{\partial}{\partial\alpha}\beta\right)\right.\\ &\quad\left.+\tfrac{1}{2}\Gamma_+\left(\frac{\partial^2}{\partial\alpha\,\partial\alpha^*}+\frac{\partial^2}{\partial\beta\,\partial\beta^*}\right)-\gamma_-\left(\frac{\partial}{\partial\beta}\alpha+\frac{\partial}{\partial\alpha}\beta\right)+\Gamma_-\frac{\partial^2}{\partial\alpha\,\partial\beta^*}+\text{C.C.}\right\}\Phi^{(A)},\end{aligned} \tag{11.43}$$

where we have again specialized to the case where $\omega_a = \omega_b$. Eq. (11.43) is in the form of a linearized Fokker-Planck equation and is easily solved by using the result given in Appendix. The steady state solution of (11.43) is

$$\begin{aligned}\Phi^{(A)}_{st} &= (\pi^2\langle n(\omega+g)\rangle\langle n(\omega-g)\rangle)^{-1}\exp\{-\tfrac{1}{2}(\langle n(\omega+g)\rangle^{-1}+\langle n(\omega-g)\rangle^{-1})\\ &\quad\times(|\alpha|^2+|\beta|^2)-\tfrac{1}{2}(\langle n(\omega+g)\rangle^{-1}-\langle n(\omega-g)\rangle^{-1})(\alpha^*\beta+\alpha\beta^*)\},\end{aligned} \tag{11.44}$$

which implies that*

$$\rho_{st}\propto\exp\{-(\omega a^+a+\omega b^+b+g(a^+b+ab^+))/(K_B T)\}. \tag{11.45}$$

This analysis shows that the system relaxes towards the correct steady state which corresponds to the total Hamiltonian $H_a+H_b+H_{ab}$, rather than the free Hamiltonian H_a+H_b. Of course in the weak coupling limit $g\to 0$, the master equation (11.43) reduces to (11.32). This result is the analog of the well known result in magnetic resonance, where a strongly driven spin

* The operator eq. (11.41) may be solved directly by using detailed balance (AGARWAL [1973b]).

system tends to relax along the instantaneous total field rather than along the static portion of the field.

§ 12. An Application of Phase Space Techniques to a Problem in Solid State Physics

In this section we consider a related problem namely the problem of anharmonic interaction in solids. This problem belongs to the domain of solid state physics, but we will discuss it here since it may be treated by methods that were specifically developed for treatments of problems in quantum optics. The classical treatment of this problem is given in detail in the book by PRIGOGINE [1962]; our quantum treatment follows closely that of CARRUTHERS and DY [1966] though our derivation of the master equation is different. We will in particular obtain the quantum analog of the BROUT-PRIGOGINE equation [1956].

We write the Hamiltonian in the form

$$H = \sum_k \omega_k a_k^+ a_k - \sum_{kk'k''} (\omega_k \omega_{k'} \omega_{k''})^{-\frac{1}{4}} (V_{kk'-k''} a_{k''}^+ a_{k'} a_k + \text{H.C.}), \quad (12.1)$$

where we have ignored the energy non-conserving terms and where $V_{kk'-k''}$ are the coupling coefficients. For our purpose their specific form is not needed. On using (12.1) and (2.20) (generalized to the case of many degrees of freedom), we find that the distribution function $\Phi^{(N)}(\{z_k\}, \{z_k^*\}; t)$ satisfies the equation

$$\frac{\partial \Phi^{(N)}}{\partial t} = \left(+i \sum_k \omega_k \left(\frac{\partial}{\partial z_k} z_k - \text{C.C.} \right) + i \sum_{kk'k''} (\omega_k \omega_{k'} \omega_{k''})^{-\frac{1}{4}} \left\{ V_{kk'-k''} \right.\right.$$
$$\left.\left. \times \left[z_{k''}^* \left(z_{k'} + \frac{\partial}{\partial z_{k'}^*} \right) \left(z_k + \frac{\partial}{\partial z_k^*} \right) - \left(z_{k''}^* + \frac{\partial}{\partial z_{k''}} \right) z_{k'} z_k \right] - \text{C.C.} \right\} \right) \Phi^{(N)}. \quad (12.2)$$

On transforming to the action angle variables J_k, φ_k defined by (2.30), eq. (12.2) reduces to

$$\frac{\partial \Phi^{(N)}}{\partial t} = \sum_k \omega_k \frac{\partial}{\partial \varphi_k} \Phi^{(N)} + \sum_{kk'k''} \left\{ \left(\frac{J_k J_{k'} J_{k''}}{\omega_k \omega_{k'} \omega_{k''}} \right)^{\frac{1}{4}} V_{kk'-k''} \exp\{i(\varphi_k + \varphi_{k'} - \varphi_{k''})\} \right.$$
$$\times \left[i \left(\frac{\partial}{\partial J_k} + \frac{\partial}{\partial J_{k'}} - \frac{\partial}{\partial J_{k''}} \right) - \frac{1}{2} \left(\frac{1}{J_k} \frac{\partial}{\partial \varphi_k} + \frac{1}{J_{k'}} \frac{\partial}{\partial \varphi_{k'}} + \frac{1}{J_{k''}} \frac{\partial}{\partial \varphi_{k''}} \right) \right.$$
$$\left. + \left(i \frac{\partial^2}{\partial J_k \partial J_{k'}} - \frac{1}{2J_{k'}} \frac{\partial^2}{\partial J_k \partial \varphi_{k'}} - \frac{1}{2J_k} \frac{\partial^2}{\partial J_{k'} \partial \varphi_k} - \frac{i}{4J_{k'} J_k} \frac{\partial^2}{\partial \varphi_k \partial \varphi_{k'}} \right) \right]$$
$$\left. + \text{C.C.} \right\} \Phi^{(N)}. \quad (12.3)$$

The energy distribution $\Phi^{(N)}(\{J_k\})$ is obtained from $\Phi^{(N)}$ by integrating over the phase variables φ_k, i.e.

$$\Phi^{(N)}(\{J_k\}; t) = \int_0^{2\pi} \ldots \int \Phi^{(N)}(\{J_k\}, \{\varphi_k\}; t) \mathrm{d}(\{\varphi_k\}). \tag{12.4}$$

The projection operator $\mathscr{P}$ is given by eq. (1.22), which can also be written as

$$\mathscr{P} \ldots = \psi(\{0\})(\psi(\{0\}), \ldots), \tag{12.5}$$

where $\psi(\{v_k\})$ are the eigenfunctions given by (2.35). The master equation for $\Phi^{(N)}(\{J_k\}; t)$ in Born and in Markovian approximations, is obtained from (4.12) and is given by

$$\frac{\partial \Phi^{(N)}}{\partial t} + \int_0^{\infty} (\psi(\{0\}), \mathscr{L}_I(t)\mathscr{L}_I(t-\tau)\psi(\{0\}))\Phi^{(N)} \mathrm{d}\tau = 0. \tag{12.6}$$

In deriving (12.6) we assumed that the distribution function at time $t = 0$ is independent of the phases φ_k so that $(1-\mathscr{P})\Phi^{(N)}(\{J_k\}, \{\varphi_k\}; 0) = 0$. We also made use of the relation $\mathscr{P}\mathscr{L}_I(t)\mathscr{P}\Phi^{(N)} = 0$, which is easily proved from the form of $\mathscr{L}_1$ as given by the second term on the right-hand side of (12.3). On introducing a complete set of states (12.6) can be written as

$$\frac{\partial \Phi^{(N)}(\{J_k\}; t)}{\partial t} + \sum_{\{v\}} \int_0^{\infty} \mathrm{d}\tau (\psi(\{0\}), \mathrm{e}^{+\mathrm{i}\mathscr{L}_0 t}\mathscr{L}_1 \mathrm{e}^{-\mathrm{i}\mathscr{L}_0 t}\psi(\{v_k\}))$$
$$\times (\psi(\{v_k\})\mathrm{e}^{+\mathrm{i}\mathscr{L}_0(t-\tau)}\mathscr{L}_1 \mathrm{e}^{-\mathrm{i}\mathscr{L}_0(t-\tau)}\psi(\{0\}))\Phi^{(N)}(\{J_k\}; t) = 0, \tag{12.7}$$

which on simplification leads to

$$\frac{\partial \Phi^{(N)}}{\partial t} + \sum_{\{v\}} (\psi(\{0\})\mathscr{L}_1\psi(\{v_k\}))(\psi(\{v_k\})\mathscr{L}_1\psi(\{0\}))\delta_-(\sum v_k \omega_k)\Phi^{(N)} = 0. \tag{12.8}$$

The equation (12.8) is not valid for time intervals given by $0 \leqq t \leqq 1/\omega_D$ where ω_D is the Debye frequency. It is easily shown from (12.3) that the only non vanishing elements of $\mathscr{L}_1$ are $(\psi(\{0\})\mathscr{L}_1\psi(\{-1_\kappa, -1_{\kappa'}, 1_{\kappa''}; 0\}))$, $(\psi(\{0\})\mathscr{L}_1\psi(\{1_\kappa, 1_{\kappa'}, -1_{\kappa''}, 0\}))$ and their complex conjugates. These matrix elements are easily computed and (12.8) then reduces to (CARRUTHERS and DY [1966])*

$$\frac{\partial \Phi^{(N)}}{\partial t} = 2\pi \sum_{kk'k''} \delta(\omega_k + \omega_{k'} - \omega_{k''}) \frac{|V_{kk'-k''}|^2}{(\omega_k \omega_{k'} \omega_{k''})} \left[\left(\frac{\partial}{\partial J_k} + \frac{\partial}{\partial J_{k'}} - \frac{\partial}{\partial J_{k''}} + \frac{\partial^2}{\partial J_k \partial J_{k'}}\right)\right.$$
$$\left. \times (J_k J_{k'} J_{k''}) \left(\frac{\partial}{\partial J_k} + \frac{\partial}{\partial J_{k'}} - \frac{\partial}{\partial J_{k''}} + \frac{\partial^2}{\partial J_k \partial J_{k'}}\right)\right] \Phi^{(N)}. \tag{12.9}$$

* One may similarly show that $\Phi^{(A)}$ satisfies (12.9) with the signs of second order derivatives in (. . .) changed.

This equation should be compared with the Brout-Prigogine equation for the classical distribution function $\Phi_{\rm cl}(\{J_k\}; t)$:

$$\frac{\partial}{\partial t}\Phi_{\rm cl} = 2\pi \sum_{kk'k''} \delta(\omega_k + \omega_{k'} - \omega_{k''}) \frac{|V_{kk'-k''}|^2}{(\omega_k \omega_{k'} \omega_{-k''})}$$
$$\times \left[\left(\frac{\partial}{\partial J_k} + \frac{\partial}{\partial J_{k'}} - \frac{\partial}{\partial J_{k''}}\right)(J_k J_{k'} J_{k''})\left(\frac{\partial}{\partial J_k} + \frac{\partial}{\partial J_{k'}} - \frac{\partial}{\partial J_{k''}}\right)\right] \Phi_{\rm cl}. \quad (12.10)$$

It is seen that (12.9) differs from (12.10) by the presence of second order derivatives which give rise to spontaneous decay of phonons. From eq. (12.9) we can easily establish the PEIERLS [1929] equation for the rate of change of phonons in the mode k:

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle n_k \rangle = \int \frac{\partial}{\partial t} \Phi^{(N)}(\{J_k\}; t) J_k \, \mathrm{d}(\{J_k\})$$
$$= 2\pi \sum_{k'k''} \frac{|V_{kk'-k''}|^2}{(\omega_k \omega_{k'} \omega_{k''})} \{2\delta(\omega_k + \omega_{k'} - \omega_{k''})[\langle (n_k+1)(n_{k'}+1)(n_{k''})\rangle$$
$$- \langle n_k n_{k'}(n_{k''}+1)\rangle] + \delta(\omega_k - \omega_{k'} - \omega_{k''})[\langle n_{k'} n_{k''}(n_k+1)\rangle$$
$$- \langle n_k (n_{k'}+1)(n_{k''}+1)\rangle]\}. \quad (12.11)$$

Finally we mention that we have not reviewed the related problems of optical pumping and resonant scattering of photons. Optical pumping has been treated by WILLIS [1970] using the master equation methods. In the theory of resonant scattering of photons methods related to master equation techniques have been employed by VDOVIN and GALITSKI [1965].

Appendix

Some Properties of the Fokker-Planck Process

We discuss in this appendix the properties of the Fokker-Planck process (for discussions of the properties of the Fokker-Planck process see, for example STRATONOVICH [1963] or LAX [1966a, b; 1968a]). Consider the Fokker-Planck equation

$$\frac{\partial P}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}(A_i P) + \sum_{ij} \frac{\partial^2}{\partial x_i \partial x_j}(D_{ij} P), \quad (A.1)$$

where A_i and D_{ij} are the drift and the diffusion coefficients respectively, given by

$$A_i(\{x_{i0}\}) = \lim_{t\to 0} \frac{1}{t} \langle x_{it} - x_{i0} \rangle$$

$$\equiv \lim_{t\to 0} \frac{1}{t} \int \ldots \int (x_{i1} - x_{i0}) P(\{x_{i1}\}, t | \{x_{i0}\}, 0) \mathrm{d}(\{x_{i1}\}), \quad \text{(A.2)}$$

$$D_{ij}(\{x_{i0}\}) = \tfrac{1}{2} \lim_{t\to 0} \frac{1}{t} \langle (x_{it} - x_{i0})(x_{jt} - x_{j0}) \rangle$$

$$\equiv \tfrac{1}{2} \lim_{t\to 0} \frac{1}{t} \int \ldots \int (x_{i1} - x_{i0})(x_{j1} - x_{j0}) P(\{x_{i1}\}, t | \{x_{i0}\}, 0) \mathrm{d}(\{x_{i1}\}). \quad \text{(A.3)}$$

In eqs. (A.2) and (A.3) $P(\{x_{i1}\}, t|\{x_{i0}\}, 0)$ is the conditional distribution function for the Fokker-Planck process and is the solution of (A.1) subject to the initial condition

$$P(\{x_{i1}\}, 0|\{x_{i0}\}, 0) = \prod_{i=1}^{N} \delta(x_{i1} - x_{i0}). \quad \text{(A.4)}$$

The N-time joint probability distributions can, of course, be determined from $P(\{x_{i1}\}, t|\{x_{i0}\}, 0)$ by using the Markoff property of the stochastic process under consideration. The drift coefficients in (A1) give the rate of change of the mean value of the variable $x_i(t)$, i.e.

$$\frac{\partial}{\partial t} \langle x_i(t) \rangle = \langle A_i(\{x_i\}) \rangle. \quad \text{(A.5)}$$

The diffusion coefficients D_{ij} give rise to the fluctuations in the mean value and form a semi-positive definite matrix. In the present article, we considered mainly quantum systems. The Fokker-Planck equation was then, an equation of motion for the phase space distribution function associated with the density operator. In such cases the diffusion matrix need not be semi-positive definite. However in most of the physical situations, the diffusion matrix is a semi-positive definite matrix and we assume it to be so throughtout this appendix.

It is known (BHARUCHA-REID [1960]) that the conditional probability distribution function $P(\{x_{i1}\}, t|\{x_{i0}\}, t_0)$ regarded as a function of $\{x_{i0}\}$ and t_0 satisfies the following Kolmogorov equation

$$\frac{\partial P}{\partial t_0} = -\sum_i A_i(\{x_{i0}\}) \frac{\partial P}{\partial x_{i0}} - \sum_{ij} D_{ij}(\{x_{i0}\}) \frac{\partial^2 P}{\partial x_{i0} \partial x_{j0}}. \quad \text{(A.6)}$$

Eq. (A.6) is the adjoint of eq. (A.1). Eq. (A.6) is also known as the "backward equation" and eq. (A.1) as the "forward equation".

Eq. (A.1) can be rewritten in the form

$$\partial P/\partial t + \sum_i \partial J_i/\partial x_i = 0, \tag{A.7}$$

where J_i is the ith component of the probability current given by

$$J_i = \left(A_i - \sum_j \frac{\partial}{\partial x_j} D_{ij}\right) P. \tag{A.8}$$

Eq. (A.7) can be interpreted as the equation for the conservation of the probability. The steady state solution of (A.7) is given by $\partial P/\partial t = 0$. In many physical problems the principle of detailed balance will be obeyed, i.e. $J_i = 0$. Then the steady state solution, denoted by P_{st}, is given by

$$\frac{\partial}{\partial x_i} \ln P_{st} = \sum_j (D^{-1})_{ij} \left(A_j - \sum_k \frac{\partial D_{jk}}{\partial x_k}\right). \tag{A.9}$$

From (A.9) and the relation

$$\partial^2 P/\partial x_i \partial x_j = \partial^2 P/\partial x_j \partial x_i, \tag{A.10}$$

it is seen that detailed balance is possible only if

$$\frac{\partial}{\partial x_j}\left[\sum_k (D^{-1})_{ik}\left(A_k - \sum_l \frac{\partial D_{kl}}{\partial x_l}\right)\right] = \frac{\partial}{\partial x_i}\left[\sum_k (D^{-1})_{jk}\left(A_k - \sum_l \frac{\partial D_{kl}}{\partial x_l}\right)\right]. \tag{A.11}$$

The condition (A.11) is known as the potential condition (STRATONOVICH [1963]). The connection of potential condition to the microscopic reversibility is well known* (see, for example LAX [1968a]). One may show that the potential condition is satisfied for the Van der Pohl oscillator model of laser light and eq. (A.9) then leads to the steady state solution given by eq. (10.33).

One may also prove an "H-theorem" for the Fokker-Planck process. We introduce the "H-function" defined by

$$H(t) = \int P(\{x_{i1}\}, t|\{x_{i0}\}, 0) \ln\left\{\frac{P(\{x_{i1}\}, t|\{x_{i0}\}, 0)}{P_{st}(\{x_{i1}\})}\right\} \mathrm{d}(\{x_{i1}\}). \tag{A.12}$$

It follows from the positive definiteness of P and the inequality $x \ln x - x \ln y - x + y \geqq 0$ that $H(t) \geqq 0$. Moreover on using (A.12) and (A.9), one may show that the rate of change of the "H-function" is given by

$$\frac{\mathrm{d}H}{\mathrm{d}t} = -\sum_{ij} \int P(\{x_{i1}\}, t|\{x_{i0}\}, 0) D_{ij} \left(\frac{\partial}{\partial x_i} \ln\left\{\frac{P(\{x_{i1}\}, t|\{x_{i0}\}, 0)}{P_{st}(\{x_{i1}\})}\right\}\right) \times \left(\frac{\partial}{\partial x_j} \ln\left\{\frac{P(\{x_{i1}\}, t|\{x_{i0}\}, 0)}{P_{st}(\{x_{i1}\})}\right\}\right) \mathrm{d}(\{x_{i1}\}). \tag{A.13}$$

* In our treatment we have assumed that the reversible part of the drift vector is zero (for generalizations see GRAHAM and HAKEN [1971]).

From eq. (A.13) and from the positive definiteness of the diffusion matrix it follows that

$$\mathrm{d}H/\mathrm{d}t \leqq 0, \tag{A.14}$$

which is the "H-theorem" for the multidimensional Fokker-Planck process.

We now discuss the eigenfunctions and the eigenvalues of the Fokker-Planck operator

$$\mathscr{L} = -\sum_i \frac{\partial}{\partial x_i} A_i + \sum_{ij} \frac{\partial^2}{\partial x_i \partial x_j} D_{ij}, \tag{A.15}$$

$$\mathscr{L}\psi_i = -\lambda_i \psi_i. \tag{A.16}$$

In general the operator $\mathscr{L}$ is not Hermitian. Eq. (A.16) can be changed into a self adjoint equation by the transformation of the eigenfunctions ψ_i

$$\psi_i = \exp\{\chi\}\varphi_i. \tag{A.17}$$

We then obtain the eigenvalue equation (LOUISELL [1969])

$$\mathscr{L}'\varphi_i = -\lambda_i \varphi_i, \tag{A.18}$$

where

$$\mathscr{L}' = \sum_{ij} \left\{ \frac{\partial}{\partial x_i} D_{ij} \frac{\partial}{\partial x_j} + \frac{\partial^2 D_{ij}}{\partial x_i \partial x_j} + D_{ij} \left(\frac{\partial \chi}{\partial x_i} \frac{\partial \chi}{\partial x_j} + \frac{\partial^2 \chi}{\partial x_i \partial x_j} \right) + 2 \frac{\partial \chi}{\partial x_i} \frac{\partial D_{ij}}{\partial x_j} \right\} - \sum_i \left\{ \frac{\partial A_i}{\partial x_i} + A_i \frac{\partial \chi}{\partial x_i} \right\}, \tag{A.19}$$

and the function χ is given by

$$\partial\chi/\partial x_i = -\tfrac{1}{2} \sum_j (D^{-1})_{ij} \Big(\sum_k \partial D_{jk}/\partial x_k - A_j \Big). \tag{A.20}$$

Since $\mathscr{L}'$ is a self-adjoint operator its eigenfunctions and eigenvalues form a complete set:

$$\sum_i \psi_i(\{x_{i1}\})\psi_i(\{x_{i0}\}) \exp(-2\chi) = \prod_i \delta(x_{i1} - x_{i0}), \tag{A.21}$$

and the conditional distribution function is then given by

$$P(\{x_{i1}\}, t | \{x_{i0}\}, t_0) = \sum_i \mathrm{e}^{-\lambda_i (t - t_0)} \psi_i(\{x_{i1}\})\psi_i(\{x_{i0}\}) \exp(-2\chi). \tag{A.22}$$

Some examples of the eigenfunction expansion are given in the book by STRATONOVICH [1963]. On comparing (A.20) and (A.9) it is seen that $\chi = \frac{1}{2} \ln P_{\mathrm{st}}$.

The conditional distribution function P, for the special case when the drift coefficients are linear in $\{x_i\}$ and the diffusion coefficients D_{ij} are

independent of random variables, is well known (WANG and UHLENBECK [1945]). If we set

$$A_i = \sum_j \beta_{ij} x_j, \tag{A.23}$$

then the solution is

$$P(\{x_{i1}\}, t|\{x_{i0}\}, 0) = [(2\pi)^N |\det \sigma(t)|]^{-\frac{1}{2}} \times \exp\{-\tfrac{1}{2}(X - b(t)X_0)^{\mathrm{T}} \sigma^{-1}(t)(X - b(t)X_0)\}. \tag{A.24}$$

where X_0 and X are the column matrices

$$X_0 = \begin{pmatrix} x_{10} \\ \vdots \\ x_{N0} \end{pmatrix}, \qquad X = \begin{pmatrix} x_{11} \\ \vdots \\ x_{N1} \end{pmatrix},$$

and the superscript T denotes the transpose of the matrix. The parameters $b(t)$ and $\sigma(t)$ are given by

$$b(t) = \mathrm{e}^{\beta t}, \qquad \sigma(t) = \sigma(\infty) - b(t)\sigma(\infty)b^{\mathrm{T}}(t), \tag{A.25}$$

and $\sigma(\infty)$ is the solution of

$$\beta\sigma(\infty) + \sigma(\infty)\beta^{\mathrm{T}} = -2D. \tag{A.26}$$

The linearized Fokker-Planck equations, i.e. the eq. (A.1) with A_i given by (A.23) and with the diffusion constants independent of $\{x_i\}$, occur in many physical examples such as in Brownian motion of an oscillator (eq. (7.7); WANG and UHLENBECK [1945], AGARWAL [1971d]); in the theories of relaxation of an oscillator (eq. (6.11); LOUISELL and MARBURGER [1967], AGARWAL [1969]), parametric frequency conversion (eq. (11.43); WALLS [1970]), parametric oscillator (GRAHAM [1968]) oscillating below or above threshold and a laser (RISKEN et al. [1966]) oscillating below or above threshold.

We next consider the Langevin treatment of the Fokker-Planck process. We will only quote the main result (for details see LAX [1966b], STRATONOVICH [1963]). Consider a nonlinear Langevin process defined by

$$\mathrm{d}x_i/\mathrm{d}t = \beta_i + \sum_j \sigma_{ij} F_j(t), \qquad (i = 1, 2, \ldots, N), \tag{A.27}$$

where $F_j(t)$ are independent delta correlated Gaussian random processes, i.e.

$$\langle F_j(t) \rangle = 0, \qquad \langle F_i(t) F_j(t') \rangle = 2\delta_{ij}\delta(t - t'), \tag{A.28}$$

and all the higher order linked moments (cumulants) vanish. Then the Langevin process (A.27) is equivalent to the Fokker-Planck process (A.1) with

$$A_i = \beta_i + \sum_{jk} \frac{\partial \sigma_{ij}}{\partial x_k} \sigma_{kj}, \tag{A.29}$$

and

$$D_{ij} = \sum_k \sigma_{ik} \sigma_{jk}. \tag{A.30}$$

Conversely if a Fokker-Planck process is given, we can construct a Langevin process since the diffusion matrix is a symmetric and positive definite matrix and necessarily possesses a square root.

It is easily seen that the linearized Fokker-Planck equation is equivalent to the Langevin equations

$$\mathrm{d}x_i/\mathrm{d}t = \sum_j \beta_{ij} x_j + F_i(t), \qquad (i = 1, 2, \ldots, N), \tag{A.31}$$

where

$$\langle F_i(t) \rangle = 0, \qquad \langle F_i(t) F_j(t') \rangle = 2D_{ij} \delta(t-t'), \tag{A.32}$$

and all the higher order linked moments of the random force $F_i(t)$ vanish. It is easily seen from (A.31) that the random variables $\{x_i\}$ constitute a Gaussian Markov process. It is also clear from (A.31) that the autocorrelation function of the process $X(t)$ is given by

$$\langle X(t) X^{\mathrm{T}}(0) \rangle = \exp(\beta t) \langle X(0) X^{\mathrm{T}}(0) \rangle. \tag{A.33}$$

We therefore obtain Doob's theorem (WANG and UHLENBECK [1945]): A random process that is stationary, Gaussian and Markovian, possesses an autocorrelation function of the form (A.33).

In the text of the present article we have written the Fokker-Planck equations in terms of the complex coordinates z and z^* in the form*

$$\frac{\partial P}{\partial t} = \left\{ \frac{\partial}{\partial z} (A(z) P) + \text{C.C.} \right\} + \left\{ D_{zz} \frac{\partial^2 P}{\partial z^2} + D_{zz^*} \frac{\partial^2 P}{\partial z \, \partial z^*} + \text{C.C.} \right\}. \tag{A.34}$$

This Fokker-Planck equation is equivalent to the following Langevin equations

$$\dot{z} = -A(z) + F(t), \qquad \dot{z}^* = -[A(z)]^* + F^*(t), \tag{A.35}$$

where $F(t)$ is a complex delta correlated Gaussian random process with zero mean, i.e.

$$\langle F(t) \rangle = 0, \qquad \langle F(t) F(t') \rangle = 2D_{zz} \delta(t-t'),$$
$$\langle F(t) F^*(t') \rangle = 2D_{zz^*} \delta(t-t'), \tag{A.36}$$

and all the higher order linked moments of the complex force $F(t)$ vanish.

* We have assumed for the sake of simplicity that D's are constants.

References

ABRAGAM, A., 1961, The Principles of Nuclear Magnetism (Oxford at the Clarendon Press).

AGARWAL, G. S., 1969, Phys. Rev. **178**, 2025.

AGARWAL, G. S., 1970, Phys. Rev **A2**, 2038.

AGARWAL, G. S., 1971a, Phys. Rev. **A3**, 828.

AGARWAL, G. S., 1971b, Phys. Rev. **A3**, 1783.

AGARWAL, G. S., 1971c, Nuovo Cimento Letters **2**, 49.

AGARWAL, G. S., 1971d, Phys. Rev. **A4**, 739.

AGARWAL, G. S., 1971e, Phys. Rev. **A4**, 1791.

AGARWAL, G. S., 1972, Z. Physik **252**, 25.

AGARWAL, G. S., 1973a, in: Proc. Third Rochester Conference on Coherence and Quantum Optics, eds. L. Mandel and E. Wolf (Plenum Publ. Co., New York) p. 157.

AGARWAL, G. S., 1973b, Z. Physik **258**, 409.

AGARWAL, G. S., 1973c, Z. Physik **258**, 401.

AGARWAL, G. S. and E. WOLF, 1968, Phys. Rev. Letters **21**, 180.

AGARWAL, G. S. and E. WOLF, 1970a, Phys. Rev. **D2**, 2161.

AGARWAL, G. S. and E. WOLF, 1970b, Phys. Rev. **D2**, 2187.

AGARWAL, G. S. and E. WOLF, 1970c, Phys. Rev. **D2**, 2206.

ARGYRES, P. N., 1966, in: Lectures in Theoretical Physics, ed. W. E. Brittin (University of Colorado Press, Boulder) Vol. VIIIA, p. 783.

ARGYRES, P. N. and P. L. KELLEY, 1964, Phys. Rev. **134A**, 98.

ARZT, V., H. HAKEN, H. RISKEN, H. SAUERMANN, C. SCHMID and W. WEIDLICH, 1966, Z. Physik **197**, 207.

BHARUCHA-REID, A. T., 1960, Elements of the Theory of Markov Processes and their Applications (McGraw Hill Co. Inc., New York) Chap. 3.

BLOCH, F., 1956, Phys. Rev. **102**, 104.

BLOCH, F., 1957, Phys. Rev. **105**, 1206.

BONIFACIO, R., P. SCHWENDIMAN and F. HAAKE, 1971a, Phys. Rev. **A4**, 302.

BONIFACIO, R., P. SCHWENDIMAN and F. HAAKE, 1971b, Phys. Rev. **A4**, 854.

BROUT, R. and I. PRIGOGINE, 1956, Physica **22**, 621.

CAHILL, K. E. and R. J. GLAUBER, 1969a, Phys. Rev. **177**, 1857.

CAHILL, K. E. and R. J. GLAUBER, 1969b, Phys. Rev. **177**, 1882.

CURRUTHERS, P. and K. S. DY, 1966, Phys. Rev. **147**, 214.

CHANDRASEKHAR, S., 1943, Rev. Mod. Phys. **15**, 1; this paper is reprinted in: Selected Papers on Noise and Stochastic Processes, ed. N. Wax (Dover Publications, Inc., New York 1954).

CHESTER, G. V. and A. THELLUNG, 1959, Proc. Phys. Soc. **73**, 745.

COMPAAN, A. and I. D. ABELLA, 1971, Phys. Rev. Letters. **27**, 23.

DAVIS, H. T., 1962, Introduction to Nonlinear Differential and Integral Equations (Dover Publications, Inc., New York) p. 364.

DIALITIS, D., 1970, Phys. Rev. **A2**, 559.

DICKE, R. H., 1954, Phys. Rev. **93**, 99.

DILLARD, M. and H. R. ROBL, 1969, Phys. Rev. **184**, 312.

EMCH, G. G. and G. L. SEWELL, 1968, J. Math. Phys. **9**, 946.

FEYNMAN, R. P., F. L. VERNON and R. W. HELLWARTH, 1957, J. Appl. Phys. **28**, 49.

FOX, R. F., 1972, J. Math. Phys. **13**, 1196.

GINZBURG, V. L. and L. LANDAU, 1950, J. Expl. Theor. Phys. (U.S.S.R.) **20**, 1064.

GLAUBER, R. J., 1963, Phys. Rev. **131**, 2766.

GLAUBER, R. J., 1965, in: Quantum Optics and Electronics, eds. C. Dewitt, A. Blandin and C. Cohen-Tannoudji (Gordon and Breach, New York) p. 165.

GORDON, J. P., 1967, Phys. Rev. **161**, 367.

GRAHAM, R., 1968, Z. Physik **210**, 319.
GRAHAM, R., F. HAAKE, H. HAKEN and W. WEIDLICH, 1968, Z. Physik **213**, 21.
GRAHAM, R. and H. HAKEN, 1970, Z. Physik **237**, 31.
GRAHAM, R. and H. HAKEN, 1971, Z. Physik **243**, 289.
GROSSMAN, S. and P. H. RICHTER, 1971, Z. Physik **242**, 458.
GROENEWOLD, H. J., 1946, Physica **12**, 405.
HAAKE, F., 1969a, Z. Physik **223**, 364.
HAAKE, F., 1969b, Z. Physik **227**, 179.
HAKEN, H., 1970, Laser Theory, in: Encylopedia of Physics, Vol. XXV/2e, ed. S. Flugge (Springer Verlag, Berlin).
HEMPSTEAD, R. D. and M. LAX, 1967, Phys, Rev. **161**, 350.
KLAUDER, J., 1960, Ann Phys. **11**, 123.
KLAUDER, J. and E. C. G. SUDARSHAN, 1968, Fundamental of Quantum Optics (W. A. Benjamin, New York).
KOHN, W. and J. M. LUTTINGER, 1957, Phys. Rev. **108**, 590.
KRAMERS, H., 1940, Physica **7**, 284.
KUBO, R., 1964, J. Phys. Soc. (Japan) **19**, 2127.
LAMA, W. L., R. JODOIN and L. MANDEL, 1972, Am. J. Phys. **40**, 32.
LAMB, W. E., 1964, Phys. Rev. **134A**, 1429.
LAMB, W. E., 1965, in: Quantum Optics and Electronics, eds. C. Dewitt, A. Blandin and C. Cohen-Tannoudji (Gordon and Breach, New York) p. 377.
LANGER, J. S., 1969, Phys. Rev. **184**, 219.
LAX, M., 1966a, Rev. Mod. Phys. **38**, 359.
LAX, M., 1966b, Rev. Mod. Phys. **38**, 541.
LAX, M., 1966c, Phys. Rev. **145**, 110.
LAX, M., 1968a, in: Statistical Physics, Phase Transitions and Superfluidity (Gordon and Breach Science Publishers, New York) Vol. II.
LAX, M., 1968b, Phys. Rev. **172**, 350.
LAX, M. and W. H. LOUISELL, 1967, I.E.E.E. J.Quant. Electron. **QE 3**, 47.
LAX, M. and W. H. LOUISELL, 1969, Phys. Rev. **185**, 568.
LEHMBERG, R. H., 1970a, Phys. Rev. **A2**, 883.
LEHMBERG, R. H., 1970b, Phys. Letters **33A**, 501.
LOUISELL, W. H., 1969, in: Quantum Optics, ed. R. J. Glauber (Academic Press, New York) p. 680.
LOUISELL, W. H. and L. R. WALKER, 1965, Phys. Rev. **137**, B204.
LOUISELL, W. H. and J. H. MARBURGER, 1967, IEEE J. Quant. Electron. **QE 3**, 348.
MCCALL, S. L., and E. L. HAHN, 1969, Phys. Rev. **183**, 457.
MEHTA, C. L., 1967, Phys. Rev. Letters **18**, 752.
MEHTA, C. L. and E. C. G. SUDARSHAN, 1965, Phys. Rev. **138B**, 274.
MEHTA, C. L. and E. C. G. SUDARSHAN, 1966, Phys. Letters **22**, 574.
MONTROLL, E. W., 1961, in: Lectures in Theoretical Physics, eds. W. E. Brittin, B. W. Downs and J. Downs (Interscience Publishers Inc., New York) p. 221.
MONTROLL, E. W., 1962, in: Fundamental Problems in Statistical Mechanics, ed. E. G. D. Cohen (North-Holland Publishing Co., Amsterdam) p. 230.
MONTROLL, E. W. and K. E. SHULER, 1957, J. Chem. Phys. **26**, 454.
MOYAL, J., 1949, Proc. Camb. Phil. Soc. **45**, 99.
OPPENHEIM, I., K. E. SHULER and G. H. WEISS, 1967, in: Advances in Molecular Relaxation Processes **1**, 13.
PAULI, W., 1928, in: Probleme der Modern Physik, Arnold Sommerfeld zum 60, p. 30–45; reprinted in: Collected Scientific Papers, eds. P. Kronig and V. F. Weisskopf (Interscience Publishers, New York, 1964) p. 549.
PEIERLS, R. E., 1929, Ann. Physik **3**, 1055.
PRIGOGINE, I., 1962, Non Equilibrium Statistical Mechanics (Interscience Publishers, New York).

PRIGOGINE, I. and P. RESIBOIS, 1961, Physica **27**, 629.
REDFIELD, A. G., 1957, IBM J. Res. Develop. **1**, 19.
REDFIELD, A. G., 1965, in: Advances in Magnetic Resonance, ed. J. S. Waugh (Academic Press, New York) Vol. **1**, p. 1.
REHLER, N. E. and J. H. EBERLY, 1971, Phys. Rev. **A3**, 1735.
RESIBOIS, P., 1961, Physica **27**, 541.
RISKIN, H., 1970, in: Progress in Optics, ed. E. Wolf (North-Holland Publishing Co., Amsterdam) Vol. VIII, p. 241.
RISKIN, H., 1972, Z. Physik **251**, 231.
RISKIN, H., C. SCHMIDT and W. WEIDLICH, 1966, Z. Physik **193**, 37.
RISKIN, H. and H. D. VOLLMER, 1967a, Z. Physik **201**, 323.
RISKIN, H. and H. D. VOLLMER, 1967b, Z. Physik **204**, 240.
SCHWEBER, S. S., 1961, An Introduction to Relativistic Quantum Field Theory (Harper and Row, New York).
SCHWINGER, J., 1961, J. Math. Phys. **2**, 407.
SCULLY, M. O. and V. DIGIORGIO, 1970, Phys. Rev. **A2**, 1170.
SCULLY, M. O. and W. E. LAMB, 1967, Phys. Rev. **159**, 208.
SCULLY, M. O. and K. G. WHITNEY, 1972, in: Progress in Optics, ed. E. Wolf (North-Holland Publishing Co., Amsterdam) Vol. X, p. 89.
SENITZKY, I. R., 1960, Phys. Rev. **119**, 670.
SENITZKY, I. R., 1961, Phys. Rev. **124**, 642.
SHULER, K. E., 1958, Phys. Fluids **2**, 442.
SLITCHER, C. P., 1963, Principles of Magnetic Resonance (Harper and Row Publishers, New York).
SNEDDON, I. N., 1957, Elements of Partial Differential Equations (McGraw Hill Book Co., New York).
STRATONOVICH, R. L., 1963, Topics in the Theory of Random Noise (Gordon and Breach Science Publishers, New York) Vol. I.
SUDARSHAN, E. C. G., 1963, Phys. Rev. Letters **10**, 277.
ULLERSMA, P., 1966a, Physica **32**, 27.
ULLERSMA, P., 1966b, Physica **32**, 56.
VAN KAMPEN, N. G., 1954, Physica **20**, 603.
VAN KAMPEN, N. G., 1971, Rep. Prog. Phys. **2**, 199.
VAN HOVE, L., 1955, Physica **21**, 517.
VAN HOVE, L., 1957, Physica **23**, 441.
VAN HOVE, L., 1962, in: Fundamental Problems in Statistical Mechanics, ed. E. G. D. Cohen (North-Holland Publishing Co., Amsterdam) p. 157.
VDOVIN, Yu. A. and V. M. GALITSKI, 1965, Sov. Phys. J. E.T.P. **21**, 904.
WALLS, D. F., 1970, Z. Physik **234**, 231.
WANG, M. C. and G. E. UHLENBECK, 1945, Rev. Mod. Phys. **17**, 323; this article is also reprinted in: Selected Papers in Noise and Stochastic Processes, ed. N. Wax (Dover Publications, Inc., New York, 1954).
WEIDLICH, W. and F. HAAKE, 1965, Z. Physik **185**, 30.
WEISSKOPF, V. and E. WIGNER, 1930, Z. Physik **63**, 54; this paper is translated in: Atomic Spectra, by W. R. Hindmarsh (Pergamon Press, New York, 1967) p. 304.
WIGNER, E., 1932, Phys. Rev. **40**, 749.
WILLIS, C. R., 1970, Phys. Rev. **A1**, 467.
WILLIS, C. R. and R. H. PICARD, 1973, in: Proc. Third Rochester Conference on Coherence and Quantum Optics, eds. L. Mandel and E. Wolf (Plenum Publ. Co., New York).
ZWANZIG, R., 1961a, in: Lectures in Theoretical Physics, eds. W. E. Brittin, B. W. Downs and J. Downs (Interscience Publishers, Inc., New York) p. 106.
ZWANZIG, R., 1961b, Phys. Rev. **124**, 983.
ZWANZIG, R., 1964, Physica **30**, 1109.

Note added in proof:

Since this article was submitted for publication, much more has been done on the properties of Fokker-Planck equations and their generalizations under the condition that detailed balance is obeyed (GRAHAM and HAKEN [1971], RISKEN [1972], AGARWAL [1972, 1973b, c], VAN KAMPEN [1971]).

The attention of the author has also been drawn to a number of Russian papers which deal with master equation methods in quantum optical problems:

ZEL'DOVICH, B. YA., A. M. PERELOMOV and V. S. POPOV, 1969, Sov. Phys. JETP **28**, 308; 1970, Sov. Phys. JETP **30**, 111;
BELAVIN, A. A., B. YA. ZEL'DOVICH, A. M. PERELOMOV and V. S. POPOV, 1969, Sov. Phys. JETP **29**, 145.

E. WOLF, PROGRESS IN OPTICS XI © NORTH-HOLLAND 1973

II

RECENT DEVELOPMENTS IN FAR INFRARED SPECTROSCOPIC TECHNIQUES

BY

HIROSHI YOSHINAGA

Department of Applied Physics, Osaka University, Suita, Osaka, Japan

CONTENTS

§ 1. Introduction

The boundary of the far infrared region is not well defined. In the region beyond 20–25 μm, suitable materials for prisms are difficult to find, and gratings have been used as dispersion devices. Measurements using optical methods have been extended up to a few mm region at present. So in this article the far infrared region is defined as the region from 20–25 μm to a few mm, and recent developments in spectroscopic techniques in this region will be described. Of course, the techniques used in the infrared region below 20 μm, which may be available in the far infrared region in principle, will be discussed in this article.

§ 2. Conventional Far Infrared Spectroscopic Instruments

Dispersion-type instruments using prisms and/or gratings have been used in the far infrared region. Newly developed elements of such instruments and interferometric instruments will be explained in this paragraph.

2.1. DISPERSION-TYPE SPECTROMETERS

In the far infrared region, prisms of KRS-5, CsBr or CsI have been used, but their resolution is poor and these prisms act now only as supplementary elements in dispersion-type instruments. Gratings of the Echelette type are the main dispersive elements, but elimination of higher order spectra which overlap the first order spectrum is troublesome for grating instruments. It is very important to find sharp cutoff filters to eliminate higher order spectra in various wavelength regions. For this purpose, grating filters, metal mesh filters and Reststrahlen powder filters have been used. The last filters are used widely in grating instruments, and recent developments in their use, especially at low temperatures, will be described.

High pressure mercury lamps are still the only radiation source in the far infrared region, and the available energy from this source is not sufficient for obtaining measurements of high resolution. So very sensitive detectors are extremely important.

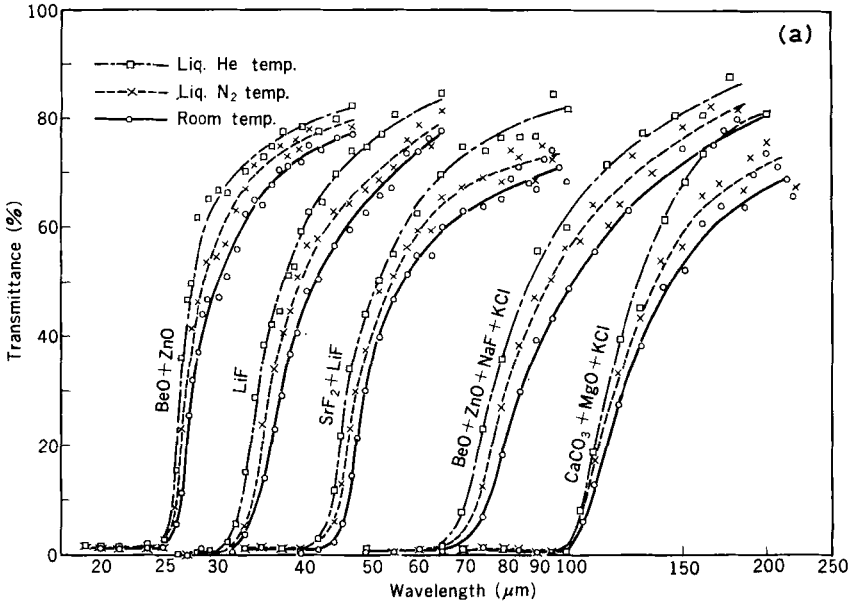


Fig. 2.1.a.

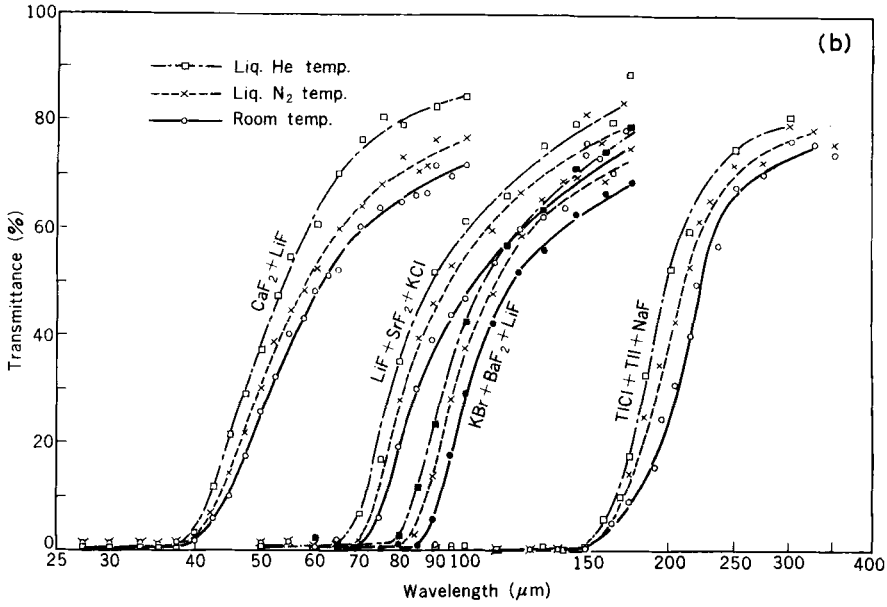


Fig. 2.1.b.

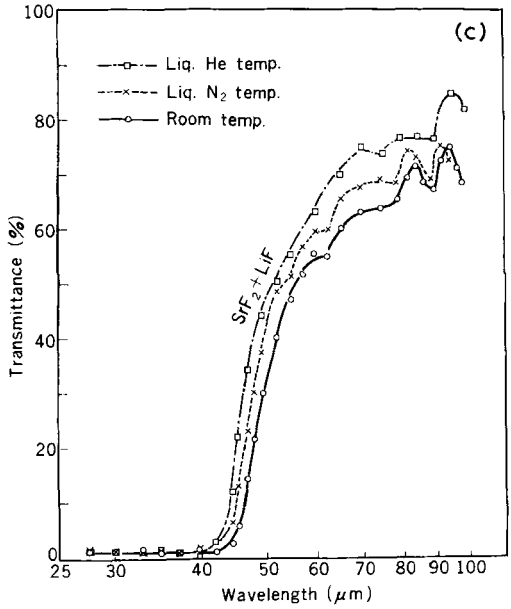


Fig. 2.1.c.

Fig. 2.1. Transmittance of Reststrahlen powder filters at various temperatures (SAKAI et al. [1965]).

2.1.1. *Reststrahlen powder filters*

Transmission filters of polyethylene sheet, containing Reststrahlen crystal powders, have been developed in our laboratory (YAMADA, MITSUISHI and YOSHINAGA [1962]). The sharp cut-off characteristics of these filters depend upon the fineness and the high concentration of the crystal powders and the thinness of the polyethylene sheet. Fig. 2.1 (SAKAI, NAKAGAWA and YOSHINAGA [1968]) shows the cut-off characteristic curves of these filters at room, liquid nitrogen and liquid helium temperatures. The data at liquid helium temperature were obtained for these filters in a light pipe inserted into a cryostat. The sharpness of cut-off characteristics of all filters increases with the decrease of temperature, and the maximum transmittance measured in a light pipe at liquid helium temperature is higher than that measured for a parallel beam at room and liquid nitrogen temperatures. Such filters are also important as cooled filters for low temperature detectors.

Beyond 250 μm, there is no Reststrahlen crystal suitable for making such filters, and metal wire-cloth meshes are used as reflection filters in the long wavelength side of the region (MITSUISHI, OHTSUKA, FUJITA and YOSHINAGA [1963]). Interference filters of multi-layers are superior to Reststrahlen crystal filters under 30 μm.

2.1.2. *Detectors*

It is important that in detectors used in far infrared spectroscopic instruments the value of the noise equivalent powder (NEP) be as low as possible. Also the size of the receiving area of detectors must be considered because it is difficult to focus the radiation falling on detectors in a light pipe to a very small area. Recently impurity doped Ge bolometers cooled by liquid helium have been commonly used in far infrared spectrometers. The value of the NEP of such bolometers decreases at lower temperatures. A typical large ($\sim 6\ \mathrm{mm}^3$) Ge (RICHARDS [1970a]) bolometer for far infrared use with a sink temperature of 1.1°K and NEP $\sim 10^{-12}\ \mathrm{W}/\sqrt{\mathrm{Hz}}$, which has been commonly used, would correspond to $\sim 3\times 10^{-14}\ \mathrm{W}/\sqrt{\mathrm{Hz}}$ at 0.37°K and $\sim 4\times 10^{-18}\ \mathrm{W}/\sqrt{\mathrm{Hz}}$ at 0.02°K. A temperature of 0.02°K or less can be reached by reliable techniques of modern low temperature physics, such as ^{3}He–^{4}He dilution refrigerators or adiabatic demagnetization. Sievers et al. have developed such a bolometer (DREW and SIEVERS [1969]). P-type Ge doped with In and Sb (6×10^{16} and 2.4×10^{16} atoms/cm^3 respectively) or Ga doped Ge is used as the detector element. Fig. 2.2 shows the construction of the cryostat for the bolometer, which is cooled with ^{3}He. The NEP value is $3\times 10^{-14}\ \mathrm{W}/\sqrt{\mathrm{Hz}}$ at 0.37°K and the response time is

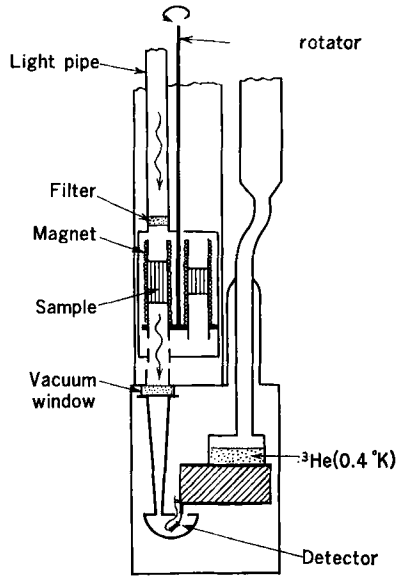


Fig. 2.2. Schematic diagram of Ge bolometer cooled by ^{3}He and sample chamber (DREW and SIEVERS [1969]).

10^{-2} sec. This bolometer is probably the most sensitive detector of far infrared radiation developed until now.

To cool detectors with liquid helium is not handy for use, especially in commercial instruments, and detectors used at room temperature are much more convenient. So a Golay detector has been used instead of its actual NEP value of $\sim 10^{-10}$ W/$\sqrt{\text{Hz}}$. Recently a pyroelectric detector using T.G.S. or S.B.N. has been developed (PUTLEY [1970]). The NEP value will approach the value of the Golay detector in the future and the response time is much faster than that of Ge bolometers.

The detectivity of those detectors described above is non-selective for wavelengths. Heterodyne detection of low-level coherent radiation will be important even for spectroscopic purposes in the future. A measurement of the signal-to-noise ratio and minimum detectable power for heterodyne detection of scattered radiation at 10.6 μm of CO_2 laser has been developed (TEICH, KEYES and KINGSTON [1966]). Using photoconductive Ge doped with Cu as the detector, the observed minimum detectable power, at a frequency of 70 kHz and in a bandwidth of 270 kHz, is 3.5×10^{-14} W. This corresponds to a minimum detectable power of 1.3×10^{-19} W in a 1-Hz bandwidth. Also heterodyne detection with a Bi thin film bolometer is reported (CONTRERAS and GADDY [1971]), but the minimum detectable power is much poorer than the Ge photoconductive detector doped with Cu.

In these developments of elements of far infrared spectroscopic instruments, resolution and rapid measurement are limited owing to the shortage of available far infrared energy. The present tendency of the development of far infrared dispersion-type spectroscopic instruments is to simplify the construction and to adapt the mechanism to a special purpose, especially for commercial instruments.

2.2. INTERFEROMETRIC SPECTROMETERS

Interferometric spectrometers compared with dispersion-type ones have advantages as follows:

a) faster optical system (Jacquinot's effect),

b) simultaneous measurement for the whole wavelength region, which increases the efficiency of utility of radiation energy (Ferrgett's effect).

Owing to these advantages, interferometric spectrometers are superior to dispersion-type ones, especially in the long wave side of the far infrared region, where available radiation energy from the source is very low. There are two types of interferometric instruments, i.e., the Michelson-type interferometer and the Laminar-type interferometer, and the latter is more efficient in the long wave side of the region, which does not need the beam

splitter. Recent tendencies in the development of far infrared interferometric spectrometers will be described below.

2.2.1. *High resolution*

As is well known, the resolution is reciprocally proportional to the maximum path difference of two beams of the interferometer. The highest resolution of far infrared interferometric spectrometers is 0.05 cm^{-1} for the maximum path difference of 20 cm (SANDERSON and SCOTT [1971]); this value is much better than that obtained with dispersion type spectrometers. Fig. 2.3 shows the spectra of water vapor at various path differences (DOWLING [1970]). Spurious spectral lines are fairly dominant in the spectrum for short path differences, but they become weaker and the spectral linewidth narrower with the increase of path difference without any apodization. Finally the spectral linewidth is very sharp and spurious spectral lines can not be seen, except noise. Fig. 2.4 shows the spectrum of water vapor in the

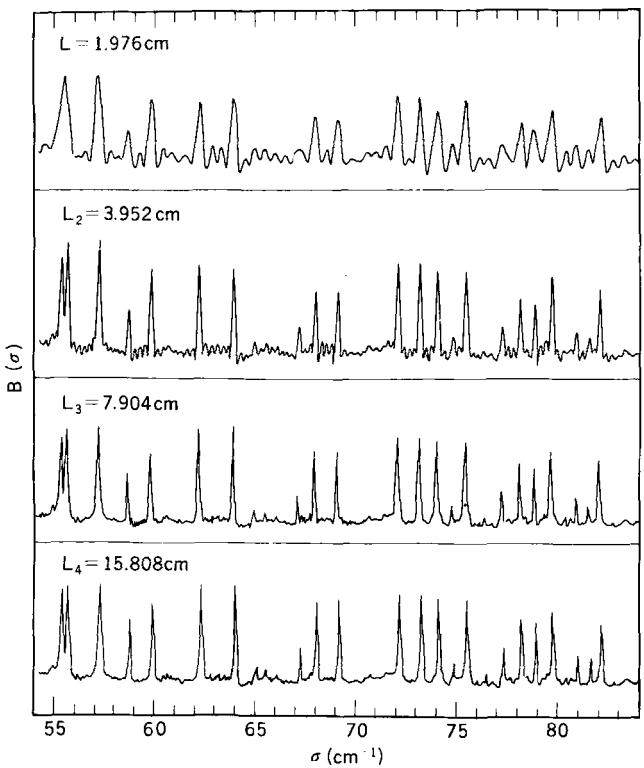


Fig. 2.3. Spectra of water vapor between 54 and 84 cm^{-1} at various path differences (DOWLING [1970]).

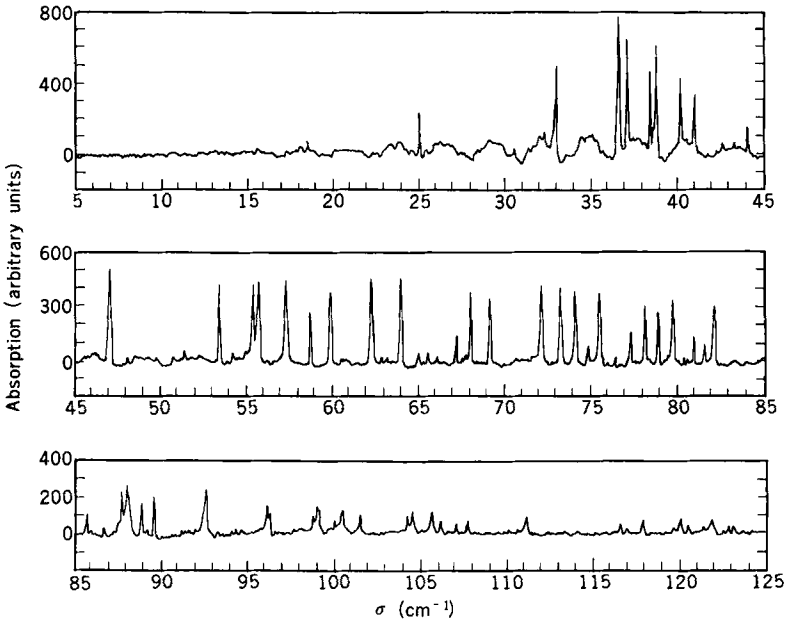


Fig. 2.4. Highly resolved water vapor absorption lines in the far infrared region (DOWLING [1970]).

far infrared region obtained with an interferometric instrument (DOWLING [1970]). There are six absorption lines near 35–40 cm^{-1}, and these lines are resolved completely in this spectrum, but the two center lines are difficult to resolve completely with dispersion-type spectrometers.

Connes and his coworker have obtained a resolution of 0.005 cm^{-1} for a path difference of 200 cm in the middle infrared region (CONNES [1970]). The mechanical accuracy of the far infrared interferometer needs to be much lower than that of the interferometer in the middle infrared region, say in the 10 μm region, because such inaccuracy needs to be proportional to the wavelength measured. The signal-to-noise ratio of the interferogram in the far infrared region is much smaller than that in the middle infrared region and the resolution does not increase with increase of path difference when the path difference becomes large. This is the reason that the highest resolution in the far infrared region is poorer than that in the middle infrared region.

An asymmetric-type interferometer developed by Bell (BELL [1966], [1970]) measures the amplitude of light instead of the strength of light. So this spectrometer can obtain the very accurate index of refraction of gases shown in Fig. 2.5. But the sample should be inserted in one of the two beams of

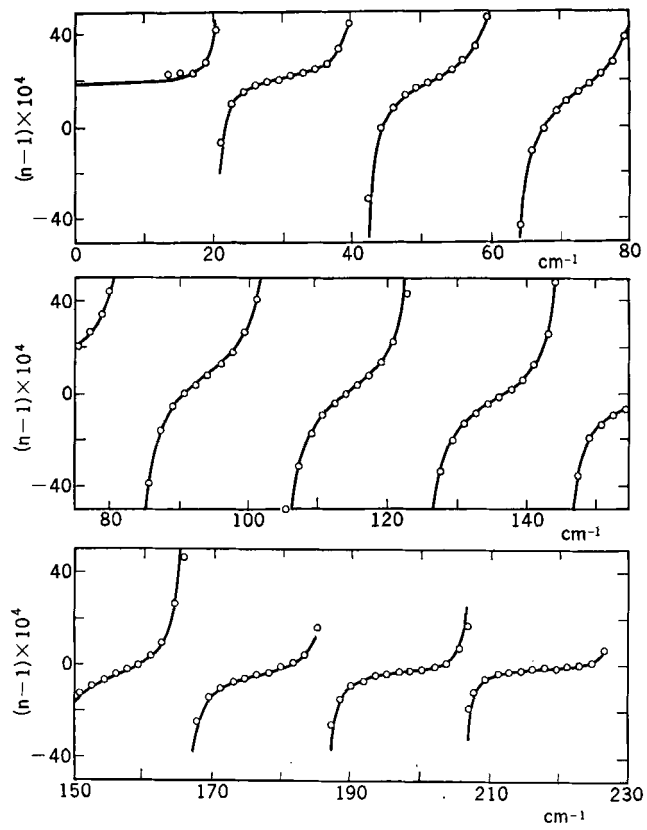


Fig. 2.5. Index of refraction of HCl absorption lines. The solid line and circles show calculated and experimentally determined values (BELL [1970]).

the interferometer. So there is no trouble for gas or liquid samples but it is in general difficult to measure solid samples.

2.2.2. *Double beam type*

Double beam spectrometers are now common among dispersion-type infrared spectrometers. For interferometric spectrometers a double beam system has not much benefit compared to dispersion type spectrometers, and is not so common. Two examples of this type will be described below.

Dowling started the optical arrangement of the sample-in and sample-out system (HALL, VRABEC and DOWLING [1966]). Let the output of the detector for the period of sample-out and sample-in be F_0 and F_s respectively, and the spectral intensity in the cases of sample-out and sample-in I_0 and I_s respectively. Then the next two formulae are obtained for the Fourier

transformation:

$$F_0(x) = \tfrac{1}{2}F_0(0) + \int_0^\infty I_0(\nu)\cos(2\pi\nu x)\mathrm{d}\nu,$$

$$F_s(x) = \tfrac{1}{2}F_s(0) + \int_0^\infty I_s(\nu)\cos(2\pi\nu x)\mathrm{d}\nu,$$

where x is the path difference. Then

$$F_0(x) - F_s(x) = \tfrac{1}{2}[F_0(0) - F_s(0)] + \int_0^\infty [I_0(\nu) - I_s(\nu)]\cos(2\pi\nu x)\mathrm{d}\nu.$$

The value of $F_0(x) - F_s(x)$ is obtained at the path difference x. Simple Fourier transformation gives the value of $I_0(\nu) - I_s(\nu)$, which shows the difference between spectral intensity in the cases of sample-out and sample-in. So the $[I_0(\nu) - I_s(\nu)] : \nu$ curve is just like a curve of emission spectrum, as shown in

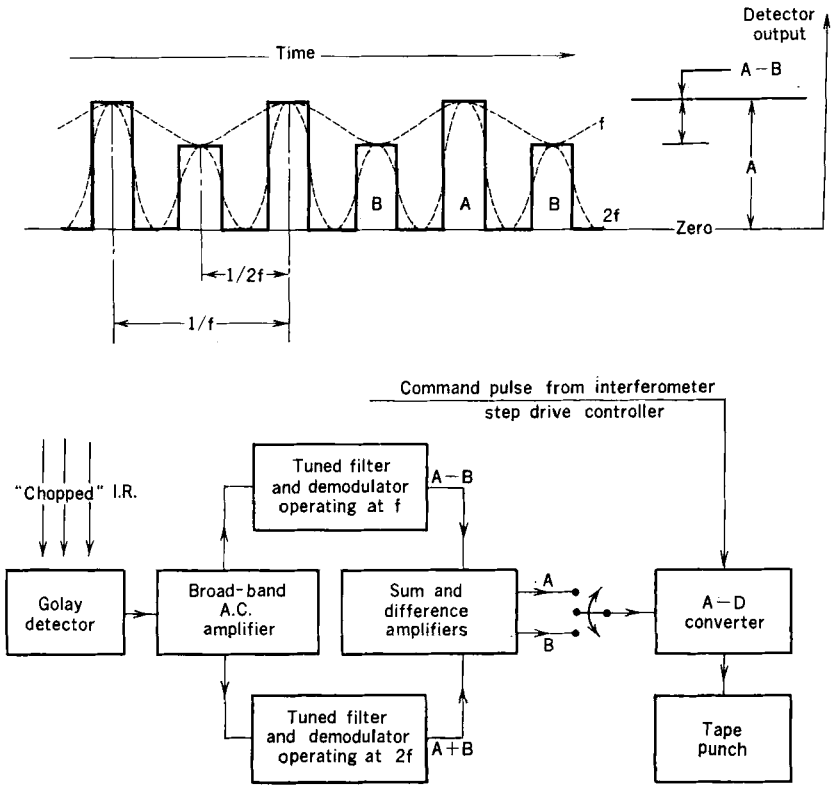


Fig. 2.6. Schematic representation of two signals and electronic devices to obtain each value of these two signals (THORPE et al. [1969]).

Fig. 2.4. Of course the transmittance of the sample can be obtained from the ratio $I_s(\nu)/I_0(\nu)$, but unwanted effects like channel spectra of the plane-parallel sample, transmission filters and windows can be avoided in the curve of $[I_0(\nu)-I_s(\nu)] : \nu$.

Milward et al. have developed a double-beam system as follows (THORPE, MILWARD, HAYWARD and YEWEN [1969]). It is arranged that reference (A) and sample (B) channels are allowed to fall on the detector for equal and alternate periods of time with an equal blank period between each, as illustrated in Fig. 2.6. If this period is of duration, say, $\frac{1}{4}f$, then it is obvious that a measure of the difference $(A-B)$ between the two channels is given by the amplitude of the periodic signal of frequency f. Similarly a measure of the sum of the two channels $(A+B)$ might be obtained from the amplitude of the periodic signal of frequency $2f$. The square waveforms shown in Fig. 2.6 may be expressed by complex Fourier series. The amplitudes of the fundamental wave and its second harmonic are related to the amplitudes A and B as follows:

$$C_1 = (\pi/\sqrt{2})^{-1}(A-B),$$
$$C_2 = (2\pi)^{-1}(A+B).$$

These simultaneous equations can be solved to give

$$A = \pi[C_2+C_1/\sqrt{2}],$$
$$B = \pi[C_2-C_1/\sqrt{2}].$$

C_1 and C_2 are proportional to the outputs from the channels f and $2f$ respectively, for the same frequency bandwidth and for a detector of uniform frequency response. The electronic system devised for this purpose is depicted in Fig. 2.6.

For interferometric spectroscopy, Fourier transform is a troublesome process and such a system as the double-beam type causes more trouble with complicated Fourier transform calculations. So interferometric spectrometers with a double-beam system are not yet widely used.

2.2.3. *Real-time on-line computers*

Interferometric instruments have the advantages described above, but have the disadvantage that the propriety of the measuring condition is found only after vast calculations of Fourier transforms.

In conventional Fourier spectroscopy spectral intensity $I(\nu)$ is calculated after the interferogram has been measured and the spectrogram, i.e., the curve of $I(\nu) : \nu$ is obtained. If a special computer is connected to the interferometer, the next operation may be possible. When the intensity $F(x)$ of

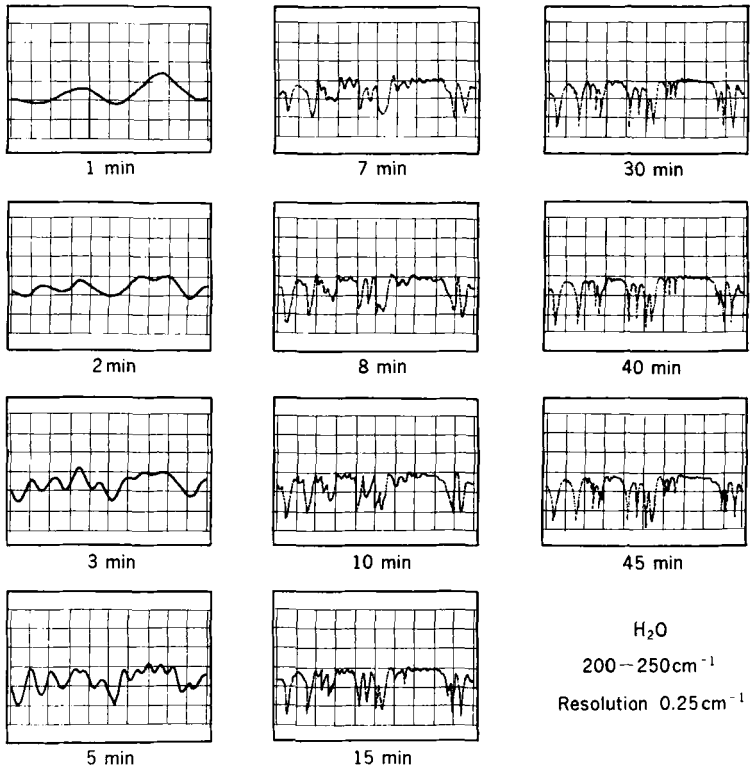


Fig. 2.7. Water vapor absorption curves displayed with a synchroscope during the scan of the interferogram (LEVY et al. [1970]).

the interferogram at the path difference x is obtained, the values of $F(x)\cos(2\pi\nu x)\Delta x$ for all ν are calculated and these values are added to the value of $\sum_{x=0}^{x-\Delta x} F(x)\cos(2\pi\nu x)\Delta x$ stored in the computer and then the results are stored again. This process is repeated for each path difference. After the interferogram has been measured, the values stored in the computer are recorded and give the spectrogram. If these values are displayed with a synchroscope while scanning the interferogram, the curve displayed gives the spectrogram for the maximum path difference x. An interferometer with such a real-time on-line computer, which may be called a kind of multichannel spectrometer, has been developed in our laboratory (YOSHINAGA, FUJITA, MINAMI et al. [1966]). Fig. 2.7 shows the spectra of far infrared water vapor absorption at various path differences found by Milward and his coworkers (LEVY, MILWARD, BRAS and LE TOULLEC [1970]) using the same method as ours. The fine structure of the spectrum can be seen more and more with

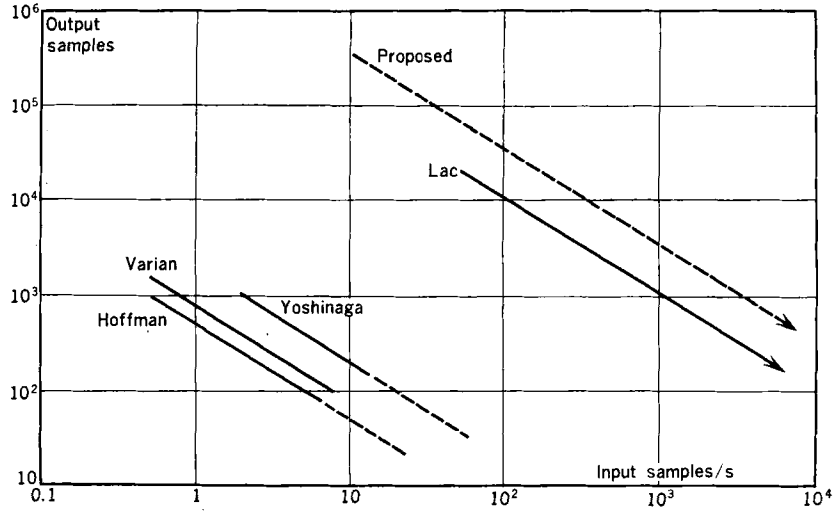


Fig. 2.8. Number of output samples for input sample rate of various real-time on-line computers (CONNES et al. [1970]).

the increase of path difference. Connes and his co-researchers have constructed a large computer of real-time on-line type and are constructing another larger one (CONNES and MICHEL [1970]). Fig. 2.8 shows the characteristics, i.e., the relation between input/sec and output of various real-time on-line computers constructed until now. The curve 'Lac' in the figure is for the computer already constructed and the broken line for the computer under construction by Connes et al. The curve given in Fig. 2.9 is the spectra of N_2O in the infrared region obtained by Connes with the computer and shows very high resolution (CONNES and MICHEL [1970]).

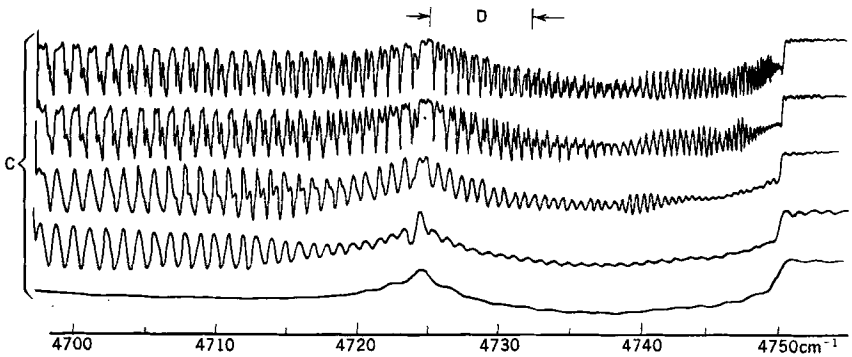


Fig. 2.9. N_2O spectra at various path differences with a real-time on-line computer (CONNES et al. [1970]).

Now we may understand that real-time on-line computers can not only avoid the disadvantage of interferometric spectrometers described above but can catch the rough spectrum immediately after starting the scan of the interferogram, compared with a dispersion type spectrometer with which it takes a long time to see the spectrum for the whole wavelength region in the far infrared region.

The computer memory needs to be large for obtaining a spectrum of high resolution. Because an interferogram can not be scanned so rapidly to avoid the decrease of signal-to-noise ratio of the interferogram, the speed of calculation does not need to be so high even for the computer with a large memory. So a big real-time on-line computer for Fourier spectroscopy seems to be not so difficult to construct in light of the recent development of elements of electronic computers. A very simple computer is satisfactory for an interferometer of the same resolution as that obtained with a dispersion-type spectrometer. As described above, interferometric spectrometers in the far infrared region have advantages which dispersion-type spectrometers have not, and become more and more useful in academic and also industrial measurements.

§ 3. New Spectroscopic Methods

In conventional spectroscopic methods, radiation sources and detectors are non-selective for some wavelength region. So some device to select monochromatic radiation in continuous radiation emitted from the radiation source, for example, a monochromator, is necessary. If a monochromatic and tunable (for wavelength) radiation source or a monochromatic and tunable detector could be developed, the device described above would not be needed, and spectroscopic methods would be very simple. Of course, such a monochromatic and tunable detector is available for emission and absorption spectroscopic methods, and such a monochromatic and tunable radiation source is available only for absorption spectroscopic methods.

In this paragraph, the present situation of the development of such detectors and radiation sources will be described, and their application for spectroscopic measurements will be explained.

3.1. MONOCHROMATIC AND TUNABLE DETECTORS

3.1.1. *InSb cyclotron resonance detector*

The conduction and valence bands of InSb split into discrete Landau levels with the application of a magnetic field. The energy of Landau levels is given by the expression

$$E = \frac{\hbar^2 k^2}{2m^*} + (N+\tfrac{1}{2})\frac{\hbar eH}{m^* c}, \tag{3.1}$$

where m^* is the effective mass of electrons, H is the magnetic field, and k and N are quantum numbers (N shows the number of Landau levels). The transition between successive Landau levels (i.e., $\Delta N = \pm 1$) gives the cyclotron resonance transition, of which the frequency ω_c is as follows:

$$\omega_c = eH/m^* c. \tag{3.2}$$

If InSb material contains impurity atoms, the energy levels of impurity atoms which belong to each Landau level, may appear. So the cyclotron resonance frequency of the transition between Landau levels or two impurity levels belonging to successive Landau levels is proportional to the magnetic field.

If the spin effect is considered, $g_c \beta H M_j$ should be added to the right side of eq. (3.1) as a third term. g_c is the effective g factor, $\beta = e\hbar/2m^*$ is the Bohr magneton, and M_j is a quantum number, being $\frac{1}{2}$ or $-\frac{1}{2}$. This spin effect is important for the tunable lasers described below.

Fig. 3.1 (YAMAMOTO and YOSHINAGA [1969]) is a schematic diagram of energy levels of InSb with hydrogen like impurity atoms in a magnetic field and of optical transitions among those levels. The transition B shows

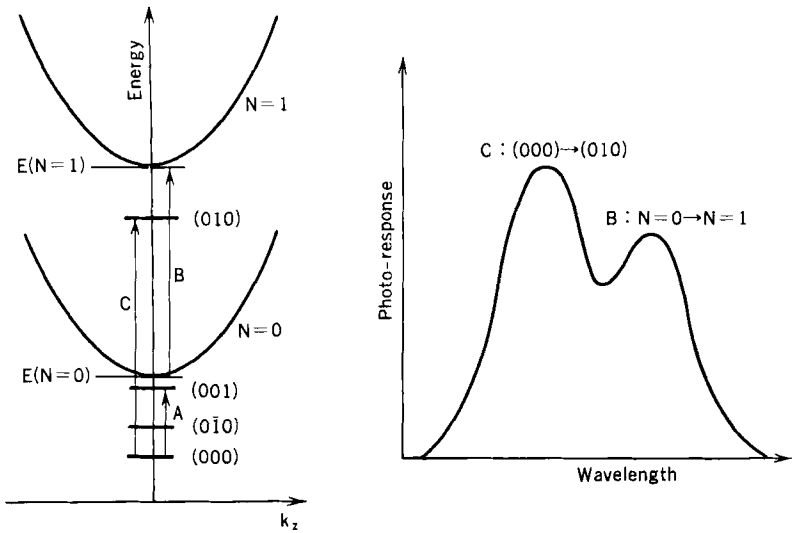


Fig. 3.1. Schematic diagram of energy levels of InSb, optical transitions (left), and spectral photo-response (right) (YAMAMOTO et al. [1969]).

the cyclotron resonance absorption between Landau levels ($N = 0$ and 1) in the conduction band, and the transition C shows the cyclotron resonance absorption from the ground state of the impurity levels. The transition A gives absorption in the millimeter wave region from the ground state to excited impurity levels. The curve on the right gives the spectral photo-response of transition B and C in the submillimeter region.

InSb detector elements must be cooled with liquid helium and be used with a magnetic field. The construction of the cryostat for such a purpose is shown in Fig. 3.2.

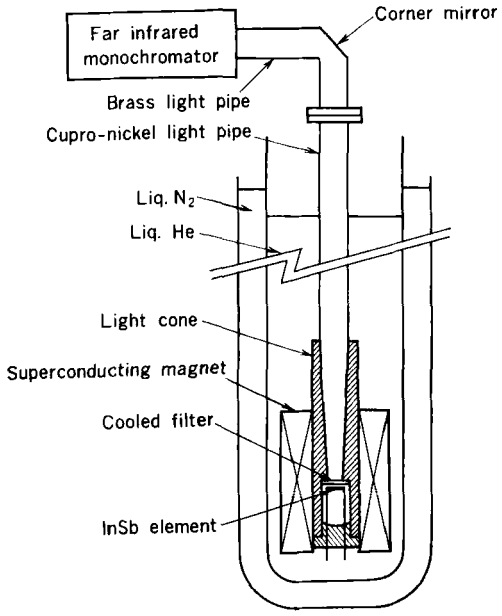


Fig. 3.2. Cryostat for InSb photoconductive detector (YAMAMOTO et al. [1969]).

Fig. 3.3(a) (YAMAMOTO and YOSHINAGA [1969]), shows the normalized photo-response for various magnetic fields. The width of photo-response becomes narrower with an increase of magnetic field, because Landau levels become sharper in a higher magnetic field. The photo-response in a weak magnetic field given in Fig. 3.3(b) shows diffuse resonance and finally the maximum of photo-response can not be seen (Putley detector). The photo-response shifts toward shorter wavelengths in a higher magnetic field. The maximum photo-response is at 35 μm for 50 kG, and at 17 μm for 100 kG. Of course an InSb crystal has lattice vibrations at about 54 μm, so an InSb detector is not available in this wavelength region.

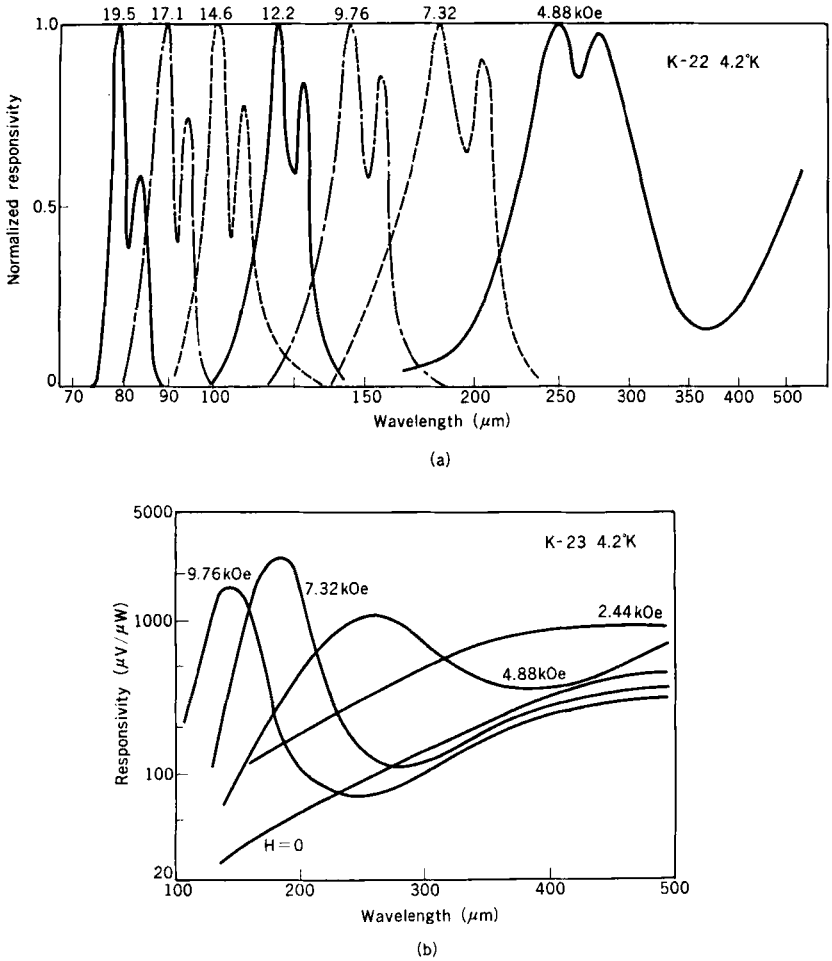


Fig. 3.3. Spectral responsivity of InSb detector at various magnetic fields. (a) strong field (normalized), (b) weak field (YAMAMOTO et al. [1969]).

In a monochromatic detector, it is desirable that the width of the spectral photo-response be as narrow as possible. Fig. 3.4 (YAMAMOTO and YOSHINAGA [1969]) shows the photo-response at various temperatures. At lower temperatures, the transition B diminishes and the photo-response curve becomes narrower. The photo-response for two InSb samples with different impurity concentrations (1.4–2.8×10^{13} and 2.4–2.7×10^{14} cm^{-3}) is given in Fig. 3.5 (YAMAMOTO and YOSHINAGA [1969]), which shows that the purer sample has a much narrower photo-response. So it is better to

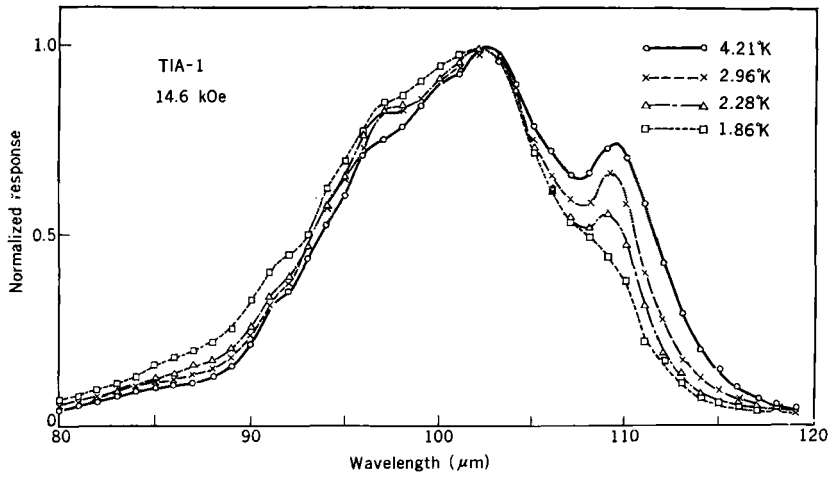


Fig. 3.4. Spectral photo-response of InSb detector at various temperatures in constant magnetic field (YAMAMOTO et al. [1969]).

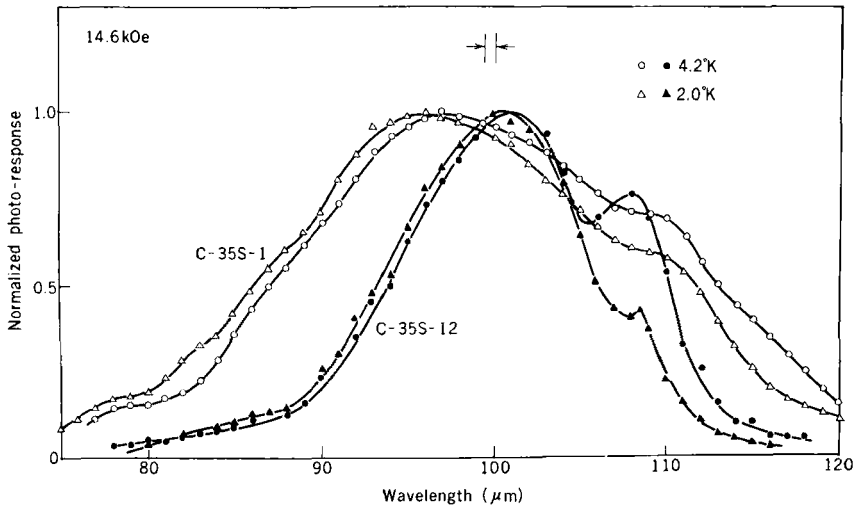


Fig. 3.5. Spectral photo-response for InSb samples of two different carrier concentrations and two temperatures (YAMAMOTO et al. [1969]).

use purer InSb material at lower temperatures for a monochromatic detector.

The present InSb cyclotron resonance detector is not so monochromatic for spectroscopic use without a monochromator. But such a detector has some benefit compared with a non-selective detector. The following two spectroscopic devices will be compared (YAMAMOTO [1971]):

(a) grating, cut-off filters of higher order spectra and non-selective detector,

(b) grating and InSb detector.

If we use a thermal radiation source, the energy at wavelength λ is given by the following Rayleigh Jeans formula:

$$B_\lambda = C_1 C_2^{-1} \lambda^{-4} T. \tag{3.3}$$

So,

$$B_m = m^4 B_1, \tag{3.4}$$

where m is the number of higher order spectra, B_1 and B_m are the energy of the first and the mth order spectra respectively. The energy passing through a monochromator is as follows:

$$E = BslT_\lambda A f^{-2} \Delta\lambda, \tag{3.5}$$

where s is the slitwidth, l is the slitlength, T_λ is the transmittance of the monochromator at λ, A is the grating area and f is the focal length. $\Delta\lambda$ is given by $sd \cos\theta/mf$, where d is the grating constant and θ is the diffraction angle. T_λ is assumed to be constant. Then,

$$E_m = m^3 E_1. \tag{3.6}$$

The output of a detector is expressed as follows:

$$V_\lambda = R_\lambda E_\lambda,$$

where R_λ is the spectral response. For a thermal detector,

$$V_m = R_\lambda E_m = R_\lambda m^3 E_1 = m^3 V_1.$$

For the total output of the detector,

$$\begin{aligned} V &= V_1 + V_2 + V_3 \\ &= (1^3 + 2^3 + 3^3 + \ldots) V_1. \end{aligned} \tag{3.7}$$

For a Putley type detector,

$$\begin{aligned} R_\lambda &= c\lambda^2 (c\text{: proportional constant}) \\ R_m &= C(\lambda_1/m)^2 = R_1/m^2. \end{aligned}$$

So, $V_m = (R_1/m^2)\, m^3 E_1 = mV_1$

$$V = (1 + 2 + 3 + \ldots) V_1. \tag{3.8}$$

An InSb detector is assumed to have some photo-conductive effect simultaneously with cyclotron resonance absorption. The output of an InSb detector is given as follows:

$$nV_1 + V_2 + V_3 + \ldots = (n+2+3+\ldots)V_1 \tag{3.9}$$

where n is the increase of photo-response by resonance absorption. To keep the influence by higher order spectra less than 1 %,

$$n \geqq 100(2+3+\ldots). \tag{3.10}$$

Now let a thermal detector, for example, a Ge bolometer, and an InSb detector be compared with each other. For the Ge bolometer,

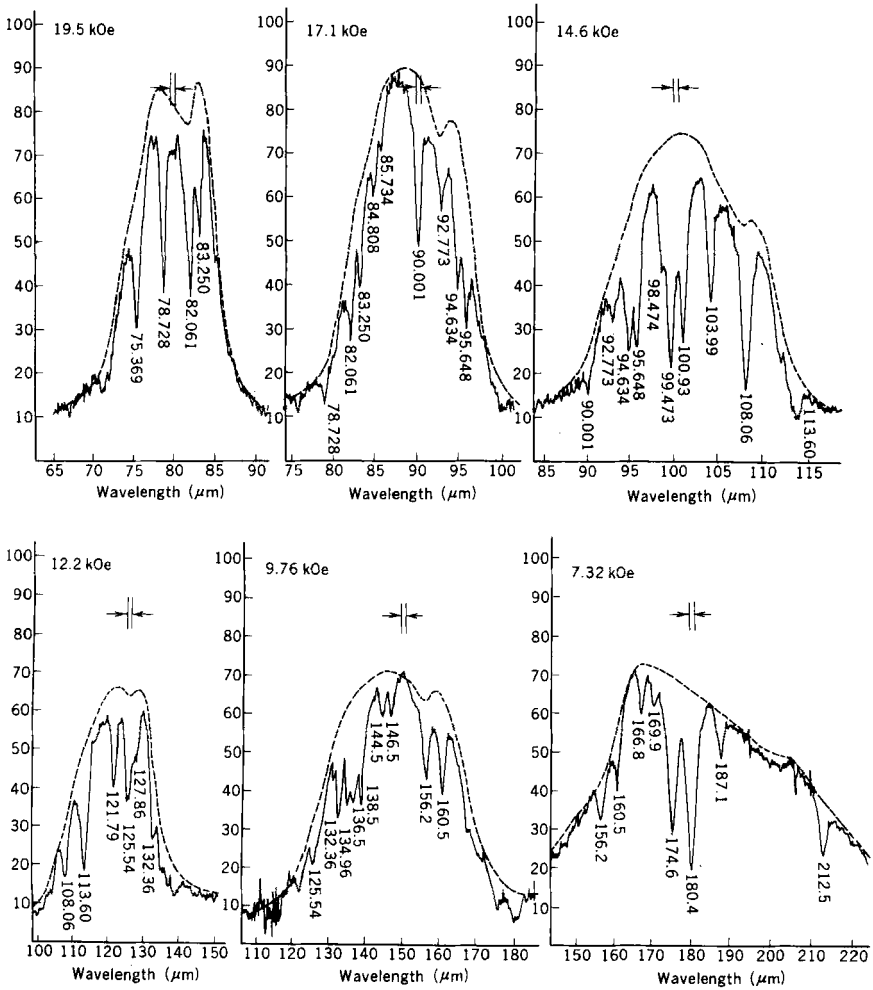


Fig. 3.6. Spectra of water vapor absorption in the far infrared region in various magnetic fields (YOSHINAGA et al. [1969]).

$$p_1 \geqq 100(p_2 2^3 + p_3 3^3 + \ldots), \tag{3.11}$$

where p_1, p_2, p_3 are the transmittance of the cut-off filter for the first, the second and the third order spectra respectively. For the InSb detector, eq. (3.10) must be satisfied. Of course, T_λ is not constant, especially for echelette gratings, and the transmittance of light pipes is lower in the shorter wavelength region. But these factors influence both detectors in the same way. As a conclusion, a Ge bolometer is desired to have as low a NEP value as possible to cover the energy loss by cut-off filters. For an InSb detector, its photo-conductivity should be as small as possible, and it is desired to use InSb material of the lowest carrier concentration at lower temperatures.

Fig. 3.6 (YOSHINAGA and YAMAMOTO [1969], YAMAMOTO [1971]) is the spectrum of water vapor obtained with a very simple monochromator and an InSb detector in constant magnetic fields. The resolution of these spectra is not so high due to the slow optical system of the monochromator, and the second and higher order spectra can not be seen at all without any cut-off filters except two scatter plates. The optical system of a new monochromator under construction, shown in Fig. 3.7, has only two scatter plates and no other cut-off filters, and is very simple compared with conventional grating spectrometers using non-selective detectors. The speed of the optical system is $F = 3$, which gives the optimum condition to collect the most radiation energy in the detector element, considering the constant spectral width and the transmittance of the light pipe. The tuning of the InSb detector by magnetic field is to be synchronized with the wavelength scanning of the monochromator. The spectrogram shown in Fig. 3.8 is an example of the spectrum of water vapor in the far infrared

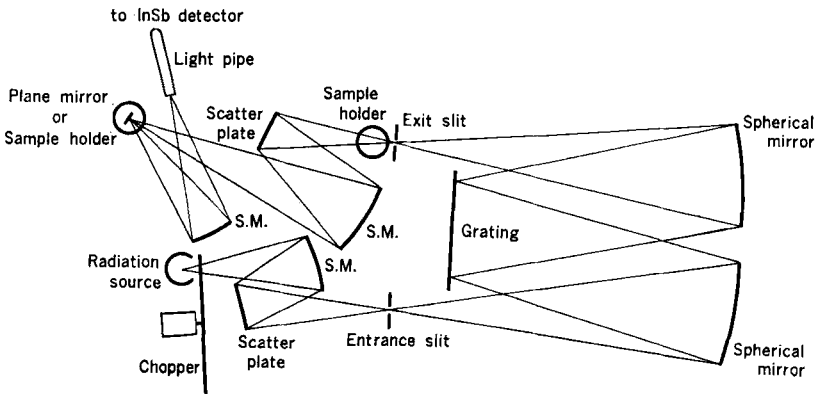


Fig. 3.7. Optical system of a simple far infrared spectrometer using InSb detector.

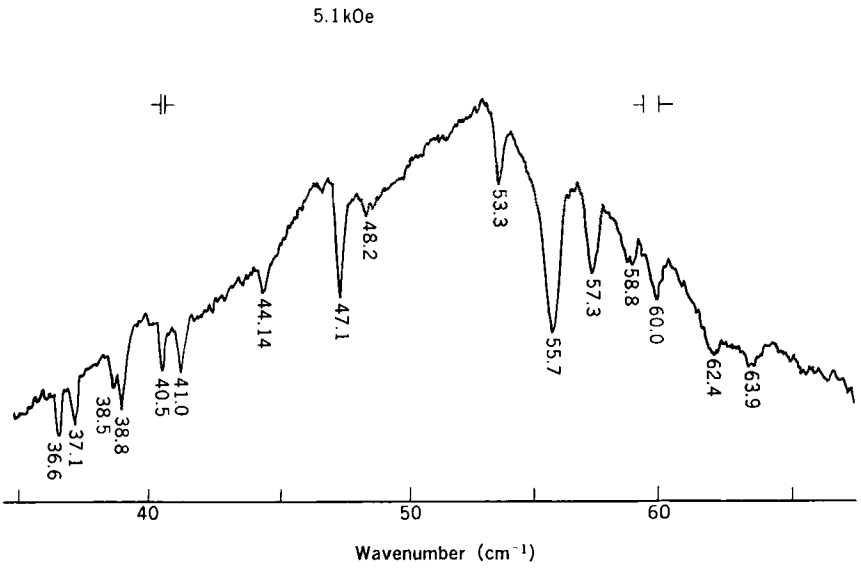


Fig. 3.8. Spectrum of water vapor absorption near 40 cm^{-1}.

region obtained with the monochromator and an InSb detector in a constant magnetic field. The highest resolution is observed to be 0.3 cm^{-1} at 40 cm^{-1} and 0.4 cm^{-1} at 90 cm^{-1}, which are very satisfactory values with such a simple far infrared spectrometer.

3.1.2. *Josephson junction detector*

This is a detector using the tunneling effect through a thin oxide barrier between two superconducting films (RICHARDS [1970b]). The tunneling current through such a junction can be divided into two parts: simple particle tunneling and Josephson tunneling. The first is the tunneling of individual electrons or quasiparticles, which usually dominates the dc current when voltages comparable to the energy gap are applied. Josephson showed that correlated superconducting pairs of electrons will also tunnel and should allow current flow with no associated voltage drop. This Josephson current is extremely sensitive to electromagnetic fields in a way which permits the construction of a variety of sensitive detectors of radiation at frequencies up to the far infrared region.

The Josephson tunneling effect is described as follows. The wave function for the superconducting state can be written in the simple form $\psi = \psi_0 e^{i\theta}$, where the phase factor is a function of both position and time. Josephson's lossless pair tunneling current depends on $\Delta\theta$, the phase difference between

the wave functions for the superconductors on the two sides of the tunnel junction

$$I = I_0 \sin \Delta\theta. \tag{3.12}$$

Here I_0 is the maximum zero-voltage current which can be carried by the junction. The final potential drop V across an oxide barrier occurs when I_0 is exceeded. By the quantum theory of tunneling, a time dependence of the phase difference is given as follows:

$$\hbar \mathrm{d}(\Delta\theta)/\mathrm{d}t = 2eV. \tag{3.13}$$

If the voltage V is constant in time, eq. (3.13) can be integrated to give $\Delta\theta = 2eV_0 t/\hbar + \theta_0$. Eq. (3.12) then predicts an alternating current flow across the barrier at the frequency $\omega_0 = 2eV_0/\hbar$. An alternating voltage $V_1 \cos \omega_1 t$ is induced across the barrier in addition to the steady voltage V_0. Eq. (3.13) then gives

$$\Delta\theta = 2eV_0 t/\hbar + 2eV_1 \sin \omega_1 t/\hbar\omega_1 + \theta_0;$$

so from eq. (3.12),

$$I(t) = I_0 \sin [2eV_0 t/\hbar + (2eV_1/\hbar\omega_1) \sin \omega_1 t + \theta_0]. \tag{3.14}$$

Choosing the value $\theta_0 = \frac{1}{2}\pi$, which corresponds to a maximum zero-voltage current ($V_0 = \omega_0 = 0$), and using standard trigonometric identities, $I(t)$ is obtained as follows:

$$I(t) = I_0 \sum_{n=-\infty}^{\infty} J_n \left(\frac{2eV_1}{\hbar\omega_1}\right) \cos (\omega_0 + n\omega_1)t. \tag{3.15}$$

Here J_n is Bessel's function of order n. The junction is thus a nonlinear device in which the Josephson currents beat with the induced ac signal. In order to operate the junction as a detector, the zero frequency beats, which occur when $\omega_0 + n\omega_1 = 0$, are measured. Then

$$I_{\mathrm{dc}} = I_0(-1)^n J_n(2eV_1/\hbar\omega_1). \tag{3.16}$$

A lossless contribution to the dc current appears whenever the voltage is adjusted so that the ac Josephson frequency $\omega_0 = 2eV_0/\hbar$ equals harmonics of the frequency ω_1.

Since $n = 0$ for $V_0 = 0$, the condition for direct current, $2eV_0/\hbar = \pm n\omega_1$ is automatically satisfied for all ω_1. So a broad band detector can be obtained. When the dc voltage is adjusted to a cavity mode, $2eV_0 = \hbar\omega_c$, where ω_c is a resonant cavity frequency, the Josephson current is found to be able to excite the cavity mode with sufficient amplitude. If the junction

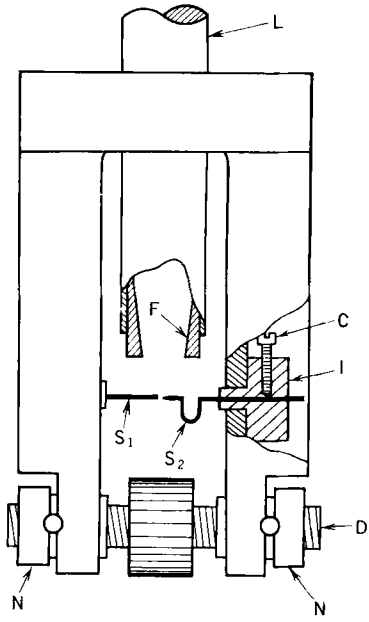


Fig. 3.9. Schematic diagram of the construction of Josephson junction detector (RICHARDS [1970b]).

is biased so that $\omega_0 \approx \omega_c$, the Josephson radiation excites the cavity mode and is fed back to the junction. This mechanism has been used by Richards and Sterling to construct a narrow band detector. Fig. 3.9 illustrates the apparatus for holding and adjusting the Josephson contact between Nb wires S_1 and S_2. The spectral response of such a detector measured using a blackbody source and a far infrared Fourier transform spectrometer is shown in Fig. 3.10. Although the peak is not resolved, the feedback narrowing is clearly seen. Richards and his coworker reported NEP $\leqq 10^{-14}\mathrm{W}/\sqrt{\mathrm{Hz}}$ and the width of spectral response $\sim 10^{-2}$ cm^{-1} (RICHARDS and STERLING [1969]). If the resonant cavity frequency ω_c could be controlled, a monochromatic tunable detector might be possible to develop. But the continuous shift of ω_c is difficult and has not yet been achieved.

The high frequency limit for the response of Josephson detectors is related to the critical temperature T_c of the superconductors. For Nb, T_c is 8° K and the frequency limit becomes 50 cm^{-1}. For Nb_3Sn, these values are 17° K and 100 cm^{-1} respectively.

As described above, ideal monochromatic and tunable detectors, available to far infrared spectroscopy without any monochromator, have not yet been developed.

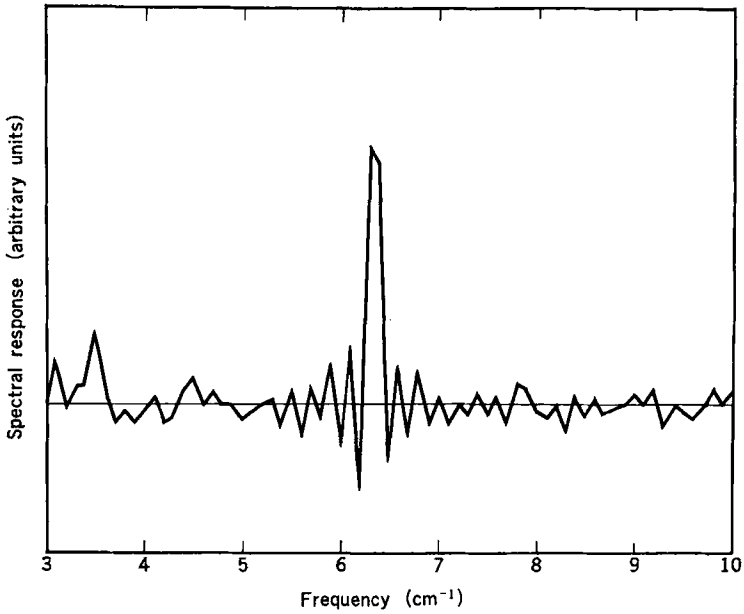


Fig. 3.10. Spectral response of Josephson junction detector (RICHARDS et al. [1969]).

3.2. MONOCHROMATIC AND TUNABLE SOURCES

Lasers are the only source to be expected as a monochromatic and tunable radiation source at present. The spectral width of laser radiation is very small, but the wavelength is fixed in general, a serious disadvantage for use as a spectroscopic source. Recently the tuning of the wavelength has been studied, and many laser wavelengths can be tuned in some wavelength regions, most of which are below 15 μm, not in the far infrared region. But such lasers are important to create a far infrared tunable laser as described below. So tunable lasers in any wavelength region in which spectral widths are very narrow, will be explained.

3.2.1. *Tuning methods of laser wavelengths*

(1) Tuning by cavity control. For gas lasers lasing in one mode, laser wavelengths can be tuned continuously through the Doppler width of the laser line by adjusting, for example, the mirror distance of a Fabry-Perot cavity. As the Doppler width is reciprocally proportional to the lasing wavelength, the tuning range is very narrow in the long wavelength region. So the tuning range is not satisfactorily wide and the laser is not available for absorption measurement, except in the case where the spectral line

to be measured falls into the Doppler width. Also, dye lasers can be tuned continuously by using prisms, gratings or etalons of Littrow arrangement as one of the two mirrors comprising a Fabry-Perot resonator. By this method tunable laser wavelengths in the whole visible region with the combination of many kinds of dye lasers can be obtained. An especially very narrow spectral width of 0.05 A has been observed for the resonator with a few etalons (WALTER and HALL [1970]).

(2) Tuning by shifting the energy levels related to laser transition. Of course, a magnetic field can shift the energy levels (Zeeman effect), but the amount of shift is too small for a spectroscopic radiation source. Diode lasers using semiconductors can be tuned in a wide range by changing the temperature, the pressure, the magnetic field or the injection current. An InSb laser using a spin flip tuned by a magnetic field is very interesting as described below.

(3) Mixing of two laser lines. The difference frequency between two laser lines, the wavelengths of which are in the visible or infrared region, comes into the far infrared region. If one of two laser lines is tunable, a tunable far infrared laser can be obtained by mixing with nonlinear material.

(4) Tuning by optical parametric method. Tunable stimulated radiation in the far infrared region can be obtained by parametric methods, for example, using the polariton effect.

(5) Semi-tunable lasers using rotational transitions of molecules. Many rotational lines in the far infrared region can be lased by optical pumping with vibration-rotational lines in the middle infrared region. Of course these rotational lines are distributed discretely.

3.2.2. *Diode lasers*

There exist diode lasers using many kinds of semiconductors. But $Pb_{1-x}Sn_xSe$ and $Pb_{1-x}Sn_xTe$ lasers are the most interesting as tunable lasers.

The wavelength of the laser line depends upon the energy gap of the semiconductor, which changes with the concentration of Sn (HARMAN, CALAWA, MELNGAILIS and DIMMOCK [1969]). The tunable range of a $Pb_{1-x}Sn_xSe$ laser ($0 \leqq x \leqq 0.276$) is from 8 to 31.2 μm. Fig. 3.11 shows the wavelength dependency on temperature. For a composition with $x \leqq 0.10$, the temperature coefficient is very low and is 5×10^{-4} eV/K below 30° K. The coefficient is negative for a composition with $x > 0.19$. Similar results are reported on a $Pb_{1-x}Sn_xTe$ laser (NORTON, CHIA, BRAGGINS and LEVINSTEIN [1971]). The tunable range of this laser

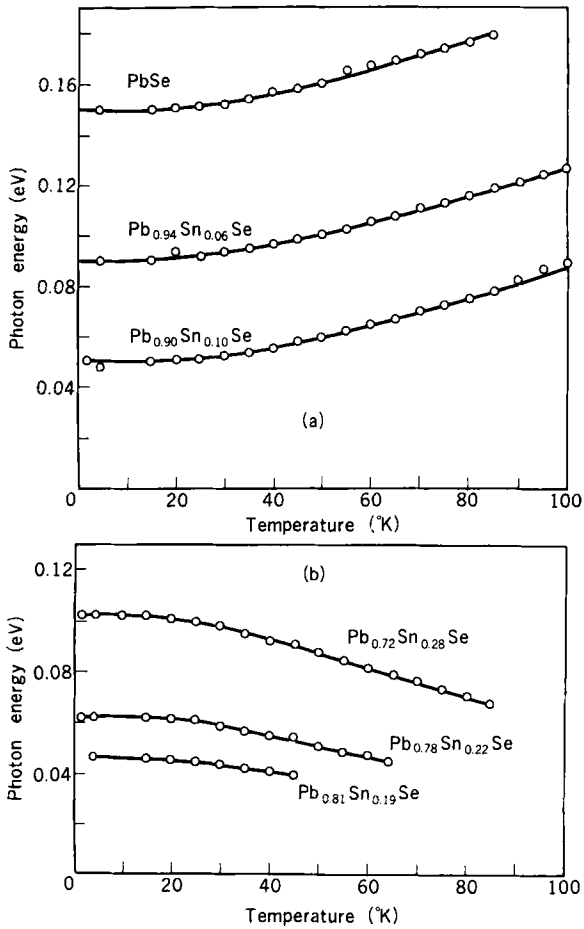


Fig. 3.11. Temperature dependence of $Pb_{1-x}Sn_xSe$ laser energy for various x values (HARMAN et al. [1969]).

($0 \leqq x \leqq 0.315$) is from 6.5 to 31 μm. For a composition with $x = 0.1$, the wavelength variation is approximately 4×10^{-4} eV/K. So the temperature must be regulated within $\pm 0.01°$ K for a wavenumber stability of ± 0.04 cm^{-1} at 10 μm. With care, $\pm 0.001°$ K can be maintained between 10° and 300° K using a liquid helium cryostat with temperature adjustments provided by a simple resistance heater wound on the diode mount. The mode spacing for a 0.3-mm cavity is about 0.03 μm at 10 μm. This resolution and tuning range makes it possible to sample a large number of points in the broadband infrared absorption spectra of gases.

A $Pb_{1-x}Sn_xTe$ laser tuned by injection current has shown very remarkable

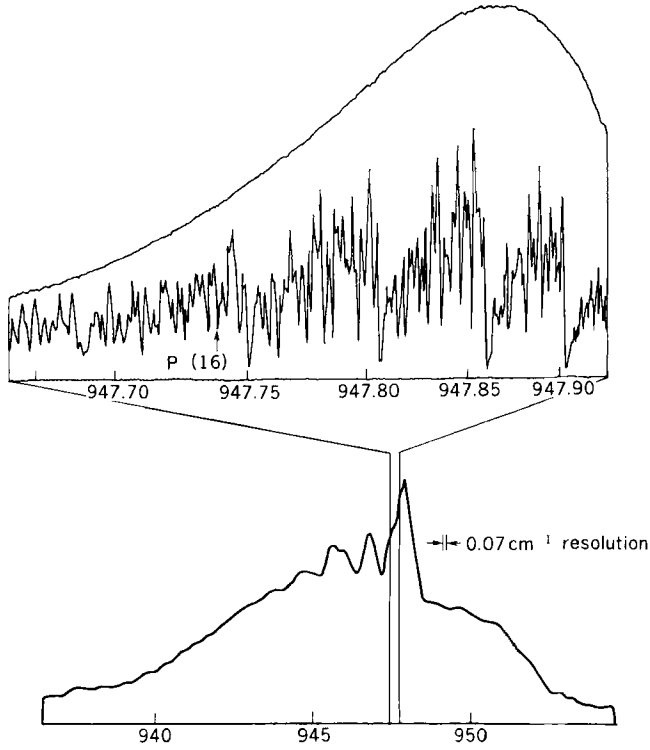


Fig. 3.12. Spectra of the ν_3 band of SF_6 molecule measured with a grating spectrometer and tunable $Pb_{0.88}Sn_{0.12}Te$ diode laser (HINKLEY et al. [1971]).

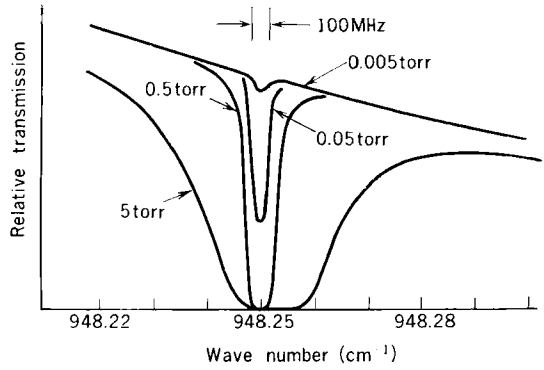


Fig. 3.13. Absorption spectra of a NH_3 line taken at various gas pressures with tunable $Pb_{0.88}Sn_{0.12}Te$ diode laser (HINKLEY et al. [1971]).

results on middle infrared spectroscopy (HINKLEY [1970], HINKLEY and KELLEY [1971]). The laser element, the size of which is $1.3 \times 0.5 \times 0.2$ mm^3, is oscillated (cw) in the 10.6 μm region with the injection current from 600 to 1800 mA under cooling by liquid helium. The current dependency of a $Pb_{0.88}Sn_{0.12}Te$ laser is from 10 to 100 MHz/mA, which depends upon the size of the laser element and the junction depth. The line width is very narrow, i.e., less than 54 kHz (1.8×10^{-6} cm^{-1}) at 10.6 μm, and the output is 0.24 mW. A SF_6 molecule has an exceedingly complex spectrum in the middle infrared region, which cannot be resolved in the 950 cm^{-1} region even with a high-quality grating spectrometer, but which is resolved by a diode laser scan. Fig. 3.12 shows the spectrum of the ν_3 band of a SF_6 molecule. The bottom curve is the absorption curve near 950 cm^{-1} at room temperature by a grating spectrometer (SF_6 pressure: 0.1 torr, cell length: 25 cm). The top curve is a part of the bottom spectrum by the heterodyne detection with P(16)CO_2 laser line at 947.738 cm^{-1} and the tunable diode laser (the pressure: 0.1 torr, cell length: 10 cm). The resolution of this curve is remarkably high, 3×10^{-6} cm^{-1}. Such high resolution measurement has the following advantage. Fig. 3.13 shows absorption profiles of the strong sP(1, 0) line of NH_3 at several pressures measured by the same detection as described above. As the NH_3 pressure is increased to 0.5 torr, the cell 10 cm long becomes essentially opaque to radiation near the line center. With a grating spectrometer of 0.1 cm^{-1} resolution, a pressure of 5 torr in a cell 200 cm long produces the same maximum absorption as that shown at 0.05 torr in a cell 10 cm long, where the amount of NH_3 in the optical path is 2000 times smaller.

Diode lasers can be tuned by the pressure applied to the laser element. A PbSe laser is tunable in the range from 7.5 to 22.3 μm by a hydrostatic pressure of 0–14 K barr (BESSON and PAUL [1968]).

Diode lasers can be tuned also by a magnetic field (CALAWA, DIMMOCK, HARMAN and MELNGAILIS [1969]). The conduction and valence bands of $Pb_{1-x}Sn_xSe$ are split by a magnetic field, and T_1, T_2 and T_3 laser oscillations occur as seen in Fig. 3.14. This figure shows also the tuning of T_1, T_2 and T_3 for $x = 0.19$, 0.22 and 0.28 by a magnetic field. T_1 for $x = 10.9$ is tunable up to 34 μm in 80 kG, of which the wavelength is the longest one of diode lasers.

3.2.3. *Dye laser*

Dye lasers are tuned by cavity control. Recently a distributed-feedback dye laser has been reported to produce a very narrow tunable laser line (SHANK, BJORKHOLM and KOGELNIK [1971]). The schematic diagram of the

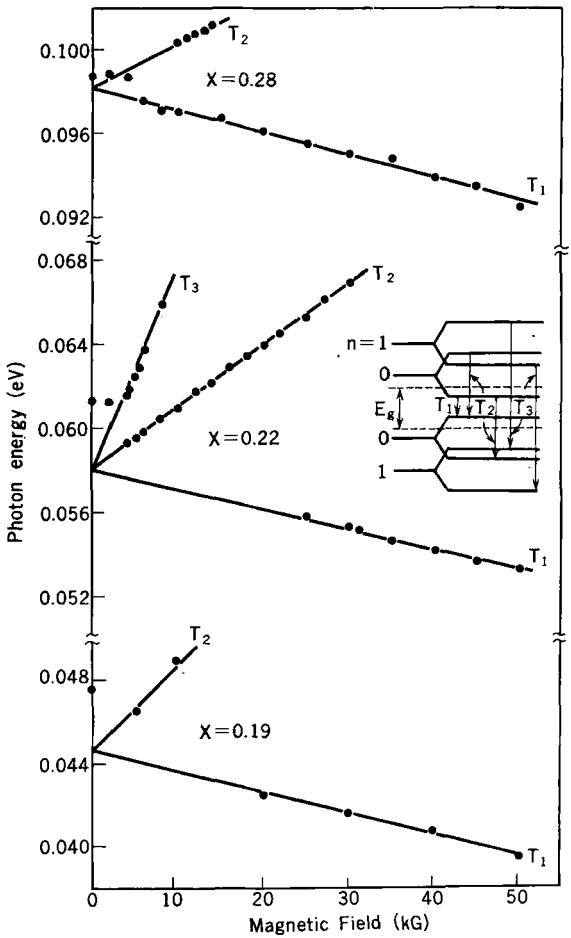


Fig. 3.14. Magnetic field dependence of $Pb_{1-x}Sn_xSe$ diode laser for various x values (CALAWA et al. [1969]).

experimental setup of mirrorless distributed-feedback laser devices is illustrated in Fig. 3.15. The wavelength of the dye laser is given by,

$$\lambda_L = n_s \lambda_p / \sin\theta, \tag{3.17}$$

where n_s is the solvent index of refraction at the lasing wavelength λ_L, and λ_p is the pump wavelength. λ_L is tuned by either n_s or θ, which is varied by changing the mirror positions.

A dye laser using 3×10^{-3} M Rhodamine 6G shows an angle dependency of $d\lambda/d\theta = 80$ A at 0.6 μm and a tuning range of 640 A. The solvent index is easily varied over the range from 1.33 to 1.55 by proportionate mixing

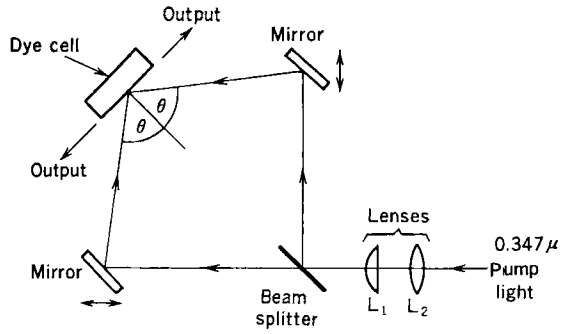


Fig. 3.15. Schematic diagram of the experimental setup of distributed-feedback dye laser devices (SHANK et al. [1971]).

of methanol and benzol alcohol. With reduced pump power, an apparent single mode operation is obtained with a linewidth of less than 0.01 A. Near 5900 A the peak output power of the laser is 36 kW, the sum of the powers out of both ends of the laser cell, with about 180 kW peak pump power. The dye laser pulse has a slight superimposed structure and a duration of approximately 10 nsec.

3.2.4. *Spin flip Raman lasers*

Raman scattering from a mobile carrier in semiconductors is also very interesting for tunable lasers. Of course, the energy and momentum conservation relations must be satisfied as follows:

$$\hbar\omega_s = \hbar\omega_0 \pm \hbar\omega_e, \tag{3.18}$$

$$\boldsymbol{k}_s = \boldsymbol{k}_0 \pm \boldsymbol{k}_e, \tag{3.19}$$

where ω_s and ω_0 are the scattered and incident light frequencies, $\hbar\omega_e$ is the energy of the elementary excitation responsible for the scattering of light, $\boldsymbol{k}_s$ and $\boldsymbol{k}_0$ are the scattered and incident light wave vectors respectively, and $\boldsymbol{k}_e$ is the wave vector of the above elementary excitation. The minus sign applies to Stokes Raman scattering and the plus sign indicates anti-Stokes Raman scattering. For the spin flip transition,

$$\omega_s = \omega_0 \pm n g_c \beta H, \tag{3.20}$$

where g_c is the effective g factor of electrons and β is the Bohr magneton. n shows the number of spin flip transitions.

An InSb spin flip Raman laser of pulse type has been reported (PATEL, SHAW and KERL [1970], PATEL and SHAW [1971]). The carrier concentration n_e of InSb used for this laser is $(1.3\text{–}3)\times 10^{16}$ cm^{-3}. Fig. 3.16 shows the

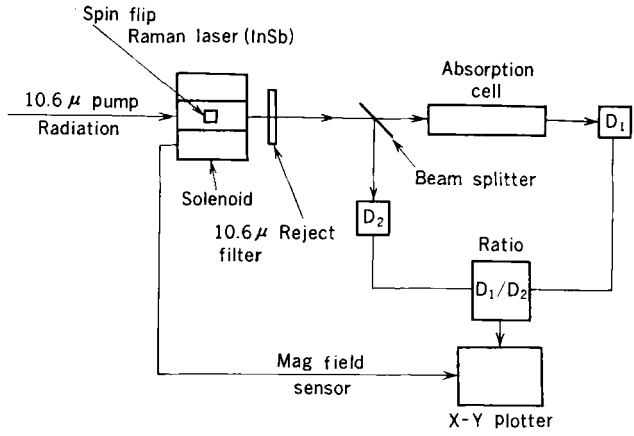


Fig. 3.16. Schematic diagram of the experimental setup for tunable InSb spin flip Raman laser (PATEL et al. [1970]).

experimental setup for using the tunable spin flip Raman laser as an infrared spectrometer source. The peak spin flip Raman laser power is 30–100 W for an input power of 1.5 kW obtained from a Q switched CO_2 laser (10.5915 μm) having a pulse repetition rate of 120 Hz. The tunable range, which depends upon the carrier concentration, is from 11.7 to 13.0 μm for $n_e = 3 \times 10^{16}$ cm^{-3} in a magnetic field from 48 to 100 kG. The pulse width is 30 nsec (3 nsec in a mode-locked operation). The laser linewidth is less than 0.16 cm^{-1} measured with a grating spectrometer of resolution 0.15 cm^{-1}, and the estimated linewidth is less than 0.03 cm^{-1}. The linearity of tunability and resettability is less than 0.026 cm^{-1} at 860 cm^{-1}.

The absoprtion of NH_3 in the range 800–900 cm^{-1} has been measured using the spin flip Raman laser as the source. Fig. 3.17 (a) shows a portion of the absorption spectrum (NH_3 pressure: ~ 10 torr and absorption length: ~ 15 cm) as a function of frequency in the range 846–855 cm^{-1}. The number on the spectrum gives the transition as follows: (1) aP(4, 0), (2) aP(4, 1), (3) aP(4, 2), (4) aP(4, 3), (5) sP(6, 1), (6) sP(6, 2), (7) sP(6, 3), (8) sP(6, 4) and (9) sP(6, 5). sP(6, 1) and sP(6, 2) are resolved and the spacing is seen to be ~ 0.05 cm^{-1}. This confirms the estimate of a spin flip Raman laser linewidth of $\leqq$ 0.03 cm^{-1}. (b) in the figure was taken with a conventional spectrometer having a 15-cm grating, of which the resolution is 0.1–0.2 cm^{-1}. The comparison between (a) and (b) in this figure leaves no doubt about the superior resolution of the spin flip Raman laser spectrometer. In addition, the spin flip Raman laser spectrometer is extremely fast (because of the relatively high, tunable monochromatic power output) and

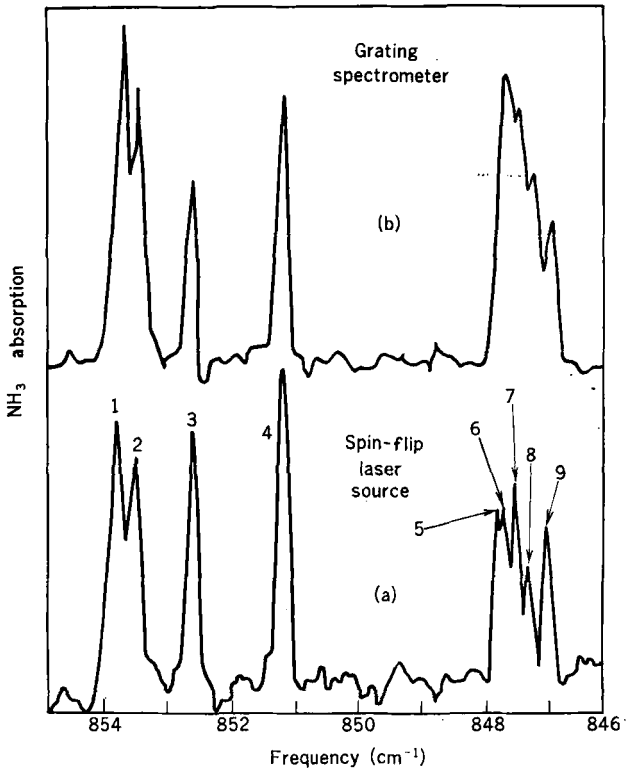


Fig. 3.17. Spectra of NH_3 absorption measured with a grating spectrometer and tunable InSb spin flip Raman laser (PATEL et al. [1970]).

the trace can be taken in < 1 min. Further improvements may be made in high resolution spectroscopy using the very narrow linewidth of the spin flip Raman laser line and heterodyne detection.

The anti-Stokes line which is tunable from 10.0 to 9.4 μm in a magnetic field of 30–65 kG has been observed (SHAW and PATEL [1971]). The output of this anti-Stokes line is about ten times smaller than that of the first Stokes line. Also the second Stokes line has been observed by pumping of a CO laser (5.3648 μm) (PATEL [1971]). The second Stokes line is weaker, but has twice as wide a tunable range as the first Stokes line. High intensity tunable spin flip Raman laser lines using InSb, with carrier concentration $n_e \approx 2 \times 10^{16}$ cm^{-3}, and a pulsed high pressure CO_2 laser (10.6 μm), with peak power of the order of 1 MW in a magnetic field of 25–100 kG, have been reported also (AGGARWAL, LAX, CHASE et al. [1971]). The tunable ranges of the first Stokes line, the second Stokes line and the anti-

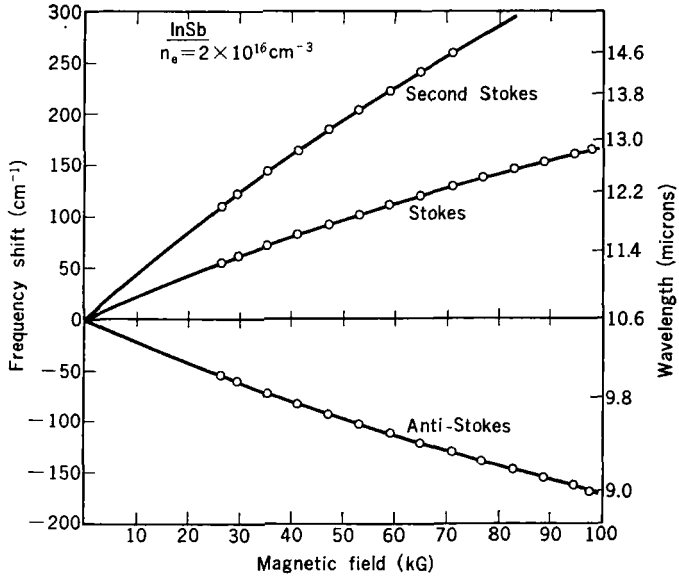


Fig. 3.18. Tuning curves for Stokes, second Stokes and anti-Stokes components of InSb spin flip Raman laser (AGGARWAL et al. [1971]).

Stokes line are 1.2–12.8 μm, 12.0–14.6 μm and 10.0–9.0 μm respectively and their maximum outputs are 1 kW, several W and ~ 30 W respectively. Fig. 3.18 shows the frequency shift of the first Stokes line, the second Stokes line and the anti-Stokes line in various magnetic fields, and Fig. 3.19 illustrates the relative output power of the first Stokes line and the anti-Stokes line at various input laser powers. The output can be seen to increase very rapidly with an increase in the magnetic field.

As a radiation source of spectroscopy, cw operation is more desirable than pulse operation. Already cw spin flip Raman laser lines (the first and second Stokes lines) pumped with a CO laser have been reported (MOORADIAN, BRUECK and BLUM [1970], BRUECK and MOORADIAN [1971]). The output of the first Stokes line is > 1 W and the threshold power is less than 50 mW.

3.2.5. *Difference frequency lasers*

Difference frequency generation in a nonlinear crystal using two laser lines is very important for obtaining tunable far infrared lasers. Most of the tunable lasers described above are in the visible and middle infrared region. Tunable far infrared lasers are possible to obtain by difference frequency generation with a fixed wavelength laser and a tunable laser.

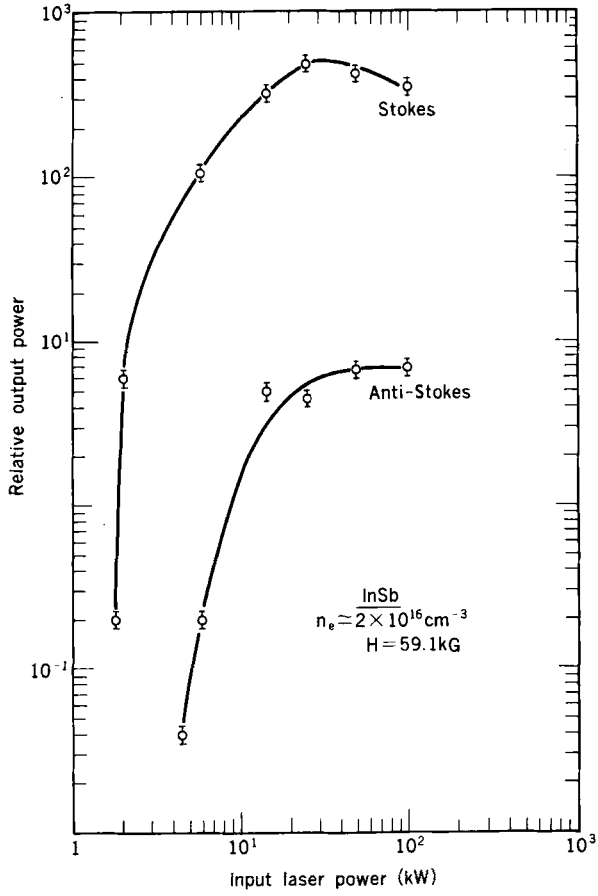


Fig. 3.19. Pump power dependence of peak output power of Stokes and anti-Stokes components of InSb spin flip Raman laser (AGGARWAL et al. [1971]).

The difference frequency power output is given by

$$P_{\nu_3} = \frac{32\pi^2\chi^2\nu_3^2}{3c^3 n_1 n_2 n_3} \frac{P_{\nu_1} P_{\nu_2} T^3}{W^2} I, \tag{3.21}$$

where ν_1 and ν_2 are the frequencies of two incident laser lights, ν_3 is the difference frequency, χ is the effective second-order nonlinearity, $n_{1,2,3}$ are the refractive indices at the frequencies ν_1, ν_2 and ν_3 respectively, W^2 is the cross sectional area of the incident beams at ν_1 and ν_2, T is the power transmission coefficient for each surface of the sample, and I is the coherence factor given by

$$I = \exp(-\alpha_3 l)\frac{1+\exp(-\Delta\alpha l)-2\exp(-\frac{1}{2}\Delta\alpha l)\cos(\Delta k l)}{(\Delta k)^2+\frac{1}{4}(\Delta\alpha)^2}, \quad (3.22)$$

where $\Delta\boldsymbol{k} = \boldsymbol{k}_1 - \boldsymbol{k}_2 - \boldsymbol{k}_3$ is the phase dismatch ($k_{1,2,3}$ are the respective propagation vectors at the three frequencies), $\Delta\alpha = \alpha_1 + \alpha_2 - \alpha_3$ with $\alpha_{1,2,3}$ being the absorption coefficients at the three frequencies of interest.

The magneto plasma effect for phase matched difference frequency generation in InSb crystals of various carrier concentrations in a magnetic field, for many lines of CO_2 laser in 9.6 and 10.6 μm regions, has been reported to produce stimulated radiation output of a few μW in the far infrared region from 10 to 150 cm^{-1} (VAN TRAN and PATEL [1969]). Of course the produced radiation is not truly tunable but is instead a group of discretely distributed lines.

The R_1 line of a ruby laser can be tuned by changing the temperature of the ruby crystal. Tunable stimulated radiation has been obtained by difference frequency generation in quartz or $LiNbO_3$ crystals with a ruby laser at constant temperature and a ruby laser at various temperatures (FARIES, GEHRING, RICHARDS and SHEN [1969]). The crystal of the latter laser is cooled by circulating ethyl alcohol. The tuning range of the stimulated radiation is from 1.2 to 8.1 cm^{-1}, the output is about 2×10^{-2} W and the linewidth is estimated to be less than 0.04 cm^{-1}. The tunable range can be extended to 20 cm^{-1} by cooling the ruby crystal with liquid nitrogen. If the warmer laser is operated on the R_2 line, then the range can be extended to 50 cm^{-1}.

Phase matched difference frequency generation in an InSb crystal with a fixed wavelength line of a CO_2 laser and a tunable spin flip Raman laser in the range from 11.7 to 13.0 μm will be able to produce tunable far infrared stimulated radiation (PATEL [1970]). So an InSb spin flip Raman laser in the 10 μm region is very important for obtaining tunable far infrared laser sources.

3.2.6. *Tunable far infrared parametric generation*

In ionic crystals transverse optical (TO) phonons and photons with nearly the same wave vector and energy are strongly coupled. The propagation state for these energies and wave vectors can no longer be described as a phonon or a photon, but is a mixture of these two elementary excitations known as a polariton. The stimulated radiation results from a parametric process by polaritons, whereby input pump photons of frequency ν_p interact with an optical vibrational mode in the crystal at frequency ν_i, producing Stokes radiation at $\nu_s = \nu_p - \nu_i$ and idler radiation at ν_i. Since the idler

frequency and wave vector must lie on the material dispersion characteristic (ν: k diagram), a unique set of allowed frequencies and wave vectors is determined for the scattering process. Simultaneous tuning of Stokes and idler is then accomplished by varying the angle between the pump and the Stokes propagation vectors (JOHNSON, PUTHOFF, SOOHOO and SUSSMAN [1971]).

From the energy conservation law,

$$h\nu_p = h\nu_s + h\nu_i,$$

and the wave vector conservation law gives

$$\boldsymbol{k}_p = \boldsymbol{k}_s + \boldsymbol{k}_i,$$

where $\boldsymbol{k}_p$, $\boldsymbol{k}_s$ and $\boldsymbol{k}_i$ are wave vectors of pump, Stokes and idler radiations respectively. Then,

$$\begin{aligned} k_i^2 &= k_p^2 + k_s^2 - 2k_p k_s \cos\theta \\ &= 4\pi^2(\nu_p n_p - \nu_s n_s)^2 + 8\pi^2 \nu_p \nu_s n_p n_s (1-\cos\theta), \end{aligned} \tag{3.23}$$

where θ is the angle between $\boldsymbol{k}_p$ and $\boldsymbol{k}_s$ vectors. n_p and n_s are the index of refraction of pump and Stokes radiation. Fig. 3.20 (YARBOROUGH, SUSSMAN,

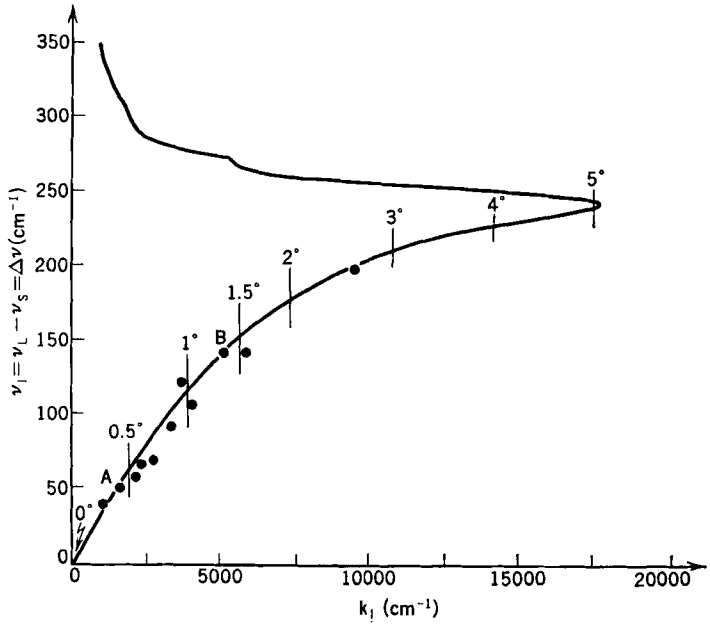


Fig. 3.20. Dispersion curve of $LiNbO_3$ in the vicinity of the A_1 symmetry 248 cm^{-1} polariton mode. The dots correspond to experimentally observed points (YARBOROUGH et al. [1969]).

PUTHOFF et al. [1969]) shows the $\nu : k$ curve of $LiNbO_3$ in the vicinity of the A_1 symmetry 248 cm^{-1} polariton mode. The lines intersecting the $\nu : k$ curve are drawn for the wave vector conservation described above. As the angle θ is varied, the idler frequency changes. The dots on the curve correspond to the observation of stimulated Stokes emission for various values of the wave vector and frequency. The experimental arrangement is shown in Fig. 3.21 (JOHNSON, PUTHOFF, SOOHOO and SUSSMAN [1971]). The pump

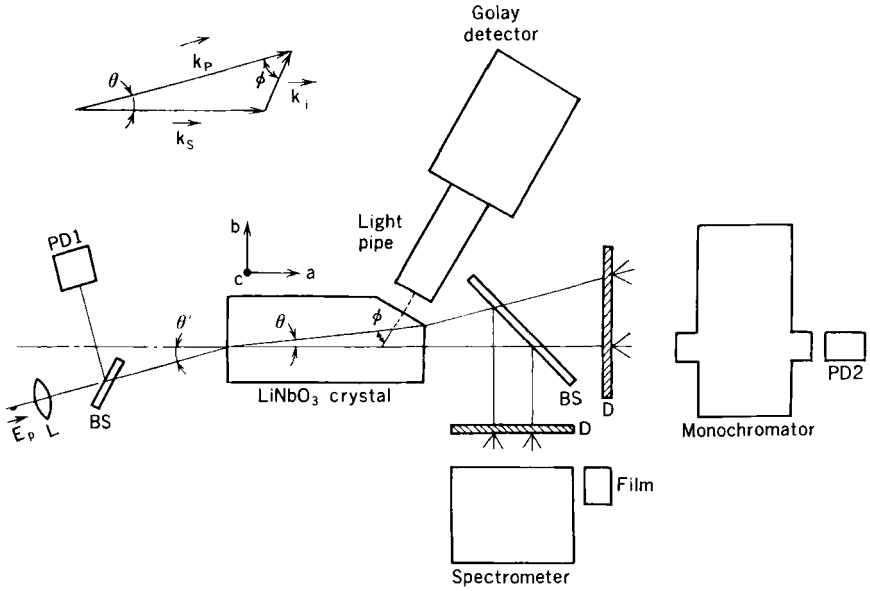


Fig. 3.21. Schematic diagram of the experimental setup of tunable far infrared parametric generation with $LiNbO_3$ (JOHNSON et al. [1971]).

is a Q switched ruby laser (6943A) emitting 20 nsec pulses with a peak of 6 MW and a beam diameter at the laser of about 2 mm. Pulse repetition rates of up to 1 pps are possible. A lens focuses the beam near the output end of a 3.4 cm a-axis $LiNbO_3$ crystal with the laser polarized along the c axis. The end faces of the crystal are polished flat and parallel. In addition, a cut is made in the corner of the crystal output end at the proper angle to allow the idler radiation to emerge approximately normal to the exit surface. Far infrared, i.e., idler radiation, is detected by a Golay cell. Wavelengths of Stokes radiation are measured with a grating spectrometer. Temporal behavior of pump and Stokes radiations is monitored on photodiodes. Fig. 3.22 shows the theoretical tuning curve and observed values for a $LiNbO_3$ crystal. The tunable range is from 66 to 200 μm.

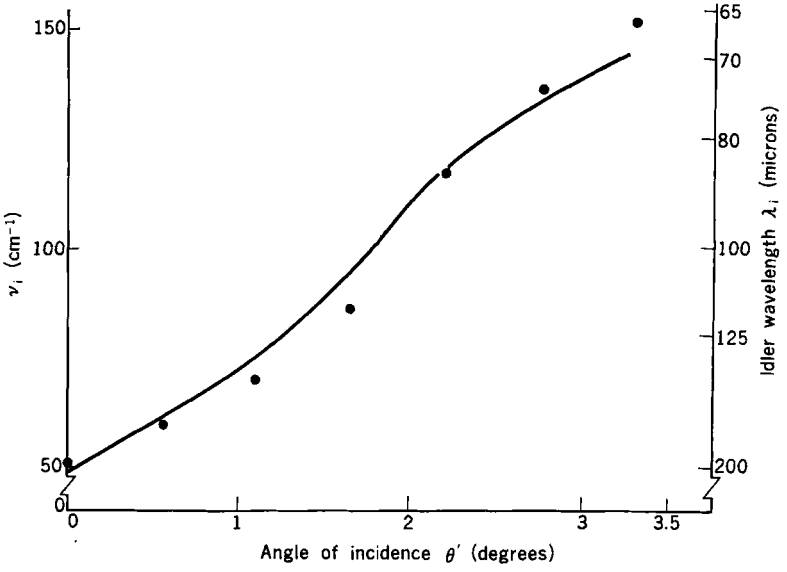


Fig. 3.22. Theoretical and observed tuning curves of far infrared parametric generation with $LiNbO_3$ (JOHNSON et al. [1971]).

The magnitude and frequency dependence of the scattered idler power is estimated theoretically as follows (JOHNSON, PUTHOFF, SOOHOO and SUSSMAN [1971]):

$$P_i = \frac{\nu_i}{\nu_s}\left(\frac{g_s \cos\varphi}{g_s \cos\varphi + \alpha_i}\right) P_s , \tag{3.24}$$

where P_i and P_s are the power at the idler and Stokes frequencies, and g_s is Stokes gain constant which satisfies

$$\partial P_s/\partial Z = g_s P_s . \tag{3.25}$$

φ is the angle between the pump and idler propagation vectors and α_i is the idler absorption constant. For a total Stokes signal power $P_s \sim 10^6$ W, the peak idler power P_i is about 3 W outside the crystal at $\nu_i = 55\ \text{cm}^{-1}$. The experimental values of peak output power are $\sim$ 3W at 200 μm and $\sim$ 0.25 W at 60 μm. About the idler and Stokes frequency bandwidths, conservation of energy for the scattering process requires that $\Delta\nu_i \leqq \Delta\nu_s$, where $\Delta\nu_i$ and $\Delta\nu_s$ are the idler and Stokes linewidths respectively. Fabry-Perot measurements indicate that $0.1 \leqq \Delta\nu_s \leqq 0.5\ \text{cm}^{-1}$. So the idler linewidth is estimated to be less than $0.5\ \text{cm}^{-1}$. Since the idler wave vector determines its propagation direction, a Δk_i results in spatial spreading of

the idler output beam. By knowing $\Delta\nu_i$, the angular divergence of the radiation as it emerges from the crystal exit face can be estimated. For a $\Delta\nu_i$ of $\sim 0.5\,\mathrm{cm}^{-1}$, the average value for the angular divergence outside the crystal is calculated to be 2.5°.

3.2.7. *Semi-tunable lasers by molecular rotational transitions*

If a vibration-rotation transition A (emission type in Fig. 3.23 (a) or absorption type in the figure (b)) between rotational levels belonging to different vibrational states of a gas molecule occurs suddenly, population inversion appears between rotational levels related to the transition as shown in the figure. Then far infrared rotational transitions B may lase. In the case (a) a Q switched laser transition A needs to introduce in a cavity for transition B. The cavity for such a purpose is shown in Fig. 3.24 (OHTSUKA [1970]). The grating acts as a diffraction grating for transition A and as a mirror for the transitions B, of which the wavelength is much bigger than the grating constant. The mirror M_1 and the rotating mirror

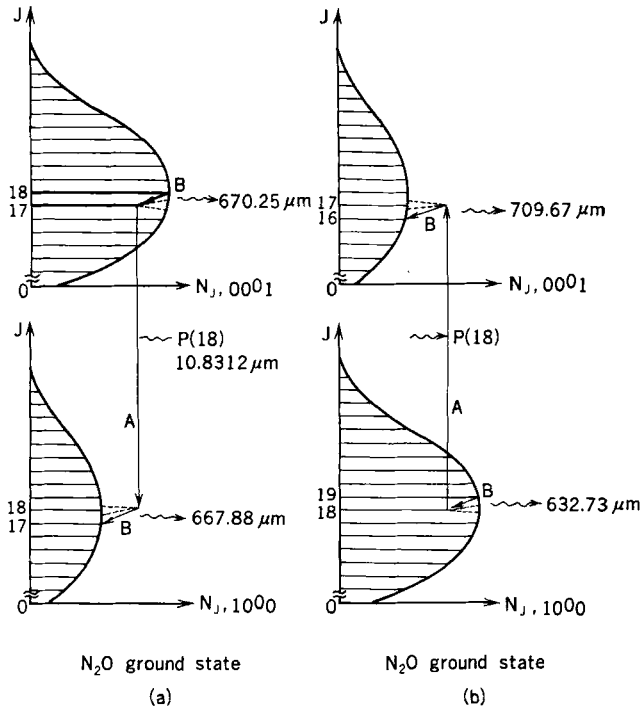


Fig. 3.23. Schematic representation of semi-tunable laser by molecular rotational transitions.

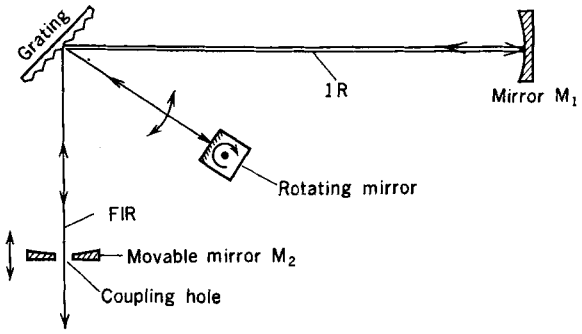


Fig. 3.24. Schematic diagram of cavities for infrared Q switched and far infrared laser transitions (OHTSUKA [1970]).

compose a cavity for a Q switched laser transition A, and the mirrors M_1 and M_2, which has a coupling hole and is movable for cavity tuning, compose a cavity for laser transitions B. These cavities are filled with gas molecules which have a dipole moment. If a Q switched laser transition occurs once, far infrared laser lines will come out from the coupling hole of mirror M_2. The transition A can be shifted among many rotational levels by rotating the cavity arm including the rotating mirror, and many far infrared laser lines can be obtained. The optimum gas pressure for the Q switched laser transition A may not be the optimum pressure for far infrared laser transition B. In the case of Fig. 3.23(b), the Q switched laser line can be produced in a different cavity from the cavity for far infrared laser lines, and the optimum pressure in those cavities does not need to be the same. Also the Q switched laser line of a different molecule can be used, when the wavelength of the laser line is just the same as that of the transition A of the molecule which produces the far infrared laser lines.

A CH_3F molecule pumped with a Q switched $CO_2P(20)$ line at 9.55 μm produces three far infrared laser lines at 496, 452 and 541 μm as shown in Fig. 3.25 (CHANG and BRIDGES [1970]). The output and the pulse width of the CO_2 laser are 1.5 kW and 0.28 μsec respectively, and the peak output and the pulse width of 496 μm are 0.1 W and 0.5 μsec respectively. The linewidth is observed to be 60 MHz at a pressure of 50 mtorr and to increase with the increase of pressure. Also CH_3F, C_2H_3Cl and CH_3OH molecules are pumped with a cw $CO_2P(20)$ laser of output 2W (CHANG, BRIDGES and BURKHARDT [1970a]). A CH_3F molecule produces 2 lines, a C_2H_3Cl molecule 3 lines and a CH_3OH molecule 23 lines in the range from 70 to 700 μm. The output power of these lines is 0.1 – a few W. Also a NH_3 molecule pumped with a N_2O laser is reported to produce 81.5 and 263.4 μm

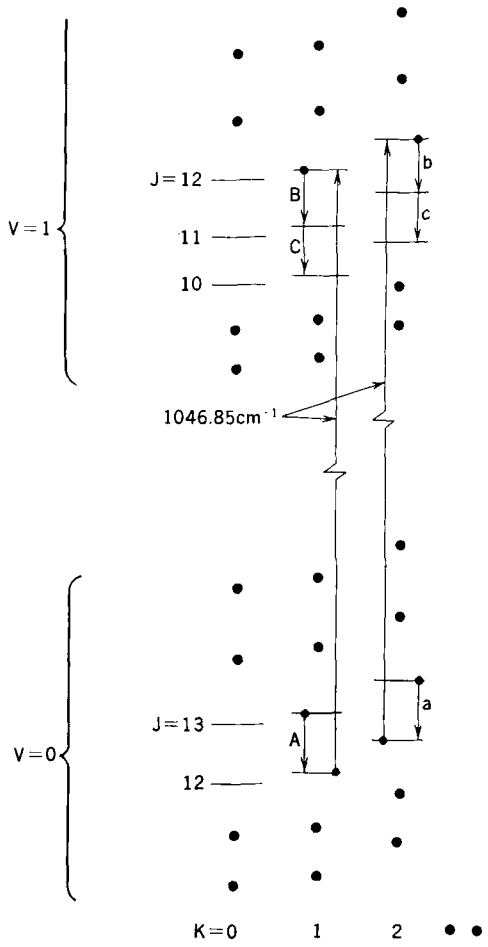


Fig. 3.25. Partial energy level diagram of CH_3F showing the observed laser transitions (CHANG et al. [1970]).

of output of 10^{-4} W (CHANG, BRIDGES and BURKHARDT [1970b]).

In these cases the molecule for pumping is different from the molecule producing far infrared laser lines. If the same molecule is used for pumping and far infrared lasing, many far infrared laser lines can be produced by shifting the pumping transitions among various rotational levels. In such a way a N_2O molecule may produce about 50 lines in the range from 100 to 300 μm and a CO molecule about 100 lines in the range from 250 to 850 μm. Since the dipole moment of these molecules is smaller than that of big molecules, like a CH_3F, C_2H_3Cl or CH_3OH molecule, it is not easy to produce laser transitions in the far infrared region.

Many laser lines are distributed discretely in the far infrared region, so that semi-tunable lasers will be expected to serve as radiation sources in far infrared spectroscopy.

3.3. TIME RESOLVED SPECTROSCOPY

Phenomena of which the spectra are to be observed may not be stable in time. In such cases the information Z obtained is given as follows:

$$Z = kI(\lambda, t), \tag{3.26}$$

where I is the intensity of radiation, which is a function of the wavelength λ and the time t, and k is a proportional constant. So two cases, i.e., (1) the continuous measurement of $I(t)$ for discrete values of λ, and (2) the continuous measurement of $I(\lambda)$ for discrete values of t, happen in time resolved spectroscopy.

In the visible and violet regions, photomultipliers with very fast response have been available for measurements of $I(\lambda)$ by very fast wavelength scanning in periods < 10 nsec. Sampling techniques are available for repeated phenomena, and $I(\lambda)$ can be measured in the period of about 1 nsec.

In the infrared, especially the far infrared region, strong radiation sources and very fast response detectors have not been practically used until now. The response time of an InSb photoconductive detector is less than 0.2 μsec. So it is possible to do fast wavelength scanning in the period of 1 μsec if a radiation source of enough intensity is available.

As described above, there have been developed monochromatic tunable pulse lasers in the middle and far infrared region, of which the pulse width is about 10 nsec and the output power is more than 1 W. Such radiation sources are much different from conventional far infrared sources like high pressure mercury lamps, of which the output power in ordinary spectral width is $10^{-7}-10^{-9}$ W. So time resolved spectroscopy in the far infrared region can be expected, especially for repeated phenomena where no fast response detector is necessary.

§ 4. Conclusion

About conventional dispersion-type spectrometers, recent fundamental developments can not be seen, except the development of elements of spectroscopic instruments. Due to their high resolution and rapid measurements interferometric spectrometers have become more and more useful for academic and industrial purposes.

Desirable monochromatic tunable detectors in the far infrared region have not yet been developed. But some monochromatic tunable laser sources have been completed, even though most of them are in the middle or visible region, except the $LiNbO_3$ polariton laser and optically pumped semi-tunable gas lasers in the far infrared region. Moreover, difference frequency generation can make far infrared tunable lasers with two strong infrared or visible lasers, if one of them is tunable.

Thus the following monochromatic tunable radiation sources will be practically used for spectroscopic use in the whole far infrared region in the near future: much higher output power of 1 W or more compared width 10^{-7}–10^{-9} W at present, much narrower spectral width of less than 10^{-2} cm^{-1} and very narrow pulse width of $\sim$ 10 nsec if necessary. Thus spectral measurements in the far infrared region, even time resolved spectroscopic measurements in the region, have been possible with much higher signal-to-noise ratio and remarkable resolution compaired with those of conventional spectroscopic measurements.

References

AGGARWAL, R. L., B. LAX, C. E. CHASE, C. R. PIDGEON, D. LIMBERT and F. BROWN, 1971, Appl. Phys. Letters **18**, 383.

BELL, E. E., 1966, Infrared Physics **6**, 57.

BELL, E. E., 1970, in: Proc. Aspen Intern. Conf. on Fourier Spectroscopy, p. 71.

BESSON, J. M. and W. PAUL, 1968, Phys. Rev. **173**, 699.

BRUECK, S. R. J. and A. MOORADIAN, 1971, Appl. Phys. Letters **18**, 229.

CALAWA, A. R., J. O. DIMMOCK, T. C. HARMAN and I. MELNGAILIS, 1969, Phys. Rev. Letters **23**, 7.

CHANG, T. Y. and T. J. BRIDGES, 1970, Opt. Commun. **1**, 423.

CHANG, T. Y., T. J. BRIDGES and E. G. BURKHARDT, 1970a, Appl. Phys. Letters **17**, 249.

CHANG, T. Y., T. J. BRIDGES and E. G. BURKHARDT, 1970b, Appl. Phys. Letters **17**, 357.

CONNES, P., 1970, in: Proc. Aspen Intern. Conf. on Fourier Spectroscopy, p. 121.

CONNES, P. and G. MICHEL, 1970, in: Proc. ASPEN Intern. Conf. on Fourier Spectroscopy, p. 313.

CONTRERAS, B. and O. L. GADDY, 1971, Appl. Phys. Letters **18**, 277.

DOWLING, J. M., 1970, in: Proc. Aspen Intern. Conf. on Fourier Spectroscopy, p. 55.

DREW, H. D. and A. J. SIEVERS, 1969, Appl. Opt. **8**, 2067.

FARIES, D. W., K. A. GEHRING, P. L. RICHARDS and Y. R. SHEN, 1969, Phys. Rev. **180**, 363.

HALL, R. T., D. VRABEC and J. M. DOWLING, 1966, Appl. Opt. **5**, 1147.

HARMAN, T. C., A. R. CALAWA, I. MELNGAILIS and J. O. DIMMOCK, 1969, Appl. Phys. Letters **14**, 333.

HINKLEY, E. D., 1970, Appl. Phys. Letters **16**, 351.

HINKLEY, E. D. and P. L. KELLEY, 1971, Science **171**, 635.

JOHNSON, B. C., H. E. PUTHOFF, T. SOOHOO and S. S. SUSSMAN, 1971, Appl. Phys. Letters **18**, 181.

LEVY, F., R. C. MILWARD, S. BRAS and R. LE TOULLEC, 1970, in: Proc. Aspen Intern. Conf. on Fourier Spectroscopy, p. 331.

MITSUISHI, A., Y. OHTSUKA, S. FUJITA and H. YOSHINAGA, 1963, Japan. J. Appl. Phys. **2**, 574.
MOORADIAN, A., S. R. J. BRUECK and F. A. BLUM, 1970, Appl. Phys. Letters **17**, 481.
NORTON, P., P. CHIA, T. BRAGGINS and H. LEVINSTEIN, 1971, Appl. Phys. Letters **18**, 158.
OHTSUKA, Y., 1970, Japan. J. Appl. Phys. **9**, 408.
PATEL, C. K. N., 1970, in: Proc. Symp. on Submillimeter Waves (New York, Polytechnic Press) p. 135.
PATEL, C. K. N., E. D. SHAW and R. J. KERL, 1970, Phys. Rev. Letters **25**, 8.
PATEL, C. K. N. and E. D. SHAW, 1971, Phys. Rev. B **3**, 1279.
PATEL, C. K. N., 1971, Appl. Phys. Letters **18**, 274.
PUTLEY, E. H., 1970, in: Proc. Symp. on Submillimeter Waves (New York, Polytechnic Press) p. 267.
RICHARDS, P. L. and S. A. STERLING, 1969, Appl. Phys. Letters **14**, 394.
RICHARDS, P. L., 1970a, in: Proc. Aspen Intern. Conf. on Fourier Spectroscopy, p. 117.
RICHARDS, P. L., 1970b, in: Physics of III-V Compounds, vol. 6 (Academic Press, N. Y.) to be published.
SAKAI, K., Y. NAKAGAWA and H. YOSHINAGA, 1968, Japan, J. Appl. Phys. **7**, 792.
SANDERSON, R. B. and H. E. SCOTT, 1971, Appl. Opt. **10**, 1097.
SHANK, C. V., J. E. BJORKHOLM and H. KOGELNIK, 1971, Appl. Phys. Letters **18**, 395.
SHAW, E. D. and C. K. N. PATEL, 1971, Appl. Phys. Letters **18**, 215.
TEICH, M. C., R. J. KEYES and R. H. KINGSTON, 1966, Appl. Phys. Letters **9**, 357.
THORPE, L. W., R. C. MILWARD, G. C. HAYWARD and J. D. YEWEN, 1969, in: Optical Instruments and Techniques, ed. J. H. Dickson (Oriel Press) p. 28.
VAN TRAN, N. and C. K. N. PATEL, 1969, Phys. Rev. Letters **22**, 463.
WALTER, H. and T. L. HALL, 1970, Appl. Phys. Letters **17**, 239.
YAMADA, Y., A. MITSUISHI and H. YOSHINAGA, 1962, J. Opt. Soc. Am. **52**, 17.
YAMAMOTO, J. and H. YOSHINAGA, 1969, Japan. J. Appl. Phys. **8**, 242 and 286.
YAMAMOTO, J., 1971, Oyobutsuri **40**, 850.
YARBOROUGH, J. M., S. S. SUSSMAN, H. E. PUTHOFF, R. H. PANTELL and B. C. JOHNSON, 1969, Appl. Phys. Letters **15**, 102.
YOSHINAGA, H., S. FUJITA, S. MINAMI, Y. SUEMOTO, M. INOUE, K. CHIBA, K. NAKANO, S. YOSHIDA and H. SUGIMORI, 1966, Appl. Opt. **5**, 1159.
YOSHINAGA, H. and J. YAMAMOTO, 1969, in: Optical Instrument and Techniques, ed. J. H. Dickson (Oriel Press) p. 41.

E. WOLF, PROGRESS IN OPTICS XI © NORTH-HOLLAND 1973

III

INTERACTION OF LIGHT AND ACOUSTIC SURFACE WAVES

BY

ERIC G. LEAN

IBM Thomas J. Watson Research Center,
Yorktown Heights, N. Y. 10598, *U.S.A.*

CONTENTS

§ 1. Introduction

The interaction of light and sound has been studied extensively since RAMAN and NATH [1935] first started their investigation. After the invention of laser which gives the ready availability of coherent light sources and the recent advances in the generation techniques of high frequency coherent elastic waves, there has been a renewed interest in the study of interaction of coherent light and sound. Such study has led to the useful applications of optical probing of elastic waves in solids and many optical signal processing devices. A recent review article by DAMON et al. [1970] has described the principles and practical devices based on interaction of light with ultrasound.

The interest in the interaction of light and acoustic surface waves began after the recent advance of acoustic surface technology in the late 1960's. IPPEN [1967] reported the first experimental observation of light diffraction by Rayleigh waves in quartz. Rayleigh waves (Lord RAYLEIGH [1885]), which is one type of acoustic surface waves commonly used, propagate on a free surface of a half infinite substrate with particle displacements decaying exponentially from the surface. These waves have bulk acoustic wave properties with additional advantages of being on the surface with the energy concentrating in a surface layer of the order of one acoustic wavelength. These unique properties, which make Rayleigh waves easily accessible along the propagation path and having high power density with a modest input power, are some of the reasons for the current interest in acoustic surface waves in signal processing applications (STERN [1969], KINO and MATHEWS [1971]).

In designing such acoustic surface wave devices, it is necessary to have detail knowledge of the propagation characteristics of these waves on any isotropic or anisotropic substrate. Optical probing of acoustic surface waves based on the light diffraction by acoustic surface waves has become a convenient tool in studying and utilizing these waves.

Light diffraction by Rayleigh waves occurs for both the reflected light due to the surface ripples in opaque substrates (MEYER et al. [1967], ADLER et

al. [1968]) and the transmitted light due to the surface ripples and the photoelastic effects in the surface layer of a transparent substrate (IPPEN [1967], LEAN [1970]). It has been used extensively as a probe to visualize the surface waves (ADLER et al. [1968]), to measure the propagation characteristics of acoustic surface waves, such as the velocity (AUTH and MAYER [1967], KROKSTAD and SVAASAND [1967]), the diffraction and the beam steering in anisotropic substrate (SLOBODNIK [1969], LEAN and POWEL [1970]), the attenuation (SLOBODNIK et al. [1970]) and the reflection due to the surface perturbations (DE LA RUE [1971]). Such optical probing technique has also been extended to study the nonlinear effects of acoustic surface waves. The spatial growth of ASW harmonics generated due to elastic nonlinearities of the substrate can be directly observed by the optical probe (LEAN et al. [1969]).

In this paper the basic principle of light diffraction by acoustic surface waves will be reviewed. The techniques of the optical probing of acoustic surface waves will be discussed together with their advantages and limitations.

The main difference between the light diffraction by bulk acoustic waves and by acoustic surface waves comes from the unique properties of acoustic surface waves. In general, acoustic surface waves include Rayleigh waves (RAYLEIGH [1887]), dispersive Rayleigh waves (EWING et al. [1957]) and Love waves (LOVE [1926]). The existence of these waves results in the boundary conditions due to a single surface on a semi-infinite substrate (Rayleigh waves), or due to a thin film on a substrate (dispersive Rayleigh waves and Love waves). Recently the acoustic surface technology has been greatly advanced and many papers concerning acoustic surface waves, Rayleigh waves in particular, have appeared in the literature. For detailed understanding of acoustic surface waves, interested readers are referred to several review articles (WHITE [1970], FARNELL [1970], DRANSFELD [1970]).

Consider Rayleigh waves propagating along the x_3 axis on a surface whose normal is in the x_2 axis. The particle displacement of such a wave can be expressed by:

$$U_i = \sum_{n=1}^{3} a_{in} \exp\{\alpha_n K x_2 + j(\omega_s t - K x_3)\}, \qquad i = 1, 2, 3 \qquad (1.1)$$

when U_i are the particle displacements, K is the acoustic wavenumber, a_{in} amplitude constants. α_n are the normalized decaying constants. The constants a_{in} and α_n can be obtained by solving the wave equation for the particle displacements in the substrate medium with a stress free boundary condition on the surface (FARNELL [1970]). In isotropic substrates, the solutions of the particle displacement for Rayleigh waves consist of only two

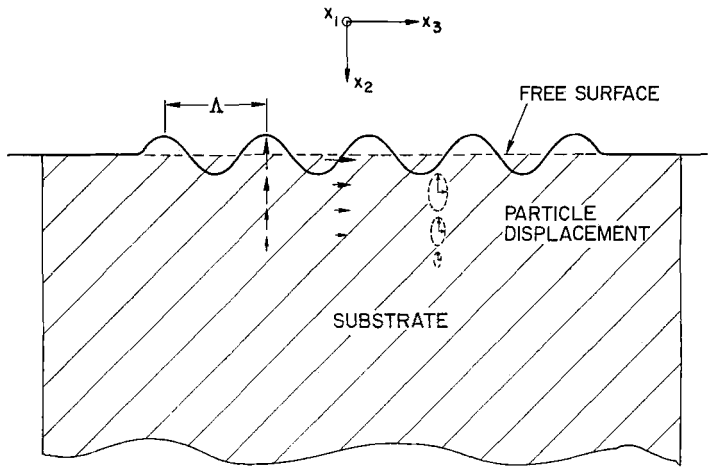


Fig. 1.1. Schematic of particle displacements of Rayleigh waves in isotropic substrate.

components (VIKTOROV [1967], WHITE [1970]) as shown in Fig. 1.1. These two particle displacements are 90° out of phase so that particle motions are ellipsoidal. The amplitude of the particle displacement decays exponentially from the substrate surface with a depth of the order of the acoustic wavelength. These waves produce a surface ripple on the surface as well as a periodic variation in the index of refraction in the medium.

The surface ripple δ is simply given by the normal component of the particular displacement at the surface ($x_2 = 0$). From eq. (1.1) the surface ripple can be written as:

$$\delta = \delta_1 \cos(\omega_s t - K x_3) \tag{1.2}$$

where

$$\delta_1 = |\sum_{n=1}^{3} a_{2n}|. \tag{1.3}$$

The periodic variation in the index of refraction in the penetration layer is caused by the photoelastic effect (NYE [1958]). The photoelastic effect, which causes a change in optical properties of a crystal in the presence of elastic waves, can be specified as a small change in the shape, size, and orientation of the indicatrix which gives a geometric representation of the crystal optical property. The indicatrix is an ellipsoid of wave normals whose coefficients are the components of the relative dielectric impermeability tensor B_{ij} at optical frequencies. If coordinates are referred to the principal dielectric axis of the crystal the indicatrix is:

$$B_{11} x_1^2 + B_{22} x_2^2 + B_{33} x_3^2 = 1 \tag{1.4}$$

where

$$B_{ii} = \left(\frac{1}{n_{ii}}\right)^2 = \frac{1}{\varepsilon_{ii}} \qquad (i = 1, 2, 3). \tag{1.5}$$

n_{ii} and ε_{ii} are the optical index of refraction and the dielectric constant along the principal crystal axes. Under elastic strain, the changes ΔB_{ij} in the impermeability is related to the strain tensor S_{kl} by the equation:

$$\Delta B_{ij} = P_{ijkl} S_{kl} \tag{1.6}$$

where P_{ijkl} is the photoelastic tensor and

$$S_{kl} = \frac{1}{2}\left(\frac{\partial U_k}{\partial x_l} + \frac{\partial U_l}{\partial x_k}\right) \tag{1.7}$$

with U_k being the particle displacement component. The change in the index of refraction Δn_{ii} and the dielectric constant $\Delta\varepsilon_{ii}$ can be obtained from (1.6), (1.5) and (1.7);

$$\Delta n_{ij} = -\tfrac{1}{4} n_{ii}^3 P_{ijkl}\left(\frac{\partial U_k}{\partial x_l} + \frac{\partial U_l}{\partial x_k}\right) \tag{1.8}$$

$$\Delta\varepsilon_{ij} = -\tfrac{1}{2}\varepsilon_{ii}^2 P_{ijkl}\left(\frac{\partial U_k}{\partial x_l} + \frac{\partial U_l}{\partial x_k}\right). \tag{1.9}$$

The usual summation convention for the tensor elements in (1.6), (1.8) and (1.9) is understood. The substitution of the particle displacements in (1.1) into (1.8) or (1.9) gives the change in the index of refraction Δn_{ij} or in the dielectric constant $\Delta\varepsilon_{ij}$.

Equations (1.8) and (1.9) are similar to those of bulk acoustic wave cases. However, the transverse dependences of acoustic surface waves in particle displacements give extra terms and complications in Δn_{ij} or $\Delta\varepsilon_{ij}$. It should be noted that the subscripts i and j in Δn_{ij} indicate the polarization of the diffracted and the incident light respectively. The light diffraction involving Δn_{ij} with $i = j$ preserves the polarization of the incident light while in the cases of Δn_{ij} with $i \neq j$ the polarization of the diffracted light is different from that of the incident light. In § 2 the light diffraction by the surface ripple and Δn_{ij} due to acoustic surface waves will be calculated theoretically and demonstrated experimentally.

Recently there has been another interesting development involving the optical guided waves which propagate in a thin film guide of the order of micrometer in thickness on a substrate (GOELL and STANDLEY [1970]). The index of refraction in the thin film needs to be higher than that in the substrate so that the light is trapped in the film by total internal reflection with

the field decaying exponentially in the substrate. Optical guided waves interact strongly by means of photoelastic effect with acoustic surface waves, which also travel in the same region and have the same form of exponentially decaying amplitudes. These guided waves which have the advantages of high power density and controllable dispersion depending on the film thickness make such acousto-optic interactions in thin films particularly important both from theoretical and practical points of view. Two recent experimental demonstrations, the efficient Bragg deflection of optical guided waves by acoustic surface waves (KUHN et al. [1970]) and the optical guided waves mode conversion based on collinear interaction of optical guided waves and acoustic surface waves in thin films (KUHN et al. [1971]) have been reported. In § 3, the theory on the interaction of optical guided waves and acoustic surface waves will be discussed in detail together with some recent experimental results and current new developments.

§ 2. Interaction of Light and Rayleigh Waves

Rayleigh waves propagate on the surface of a substrate and produce surface ripples as well as a periodic variation in the index of refraction in the medium as discussed in § 1. Depending on the width of the incident beam compared to the acoustic wavelength, the interaction of light and Rayleigh waves can be considered as either a phase grating type of diffraction or a periodic deflection of incident light. If the light beam is many wavelengths wide, the surface ripple and the periodic variation of the index of refraction provide a phase grating for the incident light. For the reflected beam only the surface ripple contributes to the light diffraction while for the transmitted beam in a transparent substrate both the surface ripple and the periodic variation in the index of refraction contribute to the diffracted light. If the light beam is focused on the surface to a spot smaller than the acoustic wavelength, the surface ripple in effect tilts the direction of the reflected beam. By properly placing a knife edge in the path of the reflected beam, the periodic beam tilting can be transformed into an amplitude modulation. In this section, the basic theory of the light diffraction and the knife's edge scheme by Rayleigh waves will be discussed together with some of their device applications.

2.1. LIGHT DIFFRACTION BY RAYLEIGH WAVES

The scattering of electromagnetic waves from a sinusoidal surface was first investigated by RAYLEIGH [1895]. Since then many workers have studied the problem (for example, BECKMANN and SPIZZICHIRO [1963]). The acoustic

surface waves produce a moving sinusoidal surface ripple. The acoustic velocity is about five order of magnitude lower than the speed of light and appears stationary to an incident light beam. The scattering of light by acoustic surface wave can thus be treated similarly to the problem of the light scattering from a periodic surface as discussed by BECKMANN et al. [1963].

2.1.1. *Reflection case*

The scattering of light from the surface ripple produced by Rayleigh waves on a solid can be treated similarly to the problem of light scattering from a periodic surface as discussed by BECKMANN et al. [1963].

The Helmholtz integral can be used to calculate the scattered electric field $E_2(\mathrm{p})$ at the observation point p;

$$E_2(\mathrm{p}) = \frac{1}{4\pi}\int_S \left(E\frac{\partial G}{\partial n} - G\frac{\partial E}{\partial n}\right) \mathrm{d}s \tag{2.1}$$

where E and $\partial E/\partial n$ are the electric field and the normal derivative of the field on the surface S; Fig. 2.1 shows the configuration and coordinates for

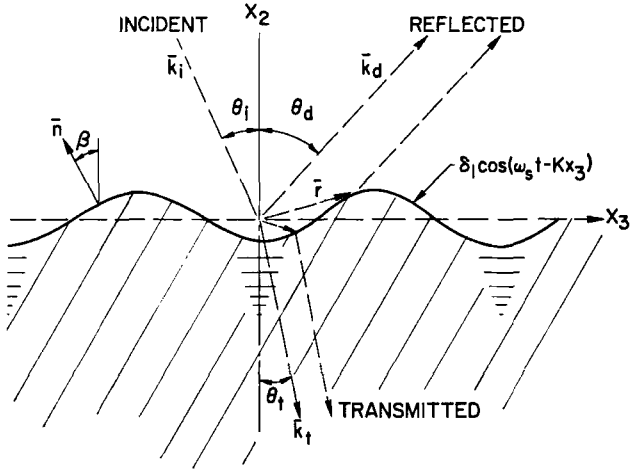


Fig. 2.1. Coordinates of the light diffraction by acoustic surface waves.

the problem we consider. The surface S is the crystal surface with the sinusoidal ripple, G is the Green function which satisfies the wave equation except at the observation point p. G can be written as:

$$G = \frac{\exp\{-\mathrm{j}k_\mathrm{d}|\boldsymbol{R}-\boldsymbol{r}|\}}{|\boldsymbol{R}-\boldsymbol{r}|} \tag{2.2}$$

where $k_d = \omega\sqrt{\varepsilon_0\mu_0}$ is the wavenumber, $\boldsymbol{R}$ is the vector from the origin to the observation point p as shown in Fig. 2.1. $\boldsymbol{r}$ is the vector from the origin to a point in the x_2x_3 plane;

$$\boldsymbol{r} = x_2\boldsymbol{x}_2 + x_3\boldsymbol{x}_3. \tag{2.3a}$$

In particular, for points on the surface S, $\boldsymbol{r}$ reduces to:

$$\boldsymbol{r}' = \delta(x_3)\boldsymbol{x}_2 + x_3\boldsymbol{x}_3 \tag{2.3b}$$

where $\delta(x_3)$ is given by (1.2), $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ on the unit vectors along the coordinate axes.

In the far field region, G can be approximated by:

$$G = \frac{\exp\{-\mathrm{j}(k_d R_0 - \boldsymbol{k}_d \cdot \boldsymbol{r})\}}{R_0} \tag{2.4}$$

where R_0 is the distance from the origin to the observation point. $\boldsymbol{k}_d$ is given by:

$$\boldsymbol{k}_d = k_d \cos\theta_d \boldsymbol{x}_2 + k_d \sin\theta_d \boldsymbol{x}_3 \tag{2.5}$$

and θ_d is the angle of the observation direction with respect to the surface normal. The surface normal $\boldsymbol{n}$ is given by:

$$\boldsymbol{n} = \cos\beta\boldsymbol{x}_3 - \sin\beta\boldsymbol{x}_3 \tag{2.6a}$$

where β is the tangential angle due to the surface ripple and is given by:

$$\tan\beta = \delta'(x_3) = \delta_1 K \sin(\omega_s t - Kx_3). \tag{2.6b}$$

The derivative of G with respect to the surface normal becomes:

$$\begin{aligned}\frac{\partial G}{\partial n} &= \mathrm{j}G\frac{\partial \boldsymbol{k}_d \cdot \boldsymbol{r}}{\partial n} \\ &= \mathrm{j}G\left[k_d \cos\theta_d \frac{\partial x_2}{\partial n} + k_d \sin\theta_d \frac{\partial x_3}{\partial n}\right] \\ &= \mathrm{j}G[k_d \cos\theta_d \cos\beta - k_d \sin\theta_d \sin\beta].\end{aligned} \tag{2.7}$$

In solving the Helmholtz integral the key point is the proper approximation of E and $\partial E/\partial n$ on S. In our case of light diffraction from the surface ripple produced by Rayleigh waves, the ripple amplitude δ_1 is usually much smaller than the acoustic wavelength Λ. Under the normal conditions, $\delta_1 < 10^{-3}\,\Lambda$, the surface ripple provides only a very small perturbation of the surface at $x_2 = 0$. The field on the surface S can be approximated as the sum of the incident field and the reflected field from the surface without the ripple. Let the incident field be:

$$E_{in} = E_0 \exp\{j(\omega t + k_i \cos\theta_i x_2 - k_i \sin\theta_i x_3)\} \tag{2.8}$$

where E_0, k_i and θ_i are the amplitude, wavenumber and incident angle of the incident light field respectively.

In cases where the surface ripple is a small perturbation, we can assume the field on the surface S is the sum of the incident and reflected fields, where the reflected field is from the substrate surface without the ripple;

$$\begin{aligned} E|_s &= (1+R)E_{in}|_s \\ &= (1+R)E_0 \exp\{j(\omega t - \boldsymbol{k}_i \cdot \boldsymbol{r}')\} \end{aligned} \tag{2.9}$$

where R is the reflection coefficient of the substrate surface without the ripple. R is dependent on the polarization of the incident light. Let R^+ and R^- be the reflection coefficients for the vertically polarized incident light (the electric field is in the plane of incidence) and the horizontally polarized light (the electric field is normal to the plane of incidence) respectively;

$$R^+ = \frac{(\varepsilon_2/\varepsilon_1)\cos\theta_i - \sqrt{\varepsilon_2/\varepsilon_1 - \sin^2\theta_i}}{(\varepsilon_2/\varepsilon_1)\cos\theta_i + \sqrt{\varepsilon_2/\varepsilon_1 - \sin^2\theta_i}} \tag{2.10}$$

$$R^- = \frac{\cos\theta_i - \sqrt{\varepsilon_2/\varepsilon_1 - \sin^2\theta_i}}{\cos\theta_i + \sqrt{\varepsilon_2/\varepsilon_1 - \sin^2\theta_i}} \tag{2.11}$$

where ε_2 and ε_1 are respectively the dielectric constants of the substrate and the medium in which the incident beam propagates.

The normal derivative of E on S can then be found as:

$$\begin{aligned} \left.\frac{\partial E}{\partial n}\right|_s &= \left.\frac{\partial E_{in}}{\partial n}\right|_s + \left.\frac{\partial E_r}{\partial n}\right|_s \\ &= j[(1-R)k_i \cos\theta_i \cos\beta + (1+R)k_i \sin\theta_i \sin\beta]E_{in}|_s . \end{aligned} \tag{2.12}$$

Substituting (2.4), (2.6), (2.8) and (2.10) into (2.1) results in:

$$\begin{aligned} E_2(p) = \frac{jE_0 B \exp\{j(\omega t - k_d R_0)\}}{4\pi R_0} \int_{-D/\cos\theta_i}^{D/\cos\theta_i} &[1 + Q \sin(\omega_s t - Kx_3)] \\ &\times \exp\{j[\alpha_1 \cos(\omega_s t - Kx_3) + \gamma x_3]\} dx_3 , \end{aligned} \tag{2.13}$$

$$B = (1+R)k_d \cos\theta_d - (1-R)k_i \cos\theta_i \tag{2.14}$$

$$Q = (-\delta_1 K)\frac{(1+R)(k_d \sin\theta_d + k_i \sin\theta_i)}{(1+R)k_d \cos\theta_d - (1-R)k_i \cos\theta_i} \tag{2.15}$$

$$\alpha_1 = (k_d \cos\theta_d + k_i \cos\theta_i)\delta_1 \tag{2.16}$$

$$\gamma = k_d \sin\theta_d - k_i \sin\theta_i \tag{2.17}$$

where $2D$ is the width of the incident beam. In obtaining eq. (2.13) the relation of $ds = dx_3/\cos\theta_i$ on the surface S has been used.

When the surface ripple reduces to zero, i.e. $\delta_1 = 0$ and $\theta_i = \theta_d$, eq. (2.13) becomes:

$$E_{20} = \frac{jE_0 RDk_i \exp\{j(\omega t - k_d R_0)\}}{\pi R_0}$$

which is the field reflected in the direction of specular reflection by a smooth surface with a reflection coefficient R.

Introducing the identity of:

$$\exp\{j\alpha_1 \cos(\omega_s t - Kx_3)\} = \sum_m j^m J_m(\alpha_1) \exp\{jm(\omega_s t - Kx_3)\}$$
$$m = 0, \pm 1, \pm 2, \ldots \quad (2.18)$$

where $J_m(\alpha_1)$ is the m^{th} order Bessel function, we can carry out the integration of (2.13) term by term and obtain:

$$E_2 = \sum_{m=-\infty}^{\infty} A_m \exp\{j[(\omega + m\omega_s)t - (k_d R_0)]\} \quad (2.19)$$

where

$$A_m = \frac{j^{m+1} E_0 BD}{2\pi R_0 \cos\theta_i} \left(1 - m\frac{Q}{\alpha_1}\right) F J_m(\alpha_1) \quad (2.20a)$$

$$F = \frac{\sin\{(\gamma - mK)D/\cos\theta_i\}}{(\gamma - mK)D/\cos\theta_i}. \quad (2.20b)$$

It is noted from (2.19) that the diffracted field consists of many orders. To each integer m there corresponds a diffraction order which has a maximum intensity along the direction θ_{dm} given by:

$$\gamma - mK = 0 \quad (2.21a)$$

or:

$$\sin\theta_{dm} = \frac{k_i}{k_d}\sin\theta_i + m\frac{K}{k_d}. \quad (2.21b)$$

It is also evident from (2.19) that the frequency of the m^{th} order is frequency shifted by an amount of $m\omega_s$.

The time averaging intensity of the m^{th} order diffracted light is given by:

$$I_{2m} = \langle E_2 E_2^* \rangle_m$$
$$= \left(\frac{E_0 D}{2\pi R_0}\right)^2 \frac{B^2}{\cos^2\theta_i} \left(1 - m\frac{Q}{\alpha_1}\right)^2 F^2 J_m^2(\alpha_1). \quad (2.22)$$

Under the normal experimental condition, $\alpha_1 \ll 1$ is usually true so that the Bessel function $J_m(\alpha_1)$ can be expanded as:

$$J_m(\alpha_1) \approx \frac{1}{m!}\left(\frac{\alpha_1}{2}\right)^m. \tag{2.23}$$

With this small argument expansion of $J_m(\alpha_1)$, it is possible to simplify the expression of (2.22) by substituting (2.14), (2.15), (2.16) and (2.20b) into (2.22). The result gives:

$$I_{2m} = I_0 F^2 H^2 \frac{1}{(m!)^2}(k_d \cos\theta_d + k_i \cos\theta_i)^{2m}\delta_1^{2m} \tag{2.24}$$

where:

$$I_0 = \left(\frac{E_0 D k_i}{\pi R_0}\right)^2 \tag{2.25}$$

$$H = \frac{(1+R)k_d \cos\theta_d - (1-R)k_i \cos\theta_i}{2k_i \cos\theta_i} + m\frac{K(1+R)(k_d \sin\theta_d + k_i \sin\theta_i)}{2k_i \cos\theta_i(k_d \cos\theta_d + k_i \cos\theta_i)}. \tag{2.26}$$

The reflection coefficient R in (2.26) is given by (2.10) or (2.11) depending on the incident light polarization.

It should be noted that under the assumption that $\lambda/\Lambda \ll 1$, $\theta_i = \theta_d$, $k_d = k_i$, eq. (2.26) for $m = 1$ becomes:

$$H = \left[R + \frac{\lambda}{\Lambda}\frac{(1+R)}{2}\frac{\sin\theta_i}{\cos^2\theta_i}\right] \approx R \tag{2.27}$$

and the peak intensity of the first order diffracted light I_{21} can be approximated by:

$$I_{21} \approx I_0 R^2 J_1^2(2k_i \cos\theta_i \delta_1). \tag{2.28}$$

Equation (2.28) has been generally used in the literature for the interpretation of the diffracted light intensity as a function of the incident angle. However, the above assumption is not always true especially around the Brewster angle region for a vertically polarized incident light where R is also very small compared to the second term in (2.27). A more detailed discussion with experimental verification is shown in Section 2.1.3.

2.1.2. *Transmission case*

In the transmission case, the transmitted beam is modulated by both the surface ripple and the periodic variation in the index of refraction in the

substrate. As discussed in § 1, the change in the index of refraction $\Delta n(x_2, x_3)$ (as given in eq. (1.8)) decays exponentially from the substrate surface with a penetration depth D of the order of the acoustic wavelength Λ. As long as the acoustic wavelength is longer than that of incident light λ, the layer of Δn provides only a phase modulation of incident light, according to WILLARD's criteria [1949] where $\lambda D/\Lambda^2 = \lambda/\Lambda < 1$.

First, let's assume that there had not been a layer of Δn. The diffracted field due to the surface ripple can be calculated similar to the reflection case with:

$$E|_s = (1+R)E_{in}|_s \tag{2.29}$$

$$\left.\frac{\partial E}{\partial n}\right|_s = -\mathrm{j}[(1-R)k_i \cos\theta_i \cos\beta + (1+R)k_i \sin\theta_i \sin\beta]E_{in}|_s \tag{2.30}$$

where $E_{in}|_s$ is the incident field on the surface S (eq. (2.8)) and R is the reflection coefficient of the substrate surface. The Green function for the transmission case is then:

$$G = \exp\{-\mathrm{j}(k_t R_0 - \boldsymbol{k}_t \cdot \boldsymbol{r})\}/R_0 \tag{2.31a}$$

where

$$\boldsymbol{k}_t = -k_t \cos\theta_t \boldsymbol{x}_2 + k_t \sin\theta_t \boldsymbol{x}_3 . \tag{2.31b}$$

θ_t is the angle of the transmitted beam with respect to $\boldsymbol{x}_2$ axis. The derivative of G with respect to the surface normal becomes:

$$\partial G/\partial n|_s = \mathrm{j}(\boldsymbol{k}_t \cdot \boldsymbol{n}')G|_s \tag{2.32}$$

where:

$$\boldsymbol{n}' = -\cos\beta \boldsymbol{x}_2 + \sin\beta \boldsymbol{x}_3 \tag{2.33}$$

is the surface normal for the transmission case and β is given by (2.7b).

Substituting the above equation into (2.1) yields the field $E_3(\mathrm{p})$ for the transmission case:

$$E_3(\mathrm{p}) = \frac{\mathrm{j}E_0 \exp\{\mathrm{j}(\omega t - k_t R_0)\}}{4\pi R_0} \int_{-D/\cos\theta_i}^{D/\cos\theta_i} B'[(1+Q')\sin(\omega_s t - Kx_3)] \times \exp\{\mathrm{j}(\boldsymbol{k}_t - \boldsymbol{k}_i)\cdot \boldsymbol{r}\}\mathrm{d}x_3 , \tag{2.34}$$

$$B' = (1+R)k_t \cos\theta_t + (1-R)k_i \cos\theta_i , \tag{2.35a}$$

$$Q' = \frac{\delta_1 K(1+R)(k_t \sin\theta_t + k_i \sin\theta_i)}{B'} , \tag{2.35b}$$

E_0 is the amplitude of the incident beam with a beam width $2D$.

The integration of (2.34) will result in a similar expression of (2.24).

However, there is a variation in the index of refraction $\Delta n(x_2, x_3)$ as given in (1.8). The effect of the surface ripple $\delta(x_3)$ and the periodic variation in the index of refraction $\Delta n(x_2, x_3)$ can be taken into account by expanding the wavevector of the transmitted beam and letting $k_t = (n+\Delta n)k_i$ in eq. (2.34);

$$\begin{aligned}(\boldsymbol{k}_t-\boldsymbol{k}_i)\cdot\boldsymbol{r} &= -[(n+\Delta n)k_i\cos\theta_t - k_i\cos\theta_i][(x_2+\delta(x_3)] \\ &\quad +[k_t\sin\theta_t - k_i\sin\theta_i]x_3 \\ &= -[(n\cos\theta_t-\cos\theta_i)k_i(x_2+\delta)+\Delta n k_i\cos\theta_t(x_2+\delta)] \\ &\quad +[k_t\sin\theta_t - k_i\sin\theta_i]x_3 . \qquad (2.36a)\end{aligned}$$

Since $\delta(x_3)$ locates at $x_2 = 0$ and Δn is a function of x_2 and x_3, to take the accumulative effect of Δn along the x_2 direction, eq. (2.36a) can be written as:

$$\begin{aligned}(\boldsymbol{k}_t-\boldsymbol{k}_i)\cdot\boldsymbol{r} \approx\ & -(n\cos\theta_t-\cos_i)k_i\,\delta(x_3) - \int_0^\infty \Delta n(x_2, x_3)k_i\cos\theta_t\,dx_2 \\ & +[k_t\sin\theta_t - k_i\sin\theta_i]x_3 , \qquad (2.36b)\end{aligned}$$

where the first term in the right hand side of (2.36b) is due to the surface ripple and the second term is due to the index variation in the substrate. Depending on the incident light polarization, the second term may add to the first term constructively or destructively. In case of constructive contributions from both terms, the substitution of (2.36b) into (2.34) and the integration of the resultant equation yield:

$$E_3(\mathrm{p}) = \sum_{m=-\infty}^{\infty} A'_m \exp\{\mathrm{j}[(\omega+m\omega_s)t - k_t R_0]\} \qquad (2.37)$$

where

$$A'_m = \frac{\mathrm{j}^{m+1}E_0 DB'}{2\pi R_0\cos\theta_i}\left(1-m\frac{Q'}{\alpha'_1}\right)F' J_m(\alpha'_1), \qquad (2.38)$$

$$B' = (1+R)k_t\cos\theta_t+(1-R)k_i\cos\theta_i , \qquad (2.35a)$$

$$Q' = \frac{\delta_1 K(1+R)(k_t\sin\theta_t+k_i\sin\theta_i)}{B'} , \qquad (2.35b)$$

$$\alpha'_1 = (n\cos\theta_t-\cos\theta_i)k_i\delta_1 + \left|\int_0^\infty \Delta n(x_2)k_i\cos\theta_t\,dx_2\right| , \qquad (2.39)$$

$$F' = \frac{\sin\{(\gamma'-mK)D/\cos\theta_i\}}{(\gamma'-mK)D/\cos\theta_i} \qquad (2.40)$$

$$\gamma' = k_t\sin\theta_t - k_i\sin\theta_i . \qquad (2.41)$$

It is evident from (2.37) that the diffracted beam in the transmission case has many discrete orders with each order shifted in frequency. The intensity of the m^{th} order diffracted light can be shown to be given by the same expression as given in (2.22) or in (2.24).

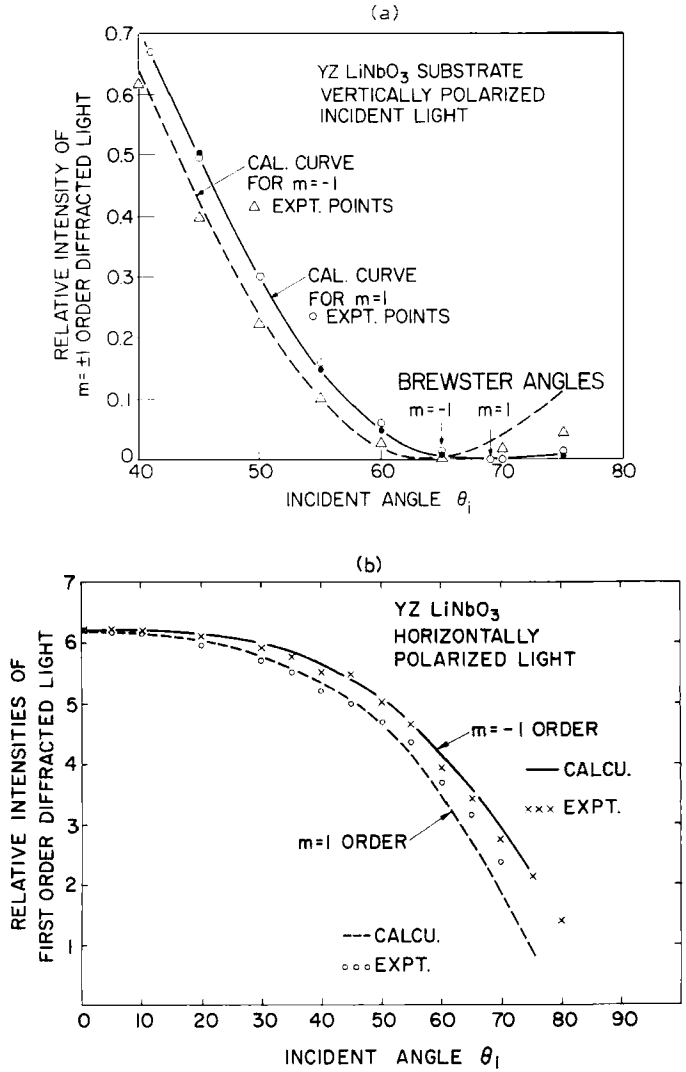


Fig. 2.2. Experimental and calculated results of the $+1$ and -1 orders of the diffracted light by a 150 MHz acoustic surface in a $LiNbO_3$ substrate. The incident light is vertically polarized (a) and horizontally polarized (b).

2.1.3. *Experimental results*

2.1.3.1. Reflection case

In order to verify the relation of the diffracted light intensities as a function of incident angle for both polarizations, as discussed in Section 2.1.1, we have monitored the $+1$ and -1 orders of the diffracted light by a 150 MHz acoustic surface wave in a lithium niobate ($LiNbO_3$) substrate. The Rayleigh waves were excited by interdigital transducers (WHITE [1965]) in a Y-cut $LiNbO_3$ substrate and propagated along the Z-axis. A 0.63μ laser with rotatable polarization provided the incident beam. Fig. 2.2a shows the experimental results for the vertically polarized incident light together with the theoretical calculation based on eq. (2.24) with R given by (2.10). With an index of refraction $n = 2.3$ for the substrate the calculated values for the $+1$ and -1 order diffracted intensities agree very well with the experimental values. For the case of horizontally polarized incident light, there is no Brewster angle for the diffracted light. The observed diffracted light for $+1$ and -1 orders incident angle are shown in Fig. 2.2b together with the calculated curves based on (2.24) with R given by (2.11).

SALZMANN and WEISMANN [1967] have reported the light diffraction by acoustic surface waves in quartz with an experimental configuration as shown in Fig. 2.3. The incident laser incident on a quartz substrate surface from

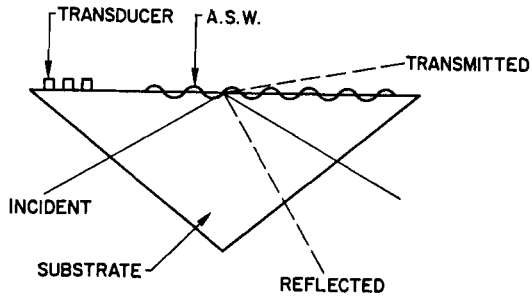


Fig. 2.3. Experimental configuration of the light diffraction by acoustic surface waves in quartz (after SALZMANN and WEISMANN [1967]).

within the solid. The $m = 1$ order diffracted light by a 316 MHz acoustic surface wave was measured as a function of the incident angle. They found that the results of their observed diffracted intensity agreed with the approximated equation given in (2.28) and concluded that the surface ripple effect alone could account for the total diffracted light intensity. ZORY and POWELL [1971] have extended the measurement and monitored both the

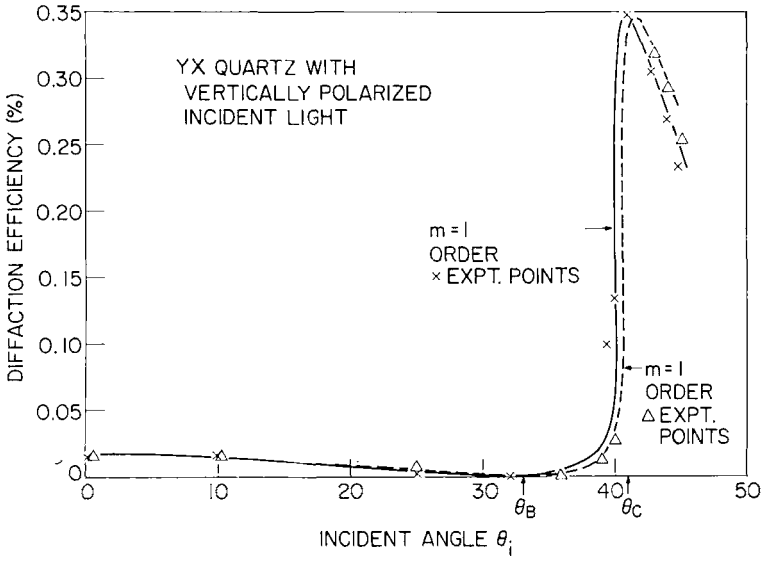


Fig. 2.4. Experimental results (from a paper by ZORY and POWELL [1971]) and calculated values (based on eq. (2.24) with R given by (2.10)) of light diffraction of $m = \pm 1$ order by acoustic surface waves in quartz.

$m = +1$ and $m = -1$ orders as a function of incident angle. Fig. 2.4 shows the measured results. It is noted that for diffraction orders $m = 1$ and $m = -1$ for vertically polarized light, the intensity reaches zero respectively before and after the Brewster angle θ_B of the crystal quartz. Also shown in Fig. 2.4 are the calculated curves based on (2.24) with R given by (2.10). The agreement between the calculation and experimental results is good.

2.1.3.2. Transmission case

In the transmission case, IPPEN [1967] and LEAN and POWELL [1970] have reported the experimental observation of light diffraction by Rayleigh waves in quartz and $LiNbO_3$ crystals.

In order to demonstrate the effects due to both the surface ripple and the change in the index of refraction, LEAN and POWELL [1971] have measured the normalized first order diffracted light for both the reflected and transmitted beams as a function of the incident light polarization angle. The sample was a YZ $LiNbO_3$ substrate with a 200 MHz acoustic surface wave propagating along the x_3 axis. A He-Ne laser beam was incident normal to the substrate. Photodetectors were placed at proper angles to detect the first

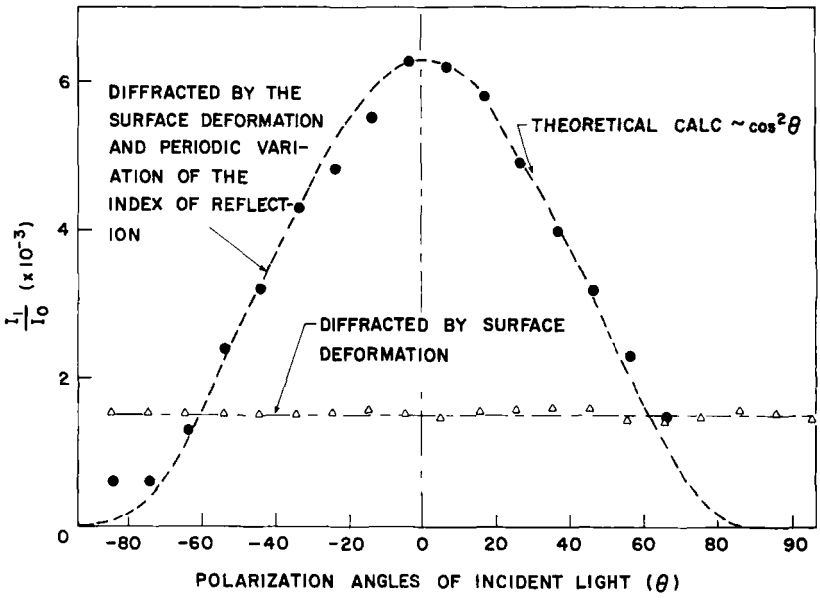


Fig. 2.5. Experimental results of the first order diffracted light I_1/I_0 as a function of incident polarization angle.

order diffracted light I_1. The experimental results of I_1/I_0 versus the incident polarization angles are shown in Fig. 2.5. I_0 is the zero order beam. For the reflection case, I_1/I_0 is independent of the incident polarization as shown by the straight line in Fig. 2.5. For the transmission case, the contribution due to the periodic variation Δn the index of refraction adds constructively to the contribution of the surface ripple when the incident light is polarized along the x_1 axis of the $LiNbO_3$ substrate. As the incident light polarization is changed along the x_3 axis, the two contributions cancel each other. Also shown in Fig. 2.5 with the experimental point is a curve of $\cos^2\theta$ (where θ is the incident polarization of angle with respect to the x_1 axis). This indicates that only the component of the incident light along the x_1 axis "sees" the constructive contributions from both the surface ripple and the index change in the medium.

ALIPPI et al. [1971] have reported recently on the contributions of the surface ripple and index change in the substrate to light diffraction by acoustic surface waves in quartz.

2.1.3.3. Diffraction efficiency versus acoustic power

Since the acoustic power density P_d is proportional to the square of acous-

tic strain, the power density P_d is proportional to $(K\delta_1)^2$. Therefore, from eq. (2.24) it is noted that I_1 is proportional to P_d. For a given power density the first order diffracted light is proportional to K^2 or to the inverse of the square of acoustic frequency (LEAN and POWELL [1970]).

To show this relationship we have measured I_1/I_0 as a function of acoustic surface wave power density in a YZ $LiNbO_3$ substrate for the transmission case at several input frequencies. The results are shown in Fig. 2.6. The in-

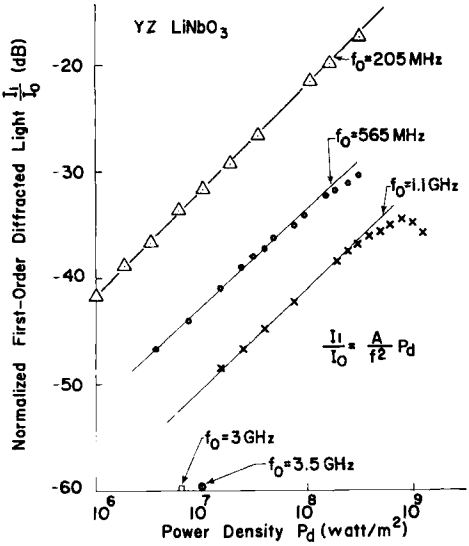


Fig. 2.6. Experimental results of I_1/I_0 as a function acoustic surface wave power density in a YZ $LiNbO_3$ substrate for the transmission case at several input frequencies.

cident light is polarized along the x_1 axis which gives a maximum diffraction efficiency in transmission. The linear relation between I_1/I_0 and input acoustic power density exists as low input power levels. As the input acoustic power density increases, one observes a saturation phenomenon in which the power of the fundamental frequency is transformed to nonlinearly generated harmonics (see Section 2.5).

2.2. DETECTION BY BEAM TILTING

The beam tilting technique to detect the surface ripple was discussed by ADLER et al. [1968] and was described in detail by WHITMAN and KORPEL [1969]. If a laser with a square cross section is focused into a spot small compared to the acoustic wavelength on the surface, the surface ripple, given by eq. (1.2) causes the illuminated portion of the surface to tilt by an angle

$$\beta = \frac{\mathrm{d}\delta}{\mathrm{d}x}\bigg|_{x=x_0} = \frac{2\pi}{\Lambda}\delta_1 \sin(\omega_s t + \phi) \tag{2.42}$$

where $\phi = -2\pi x_0/\Lambda$ is the initial phase. By placing a knife edge in the path of the reflected beam as shown in Fig. 2.7, the beam tilting can be directly

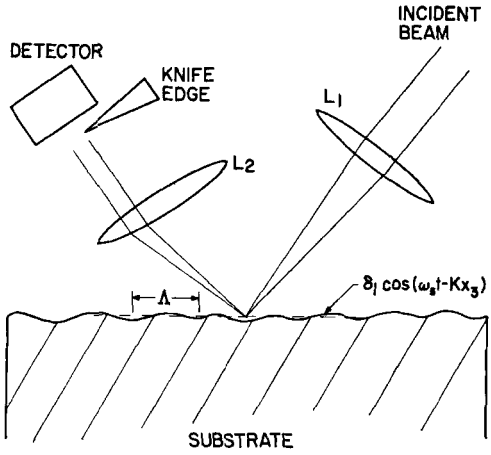


Fig. 2.7. Detection scheme of surface ripple by beam tilting techniques (after ADLER et al. [1968]).

translated to intensity variation detected by a photodetector right after the knife edge. With the laser spot focused into a $\frac{1}{2}\Lambda$ spot, the signal power P_{sig} detected by the photodetector is:

$$P_{sig} = P_0 2\pi(\delta_1/\Lambda)\cos\omega_s t \tag{2.43}$$

where P_0 is the total power of the reflected beam. This beam tilting technique has been used to visualize and display 8 MHz acoustic surface waves on steel containing a variety of surface wave deflecting obstacles (ADLER et al. [1968]). Fig. 2.8 shows the display of the acoustic surface waves on a steel surface affected by two deep grooves. Assuming that the detector is shot noise limited, the minimum detectable surface displacement by this scheme is:

$$\delta_{min} = (2eB/\alpha P_0)^{\frac{1}{2}}(\lambda/2\pi) \tag{2.44}$$

where e is the charge of an electron. B is the system bandwidth and α is the sensitivity of the photodetector. Using a laser power $P_0 = 10$ mW and $B = 1$ MHz, this system is capable of detecting surface displacement on the order of 10^{-12} m.

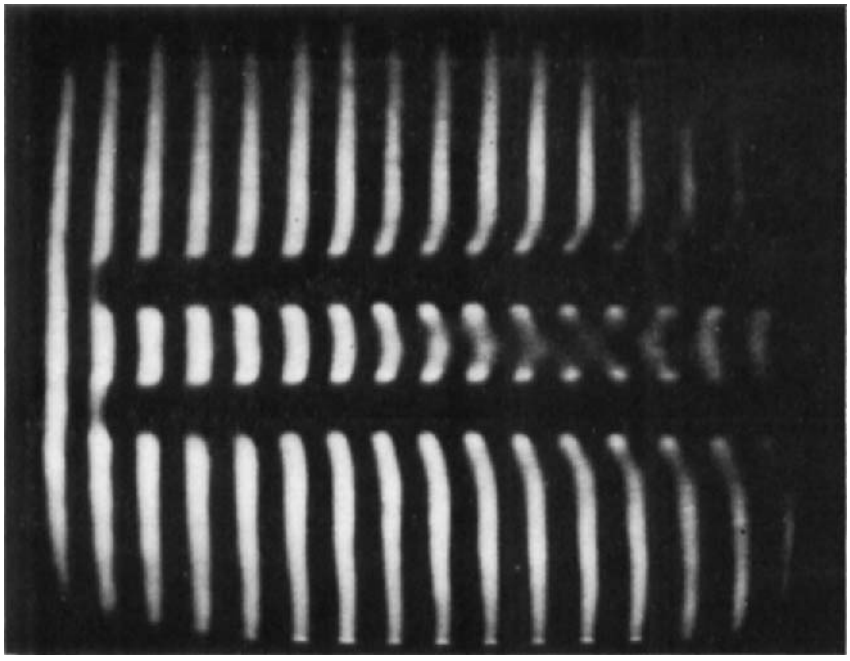


Fig. 2.8. Display of the acoustic surface waves at 8 MHz on a steel surface affected by two deep grooves (after ADLER et al. [1968]).

The same technique has also been used to detect surface perturbations of an acoustic hologram of bulk acoustic waves. The frequency limitation of this scheme depends on how well the laser spot can be focused and on the critical alignment of the knife edge. For a focused spot of 1 μm in width the beam tilting detection scheme may be used to observe acoustic surface waves of wavelengths down to approximately 2 μm.

2.3. OPTICAL PROBING OF ACOUSTIC SURFACE WAVES

Since the diffraction angle of eq. (2.21b) depends on the acoustic wavelength, the phase velocity of acoustic surface waves can be determined by measuring the diffraction angle (AUTH and MAYER [1967]; KROKSTAD and SVAASAND [1967]). The intensity of the first diffracted order is proportional to the average δ_1^2 over the laser spot as long as the spot size is many acoustic wavelengths wide. Thus monitoring the first order diffracted light is a convenient way to study the propagation characteristics of acoustic surface waves at high frequencies.

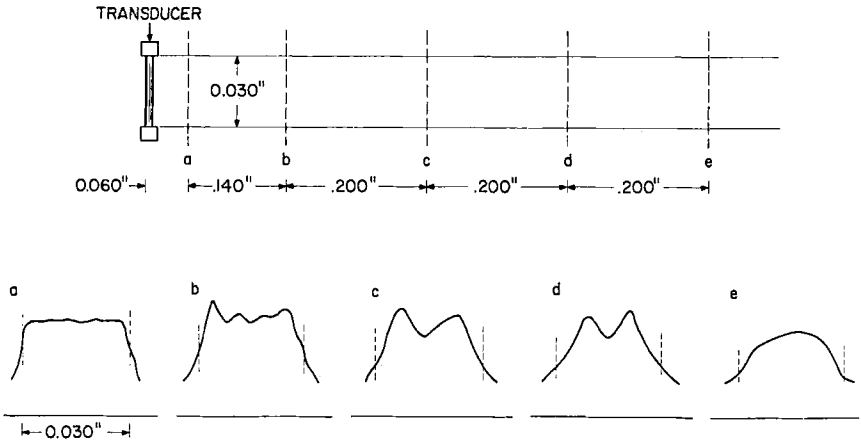


Fig. 2.9. Energy profiles of an acoustic beam as a function of distance from the input transducer to show the effect of diffraction. The transducer had an aperture of $8.8. \times 10^{-2}$ cm and operated at the acoustic wavelength of 17 μm.

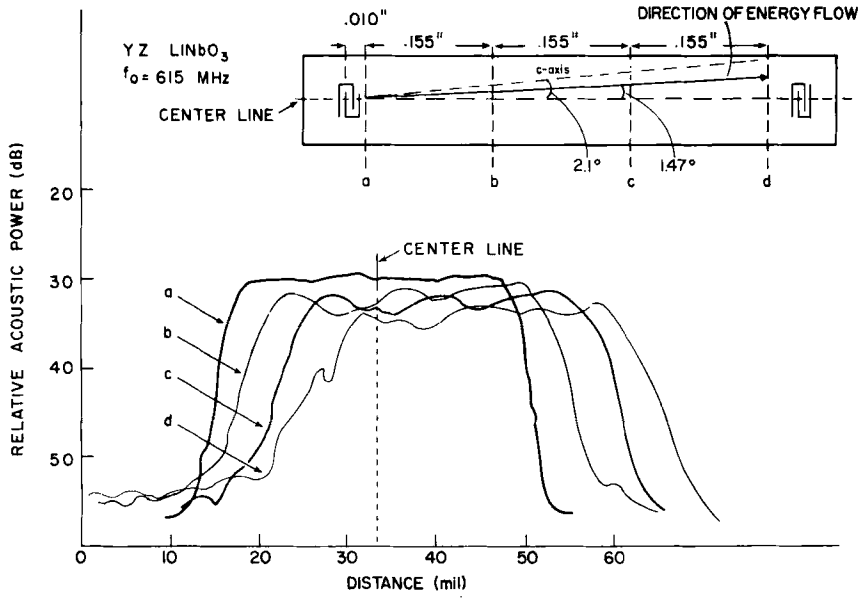


Fig. 2.10. Energy profiles of an acoustic beam as a function of distance from the input transducer to show the beam steering effect. The transducer had an aperture of 8.8×10^{-2} cm and operated at 5 μm acoustic wavelength.

2.3.1. *Beam diffraction and steering*

By focusing the laser beam to a small spot compared to the width of the ASW and then scanning across the acoustic beams, the energy profile of the ASW can be measured. Figure 2.9 shows the energy profiles of the acoustic beam as a function of distance from a transducer having an aperture of 8.8×10^{-2} cm and at the acoustic wavelength of 17 μm. The results are analogous to Fresnel diffraction of electromagnetic waves from a finite aperture in the near field region ($R < W^2/\Lambda$, where R is the observation distance from the transducer; W the transducer width; and Λ the acoustic wavelength).

In an anisotropic substrate, the energy flow direction in general does not coincide with the propagation direction. This beam steering effect can easily be evaluated by the diffraction method. The results of such beam steering effects of a 615 MHz ASW in a $LiNbO_3$ substrate is shown in Fig. 2.10. The aperture of the transducer was 8.8×10^{-2} cm and the wavelength of the waves was reduced to only 5.0 μm so that the near field region would be extended further. As a result, no diffraction was observed in this case. However, at a distance of 1.18 cm from the input transducer, the center of the surface acoustic beam had shifted from the center line of the transducer by a distance of 3×10^{-2} cm, corresponding to a beam steering angle of 1.45°. The problems of beam diffraction and beam steering on different cuts of $LiNbO_3$, have been discussed by SLOBODNIK Jr. et al. [1970] and FARNELL [1970].

2.3.2. *Attenuation measurement*

The optical probe provides a most satisfactory way of measuring the attenuation of surface waves. By translating the substrate so that the laser beam scans along the direction of Rayleigh wave propagation and by monitoring the first order diffracted light, we can measure the attenuation directly without the effects of input and output acoustic transducers. However, the beam profile measurement in Fig. 2.9 suggests that the diameter of the laser beam must be equal to or larger than the acoustic beamwidth to avoid picking up the fine structure of the acoustic beam. Fig. 2.11 shows the measurement of intensities of the first order diffracted light by Rayleigh waves in a YZ $LiNbO_3$ at frequencies varying from 1090 MHz to 3.45 GHz as a function of distance from the input transducer. The straight lines drawn through the measured points in semilogarithmic scale give the attenuation of the waves at each frequency.

The attenuation of Rayleigh waves, which is usually higher than that of

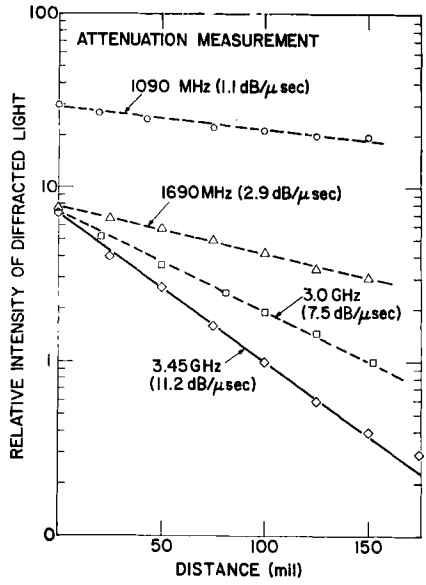


Fig. 2.11. The first order diffracted light intensity as a function of distance from the input transducer.

the corresponding bulk waves, has been one of the concerns in the development of surface acoustic wave technology. Besides the scattering by thermal phonons and crystal imperfections, Rayleigh waves suffer additional attenuation due to surface imperfections such as scratches and damaged layers, and also due to inherent nonlinear effects. Improvements in mechanical and chemical polishing techniques may soon minimize the attenuation due to surface imperfections. However, until then, the attenuation of surface acoustic waves will depend largely on the preparation of each individual substrate. Losses due to nonlinearly generated harmonics appear to be important for surface acoustic attenuation at higher frequencies, additional attenuation due to surface imperfections such as scratches and damaged layers, and also due to inherent nonlinear effects. Improvements in mechanical and chemical polishing techniques may soon minimize the attenuation due to surface imperfections. However, until then, the attenuation of surface acoustic waves will depend largely on the preparation of each individual substrate. Losses due to nonlinearly generated harmonics appear to be another important factor for surface acoustic attenuation especially at higher frequencies. In addition to thermal phonon attenuation which is temperature dependent and surface scattering (temperature independent), there is also a significant loss

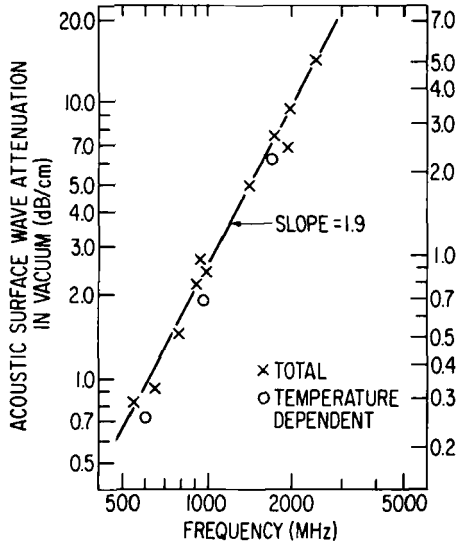


Fig. 2.12. Attenuation of acoustic surface wave in *YZ* $LiNbO_3$ substrate as a function of frequency (after SLOBODNIK Jr. et al. [1970]).

arising from air loading (CARR [1970]). SLOBODNIK Jr. et al. [1970] have used a laser probe to measure the temperature dependent and temperature independent attenuation of Rayleigh waves on *YZ* $LiNbO_3$ substrates. The results of their measurements are given in Fig. 2.12. The power levels were kept low enough to avoid any nonlinear effects. At 1GHz, they obtained attenuation values of 0.7 dB/μsec due to temperature dependent loss, 0.2 dB/μsec due to temperature independent losses and 0.2 dB/μsec due to air loading.

2.3.3. *Optical superheterodyning*

The first order diffraction intensity provides the amplitude information of the acoustic waves "seen" by the laser spot. However, if one is concerned about phase information as well, optical superheterodyning technique by mixing the diffracted beam with a reference beam is needed (WHITMAN and KORPEL [1969]; DE LA RUE et al. [1971]). One of such optical heterodyne system is shown in Fig. 2.13. A water filled Bragg cell driven at frequency f_s is used to modulate the incident light having a frequency f_0. By properly aligning the optical system, the reference beam at $f_0 - f_s$ and the probe beam at $f_0 + f_s$ strike at the same photodetector for heterodyning. The surface ripple of the acoustic surface wave, causes the probe beam to have sidebands

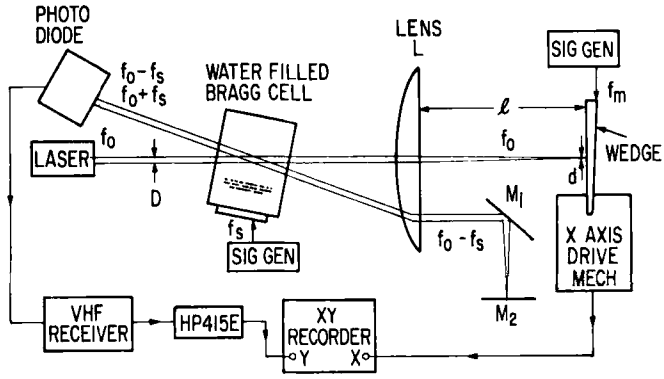


Fig. 2.13. An optical superheterodyning system (after WHITE and KORPEL [1969]).

at a frequency $f_0+f_s+f_m$, so that the heterodyne output of the photodiode contains sidebands at $2f_s+f_m$ with amplitudes proportional to the acoustic displacement. A narrow band receiver can be used to amplify and detect the signal at $2f_s-f_m$. Assuming the photodetector is shot noise limited the minimum detectable displacement is the same as that in eq. (2.44).

The system shown in Fig. 2.13 can be used to visualize and to display the surface waves by scanning the sample or the laser beam. It has also been applied to measure the phase and group velocities of acoustic surface modes in waveguide structures. The diffracted light beam radiates at an angle $\theta = \lambda/\Lambda$ with respect to incident probe beam. In order to have the optical mixing take place it is required that the angle θ is smaller than the angle ϕ of the focused prove beam. The angle ϕ is given by $\phi = D/F$, where D is the laser beam diameter and F is the focal distance of the lens. This limits the shortest acoustic wavelength Λ_{min} which can be detected by the scheme. By the condition that $\theta < \phi$, we have:

$$\Lambda_{min} > \lambda F/D \tag{2.45}$$

with $D = 1$ cm and $F = 10$ cm, Λ_{min} can be about 7 μm with a He-Ne incident laser light. At higher acoustic frequencies, the diffracted orders are discrete, the signal beam and reference beam (can be the zero order beam) have to be superimposed with critical alignment at the photodetector to obtain optical heterodyning detection. The frequency of acoustic waves of this system is limited by the response time of the available photodetector. De la Rue et al. have discussed and compared in detail the different coherent optical detection schemes (1971).

2.4. LIGHT DIFFRACTION BY NONSINUSOIDAL ACOUSTIC SURFACE WAVES

So far we have discussed the light diffraction by a sinusoidal acoustic surface wave, which results in the usual Bessel function distribution for the amplitude of the diffracted orders. If the acoustic surface waves are non-sinusoidal, and the surface ripples can be expanded as:

$$\delta = \sum_q \delta_q \sin\left[q(\omega_s t - Kx_3) + \phi_q\right] \tag{2.46}$$

where δ_q is the peak amplitude of the q^{th} harmonic and ϕ_q is the phase of the q^{th} harmonic. Each harmonic diffracts the incident light into a diffracted field with discrete orders as shown in eq. (2.19). Neglecting the multiple diffraction the total diffracted field from a non-sinusoidal acoustic surface wave is the sum of fields due to all harmonics:

$$E_2(p) = \sum_r A_r \exp\left\{j[(\omega + r\omega_s)t - k_d R_0\right\} \tag{2.47a}$$

where:

$$A_r = \sum_{q \cdot m} \frac{j^{m+1} E_0 BD}{2\pi R_0 \cos\theta_i} (1 - mQ_q/\alpha_q) F J_m(\alpha_q), \tag{2.47b}$$

$$\alpha_q = (k_d \cos\theta_d + k_i \cos\theta_i)\delta_q, \tag{2.48}$$

$$B = (1+R)k_d \cos\theta_d - (1-R)k_i \cos\theta_i, \tag{2.14}$$

$$Q_q = (-\delta_q K)(1+R)(k_d \sin\theta_d + K_i \sin\theta_i)/B, \tag{2.49}$$

$$F = \frac{\cos\theta_i}{(\gamma - rK)D} \sin\left\{\frac{(\gamma - rK)D}{\cos\theta_i}\right\}, \tag{2.50a}$$

$$\gamma = k_d \sin\theta_d - k_i \sin\theta_i. \tag{2.50b}$$

The summation in (2.47b) is for all q and m under the condition $r = q \cdot m$. The reflection coefficient R in (2.14) and (2.49) is given by (2.10) or (2.11), depending on the incident polarization. The average intensity of the r^{th} order diffracted light is simply given by:

$$I_r = |A_r|^2 = \left[\sum_{q \cdot m} \frac{E_0 BDF}{2\pi R_0 \cos\theta_i} (1 - mQ_q/\alpha_q) J_m(\alpha_q)\right]^2 \tag{2.51}$$

where the summation is for all q and m so that $r = q \cdot m$. Each term in (2.51) represents the m^{th} order diffraction from the q^{th} harmonic of the acoustic surface waves.

NEIGHBORS and MAYER [1971] have recently calculated the light diffrac-

tion by a non-sinusoidal acoustic surface wave and included multiple diffraction. However, under the normal experimental condition α_q is usually much smaller than unity. The multiple diffraction can then be neglected. For example, with a power density of 10^8 watt/m^2 at $f_0 = 565$ MHz in a $LiNbO_3$ substrate, the measured first order diffracted light has a normalized intensity of $I_1/I_0 = 10^{-4}$ in the reflection case, where I_0 is the zero order diffracted light intensity. The second order diffraction from the fundamental has a calculated value of 2.5×10^{-9}. Such a small contribution from the higher order diffraction (the terms in (2.51) with $m > 1$) may be neglected compared with the first order diffraction of higher harmonics (the terms in (2.51) with $m = 1$). By the same reason, the contributions from the multiple diffraction are small and negligible. Equation (2.51) can be approximated by letting $m = 1$, $q = r$ and $J_1(\alpha_1) = \frac{1}{2}\alpha_r$;

$$I_r \approx \left(\frac{E_0 DFB}{4\pi R_0 \cos\theta_i}\right)^2 (\alpha_r - Q_r)^2 \tag{2.52a}$$

$$\approx I_0 F^2 T^2 \delta_r^2 \tag{2.52b}$$

and

$$T = \frac{B(k_d \cos\theta_i + k_i \cos\theta_i) + K(1+R)(k_d \sin\theta_d + k_i \sin\theta_i)}{2k_i \cos\theta_i} \tag{2.53}$$

where I_0, F and B are given by equations (2.25), (2.20b) and (2.14) respectively.

It is seen from (2.52b) that the r^{th} order diffracted light is directly proportional to the ripple amplitude induced by the r^{th} harmonic of the acoustic surface wave. By monitoring the outputs of I_r/I_0 along the proper angle θ_r (2.26), the amplitude δ_r and the power density of the r^{th} harmonic can be measured directly (LEAN et al. [1970], SLOBODNIK et al. [1970]).

Due to the fact acoustic surface waves concentrate their energies in a surface layer of about one acoustic wavelength, a low input power produces a sufficiently high power density so that nonlinear effects can be easily observed. LEAN and TSENG [1970] have used this optical probing technique to study the harmonic generation and parametric mixing of Rayleigh waves in $LiNbO_3$ substrates. Unlike the optical second-harmonic generation, the lack of dispersion of the Rayleigh waves ensures that the phase matching condition is met at all frequencies. Thus, large numbers of harmonics can in principle be generated. This fact together with the complexity of two-dimensional inhomogeneous waves makes theoretical calculation extremely complicated. Experimentally it is possible to observe directly the growth of each harmonic

as a function of interaction length and input pump power level by the optical probe.

Fig. 2.14 shows a set of experimental results of the light diffractions by acoustic surface wave pulses at a fundamental frequency of 615 MHz and its harmonics as a function of distance from the input transducer. The input

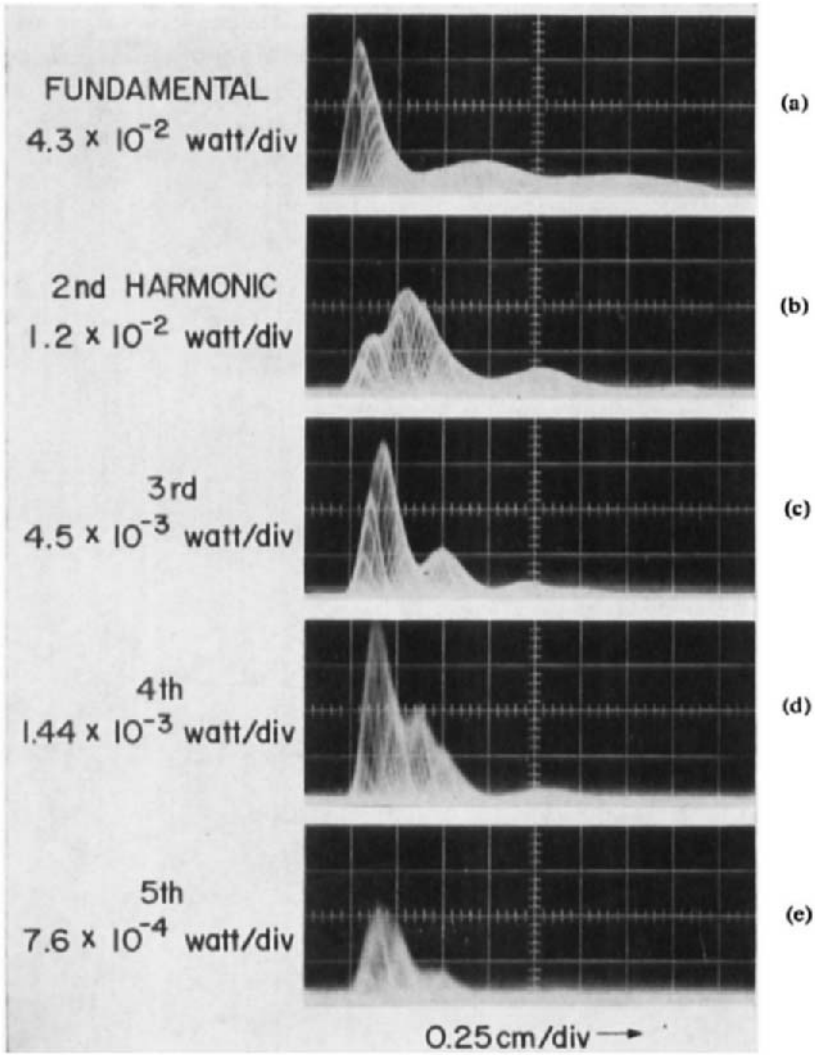


Fig. 2.14. Experimental results of the light diffraction by acoustic surface wave pulses at a fundamental frequency of 615 MHz and its harmonics as a function of distance from the input transducer.

acoustic pulse, which is less than $\frac{1}{2}$ μsec in width, has a peak power $P(f)$ at the input transducer. By placing the optical probe at the distance z away from transducer and monitoring the output pulses of the diffracted light at proper angles the intensities of the acoustic power of the fundamental frequency and each harmonic generated nonlinearly in the medium at the particular location z from the transducer can be determined. Fig. 2.14 is the multiple exposure of the diffracted pulses as the laser probe was moved along the path of propagation. The envelope of the peaks of the diffracted pulses indicates the spatial dependence of the fundamental signal and its harmonics as a function of interaction length.

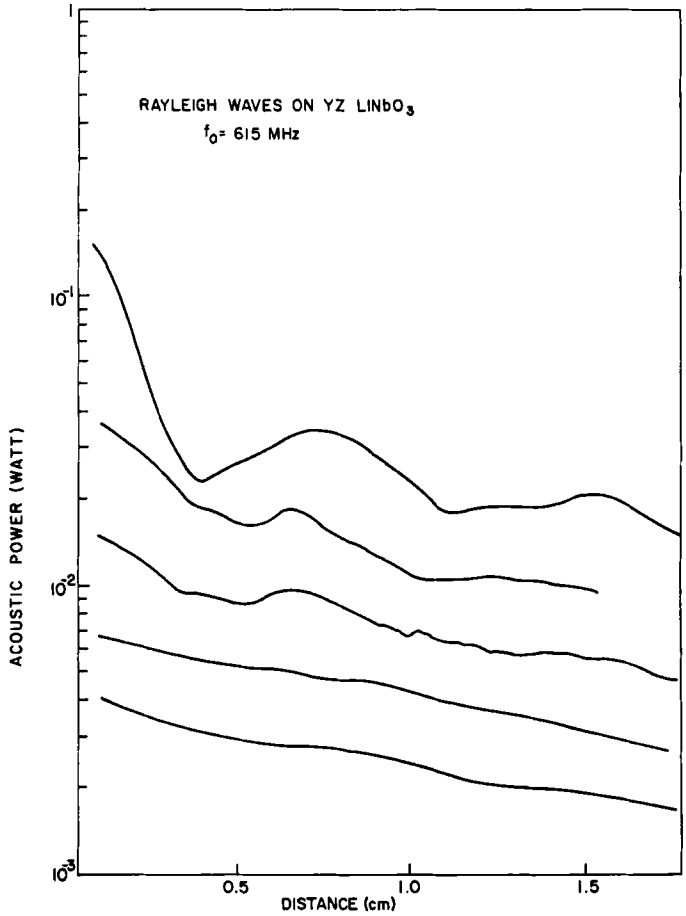


Fig. 2.15. The diffracted light intensities by acoustic waves at 615 MHz at different power levels as a function of distance from the input transducer.

It is interesting to notice that the flat portion in the curve for the second harmonic in Fig. 2.14b coincides with the peak of the third harmonic in Fig. 2.14c. At this distance the second harmonic is giving energy to the third harmonic by mixing with the fundamental. As the third harmonic decreases, the intensity of the second harmonic increases again due to the beating between the third harmonic and the fundamental. The input acoustic power at fundamental frequency $f_0 = 615$ MHz is about 1.5×10^{-1} W, which gives a power density of 3.1×10^8 W/m^2. As the input acoustic power level changes, the spatial dependences of the fundamental and its harmonics vary.

The observed spatial dependences of harmonics generated nonlinearly in $LiNbO_3$ substrate can be explained phenomenologically by a set of coupled amplitude equations which take into account both the transverse spatial variations of the Rayleigh wave harmonics and any phase mismatch between the fundamental and its harmonics due to a layer on a perfect substrate (LEAN [1971]).

At high power density, the acoustic surface waves depend on the frequency, the temperature as well as the power level of the input and the distance from the input transducer. The nonlinear effect that drains the fundamental signal into harmonics may be one of the important factors in the acoustic surface wave attenuation. For example, using the optical probe, the diffracted light intensities by the fundamental signal at 615 MHz at different power levels as a function of distance can be measured and plotted in semilogarithmic scale as in Fig. 2.15. The simple interpretation of attenuation which relates to an exponential decay in the direction of the propagation is no longer valid. It is even possible to have a local negative attenuation which is due to the increase in the fundamental signal from the parametric mixing of higher harmonic via the elastic nonlinearity of the substrate (LEAN and TSENG [1970]).

§ 3. Interaction of Optical Guided Waves and Acoustic Surface Waves

Light can be trapped in a thin film on a substrate. The index of refraction in the film must be higher than the surrounding media so that the light beam is trapped in the film by total internal reflection at interferes (SHUBERT and HARRIS [1968]). The field outside the film decays exponentially with distance from interfaces. Due to the fact that both optical guided waves and acoustic surface waves are propagating in the same thin film region, the periodic change in the index of refraction induced by acoustic surface waves (as discussed in § 1) affects the optical guided waves and results in strong interaction between them. The interaction mechanism for the guided waves are

analog to the more familiar bulk acousto-optic interaction. Depending on the acoustic wavelength and the width of the acoustic beam width, the diffraction can be either a Raman-Nath type (RAMAN and NATH [1935]) of diffraction, which is a phase grating diffraction as discussed in § 2, or a Bragg type of diffraction (QUATE et al. [1965]). However, there are also significant differences between bulk wave and surface wave interactions:

1) The guided waves are no longer TEM waves and have transverse field distribution (either TE or TM modes) while the bulk waves are normally TEM waves;

2) The guided waves have most of their energy confined to the thin film region having a thickness of the order of one micrometer so that the power densities of the guided waves can be three or four orders of magnitude higher than the corresponding bulk wave case which has a beam diameter of a few millimeter;

3) These guided waves are dispersive (both optics and acoustics) depending on the thickness of the film.

The high power densities and long interaction length of guided waves result in very efficient interaction between optical guided waves and acoustic surface waves as reported by KUHN et al. [1970]. The dispersiveness due to the finite thickness of the film provides means to adjust the phase matching conditions which are required for acousto-optic interactions or electro-optic interactions. In this section, the theoretical calculations and experimental results of the strong interactions between optical guided waves and acoustic surface waves will be discussed.

3.1. ACOUSTO-OPTIC INTERACTIONS IN THIN FILMS – ANALYSIS

Consider the interaction between two optical guided modes and one acoustic surface wave as shown in Fig. 3.1. The optical guided waves propagate in a thin film with a higher index of refraction than the index of its substrate. It is known that many discrete modes can be supported in a thin film guide. The modes can be either transverse electric (TE) or transverse magnetic (TM) depending on whether the electric or the magnetic fields lie in the plane of the film and normal to the direction of propagation (COLLINS [1965]). Their properties are well known. For example, the electric field of the TE_m mode propagating along a direction having an angle θ_i with respect to the x_2-axis in the plane of the substrate surface as shown in Fig. 3.1 is given by:

$$\boldsymbol{E}_m = (\boldsymbol{x}_2 \sin\theta_i - \boldsymbol{x}_3 \cos\theta_i) A_m U_m(x_1) \times \exp\{j(\omega_m t - k_m \cos\theta_i x_2 - k_m \sin\theta_i x_3)\} \qquad (3.1)$$

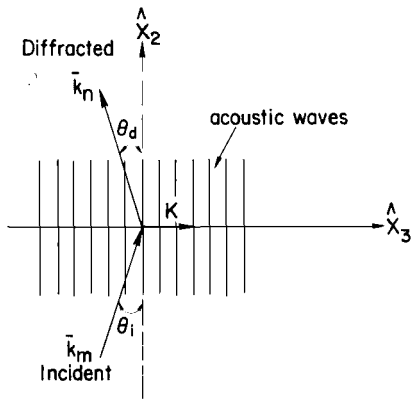


Fig. 3.1. Schematic of acousto-optic interactions in thin films, surface waves in thin films.

where $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are unit vectors; A_m is the peak amplitude; $U_m(x_1)$ is the transverse distribution function for the TE_m mode and $U_m(x_1)$ is chosen to be real; ω_m and k_m are the optical frequency and wavenumber of the mode. The functions of $U_m(x_1)$ and the dispersion relation of ω_m and k_m for both TE and TM modes can be found in a straightforward manner using the wave equation and matching boundary conditions in the interfaces (COLLINS [1965]).

Assuming that both the incident mode and diffracted mode are TE modes, in the analysis of the interaction as shown in Fig. 3.1, the incident mode has an electric field as given by eq. (3.1), while the diffracted mode has an electric field of:

$$\boldsymbol{E}_n = (\boldsymbol{x}_2 \sin\theta_d + \boldsymbol{x}_3 \cos\theta_d) A_n U_n(x_1) \times \exp\{j(\omega_n t - k_n \cos\theta_d x_2 + k_n \sin\theta_d x_3)\} \qquad (3.2)$$

where θ_d is the diffracted angle with respect to the x_2-axis. The diffracted mode can be either the same mode as the incident wave ($m = n$), or a different mode ($m \neq n$).

The acoustic surface wave with strain component S_{q3} propagates along the x_3-axis. The strain component S_{q3} is given by:

$$S_{q3} = \tfrac{1}{2} B_{q3} V_{q3}(x_1) \exp\{j(\omega_s t - K x_3)\} + \text{C.C.} \qquad q = 1, 2, 3 \qquad (3.3)$$

where B_{q3} is the peak amplitude, q indicates the particle displacement direction. All the strain components are propagating along the x_3-axis as indicated in Fig. 3.1. $V_{q3}(x_1)$ are the transverse distribution functions in the film and in the substrate. ω_s and K are the acoustic frequency and wavenumber

respectively. C.C. indicates the complex conjugate. The change in the dielectric constant induced by the strain wave through the photoelastic effect has been discussed in Section 1.1. In this case from (1.9) we have:

$$\Delta\varepsilon_{ab} = -\varepsilon_{aa}^2 P_{abq3} S_{q3}, \qquad a, b, q = 1, 2, 3 \tag{3.4}$$

where $\Delta\varepsilon_{ab}$ is the element of the change in the dielectric tensor $\Delta\boldsymbol{\varepsilon}$, ε_{aa} and P_{abq3} are the dielectric and photoelastic constants in the film and in the substrate.

The interaction of the incident light $\boldsymbol{E}_m$ and $\Delta\boldsymbol{\varepsilon}$ gives rise to an induced electric displacement $\boldsymbol{D}$ which has a component D_a given by:

$$\begin{aligned} D_a &= \Delta\varepsilon_{ab}(E_m)_b \qquad (3.5)\\ &= -\tfrac{1}{2}\varepsilon_{aa}^2 P_{abq3} B_{q3} V_{q3}(A_m)_b U_m \\ &\times \{\exp\{\mathrm{j}[(\omega_m+\omega_s)t - k_m \cos\theta_i x_2 - (k_m \sin\theta_i + K)x_3]\} \\ &+ \exp\{\mathrm{j}[(\omega_m-\omega_s)t - k_m \cos\theta_i x_2 - (k_m \sin\theta_i - K)x_3]\}\} \qquad (3.6) \end{aligned}$$

where the index a and b run from 1 to 3. $(E_m)_b$ is the b component of $\boldsymbol{E}_m$ with an amplitude component $(A_m)_b$, eqs. (3.1), (3.2), (3.3) and the relation of (3.4) have been used in obtaining the result of (3.6).

The induced displacement current $\partial\boldsymbol{D}/\partial t$ can be considered as the driving source for the diffracted TE_n mode. The power generated by the driving displacement current $\partial\boldsymbol{D}/\partial t$ can be considered as the driving source for the diffracted TE_n mode. The power generated by the driving displacement current $\partial\boldsymbol{D}/\partial t$ in a volume should be equal to the power carried away by the mode generated from the same region. This power balance leads to the following equation:

$$\oint_a (\boldsymbol{E}_n^* \times \boldsymbol{H}_n) \cdot \mathrm{d}\boldsymbol{a} = -\int_v \boldsymbol{E}_n^* \cdot \frac{\partial \boldsymbol{D}}{\partial t}\,\mathrm{d}v \tag{3.7}$$

where $\boldsymbol{E}_n$ and $\boldsymbol{H}_n$ are the electric and magnetic field associated with the generated TE_n mode. In a waveguide mode, the transverse field components E_{nt} and H_{nt} are related by a wave impedance. For TE modes, we have

$$E_{nt}/H_{nt} = \sqrt{(\mu/\varepsilon)}k/k_n. \tag{3.8}$$

The surface a encloses the volume v. KUHN [1969] has used the same argument to calculate the interaction of nonlinear optics in a finite geometry.

Consider a case with the interaction region defined by two surfaces located at ξ and $\xi+\Delta\xi$, where ξ is along the direction of propagation of the TE_n mode. Both surfaces are extended from zero to infinity in the x_1-direction. With these simple boundaries, the substitution of (3.2), (3.8) and the derivation of (3.6) into (3.7) results in the following equation:

$$A_n A_n^*|_{\xi}^{\xi+\Delta\xi} = -2\mathrm{j}\Gamma_{nm} A_m A_n^* \exp(\mathrm{j}\Delta\boldsymbol{k}\cdot\boldsymbol{r})\Delta\xi \tag{3.9}$$

where

$$\Gamma_{nm} = \frac{v_n k}{v4} \frac{(\sin\theta_i \sin\theta_d\, \varepsilon_{22} I_{22} + \cos\theta_i \cos\theta_d \varepsilon_{33} I_{33})}{\int_0^\infty U_n U_n^* \mathrm{d}x_1}, \tag{3.10a}$$

$$I_{22} = \int_0^\infty P_{22q3} B_{q3} V_{q3} U_m U_n^* \mathrm{d}x_1 \tag{3.10b}$$

$$I_{33} = \int_0^\infty P_{33q3} B_{q3} V_{q3} U_m U_n^* \mathrm{d}x_1 \tag{3.10c}$$

$$\Delta\boldsymbol{k}\cdot\boldsymbol{r} = \tag{3.11a}$$

$$\Delta k_2 x_2 + \Delta k_3 x_3 = (k_n \cos\theta_d - k_m \cos\theta_i)x_2 - (k_n \sin\theta_d + k_m \sin\theta_i - K)x_3 .$$

Eq. (3.9) is obtained under the condition that the output frequency is upshifted by ω_s, i.e.:

$$\omega_n = \omega_m + \omega_s . \tag{3.11b}$$

A similar equation can be obtained for the case of downshift in frequency in which:

$$\omega_n = \omega_m - \omega_s \tag{3.12a}$$

$$\Delta\boldsymbol{k}\cdot\boldsymbol{r} = \tag{3.12b}$$

$$\Delta k_2 x_2 + \Delta k_3 x_3 = (k_n \cos\theta_d - k_m \cos\theta_i)x_2 - (k_n \sin\theta_d + k_m \sin\theta_i + K)x_3 .$$

As $\Delta\xi$ approaches zero, eq. (3.9) becomes:

$$\frac{\mathrm{d}|A_n A_n^*|}{\mathrm{d}\xi} = -2\mathrm{j}\Gamma_{nm} A_m A_n^* \exp(\mathrm{j}\Delta\boldsymbol{k}\cdot\boldsymbol{r}). \tag{3.13}$$

A_n usually is a slowly varying function with no rapidly changing phase. Eq. (3.13) can be approximately written as:

$$\mathrm{d}A_n/\mathrm{d}\xi = -\mathrm{j}\Gamma_{nm} A_m \exp(\mathrm{j}\Delta\boldsymbol{k}\cdot\boldsymbol{r}). \tag{3.14}$$

Equation (3.14) describes the change in A_n as a function of ξ due to the input A_m and the coupling constant Γ_{nm}. The phase mismatch $\Delta\boldsymbol{k}$ is also included. If the input amplitude A_m is not depleted and can be assumed constant in the interaction region, the simple integration of (3.14) gives the result of A_n. However in case A_m depends also on A_n, a similar derivation considering $\boldsymbol{E}_n$ as the incident wave while $\boldsymbol{E}_m$ as the diffracted wave leads to the following equation:

$$\mathrm{d}A_m/\mathrm{d}\zeta = -\mathrm{j}\Gamma_{mn} A_n \exp(-\mathrm{j}\Delta\boldsymbol{k}\cdot\boldsymbol{r}) \tag{3.15}$$

where

$$\Gamma_{mn} = \frac{v_m k}{4v} \frac{(\sin\theta_i \sin\theta_d\, \varepsilon_{22} I_{22} + \cos\theta_i \cos\theta_d \varepsilon_{33} I_{33})}{\int_0^\infty U_m U_m^* \mathrm{d}x_1} \tag{3.16}$$

and $d\zeta$ is the increment in distance along the propagation direction of E_m mode. Δk is given by (3.11a) or (3.12b), and I_{22} and I_{33} are given by (3.10b) and (3.10c). Equations (3.14) and (3.15) are the two coupled amplitude equations relating A_n and A_m as a function of interaction length. One of the difficulties in solving the coupled amplitude equations in (3.14) and (3.15) is the different spatial directions ξ and ζ for both the incident and the diffracted light. However, in case the acoustic beam is wide enough and only the intensity variation of the light along the direction of acoustic wave fronts is of interest, we can simply let:

$$x_2 = \xi \cos \theta_d, \qquad x_2 = \zeta \cos \theta_i. \tag{3.17}$$

Equations (3.14) and (3.15) become:

$$\frac{dA_n}{dx_2} = -j(\Gamma_{nm} A_m / \cos \theta_d) \exp \{j(\Delta k_3 x_3 + \Delta k_2 x_2)\} \tag{3.18}$$

$$\frac{dA_m}{dx_2} = -j(\Gamma_{mn} A_n / \cos \theta_i) \exp \{-j(\Delta k_3 x_3 + \Delta k_2 x_2)\}. \tag{3.19}$$

The solution of the coupled amplitude equations (3.18) and (3.19) with boundary conditions that at $x_2 = 0$,

$$A_n = 0 \tag{3.20a}$$

and

$$A_m = A_m(0) \tag{3.20b}$$

can be readily obtained:

$$A_n(x_2) = \frac{-jA_m(0)\Gamma_{nm}}{T \cos \theta_d} \exp \{j(\tfrac{1}{2}\Delta k_2 x_2 + \Delta k_3 x_3)\} \sin (Tx_3) \tag{3.21}$$

where:

$$T^2 = |\Gamma_{nm} \Gamma_{mn} / (\cos \theta_i \cos \theta_d)| + \tfrac{1}{4}\Delta k_2^2. \tag{3.22}$$

The intensity of the diffracted light, which is proportional to $|A_n|^2$, at a distance $x_2 = L$ is given by:

$$\frac{I_n(L)}{I_m(0)} = |\Gamma_{nm} / \cos \theta_d|^2 L^2 \frac{\sin^2 (TL)}{(TL)^2} \tag{3.23}$$

when $I_m(0)$ is the incident intensity of the input TE_m mode.

Similar solution of the interaction between two TM modes or between one TE and one TM mode by the acoustic surface wave can also be obtained. The conversion between one TE mode and one TM mode requires proper photoelastic constants which will rotate the polarization of the incident mode

to the diffracted mode. The results obtained in this section are valid only for the case of collinear interaction in which $\theta_i = \theta_d = 90°$ or in cases of small incident and diffracted angles so that only one direction dependence of the optical beams either in x_2 or in x_3 is important. A more regiorous numerical calculation on guided wave theory of light diffraction by acoustic microwaves has been reported by CHU and TAMIR [1969]. It is also possible to couple the guided modes and the radiation modes which leak outside the film guide by acoustic waves as discussed by CHANG [1971].

3.2. ACOUSTO-OPTIC INTERACTIONS IN THIN FILMS – EXPERIMENTAL RESULTS

The interactions of optical guided waves in thin films and acoustic surface waves have been demonstrated in two recent experiments. One is the mode conversion between two collinear TE modes or two TM modes by a collinear acoustic surface wave (KUHN et al. [1971]); one is the Bragg diffraction of a guided mode by an acoustic surface wave (KUHN et al. [1970]). In this section, the detailed experimental results with theoretical interpretations of these two experiments will be discussed.

3.2.1. *Collinear interaction*

The experimental setup is shown in Fig. 3.2. The optical guided modes were propagating in a glass film on an aluminium coated lithium niobate ($LiNbO_3$) substrate. Grating couplers (DAKSS et al. [1970]) were used to couple both TE_1 and TE_3 modes of a 6328 Å laser beam into and out of the thin film guide. An acoustic surface wave, launched on the substrate by

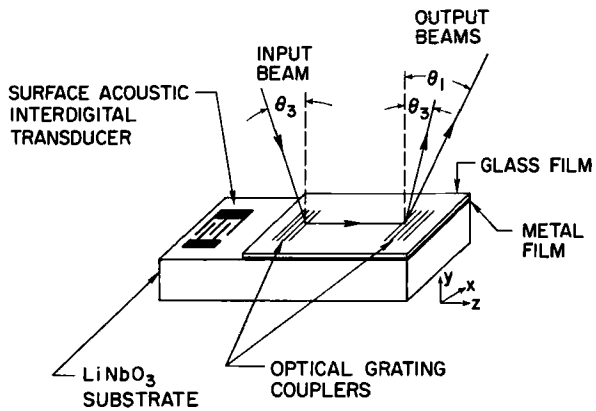


Fig. 3.2. Schematic of mode conversion experiment based on collinear interaction of optical guided waves and acoustic surface waves.

interdigital transducers propagated across the surface forming a strain wave in the film. When phase matching conditions of eq. (3.11, a and b) were satisfied as indicated in Fig. 3.3 coupling was observed from one waveguide

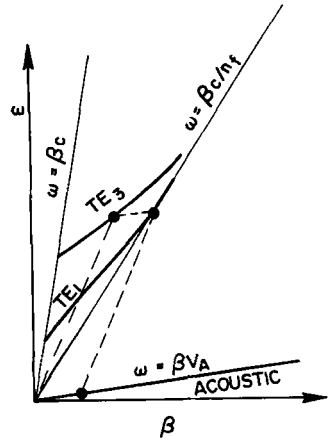


Fig. 3.3. Phase matching conditions for mode conversion.

mode to the other. Fig. 3.4 shows the experimental results. The bottom two traces show the depletion of the TE_1 mode and the augmentation of the TE_1 mode as the frequency was swept through the phase matching condition. The top trace shows the output of an optical probe (as discussed in Section

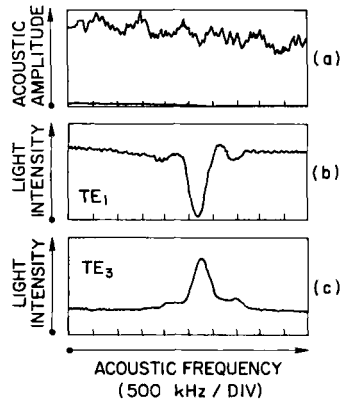


Fig. 3.4. Experimental results of mode conversion (a) optical probe of acoustic strain. The zero level is indicated at the bottom (b) and (c). Depletion of the TE_3 output and addition to the TE output as the acoustic frequency is varied over the region where phase matching is satisfied. The center frequency of the maximum interaction is $f_0 = 320$ MHz.

2.3) which was used to monitor the amplitude of the acoustic surface wave. The depletion of the TE_1 mode was about 55 % at the center frequency of 320 MHz for an acoustic surface wave with a surface displacement of 7.5 Å. The surface displacement was measured by the optical probing technique described in Section 2.3.

Assuming the spatial dependence of the TE_1 and TE_3 modes as shown in Fig. 3.2, $U_1(x_1)$ and $U_3(x_1)$ can be written as:

$$U_m(x_1) = \begin{cases} \sin h_m d \exp\left[-p_m(x_1-d)\right] & x_1 \geqq d \\ \sin h_m x_1 & 0 \leqq x_1 < d \\ 0 & x_1 \leqq 0 \end{cases} \qquad m = 1, 3 \tag{3.24}$$

where d is the thickness of the glass film. The parameters h_m and p_m are determined from the dispersion relations for TE modes;

$$p_m d = -h_m d \cot h_m d \tag{3.25a}$$

$$(p_m d)^2 + (h_m d)^2 = (n^2-1)(k_0 d)^2 \qquad m = 1, 3 \tag{3.25b}$$

where n is the index of refraction of the film.

The acoustic wave is regarded as having a constant amplitude over the thickness of the film since the thickness is only a small fraction of the acoustic wavelength, i.e.:

$$V_{q3}(x_1) = \begin{cases} 0 & x_1 \geqq d \\ 1 & 0 \leqq x_1 < d. \end{cases} \tag{3.26}$$

Substituting (3.24) and (3.26) into (3.14), (3.19), with $\theta_i = -\theta_d = 90°$ and (3.23), we can obtain the efficiency of the collinear mode conversion as a function of acoustic frequency.

$$\frac{I_1(L)}{I_3(0)} = |\Gamma_{13}|^2 \frac{\sin^2 TL}{T^2} \tag{3.27}$$

where

$$T^2 = |\Gamma_{13}\Gamma_{31}| + \tfrac{1}{4}\Delta k_3^2 \tag{3.28a}$$

$$\Gamma_{13} = \frac{\pi}{2\lambda_0} n^3 pB \frac{p_1 h_1[(h_1+h_3)\sin(h_1-h_3)d - (h_1-h_3)\sin(h_1+h_3)d]}{(h_1+h_3)(h_1-h_3)[p_1(h_1 d - \frac{1}{2}\sin(2h_1 d)) + h_1 \sin h_1 d]} \tag{3.28b}$$

$$\Gamma_{31} = \frac{\pi}{2\lambda_0} n^3 pB \frac{p_3 h_3[(h_1+h_3)\sin(h_1-h_3)d - (h_1-h_3)\sin(h_1+h_3)d]}{(h_1+h_3)(h_1-h_3)[p_3(h_3 d - \frac{1}{2}\sin(2h_3 d)) + h_3 \sin h_3 d]} \tag{3.28c}$$

$$\Delta k_3 = 2\pi(f - f_0)/v. \tag{3.28d}$$

h_m and p_m ($m = 1, 3$) are given in eq. (3.25). n and p are the index of refraction and the photoelastic constant of the film respectively; B is the peak

strain amplitude in the film; v is the acoustic velocity in the film; f_0 is the center frequency for the exact phase matching condition. Fig. 3.5 shows the results of the calculation for the depletion of TE_1 mode based on eq. (3.27). The experimental points were from the top trace in Fig. 3.4. In the calculation the measured values of the film thickness ($d = 1.6$ μm), the index of refraction ($n = 1.57$), the propagation constant from TE_3 and TE_1 mode ($k_3 = 1.51 \times 10^5$ cm^{-1} and $k_1 = 1.56 \times 10^5$ cm^{-1}) and the peak conversion efficiency of 0.55 were used. By using an interaction length $L = 0.62$ cm, the agreement between the calculated and observed results of conversion efficiency as a function of acoustic frequency is very good. The physical separation between the input and output grating couplers of the sample was about 0.8 cm. Probably due to the glass film nonuniformity, not the total available interaction length was used in the interaction.

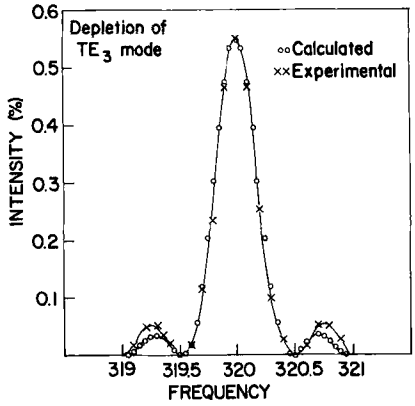


Fig. 3.5. The depletion of TE_3 mode in the mode conversion experiment with the calculation curve based on eq. (3.27).

In order to match the peak conversion efficiency of 0.55, the product of the photoelastic constant P and the strain B in the film was calculated to be about 1.62×10^{-5}. Using the measured strain value $B = 4 \times 10^{-4}$, we have the photoelastic constant calculated to be $P = 0.138$ which is a reasonable value for a glass film.

3.2.2. *Bragg deflection of optical guided waves by acoustic surface waves*

Optical Bragg diffraction by acoustic waves has been widely used in laser deflection schemes (GORDON [1966]). The same deflection occurs for optical guided waves and acoustic surface waves. Under the Bragg condition that

$$\sin \theta_B = \tfrac{1}{2}K/k \tag{3.29}$$

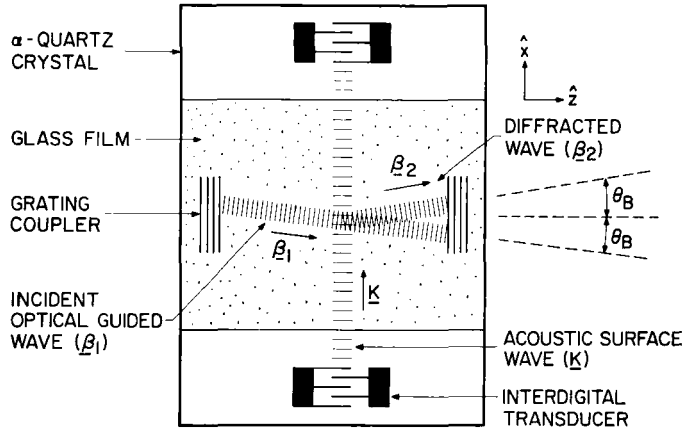


Fig. 3.6. Bragg deflection of optical guided waves by acoustic surface waves.

where θ_B is the incident and diffracted angle of the optical modes with respect to the acoustic wave fronts, the optical guided mode can be deflected. Eq. (3.29) is obtained by letting $\Delta k_3 = 0$ in eq. (3.11a). Using a configuration as shown in Fig. 3.6 in which a glass film (thickness $d = 0.8$ μm, index of refraction $n = 1.73$) was deposited in an α-quartz crystal, an incident TE mode was efficiently diffracted by an acoustic surface wave at $f = 191$ MHz under Bragg condition (KUHN et al. [1970]). The diffraction efficiency as high as 90% has been observed for an input acoustic power less than one watt.

At the perfect phase matching condition, the intensity of the diffracted light is given (from eq. (3.23) and eq. (3.10a) with $\theta_i = \theta_d = 0.037$) by:

$$I_d/I_i = \sin^2 \Gamma L \tag{3.30}$$

where

$$\Gamma = \frac{\pi}{2\lambda} n^3 P B \tag{3.31a}$$

$$L = W/\cos \theta_B \tag{3.31b}$$

where I_d is the intensity of the diffracted light; I_i is the incident light; P is the corresponding photoeleastic constant in the film; B is the peak amplitude of the strain wave; W is the width of the acoustic beam. Fig. 3.7 shows the calculated curve together with the experimental points for the diffraction efficiency as a function of acoustic strain. This efficient deflection of optical guided waves by acoustic surface waves has practical device applications as deflectors and modulators in integrated optics.

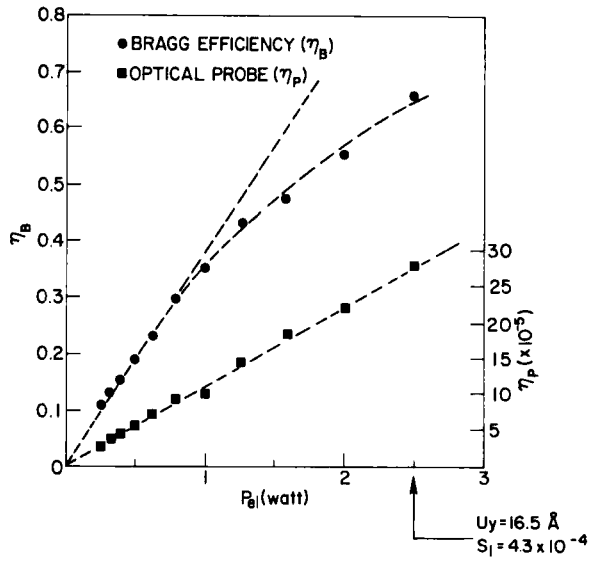


Fig. 3.7. Diffraction efficiency of Bragg deflection of optical guided waves as a function of acoustic strain.

§ 4. Conclusion

In this paper we have discussed two major areas in interaction of light and acoustic surface waves. One is the light diffraction by acoustic surface waves and the other is the interactions of optical guided waves and acoustic surface waves. The study of light diffraction by acoustic surface waves results in a convenient optical probing technique to study the propagation characteristics of acoustic surface waves which are needed for the design and utilization of acoustic surface wave devices. We have discussed the contributions due to the surface ripple and the periodic index variation in the substrate for the diffracted light in the reflection case and in the transmission case. The detail calculations of light diffraction efficiencies as a function of incident angles have been made. Experimental results for the $m = \pm 1$ order diffracted light by acoustic surface waves in a YZ $LiNbO_3$ or a quartz substrate have been obtained to compare with the calculated values. With the knowledge of the relationship of the diffracted light and the acoustic surface wave signals, some of the propagation properties of acoustic surface waves in $LiNbO_3$ substrate have been probed and measured.

In cases of nonsinusoidal acoustic surface waves, it has been shown that the diffracted intensity for each order can be correlated to the power of the

corresponding harmonic in acoustic surface waves. This leads to the study of nonlinear effects in acoustic surface waves by the optical probe. One useful application of this nonlinear effect study is the sensitive technique to measure thin film elastic properties (LEAN and POWELL [1971]).

The sensitivity of the optical probe is capable of detecting the surface ripple with a peak amplitude on the order of fraction of Å. Optical heterodyne technique as discussed in Section 2.3.3 can improve the detection sensitivity more with additional advantages of obtaining the phase information as well.

In § 3, the interactions of optical guided waves and acoustic surface waves have been calculated with the comparable of experimental results. The efficient interaction, which has important applications as efficient real time modulators for optical guided waves, is one of the reasons for the recent exciting interests in integrated optics using optical guided waves for optical signal processings and communication.

Acknowledgements

The author is deeply indebted to Professor A. Tonning for his critical comments and helpful suggestions in the theoretical calculations in § 1. The author would also like to thank his colleagues P. Zory, L. Kuhn and C. Powell for their stimulating and fruitful discussions and suggestions.

References

ADLER, R., A. KORPEL and P. DESMARES, 1968, IEEE Trans. Sonics and Ultrasonics, **SU-15**, 157.

ALIPPI, A., A. PALMA, L. PALMIERI and G. SOCINO 1971, Appl. Phys. Lett. **18**, 552.

AUTH, D. C. and W. G. MAYER, 1967, J. Appl. Phys. **58**, 5138–5140.

BECKMAN, P. and A. SPIZZICHINO, 1963, The Scattering of Electromagnetic Waves from Rough Surfaces (Macmillan, New York).

CARR, P. H. 1969, IEEE Trans. Microwave Theory Tech. **17**, 848.

CHANG, W., 1971, IEEE J. Quantum Electron. **QE-7**, No. 4, 167–170.

CHU, R. S. and T. TAMIR, 1969, IEEE Trans. Microwave Theory Tech. **MTT-17**, 1002.

COLLIN, R. E., 1960, Field Theory of Guided Waves (McGraw Hill, New York).

DAKSS, M. L., L. KUHN, P. F. HEIDRICH and B. A. SCOTT, 1970, Appl. Phys. Lett. **16**, 523.

DAMON, R. W., W. T. MALONEY and D. H. MCMAHON, 1970, Physical Acoustics, Vol. VII, Ch. 5.

DE LA RUE, R. HUMPHRYES, I. MASON and E. A. ASH, J.I.E.E. (England), to be published.

DRANSFELD, K. and E. SALZMANN, 1970, Physical Acoustics, Vol. VII, Ch. 4.

EWING, W. M., W. S. JARDETZKY and F. PRESS, 1957, Elastic Waves in Layered Media (McGraw Hill, New York).

FARNELL, G. W., 1970, Physical Acoustics, Vol. VI, Ch. 3.

FARNELL, G. W. and M. S. KHARUSI, 1971, IEEE Trans. Sonics and Ultrasonics, **SU-18**, 35–42.

GOELL, J. E. and R. D. STANDLEY, 1970, Proc. IEEE **58**, 1504.

GORDON, E. I., 1966, Proc. IEEE **54**, 1391.
IPPEN, E. P., 1967, Proc. IEEE **55**, 248.
KINO, G. and H. MATHEWS, 1971, IEEE Spectrum **8**, 22–35.
KROKSTAD, J. and L. O. SVAASAND, 1967 Appl. Phys. Lett. **11**, 155.
KUHN, L., 1969, J. Quantum Electron. **QE-5**, 383.
KUHN, L., M. L. DAKSS, P. F. HEIDRICH and B. A. SCOTT, 1970, Appl. Phys. Lett. **17**, 265.
KUHN, L., P. F. HEIDRICH and E. G. LEAN, 1971, Appl. Phys. Lett.
LEAN, E. G. H., C. C. TSENG and C. G. POWELL, 1970, Appl. Phys. Lett. **16**, 32–35.
LEAN, E. G. H. and C. C. TSENG, 1970, J. Appl. Phys. **41**, 3912–3917.
LEAN, E. G. H. and C. G. POWELL, 1971, Appl. Phys. Lett. **19**, 356.
LEAN, E. G. H. and C. G. POWELL, 1970, Proc. IEEE **58**, 1939–1947.
LOVE, A. E. H., 1911, Some Problems in Geodynamics (Cambridge University Press).
MAYER, W. G., G. B. LAMERS and D. AUTH, 1967, J. Acoust. Soc. Am. **42**, 1255–1257.
NEIGHBORS, T. H. and W. G. MAYER, 1971, J. Appl. Phys. **42**, 3670.
NYE, J. F., 1957, Physical Properties of Crystals (Oxford, Clarendon Press) part 4, p. 241.
QUATE, C. F., C. D. W. WILKINSON and D. K. WINSLOW, 1965, IEEE Proc. **53**, 1604.
RAMON, C. V. and N. S. N. NATH, 1935, Proc. Ind. Acad. Sci. **2A**, 413.
RAYLEIGH, Lord, 1885, Proc. London Math. Soc. **17**, 4–11.
SALZMANN, E. and D. WEISMANN, 1969, J. Appl. Phys. **40**, 3408–3409.
SHUBERT, R. and J. H. HARRIS, 1968, IEEE Trans. MMT **16**, 1048.
SLOBODNIK Jr., A. J., 1969, Appl. Phys. Lett. **14**, 94.
SLOBODNIK Jr., A. J., 1969, J. Acous. Soc. Am. **48**, 203.
SLOBODNIK Jr., A. J., P. H. CARR and A. J. BUDREAU, 1970, J. Appl. Phys. **41**, 4381.
SLOBODNIK Jr., A. J., 1970, Proc. IEEE **58**, 488–490.
STERN, C., 1969, IEEE Trans. Microwave Theory Tech. MTT **17**, 835.
VIKTOROV, I. A., 1967, Rayleigh and Lamb Waves, Physical Theory and Applications, Transl. from Russian (Plenum Press, New York).
WHITE, R. M. and F. M. VOLTMER, 1965, Appl. Phys. Lett. **7**, 314–316.
WHITE, R. M., 1970, Proc. IEEE **58**, 1238–1270.
WHITMAN, R. L. and A KORPEL, 1969, Appl. Optics **8**, 1967.
WILLARD, G. W., 1949, J. Acoust. Soc. Am. **21**, 101.
ZORY, P. and C. G. POWELL, 1971, Appl. Optics **10**, 2104.

E. WOLF, PROGRESS IN OPTICS XI © NORTH-HOLLAND 1973

IV

EVANESCENT WAVES IN OPTICAL IMAGING

BY

OLOF BRYNGDAHL

Xerox Palo Alto Research Center, Palo Alto, California, U.S.A.

CONTENTS

§ 1. Introduction

Solutions to the wave equation indicate the existence of different types of waves in optics. So, for example, we may have:

propagating waves with constant amplitude;

propagating waves with attenuated amplitude, either in space or time;

evanescent waves.

Two distinct classes of waves occur: homogeneous and inhomogeneous waves. A homogeneous (propagating) plane wave is a wave whose planes of constant amplitude and constant phase coincide. An inhomogeneous (evanescent) plane wave, on the other hand, is a wave whose planes of constant amplitude and constant phase do not coincide but cross each other at a finite angle.

The properties of the medium and the specific applicable boundary conditions determine the type of wave which will exist. In the following discussion, the existence and properties of the last type – the evanescent waves – in connection with the possibility of using them in optical imaging will be emphasized and discussed.

Evanescent waves have in the past frequently been regarded more as a mathematical tool than a physical phenomenon; they are encountered in wavefield analysis in order to satisfy existing boundary conditions. A common conception prevails that a direct observation of the evanescent waves is impossible because of their conversion into homogeneous waves by the interaction with the detectors used or because the field conditions are disturbed by the extraction of energy. In the following paragraphs, the physical reality and the properties of the evanescent waves will be particularly stressed upon, and it will be emphasized how their peculiar behavior can be advantageously used for optical imaging purposes.

Evanescent waves exist and are formed in optics in connection with some well-known phenomena: they are formed in the rarer medium when light is totally internally reflected, they occur in diffraction at objects with extremely fine structures, they are frequently encountered in connection with the angular spectrum representation of the electromagnetic field, they are

used for treatment of the radiation from moving charged particles, and they are used to describe fields without radiation. Evanescent waves are often also called surface waves, especially in the radio-wave field (BARLOW and BROWN [1962]), because they are frequently supported and created by interfaces between different media.

If we include in optical imaging: the object, the relay of its information, the optical system and the formation of an image, and the image, then evanescent waves may be used at several instances. We may convert evanescent waves in the object into propagating homogeneous waves, we may convert the homogeneous waves in some intermediate plane into evanescent waves which may be recorded, later reconstructed and subsequently converted into homogeneous waves (evanescent wave holography) or we may convert the light in the image into evanescent waves.

Because of their evanescence, evanescent waves cannot be used for transferring optical information over any appreciable distance. Therefore, they always have to be used in combination with homogeneous propagating waves in such a way that propagating waves are converted into evanescent waves and/or vice versa. Before conversion, we may, however, introduce a change in the state of the evanescent wave: in its amplitude, phase, frequency, and polarization. We can also record interference patterns caused by evanescent waves, and we can perform diffraction experiments with them. Of course, we may also simply absorb the energy contained in them.

§ 2. Existence and Creation of Evanescent Waves

In this paragraph, we will treat two special cases in which inhomogeneous – evanescent – waves occur. These two cases, namely diffraction and total internal reflection, can as we will see be used in connection with optical imaging. Further, we will concentrate on those properties of the inhomogeneous waves that are of importance in imaging situations.

Denoting the components of the propagation vector k along the cartesian coordinate axes k_x, k_y, k_z, a homogeneous plane wave may be expressed by

$$u = A(\omega) \exp \{j(k_x x + k_y y + k_z z - \omega t)\},$$

where A is the amplitude. The components of the propagation vector can assume any triplet of real numbers which satisfies

$$k^2 = k_x^2 + k_y^2 + k_z^2.$$

If, on the other hand, k_x, k_y, k_z are allowed to be complex

$$k_x = k'_x + \mathrm{j}k''_x, \qquad k_y = k'_y + \mathrm{j}k''_y, \qquad k_z = k'_z + \mathrm{j}k''_z,$$

but k is real, we may write

$$u = A(\omega) \exp\{\mathrm{j}(k'_x x + k'_y y + k'_z z - \omega t) - (k''_x x + k''_y y + k''_z z)\}$$

for a plane wave. However, this wave is an inhomogeneous (evanescent) wave with the following properties: its planes of constant phase are

$$k'_x x + k'_y y + k'_z z = c_1$$

and its planes of constant amplitude are

$$k''_x x + k''_y y + k''_z z = c_2,$$

where c_1 and c_2 are constants. These sets of planes are orthogonal to each other,

$$k'_x k''_x + k'_y k''_y + k'_z k''_z = 0,$$

except in the presence of absorption (k complex) as is mentioned in Section 2.3.

2.1. CREATING EVANESCENT WAVES BY DIFFRACTION

The essential information contained in the scattered field arising when a specific wave with a wavelength λ is diffracted at an object is only about object details above $\frac{1}{2}\lambda$. Thus, information about structures which are finer than half the wavelength cannot be recovered from such a wave field.

If a plane object with the complex amplitude transmission $\tau(x, y)$ in the plane $z = 0$ is normally illuminated with collimated monochromatic light

$$u(x, y, z < 0, t) = \exp\{\mathrm{j}kz\} \exp\{-\mathrm{j}\omega t\}, \tag{2.1}$$

where $k = 2\pi/\lambda = \omega/c$, the wave amplitude immediately behind the object is

$$u(x, y, +0, t) = \tau(x, y) \exp\{-\mathrm{j}\omega t\}. \tag{2.2}$$

Here x, y, z are the cartesian coordinates, t the time, ω the angular frequency, λ the wavelength, and c the velocity of light. Since the wave equation is linear, we can solve the diffraction problem independently for each Fourier component of $\tau(x, y)$ and then sum up the results to obtain the final wave field. According to Fourier's theorem, we have

$$\tau(x, y) = \int\int_{-\infty}^{\infty} \mathcal{T}(\nu, \mu) \exp\{2\pi\mathrm{j}(x\nu + y\mu)\} \mathrm{d}\nu \, \mathrm{d}\mu \tag{2.3}$$

where $\mathcal{T}$ is the Fourier transform of τ and ν and μ are the spatial frequencies

in the x and y directions of the object components. Each component gives rise to a wave

$$\exp\{2\pi j(xv+y\mu+z\eta)\}\exp\{-j\omega t\}$$

that obeys the homogeneous wave equation for $z > 0$. Thus, η is given by

$$\eta^2 = 1/\lambda^2-(v^2+\mu^2) \tag{2.4}$$

and the total wave field in the half space $z \geqq 0$ is

$$u(x, y, z, t) = \exp\{-j\omega t\}$$
$$\times \left[\iint\limits_{v^2+\mu^2 \leqq 1/\lambda^2} \mathcal{F}(v, \mu) \exp\{2\pi j(1-\lambda^2(v^2+\mu^2))^{\frac{1}{2}}z/\lambda\} \exp\{2\pi j(xv+y\mu)\}dv\,d\mu\right.$$
$$\left. + \iint\limits_{v^2+\mu^2 > 1/\lambda^2} \mathcal{F}(v, \mu)\exp\{-2\pi(\lambda^2(v^2+\mu^2)-1)^{\frac{1}{2}}z/\lambda\} \exp\{2\pi j(xv+y\mu)\}dv\,d\mu\right]. \tag{2.5}$$

From eq. (2.5) it is clear that different kinds of waves will arise. The homogeneous waves ($v^2+\mu^2 < 1/\lambda^2$) which correspond to object details with periods larger than the wavelength λ propagate in the direction α, β, γ to the positive z axis, where

$$\cos\alpha = \lambda v$$
$$\cos\beta = \lambda\mu$$
$$\cos\gamma = \{1-\lambda^2(v^2+\mu^2)\}^{\frac{1}{2}}.$$

With normal incidence to the object plane $z = 0$, as assumed here, an object period of λ results in a diffraction angle equal to $\frac{1}{2}\pi$. Of course, we may apply grazing incidence and thus diffract the light π for an object period of $\frac{1}{2}\lambda$. The other portion of the field represented in eq. (2.5) represents inhomogeneous (evanescent) waves ($v^2+\mu^2 > 1/\lambda^2$) which only propagate in directions perpendicular to the z axis but are exponentially attenuated for increasing z values.

Some characteristics of the evanescent wave may be concluded from the preceding derivations; the wavelength λ_e of the evanescent wave

$$\lambda_e = 1/(v^2+\mu^2)^{\frac{1}{2}} \tag{2.6}$$

is smaller than the wavelength λ of the homogeneous wave. Likewise, the phase velocity of the evanescent wave

$$v_e = \omega/\{2\pi(v^2+\mu^2)^{\frac{1}{2}}\} = c\lambda_e/\lambda \tag{2.7}$$

is smaller than the velocity of light c. Of course, both types of waves have the same frequency ω. The propagation vector k with the components

$$k = \{\nu\lambda, \mu\lambda, (1-\lambda^2(\nu^2+\mu^2))^{\frac{1}{2}}\}2\pi/\lambda \tag{2.8}$$

is as was stated above complex in the case of an evanescent wave.

$$k = k_e + ik_a, \tag{2.9}$$

where

$$k_e = 2\pi(\nu^2+\mu^2)^{\frac{1}{2}} \tag{2.10}$$

is the vector in the direction of propagation of the evanescent wave and $1/k_a$, where $k_a = 2\pi(\lambda^2(\nu^2+\mu^2)-1)^{\frac{1}{2}}/\lambda$, represents the distance from the plane $z = 0$ at which the wave amplitude is attenuated to 1/e of its value at $z = 0$. From

$$k^2 = (\omega n_e/c)^2 = k_e^2 - k_a^2, \tag{2.11}$$

where n_e is the refractive index of the medium in which the evanescent wave exists, it follows that

$$1/\lambda^2 = 1/\lambda_e^2 - 1/\lambda_a^2 \tag{2.12}$$

which shows that the evanescent wave can only extend to appreciable λ_a values, i.e. z values, when λ_e approaches λ.

The scattering of waves by object structures which are periodic in one dimension – diffraction gratings – is of particular interest because of the many diversified applications for which these structures are used for. From eq. (2.5) we see that the field from an infinite grating illuminated with a monochromatic plane wave consists of an infinite discrete set of plane waves. The spectrum of these waves is distributed according to the grating equation:

$$\sin i + \sin d = m\lambda/a, \tag{2.13}$$

where i and d are the angles of incidence and of diffraction, a the grating period, and m an integer. All integer values of m are allowed. Therefore, the angles i and d cannot only be real, but also imaginary. Thus, only some of the waves are propagating waves which carry energy in the direction set by the angle d. The remaining waves are evanescent waves which are exponentially (cf. eq. (2.5)) damped normal to the plane of the grating. Eq. (2.13) now tells us that in order for waves of order $\pm m$ to be evanescent,

$$|m| > (1+\sin i)a/\lambda \tag{2.14}$$

has to be fulfilled.

In order to explain the diffraction characteristics in more detail and to be able to predict certain phenomena, more rigorous approaches to the problem

have to be taken. An anomalous behavior of diffraction gratings first observed by WOOD [1902] which has stimulated numerous investigations is worth mentioning in this connection. These so-called Wood anomalies are dependent on the groove profile and on the optical properties of the grating material and appear as rapid variations in the intensity of the various diffracted orders in certain narrow wavelength regions. They are among other things associated with wavelengths λ_R that correspond to $\pm\frac{1}{2}\pi$ diffraction, i.e. for light diffracted parallel to the grating surface. Thus, they are observed in any diffraction order at wavelengths and incidence angles that satisfy

$$m_R \lambda_R = a(\sin i \pm 1). \tag{2.15}$$

The order m_R corresponding to λ_R (the Rayleigh wavelength) is called the Rayleigh order. RAYLEIGH [1907] was the first who tried to explain these anomalies by his "dynamic theory" based on expansion of the scattered field in terms of outgoing waves. A limitation of this theory is its indication of a singularity at λ_R and, thus, incapability to give the shape of the bands due to the anomaly. FANO [1938] tried to overcome this by assuming that the grating consists of a lossy dielectric material. Fano is by the way the first one to stress the importance of evanescent waves in this connection. ARTMANN [1942] further refined Rayleigh's theory, and so did among others LIPPMANN and OPPENHEIM [1954], but only in recent years a more sophisticated treatment of the scattering from different obstacles based on a multiple-scattering point of view has been developed (KARP and RADLOW [1956], MILLAR [1960], and TWERSKY [1962]). The maxima in the anomalies were demonstrated by ARTMANN [1942], but TWERSKY [1962] was first to discuss the minima. Besides early experimental investigations by WOOD [1902], INGERSOLL [1920], and STRONG [1936], the grating anomalies have more recently been studied by PALMER [1952, 1956, 1961] who also showed their polarization dependence. TWERSKY [1962] stressed that although the evanescent waves formed at the grating do not transport any energy away from it, they seem to play a significant role in redistributing the energy. They may provide a coupling between different propagating orders and in this way change the energy distribution between the directions d which give propagating waves. HESSEL and OLIVER [1965] pointed out, using a theory based on a guided wave approach, that besides a Rayleigh wavelength type of anomaly there may also exist a resonance-type anomaly (also describable by Twersky's multiple-scattering theory) which is related to guided complex waves in the grating. In case the period of the diffraction grating is so small that no diffracted orders occur when it is illuminated with a plane wave, then the guided wave along the grating would be a purely bounded surface wave.

When the period, however, is large enough so that diffracted orders arise, then the grating can no longer support a purely bounded guided wave and the wave will be a leaky wave with one or more radiating harmonics. Thus, this resonance type of Wood anomaly occurs for the wavelength at which the diffracted waves appear at just the same angles as would be taken by the leaky wave supportable by the grating. These resonance effects can occur for wavelengths far removed from λ_R.

2.2. CREATING EVANESCENT WAVES BY INTERNAL REFLECTION

Light incident on the boundary between two media is reflected and refracted according to Fresnel's formulas. If the refractive index n_s of the medium from which the light is coming is larger than the refractive index n_e of the second medium, then the incident light will be totally internally reflected for incidence angles equal to and larger than the critical angle for total reflection i_t

$$\sin i_t = n_e/n_s. \tag{2.16}$$

A rigorous treatment of Maxwell's equations for this case predicts existence of evanescent waves in the medium with refractive index n_e. These inhomogeneous waves decay exponentially with increasing distance from the interface. In case total internal reflection of a plane wave of infinite extent (as in Section 2.1) is considered, we get in accordance with eq. (2.12)

$$(n_e/\lambda_0)^2 = 1/\lambda_e^2 - 1/\lambda_a^2, \tag{2.17}$$

where λ_0 is the wavelength in vacuum. Furthermore, we have the boundary condition

$$n_s \sin i/\lambda_0 = 1/\lambda_e. \tag{2.18}$$

Thus,

$$\lambda_a = \lambda_0/\{n_s^2 \sin^2 i - n_e^2\}^{\frac{1}{2}}, \tag{2.19}$$

which indicates that the evanescent wave amplitude will be attenuated to 1/e of its value at the boundary at a distance $\lambda_a/2\pi$ below it. The penetration depth of the evanescent waves only becomes appreciable when i approaches i_t. However, in this region the amplitude of the field at the surface in the denser medium can rise to values above those of the incoming wave E_s. Depending on the polarization direction of E_s, different results will be obtained: for light polarized perpendicular to the plane of incidence

$$E_\perp/E_{s\perp} = 2 \cos i\{\cos i + \mathrm{j}(\sin^2 i - (n_e/n_s)^2)^{\frac{1}{2}}\}/\{1-(n_e/n_s)^2\} \tag{2.20}$$

and for light polarized parallel to the plane of incidence

$$E_{||}/E_{s||} = \frac{2\cos i\{(n_e/n_s)\cos i + j(n_s/n_e)(\sin^2 i - (n_e/n_s)^2)^{\frac{1}{2}}\}}{(n_e/n_s)^2\cos^2 i + (n_s/n_e)^2[\sin^2 i - (n_e/n_s)^2]}. \quad (2.21)$$

These are solutions of Maxwell's equations with proper boundary conditions (see e.g. ARZELIÈS [1946]). From these expressions is clear that at the critical angle of total reflection, we get

$$(E_\perp/E_{s\perp})_t = 2$$

and (2.22)

$$(E_{||}/E_{s||})_t = 2n_s/n_e$$

but that the electric field decreases rapidly with increasing angle of incidence for total reflection. Figure 2.1 illustrates these conditions. This possibility of magnifying the electric field has turned out to be useful, especially for maximizing interaction in local physical penomena.

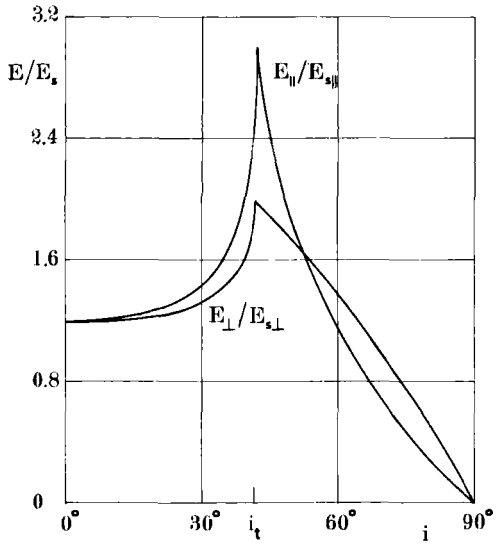


Fig. 2.1. Transmission of different polarization components of the electric field as a function of the angle of incidence when passing the interface from an optically denser to an optically rarer medium ($n_s/n_e = 1.5$).

As we here are concentrating on imaging aspects, the time-averaged flux of energy carried by the evanescent waves is of major concern. It was mentioned above that in order to fulfill the continuity requirements of the electromagnetic field at the interface between the two media an evanescent wavefield in the rarer medium has to exist. With a field of infinite extent

treated so far, a net time-averaged flow of energy is obtained only parallel to the boundary but not perpendicular to it. A more meaningful physical understanding of the phenomenon is reached by considering a wave of limited extent, as occurs in reality and especially in the case of imaging applications.

In order to achieve a broad feeling of the possible uses of evanescent waves for optical imaging, we would like to know the direction in which the energy travels. One of the first to treat this question extensively was PICHT [1929]. He came to the conclusion that the appearance of an energy flow in the rarer medium is the result of an energy flow into this medium under almost grazing incidence at some locations and a corresponding energy flow from the rarer medium at other locations. NOETHER [1931] explained Picht's conclusion in such a way that insight was obtained concerning the physical phenomena involved. SCHAEFER and PICHT [1937] explained this in an even more realistic form. They pointed out that in case the internally reflected wave is limited to a certain width the energy is flowing into the rarer medium at one edge causing the reflection to appear less than total. This energy is then travelling in the rarer medium as an evanescent wave parallel to the interface along the whole extension of the wave as is illustrated in Fig. 2.2. At the other edge of the limited wave this energy is returning to the denser medium where it adds to the reflected wave. Thus, the net effect is a displacement of the whole reflected wave a small distance along the surface.

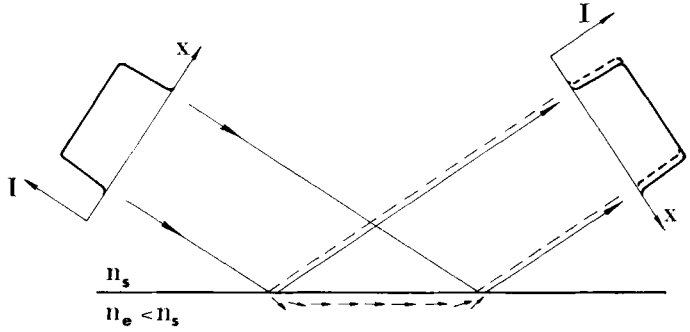


Fig. 2.2. Illustration of total reflection of a finite plane wave. It is shown how the intensity distribution $I(x)$ is shifted relative to the geometrically reflected wave. The arrows in the rarer medium indicate direction of the energy flow.

An expression for the lateral displacement d that occurs when linearly polarized light is internally reflected at an interface between two media (see Fig. 2.3) can be obtained in the following way: The time-averaged energy

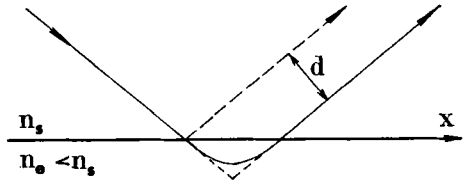


Fig. 2.3. Illustration of how light is displaced a distance d when totally reflected. The penetration into the rarer medium is here shown schematically as a parabolic path.

flow in the evanescent field across a plane perpendicular to the interface between the two media

$$\left\langle \int_0^\infty S_x(z)\mathrm{d}z \right\rangle$$

can be set equal to the energy flow of a plane wave across a surface normal to the direction of propagation with a width that is equal to the displacement d.

$$S_x = \frac{c}{4\pi}[E_y H_z - E_z H_y]$$

is the x-component of the Poynting vector which is parallel to the interface. Thus, for a polarization component that is perpendicular to the plane of incidence ($E_{s\perp} = E_y$),

$$\tfrac{1}{2}LS_x(z=+0)\int_0^\infty \exp(-4\pi z/\lambda_a)\mathrm{d}z = Ld_\perp cn_s E_{s\perp}^2/(8\pi) \qquad (2.23)$$

from which we obtain

$$d_\perp = |E_\perp||H_z|\lambda_a/\{4\pi n_s E_{s\perp}^2\}. \qquad (2.24)$$

L is the length of the considered planes perpendicular to the plane of incidence. $|H_z|$ at the boundary in the rarer medium is obtainable from Fresnel's formulas:

$$|H_z| = 2E_{s\perp} n_s \sin i \cos i/\{1-(n_e/n_s)^2\}^{\frac{1}{2}}.$$

Introducing this and the expressions in eqs. (2.19) and (2.20) in eq. (2.24) results in

$$d_\perp = (\lambda_s/\pi)\sin i \cos^2 i/[\{1-(n_e/n_s)^2\}\{\sin^2 i-(n_e/n_s)^2\}^{\frac{1}{2}}]. \qquad (2.25)$$

Similar calculations for light polarized parallel to the plane of incidence give

$$d_{||} = \frac{(\lambda_s/\pi)(n_e/n_s)^2 \sin i \cos^2 i}{\{(n_e/n_s)^4 \cos^2 i + \sin^2 i - (n_e/n_s)^2\}\{\sin^2 i-(n_e/n_s)^2\}^{\frac{1}{2}}}. \qquad (2.26)$$

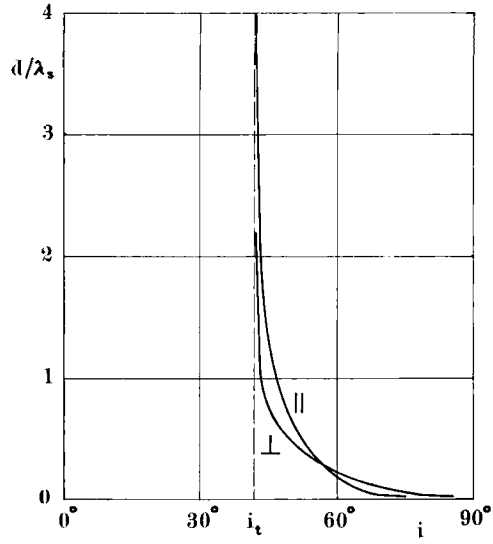


Fig. 2.4. The lateral beam displacements $d_\perp$ and $d_{||}$ as a function of the angle of incidence occurring at total reflection ($n_s/n_e = 1.5$).

In Fig. 2.4 $d_\perp$ and $d_{||}$ are shown as a function of the angle of incidence i. As is clear from this figure, $d_{||} > d_\perp$ for angles of incidence close to the critical angle of total reflection. For larger angles, on the other hand, $d_\perp > d_{||}$. Although an approximate and simplified theory in accordance with RENARD's [1964] procedure using the conservation of energy has been applied here to obtain eqs. (2.25) and (2.26), it is amazing how little they deviate from more exact and complicated approaches. For example, eq. (2.25) is identical to the one obtained by LOTSCH [1968, 1970, 1971] and eq. (2.26) only insignificantly differs from his results, which by the way still are approximations. $\sin i \cos^2 i$ in eq. (2.26) is replaced by $\{\sin i - (n_e/n_s)^2 \sin i - \sin^2 i + (n_e/n_s)^2\}$. Another method of calculating the lateral displacement that gives some more insight into the physical phenomena involved was introduced by ARTMANN [1948]. He considered total reflection of a laterally limited wave. The different plane wave components of its angular spectrum will undergo different phase shifts at reflection, because they enter the interface under different angles. In the superposition of the reflected waves, this produces a spatial variation in the light at the reflecting surface. The condition for maximum light is

$$d = -(\lambda_s/2\pi)\mathrm{d}\delta/\mathrm{d}i, \tag{2.27}$$

where δ is the phase shift at total reflection. From Fresnel's formulas, we have

$$\delta_{\perp} = 2 \arctan \{[\sin^2 i - (n_e/n_s)^2]^{\frac{1}{2}}/\cos i\}$$
$$\delta_{||} = 2 \arctan \{[\sin^2 i - (n_e/n_s)^2]^{\frac{1}{2}}/[(n_e/n_s)^2 \cos i]\}.$$

This results in somewhat simplified expressions for the lateral beam displacement:

$$d_{\perp} = (\lambda_s/\pi) \sin i/\{\sin^2 i - (n_e/n_s)^2\}^{\frac{1}{2}}$$
$$d_{||} = (\lambda_s/\pi)(n_e/n_s)^2 \sin i/\{\sin^2 i - (n_e/n_s)^2\}^{\frac{1}{2}}. \qquad (2.28)$$

Of course, we may interpret the phenomena described here not only as a lateral displacement in the plane of incidence as outlined above but also as a depth of light penetration or as a shift of the reflection center for the light in the rarer medium (cf. Fig. 2.3). Thus, NEWTON's [1717] suspicion that total reflection is not a surface but a volume effect has turned out to be correct. He expected that the path of the light energy in the less dense mediun could be described by a parabola.

The lateral displacement that occurs at total reflection has been experimentally verified by GOOS and HÄNCHEN [1947, 1949]. The effect is also commonly known as the Goos-Hänchen effect. In the former half of this century, it was commonly accepted that the flow of energy in the rarer medium could not be physically determined without disturbing the total reflection phenomenon. However, in 1943 the picture changed when Goos and Hänchen performed their cleverly devised experiment (see Section 3.2) which clearly showed what happens to the light that is totally reflected. The interpretation of the convincing experiments by Goos and Hänchen has been treated by GOOS and HÄNCHEN [1947] themselves as well as by ARTMANN [1948], v. FRAGSTEIN [1949]. WOLTER [1950] and more recently by RENARD [1964], SCHILLING [1965], LOTSCH [1968, 1970, 1971], RICARD [1970], and HOROWITZ and TAMIR [1971].

FEDOROV [1955] predicted that the reflected wave generally not only is displaced in the plane of incidence but also perpendicular to it. This behavior had already been indicated by WIEGREFE [1914, 1916]. The exit plane, however, is still parallel to the plane of incidence. FEDEROV [1955] showed that the Poynting vector is located in the plane of incidence only when the incident light is linearly polarized perpendicular or parallel to this plane. In other cases, there will also exist an energy flow normal to the plane of incidence causing a transverse displacement d_t. This displacement is maximum when the incident and reflected waves are circularly polarized in the same sense. In case the incident wave is linearly polarized, on the other hand, then the transverse displacement is maximum for a circularly polarized reflected wave. This transverse displacement which is related to the spin of the

photon has been treated by COSTA DE BEAUREGARD [1964, 1965], SCHILLING [1965], IMBERT [1968, 1972] and RICHARD [1970]. For a circularly polarized incident wave (IMBERT [1969])

$$d_t = \mp \frac{\lambda_s}{\pi} \frac{\sin^3 i \cos i}{\sin^2 i - \sin^2 i_p + \sin^4 i_p \cos^2 i}, \tag{2.29}$$

where i_p is the specific angle of incidence for which the state of polarization is preserved and the minus and plus signs represent left and right circularly polarized light. In most cases, the transverse displacement is an order of magnitude smaller than the lateral displacement.

A phenomenon similar to the lateral beam displacement at total reflection also occurs at diffraction by optical gratings. The diffracted waves are all spatially shifted with respect to the illuminating wave (TAMIR and BERTONI [1971]). This displacement is explainable in terms of a leaky-wave mechanism and is larger for grating diffraction than at total reflection. Furthermore, this displacement can be either in the forward or reverse direction with respect to the incident wave.

As mentioned earlier in this section, the energy flux in the rarer medium at reflection has in general a local component S_z perpendicular to the interface between the media. This is the case in total internal reflection as well as in the case of an absorbing medium.

2.3. CONDITIONS AT ATTENUATED TOTAL REFLECTION

In Section 2.2, a non-absorbing rarer medium was assumed. Most direct recordings of the described phenomena require, however, an absorbing material. This will modify the wavefield conditions in the rarer medium. Fresnel's formulas and Snell's law are still formally valid even though other physical interpretations are needed in this case. The refractive index n_e used in Section 2.2 has to be replaced by a complex refractive index $n_e(1-j\kappa)$, where κ is the absorption index.

In case a wave is internally reflected in a non-absorbing medium at its interface with an absorbing medium, the field in the latter medium is still a system of inhomogeneous plane waves (see STRATTON [1941]). The planes of constant amplitude are parallel to the reflecting interface as in the case of total reflection, but the planes of constant phase are no longer perpendicular to the interface. The propagation direction is normal to the planes of constant phase. Thus, in case the rarer medium is absorbing, the inhomogeneous wave will propagate in a direction inclined to the interface. The angle of inclination (refraction) depends now on the angle of incidence as does also the phase velocity. With increasing absorption index κ, the planes of con-

stant phase will incline more and more to the interface in which the reflection takes place. In the limit $\kappa \to \infty$ they will be parallel to the planes of constant amplitude and the propagation direction is normal to the surface of the absorbing medium.

Figure 2.5 illustrates the difference between the case with a perfectly transparent (Fig. 2.5(a)) and an absorbing medium (Fig. 2.5 (b)). The electric field is shown in the rarer medium at internal reflection for light polarized parallel to the plane of incidence. The magnetic field is perpendicular to the plane of the figure. Thus, the Poynting vector has the direction of the tangent to the orthogonal trajectories to the E-field shown. This means that energy enters the rarer medium at certain locations and some portion of it returns from it at others. Thus, in the time average the direction of the net energy flow is parallel to the interface when no absorption is present and inclined to the interface when the rarer medium is absorbing. As is illustrated in Fig. 2.5(b), the absorption of the rarer medium causes a progressive decrease in the amplitude of the inhomogeneous wave as it travels forward normal to the planes of constant phase.

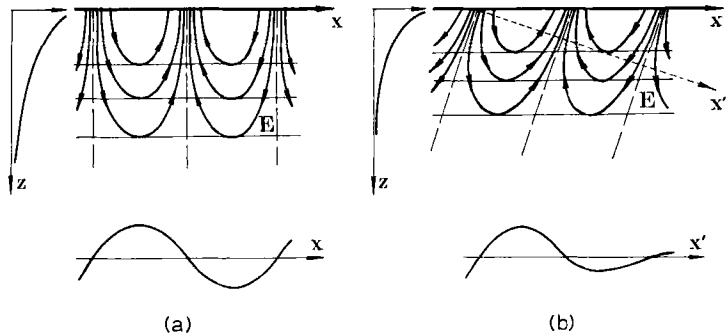


Fig. 2.5. The electric field at internal reflection in a perfect tranparent rarer medium (a) and in an absorbing one (b). Surfaces of constant amplitude are drawn solid and surfaces of constant phase dashed. The variation of the amplitude in depth is shown at the left and the variation parallel to the direction of propagation is shown at the bottom.

Thus, when $\kappa \neq 0$, we get attenuated total reflection. The absorption is largest at angles of incidence close to i_t (critical angle in case of total reflection) and depends, of course, on the value of the absorption index κ. This is illustrated in Fig. 2.6 for fairly strong absorbers. There no longer exists a particular angle for total reflection, but a transition of finite angular width (LITTMANN [1940]).

Figure 2.6 shows that when using strongly absorbing media like metals, the conditions deviate strikingly from those present at total internal reflec-

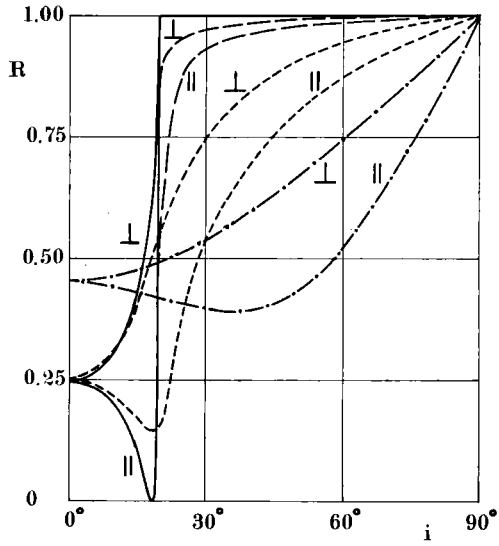


Fig. 2.6. Internal reflectivity for different polarization directions at the interface between an optically denser and an absorbing optically rarer medium ($n_s/n_e = 3.0$) as a function of the angle of incidence. Curves for different absorption coefficients $n\kappa = 0$ (———), 0.05 (— — —), 0.5 (– – – –), and 5 (—·—·—) are shown.

tion. For weak absorbers, on the other hand, the reflectivity does not deviate noticeably from that for a non-absorbing medium, except very close to the critical angle (cf. Fig. 2.6). Commercial non-processed fine grain photographic emulsions have a relatively low absorption coefficient ($n\kappa \approx 10^{-3}$). Thus, in those cases, the treatment given in Section 2.2 is a good approximation.

2.4. WAYS TO USE EVANESCENT WAVES IN OPTICAL IMAGING

In Sections 2.1 and 2.2 different ways to create evanescent wavefields in optics were described. It is also clear from these treatments that the same methods may also be used to convert an evanescent wave into a homogeneous propagating wave. According to Helmholtz' reciprocity theorem the direction of the light in the cases treated can be reversed, i.e. the evanescent waves are converted into homogeneous waves by an interface between two media or by a grating structure. This means that the geometries mentioned are not able to support evanescent waves over any appreciable lateral distances. Energy is coupled out as soon as energy has been coupled into the boundary. The lateral displacement mentioned is a measure of how long the energy remains in the boundary before it is coupled out again. However, the schemes treated can form a part of an optical imaging system as will be

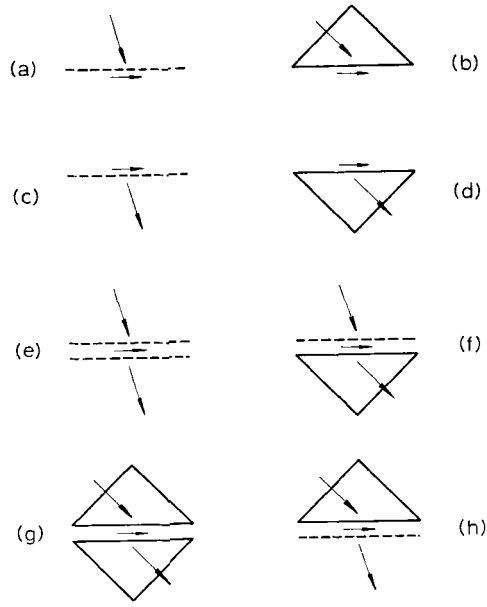


Fig. 2.7. Illustration of different ways of creating and converting evanescent waves by using and combining diffraction and internal reflection techniques.

described in detail in later chapters.

Figure 2.7 shows how, besides using these schemes (Fig. 2.7(a)–(d)) separately, we can also combine them in different ways (Fig. 2.7(e)–(h)). Fig. 2.7(g) constitutes the optical tunneling method which is commonly used in optical thin film techniques (BAUMEISTER [1967]). By introducing a spatial variation of, for example, the thickness, absorption, or refractive index of the thin film in which the evanescent waves exist, we have a simple and effective means for creating or influencing pictorial information in a light beam. The addition of a diffractive structure as in Fig. 2.7(e), (f), and (h) even further increases the flexibility of using evanescent waves for formation, recording, and processing of optical images. As is described in § 6, the schemes (f) and (h) in Fig. 2.7 are of particular value for the realization of evanescent wave holography.

§ 3. Experimental Verification of the Existence and Properties of Evanescent Waves

As is described in this paragraph, there exist different techniques among which to choose when we wish to visualize an evanescent wavefield (cf. also

§ 5). We may use photographic emulsions, diffraction gratings, geometrical boundary changes, scattering by small particles, etc., whereby the evanescent wavefield is converted into a propagating homogeneous wavefield. These methods, of course, introduce significant deviations from the unique character of total internal reflection and can as such not be considered as true experimental proof. However, as we will also see in this paragraph, indirect techniques which do not intrude on the evanescent field distribution have been devised.

3.1. PIONEERING EXPERIMENTS ON EVANESCENT WAVEFIELDS

The first experimental studies of the light that enters the less dense medium in total reflection was performed by NEWTON [1717]. He placed two slightly convex glass surfaces against each other. At and around the point where they were in contact, he could see through. In reflection, the same area appeared dark. He observed that this area seemed first to increase with increasing angle of incidence and then to decrease when the angle was increased further. From these experiments, Newton drew the conclusion that light really penetrated into the rarer medium at total reflection. QUINCKE [1866] performed experiments similar to Newton's and observed that the area around the contact point between two glass prisms seemed largest at the critical angle of total reflection. Quincke explicitly stated that the penetration depth of the light decreased with increasing angle of incidence. He also found that the penetration depth increased with increasing wavelength and that it showed a polarization dependence. Quincke observed how internally reflected light penetrates a small separation between the hypotenuse surfaces of two right angle prisms. One surface was convex and the other plane. In this way an area around the point of contact between the two surfaces appears. With this configuration Quincke studied the penetration depth of the evanescent waves as a function of the angle of incidence and obtained values of up to a couple of wavelengths for the penetration depth. Another experimental approach to study the penetration of light into the rarer medium was taken by DITSCHEINER [1870], EXNER [1889], and EDSER and SENIOR [1902]. Their idea was to place a diffraction grating at the interface where total reflection occurs. A photographically produced grating was cemented onto the hypotenuse face of a right angle prism. The direction of the lines of the grating were perpendicular to the plane of incidence. By internal reflection at the grating, it was possible to get diffracted light into the less dense medium. Different diffraction orders except the zeroth were visible. (For a review of early experiments in this field see, for example, KOROBKO-STEFANOV [1950].)

VOIGT [1899] criticized the procedures which had been used to investigate the inhomogeneous wave in the rarer medium. He rejected to the introduction of a third medium in order to detect light in the rarer medium, because when light appears in the third medium, the reflection is no longer total. Instead Voigt suggested a technique that did not require any third medium. He used a right angle prism the hypotenuse surface of which consisted of two fields slightly inclined to each other so that the edge between them was parallel to the roof of the prism. He illuminated the whole hypotenuse surface so that the light was internally reflected. Looking at the edge from a direction in the plane of one of the hypotenuse halves revealed a streak of light. The intensity of this light decreased rapidly with increasing angle of incidence. However, even Voigt's method suffers from a deviation from total reflection – some light is escaping from the edge. This was the reason why this method gave rise to a long and sharp controversy (KETTLER [1899], VOIGT [1899b, 1911] and EICHENWALD [1911]).

Another method was used by HALL [1902] to show that light penetrates into the rarer medium at total reflection. He applied a gelatine layer which had been light sensitized using extremely fine silver bromide grains onto one side of a flint-glass prism. Light internally reflected in the prism exposed the emulsion. After development Hall found that the film had only been exposed on that side which was next to the prism. He could detect penetration depths up to about 5μ. Further, he applied thin collodion films between the prism and the sensitized emulsion and was still able to expose the emulsion. WOOD [1913] describes how the same results are obtained by using a fluorescent film instead of a photographic emulsion as detector in the rarer medium.

Extensive investigations with the so-called lamella-experiment described above were performed by GOOS and HÄNCHEN [1943]. They measured the intensity distribution around the contact area between two glass prisms, one of which had a plane and the other a slightly convex hypotenuse surface. Photometric measurements were done on photographic recordings, and it was found that the measured value of the size of the contact area was maximum when the observing light entered at the critical angle for total reflection.

Some early experiments (RIGHI [1898], LAMPA [1899], SCHAEFER and GROSS [1910]) were also performed in the microwave field in order to study the wave penetration into the rarer medium at total reflection. The work by Schaefer and Gross in particular describes convincing experimental verification of the energy distribution in the evanescent wavefield. They used two paraffin prisms arranged in a variable beam-splitter configuration and mea-

sured the reflected and transmitted energies as a function of the prism separation. A wavelength of 15 cm was used and a thermoelement served as detector. Direct measurements were also made of the evanescent wavefield distribution in the rarer medium close to the hypotenuse surface where the microwaves were totally reflected.

3.2. MEASUREMENTS OF PROPERTIES IN EVANESCENT WAVEFIELDS

One of the more important experiments in this field was performed by Goos and Hänchen in the early 1940's. However their first publication on it was delayed because of the war (GOOS and HÄNCHEN [1947] and GOOS and LINDBERG-HÄNCHEN [1949]). Their experiment was devised in such a way that the total reflection was not disturbed. It relied on the theoretical prediction that the light energy at total reflection enters the rarer medium at certain locations and returns back at other locations into the optically denser medium. Light penetrates into the rarer medium as is indicated in Fig. 2.3. Goos and Hänchen were able to visualize the beam displacement at total reflection to a high accuracy. Their experimental idea was to compare the reflected beam which is drawn as a solid line in Fig. 2.3 to the one which occurs when the reflecting surface is metallized (dashed line in Fig. 2.3). This was achieved by evaporating a central strip of silver over the surface at which the total reflection occurred. In order to achieve enough accuracy in determining the beam displacement, the number of reflections was kept high by using multi-reflections in a plane parallel plate. Two edges of the plate were inclined to facilitate the entering and exit of the light in the glass plate. Goos and Hänchen's experiments showed without any doubt the existence of the beam displacement at total reflection. From observations with 20 to 70 reflections they were able to measure both from photographic recordings and from direct observations by an eyepiece micrometer the displacement in the shadow of the light that had passed two successive slits. The dependence of the displacement on the angle of incidence (GOOS and HÄNCHEN [1947]) and on polarization (GOOS and LINDBERG-HÄNCHEN [1949]) was also verified. Further, Goos and Hänchen demonstrated the beam displacement by using several wavelengths. Different beam displacements occur, of course, for different wavelengths.

The experiment by Goos and Hänchen was later repeated by WOLTER [1950] who succeeded in performing even more accurate determinations of the lateral beam displacement (Goos-Hänchen effect). Wolter observed the displacement of interference fringes formed by two slightly inclined waves (the so-called "Minimumstrahlkennzeichnung"-method) that had undergone multiple reflections on one hand at the interface glas-air and on the other

at the interface glas-silver. A plane parallel plate was used on which a silver strip was deposited. Wolter with his technique having a 16-fold improvement in resolution over that by Goos and Hänchen found good agreement between experiments and theory for both polarization directions in the entire angular range of interest.

The experiments described so far have only been concerned with the lateral beam displacement, i.e., the displacement in the plane of incidence. Transverse beam displacement, i.e., displacements orthogonal to the plane of incidence, have also been experimentally verified (IMBERT [1969, 1970a, b, 1972]). Imbert used circularly polarized light to insure maximum transverse displacement. Two configurations were used. In one the light was successively totally reflected at the sides of a long 60° triangular prism in a spiral path (IMBERT [1969, 1970a, b, 1972]). For a wavelength of 6328Å and a refractive index of the glass of 1.8 the transverse displacement is 0.29μ. With 28 reflections, this displacement was recordable. In the other setup, a 90° triangular tankprism was used. Total reflection occurred at the base side and the light was almost normally incident on the sides which had semi-reflecting coatings to change the sense of the circularly polarized light (IMBERT [1970a, b, 1972]). 20 reflections were applied in this case. To visualize the transverse displacement, a 180° phase plate (half of the field was out of phase with the other half) was used as object and placed in the light before it entered the prism arrangement. Furthermore, the illumination was arranged so that half of the field of view was right circularly and the other half left circularly polarized. The separation line between these two halves was perpendicular to the sharp edge of the phase plate. In this way the two halves of the image of the edge were displaced in opposite directions. By interchanging the states of polarization in the illuminating light, the displacements in the image were interchanged. Imbert found good agreement with theory.

The beam displacement described here can also be found at other wavelengths, e.g., in the microwave and radiowave regions. The work by WOLTER [1950, 1961] is in particular worth mentioning. Wolter was concerned with explanation of strange reflection effects above seawater in the near field of transmitting antennas. He found negative penetration depths for a wave polarized parallel to the plane of incidence. This indicates that a wave polarized in this sense is reflected at a virtual surface which is located above the water level. Several other investigations where evanescent waves were especially emphasized on have been performed using microwaves, see e.g. CULSHAW [1961], CULSHAW and JONES [1953] and BRADY et al. [1960]. Displacement of a totally reflected ultrasonic wave was successfully demonstrated by SCHOCH [1952].

SCHAFFNER and TORALDO DI FRANCIA [1949] studied the evanescent wavefield created by a diffraction grating. They illuminated a metal strip grating with microwaves of 32 mm wavelength. The grating period was chosen so that only the 0th diffraction order was leaving the grating. The other orders gave rise to evanescent waves. By placing a paraffin prism close to the grating, they were able to transform one of the first diffraction orders into a homogeneous propagating wave. The exponential attenuation of the evanescent wavefield normal to the grating was confirmed by measuring the energy in the first order as a function of the separation between the grating and the prism.

Frustrated total internal reflection experiments have been used to study different aspects of the evanescent field which occurs at total reflection. Here only two experiments will be mentioned. COON [1966] used a cooled photomultiplier to count the photons (5461Å) which had tunneled through the gap between two glass prisms. He was able to confirm the exponential attenuation of the evanescent field for prism separations ranging between 3.5 and 8.5μ. CARNIGLIA and MANDEL [1971], on the other hand, were able to verify experimentally that the phase of the light which has tunneled from one glass prism to the other is independent of the prism separation. This proves that the evanescent waves propagate parallel to the glass-air interface and that the planes of equal phase are perpendicular to this interface. Measurements were performed in He-Ne light with the prism arrangement in one arm of a modified Rayleigh interferometer. The phase was determined from the interference fringe position as a function of prism separation (up to 20 wavelengths). In both Coon's and Carniglia and Mandel's experiments, the prism separation was measured by a separate interferometer.

§ 4. Image Formation, Processing, and Transfer Using Frustrated Total Internal Reflection

Total internal reflection techniques lend themselves in a natural way to investigations of, among other things, surface topography, film thickness, refractive index, and absorption spectra. The effect used is then frustrated total internal reflection, i.e., the evanescent waves which penetrate into the rarer medium can be absorbed or diverted by bringing an absorbing or high refractive index material close enough to the reflecting surface.

4.1. APPLICATION TO PROXIMITY PROBLEMS IN METROLOGY

Systems were described by SCHMALTZ [1936] and DREYHAUPT [1939] which used frustrated total internal reflection to examine surface topography. Their interest was in particular to use the method for inspection of

the surfaces of machined parts. However, the method was little known until YOUNG and ROTHROCK's [1961] and HARRICK's [1962] treatments of its application to diversified problems in metrology. Young and Rothrock used the phenomenon to measure the film thickness between a glass and a metal interface. HARRICK [1962, 1963] discussed a number of applications in metrology and in particular its usefulness for fingerprinting. Harrick examined the film between two prisms. He also replaced the second prism by a transparent or absorbing liquid. In this way, optical contact to the film can be ensured especially when the film has a nonuniform thickness. Mercury was found practical as an absorbing second medium. The range of film thicknesses which can be measured by this method is unfortunately rather limited; usually only thicknesses from about 0.1λ to 1λ can be studied. Ways of extending this range include the use of multiple internal reflections and changing the angle of incidence. When illuminating the prism in Fig. 4.1 with

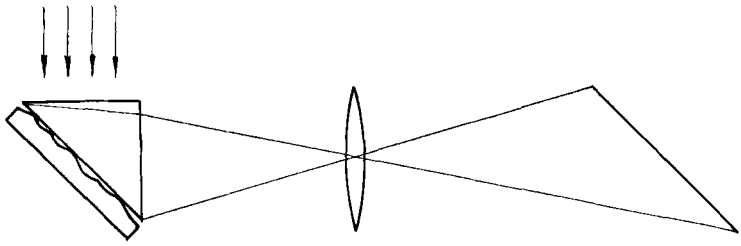


Fig. 4.1. Schematic illustration of optical system for surface topography examination using frustrated internal reflection.

collimated light at an angle of incidence close to the critical angle, the reflectivity at any position is very sensitive to the local index of refraction of the rarer medium. An obvious application of this is in visualization of phase objects like biological specimen which cannot easily be studied by conventional techniques. Because of the strong dependence of the intensity of the evanescent field on the distance from the interface, the local intensity in the reflected wave field is extremely sensitive to the proximity of the second medium to the interface. Very high contrast images can be made of surface reliefs using this method. The oblique incidence in the scheme of Fig. 4.1 will cause distortion of the image. It can be kept to a minimum by using a prism of high refractive index and by working as close to the critical angle as possible. The recording plane can also be tilted appropriately as is shown in Fig. 4.1. On a screen tilted as shown in this figure an image with minimum distortion is obtained. In case perfect focus over the entire image is desired the screen has to be tilted in the opposite sense. However with this me-

thod, it is impossible to produce an image which is simultaneously undistorted and everywhere in focus. An inconvenience with the tilted image is that it has to be projected onto a screen. With an eyepiece, it is only possible to focus on a narrow zone.

In the systems mentioned above, distortion and tilting of the Gaussian image plane occurred because the axis of the optical system was in line with the emerging light. The light that leaves the object plane is inclined to the object plane by at least the critical angle. These drawbacks do not appear if the axis of the optical system is normal to the object plane. However, the light enters the optical system under angles greater than the critical angle. McCUTCHEN [1964] used microscope objectives with numerical apertures exceeding one for this purpose. In Fig. 4.2, a setup using an oil immersion objective is shown which can be attached to a conventional microscope.

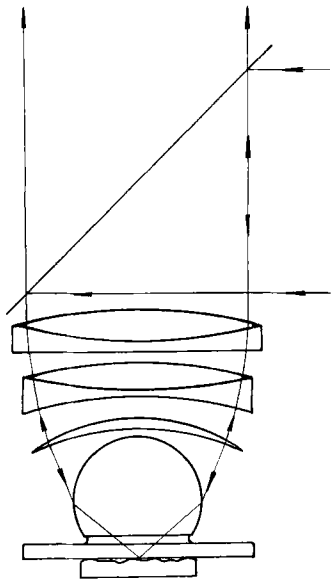


Fig. 4.2. Illustration of axially symmetric optical system suitable for surface studies using frustrated internal reflection.

McCutchen obtained satisfactory results using a conventional vertical illuminator. A cover-glass was mounted on the objective in such a way that accurate focusing could be done on the front surface of the cover-glass. The specimen has to be placed in very close proximity to the cover-glass. The amount of light reflected depends both on the separation between specimen and cover-glass and on the angle of incidence (see Fig. 4.3) and on the index of refraction of the specimen. One of the virtues of the method is the unam-

biguity in the interpretation of the image in case the specimen is uniform in refractive index. Then, there is no doubt whether a hill or valley exists in the object.

4.2. CONVERSION OF UNCONVENTIONAL RECORDINGS INTO INTENSITY DISTRIBUTIONS

With the development and availability of new recording media, a read-out scheme using frustrated total internal reflection is a valuable alternative to other methods. The schemes described in Section 4.1 appear simple and effective for conversion of a variation in the parameter in question of the medium into an image intensity distribution.

Some recording media permit only phase-only recordings. In these cases, methods such as the introduction of a spatial carrier and use of spatial filtering techniques, or Schlieren, phase contrast, or interferometric methods are commonly used to transform the phase variation into an amplitude variation. Common to all these methods is the need for relatively coherent light. Furthermore, in all these methods accurate alignment and component adjustments are required. A method similar to the one illustrated in Fig. 4.2, on the other hand, has several advantages; it works in incoherent light and does not require any special adjustments. Without any alterations, it works

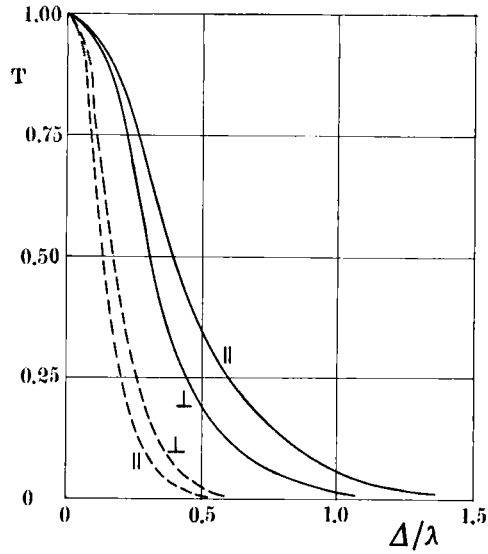


Fig. 4.3. Transmission of differently polarized light through a gap ($n = 1$) between two optically dense media ($n = 1.5$) as a function of its width. Solid curves represent an angle of incidence equal to 45° and dashed curves 60°.

as well for relief- and refractive-index change-type materials as for conventional absorption type recordings.

The refractive index changes can be achieved in a number of ways including the use of chemical modifications, electric field (Kerr effect), pressure, temperature, and strain.

Figure 4.3 shows the amount of light that is transmitted through a pair of glass prisms as a function of their separation Δ. The prisms are arranged as in Schaefer and Gross' setup (see end of Section 3.1) which they used to demonstrate controlled coupling to the penetrated evanescent wavefield. As is shown in this figure, the degree of coupling is dependent on the angles of incidence and polarization. The transmitted light is strongest for light polarized parallel to the plane of incidence. This condition is reversed for angles far away from the critical angle as is also clear from Fig. 2.4. Of course, the opposite is true for the reflected light.

The usable range of surface relief variations is obvious from Fig. 4.3. A typical linear range is 0.1–0.4 λ which is easily achievable with most phase-only recording media including bleached photographic material. The S-shaped image recording characteristic is similar to that of conventional photography. The choice between positive or negative image polarity is a matter of choice between transmitted or reflected light. This makes it possible to use several media in their different mode of operation, e.g. etched or swelled type phase images. The procedure is particularly suited for masking techniques, like unsharp masking, where positive and negative versions of the same image are superposed.

The reflection can also be frustrated by the wavelength dispersion of the refractive index. If the angle of incidence is chosen near the critical angle, preferably just below, the reflectivity will be strongly wavelength dependent if there is dispersion in the refractive index.

As pointed out above the properties of a medium influence the light reflected from it. Thus, besides being a powerful method of forming images, these techniques give us means of investigating the process of reflection and of studying how the structure of matter is manifested in the parameters of the reflected light (KIZEL [1968]). When transient signals are used as in communication type applications the buildup time of the process and its pulse shaping property has to be considered.

4.3. AFFECTING GUIDED WAVES THROUGH EVANESCENT WAVES

In cases where dielectric waveguides are transporting light energy, evanescent wavefields surround the guide. Especially when the dimensions of the guide approach the wavelength of the light, an appreciable amount of the

guided energy is contained in the evanescent field. Two types of waveguide geometry have been extensively applied: planar as used in integrated optics, and cylindrical in fiber optics. (KAPANY and BURKE [1972]; HILL, WATANABE and CHAMBERS [1972]).

The evanescent wavefield has, in this connection, an influence on the image-conveying property of the wave guiding system. Such fields are also of importance for some aspects of light switching and play a major role for coupling energy in and out of optical integrated circuits.

When fiber optics are used for imaging, the existence of the evanescent wavefield is detrimental because it establishes a leakage of light between neighboring fibers in a bundle (KAPANY and BURKE [1961], KAPANY [1967], LISITSA, BEREZHINSKII and VALAKH [1972]). This has a marked detrimental effect both on image contrast and on resolution. The smaller the diameter of the dielectric waveguide and the closer the angle at which the waves that constitute a mode are to the critical angle, the greater the fraction of the total energy of the mode that is propagated outside the fiber, In the limit, when the mode angle equals the critical angle, all the energy is guided outside the fiber. In order to reduce the coupling between neighboring fibers, they are commonly being manufactured with a core (the waveguide) surrounded by a coating (cladding) of an optically less dense medium. There is also another reason for using a cladding. Since only those modes with reflection angles larger than the critical angle propagate, the number of modes is reduced by keeping the refractive index difference small. Index differences of 1 % are typical. The effective absorption of the fiber is now a function of the core as well as of the surrounding material. The energy transferred to the neighboring fibers in the bundle is in turn transferred to their neighbors. Most of the energy that is transferred in the first place is in modes close to the critical angle. Therefore, a much higher fraction of the total energy of a secondarily excited fiber is transferred to a third, and then to a fourth, and so on. This leads to light levels of the same order of magnitude in the successive fibers surrounding an excited one.

Two different types of coupling occur between two parallel fibers, depending on the fiber diameter and on the degree of coherence in the light which excites one of the fibers: (1) for fiber diameters larger than the light wavelength, energy transfer occurs slowly in accordance with the phenomenon of frustrated total reflection until a state of equilibrium is reached, (2) for small fiber diameter and complete coherence in the light illuminating the fiber, the phenomenon of beating will occur. Almost all the energy is transferred to the initially unexcited fiber in a certain length and then returns to the excited fiber in the same length (half the beat length).

Besides passive fibers, interest has also been toward development of active fibers and thin films for integrated optics. So, for example, glass fiber lasers (SNITZER [1961]), fiber laser amplifiers (KOESTER and SNITZER [1964]), and coupled fiber lasers (using evanescent wave coupling) have been developed.

Lately integrated optics – optical circuits integrated into a common substrate – has evolved as a new field in optics. There seem to be some technological conveniences in this technique. The substrate can contain and support the optical components as well as their interconnections and there is a promise of simultaneous formation of complicated circuits by using masking techniques (MILLER [1969]). They seem to have potential value to signal processing for communication purposes (spatial variation in one dimension is possible). Further, the possible high power densities are attractive for electro-optic, acoustic-optic, and non-linear optical devices. Light couplers according to the principle of Fig. 2.7(g) (TIEN et al. [1969]) and Fig. 2.7(f) (DAKSS et al. [1970]) have been found effective. The prism-film coupler in particular has been treated in detail (TIEN and ULRICH [1970], MIDWINTER and ZERNIKE [1970]). The prism is here placed in close proximity to the wave guide. The incident light is internally reflected in the base of the prism, i.e. the waves in the prism and the film are coupled through an evanescent field. This coupler permits excitation of any one of the film modes by proper orientation of the direction of the incident wave. Coupling takes place along the entire width of the incident wave. As energy is also transferred back from the film to the prism, it is important to adjust the coupling length and the beam profile for maximum efficiency.

§ 5. Use of Evanescent Waves to Record or Transform Optical Images

Sometimes the conversion of a homogeneous wavefield into an inhomogeneous one may be of practical value in order to record images. In this connection, it is in particular the possible increase of the field strength using total reflection (cf. eqs. (2.20)–(2.22) and Fig. 2.1) and its limited penetration depth (cf. eqs. (2.12) and (2.19) and Fig. 2.4) which are of interest. This conversion may be used either as a means of increasing the optical signal detectibility or for converting the wavelength of the radiation to a more suitable range. Attractive is also the possibility of making very thin recordings in bulk materials.

5.1. THE ADVANTAGES OF THE EVANESCENT FIELD FOR IMAGE RECORDING

As described in Section 2.2, the evanescent wavefield which is created by internal reflection has two properties that are of advantage for image record-

ings: high field strength and small penetration depth. This combination is attractive in order to make high resolution recordings. The interaction length (see eqs. (2.25) and (2.26)) is in practice of the order of the wavelength, i.e., comparable to the resolution of high aperture optics. This is much smaller than the thickness of typical recording media. Further, the increased electrical field compensates to some degree for the short interaction length.

There are also recording situations where the possibility of enhancing the detector sensitivity is valuable. For example, total internal reflection schemes have been suggested in order to increase the sensitivity of photomultipliers (GUNTER et al. [1965]). Quite sophisticated schemes which may be used to increase the interaction between the light and the recording medium have been applied in connection with internal reflection spectroscopy (HARRICK [1967]; KORTÜM [1969]).

5.2. USE OF EVANESCENT WAVEFIELD FOR FREQUENCY CONVERSION

Frequency conversion of an optical wavefield may be wanted for several reasons: The detector is sensitive in another wavelength region than that in which the lightsource radiates. Local changes in frequency according to a specified pattern may be accomplished. Realization of a coherent to incoherent light conversion is possible. Another application is optical mapping in three dimensions.

One simple way to make an evanescent wavefield visible is to use fluorescent materials. Spatial variations in the field which is internally reflected in a dense medium may easily become visible by applying a fluorescent liquid in contact with the medium (SÉLÉNYI [1913]). Of course, we may also apply the fluorescent material as a spatial pattern in such a way that only in those areas where the pattern exists the light will be absorbed and re-emitted at another frequency, but the light will remain unchanged at the uncovered portions.

A recording medium with extremely high resolution in the depth direction was developed by a research group at the University of Marburg (DREXHAGE [1970]). It seems to have potential value for recording evanescent wavefields. It is built up by monomolecular dye layers consisting of long-chain fatty-acid molecules and fluorescent dye molecules. These layers, which can be made about 25Å thick, can be conveniently stacked to form multilayer media. The individual layers may be prepared with different dyes and different dye concentration to simplify investigations in the depth direction. The exponential decay of the evanescent waves away from the interface at total reflection and the dependence of this decay on the angle of incidence have

been verified using these multilayer material as the rarer medium (DREXHAGE [1970]). These layers have also been successfully applied to studies of the absorption and emission of evanescent light (CARNIGLIA et al. [1972]).

Another type of recording media, namely non-linear media, may also be used to influence or record an evanescent wavefield. The increased field strength which it is possible to achieve by applying internal reflection techniques is of advantage here. Both second- (BLOEMBERGEN and LEE [1967] and third-harmonic generation (BEY et al. [1969]) have been demonstrated.

§ 6. Evanescent Wave Holography

Holography – usually treated as a two-step lensless imaging process – was extensively developed in the 1960's (see, for example, LEITH and UPATNICKS [1967]). Information about the amplitude and phase of a wavefront is stored in a recording media in the form of an interference pattern. An exact copy of the original wavefront can be reconstructed from this recording as a diffracted wavefield. The process of wavefront reconstruction can thereby conveniently be described by: the intensity

$$I = |r+o|^2$$

which is recorded, the transmission–exposure characteristic of the recording medium, which for an ideally linear amplitude hologram is

$$t_a = t_0 - \beta I,$$

and the waves transmitted by the hologram when illuminated with r

$$rt_a = rt_0 - \beta r|r|^2 - \beta r|o|^2 - \beta r^2 o^* - \beta |r|^2 o.$$

o and r are the complex amplitudes of the object and reference waves, t_a the amplitude transmission of the hologram, t_0 and β constants determined by the processed recording medium; the asterisk denotes complex conjugate quantities. The last term of the last expression represents the desired reconstructed wavefront and can, as is now well known, be spatially separated from the rest by applying an off-set angle between the optical axes of the object and reference fields (LEITH and UPATNICKS [1967]). In 1967, STETSON [1967] made a remark that one ought to investigate the possibility to use inhomogeneous waves in holography. As will be described in this paragraph, the same techniques which are known to work for homogeneous waves can also be applied to inhomogeneous waves. o and/or r in the holographic description above may represent an evanescent wave, and we will then call the process evanescent wave holography.

6.1. DIFFERENT USES OF HOLOGRAPHY WITH EVANESCENT WAVES

One of the latest applications of evanescent waves in optical imaging has been in the field of holography (NASSENSTEIN [1968, 1969a, b, c, d, e, f, 1970a, b], BRYNGDAHL [1969]). There are in principle two ways in which holography may be used:

(1) Holographic recording of the total wavefield from an object, i.e. both the homogeneous (propagating) and inhomogeneous (damped) wavefields. This makes it possible to make recordings of objects with details below $\frac{1}{2}\lambda$. However, the recording medium has in practice to be in contact with the object, because of unavoidable disturbing boundary conditions and strong damping of the evanescent waves.

(2) Holographic recording of homogeneous wavefields in such a way that evanescent waves are created in the recording medium. Homogeneous waves are here transformed into inhomogeneous (evanescent) waves or/and vice versa in the surface of the hologram medium.

Characteristic of this type of holography is that evanescent waves are taking part both in the formation of and reconstruction from the hologram. There are several advantages and reasons to consider this type of holography, aside from the possibility of obtaining a deeper insight into the properties of the evanescent waves themselves. The in principle unlimited resolution which may be obtained from recordings according to (1) above lacks alternatives and speaks for itself. Procedure (2), on the other hand, may seem unnecessarily complicated. However, as will be described in detail, evanescent wave holography possesses some unique features which seem to be of value in several optical imaging situations.

6.2. EVANESCENT WAVE PROPERTIES USEFUL IN HOLOGRAPHY

The very nature of inhomogeneous waves implies that planes of constant amplitude do not coincide with planes of constant phase (cf. introductory section of § 2). The evanescent waves which we will consider here only exist close to a boundary between the two media where they were formed. Of course, it is important that the evanescent waves exist on that side of the boundary which contains the recording medium. One way to insure this is to allow the light to be internally reflected in a medium adjacent to the light sensitive medium. The condition for this to occur is that the medium in which the light is internally reflected has a refractive index that is higher than that of the recording medium. The evanescent waves are then propagating parallel to the surface with planes of constant amplitude parallel to and planes of constant phase perpendicular to the surface of the recording me-

dium. The penetration depth of the evanescent waves is in practical cases of the order of the wavelength of the light used.

The evanescent wave created by internal reflection has a wavelength (see eq. (2.18))

$$\lambda_e = \lambda_s/\sin i = \lambda_0/(n_s \sin i), \tag{6.1}$$

where λ_s is the wavelength in the medium of refractive index n_s adjacent to the recording medium and λ_0 is the wavelength in vacuum. Thus, the phase velocity of the evanescent wave is

$$v_e = v_s/\sin i = c/(n_s \sin i), \tag{6.2}$$

where v_s is the velocity of light in the surrounding medium. From eqs. (6.1) and (6.2) is clear that $\lambda_e > \lambda_s$ and $v_e > v_s$. However, $v_e < c/n_e$, because $\sin i > n_e/n_s$. This implies that $\lambda_e < c/(n_e \nu) = \lambda'_e$, where λ'_e is the wavelength of the corresponding homogeneous wave of frequency ν in the recording medium. The evanescent wave created by internal reflection has, thus, a wavelength within the range

$$\lambda_0/n_s < \lambda_e < \lambda_0/n_e. \tag{6.3}$$

λ_e as well as v_e are independent of the properties of the recording medium.

The electrical field of a plane polarized evanescent wave propagating in the x-direction can be expressed as

$$E_e = A_r \exp\{-2\pi n_s z/\lambda_a\} \exp\{j2\pi x/\lambda_e\}, \tag{6.4}$$

where A_r is the amplitude at the surface and λ_a is given by eq. (2.19). From this expression we conclude two things which are of importance for evanescent wave holography: The penetration depth is dependent on the angle of incidence as well as on the difference in the refractive indices between the surrounding and recording media. Maximum penetration depth is obtained at the critical angle for total reflection. The energy density in the immediate vicinity of the boundary may reach considerable amounts. As is clear from expressions (2.22), density values in excess of four times the energy density of the incident wave can be obtained at the interface for incidence at the critical angle.

According to the inequality (6.3) the smallest wavelength λ_e which we may create by internal reflection is λ_0/n_s. Even smaller λ_e values can, of course, be produced by diffraction where the spatial frequency of the object causing diffraction determines the wavelength of the evanescent wave (see eq. (2.6)). For example, if we normally illuminate a regular plane grating having a period $a < \lambda_0$, then the diffracted waves with amplitude A_d are (see e.g. SCHAEFFNER and TORALDO DI FRANCIA [1949])

$$E_e = A_d \exp\{-2\pi n_s z/\lambda_a\} \exp\{\pm j2\pi mx/a\} \tag{6.5}$$

where m is the diffraction order and λ_a is now given by

$$\lambda_a = \lambda_0/\{(m\lambda_0/a)^2 - 1\}^{\frac{1}{2}}.$$

The wavelength of the evanescent wave formed by diffraction ($\lambda_e = a/m$ for the special case mentioned here) is, as in the internal reflection case, only dependent on the boundary conditions and the incident radiation. Further, we may conclude from expression (6.5) that the penetration depth of the evanescent wave decreases with decreasing wavelength λ_e, i.e. increasing spatial frequency in the object.

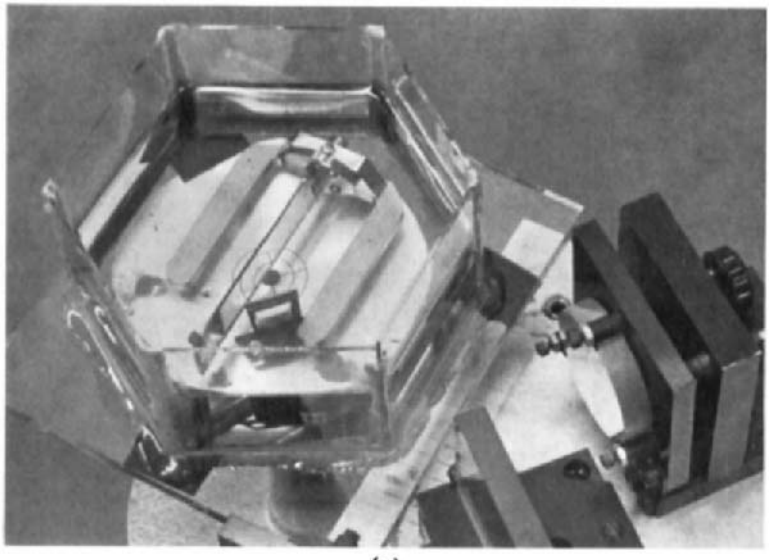
Another effect which can be used in evanescent wave holography is the possibility of influencing the propagation direction of the evanescent wave by proper choice of the state of polarization. As was described in Section 2.2, the exit and incidence planes at internal reflection coincide for light linearly polarized parallel or perpendicular to the plane of incidence, which means that the propagation direction of the evanescent wave is located in this plane. However, for other states of polarization, a propagation component with a direction transverse to this plane occurs, which is maximum for circularly polarized light.

From this section, we may suspect that application of properties such as wavelength independent of material properties, local existence of the waves, and polarization dependent propagation direction to the field of holography may result in peculiar and characteristic effects.

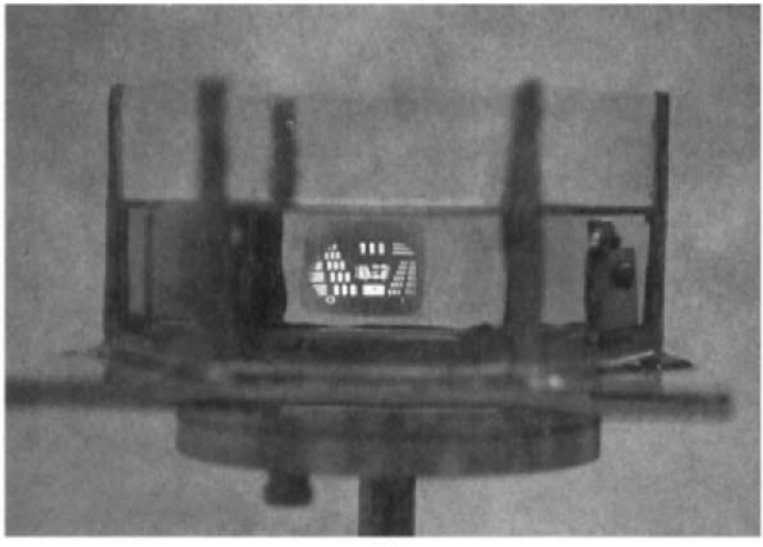
6.3. EXPERIMENTAL TECHNIQUES AND CONDITIONS

In the experiments reported on by Nassenstein and Bryngdahl, the evanescent waves in the hologram recording step were formed by using internal reflection techniques. Photographic emulsions were used as the recording medium. NASSENSTEIN [1969d, 1970a] used Agfa-Gevaert Scientia 8E75 and 10E75 emulsions, and BRYNGDAHL [1969] used Kodak spectroscopic plates, type 649F. These emulsions have a relatively high refractive index for a wavelength $\lambda_0 = 6328$Å before they are processed. However, the refractive index decreases drastically after development and fixing. Nassenstein gave the refractive index value $1.63 - j1.2 \times 10^{-3}$ for unexposed Agfa-Gevaert emulsion and Bryngdahl the value 1.61 for the real part of the refractive index of Kodak emulsion. The corresponding values given for processed emulsions were 1.54 and 1.56, respectively.

In order to be able to produce evanescent waves in these emulsions, the surrounding medium in which the internal reflection takes place has to have



(a)



(b)

Fig. 6.1. Experimental arrangements for evanescent wave holography. (a) Recording step using commercial photographic plates submerged in di-iodomethane, and (b) corresponding reconstruction configuration (BRYNGDAHL [1969]).

a refractive index higher than that of the emulsion. Two experimental techniques have been used to achieve this condition:

(1) Submerging photographic plates in a liquid with a high refractive index. Commercially available plates have substrates with lower refractive index than that of the photosensitive emulsion. Therefore, this method is preferable when we wish to use regular photographic plates. BRYNGDAHL [1969] and NASSENSTEIN [1969e, f, 1970a, b] used di-iodomethane, CH_2I_2, as immersion liquid. The refractive index of di-iodomethane is 1.73 at 6328Å. This ensures an angular range of at least 20° within which internal reflection is obtained. Further, no interaction was found between this liquid and the emulsion. The above-mentioned plates are convenient to use in experiments of this type because they are not overcoated with hardened gelatine which is usually applied in order to protect emulsions. Two optical configurations are possible. The liquid may be applied between the emulsion and a glass prism of high refractive index so that internal reflection occurs at the surface between the liquid and the emulsion. The other arrangement is to use a cell of convenient shape filled with di-iodomethane and submerge the plates. A cell with hexagonal bottom plate was found convenient for these experiments (BRYNGDAHL [1969]). Special holders for the plates and other objects have to be provided in order to keep them in place in the di-iodomethane which has a density of 3.3 g/cm^3. Figure 6.1 shows an example with this latter type of configuration. The recording geometry (Fig. 6.1(a)) is arranged so that the light from a diffuse object placed in the nearest portion of the cell will form the evanescent wavefield in the emulsion. The reference wave coming from the right penetrates the plate as a homogeneous wave. A reconstruction of a diffuse object is shown in Fig. 6.1(b) where a hologram is illuminated with an evanescent wave. In reconstruction, i.e. after processing the plates, the refractive index of the emulsion has decreased to a value about 1.55 and, thus, a wider variety of immersion liquids may be used than in the recording stage. Beside di-iodomethane, α-chloronaphthalene with a refractive index of 1.63 at 6328Å has been used with good results.

(2) Use of plates where the emulsion has been coated onto substrates with high optical density. NASSENSTEIN [1968, 1969d, 1970a] has used photographic layers on flint glass plates to make evanescent wave holograms. The substrate had then a higher refractive index than the emulsion. Here the internal reflection occurred in the glass substrate at the boundary between the glass and the photographic layer. This type of configuration seems to offer some advantages if the emulsion is bleached to obtain phase holograms. Eventual surface relief influence is avoided with this technique.

6.4. FORMATION OF AND RECONSTRUCTION FROM EVANESCENT WAVE HOLOGRAMS

As was described in the previous section, the evanescent waves only exist in the vicinity of one surface of the photographic layer irrespective of which experimental technique is used. Thus, only a portion of the emulsion thickness is filled with the hologram structure. For example, in case conventional

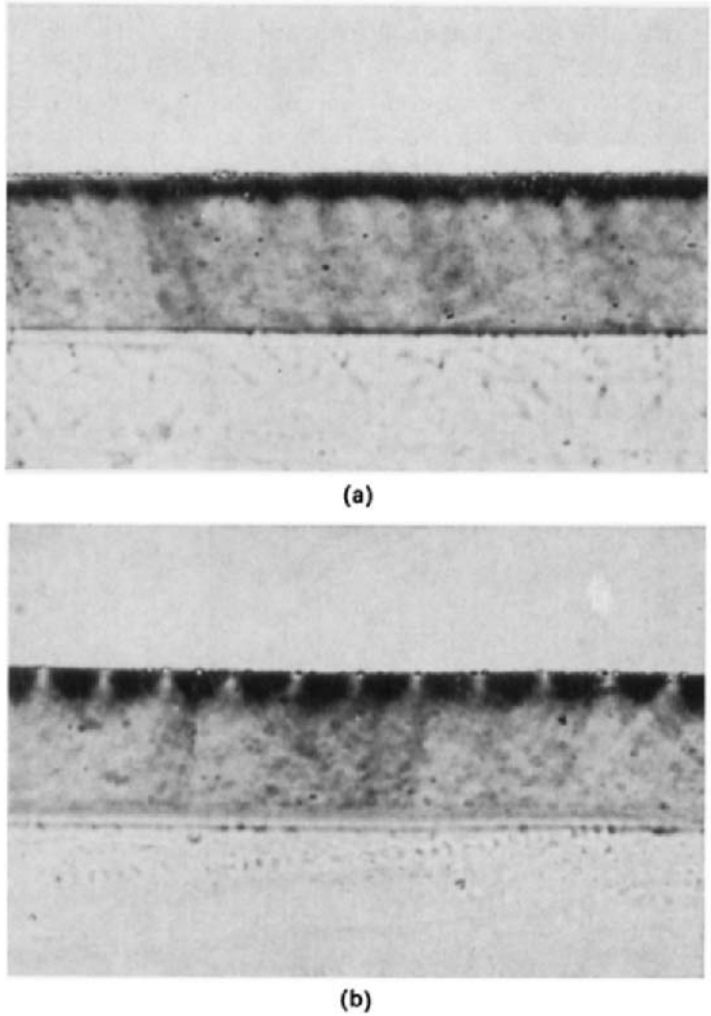


Fig. 6.2. Microphotographic microtome sections of photographic layers exposed to (a) a single evanescent wave, and (b) an interference pattern between two evanescent waves. The film base is below the photographic layer which is only blackened at its very top portion. (Courtesy of H. Nassenstein, Farbenfabriken Bayer AG, Leverkusen.)

Kodak 649F plates are used, the emulsion thickness is about 15μ but the hologram thickness is only about 1μ. This allows us to combine the two techniques (1) and (2) of Section 6.3 and use both surfaces of the emulsion for independent recordings.

In Fig. 6.2, microphotographic sections of processed photographic layers are shown which had been exposed by evanescent wavefields (NASSENSTEIN [1969b, 1970a]). The original 7μ thick gelatine layer in these illustrations had been swelled in water to about 26μ. One plane wave internally reflected in a medium optically denser than the emulsion was used to expose the film in Fig. 6.2(a) and two waves incident under different angles caused the interfering evanescent waves of Fig. 6.2(b). These microphotographs show clearly that the penetration depth of the evanescent waves is limited to about 1μ of the original thickness of the emulsion.

Two types of holograms are possible: either one or both of the interfering waves can be evanescent. For some applications also hybrid type holograms may be needed, e.g. the waves from the object are both homogeneous and

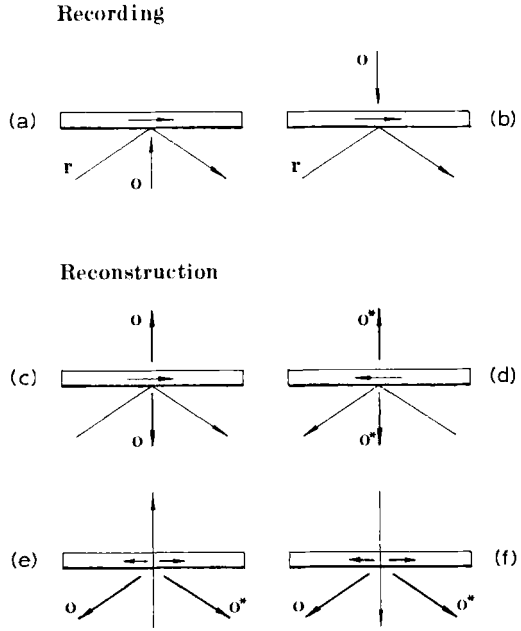


Fig. 6.3. Recording and reconstruction schemes for holograms formed by interference between a homogeneous object wave o and an evanescent reference wave r. The holograms recorded in (a) as well as (b) give the reconstructed waves – true wave field o and complex-conjugate wave field o^* – using the arrangements shown in (c), (d), (e), and (f) (BRYNGDAHL [1969]).

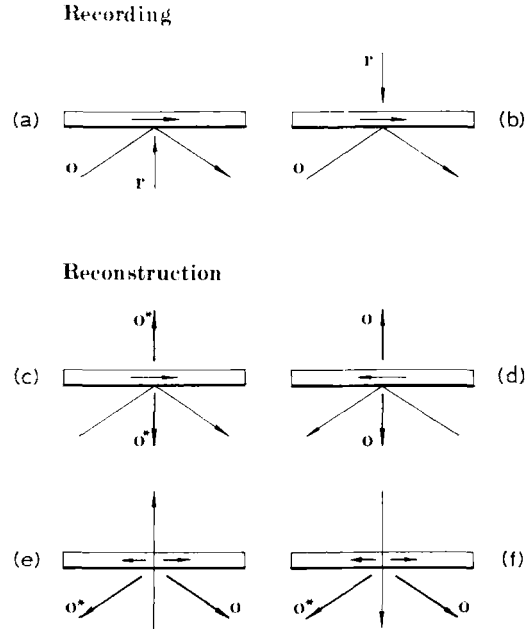


Fig. 6.4. Recording and reconstruction schemes for holograms formed by interference between an evanescent object wave o and a homogeneous reference wave r. The holograms recorded in (a) as well as (b) give the reconstructed waves – true wave field o and complex-conjugate wave field o^* – using the arrangements shown in (c), (d), (e) and (f) (BRYNGDAHL [1969]).

inhomogeneous and the reference wave may be a homogeneous or an inhomogeneous wave.

(1) Interference between a homogeneous and an evanescent wave (see Figs. 6.3 and 6.4). The evanescent field described by eq. (6.4) is

$$E_e = A_r \exp\{-2\pi n_s z/\lambda_a\} \exp\{j2\pi x/\lambda_e\}$$

and a homogeneous wave with normal incidence on the emulsion is given by

$$E_h = A \exp\{j2\pi z/\lambda_e'\},$$

where $\lambda_e' = \lambda_0/n_e$. The sign of the exponent indicates whether the homogeneous wave comes from the front or back of the plate. The interference fringe pattern between these waves is proportional to

$$|E_e+E_h|^2 = A^2+A_r^2 \exp\{-4\pi n_s z/\lambda_a\} + 2AA_r \exp\{-2\pi n_s z/\lambda_a\} \cos\{2\pi(x/\lambda_e - z/\lambda_e')\}. \quad (6.6)$$

The first term on the right indicates a uniform exposure throughout the

photographic layer and the second an exposure which rapidly attenuates away from $z = 0$, the surface of the emulsion. The last term describes the interference structure. This also is attenuated, but not as rapidly as the second term. The planes of the hologram fringes are given by

$$x/\lambda_e - z/\lambda_e' = \tfrac{1}{4}(2p+1), \qquad p \text{ integer.} \tag{6.7}$$

They are inclined to the surface of the emulsion and the distance between them is $\lambda_e\lambda_e'/(\lambda_e^2+\lambda_e'^2)^{\frac{1}{2}}$. Their separation in a plane parallel to the surface (z = const.) is λ_e.

Because of the rapid attenuation in the z-direction, holograms recorded from an exposure pattern (6.6) may in most cases be regarded as thin to a first approximation. This means that our hologram structure has a period equal to the wavelength λ_e of the evanescent wave. This also means that with the configuration described here, the same hologram is recorded regardless of which side of the plate the homogeneous wave enters. This situation is shown in Figs. 6.3 and 6.4 where (a) and (b) will produce identical recordings.

In reconstruction, these evanescent wave holograms show some peculiar features (BRYNGDAHL [1969]). These are in particular due to the collapsing thickness of the holograms and the evanescent waves propagating in the plane of the hologram. If the hologram is illuminated with the same evanescent wave that was used in the recording, then a copy of the recorded homogeneous wave is reconstructed. This is shown in Fig. 6.3(c). Because the illuminating wave here is an evanescent wave propagating along the hologram, two waves will be reconstructed. In addition to the reconstructed copy of the original wave an identical wave (mirror imaged) is reconstructed symmetrically toward the opposite side of the hologram as illustrated in the figure. The phase variation in the reconstructed wavefront is contained in the relative locations of the hologram fringes. These are sensed in the same way by the illuminating light (evanescent wave) which is diffracted symmetrically to both sides of the hologram. In order to reconstruct the complex conjugate wave, the illuminating wave has to come from the opposite direction, i.e., we have to reverse the direction of the evanescent wave. Then two reconstructions appear, both being complex conjugate with respect to the wave recorded (cf. Figs. 6.3(d) and eq. (6.6)).

There is also another reconstruction geometry possible. This is illustrated in (e) and (f) of Figs. 6.3 and 6.4. Homogeneous waves are now used to illuminate the hologram. Then, the hologram structure will diffract the light so that evanescent waves are created which propagate along the hologram. If the medium that surrounds the hologram is optically denser than the processed emulsion, the evanescent waves are converted into homogeneous

waves which leave the hologram surface at angles that exceed the critical angle of internal reflection. For commercially available plates immersed in optically dense liquids, the refractive index conditions only permit reconstructed waves on one side of the hologram as is shown in (e) and (f) of Figs. 6.3 and 6.4.

(2) Interference between two evanescent waves (see Figs. 6.5–6.7). Two evanescent waves described by eq. (6.4)

$$E_{e1} = A_{r1} \exp\{-2\pi n_s z/\lambda_{a1}\} \exp\{j2\pi x/\lambda_{e1}\}$$

and

$$E_{e2} = A_{r2} \exp\{-2\pi n_s z/\lambda_{a2}\} \exp\{j2\pi x/\lambda_{e2}\}$$

cause an interference pattern proportional to

$$|E_{e1}+E_{e2}|^2 = A_{r1}^2 \exp\{-4\pi n_s z/\lambda_{a1}\} + A_{r2}^2 \exp\{-4\pi n_s z/\lambda_{a2}\} + 2A_{r1} A_{r2} \exp\{-2\pi n_s z(1/\lambda_{a1}+1/\lambda_{a2})\} \cos\{2\pi x(1/\lambda_{e1}-1/\lambda_{e2})\}. \quad (6.8)$$

The first two terms to the right attenuate rapidly in the z-direction. The holographic periodic structure described by the third term is also evanescent perpendicular to the emulsion surface ($z = 0$), but not quite as rapidly. The period of this structure is

$$a = \lambda_{e1} \lambda_{e2}/(\lambda_{e2}-\lambda_{e1}). \quad (6.9)$$

Here, it was assumed that the two evanescent waves have the same propa-

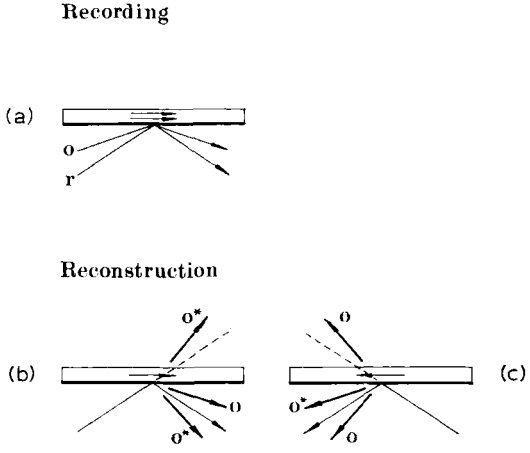


Fig. 6.5. Recording and reconstruction schemes for a hologram formed by interference between an evanescent object wave o and an evanescent reference wave r with the same propagation directions. (b) and (c) show the directions of the reconstructed true wave field o and complex-conjugate wave field o^* for opposite directions of the illuminating evanescent wave (BRYNGDAHL [1969]).

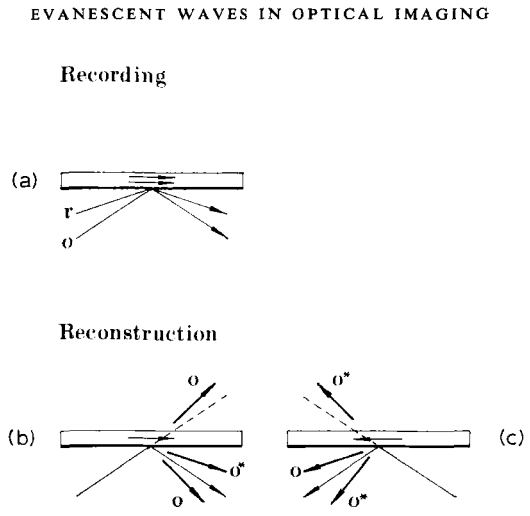


Fig. 6.6. Recording and reconstruction schemes for a hologram formed by interference between an evanescent object wave o and an evanescent reference wave r with the same propagation direction. (b) and (c) show the directions of the reconstructed true wave field o and complex-conjugate wave field o^* for opposite directions of the illuminating evanescent wave (BRYNGDAHL 1969]).

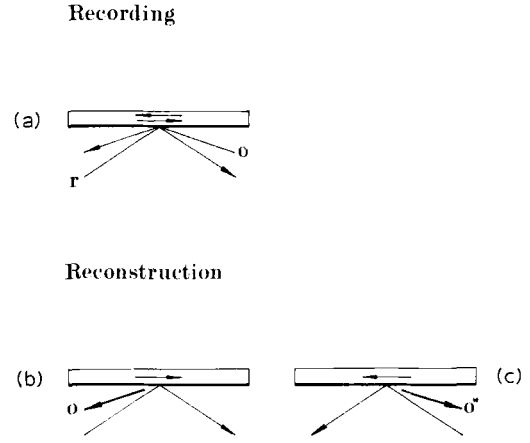


Fig. 6.7. Recording and reconstruction schemes for a hologram formed by interference between an evanescent object wave o and an evanescent reference wave r with opposite propagation directions. (b) and (c) show the directions of the reconstructed true wave field o and complex-conjugate field o^* for opposite directions of the illuminating evanescent wave.

gation direction (cf. Figs. 6.5 and 6.6). This is, of course, only a special case. The two evanescent waves can interfere under any angle provided they have a common polarization component. In the other special case, where the

waves travel in opposite directions (cf. Fig. 6.7), the period of the interference pattern is

$$a = \lambda_{e1}\lambda_{e2}/(\lambda_{e1}+\lambda_{e2}). \tag{6.10}$$

For $\lambda_{e1} = \lambda_{e2}$ eq. (6.10) becomes $a = \frac{1}{2}\lambda_{e1}$. An interference field will, of course, also exist in the surrounding medium in which the light is internally reflected. At the very surface of the emulsion (at $z = 0$) the two interference patterns will coincide. The difference between them is that the fringes of the one in the surrounding medium in general are inclined to the surface, but the fringes of the pattern existing in the emulsion are always perpendicular to the surface.

From eq. (6.10) is clear that the smallest structure in the interference pattern is $\frac{1}{2}\lambda_e$. If we allow opposed homogeneous waves to interfere in the emulsion, the pattern will have a period $\frac{1}{2}\lambda'_e$. However, $\lambda_e < \lambda'_e = \lambda_0/n_e$ which means that with evanescent wave holographic techniques we are able to record structures that are finer than those which we may form by homogeneous standing waves of the same frequency. This implies that we are able to form structures that we will not be able to observe with conventional optics using only homogeneous waves.

In the configurations of Figs. 6.5–6.7, two evanescent waves with different wavelengths but the same frequency interfere. Just as in conventional holography, the two waves will in practice originate from the same source and be separated by a beam-splitter. Another peculiarity in reconstruction from these holograms is the possible directions of the diffracted waves. Conventionally when homogeneous waves are diffracted by a thin grating with coarse structures, the wave vectors of the diffracted waves are all located in the plane of incidence (plane formed by wave vector of the illuminating wave and normal to the grating lines). However, when illuminating a thin grating with evanescent waves which propagate in the plane of the grating any direction of the diffracted waves is possible. What generally occurs is that the illuminating evanescent wave is diffracted by the hologram into other evanescent waves which when certain boundary conditions are met can be converted into propagating homogeneous waves. When the illuminating evanescent wave propagates perpendicular to the grating structure as in the special cases of eqs. (6.9) and (6.10) then the diffracted waves are all located in the plane of incidence.

In Figs. 6.5–6.7, situations described by eqs. (6.9) and (6.10) are illustrated. The main difference between these two cases concerns the spatial frequency of the recording. In Figs. 6.5 and 6.6, where the evanescent waves have the same propagation direction, the period of the grating is large. Then the

angular separation between the 0th order and the reconstructed waves is small. Similar reconstructions are formed on both sides of the hologram, because the illuminating evanescent wave is propagating in the plane of the hologram. However, the refractive indices of the substrate and emulsion do not generally allow reconstruction on the substrate side of the recording. If, as shown in Figs. 6.5 and 6.6 reconstruction is performed with a wave whose propagation direction only slightly exceeds the critical angle of total reflection, then it may be possible to extract a reconstructed image on the substrate side of the emulsion. This reconstruction is of particular interest for some applications because the 0th order (internally reflected light) does not appear on the same side of the hologram (cf. also Figs. 6.3 and 6.4 where the same situation occurs). These holograms which have low spatial frequencies can, of course, also be illuminated with a homogeneous wave. However, the reconstructed waves are now astigmatically distorted. Acceptable reconstructions can only be obtained in this case by introducing a cylinder lens after the hologram. In Fig. 6.7, on the other hand, where the evanescent waves opposite directions, the spatial frequencies of the hologram are extremely high. In this case, reconstructions can only be obtained by illumination with an evanescent wave.

6.5. SOME PECULIAR CHARACTERISTICS OF EVANESCENT WAVE HOLOGRAPHY

Evanescent wave holography possesses some characteristics that are unique for this technique of wavefront reconstruction. These properties result from the evanescent waves – their small penetration depth which tends to make the holograms extremely thin and the fact that their wavelengths are independent of the properties of the material in which they propagate.

The diffraction efficiency of evanescent wave holograms is in general higher than that obtainable with conventional absorption type holograms using homogeneous waves. This is true for holography in which the evanescent waves have been produced by internal reflection. The effect is particularly pronounced for angles of incidence of the illuminating light which are close to the critical angle of total reflection. In this region, the induced electric field can reach (and in certain cases even surpass) twice the value of the illuminating wave (cf. eq. (2.22)). NASSENSTEIN [1969e] measured diffraction efficiencies as high as 22.6 % for an absorption grating of 40 lines/mm (Agfa-Gevaert Scientia 8E75 and 10E75) illuminated with light polarized parallel to the plane of incidence at the angle of total reflection. This is appreciably more than 6.25 %, which is the maximum value that can be reached using homogeneous waves. Even for phase holograms, the efficiencies were higher

when evanescent compared to homogeneous waves were used for reconstruction. The efficiency falls off with increasing spatial frequency. For example, Nassenstein found a diffraction efficiency of 0.15 % for an absorption grating of 5220 lines/mm. This efficiency is considerably higher than can be expected from an extrapolation from the transfer function valid for homogeneous waves. BRYNGDAHL [1969] reported on efficiency of 1 % of holograms where the light from a diffusely illuminated object (Air Force target) had been internally reflected at the surface between di-iodomethane and Kodak 649F emulsion.

Section 2.2 described how the evanescent wave created by internal reflection is influenced by the state of polarization of the incident light. NASSENSTEIN [1969e] obtained a higher diffraction efficiency for light linearly polarized parallel to than perpendicular to the plane of incidence. This is in agreement with eqs. (2.25) and (2.26) and Fig. 2.4 which for angles of incidence close to the critical angle predict a larger displacement and thus a larger penetration depth and interaction length for the light with its polarization component parallel to the plane of incidence. BRYNGDAHL [1969], on the other hand, did not find any pronounced difference between the two states of linear polarization. He even found that light polarized perpendicular to the plane of incidence is preferable in some cases. This is probably because of its smaller absorption and its relation to diffraction phenomena. Another explanation is that for angles well above the critical angle of total reflection light polarized perpendicular to the plane of incidence penetrates deeper into the rarer medium (see Fig. 2.4).

A further feature of evanescent wave holography is that white light can be used for reconstruction (BRYNGDAHL [1969]). This possibility is explained by Fig. 6.8. In (a), the reconstruction using a conventional volume type hologram is shown. Constructive interference occurs for a particular wavelength only when the Bragg-condition is satisfied. Secondary scattering

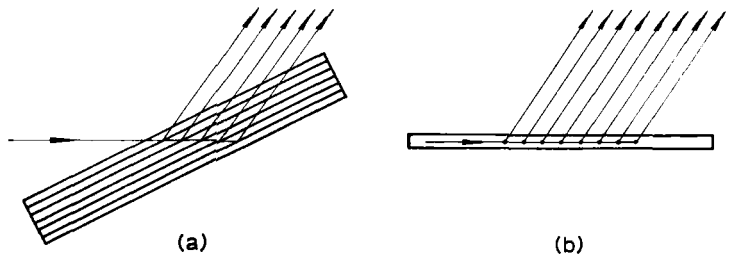


Fig. 6.8. Illustration of the similarity in reconstruction from a conventional volume-type hologram (a) and an evanescent wave hologram (b). The selectivity of the Bragg condition allows white light to be used in both cases.

which gives rise to straylight occurs when the primarily scattered light penetrates neighboring hologram fringes. In Fig. 6.8(b), the corresponding situation using evanescent wave holography is illustrated. Here it is shown how the illumination of a thin hologram with a wave propagating in the plane of the hologram is equivalent to the situation in Fig. 6.8(a). Furthermore, because the primarily scattered light does not have to penetrate neighboring fringes, evanescent wave holography is not disturbed by secondary scattering. In the cases described in Section 6.4, besides using monochromatic coherent light, it is also possible to reconstruct in white light. Shrinkage in depth of the processed emulsion does not influence the reconstruction in the case of Fig. 6.8(b), because the hologram is only confined to the very surface of the photographic emulsion.

The possibility of changing the wavelength of the evanescent wave without changing its frequency or direction (or changing the frequency without changing its wavelength) gives flexibility, especially to the white light reconstruction. In general, a wavelength change occurs when an evanescent wave is diffracted. In the case of a three-dimensional structure, this wavelength change requires a modified Bragg-condition (NASSENSTEIN [1970a]); the angle of reflection from an interference plane is no longer equal to the angle of incidence. However, the three-dimensional case is only of secondary importance because of the small penetration depth of the evanescent waves.

Evanescent waves are often formed in the reconstruction process of conventional holograms (WOLF and SHEWELL [1970]). Especially in cases where the illuminating wave is not normally incident on the hologram, one of the reconstructed images, or a portion of one image, is sometimes absent. This happens because of diffraction of the illuminating wave into evanescent waves by the hologram. The more the illuminating wave is inclined toward the hologram, the more of the reconstructed image in question will disappear into the plane of the hologram. In the type of evanescent wave holography where reconstructed evanescent waves are converted into homogeneous waves, the situation is similar. Only evanescent waves with wavelengths $\lambda_0/n_s < \lambda_e < \lambda_0/n_e$ (cf. (6.3)) will be converted into homogeneous waves using the techniques of Section 6.4. Information carried by evanescent waves with $\lambda_e < \lambda_0/n_s$ cannot be recovered. These evanescent waves can no longer be converted into homogeneous waves. Thus, the angular region $0° - i_t$ corresponds to diffracted homogeneous waves and the region $i_t - 90°$ to conversion of diffracted evanescent waves into homogeneous waves.

6.6. APPLICATIONS OF EVANESCENT WAVE HOLOGRAPHY

Evanescent wave holography appears to be one of our most powerful tools

to study and to deepen our insight into the properties of these waves.

Holography with evanescent waves is one possible way to store information contained in these waves. As such, it may be considered as an evolution of holography in order to be able to store information about the total optical wave field. One of its more evident applications in this respect is in the formation of high-aperture and high-resolution optical images. Evanescent wave holography offers two major advantages here. An aperture filling the whole half space can be used in recording and making reconstructions from holograms. The reference and illuminating waves can be confined to that side of the hologram which is opposite to the object and reconstructed image (cf. (b), (c), and (d) in Figs. 6.3 and 6.4). Further, this type of hologram shows high efficiency and insensitivity to shrinkage.

All the unconventional as well as conventional types of holography with homogeneous waves may also be carried out using evanescent waves. Thus, incoherent (spatial or temporal incoherence), synthetic and computer-generated holography can be applied. Extremely fine structures recorded with electron beam techniques can be read out using light waves. Optical and ultrasonic techniques may also be combined.

Another promising application lies in the possibility of using evanescent waves to obtain information about portions of object spectra that are located beyond the frequency set by the conventional resolution limit in optics, which is determined by the aperture and wavelength of the light. As was described in Section 2.1, information about object structures smaller than half the wavelength is only contained in the evanescent wave field. The first technique described in Section 6.1 is one possible approach to recording such information. As the evanescent waves have a shorter wavelength than the corresponding homogeneous waves of the same frequency, an increase in resolution ought to be obtained by illuminating our object with evanescent waves. NASSENSTEIN [1969a, f, 1970b] has shown how illumination of an object with evanescent waves results in diffracted homogeneous waves which contain information of high spatial frequencies of the object. This means that high spatial frequencies are transformed into low spatial frequencies. However, a magnification and a shift of the spectrum is also necessary to obtain a magnified image of the original object. Illumination of an object of spatial wave number k_0 with an evanescent wave with propagation vector k_e results in $k_r = k_0 - k_e$. Magnification by a factor M gives $k_r/M = k_0/M - k_e/M$, which after a shift of k_e/M results in k_0/M. NASSENSTEIN [1970b] has experimentally verified this procedure using evanescent wave holography, which clearly shows that information about object details smaller than the value determined by the classic resolution limit can be obtained using eva-

nescent wave illumination. The resolution is now set by the wavelenhgt of the evanescent wave. This in turn is determined by the conditions for internal reflection or other techniques that may be used to create the evanescent light.

Even finer details may be examined with other configurations. The scheme of Fig. 2.7(e) is particularly powerful in regard to spatial frequency transformation. The principle of this coupling between gratings of extremely short periods can conveniently be treated as a Moiré-effect. Of course, the distance between the gratings has to be shorter than the extension range of the evanescent waves.

Evanescent wave holography may also have some potential use in connection with integrated optics – either for component fabrication or as a mean of deflecting or transfering energy. Further, when optical signals are transmitted through a thin film, a strong evanescent wave field generally exists in the medium surrounding this film. This makes evanescent wave holographic techniques especially suited for introduction and extraction of spatial information in integrated optical circuits.

§ 7. Lateral Waves in Optical Imaging Situations

In § 2, it was mentioned that when a homogeneous wave is converted into an evanescent wave and then reconverted into a homogeneous wave, either by using total reflection or grating techniques, the wave will encounter a lateral displacement. As also mentioned, this displacement is dependent on several parameters such as angle of incidence, frequency and state of polarization of the light as well as the refractive indices of the medium in which the light is internally reflected and of the rarer medium. In § 2, we only treated what happens for one plane wave. If, on the other hand, we regard a convergent or divergent wave, the amount of displacement will vary over the wavefront, which will result in aberrations. Spherical aberration and astigmatism will in general be present in the reflected light and vary with the state of polarization (PICHT [1955]).

7.1. CREATION OF LATERAL WAVES IN TOTAL REFLECTION

Evanescent waves arise in the rarer medium when light is internally reflected at an interface between two media. This means that evanescent waves occur for all angles of incidence which are larger than the critical angle of total reflection. For angles smaller than the critical angle light will be refracted into the rarer medium. In the transition region between these two cases, the refracted light will travel along the interface where it will cause a disturbance which in turn will give rise to light – a lateral wave – in the

optically denser medium. As is illustrated in Fig. 7.1, a wave traveling along the interface leaks continuously back into the denser medium. There it acts like an inhomogeneous wave (cf. locations of planes with constant phase

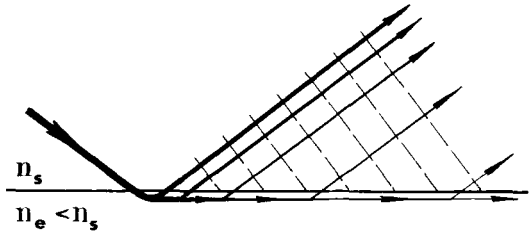


Fig. 7.1. Illustration of the lateral wave field that accompanies the reflected part of a wave incident at the critical angle of total reflection. Solid lines represent surfaces of constant amplitude and dashed lines surfaces of constant phase.

and amplitude in Fig. 7.1). The irradiance corresponding to this wave is usually small compared to that of the reflected wave.

As is clear from Fig. 7.1, inhomogeneous waves may exist at total reflection in both the medium in which the light is internally reflected and the rarer one. The inhomogeneous wave in the rarer medium is the evanescent wave treated in the previous paragraphs. The energy of this wave is coupled in as well as out from the rarer medium through the interface which means that the propagating evanescent waves will be localized to regions specified by the spatial distribution of the homogeneous wave incident on the interface (cf. Section 2.4). The inhomogeneous wave propagating along with the reflected wave in the optically denser medium, on the other hand, is called a lateral wave and in contrast to the evanescent wave, it is not confined to an interface and, thus, it will not be converted back into a homogeneous wave in close proximity to where it was formed.

Lateral waves which are more pronounced in other fields like elasticity and acoustics, e.g., they occur as boundary layer waves at earthquakes and give the first seismic response, were given a physical interpretation by V. SCHMIDT [1934, 1938]. V. SCHMIDT [1938] was able to show that the reflection of spherical waves at an acoustically less dense medium gives rise to peculiar additional waves accompanying the regular reflected ones. By means of Schlieren photographs, he showed their existence in acoustical spark experiments arranged in the way illustrated in Fig. 7.2. Of course, the situation illustrated in Fig. 7.2 is equally valid for the optical case. A point source emits spherical waves in the medium with the lower velocity. Portions of these waves will be reflected and the rest will be refracted (the wavefronts

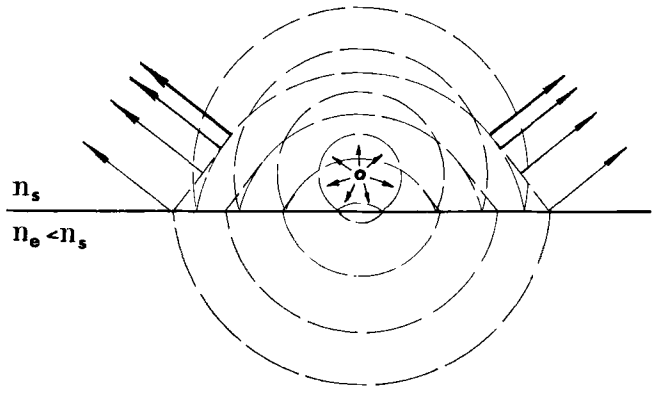


Fig. 7.2. Illustration of the formation of a conical wave in addition to the regular reflected and refracted ones when spherical waves are reflected at an optically less dense medium.

are shown dashed in Fig. 7.2). As long as the angle of incidence is smaller or equal to the critical angle of total reflection, the wavefronts of the incident, reflected, and refracted waves cross at the interface. For angles larger than the critical angle, the incident wave causes an evanescent wave in the rarer medium, i.e. the medium, with higher velocity, and is internally reflected. The refracted wave, on the other hand, that propagates along the interface will propagate on that side which gives it the higher velocity. Due to continuity requirements, this wave will give rise to a new wave in the medium with the lower velocity and its propagation direction is determined by the law of refraction, i.e., the light will leave the interface at the critical angle of total reflection. The specific geometry of Fig. 7.2 results in a lateral wavefield of conical shape which connects tangentially with the reflected wave and intersects the refracted wave at the interface. Thus, the lateral waves will cause the image of a point source to appear comet-shaped (MAECKER [1949]). The amplitude ratio between the waves polarized parallel and perpendicular, respectively, to the plane of incidence and which have penetrated the interface at the critical angle is n_s/n_e (see eqs. (2.22)). At the conversion of the refracted wave into a lateral wave, this amplitude ratio will be multiplied by the factor n_s/n_e. Thus, $|E_{||}/E_{\perp}|^2 = (n_s/n_e)^4$ which shows that the lateral waves are much weaker for incident light that is polarized perpendicular to the plane of incidence than parallel to it (MAECKER [1949]).

MAECKER [1949] showed by using a ray-tracing approach that there exists a close relationship between lateral waves and the lateral beam displacement that occurs at total reflection. Figure 7.3 illustrates this connection. Rays at angles exceeding the critical angle for total reflection from a point source

are shown. These rays are differently inclined to the interface and encounter different amounts of lateral displacement. It is clear from the figure that for angles close to the critical angle a lateral wave (conical wavefront) is obtained and for angles further away from the critical angle, a reflected wave with a curved wavefront occurs. A similar ray-tracing procedure for a convergent incident wave reveals that lateral waves are not formed in that case. Thus, only when the incident wave is divergent or plane and limited (as mentioned below) will lateral waves exist.

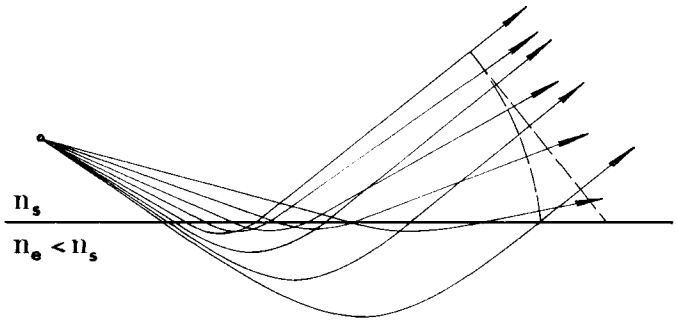


Fig. 7.3. Illustration of the existence of a lateral wave resulting from the lateral beam displacement when a diverging wave is totally reflected.

In case we have a parallel beam of limited extension, the diffracted waves caused by the edges of the field will be superposed on the plane wave field. Since these diffracted waves are divergent, they can generate lateral waves when reflected at an interface. In general, the irradiance of the lateral waves is quite small. However, in special cases such as at grazing incidence, absorbing denser medium, and collimated light incident at the critical angle, i.e., in cases where an interference or an absorption mechanism reduces the reflected field, the lateral wave (the diffracted field) may be strong and may even be the dominant one (TAMIR and OLINER [1969]).

As described above, a conical wavefront is created as a second order effect when light from a point source is internally reflected at a plane interface (see Fig. 7.2). A similarly shaped wavefront also occurs in another phenomenon in which evanescent waves play a major role, namely the conical wavefront associated with Čerenkov radiation (TORALDO DI FRANCIA [1960], ASBY and WOLF [1971]). This wave appears when an electron moves in a straight line with a velocity that is greater than the phase velocity of light in the medium. Toraldo di Francia assumed in order to explain the Čerenkov effect an electron that moves in close vicinity to a dielectric material similar to the situation illustrated in Fig. 7.1. The electric field caused by the elec-

tron may be regarded as being made up of evanescent waves. Thus, if the electron is close enough to the interface, some of these evanescent waves will be refracted and converted to homogeneous propagating waves. With the same kind of reasoning TORALDO DI FRANCIA [1960] was also able to explain the Smith-Purcell effect, i.e., the effect observed when an electron moves close to a metallic grating. Radiation occurs then in certain directions which are specified by the velocity of the electron and the period of the grating.

7.2. EXPERIMENTAL VERIFICATION OF OPTICAL LATERAL WAVES

Since v.Schmidt performed his experiments in the 1930's, several theoretical treatments of optical lateral waves have appeared (see e.g. OTT [1942, 1949], BREKHOVSKIKH [1960], FELSEN [1967], TAMIR and OLINER [1969] and TAMIR [1972]).

The first observations of lateral waves in optics were made by MAECKER [1949]. Later experiments have been performed by ACLOQUE and GUILLEMET [1960] and OSTERBERG and SMITH [1964]. These experiments use in essence a beam of light which, after it is internally reflected in a glass prism, is blocked by an absorbing medium. The light trailing the totally reflected beam was studied and found to be well observable at angles of incidence close to (within 2 sec of arc) the critical angle of total reflection. It was found that the trailing light had the same propagating direction as the totally reflected light. The lateral wave field was found to decrease with the distance away from the reflected wave. The calculations by TAMIR and OLINER [1969] for a plane incident wave indicate a decrease proportional to $x_1^{-\frac{3}{2}}$, where x_1 is the distance the wave travels along the interface. The lateral wave field in the less dense medium decays exponentially with distance normal to the interface (BREKHOVSKIKH [1960]). Thus, the power flow of the lateral wave in the less-dense medium is confined to a narrow layer close to the interface, much as are evanescent waves in the same medium.

References

ACLOQUE, P. and C. GUILLEMET, 1960, Comt. Rend. **250B**, 4328.
ARTMANN, K., 1942, Z. Physik **119**, 529.
ARTMANN, K., 1948, Ann. Physik (6) **2**, 87.
ARZELIÈS, H., 1946, Ann. Physique (12) **1**, 5.
ASBY, R. J. and E. WOLF, 1971, J. Opt. Soc. Am. **61**, 52.
BARLOW, H. M. and J. BROWN, 1962, Radio Surface Waves (Oxford University Press, Oxford).
BAUMEISTER, P. W., 1967, Appl. Opt. **6**, 897.
BEY, P. P., J. F. GIULIANI and H. RABIN, 1969, Phys. Rev. **184**, 849.
BLOEMBERGEN, N. and C. H. LEE, 1967, Phys. Rev. Letters **19**, 835.

BRADY, J. J., R. O. BRICK and M. D. PEARSON, 1960, J. Opt. Soc. Am. **50**, 1080.
BREKHOVSKIKH, L. M., 1960, Waves in Layered Media (Academic Press, New York).
BRYNGDAHL, O., 1969, J. Opt. Soc. Am. **59**, 1645.
CARNIGLIA, C. K. and L. MANDEL, 1971, J. Opt. Soc. Am. **61**, 1035 and 1423.
CARNIGLIA, C. K., L. MANDEL and K. H. DREXHAGE, 1972, J. Opt. Soc. Am. **62**, 479.
COON, D. D., 1966, Am. J. Phys. **34**, 240.
COSTA DE BEAUREGARD, O., 1964, Cahiers Phys. **18**, 471.
COSTA DE BEAUREGARD, O., 1965, Phys. Rev. **139**, B1446.
CULSHAW, W., 1961, Advances in Electronics and Electron Physics, Vol. 15 (Academic Press, New York) pp. 197–263.
CULSHAW, W. and D. S. JONES, 1953, Proc. Phys. Soc. **B66**, 859.
DAKSS, M. L., L. KUHN, P. F. HEIDRICH and B. A. SCOTT, 1970, Appl. Phys. Letters **16**, 523.
DITSCHEINER, L., 1870, Sitzber. k. Gesellsch. Wiss. Wien (2) **60**, 584.
DREXHAGE, K. H., 1970, Sci. Amer. **222**, No, 3, 108.
DREYHAUPT, W., 1939, Werkstattstechnik Werksleiter **33**, 321.
EDSER, E. and E. SENIOR, 1902, Phil. Mag. (6) **4**, 346.
EICHENWALD, A., 1911, Ann Physik (4) **35**, 1037.
EXNER, K., 1889, Wien. Sitzungsber. (IIa) **98**, 51.
FANO, U., 1938, Ann. Physik (5) **32**, 393.
FEDOROV, F. I., 1955, Dokl. Akad. Nauk SSSR **105**, 465.
FELSEN, L., 1967, in: Electromagnetic Wave Theory, Part 1, ed. J. Brown (Pergamon Press, Oxford) pp. 11–44).
FRAGSTEIN, C. v., 1949, Ann. Physik (6) **4**, 271.
GOOS, F. and H. HÄNCHEN, 1943, Ann. Physik (5) **43**, 383.
GOOS, F. and H. HÄNCHEN, 1947, Ann. Physik (6) **1**, 333.
GOOS, F. and H. LINDBERG-HÄNCHEN, 1949, Ann. Physik (6) **5**, 251.
GUNTER Jr., W. D., E. F. ERICKSON and G. R. GRANT, 1965, Appl. Opt. **4**, 512.
HALL, E. E., 1902, Phys. Rev. **15**, 73.
HARRICK, N. J., 1962, J. Appl. Phys. **33**, 2774.
HARRICK, N. J., 1963, Philips Tech. Rev. **24**, 271.
HARRICK, N. J., 1967, Internal Reflection Spectroscopy (John Wiley, New York).
HESSEL, A and A. A. OLINER, 1965, Appl. Opt. **4**, 1275.
HILL, K. O., A. WATANABE and J. G. CHAMBERS, 1972, Appl. Opt. **11**, 1952.
HOROWITZ, B. R. and T. TAMIR, 1971, J. Opt. Soc. Am. **61**, 586.
IMBERT, C., 1968, Comt. Rend. **267B**, 1401.
IMBERT, C., 1969, Comt. Rend. **269B**, 1227.
IMBERT, C., 1970a, Comt. Rend. **270B**, 529.
IMBERT, C., 1970b, Phys. Letters **31A**, 337.
IMBERT, C., 1972, Phys. Rev. **5**, D787.
INGERSOLL, L. R., 1920, J. Astrophys. **51**, 129.
KAPANY, N. S., 1967, Fiber Optics (Academic Press, New York).
KAPANY, N. S. and J. J. BURKE, 1961, J. Opt. Soc. Am. **51**, 1067.
KAPANY, N. S. and J. J. BURKE, 1972, Optical Waveguides (Academic Press, New York).
KARP, S. N. and J. RADLOW, 1956, IRE Trans. **AP-4**, 654.
KETTLER, E., 1899, Ann. Physik (3) **67**, 879.
KIZEL, V. A., 1968, Sov. Phys. Usp. **10**, 485.
KOESTER, C. J. and E. SNITZER, 1964, Appl. Opt. **3**, 1182.
KOROBKO-STEFANOV, A. A., 1950, Usp. Fiz. Nauk **42**, 433.
KORTÜM, G., 1969, Reflectance Spectroscopy (Springer Verlag, New York).
LAMPA, A., 1899, Wien. Sitzungsber. (IIa) **108**, 786.
LEITH, E. N. and J. UPATNIEKS, 1967, in: Progress in Optics, Vol. VI, ed. E. Wolf (North-Holland, Amsterdam) pp. 1–52.

LIPPMANN, B. A. and A. OPPENHEIM, 1954, Tech. Research Group, N.Y.
LISITSA, M. P., L. I. BEREZHINSKII and M. YA. VALAKH, 1972, Fiber Optics (Keter, New York).
LITTMANN, H , 1940, Ann. Physik (5) **38**, 139.
LOTSCH, H. K. V., 1968, J. Opt. Soc. Am. **58**, 551.
LOTSCH, H. K. V., 1970, Optik **32**, 116 and 189.
LOTSCH, H. K. V., 1971, Optik **32**, 299 and 553.
MAECKER, H., 1949, Ann. Physik (6) **4**, 409.
MCCUTCHEN, C. W., 1964, Rev. Sci. Instr. **35**, 1340.
MIDWINTER, J. E. and F. ZERNIKE, 1970, Appl. Phys. Letters **16**, 198.
MILLAR, R. F., 1960, Can. J. Phys. **38**, 272.
MILLER, S. E., 1969, Bell Syst. Tech. J. **48**, 2059.
NASSENSTEIN, H., 1968, Phys. Letters **28A**, 249.
NASSENSTEIN, H., 1969a, Phys. Letters **29A**, 175.
NASSENSTEIN, H., 1969b, Optik **29**, 456.
NASSENSTEIN, H., 1969c, Optik **29**, 597.
NASSENSTEIN, H., 1969d,Optik **30**, 44.
NASSENSTEIN, H., 1969e, Optik **30**, 201.
NASSENSTEIN, H., 1969f, Opt. Commun. **1**, 146.
NASSENSTEIN, H., 1970a, Naturwissenschaften **57**, 468.
NASSENSTEIN, H., 1970b, Opt. Commun. **2**, 231.
NEWTON, I., 1717, Opticks, 2nd ed., book III, part 1, query 29 and book II, part 1, observation 8 (Dover Publications, New York, 1952).
NOETHER, F., 1931, Ann. Physik (5) **11**, 141.
OSTERBERG, H. and L. W. SMITH, 1964, J. Opt. Soc. Am. **54**, 1073 and 1078.
OTT, H., 1942, Ann. Physik (5) **41**, 443.
OTT, H., 1949, Ann. Physik (6) **4**, 432.
PALMER, C. H., 1952, J. Opt. Soc. Am. **42**, 269.
PALMER, C. H., 1956, J. Opt. Soc. Am. **46**, 50.
PALMER, C. H., 1961, J. Opt. Soc. Am. **51**, 1438.
PICHT, J., 1929, Ann. Physik (5) **3**, 433.
PICHT, J., 1955, Optik **12**, 41.
QUINCKE, G., 1866, Ann. Phys. Chem. **127**, 1 and 199.
RAYLEIGH, Lord, 1907, Proc. Roy. Soc. (London) **A79**, 399.
RENARD, R. H., 1964, J. Opt. Soc. Am. **54**, 1190.
RICHARD, J., 1970, Nouv. Rev. d'Opt. Appl. **1**, 273.
RIGHI, A., 1898, Die Optik der Elektrischen Schwingungen (Leipzig) p. 161.
SCHAEFER, C. and G. Gross, 1910, Ann. Physik (4) **32**, 648.
SCHAEFER, C. and R. PICHT, 1937, Ann. Physik (5) **30**, 245.
SCHAFFNER, M. and G. TORALDO DI FRANCIA, 1949, Nuovo Cimento **6**, 125.
SCHILLING, H., 1965, Ann. Physik (7) **16**, 122.
SCHMALTZ, G., 1936, Technische Oberflächenkunde (Springer Verlag, Berlin).
SCHMIDT, O. V., 1934, Ann. Physik (5) **19**, 891.
SCHMIDT, O. V., 1938, Z. Tech. Physik **19**, 554.
SCHOCH, A., 1952, Acustica **2**, 18.
SÉLÉNYI, P., 1913, Compt. Rend. **157**, 1408.
SNITZER, E., 1961, Phys. Rev. **7**, 444.
STRATTON, J. A., 1941, Electromagnetic Theory (McGraw-Hill, New York) pp. 500–505.
STETSON, K. A., 1967, Laser Focus **3**, No. 3, 25.
STRONG, J., 1936, Phys. Rev. **49**, 291.
TAMIR, T., 1972, Optik **36**, 209.
TAMIR, T. and H. L. BERTONI, 1971, J. Opt. Soc. Am. **61**, 1397.
TAMIR, T. and A. A. OLINER, 1969, J. Opt. Soc. Am. **59**, 942.

TIEN, P. K. and R. ULRICH, 1970, J. Opt. Soc. Am. **60**, 1325.
TIEN, P. K., R. ULRICH and R. J. MARTIN, 1969, Appl. Phys. Letters **14**, 291.
TORALDO DI FRANCIA, G., 1960, Nuovo Cimento **16**, 61.
TWERSKY, V., 1962, J. Opt. Soc. Am. **52**, 145.
VOIGT, W., 1899a, Ann. Physik (3) **67**, 185.
VOIGT, W., 1899b, Ann. Physik (3) **68**, 135.
VOIGT, W., 1911, Ann. Physik (4) **34**, 797 and (4) **36**, 866.
WIEGREFE, A., 1914, Ann. Physik (4) **45**, 465.
WIEGREFE, A., 1916, Ann. Physik (4) **50**, 277.
WOLF, E. and J. R. SHEWELL, 1970, J. Math. Phys. **11**, 2254.
WOLTER, H., 1950, Z. Naturforschg. **5a**, 143.
WOLTER, H., 1961, in: Progress in Optics, Vol. I, ed. E. Wolf (North-Holland, Amsterdam) p. 180.
WOOD, R. W., 1902, Phil. Mag. **4**, 396.
WOOD, R. W., 1913, Phys. Z. **14**, 270.
YOUNG, T. R. and B. D. ROTHROCK, 1961, J. Opt. Soc. Am. **51**, 1038.

This Page Intentionally Left Blank

E. WOLF, PROGRESS IN OPTICS XI © NORTH-HOLLAND 1973

V

PRODUCTION OF ELECTRON PROBES USING A FIELD EMISSION SOURCE *

BY

ALBERT V. CREWE

Departments of Physics and Biophysics and the Enrico Fermi Institute, The University of Chicago, Chicago, Ill. 60637, U.S.A.

* Work supported by the U.S. Atomic Energy Commission.

CONTENTS

§ 1. Introduction – Consideration of Brightness

Electron probes have a large variety of uses ranging from electron beam evaporators and welders to oscilloscopes, microprobes and scanning microscopes. These various uses place a variety of different demands upon the electron optical system and upon the electron source itself. By far the majority of these instruments use a hot tungsten filament as the source of electrons because it is simple and easy to use and because it is a well developed technique. It does, however, have its limitations, and the purpose of this article is to investigate the useful range of applicability of another source of electrons – field emission – which is superior to the hot filament in brightness but inferior in the total current which can be extracted. We will not consider some of the more exotic applications of field emission such as pulsed sources, but will concentrate upon the production of electron probes.

A hot tungsten filament can have dimensions of a few microns and emit many milliamperes of current. A source brightness of 10^5 A/cm^2 sterad can be achieved (HAINE and COSSLETT [1961]). In cases where the final probe is comparable to the source size high probe currents can be obtained. When the final probe size must be much smaller than the source size the probe current becomes much lower. The reason for this is that lenses must be used to demagnify the source and all electron lenses have aberrations. The existence of spherical aberration in particular places an upper limit on the convergence angle of the beam leaving the lens, thereby reducing the probe current. When the demagnification must be very large the probe current may be reduced to unusable levels. A well known example of this is the scanning electron microscope, where a probe radius of 50 Å reduces the probe current to 10^{-12} A or less, with the result that inordinately long exposures are required to produce a micrograph (PEASE and NIXON [1965]).

A field emission source has completely different characteristics. The effective source radius is about 10 Å and currents up to a few hundred μA are attainable. Typical brightness levels range from 5×10^8–10^{10} A/cm^2 sterad (see later discussion).

The limitation on total current means that such sources are not useful when high probe currents are required. The small source size, however, means that little demagnification is required for the production of small probes. As we shall show later, this characteristic, together with the high brightness, means that field emission sources are much superior to hot filament sources when small probes are required.

§ 2. Field Emission Sources

We will not attempt to provide here any detailed discussions of the field emission process. The reader can be referred to several excellent reviews and articles (GOMER [1961]). We give here only the briefest review and will concentrate on the practical features of the process as a source of electrons.

The phenomenon of field emission is essentially a tunneling process. Electrons in a metal are effectively prevented from leaving the surface by the work function ϕ. If the metal has a sharp point of radius r and a voltage V_1 is applied to it, an electric field of value V_1/r exists close to the surface. The shape of this field is such as to allow tunneling through it. The tunneling process was calculated by FOWLER and NORDHEIM [1928], and their expression for the total emission current I as a function of the applied voltage V_1 can be written

$$\log (I/E^2) \propto 1/E.$$

Such a graph is called a Fowler-Nordheim plot and can be used to verify the existence of field emission.

The field strengths required for field emission are rather large – a few volts per Ångstrom. This imposes mechanical stress on the material of the source so that the only suitable materials are those with high tensile strengths. At room temperatures not many materials possess the necessary strength.

Field emission is not uniform over the surface, but depends upon crystal orientation. For a probe-forming system we need emission along the axis, and the (111) and (310) directions of tungsten are suitable.

The state of the surface is also important. A monolayer of gas molecules can have a profound effect on the emission current. A dielectric monolayer has the effect of increasing the potential barrier width by the thickness of the layer. This can reduce the emission current by a factor of 10–100. The time for formation of a monolayer depends upon the ambient pressure, but at 10^{-10} torr it is about 30 minutes. It is clear, then, that field emission

from a clean surface is only practicable in systems having substantially lower pressures than this. In the present state of technology this is difficult, but not impossible. A more attractive alternative is to allow time for the monolayer to form and then operate in that condition. We have found that stable operation for many hours can be obtained at pressures in the range $(1\text{–}3)\times 10^{-10}$ torr.

It is not practicable to operate at pressures much greater than this because the emission rapidly becomes erratic due to the bombardment of the tip by ionized gas molecules.

Practical field emission sources are obtained by etching a fine wire of single crystal tungsten in NaOH until it attains a diameter of about 1000 Å. Heating in vacuum "forms" the end of the wire into a hemisphere. Practical details for the fabrication of such sources are readily available in the literature (CREWE, EGGENBERGER, WALL and WELTER [1968]).

These tips are normally welded to a tungsten hairpin filament. When a pulse of current is applied to this filament it (and the tip) are heated and the occluded gas molecules are removed. No heating current is necessary during normal operation, and the tip can be operated at room temperature.

Our experience with the use of such sources can be summarized as follows.

When operated in the range $(1\text{–}3)\times 10^{-10}$ torr it takes about 30 minutes for the gas monolayer to form. During this time the emission current at fixed voltage drops by a factor of 50–100. Stable emission of 5–10 μA can then be obtained for periods of 10 hours or more. Normal operating voltages are 3–4 kV. The lifetime of such tips is unknown, but certainly exceeds 1000 hours. The energy spread of electrons from the source is about 0.2 V even when operated with the gas monolayer in place (SWANSON and CRONSER [1967]).

The total emission usually occurs in a cone of half-angle about 17°, and the emission is approximately uniform when the tip is in its "dirty" condition.

There remains the question of source size. A rough idea can be obtained by assuming electrons are emitted isotropically with an energy spread V_T ($= 0.2$ V). Conserving transverse momentum we obtain an apparent source radius

$$\Delta = \Delta_1 \sqrt{V_T/V_1}$$

where Δ_1 is the tip radius.

Taking $\Delta_1 = 1000$ Å, $V_1 = 3$kV, we obtain

$$\Delta = 8.2 \text{ Å}.$$

More accurate calculations have been performed by WIESNER [1970], who obtains values less than this. MOLLENSTEDT [1971] obtained an experimental upper limit of 10 Å for the source radius.

In this article we will take $\Delta = 10$ Å as the basis for our calculations. If, in the future, better values are obtained experimentally, the necessary modifications to the ensuing calculations can easily be made. It should be noted that all our calculations of probe current are independent of the assumption of 10 Å for the apparent source radius.

§ 3. An Electron Gun System

Electron guns which have been developed for use with hot filaments are not particularly suited for use with field emission sources owing to the very different characteristics of these sources.

It is clear that a triode system is, in any case, required because field emission requires an applied voltage of about 3 kV, whereas the required probe voltage may range from a few volts to a few hundred kV or more. Therefore two anodes are needed. Each of these must have an aperture to allow the electrons to pass through.

The simplest form of suitable electron gun therefore consists of a field emission tip and two flat anodes, each having a circular aperture. Spacings should be compatible with the various voltages involved, and the tip and the two apertures should be colinear.

Such a simple design has some disadvantages, however. Apertures in anodes act as electrostatic lenses with a focal length which is approximately

$$f = V/\Delta E$$

where V is the electron beam voltage and ΔE is the difference in electric field on the two sides of the aperture. The spherical aberration of these lenses is one or two orders of magnitude greater than the focal length. In addition, because the focusing action is produced by the field change in the neighborhood of the aperture, it must be very carefully made; deviations from symmetry will produce additional aberrations.

It appears that such a gun can be improved by using thick anodes. In this case the electric field near the apertures can be made zero. This can be done by placing the aperture at the apex of a conical hole whose angle is $\arctan \sqrt{2}$. If this is done the problem of mechanical tolerances is transferred to the fabrication of this cone.

If the electric field is zero near the two apertures it must be allowed to change in the space between them so that the electron beam can be

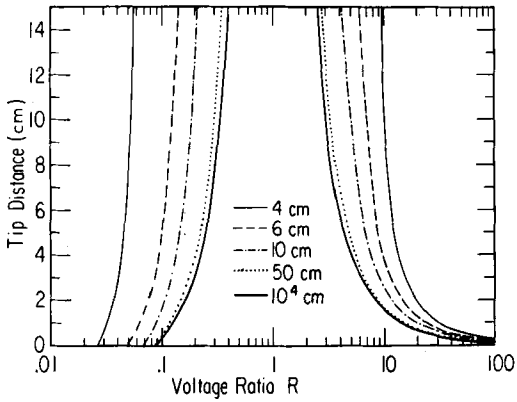


Fig. 1. Results of the computer calculation on the 3 cm electron gun indicating the distance which the field emission tip must be from the first anode in order to obtain a focus 4, 6, 10, 50 and 10^4 cm measured from the first anode. Tip distance is given as a function of the energy ratio $R = V_0/V_1$. Tip distances greater than 15 cm are not indicated because this would lead to an unwieldy system.

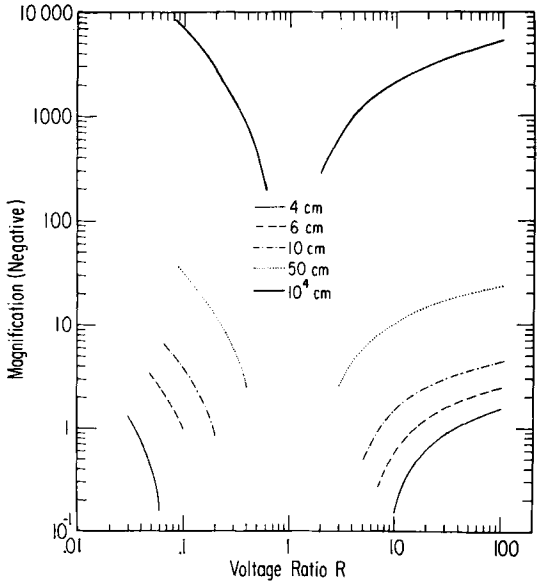


Fig. 2. Magnification of a 3 cm gun for image distances of 4, 6, 10, 50 and 10^4 cm. Only portions of the full curve are shown, but these portions are relevant to tip distances from 0 to 15 cm.

accelerated. This can be accomplished by suitable shaping of the thick anodes. In principle, at least, this can be done in such a way as to minimize the spherical aberration of the system. BUTLER [1966] found that the simplest possible shape, namely a parabolic field shape, produced the lowest aberrations.

Specifically, if the electric field has two zeros, the potential distribution along the axis should contain a cubic term $V = aZ^2 + bZ^3$. Applying the boundary conditions and solving Laplace's equation we can calculate the electrode shape. KOMODA [1971] has extended these calculations by adding fifth order terms in the potential. Some improvement in the coefficient of spherical aberration (about 30 %) was obtained. However, this improvement is a small effect because the spherical aberration constant usually occurs as the one quarter power in any expression for probe size. We will therefore not add this complication.

THOMSON [1971] has used this field shape to calculate first order lens properties and chromatic and spherical aberration coefficients. The results of this computer calculation are shown in Figs. 1 to 4.

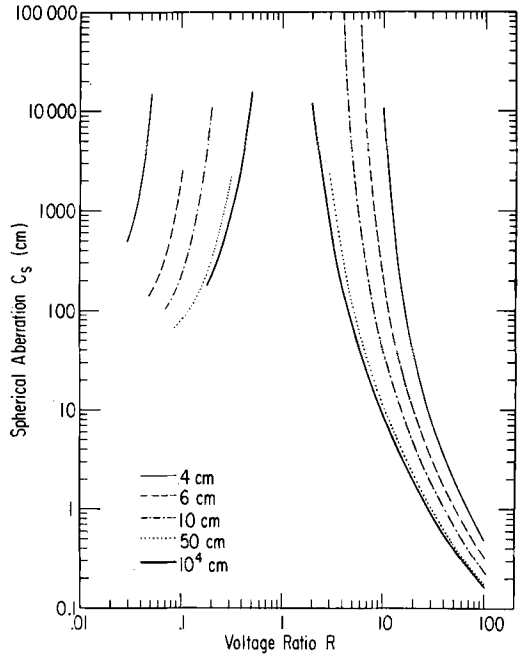


Fig. 3. Spherical aberration of a 3 cm gun for image distances of 4, 6, 10, 50 and 10^4 cm. Portions of the curve shown are for tip distances 0 to 15 cm. The spherical aberration which is shown by these curves is referred back to the source (see text).

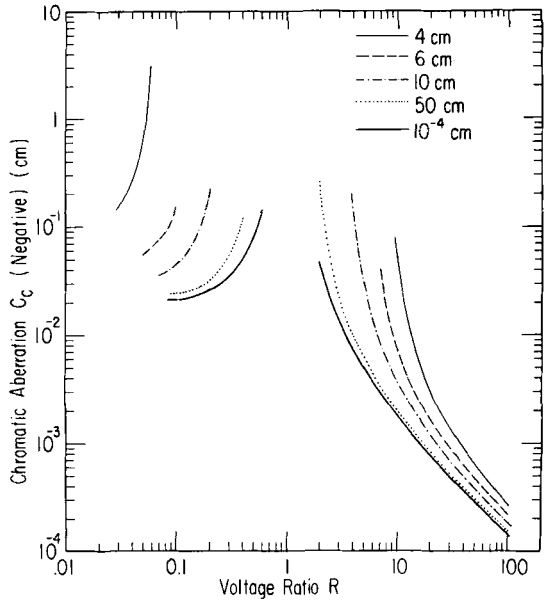


Fig. 4. Chromatic aberration of a 3 cm gun for image distances of 4, 6, 10, 50 and 10^4 cm. Portions of the curve shown are for tip distances 0 to 15 cm. The chromatic aberration which is shown by these curves is referred back to the source (see text).

These results are given for an electron gun which is 3 cm long (measured between the apices of the two cones). Distance measurements are from the first anode and are positive in the direction of the electron beam. Aberrations are referred back to the source. In the case of spherical aberration, the effective source radius is

$$\Delta_s = C_s \alpha^3$$

where C_s is given in the graphs and α is the semi-angle subtended by the defining aperture at the tip.

In the case of chromatic aberration the effective source radius is defined as

$$\Delta_c = C_c \alpha (\Delta V / V_1) \times 1000$$

where ΔV is the energy spread from the source (typically 0.2 V) and V_1 is the voltage on the tip in volts. C_c is given in the graphs.

In the following calculations we shall assume that all power supplies are stable to better than 0.2 V so that the effect of the electron energy spread from the source is the dominant term in the chromatic aberration effect.

All calculations are given in terms of the ratio R where $R = V_0/V_1$.

V_0 is the final voltage of the electron beam. If it is preferred to use angles on the exit side of the gun (α_2) these can be calculated from $\alpha = \alpha_2 M\sqrt{R}$, where M is the magnification of the gun and is also given in the graphs.

We have not shown values for these parameters when the tip is more than 15 cm from the first anode since this would make a clumsy system and also stray magnetic fields, both static and dynamic, would be very troublesome.

We have found these calculations to be extremely reliable. All measurements performed so far agree with the calculations so that in the remainder of this paper we will base all our estimates of probe size on these results.

Several electron guns of this type have been built and tested in a variety of systems (CREWE, EGGENBERGER, WALL and WELTER [1968], CREWE, ISAACSON and JOHNSON [1969], CREWE and WALL [1970], and SAXON [1972]). All such guns have used a 2 cm length rather than the 3 cm indicated in the graphs. We have recently chosen to construct 3 cm guns because this makes the electrical problems somewhat simpler. The curves can easily be scaled to any gun length, however. All distances scale with gun size, but the magnification does not.

§ 4. Types of Imaging Systems to be Considered

The most general probe-focusing system consists of a source of electrons, an acceleration system and a system of lenses. In this way the source can be adjusted to provide the required beam current, the acceleration system will provide the required voltage and the lens will provide the desired probe size.

Each of these components has its limitations so that a careful choice of components must be made to achieve the desired goal, and in some cases it may not be possible to achieve it. We will therefore explore a number of possible systems within this general description in order to assess their range of usefulness.

In order to establish a uniform description we consider the general case shown in Fig. 5. The field emission source has an apparent radius Δ and emits electrons over an angular range α. Some of the electrons are accelerated by the electron gun and in general they will also be focused by the gun. The Gaussian image of Δ is a spot of radius δ_2 and the convergence angle at the focus is α_2. Note that this image can be either real or virtual.

The lens system acts on the electron beam to provide a real probe whose Gaussian image radius is δ_0 and whose convergence angle is α_0.

We will consider only one defining aperture to exist and will imagine it to be located in the electron gun. In practice the physical aperture can be

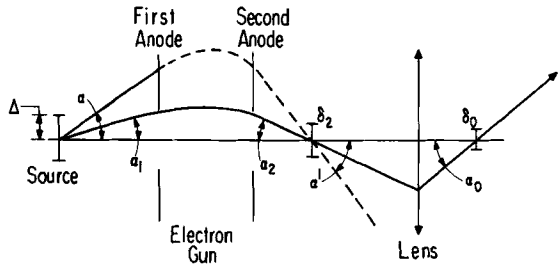


Fig. 5. Definition of the parameters used in this paper. The object and image sizes shown are the apparent object size and the Gaussian images of that object. This diagram indicates the definition of the various angles which enter into the calculation.

in any location and its placement is a matter of convenience.

Aberrations can occur in the electron gun and lens system so that the actual probe radius is not δ_0 but some greater value δ.

The correct method of combining the effects of aberrations with diffraction and Gaussian radius is a matter for debate, but we will adopt a uniform method of taking the R.S.S. value (the square root of the sum of the squares). This may very well cause errors, but not by an order of magnitude. The only exception to this rule which we will use is the combination of spherical aberration and diffraction. This is a well known case in electron microscopy.

The central problem is then to compute the actual spot radius and expected intensity for all the probe-forming systems.

§ 5. General Considerations of Brightness and Intensity

Electrons from the source of radius Δ, and semi-angle of emission α are accelerated and focused by the electron gun. The Gaussian image has a radius δ_2 and in the absence of aberrations the semi-angle of convergence at the image is α' where

$$\delta_2 = M\Delta, \qquad \alpha' = \alpha/M\sqrt{R},$$

where M is the magnification of the accelerating system and R is the ratio of the total acceleration voltage to the voltage on the field emission tip.

If there were no defining apertures in the gun, the intensity at the image would be identical to that at the source

$$I = \beta\pi^2\Delta^2 \sin^2\alpha = \beta\pi^2\delta_2^2\alpha'^2 R.$$

β is the brightness of the source in A/cm^2 sterad. It is considered to be the brightness measured at the tip voltage V_1. We take the approximation

that $\sin\alpha = \alpha$, which is quite accurate, since the value of α for a field emission source is seldom more than 20 degrees.

If there *are* defining apertures in the gun the angle of convergence at the image will not be α' but some smaller value α_2. The value of δ_2 will not be affected so that the intensity in the image will be

$$I_2 = \beta\pi^2\delta_2^2\alpha_2^2 R.$$

The image of the tip produced by the gun is now refocused by the lens system. The Gaussian image produced by the lens system has a radius δ_0 and the semi-angle of convergence at the new image is α_0 and we have

$$\delta_0\alpha_0 = \delta_2\alpha_2$$

by Liouville's theorem. We need not consider any additional defining apertures because there can be only one in a cylindrically symmetric system.

The intensity of the probe is therefore equal to the intensity at the intermediate image of the tip:

$$I_0 = I_2 = \beta\pi^2 R \cdot \delta_0^2\alpha_0^2.$$

This is the total intensity in a spot of radius δ where the value of δ is determined by δ_0 together with the effect of all the aberrations of the system.

We can calculate β from the values already given, that is $\Delta = 10$ Å, $I = 10\ \mu$A, $\alpha = 17°$,

$$\beta \sim 3\times10^8\ \text{A/cm}^2\ \text{sterad}$$
$$\sim 1.8\times10^{11}\ \text{electrons/Å}^2\ \text{sterad sec.}$$

The expression for I_0 can now be written

$$I_0 = R\alpha_0^2\delta_0^2 \times 1.8\times10^{11}\ \text{electrons/sec.}$$

It is clear from the equation above that δ_0 should be as large as possible if high intensity is required. If we assume that all contributions to probe size can be taken in quadrature to calculate the probe radius one can make δ_0 comparable to δ.

α_0 should also be as large as possible, and this corresponds to the use of a lens of the shortest possible focal length for any given application.

In order to proceed further, we will make some assumptions which appear to have general applicability. First of all we will assume that the effect of gun aberrations is small compared to that of the lens aberrations (this is usually, but not always, a very good assumption). Secondly we will assume that the lens is used in its "optimum" configuration. That is, corresponding to the operational conditions of an electron microscope we adjust the

aperture in the lens to give the smallest probe size. This will usually occur when the effects of spherical aberration and diffraction are comparable. This condition has been analyzed by many authors (GRIVET [1965]) and we will take

$$\delta = 0.5 C_{sL}^{\frac{1}{4}} \lambda^{\frac{3}{4}},$$

$$\alpha_0 = 1.4[\lambda / C_{sL}]^{\frac{1}{4}},$$

where C_{sL} is the spherical aberration coefficient of the lens.

It can be readily seen that the product $\delta\alpha_0$ can readily be written solely in terms of λ, the electron wavelength, and independently of the particular lens which is used. If the lens were perfect, the product $\delta\alpha_0$ would be equal to 0.6 λ, the Rayleigh criterion. In the presence of spherical aberration this is modified slightly and becomes 0.7 λ.

With the assumptions we are making, the value of the probe radius is now

$$\delta = \sqrt{\delta_0^2 + (0.7\lambda/\alpha_0)^2}.$$

From this equation we can see that when $\delta_0 \ll 0.7\lambda/\alpha_0$ the intensity will be small but the probe radius will have its smallest value. If $\delta_0 = 0.7\ \lambda/\alpha_0$ the probe radius will increase to approximately $\delta = \lambda/\alpha_0$ and the intensity will be high. This represents a 40 % increase in probe radius and may not be a high price to pay for an increase in intensity.

We call this the "optimum" condition and in this case the intensity in the spot will be

$$I_{opt} = \beta\pi^2 R\lambda^2.$$

Now we take $\lambda = 12.5\ V_0^{-\frac{1}{2}} \times 10^{-8}$ cm and this gives

$$I_{opt} = (1.5\beta/V_1) \times 10^{-13}.$$

β/V_1 is a constant of the system, depending only upon the particular tip which is used. We therefore see that I_0 has a constant value independent of all the other parameters of the system, including the accelerating voltage.

Inserting values ($\beta = 3 \times 10^8$ A/cm^2 sterad, $V_1 = 3$ kV) we obtain

$$I_{opt} = 1.5 \times 10^{-8} \text{A} = 9 \times 10^{10} \text{ electrons/sec.}$$

This estimate provides us with one of the limits of usefulness of a field emission source. It states that a reasonable upper limit to the probe current is 1.5×10^{-8} A. For probe currents substantially greater than this a hot filament source can be used.

It might appear that probe currents could be increased by making δ_0 the dominant term and large, but as we shall see later this does not appear to be possible.

§ 6. Examples of Electron Optic Systems

We now consider various possible electron optic systems starting with the simplest ones. In each case we calculate the range of probe sizes which are possible and the current to be expected in these probes.

6.1. QUASI-PARALLEL BEAM

We consider the possibility of using the gun alone to form a probe at large distances.

The contributions to probe radius are

$$\begin{aligned}
\delta_2 &= 10M && \text{(Gaussian image)}\\
\delta_s &= \tfrac{1}{4}C_s\alpha_2^3 M^4 R^{\frac{3}{2}} \times 10^8 && \text{(Spherical aberration)}\\
\delta_c &= C_c\alpha_2 M^2 R^{\frac{3}{2}}(\Delta V/V_0)\times 10^{11} && \text{(Chromatic aberration}\\
\delta_D &= 0.61\lambda/\alpha_2 && \text{(Diffraction).}
\end{aligned}$$

All values of δ are given in Ångstrom. The parameter $\frac{1}{4}$ appearing in the expression for the spherical aberration term is the value of the beam radius at the narrowest point of the caustic.

We can rewrite these expressions in terms of the radius r of a defining aperture which is placed at the gun exit. In so doing we will also write the expression in terms of the value of M/S since M/S varies only slowly with S:

$$\begin{aligned}
\delta_2 &= 10\times(M/S)S\\
\delta_s &= \tfrac{1}{4}C_s r^3(M/S)^4 R^{\frac{3}{2}} S\times 10^8 = \tfrac{1}{4}Ar^3 S\\
\delta_c &= C_c r(M/S)^2 R^{\frac{3}{2}} S(\Delta V/V_0)\times 10^{11} = B(\Delta V/V_0)rS\\
\delta_D &= 0.61(\lambda S/r) = CS/r.
\end{aligned}$$

We see that the probe radius is proportional to S and the only problem is to optimise r for any given value of R.

Since S is referred to the first anode and r refers to the radius of the beam at the second anode, we can expect some errors to occur when S is small. However, these errors will not be large and will be ignored in subsequent discussions in order to avoid undue complications.

The parameter $A = C_s(M/S)^4 R^{\frac{3}{2}}\times 10^8$ is a new spherical aberration coefficient, $B = C_c(M/S)^2 R^{\frac{3}{2}}\times 10^{11}$ is a chromatic aberration coefficient and C is a diffraction coefficient.

A and B are plotted in Figs. 6 and 7 as a function of R. It can be seen that when $R > 1$ A and B can almost be represented by two universal curves. This is a reasonable behavior since this is the property possessed by thin

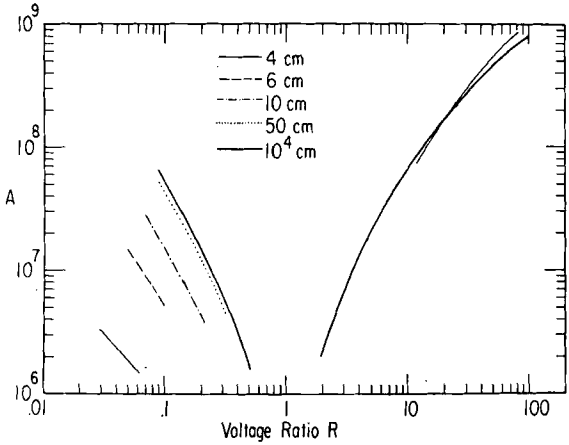


Fig. 6. Parameter A for image distances of 4, 6, 10, 50 and 10^4 cm. For voltages ratios greater than 1 a single curve will almost suffice for the whole range of object distances. The parameter A is a spherical aberration coefficient referred to image space such that the image size due to spherical aberration equals $A \times r^3 \times S$, where r is the radius of the exit aperture of the gun and S is the image distance measured from the first anode. When r and S are in cm, the image radius is in Ångstrom.

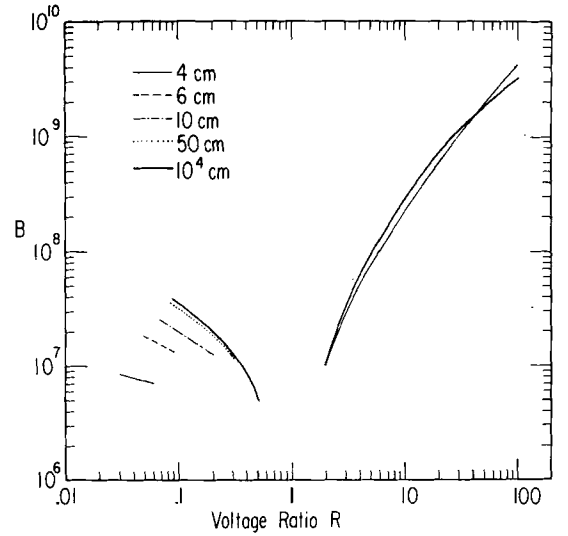


Fig. 7. Parameter B. Chromatic aberration. The chromatic aberration coefficient B is given in image space such that the image size equals $B \times r \times S\Delta V/V_0$. When r and S are in cm the image size is in Ångstrom.

lenses. For values of $R < 1$ no universal curves can be given since the values of A and B are strong functions of S as well as R. Presumably the gun can no longer be considered to be thin.

These curves for A and B are much more convenient in use than using the values of aberration coefficients referred back to the source. We will now estimate some probe sizes and intensities. Table 1 shows actual values of A as a function of R. In this table we have used values of A (and B) for the case $S = 10^4$ cm although as we have previously noted these values are reasonably accurate for any value of S.

TABLE 1

R	M/S	Tip distance (cm)	$\frac{1}{4}A$	C	B_0	δ/S (Å)	$r(\mu)$	I (electrons/sec)
2	0.0302	26.4	5.6×10^5	.098	3.46×10^2	8.6	168	1.1×10^8
4	0.101	5.8	3.4×10^6	.0695	1.1×10^3	11.9	79.4	6.5×10^8
6	0.153	3.12	7.4×10^6	.0568	1.42×10^3	12.8	63.1	1.3×10^9
8	0.192	2.11	1.18×10^7	.0491	1.69×10^3	13.2	53.9	1.9×10^9
10	0.223	1.59	1.654×10^7	.0439	1.86×10^3	13.2	48.6	2.6×10^9
12	0.247	1.27	2.116×10^7	.0401	1.99×10^3	13.0	44.9	3.3×10^9
14	0.269	1.06	2.66×10^7	.0372	2.09×10^3	12.9	42.1	3.9×10^9
16	0.286	0.91	3.15×10^7	.0348	2.16×10^3	12.8	40.1	4.5×10^9
18	0.301	0.79	3.60×10^7	.0328	2.21×10^3	12.6	38.4	5.2×10^9
20	0.315	0.70	4.14×10^7	.0311	2.28×10^3	12.5	37.0	5.7×10^9
24	0.338	0.57	5.14×10^7	.0284	2.31×10^3	12.1	34.9	7.0×10^9
30	0.364	0.44	6.56×10^7	.0254	2.37×10^3	11.8	32.7	8.6×10^9

In order to calculate values for the effect of chromatic aberration we need to know the value of $\Delta V/V_0$. ΔV we have previously given as 0.2 V and our experience shows that V_1 is approximately 3 kV. Therefore we write

$$B_0 = BR^{-1} \times \Delta V/V_1 = 6.67 \times 10^{-5} BR^{-1}$$

and

$$\delta_c = B_0 rS.$$

Values of B_0 are also given in Table 1. We also make the assumption that $V_1 = 3$ kV in calculating C:

$$C = 0.61\lambda = 0.139R^{-\frac{1}{2}}.$$

The values of A, B_0 and C can now be used to calculate probe sizes. From this table it can be seen that spherical aberration and chromatic aberration will be equal when $r = 250\,\mu$ for $R = 2$ and $r = 60\,\mu$ for $R = 30$.

In each case the value of C/r is less than Ar^3 or Br so that the system is not optimised for the smallest probe, because r is too large.

For values of r less than these the effect of spherical aberration becomes very small. We therefore neglect it and optimise the effects of diffraction and chromatic aberration alone.

The optimised value of δ/S and the appropriate value of r are both given in Table 1.

It can be seen that the optimized spot radius is remarkably constant at about 13 S Å. This is to be compared with the Gaussian radius δ_2/S which varies by a factor of 10 for a change of R from 2 to 30 but is always less than 4 S Å.

The intensity can be calculated from the values of the Gaussian image size δ_2/S and the real image size δ/S,

$$I = I_{\text{opt}} \times (\delta_2/\delta)^2 = 9 \times 10^{10} (\delta_2/\delta)^2.$$

It can be seen that the intensities are quite reasonable.

While these values have all been calculated for the case $S = 10^4$ cm it is clear from the curves of A and B that no great error will be made even when going to the case $S = 4$ cm. Therefore Table 1 can be used for the general case of gun use whenever a real image is required outside the gun.

6.2. GUN EXIT FOCUS

The calculations given above cannot be expected to be accurate when the real image is much closer to the gun than 1 cm from the second anode ($S = 4$). In particular, when the image is at the gun exit an aperture cannot be placed there, and the useful pencil of rays must be defined elsewhere, for example at the first anode. This mode of operation generally requires rather high values of R, so if there is any value in such operating conditions it would be for high voltage probes.

TABLE 2

R	Tip distance (cm)	M	C_s(Å)	C_c(Å)	A'	B'	C	δ(Å)	α(mr)	I (electrons/sec)
12	53.0	.022	5.9×10^6	1.15	14.4	770	.0401	12.7	3.64	2.7×10^7
14	10.7	.103	1.38×10^4	5.58×10^{-2}	20.34	148	.0372	13.0	3.27	5.6×10^8
16	5.96	.173	1.75×10^3	1.99×10^{-2}	25.1	159	.0348	13.1	3.06	1.6×10^9
18	4.12	.235	5.18×10^2	1.08×10^{-2}	30.2	169	.0328	13.1	2.87	2.4×10^9
20	3.14	.29	2.22×10^2	7.07×10^{-3}	35.1	176	.0311	13.1	2.73	4.4×10^9
24	2.13	.386	70.6	3.94×10^{-3}	46.1	192	.0284	13.1	2.49	7.8×10^9
30	1.44	.503	24.0	2.27×10^{-3}	63.1	210	.0254	13.0	2.24	1.3×10^{10}

A table of values of parameters for this condition is given in Table 2.

Using these values we can calculate the various contributions to probe radius:

$$\delta_2 = 10\,M$$
$$\delta_C = \tfrac{1}{4}C_s\alpha_2^3 M^4 R^{\frac{1}{2}} \times 10^8 = A'\alpha_2^3$$
$$\delta_S = C_c\alpha_2 M^2 R^{\frac{1}{2}}(\Delta V/V_0) \times 10^{11} = B'\alpha_2$$
$$\delta_D = 0.6\lambda/\alpha_2 = C/\alpha_2 .$$

The values of A', B' and C are given in Table 2 for our usual assumptions of tip operating conditions.

It can easily be seen that the effect of chromatic aberration is always small so that the value of δ can be obtained by minimizing the spherical aberration and diffraction terms. We will do this in this usual way

$$\delta = 0.5C'^{\frac{1}{4}}\lambda^{\frac{3}{4}},$$
$$\alpha_0 = 1.4[\lambda/C']^{\frac{1}{4}}.$$

This value must be combined with δ_2 to estimate the final spot size. The intensity can be calculated knowing δ_2 and δ. The values of δ, α_2, δ_2 and I are given in Table 2.

Again the value of the probe radius is remarkably constant at about 13 Å.

§ 7. Field Emission Source and Electron Gun and Magnetic Lens

The use of a magnetic lens in conjunction with the field emission gun can provide greater flexibility both in probe size and in the location of the focus. We can conveniently study this system by using the results of the previous section.

7.1. QUASI-PARALLEL BEAM

We take the focal length of the magnetic lens to be f and f to be much smaller than the image distance S. We do not consider here the case where an intermediate cross-over exists between the gun and the lens (as in Fig. 5) but the calculations can readily be extended to such a case. One additional parameter would be required to specify this system.

The terms contributing to spot size originating in the electron gun should now be multiplied by f/S so that the values given in Table 1 can be used by substituting f for S. It can be seen that these values will always be small for any reasonable focusing lens. For example if $f = 1$ cm these terms will

contribute 18 Å or less. However, it will be remembered that the values of δ/S in Table 1 were obtained by optimising the effects of chromatic aberration and diffraction. In the present case we have an additional lens in the system which will introduce two new aberration terms – chromatic and spherical aberration. We will first compare these two aberrations with those of the same type which are produced by the gun itself. Let us take, for example, the two spherical aberration terms

$$Ar^3S \times f/S \quad \text{and} \quad C_{sL}\alpha_0^3 \times 10^8$$

from the electron gun and the lens, respectively, where C_{sL} is the spherical aberration coefficient of the lens.

These can be written approximately as

$$Af^4\alpha_0^3 \quad \text{and} \quad C_{sL}\alpha_0^3 \times 10^8.$$

The ratio of the two terms is then

$$(Af^4/C_{sL}) \times 10^{-8}.$$

We can take C_{sL}/f as about 1 so that the ratio becomes

$$Af^3 \times 10^{-8}.$$

Looking at the values of A in Table 1 we can see that for $Af^3 = 10^8$, f must be greater than 1 cm. Therefore the spherical aberration of the lens will predominate over that of the gun for focal lengths shorter than this.

Similarly the chromatic aberration coefficients are in the ratio of

$$(Bf/C_{cL}) \times 10^{-8}$$

where C_{cL} is the chromatic aberration coefficient of the lens. C_{cL} is approximately equal to f so that the ratio becomes $Bf \times 10^{-8}$. Therefore the chromatic aberration effect of the gun almost always predominates, but it is usually small.

We can therefore conveniently classify lenses into short and long focal lengths,

$$f < 1 \text{ cm, resp. } f > 1 \text{ cm.}$$

7.1.1. *Short focal lengths*

In these systems the effect of the lens predominates so that the final probe size is given by the usual equation for the resolution in an electron microscope – to which must be added the contribution from the Gaussian image which will be $\delta_0 = 10 \times (M/S) \times f$.

The intensity in the spot can be calculated from the values of δ_0 and α_2.

As an example we will take the case of a lens of focal length 2 mm and a spherical aberration coefficient of 1 mm operating at $R = 30$ (90 kV). (We will call this our "standard" lens.)

The optimum resolution and convergence angle can be calculated as

$$\delta = 2.56 \text{ Å}, \qquad \alpha_0 = 11.25 \text{ mr}.$$

The calculated value of δ_0 is $\delta_0 = 0.72$ Å from Table 1. From these we can calculate the final probe radius

$$\delta = 2.66 \text{ Å}.$$

The intensity can be most conveniently calculated from

$$I_0 = R\alpha_0^2 \delta_0^2 \times 1.8 \times 10^{11} \text{ electrons/sec}.$$

In this particular case

$$I = 3.5 \times 10^8 \text{ electrons/sec} = 5.8 \times 10^{-11} \text{ A}.$$

While we have not used this exact system, we have used one which is similar. A 2 cm electron gun has been used in conjunction with a lens with a spherical aberration coefficient of about 0.5 mm. The probe size in this case is approximately 3 Å and the intensity in the probe is 3×10^{-11} A (CREWE and RETSKY [1972]).

7.1.2. *Long focal lengths*

In these systems the effect of the gun predominates over that of the lens so that the probe radius will be simply $\delta = (\delta/S) \times f$ and the intensity will be that given in the last column of Table 1. The aperture used should be the one given in Table 1.

This analysis could be carried out for cases where f is not small compared to S or to cases where the focal point for the gun is close to the lens or where there exists an intermediate cross-over. There are, however, simple extensions of the arguments given above.

A special case of an intermediate cross-over is given below.

7.2. GUN EXIT FOCUS

In this case we can distinguish two different cases of the use of an auxiliary lens – magnifying and demagnifying.

7.2.1. *Demagnifying*

Again we can compare the effects of the gun and the lens above.

In the case of spherical aberration, the two terms are

$$A'\alpha_1^3 M_L = A'\alpha_0^3 M_L^4 \quad \text{and} \quad C_{sL}\alpha_0^3 .$$

The ratio of those two is

$$A'M_L^4/C_{sL} .$$

We are considering demagnifying systems so that we expect $\delta < 13$ Å. This requires a short focal length lens with $f < 1$ cm. The physical distance from the gun to the lens can hardly be less than about 5 cm for practical reasons. In this case the effect of the gun will be negligible. Therefore the spot radius and intensity can be calculated in the same manner as the previous section.

As an example we consider the same lens as before ($f = 2$ mm, $C_{sL} = 1$ mm, $R = 30$) and obtain

$$\delta = 2.6 \text{ Å}.$$

With the assumption that the lens is 5 cm from the gun, δ_0 would have a value of 0.2 Å and the intensity would be $I_0 = 2.7 \times 10^7$ electrons/sec.

7.2.2. *Magnifying*

The two spherical aberration terms are

$$A'\alpha_2^3 M_L = A'\alpha_0^3 M_L^4$$

and

$$C_{sL}\alpha_2^3 M_L = C_{sL}\alpha_0^3 M_L^4 .$$

The ratio of these two terms is now

$$A'/C_{sL} .$$

In order to obtain substantial magnification a lens of short focal length would be used and in that case the effect of the gun will predominate. Therefore the spot radius and intensity can be calculated immediately from Table 2. Values of δ should be multiplied by M_L, the magnification of the lens. The values of intensity will be those given in the last column of Table 2.

§ 8. Generalized System

Here we consider a probe-forming system consisting of a field emission source, an accelerating gun and a system of several lenses. Such a system could be used either to magnify the intermediate image of the tip or to demagnify it.

8.1. DEMAGNIFYING SYSTEM

The main purpose of using a series of lenses instead of a single lens would be to increase the flexibility of the system and perhaps also the intensity.

An electron gun and a single demagnifying lens such as we have used already (CREWE and WALL [1970]) is inflexible in that the position of the image of the tip can be uncomfortably close to the focal plane of the lens for some value of the ratio R. If the aperture in the gun is fixed (as in our system) this means that the system can only be operated over a very small range of R. An additional lens in the system can be used as a condenser lens to prepare the electron beam for insertion in the final lens.

The final demagnifying lens would normally be operated in its optimum configuration, and in that case the only effect of a condenser lens on the resolution would be to change the value of δ_0, the Gaussian image contribution to the probe size.

The number of possible configurations of a multiple lens system is very large, and we cannot consider them all here.

The available parameters include the number and disposition of the lenses and their focal properties and the chosen operating condition of the electron gun. There may be no intermediate foci in the system, or there may be many.

We will, therefore, only consider a very general case in order to illustrate the general characteristics. We describe the complete lens system as having an overall magnification M_L. The final probe-forming lens we take to be our standard lens.

The contributions to image size from the gun are

$$\delta_s = Ar^3SM_L$$
$$\delta_c = B(\Delta V/V_0)rSM_L$$
$$\delta_0 = 10(M/S)SM_L.$$

The spherical aberration of the lens and diffraction can be combined to give

$$\delta = 0.5C_{sL}^{\frac{1}{4}}\lambda^{\frac{3}{4}}$$

and

$$\alpha_0 = 1.4(\lambda/C_{sL})^{\frac{1}{4}}$$

as usual.

However, there is a relationship between α_0 and r,

$$r = M_L S\alpha_0$$

when $S \gg 3$ cm, and we now have

$$\delta_s = AS^4M_L^4\alpha_0^3$$
$$\delta_c = B(\Delta V/V_0)S^2M_L^2\alpha_0$$
$$\delta_0 = 10(M/S)SM_L.$$

The intensity of the probe can be calculated from the values of δ_0 and α_0. It can be seen that the design problem reduces to that of choosing an appropriate value of SM_L.

As an example we take our standard lens operating at 90 kV. The ratio of the spherical aberration of the gun to that of the lens at the image is 26 $(SM_L)^4$. The ratio of the effects of the two chromatic aberrations is $50(SM_L)^2$, and the Gaussian image size is 3.64 (SM_L). The intensity of the probe is proportional to $(SM_L)^2$.

The consequences of changing SM_L is as follows. At low values of SM_L, say 0.1–0.2, the size of the probe is determined essentially by the final lens and is 2.56 Å. The intensity will be $(1\text{–}4)\times 10^8$ electrons/sec. For values of $SM_L > 0.14$ the chromatic aberration of the gun becomes greater than that of the lens, but the effect is still small. When SM_L reaches 0.3 the chromatic aberration of the gun will begin to increase the spot size to about 3 Å and at this point the intensity will be 10^9 electrons/sec. When SM_L reaches 0.44 the spherical aberration of the gun will equal that of the lens and the probe size will begin to increase rapidly without a corresponding gain in intensity.

To convert these calculations to a specific design we could use a single condenser lens before the final probe-forming system. In that case, if the electron beam is parallel to the axis between the lenses M_L is equal to the ratio of the focal lengths f_0/f_1. With $f_0 = 2$ mm we therefore conclude that we obtain the smallest probe with the highest intensity when

$$SM_L = 0.1 = (S/f_1)\times 0.2.$$

A reasonable choice would then be $f_1 = 2S$. This corresponds to the proposed operating conditions of a 100 kV scanning microscope which we are now constructing. In this case we chose $f_1 = 5$ cm (CREWE and RETSKY [1972]).

8.2. MAGNIFICATION SYSTEM

A similar analysis can be made for a magnifying system. It would appear from all the previous cases that it would be unreasonable to expect a high probe current from a field emission source and therefore large probe sizes are better obtained by using a hot tungsten filament. However, for probe sizes in the range 100–1000 Å, the use of a field emission source may be superior and estimates of performance can be made using the results of the previous section.

In this case the first lens in the system would have the shortest focal length so that the ideal performance of the system would be

$$\delta = 0.5 C_{sL}^{\frac{1}{4}} \lambda^{\frac{3}{4}} M_L'$$
$$\alpha_0 = 1.4(\lambda/C_{sL})^{\frac{1}{4}}/M_L'$$

where M_L' may be different from M_L if a condenser lens is used before the lens of shortest focal length as if the field of this lens has a condensing effect. Such a system is too complex to be analyzed generally, and each case should be examined individually. The principles are clear, however.

§ 9. Discussion

A field emission source can be used for the production of electron probes ranging in size from a few Å to a few microns, and in energy from a few kV to a few hundred kV. The intensity of such probes can attain values of 10^{10} electrons/sec assuming a total emission from the tip of 10 μA. This is a very conservative value, and there is little doubt that it can easily be increased to a few hundred μA giving a probe current of perhaps 5×10^{-8} A.

Such probes find uses in scanning microscopes and microanalysis. It is hoped that this general survey will be useful to designers of such instruments and they will then be able to improve their performance with respect to probe size, current and working distance.

References

BUTLER, J. W., 1966, Proceedings 6th International Conference for Electron Microscopy (Kyoto, Japan) p. 191.

CREWE, A. V., E. N. EGGENBERGER, J. WALL and L. WELTER, 1968, Rev. Sci. Inst. **39**, 576.

CREWE, A. V., M. ISAACSON and D. JOHNSON, 1969, Rev. Sci. Inst. **40**, 241.

CREWE, A. V. and M. RETSKY, 1972, Proceedings 30th Annual EMSA Conference, Los Angeles, to be published.

CREWE, A. V. and J. WALL, 1970, J. Molec. Biol. **48**, 375.

FOWLER, R. H. and L. NORDHEIM, 1928, Proc. Roy. Soc. London **A19**, 173.

GOMER, R., 1961, Field Emission and Field Ionization (Harvard University Press, Cambridge).

GRIVET, P., 1965, Electron Optics (Pergamon Press, London) p. 421.

HAINE, M. E. and V. E. COSSLETT, 1961, The Electron Microscope (E. & F. Spon, London) p. 125.

KOMODA, T., 1971, private communication.

MOLLENSTEDT, G., 1971, Proceedings International Microprobe Conference, Osaka, Japan.

PEASE, R. F. W. and W. C. NIXON, 1965, J. Sci. Inst. **42**, 81.

SAXON, J., 1972, Optik, to be published.

SWANSON, L. W. and L. C. CRONSER, 1967, Phys. Rev. **163**, 622.

THOMSON, M. G. R., 1971, private communication.

WIESNER, C., 1970, Ph.D. Thesis, *University of California, Berkeley, California.*

E. WOLF, PROGRESS IN OPTICS XI © NORTH-HOLLAND 1973

VI

HAMILTONIAN THEORY OF BEAM MODE PROPAGATION

BY

J. A. ARNAUD

Bell Telephone Laboratories, Incorporated,
Crawford Hill Laboratory, Holmdel, N. J. 07733, U.S.A.

CONTENTS

§ 1. Introduction

Theoretical developments in the study of the propagation of optical beams have been stimulated during the last decade by the invention of coherent sources of electromagnetic radiation of short wavelength and the growing need for communication systems of high capacity. These theoretical developments have been reviewed in a few excellent papers such as those authored by KOGELNIK and LI [1966], DI FRANCIA [1966], KOPPELMAN [1967], GOUBAU [1969] and HARVEY [1970]. There is therefore little need for reviewing again in detail the best-known aspects of Beam Optics. However, the close relationship existing between Beam Optics and other parts of physics, especially Wave Mechanics, has perhaps not been sufficiently emphasized in the past. For that reason the analogy existing between Scalar Wave Optics and Wave Mechanics will be discussed in a section of this paper. We shall also point out that various observations concerning the representation of Gaussian beams by ray packets are best understood in the framework of the Geometrical Theory of Diffraction. Consideration will be given to optical systems that have not been discussed in previous reviews, such as optical systems lacking meridional planes of symmetry and optical systems incorporating anisotropic materials. The main conclusion of this paper is that the laws of propagation of beam modes are most easily obtained from an extension to the complex plane of the methods of Hamiltonian optics.

Let us now motivate our interest in optical beams, i.e., in electromagnetic radiations having small angular divergences, and explain under what condition optical beams can be described by "beam modes". Conventional metallic waveguides whose transverse dimensions are of the order of a wavelength can hardly be used in optics because of their high losses and small size. It was therefore proposed in the 1960 to give consideration to guiding systems whose transverse dimensions are large compared with the wavelength. When optical beams are launched into free space they expand because of diffraction. However, if their transverse dimensions are of the order of, say, 1000 wavelengths (roughly 1 mm in the optical range), the divergence angle is only 10^{-3} radians. This small divergence angle indicates

that the force of diffraction is small and can be balanced by a relatively weak focusing force. This focusing force can be applied either continuously along the system axis as in the case of glass fibers with graded index of refraction, or in a discrete manner as in the case of optical waveguides incorporating conventional lenses. In the former case a small reduction in refractive index off-axis ensures beam confinement. For example, a change in refractive index of 10^{-6} is sufficient in principle to keep confined an optical beam with a diameter of 1 mm. The requirements to be met concerning the guide uniformness and straightness, unfortunately, are much more stringent than in the case of conventional waveguides.

Some important theoretical aspects of optical beam propagation have been known for a long time. Indeed, except for a change in scale, the laws of propagation of optical beams in continuously focusing media are the same as the laws of propagation of radio waves in atmospheric ducts, a problem which received considerable attention during the first half of this century (see, for instance, BOOKER and WALKINSHAW [1946]). The concept of trapped modes and leaky modes, and the quasigeometrical optics (J.W.-K.B.) methods which were introduced in these early investigations remain quite valuable.

As indicated before, we shall restrict our attention to beams that have small angular divergences and to media whose refractive index does not vary much in transverse directions. It turns out that in such cases the vector wave equation obeyed by the electromagnetic field (Maxwell's equations) can be reduced to a scalar wave equation. This does not mean that all polarization effects can be ignored, but it means that such effects can be considered independently of the focusing properties of the medium. Because of this simplification a close relationship can be found between the laws of propagation of optical beams and the laws of propagation of scalar waves such as sound waves and waves associated with material particles. As FEYNMAN [1948] points it out, Wave Mechanics "is easily interpreted physically as the expression of Huygens principle for matter waves". A comparison between Scalar Optics and Wave Mechanics at an elementary level can be found in GOUDET and MEULEAU [1957].

A great variety of approaches have been considered in the 1960's to study the laws of propagation of optical beams in open resonators, sequences of lenses, or lens-like media. KELLER and RUBINOW [1960] generalized the quasi-geometrical optics approach previously mentioned to systems with more than one dimension. Other investigators (GOUBAU and SCHWERING [1961], MARCATILI [1964], KURTZ and STREIFER [1969]) used Maxwell's equations and introduced suitable approximations in the course of their

calculations. PIERCE [1961] made use of the scalar Helmholtz equation that each cartesian component of the field obeys approximately. VAINSHTEIN [1965], [1966] and KOGELNIK [1965] based their calculations on the parabolic approximation of Helmholtz equation introduced by LEONTOVICH and FOCK [1946]. In its integral form, the parabolic wave equation is equivalent to the Fresnel approximation of the Kirchhoff integral, a formalism used by FOX and LI [1960] and BOYD and GORDON [1961]. More recently, quasi-classical (Hamiltonian) approaches have been proposed that we shall discuss in detail. Within the approximations considered in this paper, these various methods lead to the same conclusions. They suggest a generalization of the concept of mode that we shall now explain.

Let us consider a uniform fiber whose refractive index decreases as a function of the distance from axis according to some law $n(x) \leqq n(0)$ up to a distance a beyond which n remains constant. Optical beams with wavelength λ can be trapped in such a fiber if the change in refractive index is sufficiently large (roughly, if $[n(0)-n(a)]^{\frac{1}{2}}a \gtrsim \lambda$). Assuming that this condition is met, fields can be found that have invariant irradiance patterns. Such field configurations are called stationary states, or modes. These stationary states are real and the wavefronts associated with them are plane and perpendicular to the system axis. Note incidentally that because the duct is finite in depth, the number of modes is also finite. The field of the modes, numbered according to the number of nodes in a transverse plane, extends to a distance which usually increases with the mode number.

When the refractive index law $n(x)$ is not independent of the axial coordinate z as assumed before, but does not vary much over a period of oscillation of the rays, the modes are slowly varying functions of z. If a beam in one mode is launched at the end of a slightly nonuniform fiber, the number of nodes of the field (i.e., the mode number) remains the same along the fiber. This law of invariance of the mode number unfortunately is not applicable in the case of fast changes in the axial direction; in particular, it is not applicable to a sequence of conventional lenses. The field transformation becomes complicated in that case and the concept of mode loses its usefulness.

The situation simplifies considerably when the expansion of the square of the refractive index in power series of the transverse coordinates x_1, x_2 can be limited to the first-order terms, i.e., when $n^2(x_1,x_2) = a+bx_1+cx_2+dx_1^2+ex_1x_2+fx_2^2$, where a, b, c, d, e, f are arbitrary functions of z. In analogy with the "weak potentials" considered in Mechanics such media will be called *weak media.* Clearly, a medium can be weak only within a finite domain of the x_1, x_2 plane because n has upper and lower

bounds. Explicit calculations need to be made to see whether the field irradiance outside this domain is small enough to be neglected. This is not always easy to decide because some effects (aberration, radiation losses, mode conversion) can be small in absolute value and yet be significant over long distances. We shall ignore these difficulties and consider only ideally weak media.

In uniform weak fibers the coefficients a, b, c, d, e, f are constant. It can be shown that in that case the stationary states (modes) are described by the product of a function of Gauss and a Hermite polynomial whose arguments are real functions of x_1, x_2. (In the case of systems with rotational symmetry the Hermite polynomials can be replaced by Laguerre polynomials.) The fundamental property of weak media is that, even if a, b, c, d, e, f

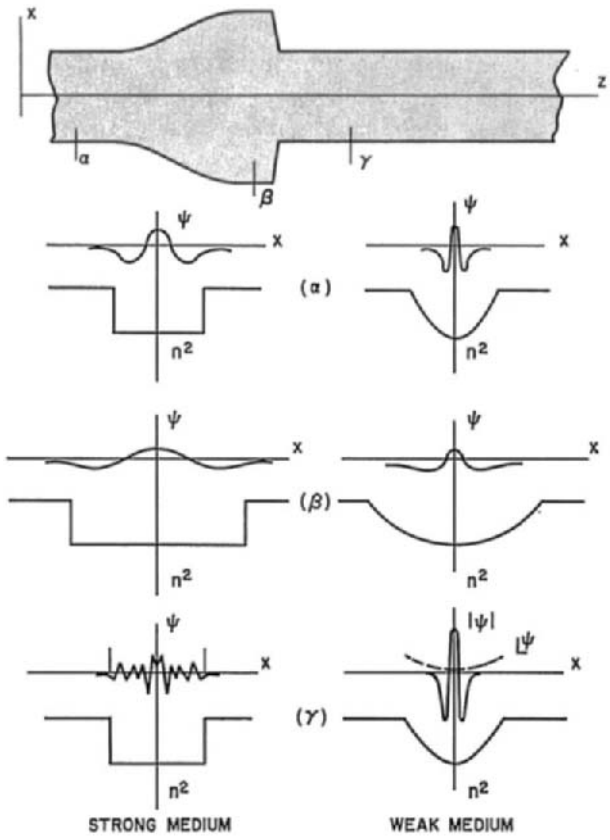


Fig. 1. Field transformation in aberrated media (left) and square-law media (right) for the case of adiabatic (β) and nonadiabatic (γ) changes in the refractive index law along the fiber axis.

are rapidly varying, perhaps discontinuous, functions of z, beams can be defined whose transverse field distribution is described in any transverse plane by the product of a function of Gauss and a Hermite polynomial. However the arguments of these functions are in general complex functions of x_1, x_2 and the wavefronts are not plane as before. To distinguish these field configurations from the stationary (or quasi-stationary) states previously defined, the former have been called *beam modes* by GOUBAU and SCHWERING [1961]. Figure 1 illustrates the difference in behavior of optical beams propagating in weak (square-law) and strong (aberrated) media for the case of adiabatic and non-adiabatic changes in refractive index. The transformation of the field of a mode of order 2 in the case of strong media is illustrated on the left of Fig. 1. Non-adiabatic changes cause some of the power of the launched mode to be transferred to other modes. The transformation of the field in the case of weak media is illustrated on the right of Fig. 1. In that case, the general form of the field amplitude $|\psi|$ is preserved. The phase $\underline{|\psi}$ of the field, however, does not remain a constant.

In the special case where the fiber is uniform the irradiance pattern of beam modes varies periodically (Fig. 2a) the period of oscillation being equal to half the period of ray oscillation. For properly chosen initial conditions the irradiance patterns are invariant; in that case beam modes

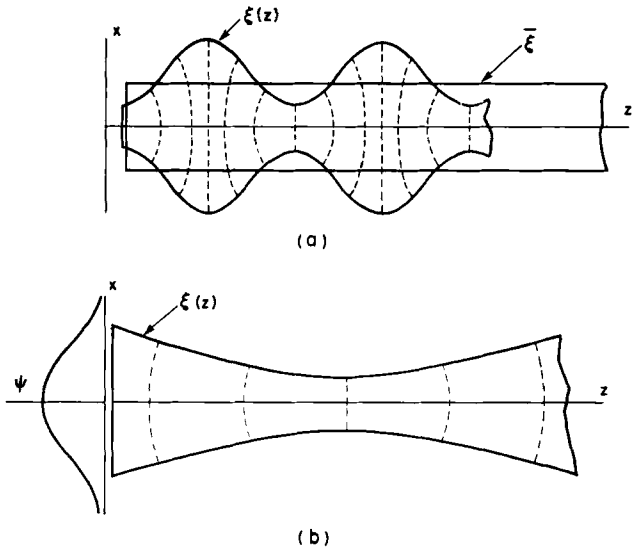


Fig. 2. (a) Mode (straight profile) and beam mode (ondulating profile) in square-law media. (b) In free space, a beam mode reaches its minimum size, the so-called beam waist, only once.

coincide with ordinary modes. In free space, the period of oscillation of the irradiance pattern of a beam mode is infinite. Thus, the beam reaches a minimum width, its so-called *beam waist*, only once (see Fig. 2b). Note that ordinary modes do not exist in free space, because the field of plane waves do not belong to the space of magnitude square integrable functions that we are implicitly considering. The concept of beam mode will be clarified further by examples given in subsequent sections.

Sections 2 to 8 are aimed at giving an overall view of the subject matter. Most results are given there without proof. Sections 9 to 12 are more formal; they are aimed at giving the background material needed to derive the results stated in preceding sections.

§ 2. Geometrical Optics Fields

The purpose of this section is to recall a few elementary results of Geometrical Optics and to indicate how fields can be associated with ray manifolds. DESCARTES [1637] observed that the refraction of optical rays at plane interfaces can be understood by analogy with the behavior of material particles ("L'action de la lumière suit les mêmes lois que le mouvement de cette balle. . ."). Descartes assumes that the particle velocity v is equal to the refractive index n of the medium ($v = n$), and notes that, when a particle encounters a plane surface, only the normal component of the velocity is affected. Equivalently, we can say that rays follow the same trajectories as material particles of zero total energy in a gravitational potential $-\frac{1}{2}n^2$ (see for instance LUNEBURG [1944] p. 84). Because the acceleration experienced by (non-relativistic) particles is the negative of the gradient of the potential, light rays $\mathbf{r}(t)$ obey the equation

$$\mathrm{d}^2\mathbf{r}/\mathrm{d}t^2 = \nabla(\tfrac{1}{2}n^2) \tag{2.1}$$

where $\mathrm{d}l/\mathrm{d}t \equiv n$, l being the arc length measured along the ray. From this equation Fermat's principle can be obtained. This principle states that the variation of

$$S(\mathbf{r}; \mathbf{r}') \equiv \int_{\mathbf{r}'}^{\mathbf{r}} n \, \mathrm{d}l \tag{2.2}$$

is at most of second order with respect to small departures of the path of integration from a ray path. The stationary value of S is called the *point-eikonal*. Fermat's principle is equivalent to Maupertuis principle applicable to particles of constant energy (see for instance LANDAU and LIFSHITZ [1960] p. 140) with the substitution $n \to v$.

This mechanical analogy has been formalized by HAMILTON [1835] who introduced an equation of the form

$$H(\nabla S, \mathsf{r}) = 0. \tag{2.3}$$

Eq. (2.3) is called the *eikonal equation* in Geometrical Optics. ∇S denotes the gradient of the eikonal S at some point r in space; it describes the local direction of propagation of the waves. In Mechanics, ∇S is to be interpreted as the four-dimensional gradient of the *action* at some space-time point; it is related to the energy and momentum of the particle. For some problems, it proves convenient to solve (2.3) for a particular component of ∇S, namely $\partial S/\partial z$ in Optics and $\partial S/\partial t$ in Mechanics. The solution is then assumed to be unique.

Equation (2.3) is a partial differential equation for S whose solutions $S(\mathsf{r}) =$ constant can be viewed as wavefronts. Thus, as SYNGE [1954] points it out, eq. (2.3) is really a wave equation, although the wavelength remains unspecified. Specification of the wavelength: λ/n in Optics, h/v in Mechanics (h denotes Planck's constant; the mass of the particle is assumed to be unity) results in a quantification of the ray trajectories which is now discussed.

One often makes use in Geometrical Optics of manifolds of rays called *normal congruences*. Ray congruences are defined by the condition that no more than one ray passes through any point in space. These congruences are called normal if a surface $S(\mathsf{r}) =$ constant exists which is perpendicular to all of the rays of the manifold. The theorem of Malus and Dupin grants that a normal congruence remains normal after an arbitrary number of refractions (see BORN and WOLF [1965] p. 130). Geometrical Optics fields can be associated with such manifolds, the phase of the field being obtained by integrating $n\mathrm{d}l$ along a ray from some reference surface, and the field amplitude being obtained by specifying that the power flowing in a narrow tube of rays be invariant.

It often happens that the rays of the manifold are tangent to a curve called a *caustic*. (For simplicity, we now restrict ourselves to the two-dimensional case, exemplified in Fig. 3a.) Two neighboring rays intersect each other near the caustic. If ε denotes the distance between two such rays, the field amplitude must be proportional to $\varepsilon^{-\frac{1}{2}}$ to satisfy the law of conservation of power. Thus, the field amplitude is infinite at the caustic where ε vanishes. We further observe that ε is an algebraic quantity which changes sign at the caustic. Because of the exponent $\frac{1}{2}$ in the expression of the field amplitude, the change in sign of ε results in a phaseshift equal to $\frac{1}{2}\pi$. (This phaseshift turns out to be a phase *retardation*.) The difficulty that we now encounter is that two rays instead of one are passing through any

given point. To overcome this difficulty KELLER [1958] suggests that we view the plane of the figure as double sided and assume that at the caustic the rays actually change side, i.e., pass *behind* the plane. Rays are represented in Fig. 3b by dotted lines when they are located behind the plane

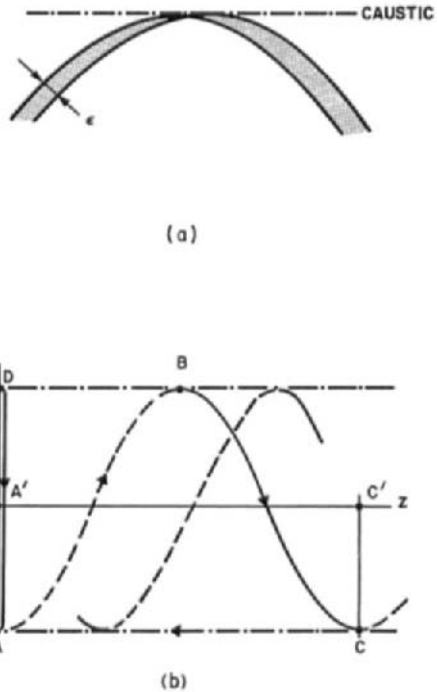


Fig. 3. (a) This figure illustrates the fact that two neighboring rays meet at the caustic when such a caustic is present. (b) Ray congruences in square-law media.

and by plain lines when they are in front of it. The manifold of rays is now a normal congruence with respect to the double-sided surface.

2.1. LENS-LIKE MEDIA

Let us consider as an example the case of a square law medium with refractive index

$$n(x) = 1 - \tfrac{1}{2}\Omega^2 x^2, \tag{2.4}$$

where Ω is a constant. Equation (2.1) shows that rays $\bar{q}(z)$ obey the equation

$$\ddot{\bar{q}} + \Omega^2 \bar{q} = 0 \tag{2.5}$$

for small values of $\bar{q}$ (dots denote differentiation with respect to z). The

solution of (2.5) is

$$\bar{q}(z) = \bar{\xi} \cos(\Omega z + \theta), \tag{2.6}$$

where $\bar{\xi}$ and θ are constants of integration. Consider now the manifold of rays obtained by letting θ assume all values from 0 to 2π. According to Keller's representation, we view the strip $-\bar{\xi} < x < \bar{\xi}$, bounded by the two caustic lines $x = \bar{\xi}$, $x = -\bar{\xi}$, as double sided. The variation of the eikonal between two points such as A and C (see Fig. 3b) can be evaluated along two independent paths. Fermat's principle readily shows that the optical length of the ray ABC is equal to the corresponding length $A'C' = 2\pi/\Omega$ on axis. Consider next the path AC which coincides with the caustic line. Because each elementary section of the caustic coincides with a ray, the variation of the eikonal along the caustic line is $(2\pi/\Omega)n(\bar{\xi}) = (2\pi/\Omega)\times(1-\frac{1}{2}\Omega^2\bar{\xi}^2)$. Thus, the difference $\overline{ABC}-\overline{AC}$ is equal to $\pi\Omega\bar{\xi}^2$. To evaluate the total phaseshift along the closed path $\overline{ABCA}$ we must take into account the fact that two caustic lines are crossed and introduce a phase retardation equal to π. Thus the eikonal is single valued if $k\pi\Omega\bar{\xi}^2-\pi = 2m\pi$, where $k \equiv 2\pi/\lambda$ and m denotes an integer, i.e., if

$$\tfrac{1}{2}k\Omega\bar{\xi}^2 = m+\tfrac{1}{2}. \tag{2.7}$$

Only discrete values $\bar{\xi}_m$ of $\bar{\xi}$ are therefore permissible, those corresponding to $m = 0, 1, 2\ldots$ in (2.7). For these values of $\bar{\xi}$, a geometrical optics field can be associated to the ray manifold by adding the contributions of the two congruences shown by plain lines and dotted lines, respectively, in Fig. 3b. The field is easily found to be

$$e_m(x, z) = (1-\chi^2)^{-\frac{1}{4}} \cos\{(m+\tfrac{1}{2})[\arccos(\chi)-\chi(1-\chi^2)^{\frac{1}{2}}]-\tfrac{1}{4}\pi\} \times \exp\{i[k-(m+\tfrac{1}{2})\Omega]z\} \tag{2.8}$$

where $\chi \equiv x/\bar{\xi}_m$. The function $e(\chi)$ is represented by a dotted line in Fig. 4 for $m = 6$. $e(\chi)$ resembles the exact mode field [given in (3.1)] shown by plain lines, except for the fact that, unlike the exact field, $e(\chi)$ tends to infinity as the caustic line is approached. This defect is remedied by the J.W.K.B. method. With the help of this method the solutions given by (2.8) on either side of the caustic ($|x| < \bar{\xi}_m$ and $|x| > \bar{\xi}_m$) can be matched smoothly to one another. The J.W.K.B. solution, which involves Airy functions, is obtained by assuming that the variation of the refractive index is linear in the neighborhood of the caustics. For $m = 0$ the geometrical optics field (2.8) departs sharply from the exact modal solution; however it turns out that the caustic $\chi = \bar{\xi}_0$ coincides with the beam profile (to be defined in Section 3). Furthermore, the expression for the propagation constant of a mode of order m given in (2.8) turns out to be exact for any m.

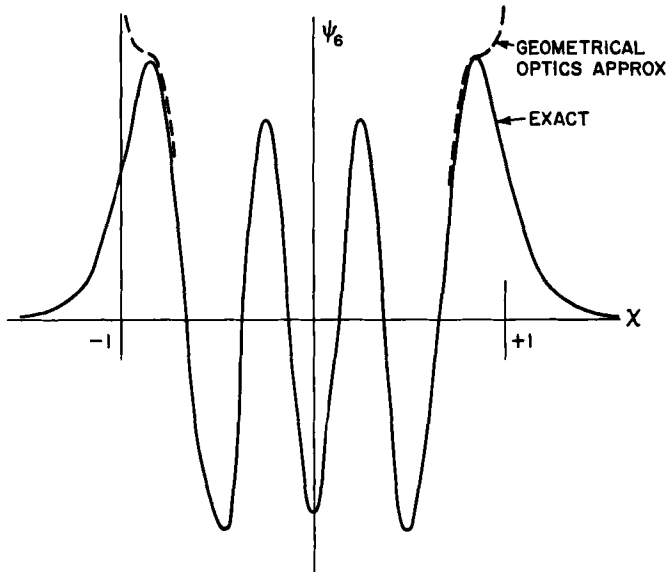


Fig. 4. Comparison between the exact solution for the field $\psi \equiv e$ of a mode with mode number 6 (plain line) and the geometrical optics approximation (dotted line). The latter departs from the former only in the close neighborhood of the caustics $\chi = \pm 1$.

Note that, instead of considering the closed path $\overline{\text{ABCA}}$, we could have considered as well the path $\overline{\text{ADA}}$, shown in Fig. 3b, taken at a fixed value of z. Setting $p \equiv \partial S/\partial x$, the variation of S along this path is

$$\int_{\overline{\text{ADA}}} p \, \mathrm{d}q = (m + \tfrac{1}{2})\lambda. \tag{2.9}$$

Equation (2.9), with λ changed to h, is known in Quantum Mechanics as the quantum condition of Bohr-Sommerfeld.

To evaluate the left hand side of (2.9) it is convenient to represent the position of a ray at some plane z in phase space with coordinates p, q. In the paraxial approximation, and for n close to unity, p coincides with the slope $\dot{q}$ of the ray q, whose maximum value is $\Omega\xi$. For the case of square-law media presently discussed it proves convenient to use the coordinate system $\Omega^{\frac{1}{2}}q$ and $\Omega^{-\frac{1}{2}}p$ instead of q and p. The rays of the ray manifold representing a mode m are located, in phase space, on a circle with radius $\Omega^{\frac{1}{2}}\xi_m$ (see Fig. 5). The left hand side of (2.9) is equal to the area $\pi\Omega\xi_m^2$ enclosed by that circle. Thus (2.9) leads again to the condition (2.7). The rays of the ray manifold which describes a beam mode, on the other hand, are located on an ellipse rather than on a circle. Because this ellipse rotates

at a uniform rate Ω as z varies, the beam width varies as illustrated in Fig. 2a.

The phase-space representation just described is particularly useful when Ω is a slowly varying function of z. It can be shown that the area covered by any continuous set of rays remains the same as z varies. This result, known in Optics as the law of invariance of the luminance (see BROUWER and WALTHER [1967]) and, in Mechanics, as Liouville's theorem (see LANDAU and LIFSHITZ [1960] p. 147), is a consequence of the existence of eikonal functions. Thus, as Ω varies, the ray manifold expands or contracts but $\Omega\bar{\xi}_m^2$ remains the same. This is the law of adiabatic invariance for rays.

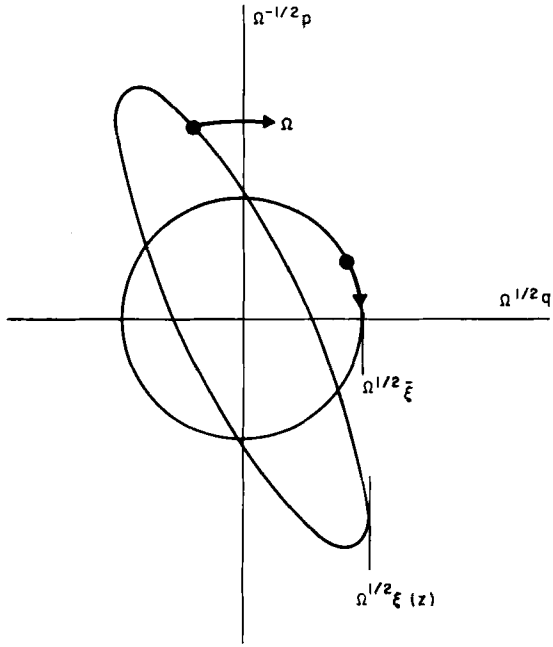


Fig. 5. Representation in phase-space of the ray congruences associated with modes (circle) and beam modes (ellipse).

According to (2.7), the constancy of $\Omega\bar{\xi}_m^2$ for slow variations of Ω implies that if a beam is at some plane z in a single mode m it will remain in that same mode as Ω varies (see for instance MILDER [1969]). If Ω varies rapidly with z, mode conversion does occur. However it turns out that, as we have indicated in the introduction, the general form of the field remains the same (e.g., Gaussian). What is happening is that the beam wavefront does not remain plane and its half-width $\bar{\xi}_m$ ceases to follow the law of adiabatic invariance.

2.2. RESONATORS

Let us now consider closed systems called resonators. RAYLEIGH [1894] noted that in a large room sound waves tend to follow closed curves, either clinging to the boundaries (whispering gallery modes), or bouncing back and forth between opposite walls (bouncing ball modes). KELLER and RUBINOW [1960] considered the case of perfectly reflecting elliptical boundaries, for which exact solutions are known. Using the Geometrical Theory of Diffraction (or asymptotic forms of the exact solution), they obtained

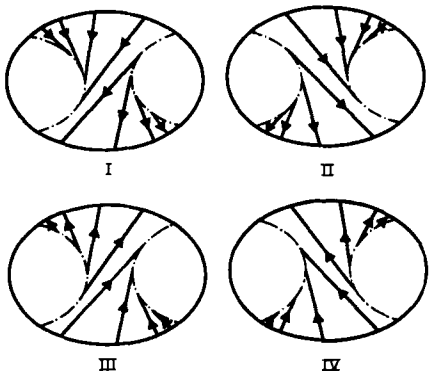


Fig. 6. Ray manifolds associated with bouncing ball modes in resonators with elliptical boundaries. The normal congruences shown in I, II, III and IV are to be viewed as superimposed and "sewed" together at the boundaries (after KELLER and RUBINOW [1960], by permission).

an approximate expression for the resonant frequencies associated with beams bouncing back and forth along the small axis of the ellipse (see Fig. 6) in the form

$$k_{lm}L = 2l\pi + 8(m+\tfrac{1}{2})[\arctan(\exp R_0) - \tfrac{1}{4}\pi], \tag{2.10}$$

where $\frac{1}{2}L$ denotes the length of the ellipse minor axis and $\tanh(R_0)$ denotes the minor to major axis length ratio. k_{lm} is the free-space propagation constant associated with a mode with axial number l and transverse number m. If we introduce the radius of curvature R of the ellipse at the intersection with the small axis, (2.10) becomes

$$k_{lm}L = 2l\pi + 2(m+\tfrac{1}{2})\arccos(1 - L/2R), \tag{2.11}$$

an expression which coincides with a result obtained independently by BOYD and KOGELNIK [1962] for the resonance of an open cavity incorporating two circular mirrors of radius R facing each other. This agreement

proves that within the paraxial approximation no distinction should be made between circular and elliptic mirrors having the same curvature.

The Geometrical Theory of modes shows that any ray launched in an optical cavity generates, after a large number of round trips, a double congruence of rays, bounded by two caustics. These caustics outline the beam profile. This was noted by BYKOV and VAINSHTEIN [1965] and KAHN [1965]. Similar remarks have been applied to the case of square-law media by KURAUCHI and KAHN [1966]. STEIER [1966] observed that the bisectrix of two rays of the ray congruence intersects the axis at the wavefront center.

It should be noted that if a ray happens to retrace its own path after a number N of round trips, this ray can be chosen as the closed path defining the quantum condition (2.9). The resonator is then said to be N-fold degenerate. This point of view, first suggested by DI FRANCIA [1964], has been developed recently by RAMSAY and DEGNAN [1970].

According to the laws of conventional geometrical optics, the optical field vanishes beyond the caustics. By giving consideration to complex rays, KELLER [1958] was able to evaluate the penetration of the field beyond these classical limits, thereby taking into consideration diffraction effects. Keller's theory suffers of the defect that the field obtained is infinite at the caustic. A number of refinements have been proposed to "flesh a geometrical optics skeleton with diffraction substance", as KRAVTSOV [1967] vividly pictures it, while avoiding the singularities at the caustics. One approach consists of assuming that the refractive index varies linearly in the neighborhood of the caustic and in matching the solutions obtained in various regions; this is an extension to three dimensions of the J.W.K.B. method alluded to before. KRAVKSOV [1965] and LUDWIG [1966] succeeded in making the asymptotic expansion of the field uniform in the neighborhood of caustics. These modifications of the Geometrical Theory of Diffraction will not be discussed further in this paper.

§ 3. Wave Optics and Wave Mechanics

As is well known, Geometrical Optics gives an adequate description of most optical phenomena when the wavelength, λ, is very small. Similarly, Classical Mechanics adequately describes mechanical phenomena when the mass of the particles under consideration is very large or, equivalently, if we let the Planck constant, h, tend to zero. Because of the analogy existing between Geometrical Optics and Classical Mechanics we expect that the procedure commonly used to go from the equations of Classical Mechanics to those of Wave Mechanics, which consists of replacing in (2.3) $ik\nabla S$ by

the differential operator $\mathbf{V}$, is also applicable to the optical problem. Comparison with exact scalar wave equations derived from Maxwell's equations shows that this is indeed the case provided the medium is homogeneous or weakly inhomogeneous (see Section 9).

The theory of beam propagation in weak (square-law) media effectively coincides with the quantum theory of harmonic oscillators as noted by BUDDEN [1961] p. 198. (See also GORDON [1966] and the recent contributions of POPOV [1968], MILDER [1969], GLOGE and MARCUSE [1969], EICHMANN [1971], FELDMANN [1971] and BERGMANN [1972].) To exemplify this relationship, let us consider an arrangement of practical significance. Glass fibers with quadratic refractive index profiles are presently under active consideration for long-distance optical communication. In the arrangement shown in Fig. 7, an optical beam generated by a laser is launched at one

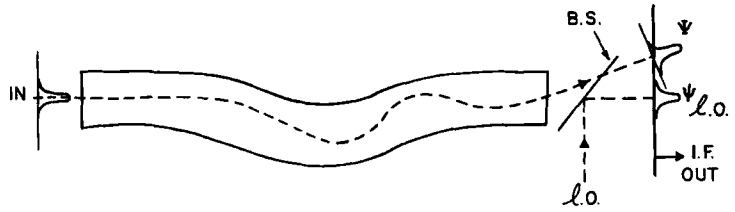


Fig. 7. Beam deflection in bent square-law fiber. The beam at the right end of the fiber is detected by optical heterodyning.

end of the fiber and detected at the other end by a heterodyne optical receiver. To discuss the transmission properties of this system we shall make use of formulas borrowed from GOL'DMANN and KRIVCHENKOV [1957], referred to in the future as GK. Only trivial changes of notation are required: In principle, we should set $dt = dz/n$ and $\psi = n^{\frac{1}{2}}e$, ψ being the wave function and e the electric field; for simplicity, we assume that the on-axis refractive index is close to unity; no distinction need therefore be made between t and z, and between ψ and e. We assume that the mass of the particle is unity, replace the Planck constant h by the wavelength λ, and take into account the geometrical-optics phaseshift kz suffered by the optical field.

3.1. FIBER TRANSMISSON

Let us consider a uniform square-law fiber, whose refractive index is given in (2.4). The mechanical analog of this medium is an undamped harmonic oscillator with spring constant Ω^2. The modes (stationary states) are (GK p. 103)

$$\psi_m(x,z) = \exp(-\tfrac{1}{2}x^2/\bar{\xi}_0^2)\mathrm{H}_m(x/\bar{\xi}_0)\exp\{\mathrm{i}[k-(m+\tfrac{1}{2})\Omega z]\}, \qquad (3.1)$$

where $\bar{\xi}_0 = (k\Omega)^{-\frac{1}{2}}$ [see (2.7)] and $\mathrm{H}_m(x)$ denotes a Hermite polynomial of order m.

Let us now assume that the fiber, instead of having a straight axis, is bent with a variable curvature $-C(z)$ from, say, $z = z_1$ to z_2. It is not difficult to show that, within the paraxial approximation, a curvature $-C(z)$ of the guide axis is equivalent to an increment $C(z)x$ in refractive index [i.e., $n(x) \to n(x)+C(z)x$]. This equivalence, incidentally, is not restricted to weak media (BOOKER and WALKINSHAW [1946]). In the mechanical analog previously considered, $C(z)$ represents a driving force acting on the mass (BERREMAN [1965]). The ray equation (2.5) is therefore replaced by the inhomogeneous equation

$$\ddot{\bar{q}} + \Omega^2\bar{q} = C(z). \qquad (3.2)$$

A ray which initially coincides with the fiber axis is driven into oscillations of amplitude $\bar{q}$ and slope $\dot{\bar{q}}$ with a total energy (GK p. 108)

$$E \equiv \tfrac{1}{2}(\dot{\bar{q}}^2 + \Omega^2\bar{q}^2) = \frac{1}{2}\left|\int_{z_1}^{z_2} C(z)\mathrm{e}^{\mathrm{i}\Omega z}\mathrm{d}z\right|^2, \qquad (3.3)$$

where the vertical bars denote: modulus. This result, stated without proof in GK, is most conveniently obtained by introducing the normal-mode amplitude $a \equiv \Omega^{\frac{1}{2}}\bar{q} + i\Omega^{-\frac{1}{2}}\dot{\bar{q}}$ which obeys the differential equation: $\dot{a} + \mathrm{i}\Omega a = \mathrm{i}\Omega^{-\frac{1}{2}}C$. Thus $b \equiv a\exp(\mathrm{i}\Omega z)$ obeys the equation $\dot{b} = \mathrm{i}\Omega^{-\frac{1}{2}}C \times \exp(\mathrm{i}\Omega z)$ whose solution is straight-forward. Eq. (3.3) is obtained by noting that $E = \frac{1}{2}\Omega b^* b$, see LANDAU and LIFSHITZ [1960] p. 63. In establishing (3.3) it was of course assumed that the fiber radius is larger than the maximum excursion of the rays. As an example of application of (3.3) let us assume that the curvature $C(z)$ follows the law $C(z) = \bar{C}\exp(-z^2/\tau^2)$. We have in that case (GK p. 108)

$$E = \tfrac{1}{2}\pi\,\bar{C}^2\tau^2\exp(-\tfrac{1}{2}\Omega^2\tau^2). \qquad (3.4)$$

Let us now consider what is happening to a fundamental (Gaussian) mode launched into the distorted fiber. A well known result of the Quantum Theory of harmonic oscillators is that the center of gravity of a wave packet follows a classical trajectory (see for instance MESSIAH [1961]). A related result has been obtained in the language of Optics by TIEN, GORDON and WHINNERY [1965] and UNGER [1965]. If $\psi(x,z)$ denotes a solution of the unperturbed wave equation and $\bar{q}(z)$ the equation of a ray, the wave function

$$\Psi(x, z) = \psi(x-\bar{q}, z) \exp\left\{ik\left[\dot{\bar{q}}(x-\bar{q})+\int_{z'}^{z}\bar{L}\,dz\right]\right\} \tag{3.5}$$

is a solution of the complete wave equation (GK p. 109). In the above expression, $\bar{L}$ denotes the Lagrangian associated with $\bar{q}(z)$:

$$\bar{L}(z) = 1+\tfrac{1}{2}\dot{\bar{q}}^2-\tfrac{1}{2}\Omega^2\bar{q}^2+C\bar{q}. \tag{3.6}$$

Note, incidentally, that the integration in (3.5) can be partially carried out. We have, using (3.2)

$$\int_{z'}^{z}\bar{L}\,dz = (z-z')+\frac{1}{2}\left[\bar{q}(z)\dot{\bar{q}}(z)-\bar{q}(z')\dot{\bar{q}}(z')+\int_{z'}^{z}C\bar{q}\,dz\right]. \tag{3.7}$$

Because the term involving $\bar{L}$ in (3.5) is a pure phase term, it can be overlooked in the present discussion. The important fact is that the solution $\Psi(x,z)$ is obtained essentially by off-setting the solution $\psi(x,z)$ of the homogeneous equation by a distance equal to the ray position $\bar{q}(z)$.

Let us assume that, at the exit of the fiber, the optical beam is detected with the help of a heterodyne optical receiver (see Fig. 7). If $\psi_{\text{l.o.}}$ and Ψ denote the local oscillator and signal fields, respectively, at the detector, the intermediate frequency current is proportional to

$$\int_{\text{Det.}}\psi^*_{\text{l.o}}\,\Psi\,dx. \tag{3.8}$$

The optical beams being assumed to be concentrated on the detector the integration in (3.8) can be extended to infinity. Assuming now that the field of the local oscillator has been shaped for optimum detection of the unperturbed beam ψ, the received signal is proportional to

$$\int_{-\infty}^{+\infty}\psi^*\Psi\,dx. \tag{3.9}$$

Taking for ψ and Ψ the expressions given before in (3.1) (with $m = 0$) and (3.5), the received signal is found to be proportional to (GK p. 108)

$$\exp(-E/\lambda v), \tag{3.10}$$

E being given by (3.3) or (3.4), and $v \equiv \Omega/2\pi$. (3.10) shows how fast the received signal drops when the fiber is bent according to some curvature law. For instance, a 20 dB drop is experienced if a 1 km long fiber is bent sinusoidally with an amplitude of 1Å (10^{-10}m) and a period equal to $2\pi/\Omega$, assuming that the period of oscillation of the rays is 6 mm and the wavelength 3 μm.

3.2 GREEN FUNCTIONS

Green functions are of great importance in Beam Optics as in other parts of physics. For later use and comparison let us quote a few expressions of the Green functions associated with waves propagating in weak media.

Consider a nonuniform fiber with an index profile

$$n(x, z) = 1 - \tfrac{1}{2}\Omega^2(z)x^2, \tag{3.11}$$

where $\Omega(z)$ is an arbitrary function of z. The field radiated by a point source located at 0, z' is (GK p. 111)

$$G(x, z; 0, z') = \pm(i\lambda q)^{-\frac{1}{2}} \exp(\tfrac{1}{2}i\, k\, \dot{q}q^{-1}x^2) \exp[ik(z-z')], \tag{3.12}$$

where $q(z)$ represents a ray (i.e., a solution of the differential equation: $\ddot{q}+\Omega^2 q = 0$) which satisfies the initial conditions: $q(z') = 0$, $\dot{q}(z') = 1$.

The physical significance of (3.12) is clear. Because the invariant power in the ray pencil bounded by the ray $q(z)$ is proportional to $\psi^*\psi q$, the modulus of $\psi(x, z) \equiv G(x, z)$ is proportional to $q^{-\frac{1}{2}}$. The exponential term in (3.12), on the other hand, expresses the departure of the circular wavefront from the tangent plane. Note that, because $\Omega(z)$ is an arbitrary, perhaps discontinuous, function of z, (3.12) is applicable to sequences of lenses as well as to lens-like media. As we shall see later this simple formula contains all the information needed for understanding the propagation of Gaussian beams through weak media and the resonance properties of unaberrated optical resonators, in two dimensions. The generalizations considered in this paper essentially amount to replacing the scalar ray q by a matricial ray Q for the case of non-orthogonal astigmatic systems (see Section 7) and the ray slope $\dot{q}$ by the variable p canonically conjugate to q for the case of anisotropic media (see Section 8). The higher order modes of propagation are obtained by expanding the complete expression of the Green Function, to be given next, in power series of x'.

Let us assume for generality that the fiber is bent with a curvature $-C(z)$; the equivalent refractive index $n(x)$ is, as we have seen: $1-\frac{1}{2}\Omega^2(z)x^2+C(z)x$. The Green function associated with waves propagating in such a medium is (GK p. 111)

$$G(x, z; x', z') = \pm(i\lambda q)^{-\frac{1}{2}} \exp\left\{\tfrac{1}{2}ik\left[\dot{q}q^{-1}(x-\bar{q})^2 - 2q^{-1}(x-\bar{q})x' + q^{-1}q^{\dagger}x'^2 + \dot{\bar{q}}(x-\bar{q}) + \int_{z'}^{z}\bar{L}\,dz\right]\right\} \exp[ik(z-z')], \tag{3.13}$$

where q and $q^{\dagger}$ are solutions of the homogeneous ray equations

$$\ddot{q}+\Omega^2 q = 0; \qquad q(z') = 0, \qquad \dot{q}(z') = 1$$
$$\ddot{q}^{\ddagger}+\Omega^2 q^{\ddagger} = 0; \qquad q^{\ddagger}(z') = 1, \qquad \dot{q}^{\ddagger}(z') = 0,$$

and $\bar{q}$ is a solution of the complete ray equation

$$\ddot{\bar{q}}+\Omega^2\bar{q} = C; \qquad \bar{q}(z') = 0, \qquad \dot{\bar{q}}(z') = 0.$$

Eq. (3.13) can be used to evaluate the transformation of the field of an arbitrary incident optical beam $\psi'(x')$:

$$\psi(x, z) = \int_{-\infty}^{+\infty} G(x, z; x', z')\psi'(x', z')\mathrm{d}x'. \tag{3.14}$$

Eqs. (3.13) and (3.14) exhibit the fact that the transformation of the field in weak media depends only on the geometrical optics properties of the system (except perhaps for an ambiguity in sign), and more specifically on the trajectories of three independent rays.

§ 4. Complex Ray Representation of Gaussian Beams

We have seen in Section 2 that beam modes can be represented approximately by manifolds of rays bounded by caustics, the rays being real between the caustic curves and imaginary beyond them. A distinctly different description is considered in this section. Fundamental beam modes (Gaussian beams) are described by a single complex ray, or two real rays.

4.1. HISTORICAL BACKGROUND

DESCHAMPS and MAST [1964] observed that the transformation of the radius and wavefront curvature of Gaussian beams through a sequence of lenses is equivalent to an impedance transformation through a reciprocal two-port network, lenses and sections of free-space being analogous to parallel and series reactances, respectively. This analogy shows that it is unnecessary to follow the transformation of the beam step by step through each element of the optical system. The transformation of the beam parameters can be evaluated from the transfer matrix of the equivalent circuit. The transformation of the beam radius and wavefront curvature can alternatively be expressed in terms of the elements of the ray-matrix relating ray positions and slopes at the input and output planes (KOGELNIK [1965] and SUEMATSU and FUKINUKI [1965]). Let

$$\begin{bmatrix} q \\ \dot{q} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} q' \\ \dot{q}' \end{bmatrix} \tag{4.1}$$

denote the ray transformation, primed quantities referring to the input plane. The beam half-width ξ, defined as the distance from axis where the irradiance is reduced by a factor e = 2.718, and the wavefront curvature ρ^{-1} can be combined into a single complex quantity μ, called the complex wavefront curvature; it is defined by

$$\mu = \rho^{-1} + \mathrm{i}/k\xi^2. \tag{4.2}$$

The complex wavefront curvature μ is essentially the "variance" introduced by DESCHAMPS and MAST [1964] and the "complex beam parameter" introduced by KOGELNIK [1965a] who, incidentally, denotes it q. In our work, q denotes ray positions to conform with the notation used in Mechanics. The "ABCD law" of transformation of the curvature μ is (KOGELNIK [1965a])

$$\mu = (C + D\mu')(A + B\mu')^{-1}. \tag{4.3}$$

KOGELNIK [1965b] observed that if we set $\mu \equiv \dot{q}/q$ in (4.3), the quantity $q(z)$ formally obeys the ray equation. He also noted that, in Deschamps and Mast circuit analogy, what corresponds to the electric potential and electric current are the quantities q and $\mathrm{i}\dot{q}$, respectively. No connection was established, however, between the phase of the field and the value of q. The first indication that the phase of the on-axis field is related to the geometrical optics properties of the system is due to VLASOV and TALANOV [1965] who demonstrated that the on-axis phaseshift experienced by a Gaussian beam matched to a periodic system (i.e., a Gaussian beam such that $\mu(z+L) = \mu(z)$, where L denotes the period) is related to the trace $A+D$ of the ray matrix associated with one period of the system. In two dimensions, the on-axis phaseshift is

$$kL - \tfrac{1}{2}\theta = kL - \tfrac{1}{2} \arccos\,[\tfrac{1}{2}(A+D)]. \tag{4.4}$$

DESCHAMPS [1968] noted that the field of a Gaussian beam can be viewed as the field of a ray pencil whose center has a complex location. ARNAUD [1969] observed that the field of a Gaussian beam formally coincides with the Green function (3.12). This observation is clarified in the next section.

4.2. TRANSFORMATION OF GAUSSIAN BEAMS

The Green function (3.12) describes a Gaussian beam if the initial value given to $q(z)$ is complex. Then $q(z)$ is a complex function of z whose real and imaginary parts obey separately the ray equation (2.5). Let us show how we can relate this complex ray $q(z)$ to quantities of direct physical significance such as the beam half-width: ξ and the on-axis phase: $kz - \frac{1}{2}\theta$.

To within an unimportant constant factor (3.12) is

$$\psi(x, z) = q^{-\frac{1}{2}} \exp\left(\tfrac{1}{2}\mathrm{i}\, k\, \dot{q} q^{-1} x^2\right) \exp(\mathrm{i}kz). \tag{4.5}$$

The quantity $\dot{q}q^{-1}$ in (4.5) is the complex wavefront curvature μ. (4.5) readily shows that the phase of the on-axis field $\psi(0, z)$ is equal to the phase of q except for a factor $-\frac{1}{2}$. The phase-shift experienced by a beam with input half-width ξ' and wavefront curvature ρ'^{-1} through an optical system of optical length L is therefore easily obtained. It reads (ARNAUD [1969])

$$kL - \tfrac{1}{2}\theta + \tfrac{1}{2}\theta' = kL - \tfrac{1}{2}\,\text{Phase}\,(q/q') = kL - \tfrac{1}{2}\,\text{Phase of}\,(A + B\mu'),$$

or, making use of (4.2)

$$\text{Phaseshift} = kL - \tfrac{1}{2}\,\text{arccot}\,[k\xi'^2(\rho'^{-1} + A/B)]. \tag{4.6}$$

This expression, together with the "*ABCD* law" (4.3), completely defines the transformation of Gaussian beams through two-dimensional optical systems. Note that (4.6) reduces to (4.4) in the special case where $\mu = \mu'$.

It proves convenient to normalize the complex ray $q(z)$ by the condition

$$(k/2\mathrm{i})\,(q^*\dot{q} - \dot{q}^* q) = 1. \tag{4.7}$$

This is possible because in lossless media the quantity $(q^*\dot{q} - \dot{q}^* q)$ is independent of z. Indeed, setting $q \equiv q_\mathrm{r} + \mathrm{i}q_\mathrm{i}$, the left hand side of (4.7) is $k(q_\mathrm{r}\dot{q}_\mathrm{i} - q_\mathrm{i}\dot{q}_\mathrm{r})$. This quantity, known as a Lagrange ray-invariant (LUNEBURG [1944] p. 251), does not depend on the plane where it is evaluated. When (4.7) holds the beam half-width ξ is precisely equal to the modulus of q.

Let us now consider the propagation of Gaussian beams in free space. In free space, $q(z)$ is a linear function of z which can be written, taking (4.7) into account

$$q(z) = \xi_0 + \mathrm{i}z/k\xi_0. \tag{4.8}$$

This complex ray represents a Gaussian beam whose half-width $\xi(z)$ is

$$\xi^2(z) = q(z)q^*(z) = \xi_0^2 + (z/k\xi_0)^2. \tag{4.9}$$

Thus ξ_0 is the half-width of the beam waist, located at $z = 0$. The phase of the on-axis field, on the other hand, is

$$kz - \tfrac{1}{2}\theta = kz - \tfrac{1}{2}\,\text{Phase}\,(q) = kz - \tfrac{1}{2}\arctan\,(z/k\xi_0^2). \tag{4.10}$$

We observe that the beam experiences a phase retardation equal to $\frac{1}{2}\pi$ from $z = -\infty$ to $z = +\infty$. This phase retardation occurs in the neighborhood of the beam waist. In the limit where $\xi_0 \to 0$, it corresponds to the so-called

"anomalous" phase retardation experienced by ray pencils at focal points (see BORN and WOLF [1965] p. 445). Note also that the complex ray $q(z) \equiv q_r(z) + iq_i(z)$ can be represented as a real skew ray in a three dimensional coordinate system: q_r, q_i, z, as shown in Fig. 8a. In this three-dimensional representation, changing the phase θ of q amounts to rotating the

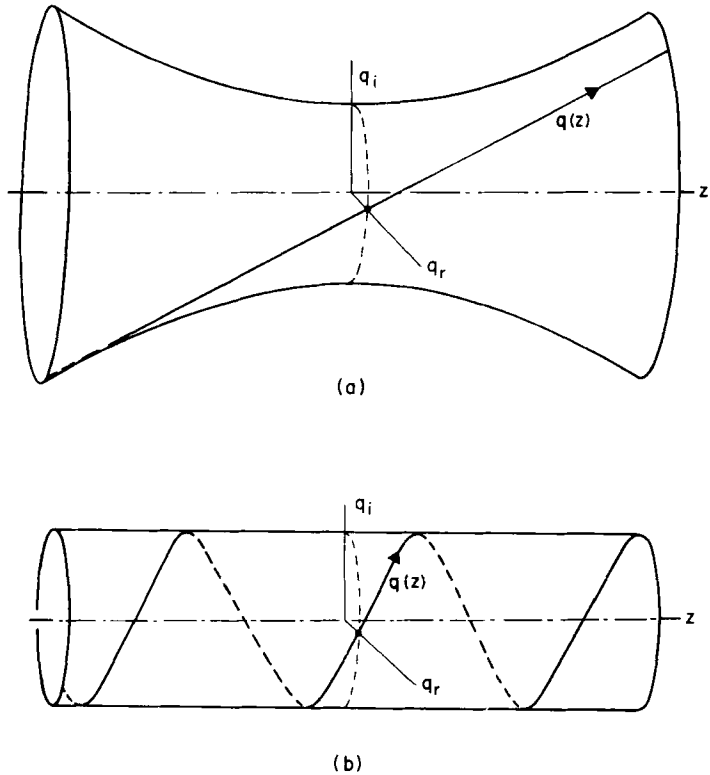


Fig. 8. (a) A two-dimensional Gaussian beam can be represented by a real skew ray in a three-dimensional space: q_r, q_i, z. (b) Three-dimensional ray representation of the Gaussian beam propagation in square-law media.

skew ray about the z-axis. As is well known in geometry, this operation generates a hyperboloid of revolution. Thus the beam profile $\xi(z)$ is a hyperbola, as (4.9) also indicates. Finally, we note that the manifold of rays $q_r(z, \theta)$ coincides with the ray congruence considered in Section 2.

The fact that the beam half-width is equal to the modulus of q suggests a convenient beam tracing method. If we set $q(z) = q_r(z) + iq_i(z)$, $q_r(z)$ and $q_i(z)$ represent two real rays that can be obtained from given initial conditions by ordinary ray tracing methods. At any plane z, the beam half-

width ξ is equal to $(q_r^2+q_i^2)^{\frac{1}{2}}$. Let us illustrate this technique by evaluating the half-width of an incident Gaussian beam with waist half-width ξ_0, at the image focal plane of a lens with focal length f. The tracing of the two rays $q_r(z)$ and $q_i(z)$ shown in Fig. 9a is self-explanatory. It shows that the beam half-width at the image focal plane of the lens is equal to $f/k\xi_0$; it is therefore independent of the location of the beam waist, as shown by KOGELNIK [1965a] with the help of (4.3). Another example of beam tracing is shown in Fig. 9b.

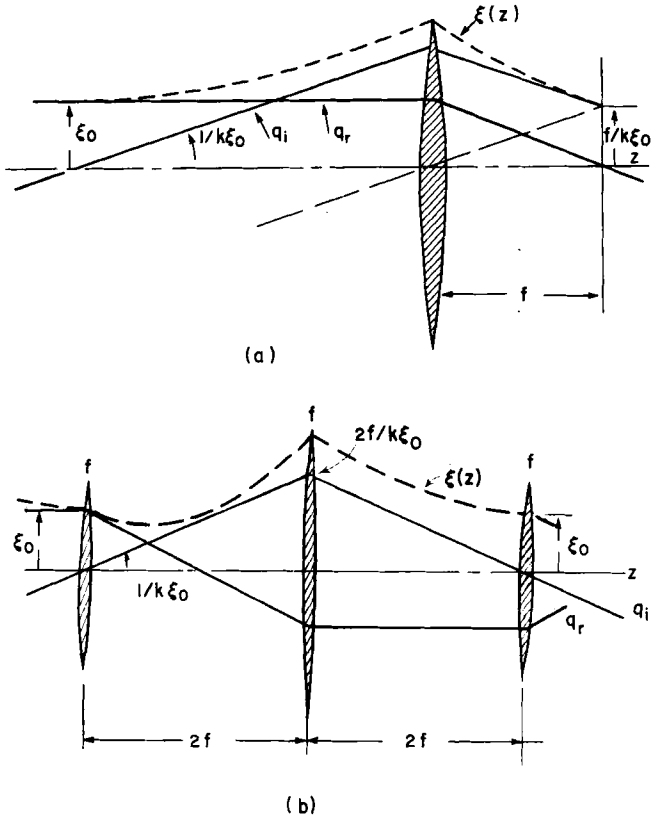


Fig. 9. (a) Beam tracing through a lens. The beam radius at the focal plane of the lens does not depend on the location of the incident beam waist. (b) Beam tracing in a system of three confocal lenses.

Let us now consider the case of a uniform square-law fiber with the refractive index law (2.4). The general solution (2.6) of the ray equation is rewritten, for convenience

$$q(z) = q^+ \exp(i\Omega z) + q^- \exp(-i\Omega z), \tag{4.11}$$

where q^+ and q^- are constants. In the special case where $q^- = 0$, the beam half-width ξ is equal to the modulus of q^+, a constant. Such a beam, with constant half-width $\bar{\xi}_0 = (k\Omega)^{-\frac{1}{2}}$ is, as we have seen before, a mode of the fiber. In the three-dimensional representation (q_r, q_i, z), the ray $q(z) = q^+ \exp(i\Omega z)$ follows a helical path, as shown in Fig. 8b. The real component of q generates the ray manifold considered in Section 2, as the phase of q is varied.

In the more general case where the beam is launched with a waist half-width ξ_0 different from the mode half-width $\bar{\xi}_0$, the initial conditions are $q(0) = q^+ + q^- = \xi_0$ and $\dot{q}(0) = i\Omega(q^+ - q^-) = i/k\xi_0$. The beam half-width $\xi(z) = [q(z)q^*(z)]^{\frac{1}{2}}$ is therefore

$$(\xi/\bar{\xi}_0)^2 = (\bar{\xi}_0/\xi_0)^2 + [(\xi_0/\bar{\xi}_0)^2 - (\bar{\xi}_0/\xi_0)^2] \cos^2 \Omega z. \tag{4.12}$$

This expression shows that the beam half-width ξ oscillates above and below the mode half-width $\bar{\xi}_0$, the maximum and minimum half-widths satisfying the relation $\xi_{min}\xi_{max} = \bar{\xi}^2{}_0$ (TIEN, GORDON and WHINNERY [1965]). Because of the electrical analogy pointed out before, the spatial variations of the beam half-width are the same as the variations of the voltage along an unmatched coaxial line.

4.3. OPTICAL RESONATORS

Let us discuss briefly the general properties of two-dimensional optical resonators. An optical resonator is a section of optical waveguide of the type considered before, closed on itself. The resonance condition is obviously that the field reproduces itself exactly after a round trip. This general condition means that the field must in the first place reproduce itself to within some constant factor α. Secondly the constant α must be unity. In the case of two-dimensional resonators it is sufficient to consider the transformation of the fundamental beam mode. The field of Gaussian beams reproduces itself, to within some constant factor α, if the complex wavefront curvature μ assumes after a round trip its original value μ'. Thus, setting $\mu = \mu'$ in (4.3) and solving for μ one obtains the parameters of the resonating mode at the reference plane. To get the resonance frequencies one needs specify that the factor α multiplying the field is unity. Using the expression of the fundamental phaseshift given in (4.4) and the quantum condition (2.9) we get the resonance condition

$$k_{lm}L = 2l\pi + (m + \tfrac{1}{2})\theta, \tag{4.13}$$

where

$$\theta \equiv \arccos[\tfrac{1}{2}(A + D)].$$

The choice of the reference plane is unimportant because changing the location of the reference plane amounts to performing a similitude transformation on the ray matrix, an operation which preserves the trace $A+D$.

Eq. (4.13) shows that the resonator is lossless (i.e., k_{lm} is real) if

$$-2 < A+D < 2. \tag{4.14}$$

This relation is called the *stability condition* of the resonator (a similar problem is discussed in PIERCE [1954]). When (4.14) is satisfied, there exists one and only one Gaussian beam whose transverse field configuration reproduces itself after a round trip in the resonator (see Fig. 10a).

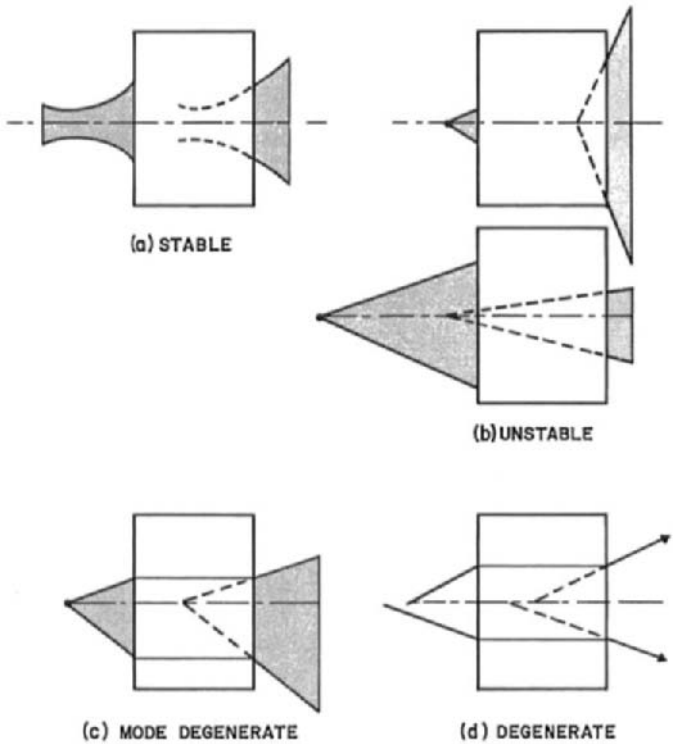


Fig. 10. Self-reproducing beams and rays in periodic systems (only one period is shown). (a) Self-reproducing Gaussian beam in a stable system. (b) Self-reproducing ray pencils in unstable systems. (c) Self-reproducing rays in mode-degenerate resonators. (d) In degenerate resonators all rays recycle.

If $A+D$ is either larger than 2 or smaller than -2, the resonator is *unstable*. The expression obtained from k_{lm} possesses in that case an imaginary part which is interpreted as a loss. This loss does not result from any actual

dissipation in the resonator but from a steady expansion of the circulating beam. The beam eventually reaches the edges of the mirrors, no matter how large the mirrors are. A resonator incorporating two circular mirrors with curvature R^{-1} separated by a distance $\frac{1}{2}L$, for instance, is unstable if $g \equiv 1-L/2R$ is larger than unity, or less than -1. The round-trip power loss for the fundamental mode is, setting $m = 0$ in (2.11)

$$\mathscr{L} = \exp\left[2\,\mathrm{Im}\,(k_{l0}\,L)\right] = \frac{1-(1-g^{-2})^{\frac{1}{2}}}{1+(1-g^{-2})^{\frac{1}{2}}}, \tag{4.15}$$

where Im stands for: imaginary part. This result was first obtained by SIEGMAN [1965] from geometrical considerations. Siegman noted that, when a resonator is unstable, two homocentric ray pencils can be found which reproduce themselves after a round trip (see Fig. 10b). The individual rays of these ray pencils, however, do not recycle because their slopes are multiplied at each passage by real constants that differ from unity, namely $\exp(\mathrm{i}\theta)$ and $\exp(-\mathrm{i}\theta)$ (remember that θ is imaginary in the case of unstable resonators).

In the special case where $A+D$ is precisely equal to either 2 or -2 (e.g., the plane parallel Fabry-Perot) the resonator is called *mode-degenerate* because the resonance frequencies of transverse modes of different orders coincide. The two self-reproducing ray pencils previously mentioned for the case of unstable systems coalesce into a single ray pencil whose individual rays recycle (see Fig. 10c). Arbitrary rays, however, do not recycle, except in the very special case where $A = D = 1$, $B = C = 0$. If the latter conditions are satisfied, the cavity is *degenerate* (see Fig. 10d).

Degenerate cavities are of great practical interest because they provide frequency filtering of incident optical signals without introducing at the same time spatial filtering. A medium with spherical symmetry and refractive index law $n = n_0/(1+r^2)$, known as the Maxwell fish-eye (see LUNEBURG [1944] p. 172, and DEMKOV and OSTROVSKII [1971]) is an example of degenerate optical cavity free of aberration. POLE [1965] proposed a more practical configuration, the "Conjugate Concentric Cavity". A detailed discussion of this class of resonators can be found in ARNAUD [1969] and [1970a].

§ 5. Systems with Non-Uniform Losses

The gain of many gas lasers is not uniform over the cross-section of the discharge tube and decreases off-axis approximately according to a quadratic law (in decibels). This effect may significantly affect the focusing

properties of the resonator in which the active medium is incorporated. Consideration needs sometimes be also given to apertures whose transmissivity decreases according to a Gaussian law. Introduction of such apertures (called for brevity Gaussian apertures) in laser oscillators forces the laser to oscillate in a pure Gaussian mode. Lasers incorporating apertures with sharp edges, in contrast, generate optical beams whose field distribution is only approximately Gaussian. In the case of confocal resonators for instance, the field is described by prolate spheroidal wave functions as shown by BOYD and GORDON [1961].

As is well known, dissipation and gain in a medium are expressed, phenomenologically, by adding an imaginary part to the dielectric constant, i.e., by allowing the refractive index to be complex. Application of this general idea to optical resonators was made by VAKHINOV [1964] who noted that the resonant frequencies and losses of optical resonators incorporating mirrors with Gaussian reflectivity profiles can be obtained by giving complex values to the radius of curvature of the mirrors. To clarify this point of view, let us observe than an ordinary lens with focal length f has a field transmissivity (output field divided by input field) equal to $\exp(-ikx^2/2f)$ at some distance x from axis. A Gaussian aperture with effective half-width a, on the other hand, has by definition a field transmissivity $\exp(-\frac{1}{2}x^2/a^2)$. By comparing these two expressions we see that a Gaussian aperture is formally equivalent to a lens with imaginary focal length

$$f = ika^2. \tag{5.1}$$

Using a similar argument, we easily find that a curved mirror with curvature R^{-1} and power reflectivity: $\exp[-(x/a)^2]$ is equivalent to a mirror with complex curvature $R'^{-1} = R^{-1} - i/2ka^2$.

Let us consider, as an example, a resonator incorporating two identical circular mirrors with Gaussian reflectivity profile. Let the mirror separation be denoted $d \equiv \frac{1}{2}L$, the mirror curvature R^{-1}, and the reflectivity half-width a. The round-trip loss is obtained by replacing R^{-1} in (2.11) by the complex curvature $R'^{-1} = R^{-1} - i/2ka^2$. The round-trip power loss is therefore, for the fundamental mode of resonance

$$\mathscr{L} = \exp[2\,\mathrm{Im}\,(k_{i0}L)] = \exp(2\,\mathrm{Im}\,\{\arccos[1 - d(R^{-1} - i/2ka^2)]\}). \tag{5.2}$$

Curves of constant loss are easily found to be ellipses with equation

$$\left(\frac{g}{\cosh y}\right)^2 + \left(\frac{d/2ka^2}{\sinh y}\right)^2 = 1 \tag{5.3}$$

where we have defined $g \equiv 1 - d/R$ and $\exp(2y) \equiv \mathscr{L}$. These curves are

shown in Fig. 11 for round-trip losses of 1.5 and 5 dB, respectively. The segment $-1 < 1-d/R < 1$ on the g axis corresponds to stable lossless resonators. It should be noted that arccos (x) has an undefined sign. The proper sign in (5.2) is obtained by specifying that the power flowing in the resonator is finite. The upper part ($a^2 > 0$) of the curves corresponds to a loss while the lower part ($a^2 < 0$) corresponds to a gain because, in the latter case, the mirror reflectivity increases as a function of the distance from axis and therefore exceeds unity.

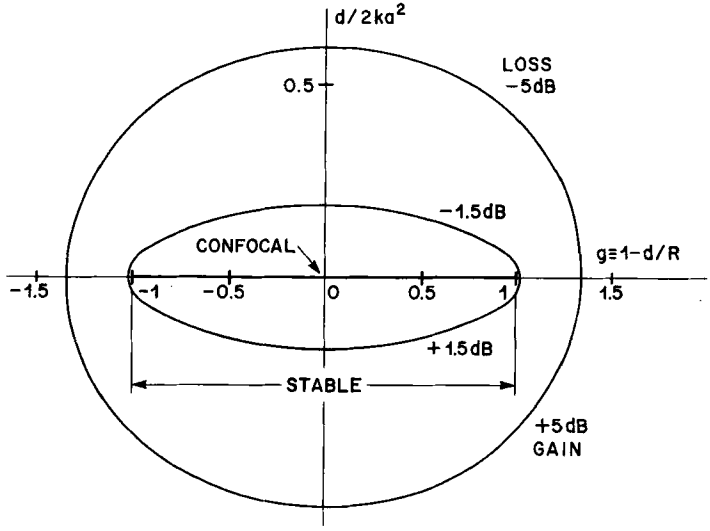


Fig. 11. Curves of constant loss in resonators with Gaussian mirror reflectivity. The segment $-1 < g < +1$ corresponds to stable resonators.

KOGELNIK [1965b] considered the propagation of Gaussian beams through media with non-uniform losses or gain and noted that the transformation of such beams can be understood by giving complex value to the focusing constant in expressions obtained for the lossless case. The above observations are summed up by saying that the expression (3.12) for the field of a Gaussian beam is applicable to lossy media. In this expression, $q(z)$ remains formally a solution of the ray equation $\ddot{q}+\Omega^2 q = 0$. However, because Ω is now complex, it is no longer true that the real and imaginary parts q_r, q_i of q obey separately the ray equation. This circumstance makes beam tracing procedures somewhat more complicated than in the case of lossless media, but all the algebraic results given before are unaffected. Related techniques have been considered for radio ray tracing

in lossy media (see BUDDEN and TERRY [1970] and a discussion by BERTONI, FELSEN and HESSEL [1971]).

§ 6. Mode-Generating Systems

We have seen in Section 4 that fundamental beam modes (Gaussian beams) can be represented by complex rays. This complex ray representation can be extended to higher order modes of propagation with the help of the concept of "mode generating system" which is discussed in this section.

It will perhaps bring some clarity to the subject if we describe first the representation of Gaussian beams by manifolds of complex rays proposed

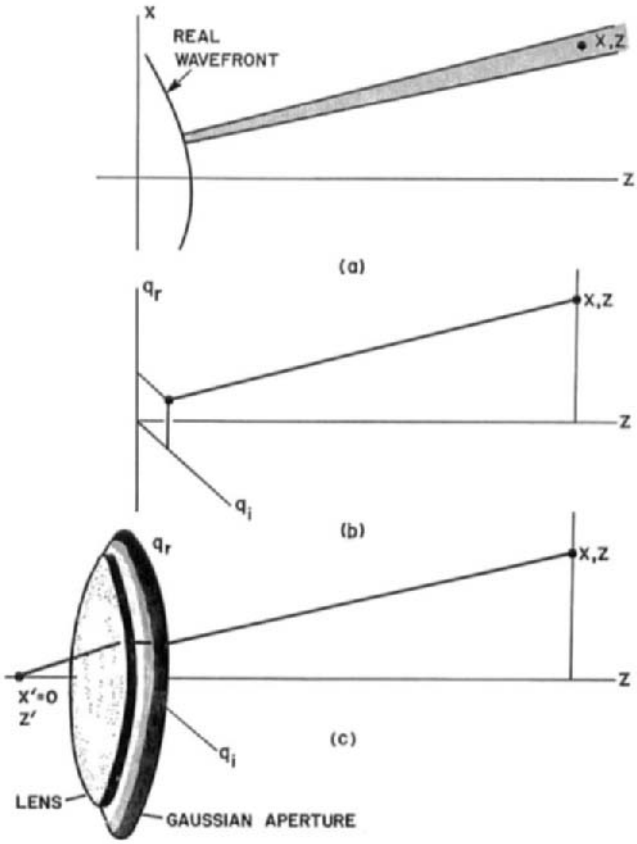


Fig. 12. (a) Evaluation of the field at x, z for the case of a real wavefront. (b) Case of a complex wavefront. (c) Beam generating system, incorporating a lens and an aperture with Gaussian transmissivity.

by KRAVTSOV [1967] (see also KELLER and STREIFER [1971]). Given a real wavefront at plane $z = 0$ with uniform amplitude, the field at some point x,z is easily obtained from the laws of geometrical optics by finding the ray, normal to the wavefront, which passes through the observation point (see Fig. 12a). Kravtsov notes that a Gaussian amplitude distribution $\exp[-\frac{1}{2}(x^2/\xi^2)]$ can be written alternatively $\exp[\frac{1}{2}ik(ix^2/k\xi^2)]$, and the curve $z = ix^2/k\xi^2$ viewed as an imaginary wavefront. The laws of transformation of Gaussian beams in free space are obtained by applying formally to imaginary wavefronts the laws of geometrical optics. The ray going from the imaginary wavefront to the real observing point is of course a complex ray. We observe that the rays perpendicular to the imaginary wavefront are of the form $\alpha q(z)$, where $q(z)$ denotes the complex ray considered before in Section 4, and α is an arbitrary complex number. These rays thus form a 2-parameter ray manifold which can be represented, if we so desire, as a manifold of real rays in the q_r, q_i, z three-dimensional space (see Fig. 12b). This ray manifold forms a skew congruence.

Let us now introduce a slightly different point of view. Let an isotropic point source be located on axis at z' and assume that there is, at plane $z = 0$, a thin lens with focal length $-z'$ and a Gaussian aperture with effective radius a. This aperture is assumed to have a power transmissivity $\exp[-(x/a)^2]$, (see Fig. 12c). The role of the thin lens is to transform the wave diverging from the point source, into a plane wave. The Gaussian aperture, on the other hand, shapes the field amplitude distribution into a Gaussian field distribution: $\psi = \exp[-\frac{1}{2}(x/a)^2]$. Just after the aperture, the field therefore coincides with the field of a Gaussian beam. (This lens-aperture combination is capable of synthesizing Gaussian beams, but it has in practice the disadvantage of absorbing a significant part of the power radiated by the source. The motivation for describing this arrangement is therefore essentially theoretical.) This example shows that the problem of finding the transformation of the field of a Gaussian beam can be reduced to the problem of finding the field radiated by a point source. What is now needed is the Green function for lossy inhomogeneous media. The expression for such a Green function was formally given in (3.12). Generalization to the case of lossy media is straightforward if the medium parameters are analytic functions of their arguments, an assumption which can almost always be made. The algebra will be carried out explicitly in Section 12 for the more general case of non-orthogonal anisotropic media.

The laws of propagation of the higher order modes are also easily obtained along the same lines, by considering the field radiated by off-set point sources given in (3.13); the higher order modes of propagation are

the coefficients of the expansion of the radiated field in power series of the transverse coordinates of the source (for a general discussion see FRIEDMAN [1960]). More precisely, it can be stated that beam modes are the fields radiated by imaginary multipoles through a properly chosen optical system.

The most convenient way of evaluating the coupling between two radiators consists of making use of the mode-generating system concept because application of this concept avoids the need of performing integrations. Let us consider, for concreteness, a transmitting antenna (α) and, at some distance away from it, a receiving antenna (β) as shown in Fig. 13a, and let us ask for the voltage generated across the load of antenna (β) when a

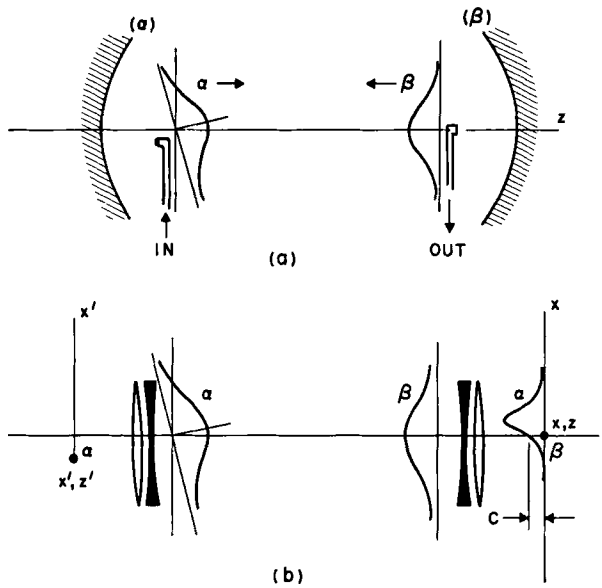


Fig. 13. This figure illustrates the mode-generating system approach to the evaluation of the coupling between two antennas. (a) Actual arrangement. (b) Equivalent arrangement. The coupling C is proportional to the field radiated at (β) by (α), or vice-versa.

voltage unity is applied to antenna (α). Assuming that the two antennas are matched to their loads, the received voltage is proportional to the integral of the product of the field ψ_α radiated by antenna (α) and the field ψ_β that (β) would radiate if it were used as a transmitter. This integral can be evaluated over any plane located between the two antennas. Let us now assume that the field distributions generated by the antennas are Gaussian; this is often the case in practice, strong illumination tapers being

usually introduced to minimize the level of the side lobes. Because each Gaussian field distribution can be represented by a point source, as we have seen, the problem of evaluating the coupling between two antennas is reduced to the problem of evaluating the coupling between two point sources. This coupling is simply the field radiated at β by α (or vice-versa), because the field radiated by β is a delta function (see Fig. 13b). Thus, the coupling can be obtained without integration, by application of (3.13). (For detailed algebraic expressions see ARNAUD [1971a].)

§ 7. Non-Orthogonal Optical Systems

Astigmatic optical systems are frequently encountered in Beam Optics. Brewster angle windows, for instance, often used in laser technology, introduce a small degree of astigmatism. Because spherical mirrors under oblique incidence exhibit different focal lengths for ray manifolds lying in the incidence plane and in the perpendicular plane, most ring lasers are astigmatic (see COLLINS [1964], BAUES [1969], HERTZ and MINKWITZ [1969]). Such systems possess meridional planes of symmetry and their properties can be discussed by considering separately what is happening in the two mutually perpendicular meridional planes of symmetry. The analysis thus reduces to the two-dimensional case discussed before and need not be repeated. It may happen, however, that the optical system considered does not possess meridional planes of symmetry. This is the case, for instance, for a sequence of two cylindrical lenses oriented at an angle of, say, 45° to one another (see Fig. 14a). Such optical systems are called "non-orthogonal". The transformation of paraxial rays can be described by a ray matrix

$$\begin{bmatrix} q \\ \dot{q} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} q' \\ \dot{q}' \end{bmatrix} \equiv \mathfrak{M} \begin{bmatrix} q' \\ \dot{q}' \end{bmatrix}, \tag{7.1}$$

which has the same formal appearance as the ray matrix (4.1) applicable to two-dimensional systems. In the present case, however, q and $\dot{q}$ denote 2-vectors, and A, B, C, D denote 2×2 matrices. The inverse of the 4×4 matrix $\mathfrak{M}$ is (LUNEBURG [1944] p. 216)

$$\mathfrak{M}^{-1} = \begin{bmatrix} \tilde{D} & -\tilde{B} \\ -\tilde{C} & \tilde{A} \end{bmatrix}. \tag{7.2}$$

This result (7.2) rests on the existence of a point-eikonal function. A, B, C, D satisfy a number of relations that are readily obtained from (7.2). They will be given explicitly and discussed in Section 11.

Modal solutions in non-orthogonal optical systems were investigated by KAHN and NEMIT [1967] who considered the case of periodically rotated astigmatic lenses. Using a real ray technique (see Section 2) they defined the conditions for stability and unstability of the system. The transformation of Gaussian beam irradiances and wavefronts in nonorthogonal optical systems was obtained by SUEMATSU and FUKINUKI [1968], (see also COLLINS [1970]). The expressions given by these authors are rather complicated. Actually, the *ABCD* law is applicable to non-orthogonal systems provided A, B, C, D, μ and μ' are understood as 2×2 matrices rather than scalars and the terms are ordered as shown in (4.3). Experimental observation of the transformation of Gaussian beams through non-orthogonal systems was made by ARNAUD and KOGELNIK [1969]. The beam irradiance pattern is elliptical in shape. It is interesting to note that this elliptical pattern

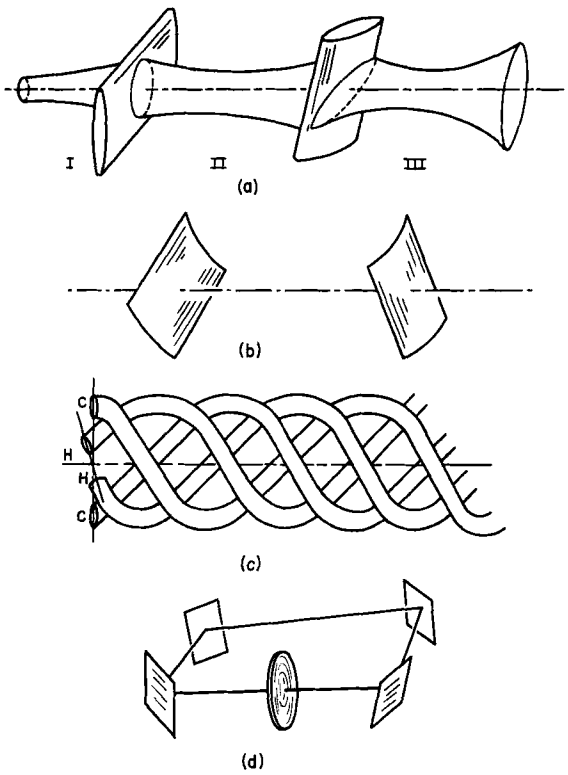


Fig. 14. Non-orthogonal optical systems. (a) Gaussian beam transformation in a sequence of two cylindrical lenses with different orientations. (b) Resonator incorporating cylindrical mirrors. (c) Helical gas lens. (d) Cavity with image rotation.

rotates continuously as the beam propagates, the total rotation angle from $z = -\infty$ to $z = +\infty$ being equal to π in free space.

Far-reaching results concerning the theory of non-orthogonal resonators were obtained by POPOV [1968], who gave a general expression for the resonant frequencies of lossless non-orthogonal resonators and applied it to the case of a resonator incorporating two cylindrical mirrors whose generatrices make arbitrary angles to one another (see Fig. 14b). This expression has been generalized to the case of lossy systems (ARNAUD [1971a]). Let λ_1, λ_2, λ_3, λ_4 denote the four eigenvalues of the matrix $\mathfrak{M}$. Because of the special properties of $\mathfrak{M}$ exhibited by the Luneburg rule of inversion (7.2), it turns out that $\lambda_3 = \lambda_1^{-1}$ and $\lambda_4 = \lambda_2^{-1}$. It can be shown that, after a round-trip in the resonator the field of a mode m_1, m_2 reproduces itself except for a constant factor $\pm\lambda_1^{m_1+\frac{1}{2}}\lambda_2^{m_2+\frac{1}{2}}$ and a term expressing the geometrical-optics round-trip phase shift: $\exp(ikL)$. Thus the resonance frequencies are given by

$$\pm\exp(ik_{lm_1m_2}L)\lambda_1^{m_1+\frac{1}{2}}\lambda_2^{m_2+\frac{1}{2}} = \exp(2il\pi) \tag{7.3}$$

where l is the axial mode number. Note that there is some arbitrariness in the choice of λ_1 and λ_2 which, according to our previous discussion, could be replaced by their inverses λ_1^{-1} and λ_2^{-1}, respectively. This ambiguity can be lifted by specifying that the power in the mode is finite. Setting now $\lambda_1 \equiv \exp(i\theta_1)$ and $\lambda_2 \equiv \exp(i\theta_2)$, the resonance condition (7.3) can be rewritten

$$k_{lm_1m_2}L = 2l\pi + (m_1+\tfrac{1}{2})\theta_1 + (m_2+\tfrac{1}{2})\theta_2 \pm \pi. \tag{7.4}$$

It can be shown that if the resonator is free of dissipation losses ($\mathfrak{M}$ real), θ_1 and θ_2 are either real or imaginary. If θ_1 and θ_2 are both real, $k_{lm_1m_2}$ is real and the resonator is stable. This is the case explicitly considered by POPOV [1968]. When either θ_1 or θ_2 are imaginary, $k_{lm_1m_2}$ has an imaginary positive part; the resonator is unstable and suffers from the geometrical losses discussed in Section 4 for the two-dimensional case.

Let us observe further that (7.4) reduces to (4.13) in the case of two-dimensional resonators. For most resonators that have rotational symmetry $\theta_2 = \theta_1 \equiv \theta$. Then (7.4) becomes

$$k_{lm_1m_2}L = 2l\pi + (m_1+m_2+1)\theta. \tag{7.5}$$

Such resonators with rotational symmetry are obviously mode-degenerate.

An interesting result concerning optical resonators with folded optical-axis is that, in general, the wavefronts of the resonating modes coincide with the surface of the end mirrors. (Furthermore the field at the end mir-

rors assumes the same form as in conventional resonators in an oblique coordinate system.) This result is most easily verified for the case of weak media by making use of the symmetry property of the round-trip point-eikonal. More generally, this property results from the fact that the operator expressing the round-trip transformation of the field evaluated from and to one of the two end mirrors is symmetric if the medium is free of internal reflection and reciprocal, and unitary if the medium is lossless. As one easily proves, the eigenvectors of symmetric unitary operators corresponding to distinct eigenvalues are real. Because the modal functions are real the wavefronts coincide with the mirror surface.

POPOV [1969a, b] and ARNAUD [1970b] gave descriptions of Gaussian beams in terms of complex matricial rays Q. With the formal substitution $q \to Q$, results obtained for the two-dimensional case are applicable to non-orthogonal systems as well. For instance the field of Gaussian beams propagating in a medium with refractive index

$$n(x, z) = 1 + \tfrac{1}{2}\tilde{x}Nx \tag{7.6}$$

where x denotes a vector with components x_1, x_2 and N a 2×2 matrix, is obtained essentially by replacing q by Q in (4.5). We have

$$\psi_{00}(x, z) = |Q|^{-\frac{1}{2}} \exp\left(\tfrac{1}{2}ik\,\tilde{x}\dot{Q}Q^{-1}x\right) \exp(ikz) \tag{7.7}$$

where $|Q|$ denotes the determinant of the matrix Q. Q denotes a solution of the matrix ray equation $\ddot{Q} = NQ$ with initial conditions so chosen that ψ assumes the specified form at the input plane.

POPOV [1968] gave an expression for the fields of modes of resonance of any order (m_1, m_2) in the form

$$\psi_{m_1 m_2} = \Lambda_1^{m_1} \Lambda_2^{m_2} \psi_{00}, \tag{7.8}$$

where ψ_{00} denotes the fundamental mode given in (7.7) and the Λ_i, $i = 1, 2$, denote raising operators of the form: $(ik)^{-1}\tilde{q}_i\partial/\partial x - \dot{\tilde{q}}_i x$, were the q_i denote complex rays. ARNAUD [1970b] has shown that $\psi_{m_1 m_2}$ is, explicitly, a product of a Gauss function and a Hermite polynomial in two complex variables. In general, however, the fields $\psi_{m_1 m_2}$ defined by (7.8) are not orthogonal. Special conditions must be met by the generating rays q_1, q_2 if the orthogonality condition is to hold. An orthogonal set of modes is easily obtained with the help of the mode generating system concept (see Sec. 12). The modes of resonance of a resonator, on the other hand, are generated by the modal matrix of the resonator round-trip ray matrix, itself a ray matrix.

The helical gas lens is an example of non-orthogonal optical system of practical interest. The focusing properties of electrostatic lenses incorporating four coaxial helices at potentials $+V$, $-V$, $+V$ and $-V$, respectively, have been known for a long time in the technology of particle accelerators. These electrostatic lenses evolved from the concept of strong-focusing according to which a periodic sequence of converging and diverging lenses with equal absolute powers has a net focusing effect, provided the period

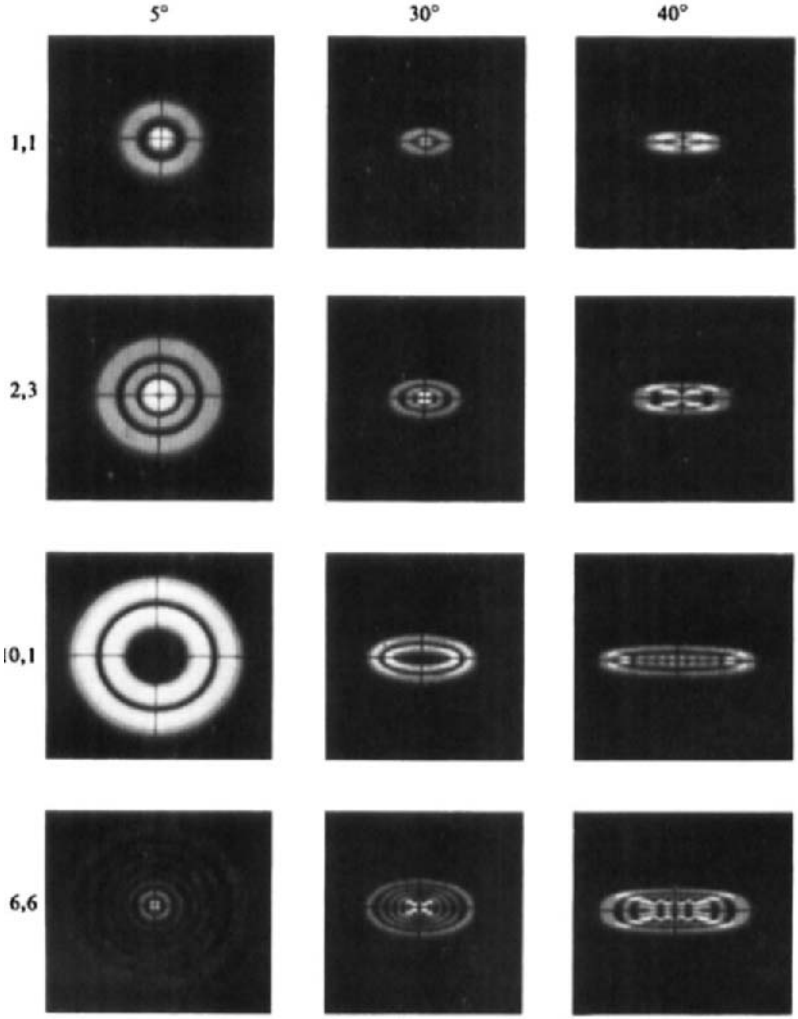


Fig. 15. Computed beam patterns in helical gas lenses. The numbers on the left are the mode numbers. The numbers on top (5°, 30°, 40°) are related to the helix temperature (after ARNAUD [1972b], computed by W. Mammel. By permission of the Optical Society of America).

does not exceed a certain critical value. TIEN, GORDON and WHINNERY [1965], UNGER [1965] and MARIÉ [1970] have discussed the application of this technique to the guidance of optical beams, the four helices being raised at alternately high and low temperatures (see Fig. 14c). Because of the difference in temperature, gradients of refractive index are created in the gas filling the space inside the helices. The gas thus acts as a quadrupole lens whose principle axes rotate along the system axis. Alternatively, the gas can be replaced by a liquid with low optical losses. Refractive index gradients of the type considered can also be induced in electrooptic materials by d.c. electric fields. TIEN, GORDON and WHINNERY [1965] gave an approximate expression for the field of the fundamental mode of propagation, applicable when the temperature difference between the helices is small. MARIÉ [1970] gave an exact description of the fundamental mode of propagation. The general modal solution was obtained by ARNAUD [1972b]. Fig. 15 shows irradiance patterns calculated for various mode numbers and temperatures of the helices. These patterns rotate along the system axis together with the helices.

Another non-orthogonal system of possible practical interest is the cavity with image rotation (ARNAUD [1970b], BERGMANN [1972]). This three-dimensional optical resonator possesses the property that the irradiance of all the modes is rotationally symmetric. In addition, the usual polarization degeneracy is lifted as a result of the twist of the path. This feature is of interest for reducing the coupling between clockwise and counterclockwise waves in laser gyroscopes.

§ 8. Anisotropic Media

In the previous discussions we were exclusively concerned with isotropic media, i.e., with media in which the wavelength is independent of the direction of propagation. Many lasers, however, incorporate materials that have either natural or induced birefringence.

Let us first consider anisotropic effects that are too small to influence significantly the focusing power of the medium. For instance, if we let a ring laser spin about an axis perpendicular to its plane, the resulting Doppler effect can be accounted for by assuming that the medium is slightly nonreciprocal. To evaluate this effect let us consider a perfectly conducting cylinder in free space (see Fig. 16). Clockwise (CW) and counterclockwise (CCW) waves can cling to the concave side of this boundary with a velocity almost equal to the speed of light c (the phase velocity slightly exceeds c at the boundary; this is of no consequence for the present discussion). Let

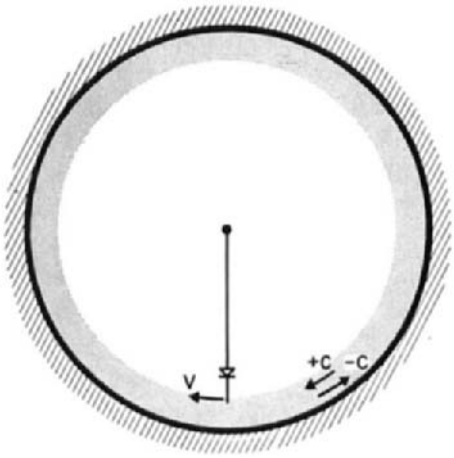


Fig. 16. This figure illustrates the principle of laser gyroscopes. The rotation of the circular boundary being immaterial, the beat frequency between clockwise and counterclockwise waves can be seen as resulting from the rotation of the mixer (Doppler effect).

us now assume that the fields of the CW and CCW waves are mixed in a rotating nonlinear device having a linear velocity v. Because of the Doppler effect, the relative beat frequency between CW and CCW waves is

$$\Delta f/f = 2\, v/c. \tag{8.1}$$

The rotation of the cylinder itself being immaterial, (8.1) is applicable to the case where the cylinder and the detector are held fixed with respect to one another and rotate at the same angular velocity; such an arrangement is called a laser gyroscope. If the gyroscope is immersed in a medium with refractive index n moving with it, the drag of the CW and CCW waves resulting from the rotation of the medium has to be taken into account. It is evaluated by using the formula of addition of velocities of special relativity: $u = (v_1+v_2)/(1+v_1v_2/c^2)$ with $v_1 = v$ and $v_2 = \pm c/n$. The relative beat frequency is found to be, after a few rearrangements

$$\Delta f/f = 2v/cn, \tag{8.2}$$

to first order in v/c. Note, incidentally, that the gyroscopic effect vanishes in the non-relativistic limit of large n. These results (8.1), (8.2) can be obtained alternatively in a non-inertial frame of reference rotating with the gyroscope, on the basis of the theory of General Relativity (HEER

[1964])*. We now observe that the rotation can be ignored provided the isotropic medium is replaced by a fictitious non-reciprocal medium with refractive indices $n_{CW} = n+v/c$ and $n_{CCW} = n-v/c$ applicable to waves propagating in the CW and CCW directions, respectively. The lack of isotropy resulting from spinning is usually very small ($v \ll c$) and does not affect significantly the focusing properties of the medium.

The propagation of beam modes in strongly anisotropic materials has been investigated by BERGSTEIN and ZACHOS [1966], BHAWALKAR, GONCHARENKO and SMITH [1967], SCHAEDLA and BEYER [1968], WUNSCHE [1970], TANAKA, SUZUKI and MATSUMOTO [1970], SCHACHTER and CHANGHWI [1970], SUEMATSU [1971] and ERMERT [1971]. These authors gave consideration to uniaxial, biaxial or gyrotropic media and based their calculations on approximate forms of Maxwell's equations.

MASON [1971] has shown that if the surface of wave normals can be approximated by a parabola, i.e., if

$$k\partial S/\partial z = \phi_0+\phi_1 kp+\phi_2 k^2p^2 \equiv -kH(p), \tag{8.3}$$

where $p \equiv \partial S/\partial x$ and ϕ_0, ϕ_1, ϕ_2 are constants, the law of transformation of the half-width ξ of a Gaussian beam with a half-width ξ_0 at the input plane is

$$\xi^2 = \xi_0^2+(2z\phi_2/\xi_0)^2. \tag{8.4}$$

This result can be obtained alternatively from a Hamiltonian approach (ARNAUD [1972a]). Because we are interested only in the half-width of the beam, consideration need be given only to the term in x^2 in the expression of the point-eikonal. Thus, setting $S \equiv \frac{1}{2}Ux^2$, substituting in (8.3), and comparing the terms in x^2, we find that U obeys the differential equation $\dot{U} = 2\phi_2 kU^2$ whose solution is $U = (a-2\phi_2 kz)^{-1}$, where a denotes an integration constant. Because the field is proportional to $\exp(ikS)$, the field amplitude is $\exp\{\mathrm{Re}\,[\frac{1}{2}ik(a-2\phi_2 kz)^{-1}x^2]\} \equiv \exp[-(x/\xi)^2]$ where ξ is given by (8.4), a being obtained by specifying that $\xi = \xi_0$ at $z = 0$.

In the example just discussed the solution of the Riccati equation $\dot{U} = 2\phi_2 kU^2$ was straightforward. However, in more complicated cases it is easier to solve linear ray equations. If we set $U = pq^{-1}$ we find that q, p obey Hamilton's equations for light rays without misalignment terms [i.e., with $\phi_1 = 0$ in (8.3)]. Thus, we have, from (8.3)

* Note that if the laser is located in a gravitational potential ϕ ($\phi < 0$), the frequency measured by a distant observer where $\phi = 0$ is red-shifted as a result of the loss of momentum of the photons. The relative change in frequency is $\Delta f/f = \phi/c^2$ as one easily finds from the principle of equivalence of General Relativity. This effect is of course the same for CW and CCW waves.

$$\dot{q} = \partial H/\partial p = -2\phi_2 kp, \qquad (8.5)$$
$$\dot{p} = -\partial H/\partial q = 0.$$

The second equation (8.5) shows that p is a constant; this is a consequence of the homogeneity of the medium. The solution of the first equation (8.5) is clearly $q = b - 2\phi_2 kpz$, where b is a constant of integration. Setting $a \equiv b/p$ we find $U = pq^{-1} = (a - 2\phi_2 kz)^{-1}$ as before. This simple example will help clarify the more general derivations given in Sections 10 to 12.

It is interesting to note that in the realm of Wave Mechanics anisotropies usually arise as a result of the presence of d.c. magnetic fields. The solutions obtained in the case of harmonically bound particles immersed in a magnetic field, for instance (JONES and PAPADOPOULOS [1972]), are related to those discussed in the present section. The case of general quadratic Lagrangians has been discussed in an important paper by CHERNIKOV [1968].

§ 9. The Scalar Wave Equation

The purpose of this section is to show that under certain conditions Maxwell's equations can be reduced to a scalar wave equation. The approximations needed are rather drastic; yet they appear to be well supported by experiment for most systems of practical interest.

9.1. THE EIKONAL EQUATION

The optical field e, h obeys Maxwell's equations. In a time-invariant linear medium, free of spatial and temporal dispersion and free of sources, these equations are conveniently written

$$M(r) \begin{bmatrix} \partial \\ \nabla \times \end{bmatrix} \mathsf{e}(r, t) = \begin{bmatrix} -\nabla \times \\ \partial \end{bmatrix} h(r, t), \qquad (9.1)$$

where $M(r)$ denotes a 6×6 matrix which characterizes the macroscopic electromagnetic properties of the medium at some point r in space. ∂ stands for $\mathbf{1}\partial/\partial t$ where $\mathbf{1}$ denotes the 3×3 unit matrix and $\nabla \times$ denotes the rotational operator.

An equation for e alone is obtained by multiplying (9.1) on the left by $[\partial \; \nabla \times]$. We get

$$[\partial \; \nabla \times] M \begin{bmatrix} \partial \\ \nabla \times \end{bmatrix} \mathsf{e} = 0. \qquad (9.2)$$

When the medium is homogeneous, i.e., when M does not depend on r, the three components of e obey the same scalar wave equation

$$\det. \left\{ [\partial \; \nabla \times] M \begin{bmatrix} \partial \\ \nabla \times \end{bmatrix} \right\} e_j = 0, \qquad j = 1, 2, 3. \tag{9.3}$$

Let us now consider the quasi-classical approximation and assume that the field has the form

$$\mathsf{e} = \bar{\mathsf{e}} \exp (\mathrm{S}), \tag{9.4}$$

where the space-time variations of $\bar{\mathsf{e}}$ are slow compared with those of S. Then $\partial \mathsf{e} \sim (\partial \mathrm{S})\mathsf{e}$. Substituting in (9.2) fields of the form (9.4) we get an equation for $\bar{\mathsf{e}}$ which reads

$$[\partial \mathrm{S} \; \nabla \mathrm{S} \times] M \begin{bmatrix} \partial \mathrm{S} \\ \nabla \mathrm{S} \times \end{bmatrix} \bar{\mathsf{e}} = 0. \tag{9.5}$$

Eq. (9.5) is a system of three linear equations in $\bar{e}_1, \bar{e}_2, \bar{e}_3$ which admits nontrivial solutions ($\bar{\mathsf{e}} \neq 0$) only if

$$\det. \left\{ [\partial \mathrm{S} \; \nabla \mathrm{S} \times] M \begin{bmatrix} \partial \mathrm{S} \\ \nabla \mathrm{S} \times \end{bmatrix} \right\} = 0. \tag{9.6}$$

Equation (9.6) is a partial differential equation for S, called the eikonal equation; it can be obtained formally by the substitution $\partial \to \partial \mathrm{S}$ in the scalar wave equation (9.3) applicable to homogeneous media. Note that (9.6) is un-affected if $\nabla \mathrm{S}$ is changed to $-\nabla \mathrm{S}$ and M to its transposed $\tilde{M}$, because the transposed of the antisymmetrical matrix $\nabla \mathrm{S} \times$ is $-\nabla \mathrm{S} \times$. Thus, to any ray trajectory in the first medium (M), there corresponds an identical ray trajectory in the transposed medium ($\tilde{M}$), described in the opposite direction.

9.2. RECIPROCITY

Let us introduce an adjoint field $\mathsf{e}^\dagger$, $h^\dagger$ propagating in the transposed medium, characterized by the matrix $\tilde{M}$. A generalized form of Lorentz reciprocity theorem (KONG [1970]) states that, for any closed surface $\mathscr{S}$

$$\int_{\mathscr{S}} (v^\dagger \cdot i - i^\dagger \cdot v) \mathrm{d}\mathscr{S} = 0. \tag{9.7}$$

In (9.7) v denotes the tangential component of e on $\mathscr{S}$ and $i \equiv h \times \hat{n}$ ($\hat{n}$ being the unit vector normal to $\mathscr{S}$ pointing outward) is also tangent to $\mathscr{S}$. Similar definitions are applicable to $v^\dagger$, $i^\dagger$.

If we choose for $\mathscr{S}$ two planes perpendicular to some axis z at z and $z + \mathrm{d}z$ ($\mathscr{S}$ is closed at infinity where the fields are assumed to be negligibly small),

(9.7) expresses the fact that the integral over a transverse plane of $(v^\ddagger \cdot i - i^\ddagger \cdot v)$ is independent of z.

For the clarity of the discussion, let us consider now lossless birefringent crystals. In such crystals, the electric field, the magnetic field and the Poynting vector are, for each eigenstate of polarization, three mutually perpendicular real vectors (see BORN and WOLF [1965] p. 665). Thus, assuming that the beam (e, h) propagates in a direction close to the z axis we have approximately, to within an unimportant factor, $i = nv$, where n denotes the ray index of refraction. Let us assume further that $\mathsf{e}^\ddagger, h^\ddagger$ describes a beam propagating in a direction close to the $-z$ axis with a state of polarization corresponding to the state of e, h. We have approximately $i^\ddagger = -nv^\ddagger$. Therefore Lorentz reciprocity theorem (9.7) can be written

$$\frac{\mathrm{d}}{\mathrm{d}z}(\psi^\ddagger, \psi) = 0, \tag{9.8}$$

where we have defined $\psi \equiv n^{\frac{1}{2}}v$, $\psi^\ddagger \equiv n^{\frac{1}{2}}v^\ddagger$ and $(\psi^\ddagger, \psi)$ denotes a scalar product in both real and function space

$$(\psi^\ddagger, \psi) \equiv \int\int \psi^\ddagger \cdot \psi \, \mathrm{d}x_1 \, \mathrm{d}x_2 . \tag{9.9}$$

We now assume that the state of polarization of the waves is independent of x_1, x_2, and set $\psi(x, z) \equiv \psi(x, z)\,\phi\,(z)$, $\psi^\ddagger(x, z) \equiv \psi^\ddagger(x, z)\phi^\ddagger(z)$. The transformation of $\phi, \phi^\ddagger$ is assumed to be the same as if the medium parameters were independent of x_1, x_2. The problem of finding the transformation of the vector wave function $\psi(x, z)$ is consequently split into two independent problems: first, find the transformation of the state of polarization $\phi(z)$. Secondly, find the transformation of the scalar wave function $\psi(x, z)$. (9.8) is applicable separately to $\phi(z)$, $\phi^\ddagger(z)$ and to $\psi(x, z)$, $\psi^\ddagger(x, z)$; scalar products in real space are involved in the first case and scalar products in function space are involved in the second case.

9.3. TRANSFORMATION OF THE POLARIZATION

We are dealing in this subsection with the problem of finding the transformation of plane waves propagating along the z-axis. The medium parameters are assumed to depend only on the z coordinate, i.e., the medium is stratified. Because the scale of the inhomogeneities is large compared with the wavelength, the medium can be considered free of reflection and only two waves, both propagating in the forward direction, need be considered. The vector $\phi(z)$ therefore obeys a first order differential equation

of the form $\dot{\phi}(z) = A(z)\phi(z)$ where $A(z)$ denotes a 2×2 matrix. The adjoint equation is clearly $\dot{\phi}^\dagger = -\tilde{A}\phi^\dagger$ if (9.8) is to be satisfied. It is convenient to diagonalize A and write it $R\dot{S}R^{-1}$ where $\dot{S}$ denotes a diagonal matrix whose elements $\dot{S}_1$ and $\dot{S}_2$ are essentially the local propagation constants of the two waves. The columns of the modal matrix R, on the other hand, are the local eigenstates of polarization. Thus ϕ and $\phi^\dagger$ obey the differential equations,

$$\dot{\phi} = R\dot{S}R^{-1}\phi, \tag{9.10a}$$

$$\dot{\phi}^\dagger = \tilde{R}^{-1}(-\dot{S})\tilde{R}\phi^\dagger, \tag{9.10b}$$

respectively. When the medium is isotropic we have $\dot{S}_1 = \dot{S}_2 \equiv \dot{S}$, and (9.10a) has the obvious solution $\phi(z) = \phi(0) \exp[S(z)]$ indicating that the state of polarization is invariant. If, on the other hand, the medium is anisotropic ($\dot{S}_1 \neq \dot{S}_2$) and R varies slowly with z, (9.10a) has the solution $\bar{\phi}(z) = \exp[S(z)]\bar{\phi}(0)$, where $\bar{\phi}(z) \equiv R^{-1}(z)\phi(z)$. Because S is diagonal, this result shows that the components of the wave function on each eigenstate of polarization can be dealt with independently when the eigenstates are not degenerate and vary slowly with z. We shall restrict ourselves to these two cases (isotropic or weakly inhomogeneous media).

It sometimes happens that the z axis is curved rather than straight. In that case, a natural choice for the x_1 and x_2 axes is the binormal and the principal normal to the curve z, respectively. This coordinate system, however, rotates about the z axis at a spatial rate equal to the torsion τ of the axis. Thus, in order to preserve the applicability of our previous results, it is preferable to define x_1, x_2 as the coordinate system which rotates about z with respect to the binormal at a rate $-\tau$ (see RYTOV [1937], LUNEBURG [1944] p. 57).

It is instructive to consider the case where the z axis is a closed nonplanar curve and the medium is isotropic. After a round trip, ϕ experiences, with respect to its original state, a rotation equal and opposite to the integrated torsion of the closed curve. Thus, only circular states of polarization (clockwise or counterclockwise) reproduce themselves after a roundtrip, except for a phase factor. The same conclusion is reached for the case where the closed path is defined by an even number of perfectly conducting mirrors. Because the tangential component of the total electric field vanishes on the mirror surface, the vector ϕ experiences upon reflection a symmetry with respect to the normal to the mirror; this transformation can be viewed as the product of a symmetry with respect to the plane of the mirror and an inversion with respect to the point of incidence. When the number of mirrors is even, all inversions cancel out. An even number of symmetries, on

the other hand, is equivalent to a rotation. The transformation is therefore the same as in the case of isotropic continuous media. In both cases the resonating fields are circularly polarized (clockwise and counterclockwise).

9.4. THE ORDERED SCALAR WAVE EQUATION

We now give consideration to the scalar part $\psi(x, z)$ of the vector wave function. The eikonal equation (9.6) can be solved, in principle, for $\partial S/\partial z$ and written, for time-harmonic sources (the $-i\omega t$ time dependence of S being omitted)

$$H(p, q, z)+\partial S/\partial z = 0, \tag{9.11a}$$

where we have set $p \equiv \partial S/\partial x$ and allowed x to assume complex values q. H is called the Hamiltonian. For rays in the transposed medium, the eikonal equation is, changing ∇S to $-\nabla S$,

$$H(-p, q, z)-\partial S/\partial z = 0. \tag{9.11b}$$

As we have seen, these equations can be obtained from the scalar wave equation applicable to the case of homogeneous media through the substitution $\nabla \to \nabla S$. We wish to investigate whether, inversely, approximate yet physically meaningful scalar wave equations can be obtained from the substitution $\nabla S \to \nabla$ in (9.11) when the medium lacks homogeneity. Specifically, we wish to see under what conditions the reciprocity relation (9.8)

$$\frac{d}{dz}(\psi^\dagger, \psi) = (\partial\psi^\dagger/\partial z, \psi)+(\psi^\dagger, \partial\psi/\partial z) = 0 \tag{9.12}$$

holds.

Let the Hamiltonian $H(p, q, z)$ be expanded in power series of p, each term being arranged in symmetrical form

$$H(p, q, z) = f(q)+[g(q)p+pg(q)]+\tfrac{1}{2}pF(q)p+\ldots. \tag{9.13}$$

Tildes on the first vectors of matrix products are omitted; no ambiguity should result from this simplification. f, g and F are arbitrary functions of q and z, and F is symmetrical. With the substitution $\nabla S \to \nabla$ (i.e. $p \to \partial/\partial x$, $\partial S/\partial z \to \partial/\partial z$) in (9.11a) and (9.11b) we get the two following differential equations for ψ and $\psi^\dagger$, respectively

$$\frac{\partial\psi}{\partial z} = \left\{-f(x)-\left[g(x)\frac{\partial}{\partial x}+\frac{\partial}{\partial x}g(x)\right]-\frac{1}{2}\frac{\partial}{\partial x}F(x)\frac{\partial}{\partial x}+\ldots\right\}\psi, \tag{9.14a}$$

$$\frac{\partial\psi^\dagger}{\partial z} = \left\{f(x)-\left[g(x)\frac{\partial}{\partial x}+\frac{\partial}{\partial x}g(x)\right]+\frac{1}{2}\frac{\partial}{\partial x}F(x)\frac{\partial}{\partial x}+\ldots\right\}\psi^\dagger. \tag{9.14b}$$

These expressions for $\partial\psi/\partial z$ and $\partial\psi^\dagger/\partial z$ can be introduced in (9.12). Rather than performing directly the integrations involved, let us note that each component of $\mathsf{p} \equiv \partial/\partial \mathsf{x}$ is an antisymmetrical operator. For $p_1 \equiv \partial/\partial x_1$ for instance, we have

$$(\psi^\dagger, p_1\psi)+(p_1\psi^\dagger, \psi) = \int\int (\psi^\dagger \frac{\partial}{\partial x_1}\psi + \psi \frac{\partial}{\partial x_1}\psi^\dagger)\mathrm{d}x_1\,\mathrm{d}x_2$$
$$= \int \mathrm{d}x_2 \int_{-\infty}^{+\infty} \frac{\partial}{\partial x_1}(\psi^\dagger\psi)\mathrm{d}x_1 = 0 \qquad (9.15)$$

because the wave functions vanish rapidly at infinity. An arbitrary function $f(q)$ of q, on the other hand, is symmetrical. Thus, terms such as $g(q)p+pg(q)$ which change sign with p are antisymmetrical operators, while terms such as $pF(q)p$ involving even powers of p are symmetrical operators. We now see readily, by considering each term in the expansion of H at a time, that reciprocity relation (9.12) is satisfied. Conservation of power could be demonstrated along similar lines by giving consideration to Hermitian products instead of simple products.

In conclusion, a physically acceptable scalar wave equation is obtained by the substitution $\nabla S \to \nabla$ in the eikonal equation provided the latter is arranged in symmetrical form. Note that a term such a $\frac{1}{2}\,pF(q)p$ could be written as well $\frac{1}{4}[p^2F(q)+F(q)p^2]$. This ambiguity is expected; we cannot hope to obtain a well-defined scalar wave equation because the components of the electric or magnetic field do not even satisfy the same wave equation when the medium lacks homogeneity. However, within the approximation of Gauss, F does not depend on q. The wave equation is therefore free of ambiguity in the case that we shall consider in detail.

§ 10. Hamiltonian Optics

The main concepts and results of Hamiltonian Optics are recalled in this section. For a detailed exposition the reader should see KLINE and KAY [1965]. The motivation for giving here these well-known results is to clarify the fact that they are, in general, applicable to complex spaces as well as to real spaces.

The eikonal equation (2.3) is most conveniently written

$$H(p_\alpha, q_\alpha) = 0, \qquad (10.1)$$

where $p_\alpha \equiv \partial S/\partial q_\alpha$, $\alpha = 1, 2, 3$; $q_3 \equiv z$. In general, the left-hand side of

(10.1) involves complex coefficients, and the eikonals S/ik are complex. Even in the case of lossless media, complex eikonals are of interest to account for diffraction effects. We assume that all relevant quantities are analytic functions of their arguments.

In order to solve (10.1) it proves useful to introduce complex rays $q_\alpha(t)$, where t denotes now an arbitrary real parameter, defined by the equation

$$dq_\alpha/dt = \partial H(p_\alpha, q_\alpha)/\partial p_\alpha. \tag{10.2}$$

These equations can be solved, in principle, for the p_α; we assume that the p_α are single valued functions of the q_β and dq_β/dt. Differentiating (10.1) totally with respect to q_β ($\beta = 1, 2, 3$) we get

$$\frac{\partial H}{\partial p_\alpha}\frac{\partial p_\alpha}{\partial q_\beta} + \frac{\partial H}{\partial q_\beta} = 0, \tag{10.3}$$

where the summation sign over repeated indices is omitted. Because $p_\alpha \equiv \partial S/\partial q_\alpha$, we have

$$\partial p_\alpha/\partial q_\beta = \partial p_\beta/\partial q_\alpha. \tag{10.4}$$

Thus, with the help of (10.2), (10.3) becomes

$$dp_\beta/dt = -\partial H/\partial q_\beta. \tag{10.5}$$

Equations (10.2) and (10.5) are Hamilton's equations for light rays. Together with some initial conditions, they completely define the complex ray trajectories.

Going back to the notation used in (9.11a) with $t \equiv z$, Hamilton equations (10.2), (10.5) are

$$\dot{q} = \partial \mathrm{H}/\partial \mathrm{p} \tag{10.6a}$$

$$\dot{p} = -\partial \mathrm{H}/\partial q \tag{10.6b}$$

where the upper dots denote differentiation with respect to z. These equations show that $dH/dz = \partial H/\partial z$. Thus, if the medium parameters are independent of z, H remains the same along a ray. The point-eikonal $S(r; r')$ is defined by

$$\mathrm{S}(q, z; q', z') = \int p_\alpha dq_\alpha = \int_{z'}^{z} [p\dot{q} - \mathrm{H}(p, q, z)]dz = \int_{z'}^{z} \mathrm{L}(\dot{q}, q, z)dz. \tag{10.7}$$

The quantity

$$\mathrm{L}(\dot{q}, q, z) \equiv p\,\dot{q} - \mathrm{H}(p, q, z), \tag{10.8}$$

where p is expressed as a function of $\dot{q}$, q and z, is called the Lagrangian. Eqs. (10.6a) and (10.6b) show that the ray trajectories can be obtained from the Lagrangian by means of

$$\partial L/\partial q = \dot{p}, \tag{10.9a}$$

$$\partial L/\partial \dot{q} = p. \tag{10.9b}$$

It is not difficult to show from the previous expressions that, for any small variation of path between r and r', $\delta S(r; r') = 0$. This is the mathematical expression of the well-known Fermat's principle. This principle can be used to evaluate the point-eikonal of a sequence of two optical systems whose individual point-eikonals $S_1(q''; q')$ and $S_2(q; q'')$, respectively, are known. It is obtained by addition

$$S(q; q') = S_2(q; q'') + S_1(q''; q'), \tag{10.10}$$

the intermediate variable q'' being eliminated by application of Fermat's principle:

$$\frac{\partial}{\partial q''}[S_2(q; q'') + S_1(q''; q')] = 0. \tag{10.11}$$

(10.11) can be solved, in principle, for q'', and the solution substituted in (10.10).

We are now in position to state the main result of this section: the Van Vleck propagator

$$G(r; r') = \pm \frac{1}{2\pi} |-\partial^2 S/\partial x_i \partial x'_j|^{\frac{1}{2}} \exp(S), \tag{10.12}$$

where the vertical bars denote a determinant, is the asymptotic form of the Green function of (9.14a). VAN VLECK [1928] has shown that if the right hand side of (10.12) is substituted in (9.14a) and the result ordered in decreasing powers of k, the two first terms vanish, in agreement with the correspondence principle. The physical interpretation of (10.12) is simple when the eikonal is real. The term exp (S) expresses the phase shift resulting from the optical length S/ik. The term in front of it, on the other hand, can be obtained from power conservation requirements. Clearly, the propagator for the transposed medium is obtained by changing the sign of S in (10.12).

We note further that if we limit ourselves to the terms exhibited in (9.14) (parabolic approximation) and assume that S is, at most, quadratic in x, x' (approximation of Gauss), all terms vanish, and (10.12) gives the exact expression for the Green function (see CHOQUARD [1955]). Because

only the approximation of Gauss is considered in subsequent sections, (10.12) can be used without restrictions. The field transformation may alternatively be expressed in terms of the mixed eikonal, as shown by WALTHER [1969].

§ 11. Properties of the Point-Eikonal

In order to make use of the expression of the Green function given in the previous section we must be able to evaluate the point-eikonal S/ik. This evaluation is in general easy to carry out because we limit ourselves to the approximation of Gauss. The algebraic details are displayed in this section.

11.1. RAY MATRICES

We have indicated in Section 10 that the point-eikonal of a sequence of two optical systems can be obtained by adding the individual point-eikonals and eliminating the intermediate variables with the help of Fermat's principle. It turns out that it is sometimes more convenient to characterize optical systems by ray matrices rather than by point-eikonals because the ray matrix of a sequence of optical systems can be obtained by matrix multiplication. In this subsection we indicate how the ray matrix is related to the point-eikonal.

Within the approximation of Gauss, S assumes the general form

$$S(q, z; q', z') = d + uq + u'q' + \tfrac{1}{2}\, qUq + qVq' + \tfrac{1}{2}q'Wq', \qquad (11.1)$$

where d is a scalar, u, u' are 2-vectors, U, W are 2×2 symmetric matrices, and V is a 2×2 matrix.

d, u, u', U, V and W are complex functions of z and z'. Equation (11.1) shows that an optical system is defined by 11 (possibly complex) constants. By definition we have

$$\begin{aligned} p &= \partial S/\partial q = u + Uq + Vq', \qquad (11.2)\\ -p' &= \partial S/\partial q' = u' + \tilde{V}q + Wq'. \end{aligned}$$

These relations can be written

$$\begin{bmatrix} q \\ p \\ 1 \end{bmatrix} = \begin{bmatrix} A & B & a \\ C & D & c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} q' \\ p' \\ 1 \end{bmatrix} \equiv \mathfrak{M} \begin{bmatrix} q' \\ p' \\ 1 \end{bmatrix}. \qquad (11.3)$$

By comparing (11.2) and (11.3) we readily find that

$$\begin{aligned} U &= DB^{-1}, & W &= B^{-1}A, \\ V &= C - DB^{-1}A, & u &= c - DB^{-1}a, \\ \tilde{V} &= -B^{-1}, & u' &= B^{-1}a. \end{aligned} \tag{11.4}$$

LUNEBURG [1944] relations are easily obtained from (11.4) by noting that U and W are symmetrical matrices. We have

$$\begin{aligned} A\tilde{B} &= B\tilde{A}, & \tilde{B}D &= \tilde{D}B, & D\tilde{C} &= C\tilde{D}, \\ \tilde{C}A &= \tilde{A}C, & \tilde{D}A - \tilde{B}C &= 1, & D\tilde{A} - C\tilde{B} &= 1, \\ & & |\mathfrak{M}| &= 1, \end{aligned} \tag{11.5}$$

and the inverse of $\mathfrak{M}$ is

$$\mathfrak{M}^{-1} = \begin{bmatrix} \tilde{D} & -\tilde{B} & \tilde{B}c - \tilde{D}a \\ -\tilde{C} & \tilde{A} & \tilde{C}a - \tilde{A}c \\ 0 & 0 & 1 \end{bmatrix}. \tag{11.6}$$

$\mathfrak{M}^{-1}$ is the matrix relating the input quantities $(q', p', 1)$ to the output quantities $(q, p, 1)$. Luneburg's relations (11.5) are effectively equivalent to six independent scalar equations.

An optical system can alternatively be described by the position and slope assumed at the output plane by five rays with given initial conditions. Let these rays be denoted $\bar{q}$, $\bar{p}$, Q, P and $Q^{\ddagger}$, $P^{\ddagger}$. The pairs of matrices (Q, P) and $(Q^{\ddagger}, P^{\ddagger})$ actually represent two rays each, because they are defined by the association of two ordinary rays: $Q \equiv [q_1 q_2]$ and $Q^{\ddagger} \equiv [q_1^{\ddagger} q_2^{\ddagger}]$. The ray matrix of the optical system can clearly be expressed in term of these rays

$$\mathfrak{M} \equiv \begin{bmatrix} A & B & a \\ C & D & c \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} Q^{\ddagger} & Q & \bar{q} \\ P^{\ddagger} & P & \bar{p} \\ 0 & 0 & 1 \end{bmatrix} \tag{11.7}$$

if the initial conditions for $Q^{\ddagger}$, Q and $\bar{q}$ are, respectively

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{11.8}$$

Note that, because the rays Q and $Q^{\ddagger}$ satisfy (11.5), PQ^{-1} and $Q^{-1}Q^{\ddagger}$ are symmetrical matrices.

We can now use (11.4) and (11.7) to obtain an expression of S in terms of the trajectories of five rays. We have

$$S(q, z; q', z') = d + (\bar{p} - PQ^{-1}\bar{q})q + \bar{q}\tilde{Q}^{-1}q' + \tfrac{1}{2}qPQ^{-1}q - q\tilde{Q}^{-1}q' + \tfrac{1}{2}q'Q^{-1}Q^{\dagger}q'. \quad (11.9)$$

It remains to express d in terms of $\bar{q}$. We have, from (11.9) and (10.7)

$$S(\bar{q}, z; 0, z') = d + \bar{p}\bar{q} - \tfrac{1}{2}\bar{q}PQ^{-1}\bar{q} = \int_{z'}^{z} L(\dot{\bar{q}}, \bar{q}, z)dz \quad (11.10)$$

where L denotes the Lagrangian. This relation provides the desired expression for d. Introducing (11.10) in (11.9), S can be written

$$S(q, z; q', z') = \int_{z'}^{z} L(\dot{\bar{q}}, \bar{q}, z)dz + \bar{p}(q - \bar{q}) + \tfrac{1}{2}(q - \bar{q})PQ^{-1}(q - \bar{q}) - (q - \bar{q})\tilde{Q}^{-1}q' + \tfrac{1}{2}q'Q^{-1}Q^{\dagger}q'. \quad (11.11)$$

Eq. (11.11) allows us to generalize (3.13) to the case of non-orthogonal systems. This expression sometimes proves more convenient to use than (11.1) because rays obey linear equations. The parameters $U, V, W\ldots$, on the other hand, obey nonlinear equations, as we shall see in the next subsection.

11.2. EVALUATION OF THE POINT-EIKONAL

Within the approximation of Gauss $\partial S/\partial z$ is, at most, quadratic in $p \equiv \partial S/\partial q$ and q. Thus, the Hamiltonian assumes the form

$$\partial S/\partial z = -H(p, q, z) = n + nq + \tfrac{1}{2}qNq - \tfrac{1}{2}(p - g - Gq)F(p - g - Gq), \quad (11.12)$$

where n, n, N, g, G and F are functions of z that can be obtained from knowledge of the material matrix $M(r)$. Hamilton equations (10.6a) and (10.6b) are, explicitly

$$\dot{q} = \partial H/\partial p = F(p - g - Gq), \quad (11.13a)$$

$$\dot{p} = -\partial H/\partial q = n + Nq + \tilde{G}F(p - g - Gq). \quad (11.13b)$$

Using for p the expression given in (11.13a) we find the Lagrangian

$$L(\dot{q}, q, z) \equiv p\dot{q} - H = n + nq + \tfrac{1}{2}qNq + q(g + Gq) + \tfrac{1}{2}\dot{q}F^{-1}\dot{q}, \quad (11.14a)$$

and

$$\int_{z'}^{z} L\,dz = \int_{z'}^{z} n\,dz + \frac{1}{2}\left[q(z)p(z) - q(z')p(z') + \int_{z'}^{z}(nq + g\dot{q})dz\right], \quad (11.14b)$$

a generalization of (3.7).

Let us now introduce the expression (11.1) for S in (11.12) and compare

terms that have the same powers in q and q'. We get a set of ordinary differential equations for U, V, W, u, u' and d which reads

$$\begin{aligned} U+(U-G)F(U-G) &= N, & \dot{u}+(U-G)F(u-g) &= n, \\ \dot{V}+(U-G)FV &= 0, & u'\tilde{V}F(u-g) &= 0, \\ \dot{W}+\tilde{V}FV &= 0, & \dot{d}+\tfrac{1}{2}(u-g)F(u-g) &= n. \end{aligned} \tag{11.15}$$

These equations are most easily solved in the order shown above because each equation involves only the solutions of previous equations. Note that these equations are of first order; the solutions are therefore uniquely defined by the values assumed by d, u, u', U, V, W at some plane.

Because the differential equations (11.15) are nonlinear it is sometimes simpler to use the form (11.11) of the point-eikonal. In this equation $\bar{q}$, $\bar{p}$ are solutions of the ray equations (11.13) with the initial conditions $\bar{q}(0) = 0$, $\bar{p}(0) = 0$. Q, P obey the ray equation without misalignment terms, namely

$$\dot{Q} = F(P-GQ), \qquad \dot{P} = NQ+\tilde{G}F(P-GQ) \tag{11.16}$$

with the initial conditions $Q(0) = 0$, $P(0) = 1$. $Q^{\ddagger}$ and $P^{\ddagger}$ are solutions of (11.16) with the initial conditions $Q^{\ddagger}(0) = 1$, $P^{\ddagger}(0) = 0$.

Assuming now that the medium parameters are independent of z, we obtain by differentiation of (11.13a) and use of (11.13b) the ray equation

$$F^{-1}\ddot{q}+(G-\tilde{G})\dot{q}-Nq = n. \tag{11.17}$$

Many optical systems incorporate only homogeneous media. When the medium is homogeneous, $G = 0$, $N = 0$ and $n = 0$. Thus (11.17) reduces to $\ddot{q} = 0$, the rays being (complex) straight lines. The point-eikonal is obtained by multiplying the right-hand side of (11.14a) by $z-z'$

$$S(q, z; q', z') = (z-z')(n+\dot{q}g+\tfrac{1}{2}\dot{q}F^{-1}\dot{q}) \tag{11.18}$$

where

$$\dot{q} = (q-q')/(z-z').$$

It remains to obtain the effective optical thickness associated with curved surfaces of discontinuity. We shall assume for simplicity that the discontinuity is not really abrupt, but tapered in such a way that the law of adiabatic invariance of the polarization is obeyed. This matching layer is therefore assumed to be thick compared with $(\lambda_o^{-1}-\lambda_e^{-1})^{-1}$, where λ_o and λ_e denote the ordinary and extraordinary wavelengths, yet thin from a ray-optics point of view. Let us first assume that the surfaces of discontinuity are plane and that the z-axis is a ray of the system. At each plane of refraction the position of a ray can be defined by its distance x_2' from the

z-axis in the plane of incidence and its position x_1' in the direction perpendicular to the plane of incidence. Such a ray reaches the plane of refraction at $y_1 = x_1'$, $y_2 = x_2'/\cos i'$, if i' denotes the angle of incidence (see Fig. 17). Let us now go back to the situation often encountered where the surface of discontinuity is curved. Let the equation of the surface of discontinuity be denoted $\zeta(y_1, y_2)$, the axis ζ being perpendicular to the reference surface. The effective optical thickness resulting from this deformation of the surface is $(H-H')\zeta \equiv (\partial S'/\partial\zeta - \partial S/\partial\zeta)\zeta$. In the special case of isotropic media for instance, we have (see LUNEBURG [1944] p. 90) $H' = -ikn' \times \cos(i')$, $H = -ikn\cos(i)$, where n', n denote the refractive indices in the

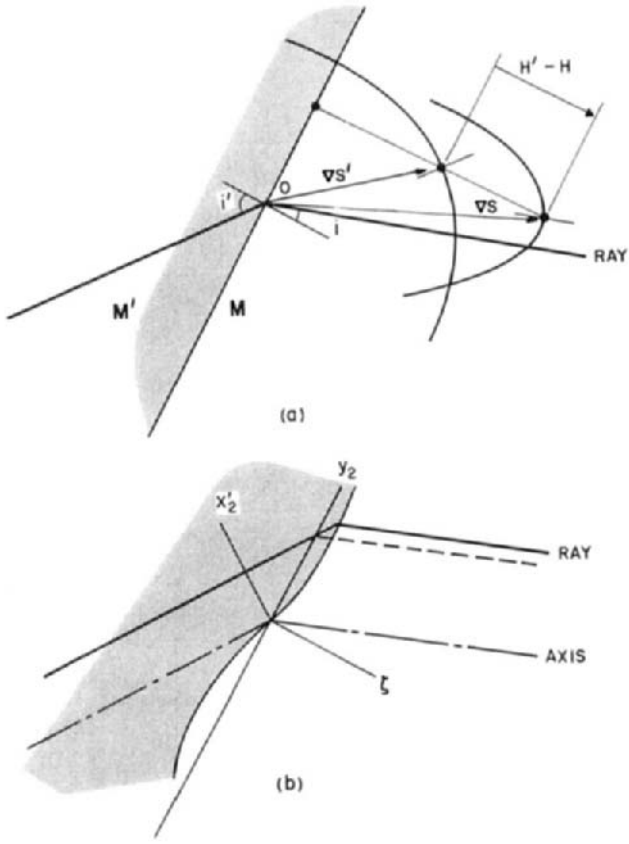


Fig. 17. Evaluation of the effective optical thickness associated with a deformation of the surface of discontinuity between two homogeneous media. (a) Refraction by a plane interface. $\nabla S'$ and ∇S describe the surfaces of wave normals for the two media, the origin being at point O. (b) A displacement ζ of the surface introduces an effective optical thickness $(H-H')\zeta$.

first and second media, respectively. The rest of the calculation consists in performing first-order expansions consistent with the approximation of Gauss. This procedure provides closed form expressions for the constants d, u, u', U, V, W introduced in (11.1) in the case of arbitrary sequences of curved surfaces of discontinuity.

§ 12. Beam Modes

We are now in position to give a general expression for the field of beam modes. Substituting the expression (11.1) given for S in the Van Vleck propagator (10.12), the Green function is found to be

$$G(x, z; q', z') = \frac{1}{2\pi} |V|^{\frac{1}{2}} \exp\left(d + ux + u'q' + \tfrac{1}{2}xUx + xVq' + \tfrac{1}{2}q'Wq'\right). \quad (12.1)$$

Note that, because V is in general complex and varies with z, the determinant $|V|$ contributes to the variation of both the phase and amplitude of G.

We define now the modes of propagation as the coefficients of the expansion of $G(r; q', z')$ in power series of q_1' and q_2'. Explicit expressions are easily obtained if we recall the definition of Hermite polynomials in terms of their generating functions

$$\exp\left(\xi q' + \tfrac{1}{2}q'Wq'\right) = \sum_{m_1, m_2 = 0}^{\infty} \frac{q_1'^{m_1}}{m_1!} \frac{q_2'^{m_2}}{m_2!} \mathscr{H}_{m_1 m_2}(\xi; W), \quad (12.2)$$

where ξ denotes a vector. Explicitly we have (ARNAUD [1971])*

$$\mathscr{H}_{m_1 m_2}(\xi; W) = \sum_{\alpha, \beta, \gamma = 0}^{\text{exp.}} \frac{m_1! m_2! W_{11}^{\alpha} W_{22}^{\beta} W_{12}^{\gamma - \alpha - \beta} \xi_1^{m_1 - \gamma - \alpha + \beta} \xi_2^{m_2 - \gamma - \beta + \alpha}}{2^{(\alpha+\beta)} \alpha! \beta! (\gamma - \alpha - \beta)! (m_1 - \gamma - \alpha + \beta)! (m_2 - \gamma - \beta + \alpha)!}, \quad (12.3)$$

where exp. means that the series terminates when one of the exponents becomes equal to zero. Notice that, for any diagonal matrix D with elements λ_1, λ_2 we have

$$\mathscr{H}_{m_1 m_2}(D\xi; DWD) = \lambda_1^{m_1} \lambda_2^{m_2} \mathscr{H}_{m_1 m_2}(\xi; W), \quad (12.4)$$

a relation which readily results from the definition (12.2).

Let us make use of these mathematical results, setting $\xi \equiv u' + \tilde{V}x$ and expanding the right-hand side of (12.1) in power series of q_1', q_2'. The coefficients are, to within unimportant factors,

$$\psi_{m_1 m_2}(x, z) = |V|^{\frac{1}{2}} \exp\left(d + ux + \tfrac{1}{2}xUx\right) \mathscr{H}_{m_1 m_2}(u' + \tilde{V}x; W), \quad (12.5)$$

* Because a minus sign has been dropped for simplicity on the left-hand side of (12.2), the polynomial $\mathscr{H}$ differs slightly from a Hermite polynomial.

or, alternatively, in terms of the rays $\bar{q}$, Q and $Q^{\ddagger}$ [see (11.11)]

$$\psi_{m_1m_2}(x, z) = |Q|^{-\frac{1}{2}} \exp\left[\bar{p}(x-\bar{q})+\tfrac{1}{2}(x-\bar{q})PQ^{-1}(x-\bar{q})\right] \times \mathscr{H}_{m_1m_2}[Q^{-1}(x-\bar{q}); Q^{-1}Q^{\ddagger}] \exp\left[\int^{z} L(\dot{\bar{q}}, \bar{q}, z)dz\right]. \quad (12.6)$$

These expressions give the modes of propagation in their most general form.* The initial values of the parameters $\bar{q}$, $\bar{p}$, Q, P and $Q^{\ddagger}$, $P^{\ddagger}$ are found by comparing (12.6) with the incident field at $z = 0$. Eq. (12.4) exhibits the fact that there is some arbitrariness in the determination of these parameters. This arbitrariness can be removed, if desired, by introducing normalization conditions such as those given in (11.8). Note incidentally that this normalization condition (11.8) differs by a complex numerical factor from the normalization condition used in Section 4. Once the initial values of the rays have been obtained, the field at any point in space is obtained by solving the ray equations (11.13) or (11.17) and substituting in (12.6).

§ 13. Conclusion

We have shown that the problem of finding the modes of propagation in unaberrated systems can be reduced to a routine application of the methods of Hamiltonian Optics because the continuation in the complex plane that is needed to take into account diffraction effects does not involve any formal change in the algebra. The approach to the theory of modes discussed in detail in this paper is based on a power series expansion of the Green function. This approach is simple, both conceptually and algebraically, even in the case of optical systems incorporating astigmatic and anisotropic media; it requires only the use of matrix algebra.

Because of space limitation, application of these general results to particular optical systems has not been made. Also, a few important topics pertaining to the theory of beam modes had to be left out: Mode orthogonality, mode coupling, and a general theory of optical resonators.

For a generalization of the results discussed in this paper one should take into account aberration effects, perhaps with the help of a perturbation method. One should also try to account more fully for the vectorial character of the optical field and the coupling between the local eigenstates of polarization. Many optical systems have already been analyzed with due consideration to such effects. New significant results, however, will probably appear in the future in this relatively new field.

* For a comparison with the results given in Section 3, all lengths should be multiplied by the factor: ik.

References

ARNAUD, J. A., 1969, Appl. Opt. **8**, 189 and 1090.

ARNAUD, J. A. and H. KOGELNIK, 1969, Appl. Opt. **8**, 1687.

ARNAUD, J. A., 1970a, Appl. Opt. **9**, 1192.

ARNAUD, J. A., 1970b, Bell Syst. Tech. J. **49**, 2311.

ARNAUD, J. A., 1971a, J. Opt. Soc. Am. **61**, 751.

ARNAUD, J. A., 1971b, Proc. IEEE Letters **59**, 1378.

ARNAUD, J. A., 1972a, J. Opt. Soc. Am. **62**, 290.

ARNAUD, J. A., 1972b, Appl. Opt. **11**, 2514.

BAUES, P., 1969, J. Opto-Electronics **1**, 37 and 103.

BERGMANN, E. E., 1972, Appl. Opt. **11**, 113.

BERGSTEIN, L. and T. H. ZACHOS, 1966, IEEE J. Quant. Elect. **QE-2**, 677.

BERREMAN, D. W., 1965, Bell Syst. Tech. J. **44**, 2117.

BERTONI, H. L., L. B. FELSEN and A. HESSEL, 1971, IEEE Trans. on Antennas and Propagation **AP-19**, 226.

BHAWALKAR, D. D., A. M. GONCHARENKO and R. C. SMITH, 1967, Brit. J. Appl. Phys. **18**, 1431.

BOOKER, H. G. and W. WALKINSHAW, 1946, in: Meteorological Factors in Radio Wave Propagation (London, Physical Society) pp. 80-127.

BORN, M. and E. WOLF, 1965, Principles of Optics (Pergamon Press, Oxford).

BOYD, G. D. and J. P. GORDON, 1961, Bell Syst. Tech. J. **40**, 489.

BOYD, G. D. and H. KOGELNIK, 1962, Bell Syst. Tech. J. **41**, 1347.

BROUWER, W. and A. WALTHER, 1967, in: Advance Techniques in Optical Design, ed. A. C. S. Van Heel (North-Holland, Amsterdam).

BUDDEN, K. G., 1961, The Wave-Guide Mode Theory of Wave Propagation (Logos Press/ Prentice Hall, Inc., London).

BUDDEN, K. G. and P. D. TERRY, 1970, Proc. Roy. Soc. London A **321**, 275.

BYKOV, V. P. and L. A. VAINSHTEIN, 1965, Soviet Phys. JETP **20**, 338.

CHERNIKOV, N. A., 1968, Soviet Phys. JETP **26**, 603.

CHOQUARD, P., 1955, Helv. Phys. Acta **28**, 89.

COLLINS, S. A., 1964, Appl. Opt. **3**, 1263.

COLLINS, S. A., 1970, J. Opt. Soc. Am. **60**, 1168.

DEMKOV, Y. N. and V. N. OSTROVSKII, 1971, Soviet Phys. JETP **13**, 1083.

DESCARTES, R., 1637, Dioptrique, see: Oeuvres Complètes, Coll. la Pléiade (Gallimard, Paris, 1963) pp. 188–199.

DESCHAMPS, G. and MAST, 1964, Beam Tracing and Applications, in: Proc. Symp. on Quasi-Optics (Polytechnic Press, New York) p. 379.

DESCHAMPS, G., 1968, U.R.S.I. Symp. on Electromagnetic Theory, Stresa (unpublished).

DESCHAMPS, G., 1972, Proc. IEEE. **60**, 1022.

DI FRANCIA, G. T., 1964, in: Symp. on Quasi-Optics, Polytech. Inst. Brooklyn.

DI FRANCIA, G. T., 1966, Opt. Acta **13**, 323.

EICHMANN, G., 1971, J. Opt. Soc. Am. **61**, 161.

ERMERT, H., 1971, A.E.U. **25**, 17.

FELDMAN, M., 1971, J. Opt. Soc. Am. **61**, 446.

FEYNMAN, R. P., 1948, Rev. Modern Phys. **20**, 367.

FOX, A. G. and T. LI, 1960, Proc. Inst. Radio Engrs. **48**, 1904.

FRIEDMAN, B., 1960, Principles and Techniques of Applied Mathematics (Wiley, London).

GLOGE, D. and D. MARCUSE, 1969, J. Opt. Soc. Am. **59**, 1629.

GOL'DMAN, I. I. and V. D. KRIVCHENKOV, 1957, Problems in Quantum Mechanics (Moscow, Transl.: Pergamon Press, Oxford, 1961).

GORDON, J. P., 1966, Bell Syst. Tech. J. **45**, 321.

GOUBAU, G. and F. SCHWERING, 1961, I.R.E. Trans. on Antennas and Propagation **AP-9**, 248.
GOUBAU, G., 1969, in: Millimetre and Submillimetre Waves, ed. F. A. Benson (Iliffe Books Ltd., London) p. 403.
GOUDET, G. and C. MEULEAU, 1957, Les Semiconducteurs (Eyrolles, Paris).
HAMILTON, W. R., 1835, see: Mathematical Papers (Cambridge University Press, 1941).
HARVEY, A. F., 1970, Coherent Light (Wiley-Interscience, London).
HEER, C. V., 1964, Phys. Rev. **134A**, 799.
HERTZ, J. H. and G. MINKWITZ, 1969, Optica Acta **16**, 593.
JONES, A. V. and G. J. PAPADOPOULOS, 1972, J. Phys. A: Gen. Phys., L86.
KAHN, W. K., 1965, Appl. Opt. **4**, 758.
KAHN, W. K. and J. T. NEMIT, 1967, in: Proc. Symp. on Modern Optics, ed. J. Fox (Polytechnic Press, Brooklyn, N.Y.).
KELLER, J. B., 1958, in: Proc. Symp. Applied Mathematics, Vol. 8 (McGraw-Hill, New York) p. 27.
KELLER, J. B. and S. J. RUBINOW, 1960, Ann. Phys. **9**, 24.
KELLER, J. B. and W. STREIFER, 1971, J. Opt. Soc. Am. **61**, 40.
KLINE, M. and I. W. KAY, 1965, Electromagnetic Theory and Geometrical Optics (Wiley-Interscience, New York) Ch. 3.
KOGELNIK, H., 1965a, Bell Syst. Tech. J. **44**, 455.
KOGELNIK, H., 1965b, Appl. Opt. **4**, 1562.
KOGELNIK, H. and T. LI, 1966, Proc. IEEE **54**, 1312.
KONG, J. A., 1970, Proc. IEEE **58**, 1966.
KOPPELMAN, G., 1967, in: Progress in Optics, Vol. VII, ed. E. Wolf (North-Holland, Amsterdam).
KRAVTSOV, Y. A., 1965, Soviet Radiophys. **8**, 659.
KRAVTSOV, Y. A., 1967, Complex Rays and Complex Caustics, 4th All-Union Symp. on Diffraction of Waves, Kharkov (Transl. Bell Tel. Lab., Oct. 1969).
KURAUCHI, N. and W. K. KAHN, 1966, Appl. Opt. **5**, 1023.
KURTZ, C. N. and W. STREIFER, 1969, IEEE Trans. on Microwave and Tech. **MTT-17**, 11.
LANDAU, L. and E. LIFSHITZ, 1960, Mechanics (Addison-Wesley, Reading).
LANDAU, L. and E. LIFSHITZ, 1961, The Classical Theory of Fields (Addison-Wesley, Reading).
LEONTOVICH, M. A. and V. A. FOCK, 1946, Zh. Eksperim. i Teor. Fiz. **16**, 557.
LUDWIG, D., 1966, Comm. Pure and Appl. Math. **19**, 215.
LUNEBURG, R. K., 1944, Mathematical Theory of Optics, Lectures at Brown University (California Press, Los Angeles, 1964).
MARCATILI, E. A. J., 1964, Bell Syst. Tech. J. **43**, 2887.
MARIÉ, P., 1970, Ann. Télécom. **25**, 320.
MASON, I. M., 1971, Electron. Letters **7**, 344.
MESSIAH, A., 1961, Quantum Mechanics (Noth-Holland, Amsterdam) pp. 218 and 446.
MILDER, D. M., 1969, J. Acoust. Soc. Am. **46**, 1259.
PIERCE, J. R., 1954, Theory and Design of Electron Beams (Van Nostrand, New York).
PIERCE, J. R., 1961, Proc. Natl. Acad. Sci. **47**, 1808.
POLE, R. V., 1965, J. Opt. Soc. Am. **55**, 254.
POPOV, M. M., 1968, Opt. and Spectr. **25**, 170 and 213.
POPOV, M. M., 1969a, Vestn. Leningr. Univ. **22**, 42.
POPOV, M. M., 1969b, Dokl. Akad. Science USSR **184**, 1076.
RAMSAY, J. A. and J. J. DEGNAN, 1970, Appl. Opt. **9**, 385.
RAYLEIGH, Lord, 1894, Theory of Sound, Vol. 1 (London; Dover Publications, New York, 1945).
RYTOV, S. M., 1937, Compt. Rend. (Doklady) Ac. Sc. URSS **18**, 263.
SCHACHTER, and H. C. CHANGHWI, 1970, Opt. Acta **17**, 801.

SCHAEDLA, W. H. and J. B. BEYER, 1968, IEEE Trans. on Antennas and Propagation **AP-16**, 108.
SIEGMAN, A. E., 1965 Proc. IEEE **53**, 277.
STEIER, W. H., 1966, Appl. Opt. **5**, 1229.
SUEMATSU, Y. and H. FUKINUKI, 1965, I.E.C.E. **48**, 64.
SUEMATSU, Y. and H. FUKINUKI, 1968, Bull. Tokyo Inst. Technol. **88**, 33.
SUEMATSU, Y., 1971, Japan J. Appl. Phys. **10**, 1060.
SYNGE, J. L., 1954, Geometrical Mechanics and DeBroglie Waves (Cambridge at the Univ. Press).
TANAKA, T., M. SUZUKI and T. MATSUMOTO, 1970, Elec. and Comm. in Japan **53-B**, 60.
TIEN, P. K., J. P. GORDON and J. R. WHINNERY, 1965, Proc. IEEE **53**, 129.
UNGER, H. G., 1965, Arch. Elektr. Ubertrajung **19**, 189.
VAINSHTEIN, L. A., 1965, Soviet Phys. JETP **18**, 471.
VAINSHTEIN, L. A., 1966, Open Resonators and Open Waveguides (Transl. Golem Press, Boulder, Colorado, 1969).
VAKHIMOV, N. G., 1964, 3rd All Union Symp. on Wave Diffraction, Tiflis [Radio Eng. Electron Phys. **10** (1965) 1439].
VAN VLECK, J. H., 1928, Proc. Natl. Acad. Sci. (U.S.) **14**, 178.
VLASOV, S. N. and V. I. TALANOV, 1965, Izv. Vysshik Uchebn. Zavedenii Radiofiz. **8**, 195.
WALTHER, A, 1969, J. Opt. Soc. Am. **59**, 1325.
WUNSCHE, A., 1970, Ann. Phys. **25**, 113.

E. WOLF, PROGRESS IN OPTICS XI © NORTH-HOLLAND 1973

VII

GRADIENT INDEX LENSES

BY

E. W. MARCHAND

Eastman Kodak Research Laboratories, Kodak Park, Rochester, N.Y. 14650, U.S.A.

CONTENTS

§ 1. Historical Background

1.1. INTRODUCTION

Inhomogeneous or gradient-index media have long been of interest to optical workers, and rightly so, since these media occur frequently in nature. The lens of the human eye and the atmosphere of the earth are familiar examples.

It is known that gradient-index media offer attractive theoretical possibilities in the design of optical instruments. For example, MAXWELL [1854] showed that an inhomogeneous medium of a certain type has the property of forming a sharp image of every point in its interior. He thus demonstrated that an optical instrument that is perfect (in a limited sense) is theoretically possible.

WOOD [1905] described a method of forming a circular slice of gelatin in which the refractive index is a function of the distance from the center. He thus showed that, with the help of a gradient-index medium, it is possible for a plane-parallel plate to act like a lens. More recently, LUNEBURG [1964] explored mathematically the possibilities of using inhomogeneous media in optical systems.

Until recent times these and other similar examples were regarded by lens designers as mathematical curiosities since there seemed to be no way to fabricate gradient-index lens elements for practical applications. Within the last few years, however, some advances in the pertinent technology have been made. These developments have created a new interest in both theoretical and practical aspects of inhomogeneous media, and a number of papers on this subject have appeared recently.

One promising application of the new technology consists of the so-called GRIN (gradient-index) rods. A thin glass rod can be treated so as to form an index gradient in the rod, the index of refraction then being a decreasing function of the distance from the axis of the rod. Such a rod has a focusing effect on rays entering one end, and images can be transmitted along the rod.

A second application, possibly of greater importance, consists of the use of classical-type lens elements made with inhomogeneous media. From the

mathematical viewpoint, such elements provide the lens designer with several more design parameters, namely, the parameters required to specify the gradient functions. The new degrees of freedom should make it possible, in many cases, to match the quality of a given instrument with fewer lens elements, thus reducing space, weight, and possibly cost.

1.2. MAXWELL'S FISHEYE

Optical theoreticians have long been interested in the possibility of a perfect instrument. An instrument, for example, is called "absolute" if every point of a three-dimensional region is imaged stigmatically (sharply). This means that all rays from a single point of this region (object point) converge to a single, corresponding image point.

MAXWELL [1854] showed that an absolute instrument is possible in principle, and, in fact, proved that an inhomogeneous medium having an index function of the form

$$n(r) = \frac{n_0}{1+(r/a)^2} \tag{1.1}$$

has the desired property. Here r represents the distance from a fixed point O, and n_0 and a are constants. This index distribution evidently has spherical symmetry about a point, and a lens described by eq. (1.1) is known as a Maxwell fisheye lens. Rays in such a medium follow plane curves each in the form of a circular arc, as shown, for example, by HERZBERGER [1958] and BORN and WOLF [1970].

The fisheye lens is perfect only in a limited sense. Although every object point is sharply imaged, the images of extended objects suffer severely from various aberrations such as distortion. Furthermore, only points within the lens itself are sharply imaged.

1.3. LUNEBURG LENS

LUNEBURG [1964] solved the problem of imaging stigmatically all points infinitely distant from a certain lens. This lens, however, does not represent an absolute instrument since the points at infinity can be regarded only as points of a surface rather than a region of three-dimensional space.

The Luneburg lens consists of an inhomogeneous medium indicated by the index function

$$n(r) = (2-r)^{\frac{1}{2}}, \tag{1.2}$$

where again r is the distance from the origin. The region around the lens is assumed to have an index of unity. Every parallel bundle of rays is imaged sharply.

It is doubtful if such a lens can be fabricated for use in visible light since the index given by eq. (1.2) must vary from 0 to $\sqrt{2}$. However, the Luneburg lens has proven useful in applications to radar.

1.4. WOOD LENS

WOOD [1905] showed how to construct a simple lens in the form of a plane parallel plate of gelatin. His procedure was to use a dipping technique to make a cylinder of gelatin in which the refractive index is a function of the distance from the axis. Then slicing the cylinder by cuts perpendicular to the axis results in a number of so-called Wood lenses.

A typical gradient in a Wood lens is of the form

$$n(r) = N_0 + N_1 r^2, \tag{1.3}$$

where now r represents the distance from the axis of rotational symmetry. If the thickness d of the slice is small, it is found that the Wood lens with index given by eq. (13) has a focal length of

$$f' = -1/(2N_1 d), \tag{1.4}$$

approximately. This result shows that the plate acts like a converging or diverging lens according to whether $N_1 \gtrless 0$.

Wood mentions that Schott has prepared glass lenses of this type by pouring molten glass into an iron tube and then suddenly chilling the tube on the outside. The tension in the glass cylinder creates a radial variation in the refractive index. Plane parallel plates can then be obtained by cutting the glass cylinder into slices.

§ 2. Recent Developments

For many years glass technologists have experimented with the effects of diffusing ions into glass in order to modify the refractive index. Only recently, however, was it found that a controlled variation in the index could be produced by ion diffusion in certain types of glass in such a way as to leave the glass strain-free and colorless (HAMBLEN [1969] and PEARSON, FRENCH and RAWSON [1969]).

By the ion-diffusion technique an inhomogeneous layer can be formed just below the surface of the glass and extending several millimeters beneath the surface. The index may be either an increasing or a decreasing function of depth, and index variations of 0.08 or more can be achieved.

The above method of treatment lends itself to the fabrication of the so-called GRIN rods mentioned in § 1 provided the diameters of the rods are

small. If a rod is cylindrical and the ion diffusion is applied to the curved surface, it is evident that the resulting index will be a function of the radial distance from the axis. A gradient with this type of symmetry is called a cylindrical or radial gradient. The Wood lens is another example involving a gradient of this type.

Another potential use of the ion diffusion method consists of the treatment of any conventional type of lens element on one side (or both) so as to form a gradient-index layer. If the surface treated is spherical, the index becomes a function of the perpendicular distance from the surface or a function of the radial distance from the center of curvature. The index function is then said to have spherical symmetry and is referred to as a "spherical gradient". Maxwell's fisheye and the Luneburg lens are examples of this type.

A surface containing a spherical gradient can be ground and polished, and, after this is done, the physical surface may no longer be concentric with the spheres of constant index. Thus considerable latitude is possible in constructing lens elements of various types.

In the limiting case in which the center of symmetry of the spherical gradient is at infinity, the index function is simply a function of the distance from a reference plane. Such a gradient (axial gradient) can evidently be produced by using the ion-diffusion technique starting with a plane glass surface. When an axial gradient element is used as part of an optical system, there is some evidence to indicate that the effect is similar to the presence of an aspheric surface. MOORE [1971] has found that the third-order contributions to the aberrations of the system are, indeed, the same as if one replaced the axial gradient by an aspheric surface.

At present it is not clear how useful spherical gradients will be in practical lens design. The fact that the gradient layer produced by ion diffusion extends only a small distance below the surface does appear to be an important limitation. The question then arises whether a cylindrical gradient can be produced with index variations extending over a large distance from the optical axis and being effective through the entire thickness of the lens element. If something like a Wood lens could be made with glass, the desired result would be achieved.

SINAI [1971] describes a method by which the above can be accomplished by irradiation of glass with high-energy neutrons. By suitable masking, the radiation dose can be controlled over selected regions of the lens surface, so that a two-dimensional index distribution can be produced within the lens according to desired specifications. In particular, a cylindrical gradient with a prescribed profile can be obtained. Sinai has shown that a single-

element lens produced in this manner may exhibit substantially lower spherical aberration than comparable homogeneous lenses.

Thus it appears that at least two methods of fabricating gradient-index optical elements are known. Admittedly, these methods are somewhat difficult and costly at present. But, doubtless, easier and cheaper methods will be found. Already the developments cited above have generated renewed interest in the design of optical instruments involving inhomogeneous elements.

§ 3. Ray Tracing

3.1. GENERAL NUMERICAL RAY TRACE

In order to make use of gradient index elements in lens design (or in GRIN rods), the first step is to prepare a suitable method of tracing rays in inhomogeneous media. The basic theory for doing this is well known, the differential equation of the ray paths being given in the form (BORN and WOLF [1970])

$$(n\,\boldsymbol{r}')' = \nabla n, \tag{3.1}$$

where $\boldsymbol{r}$ is the variable position vector for points on the ray, ∇n is the gradient of the refractive index n, and a prime denotes differentiation with respect to the arc length s.

In principle, eq. (3.1) determines the paths of all the rays. However, this equation, which actually represents three coupled scalar differential equations of second order, cannot be solved analytically in general. A computer routine can be devised for solving for the ray paths when the index function $n(x, y, z)$ is known and the starting data for each ray are given. MONTAGNINO [1968] has suggested the following method for doing this.

Consider the Taylor series expansion of $\boldsymbol{r}$ in terms of the arc length s in the neighborhood of a starting point P_0, i.e.,

$$\boldsymbol{r}(s) = \boldsymbol{r}(s_0)+\boldsymbol{r}'(s_0)\Delta s+\tfrac{1}{2}\boldsymbol{r}''(s_0)(\Delta s)^2+ \ldots . \tag{3.2}$$

Likewise, the expansion of the derivative $\boldsymbol{r}'(s)$ is

$$\boldsymbol{r}'(s) = \boldsymbol{r}'(s_0)+\boldsymbol{r}''(s_0)\Delta s+ \ldots . \tag{3.3}$$

As usual in differential geometry

$$\boldsymbol{t} = \boldsymbol{r}', \qquad \boldsymbol{K} = \boldsymbol{r}'' = \boldsymbol{t}', \tag{3.4}$$

where $\boldsymbol{t}$ is the unit tangential vector and $\boldsymbol{K}$ is the curvature vector of the ray path. Thus eq. (3.3) can be written

$$\boldsymbol{t}(s) = \boldsymbol{t}(s_0)+\boldsymbol{K}(s_0)\Delta s+ \ldots . \tag{3.5}$$

To obtain a formula for $\boldsymbol{K}$ we write eq. (3.1) in the form

$$n'\boldsymbol{t}+n\boldsymbol{K} = \nabla n. \tag{3.6}$$

We then apply scalar multiplication by $\boldsymbol{t}$, noting that

$$\boldsymbol{t}^2 = 1, \qquad \boldsymbol{t}\cdot\boldsymbol{t}' = \boldsymbol{t}\cdot\boldsymbol{K} = 0, \tag{3.7}$$

which gives

$$n' = \boldsymbol{t}\cdot\nabla n. \tag{3.8}$$

Hence, from eq. (3.6),

$$\boldsymbol{K} = [\nabla n - \boldsymbol{t}(\boldsymbol{t}\cdot\nabla n)]/n. \tag{3.9}$$

The numerical computation then proceeds as follows. Assuming $\boldsymbol{r}(s_0)$ and $\boldsymbol{t}(s_0) = \boldsymbol{r}'(s_0)$ known at the starting point P_0, calculate $\boldsymbol{K}$ at P_0 from eq. (3.9). Select an arbitrary small value for Δs and compute $\boldsymbol{t}(s)$ from eq. (3.5) using only two terms on the right-hand side. Also calculate $\boldsymbol{r}(s)$ from eq. (3.2) using only two terms on the right-hand side. By this procedure approximate values of $\boldsymbol{r}$ and $\boldsymbol{t}$ are found for a point P a small distance Δs away from the starting point P_0. Iterating this routine determines (approximately) a succession of points of the ray and also the direction of the ray at each point.

Ray tracing by the numerical method outlined above can be done effectively with a high-speed computer. By taking small values of Δs and carrying many significant figures, very accurate results can be obtained. Furthermore, this approach is essentially independent of any symmetry in the function $n(x, y, z)$.

On the other hand, when possible, it is desirable to take account of special symmetries. In fact, in certain cases, exact analytical solutions can be found and, in other cases, simplified ray-tracing equations can be devised so as to reduce the time of computation without loss of accuracy. Also, analytic or semi-analytic ray-tracing formulas can be helpful in describing the imaging properties of the rays. We shall therefore consider some alternative ray-tracing methods.

3.2 SPHERICAL MEDIUM

If the index n is a function ot the distance r from a fixed point (spherical gradient) the solution of the differential equation (3.1) can be reduced to the evaluation of a single quadrature (LUNEBURG [1964]). For use in practical lens design, it is important to arrange the tracing formulas in a form which is convenient for use in analyzing conventional-type optical systems modified by gradients (MARCHAND [1970]).

It is natural to introduce spherical coordinates (r, θ, ϕ) with origin O at the center of symmetry of the index function and with polar axis along the line OP_0, where P_0 is the starting point of the ray (see Figs. 1 and 2). Then Fermat's principle requires stationarity of the integral

$$L = \int_{s_0}^{s} n \, ds = \int_{r_0}^{r} F \, dr, \tag{3.10}$$

where

$$F = n(r)(1 + r^2\dot{\theta}^2 + r^2 \sin^2\theta \, \dot{\phi}^2)^{\frac{1}{2}}, \tag{3.11}$$

the dot indicating differentiation with respect to r.

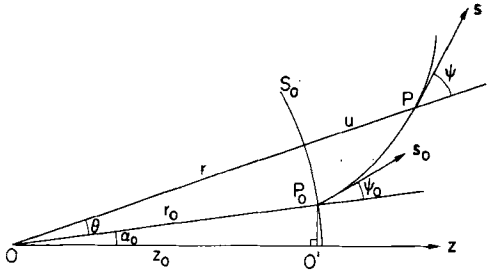


Fig. 1. Ray in a concentric medium (case with $z_0 > 0$). Figure is schematic in that planes OP_0P and $OO'P_0$ need not coincide.

Since $\partial F/\partial\phi = 0$, the first Euler equation reduces to

$$nr^2 \sin^2\theta \, \dot{\phi}(1 + r^2\dot{\theta}^2 + r^2 \sin^2\theta \, \dot{\phi}^2)^{-\frac{1}{2}} = \text{const.}$$

However, at P_0, $\theta = 0$, showing that the constant is 0. For a nonradial ray, $\sin\theta \neq 0$, except at P_0, and hence $\dot{\phi} \equiv 0$ along the ray. This verifies the well-known fact that all rays in a concentric medium are plane curves.

With $\dot{\phi} = 0$, we have $\partial F/\partial\theta \equiv 0$, so that the second Euler equation reduces to

$$nr^2\dot{\theta}(1 + r^2\dot{\theta}^2)^{-\frac{1}{2}} = e, \tag{3.12}$$

where e is constant along any ray. It is not difficult to show the geometrical meaning of e as indicated by the formula

$$e = \pm nr \sin\psi, \tag{3.13}$$

ψ being the angle shown in Figs. 1 and 2, and the $\pm$ being chosen as $\psi \gtrless \frac{1}{2}\pi$. It is convenient for conventional-type optical systems to assume that θ is in the interval $0 \leqq \theta < \frac{1}{2}\pi$, and to consider two cases according to whether $z_0 > 0$ (Fig. 1, with $\dot{\theta} > 0$ and $\psi < \frac{1}{2}\pi$) or $z_0 < 0$ (Fig. 2, with $\dot{\theta} < 0$ and

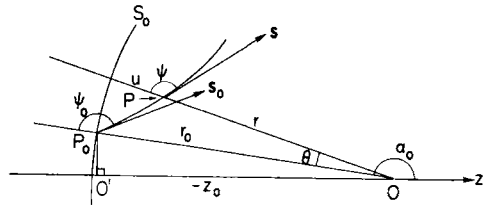


Fig. 2. Similar to Fig. 1, but with $z_0 < 0$.

$\psi > \frac{1}{2}\pi$). We shall assume $r > 0$ throughout. The value of e can be determined for a given ray by applying eq. (3.13) at the starting point, i.e.,

$$e = \pm n_0 r_0 \sin\psi_0. \tag{3.14}$$

Rearranging eq. (3.12) we obtain

$$r\dot{\theta}(n^2r^2 - e^2)^{\frac{1}{2}} = e, \tag{3.15}$$

the positive root being chosen since e and $\dot{\theta}$ must have the same sign [see eq. (3.12)]. From this we have

$$\theta = e\int_{r_0}^{r} \frac{\mathrm{d}r}{r\sqrt{(n^2r^2 - e^2)}}. \tag{3.16}$$

It now appears to be a simple matter to obtain the Cartesian coordinates of a general point on the ray. However, the polar axis was chosen in the direction OP_0 making an angle α_0 with the z-axis of the optical system (see Figs. 1 and 2). In order to obtain convenient tracing formulas it is necessary to employ Cartesian coordinates (x, y, z) with the z-axis along the optical axis and optical direction cosines (p, q, l) relative to these axes. Therefore we introduce the following vectors, which are illustrated geometrically in Fig. 3:

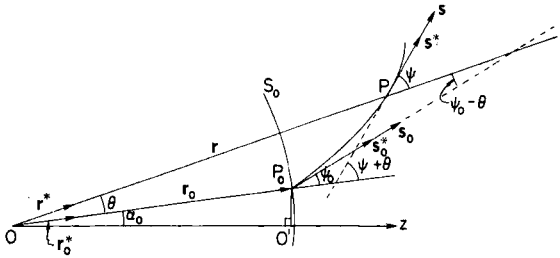


Fig. 3. Similar to Fig. 1, but showing vectors.

$$\begin{aligned} \boldsymbol{r} &= (x, y, z) & \hat{\boldsymbol{r}} &= \boldsymbol{r}/r \\ \boldsymbol{r}_0 &= (x_0, y_0, z_0) & \hat{\boldsymbol{r}}_0 &= \boldsymbol{r}_0/r_0 \\ \boldsymbol{s} &= (p, q, l) & \hat{\boldsymbol{s}} &= \boldsymbol{s}/n \\ \boldsymbol{s}_0 &= (p_0, q_0, l_0) & \hat{\boldsymbol{s}}_0 &= \boldsymbol{s}_0/n_0 . \end{aligned} \tag{3.17}$$

The vectors in the second column are simply the unit vectors corresponding to those in the first column.

Since each ray lies entirely in one plane, we can write

$$\hat{\boldsymbol{r}} = \alpha\hat{\boldsymbol{r}}_0 + \beta\hat{\boldsymbol{s}}_0 , \tag{3.18}$$

where α and β are coefficients to be determined. Scalar multiplication by $\hat{\boldsymbol{r}}_0$ and $\hat{\boldsymbol{s}}_0$ in turn gives

$$\begin{aligned} \hat{\boldsymbol{r}} \cdot \hat{\boldsymbol{r}}_0 &= \alpha\hat{\boldsymbol{r}}_0^2 + \beta(\hat{\boldsymbol{s}}_0 \cdot \hat{\boldsymbol{r}}_0) \\ \hat{\boldsymbol{r}} \cdot \hat{\boldsymbol{s}}_0 &= \alpha(\hat{\boldsymbol{r}}_0 \cdot \hat{\boldsymbol{s}}_0) + \beta\hat{\boldsymbol{s}}_0^2 . \end{aligned} \tag{3.19}$$

As seen from Fig. 3, these relations lead to

$$\begin{aligned} \cos\theta &= \alpha + \beta\cos\psi_0 \\ \cos(\psi_0 - \theta) &= \alpha\cos\psi_0 + \beta, \end{aligned} \tag{3.20}$$

after which it is easily found that

$$\beta = \sin\theta/\sin\psi_0, \qquad \alpha = \cos\theta - \beta\cos\psi_0. \tag{3.21}$$

From eqs. (3.17) and (3.18) we obtain

$$\begin{aligned} x &= r(\alpha\, x_0/r_0 + \beta\, p_0/n_0) \\ y &= r(\alpha\, y_0/r_0 + \beta\, q_0/n_0) \\ z &= r(\alpha\, z_0/r_0 + \beta\, l_0/n_0). \end{aligned} \tag{3.22}$$

In principle these are the desired ray-tracing equations provided α and β are eliminated by eqs. (3.21) and θ is evaluated by eq. (3.16). The components (p, q, l) of the ray direction vector $\boldsymbol{s}$ can be found by differentiating eqs. (3.22) with respect to r and then normalizing to a length n.

In practice, eqs. (3.22) involve three difficulties. First, for radial or nearly radial rays, β becomes indeterminate since both $\sin\theta$ and $\sin\psi_0$ tend to 0. Second, when OP_0 is large, the quantities z, z_0, r and r_0 are large and may even be infinite. Finally, after elimination of α, β and θ, eqs. (3.22) give the ray-tracing equations in terms of the inconvenient parameter r instead of z. We therefore revise the tracing formulas to remove these difficulties.

Equation (3.16) can be written

$$\theta = eM_0, \tag{3.23}$$

where

$$M_0 = \int_{r_0}^{r} \frac{\mathrm{d}r}{r\sqrt{(n^2r^2 - e^2)}}. \tag{3.24}$$

We multiply and divide by θ in the first of eqs. (3.21) and note eqs. (3.14) and (3.23) to obtain

$$\beta = \pm\, n_0 r_0\, M_0 \operatorname{sinc}\theta \qquad (\pm \text{as } z_0 \gtrless 0). \tag{3.25}$$

This removes the difficulty with nearly radial rays. Numerical computation of $\operatorname{sinc}\theta = (\sin\theta)/\theta$ can be done by division when $\theta > 0.01$ (say) and by the series

$$\operatorname{sinc}\theta = 1 - \theta^2/3! + \theta^4/5! - \ldots, \tag{3.26}$$

when $\theta \leqq 0.01$.

We consider next the case where OP_0 is large, as is often true in practice. In fact, when $\mathrm{OP}_0 \to \infty$, the surfaces of constant index become planes perpendicular to the optical axis and we have an axial gradient. To deal with large values of OP_0 we introduce a new parameter u in place of r, i.e.,

$$u = \pm(r - r_0) \qquad (\pm \text{ as } z_0 \gtrless 0). \tag{3.27}$$

The geometrical meaning of u, as illustrated in Figs. 1 and 2, is simply the penetration depth of the point P beyond the surface S_0 of constant index passing through P_0. It is convenient to introduce the symbol ρ_0 for the curvature of the sphere S_0, with ρ_0 considered negative or positive depending on whether S_0 is concave to the left or the right, i.e., as $z_0 \gtrless 0$. This agrees with the usual optical convention. In both cases this convention leads to

$$r/r_0 = 1 - \rho_0 u. \tag{3.28}$$

With help of eqs. (3.27) and (3.28), we find that eq. (3.24) can be written

$$M_0 = \pm\, M\rho_0^2 \qquad (\pm \text{ as } z_0 \gtrless 0), \tag{3.29}$$

with

$$M = \int_0^u \frac{\mathrm{d}u}{(1 - \rho_0 u)\sqrt{\{n^2(1 - \rho_0 u)^2 - (e\rho_0)^2\}}}. \tag{3.30}$$

This integral is seen to be well-behaved when OP_0 is large since ρ_0 is small, $e\rho_0$ is always finite, and u is small in cases of practical interest. It should be understood that, in eq. (3.30), n is to be expressed as a function of u, the distance of P from the surface S_0 measured radially. The sphere S_0

need not coincide with the glass surface since the grinding and polishing done after the ion-diffusion treatment may produce a different curvature.

It is helpful to introduce the symbol

$$\bar{e} = |e\rho_0| = n_0 \sin\psi_0, \tag{3.31}$$

after which we find easily

$$\begin{aligned} \theta &= \bar{e}\,|\rho_0|\,M, \qquad \beta = n_0\,|\rho_0|\,M \operatorname{sinc}\theta \\ \alpha &= \cos\theta - \beta\cos\psi_0, \end{aligned} \tag{3.32}$$

and

$$\begin{aligned} x &= (1-\rho_0 u)(\alpha x_0 + p_0\, M \operatorname{sinc}\theta) \\ y &= (1-\rho_0 u)(\alpha y_0 + q_0\, M \operatorname{sinc}\theta) \\ z &= (1-\rho_0 u)(\alpha z_0 + l_0\, M \operatorname{sinc}\theta). \end{aligned} \tag{3.33}$$

Here M is given by eq. (3.30) with $(e\rho_0)^2$ replaced by $\bar{e}^2$.

Equations (3.32) and the first two of eqs. (3.33) are well-behaved, but the equation for z breaks down when OP_0 is large since z and z_0 become large. This difficulty can be removed by introducing a new origin obtained by projecting the point P_0 onto the z-axis. The coordinates of P relative to this origin are

$$\bar{x} = x, \qquad \bar{y} = y, \qquad \bar{z} = z - z_0. \tag{3.34}$$

After some manipulation (MARCHAND [1970]) a well-behaved formula for $\bar{z}$ can be found. The resulting tracing formulas are still not entirely convenient because they are expressed in terms of the parameter u. It would be helpful to solve the last of eqs. (3.33) for $\bar{z}$ in terms of u so that u could be eliminated from the first two equations. This cannot be done explicitly because of the implicit form of the third equation. However, the equation can be solved numerically, and a convenient procedure for doing so is to arrange it in the form

$$\begin{aligned} u = \pm\sec\theta\{\bar{z}/a + M[(1-\rho_0 u)(b - l_0/a)\operatorname{sinc}\theta \\ + \bar{e}\sin(\tfrac{1}{2}\theta)\operatorname{sinc}(\tfrac{1}{2}\theta)]\} \qquad (\pm \text{ as } z_0 \gtrless 0), \end{aligned} \tag{3.35}$$

where

$$\begin{aligned} b &= a l_0 + |\rho_0|(p_0 x_0 + q_0 y_0) \\ a &= \pm[1 - \rho_0^2(x_0^2 + y_0^2)]^{\frac{1}{2}} \end{aligned} \tag{3.36}$$

(MARCHAND [1970]).

Equation (3.35) can be used effectively for numerical ray tracing in conjunction with eqs. (3.36) and eqs. (3.30)–(3.33). With $\bar{z}$ given we put a trial value of u, and the corresponding values of M and θ, on the right-hand side

of eq. (3.35) and so calculate an improved approximate value of u. Iteration of this procedure gives u to the desired degree of accuracy.

It is found (MARCHAND [1970]) that the optical direction cosines of the ray at a typical point P are given by

$$\begin{aligned} p &= A_2 \mid \rho_0 \mid x_0 + B_2 p_0 \\ q &= A_2 \mid \rho_0 \mid y_0 + B_2 q_0 \\ l &= A_2\, a + B_2 l_0, \end{aligned} \tag{3.37}$$

where

$$\begin{aligned} A_2 &= (\alpha t - \bar{e} \sin\theta - b \cos\theta)/(1-\rho_0 u) \\ B_2 &= (\beta t/n_0 + \cos\theta)/(1-\rho_0 u) \\ t &= \pm[n^2(1-\rho_0 u)^2 - \bar{e}^2]^{\frac{1}{2}} \qquad (\pm \text{ as } z_0 \gtrless 0). \end{aligned} \tag{3.38}$$

These equations have also been arranged so as to be valid when OP_0 is large and for radial rays. As before, with $\bar{z}$ given, the corresponding value of u must be the one determined from eq. (3.35) by iteration.

It is not difficult to show that the optical path from P_0 to P is given by

$$L = \pm \int_0^u \frac{n^2(1-\rho_0 u)\mathrm{d}u}{t} \qquad (\pm \text{ as } z_0 \gtrless 0). \tag{3.39}$$

The proof follows from eqs. (3.11) and (3.12) with $\dot{\phi} \equiv 0$ and eqs. (3.15), (3.28), (3.31) and (3.38). It is necessary to replace F by $-F$ in eq. (3.11) in the case when s is a decreasing function of r.

3.3. CYLINDRICAL MEDIUM

In a cylindrical medium the refractive index n is a function of the radial distance r measured from the axis of symmetry. It is therefore convenient to introduce cylindrical coordinates

$$x = r\cos\theta, \qquad y = r\sin\theta, \qquad z = z. \tag{3.40}$$

Then, to apply Fermat's principle, we write

$$L = \int_0^z F\mathrm{d}z \tag{3.41}$$

with

$$F = n(r)(1+\dot{r}^2 + r^2\dot{\theta}^2)^{\frac{1}{2}}, \tag{3.42}$$

in which the dot indicates differentiation with respect to z.

Since $\partial F/\partial\theta = 0$, the first Euler equation reduces to

$$nr^2\dot{\theta}(1+\dot{r}^2 + r^2\dot{\theta}^2)^{-\frac{1}{2}} = c, \tag{3.43}$$

c being constant along any ray and having the same sign as $\dot{\theta}$ at every point. The value of c can be found from the initial conditions of the ray.

Solving for $\dot{\theta}$ we find

$$mr^2\dot{\theta} = c(1+\dot{r}^2)^{\frac{1}{2}}, \tag{3.44}$$

where

$$m = (n^2 - c^2/r^2)^{\frac{1}{2}}. \tag{3.45}$$

With the help of eqs. (3.44) and (3.45) it is found that the second Euler equation reduces to the form

$$\frac{\mathrm{d}}{\mathrm{d}z}[m\dot{r}(1+\dot{r}^2)^{-\frac{1}{2}}] = m'(1+\dot{r}^2)^{\frac{1}{2}}, \tag{3.46}$$

where the prime denotes differentiation with respect to r. This equation can be written

$$[m\dot{r}(1+\dot{r}^2)^{-\frac{1}{2}}]\mathrm{d}[m\dot{r}(1+\dot{r}^2)^{-\frac{1}{2}}] = m\,\mathrm{d}m \tag{3.47}$$

and integrated to give

$$[m\dot{r}(1+\dot{r}^2)^{-\frac{1}{2}}]^2 + k = m^2 \tag{3.48}$$

or

$$m^2/(1+r^2) = k. \tag{3.49}$$

This relation shows that the expression on the left-hand side is invariant along any given ray. The optical significance of this result can be seen as follows. The optical direction cosines of a ray at any point are given by

$$p = n\,\mathrm{d}x/\mathrm{d}s, \qquad q = n\,\mathrm{d}y/\mathrm{d}s, \qquad l = n\,\mathrm{d}z/\mathrm{d}s. \tag{3.50}$$

The third equation can be written

$$l = n/\dot{s} = n(1+\dot{r}^2+r^2\dot{\theta}^2)^{-\frac{1}{2}}. \tag{3.51}$$

Then, from eqs. (3.43), (3.44) and (3.49), it follows that

$$l^2 = \frac{m^2}{1+\dot{r}^2} = k. \tag{3.52}$$

Hence, in any cylindrical medium, the third optical direction cosine is constant along any ray:

$$l = l_0 = n_0 \cos\gamma_0, \tag{3.53}$$

the zero subscript denoting evaluation at P_0. This invariance of l can also be deduced at once from the z-component of eq. (3.1).

Equation (3.49) can now be written

$$\dot{r}^2 = m^2/l_0^2 - 1 \tag{3.54}$$

or

$$\dot{r} = \mathrm{d}r/\mathrm{d}z = g/l_0 \tag{3.55}$$

with

$$g = \pm(m^2 - l_0^2)^{\frac{1}{2}}. \tag{3.56}$$

Here the $\pm$ is chosen depending on whether r is an increasing or decreasing function of z.

In principle, the entire ray trace can now be reduced to two quadratures. From eq. (3.55) we have

$$z = l_0 \int_{r_0}^{r} \frac{\mathrm{d}r}{g} \tag{3.57}$$

and from eqs. (3.44) and (3.52),

$$\dot{\theta} = c/(l_0 r^2) \tag{3.58}$$

giving

$$\theta = \theta_0 + (c/l_0) \int_0^z \frac{\mathrm{d}z}{r^2}. \tag{3.59}$$

Unfortunately, eq. (3.57) gives z as a function of r and thus gives r only as an implicit function of z. So it is necessary to do a numerical inversion to obtain r in terms of z, after which θ can be determined from eq. (3.59). Furthermore, the denominator in eq. (3.57) is often small, so that the evaluation of the integral becomes highly inaccurate.

In view of the above difficulties, practical ray tracing is best done by a different method. From eqs. (3.52) and (3.53) it follows that eq. (3.46) reduces to

$$\ddot{r} = mm'/l_0^2. \tag{3.60}$$

This equation, together with eq. (3.58) and appropriate initial conditions, provides an efficient method for tracing rays in a cylindrical medium. Standard subroutines for solving such differential equations are well known. For meridional rays, $c = 0$, in which case only eq. (3.60) is required. In the case of a skew ray passing near the axis, some care may be needed since the denominator in eq. (3.58) becomes small.

Once r and θ are known, for a given value of z, x and y are found by eqs. (3.40). To find the direction of the ray we first note from eqs. (3.50) and (3.53) that

$$\begin{aligned} p &= n\dot{x}/\dot{s} = l\dot{x} = l_0\,\dot{x} \\ q &= n\dot{y}/\dot{s} = l\dot{y} = l_0\,\dot{y} \\ l &= l_0\,. \end{aligned} \tag{3.61}$$

Differentiation of eqs. (3.40) gives

$$\begin{aligned} \dot{x} &= \dot{r}\cos\theta - r\sin\theta\,\dot{\theta} \\ \dot{y} &= \dot{r}\sin\theta + r\cos\theta\,\dot{\theta}. \end{aligned} \tag{3.62}$$

Hence, we find with the help of eqs. (3.55), (3.58) and (3.40)

$$\begin{aligned} p &= g\cos\theta - (c/r)\sin\theta \\ q &= g\sin\theta + (c/r)\cos\theta \\ l &= l_0\,. \end{aligned} \tag{3.63}$$

From the preceding formulas it is not difficult to show that

$$xq - yp = c \tag{3.64}$$

at each point of a ray, showing that c is the well-known skewness invariant. Equation (3.64) is a useful check formula. Also, by applying this equation at P_0, we have

$$c = x_0 q_0 - y_0 p_0, \tag{3.65}$$

which serves to determine the value of c at the start of a ray trace. A second check formula is the relation

$$p^2 + q^2 + l^2 = n^2, \tag{3.66}$$

which holds at each point of a ray.

For some problems it is important to compute the optical path from P_0 to P. This is obtained by noting that

$$L = \int_0^s n\,\mathrm{d}s = \int_0^z n\dot{s}\,\mathrm{d}z = \int_0^z \frac{n^2\mathrm{d}z}{l}\,. \tag{3.67}$$

Since l is invariant, this gives

$$L = \frac{1}{l_0}\int_0^z n^2\mathrm{d}z. \tag{3.68}$$

To evaluate this integral it suffices to know n as a function of r and, in turn, r as a function of z.

3.4. AN IMPORTANT SPECIAL CASE

In the Wood lens a quadratic function was assumed for $n(r)$ as given by eq. (1.3). More generally, cylindrical media have often been described by a series of the form

$$n(r) = N_0 + N_1 r^2 + N_2 r^4 + \ldots . \tag{3.69}$$

However, it is found (MARCHAND [1972]) that, if an index function of the form

$$n^2(r) = N_0^2 \pm b^2 r^2, \qquad b > 0, \tag{3.70}$$

is assumed, the ray-tracing equations can be integrated completely in closed form. This can be seen as follows.

Differentiation of eqs. (3.62) gives

$$\ddot{x} = \ddot{r} \cos\theta - 2\,\dot{r}\,\dot{\theta} \sin\theta - r\,\dot{\theta}^2 \cos\theta - r\,\ddot{\theta} \sin\theta \tag{3.71}$$

with a similar equation for $\ddot{y}$. From eqs. (3.60), (3.45) and (3.70), we find

$$\ddot{r} = (nn' + c^2/r^3)/l_0^2 = (c^2/r^3 \pm b^2 r)/l_0^2, \tag{3.72}$$

and from eq. (3.58),

$$\dot{\theta} = c/l_0 r^2, \qquad \ddot{\theta} = -2c\dot{r}/l_0 r^3. \tag{3.73}$$

Thus, eq. (3.71) reduces to form

$$\ddot{x} = \pm (b/l_0)^2\, r \cos\theta, \tag{3.74}$$

with a similar equation for $\ddot{y}$. It is convenient to introduce a scale factor for the z-coordinate, i.e.,

$$\bar{z} = z\, b/l_0, \tag{3.75}$$

after which the differential equations can be written

$$\begin{aligned} \mathrm{d}^2x/\mathrm{d}\bar{z}^2 + x &= 0 \\ \mathrm{d}^2y/\mathrm{d}\bar{z}^2 + y &= 0, \end{aligned} \tag{3.76}$$

provided the negative sign is taken in eq. (3.70).

These equations have the solution

$$\begin{aligned} x &= x_0 \cos\bar{z} + (p_0/b) \sin\bar{z} \\ y &= y_0 \cos\bar{z} + (q_0/b) \sin\bar{z}, \end{aligned} \tag{3.77}$$

where the constants of integration have been chosen so that at P_0, x and y reduce to x_0, y_0, respectively, and the optical direction cosines of the ray are given by p_0, q_0 at that point.

The corresponding formulas for the direction of the ray at any point are found to be

$$\begin{aligned} p &= p_0 \cos \bar{z} - b x_0 \sin \bar{z} \\ q &= q_0 \cos \bar{z} - b y_0 \sin \bar{z}. \end{aligned} \tag{3.78}$$

When the positive sign is taken in eq. (3.70), the differential equations reduce to the form

$$\begin{aligned} \mathrm{d}^2 x/\mathrm{d}\bar{z}^2 - x &= 0 \\ \mathrm{d}^2 y/\mathrm{d}\bar{z}^2 - y &= 0, \end{aligned} \tag{3.79}$$

and the corresponding ray-tracing formulas are found to be

$$\begin{aligned} x &= x_0 \cosh \bar{z} + (p_0/b) \sinh \bar{z} \\ y &= y_0 \cosh \bar{z} + (q_0/b) \sinh \bar{z} \\ p &= p_0 \cosh \bar{z} + b\, x_0 \sinh \bar{z} \\ q &= q_0 \cosh \bar{z} + b\, y_0 \sinh \bar{z}. \end{aligned} \tag{3.80}$$

Inspection of eqs. (3.77), (3.78) and (3.80) shows that they can be used without difficulty for numerical ray tracing of all rays. In case b is very small, division by b can be dealt with by the simple device of writing

$$(1/b) \sin \bar{z} = (\bar{z}/b) \operatorname{sinc} \bar{z} = (z/l_0) \operatorname{sinc} \bar{z},$$

where eq. (3.75) was used. The same device can be used for the computation of $(1/b)\sinh \bar{z}$ when b is small.

Owing to the simplicity of the solutions of eq. (3.70), it seems reasonable to try gradients of this type in design work. There are two parameters N_0 and b available to optimize aberrations. The choice of sign in eq. (3.70) allows the gradient to have either a positive or a negative power. Once a preliminary design has been achieved, fine correction can be sought by adding higher terms in eq. (3.70).

3.5. ANALYTICAL RAY TRACING

Equations (3.57) and (3.59) represent an attempt to integrate the ray-tracing equation as far as possible analytically in the case of a cylindrical medium. But, of course, nothing more can be done unless the function $n(r)$ is specified. Furthermore, these equations involve difficulties in practical calculations, and STREIFER and PAXTON [1971] developed an analytic approach to this problem by assuming that the index function is described by an expression of the form

$$n^2(r) = N_0^2 [1 - \delta (r/r_0)^2 + \alpha_2\, \delta^2 (r/r_0)^4 + \ldots], \tag{3.81}$$

where N_0, δ, α_2, α_3, ... are constants that determine the index function. In their treatment it was assumed that $\delta > 0$, which is certainly valid for the GRIN rod application where a converging medium is required.

It should be noted that, in eq. (3.81), n^2 instead of n is expanded in a series, so that a conversion is required in order to compare coefficients with those used by most other writers.

Streifer and Paxton show that, for $\delta > 0$, an exact solution can be found if a finite number of terms are used in eq. (3.81). However, the algebraic expressions for the solutions become very complicated if more than two or three terms of eq. (3.81) are included. It is interesting to note that the derivation of the intricate algebraic formulas can be done with modern computer software techniques.

The form chosen for eq. (3.81) makes it convenient for consideration of asymptotic approximations based on small values of δ.

3.6. GENERAL ROTATION-SYMMETRIC MEDIUM

In case n is a function of both r and z, much of the development in § 3.3 remains valid. However, the third optical direction cosine l is no longer invariant. Equation (3.46) must now be replaced by

$$\frac{\mathrm{d}}{\mathrm{d}z}\left[m\dot{r}(1+\dot{r}^2)^{-\frac{1}{2}}\right] = \frac{\partial m}{\partial r}(1+\dot{r}^2)^{\frac{1}{2}}, \tag{3.82}$$

and this cannot be integrated without knowing how m depends on both r and z.

Setting

$$u = m\dot{r}(1+\dot{r}^2)^{-\frac{1}{2}}, \tag{3.83}$$

we find that

$$(1+\dot{r}^2)^{\frac{1}{2}} = m/\sqrt{(m^2-u^2)}, \tag{3.84}$$

$$\dot{r} = u/\sqrt{(m^2-u^2)}, \tag{3.85}$$

and

$$\dot{u} = m(\partial m/\partial r)/\sqrt{(m^2-u^2)}. \tag{3.86}$$

Also, since eq. (3.44) is still valid in the present case, we obtain

$$\dot{\theta} = \frac{c}{r^2\sqrt{(m^2-u^2)}}. \tag{3.87}$$

Equations (3.85), (3.86) and (3.87) form a set of first-order differential equations that can be solved by standard computer subroutines. Here it is assumed that n is a known function of r and z and that n is given by eq.

(3.45), i.e.,

$$m^2 = n^2 - c^2/r^2. \tag{3.88}$$

Thus, one obtains numerically the functions $r(z)$, $u(z)$ and $\theta(z)$ and, from these, the ray-tracing equations in Cartesian form as given by eqs. (3.40).

For the present case we find, with the help of eq. (3.51), that

$$l = \sqrt{(m^2 - u^2)}. \tag{3.89}$$

The first two optical direction cosines of the ray at each point are found by

$$\begin{aligned} p &= u\cos\theta - (c/r)\sin\theta \\ q &= u\sin\theta + (c/r)\cos\theta, \end{aligned} \tag{3.90}$$

which are essentially the same as the first two of eqs. (3.63). The optical path along a ray is still given by eq. (3.67), but not by eq. (3.68), since l is no longer constant along a ray.

It is evident that the formulas of this section provide an alternative method for tracing rays in a spherical medium.

§ 4. Lens Design

4.1. PARAXIAL RAYS IN A CYLINDRICAL MEDIUM

In a gradient-index medium the rays travel in curved paths, so that the classical formulas for lens aberrations must be generalized for such media. In defining aberrations the well-established approach is to begin by considering the first-order approximations to the exact ray-tracing formulas, i.e., the paraxial tracing formulas. This name derives from the fact that these rays describe well the rays that lie near the optical axis.

It is known that, if all the rays obeyed the paraxial formulas exactly, the image formation of the optical system would be ideal. Therefore, it is customary to use the paraxial rays as reference for the real rays. The deviations of the real rays from the paraxial ones serve as definitions of the aberrations of the system. It is therefore important to obtain the appropriate paraxial ray-tracing formulas for gradient-index media.

In a cylindrical medium, such as one described by eq. (3.69), it is found that the coefficients N_0 and N_1 suffice to determine the paraxial formulas. It is not possible to disregard the coefficient N_1 since the differential equations determining the ray trace involve the derivative

$$n'(r) = 2r(N_1 + 4N_2 r^2 + \ldots). \tag{4.1}$$

However, the second term and higher terms here can be discarded in the first-order theory.

A convenient way to obtain the paraxial formulas is to consider the special case described in § 3.4. When r is small, eq. (3.70) can be expanded as

$$\begin{aligned} n(r) &= N_0(1 \pm b^2r^2/N_0^2)^{\frac{1}{2}} \\ &= N_0[1 \pm \tfrac{1}{2}(br/N_0)^2 - \tfrac{1}{4}(br/N_0)^4 \pm \ldots]. \end{aligned} \tag{4.2}$$

From the viewpoint of paraxial optics any cylindrical medium having

$$N_1 = \pm\, b^2/(2N_0) \tag{4.3}$$

will behave the same as the special medium with parameters N_0 and b. Therefore, the tracing formulas (3.77), (3.78) and (3.80) can be used as the paraxial tracing formulas for any cylindrical medium provided we take

$$b = (2N_0|N_1|)^{\frac{1}{2}} \tag{4.4}$$

and also replace several quantities by their appropriate paraxial expressions.

The scale factor b/l_0 becomes

$$b/l_0 = (2N_0|N_1|)^{\frac{1}{2}}/(n_0 \cos \gamma_0) \approx (2|N_1|/N_0)^{\frac{1}{2}}, \tag{4.5}$$

in the paraxial limit. In this case

$$\bar{z} = kz, \qquad k = (2|N_1|/N_0)^{\frac{1}{2}}. \tag{4.6}$$

In aberration theory, for the most part, only meridional paraxial rays are of interest. For a meridional ray in the xz-plane

$$q = n \cos \beta = n\, \mathrm{d}y/\mathrm{d}z = -nu, \tag{4.7}$$

where u is the negative of the slope, as often used in describing meridional rays. Taking note of eqs. (4.4) and (4.6) we thus find in the paraxial limit

$$\begin{aligned} q &= -N_0u, \qquad q_0 = -N_0u_0 \\ q_0/b &= -u_0/k. \end{aligned} \tag{4.8}$$

We now obtain the paraxial tracing formulas from eqs. (3.77), (3.78) and (3.80), the results being

$$\begin{aligned} y &= y_0 \cos \bar{z} - (u_0/k) \sin \bar{z} \\ u &= u_0 \cos \bar{z} + (y_0k) \sin \bar{z} \end{aligned} \tag{4.9}$$

when $N_1 < 0$, and

$$\begin{aligned} y &= y_0 \cosh \bar{z} - (u_0/k) \sinh \bar{z} \\ u &= u_0 \cosh \bar{z} - (y_0k) \sinh \bar{z} \end{aligned} \tag{4.10}$$

when $N_1 > 0$. It is interesting to observe that the paraxial rays are not straight lines.

4.2. FOCAL LENGTH OF A SINGLET

The paraxial tracing formulas are used to determine the focal length of the optical system. For this purpose we consider a meridional ray entering the system parallel to the axis. Using the paraxial tracing formulas we must find out where the ray meets the z-axis after passing through the system. To illustrate the procedure we consider a single lens element having a cylindrical gradient index and surrounded by air.

Let y_0 and y_1 be the ray heights at the first and second surfaces, respectively, and u_0, u_1, u_2, u_3 the negative slope values before and after the first and second surfaces, respectively. For paraxial refraction formulas we use the axial value N_0 for the index inside the lens and so obtain

$$\begin{aligned} N_0 u_1 &= u_0 + y_0 \rho_0 (N_0 - 1) \\ u_3 &= N_0 u_2 + y_1 \rho_1 (1 - N_0), \end{aligned} \tag{4.11}$$

where ρ_0, ρ_1 are the two surface curvatures of the singlet.

Adjusting the notation in eqs. (4.9) and (4.10), we have

$$\begin{aligned} u_2 &= u_1 \begin{Bmatrix} \cos(kd) \\ \cosh(kd) \end{Bmatrix} + y_0 k \begin{Bmatrix} \sin(kd) \\ -\sinh(kd) \end{Bmatrix} \\ y_1 &= y_0 \begin{Bmatrix} \cos(kd) \\ \cosh(kd) \end{Bmatrix} - (u_1/k) \begin{Bmatrix} \sin(kd) \\ \sinh(kd) \end{Bmatrix}, \end{aligned} \tag{4.12}$$

in which the upper or lower quantities are used as $N_1 \lesseqgtr 0$, and d is the axial thickness.

The focal length f^* is obtained by setting $u_0 = 0$ and using the definition

$$1/f^* = u_3/y_0. \tag{4.13}$$

After elimination of u_1, u_2 and y_1 we have

$$\begin{aligned} 1/f^* = &\begin{Bmatrix} \cos(kd) \\ \cosh(kd) \end{Bmatrix} (N_0 - 1)(\rho_0 - \rho_1) \\ &+ \begin{Bmatrix} \sin(kd)/kd \\ \sinh(kd)/kd \end{Bmatrix} \frac{(N_0 - 1)^2 d \rho_0 \rho_1}{N_0} \\ &+ \begin{Bmatrix} \sin(kd) \\ -\sinh(kd) \end{Bmatrix} N_0 k, \qquad N_1 \lesseqgtr 0. \end{aligned} \tag{4.14}$$

Here

$$k = (2|N_1|/N_0)^{\frac{1}{2}}, \tag{4.15}$$

as in eq. (4.6). It is evident that, when $N_1 = 0$, eq. (4.14) reduces to the customary formula for the focal length of a homogeneous singlet,

$$1/f_0 = (N_0-1)(\rho_0-\rho_1)+(N_0-1)^2 d\rho_0\rho_1/N_0. \tag{4.16}$$

By the same general procedure we can derive formulas for locating the paraxial image plane for a given object plane.

Equation (4.14) can often be replaced by an approximate formula, valid when $kd \ll 1$, obtained by approximating the terms in braces by the leading terms of their series expansions. This leads to

$$1/f^* = 1/f_0 - 2N_1 d. \tag{4.17}$$

From eq. (4.14) we can find the focal length of a Wood lens by letting $\rho_0 \to 0$ and $\rho_1 \to 0$. In this case, the approximate formula, eq. (4.17), reduces to

$$1/f^* = -2N_1 d, \tag{4.17a}$$

showing that a cylindrical gradient does, indeed, produce power in a plane parallel plate.

4.3. BUCHDAHL THEORY

As pointed out, the various geometrical aberrations can be computed for a given optical system by tracing real rays and paraxial rays. This makes it possible to predict the performance of the system from its construction parameters. However, the central problem in lens design is to decide how to change the construction parameters so as to reduce the aberrations.

Several years ago BUCHDAHL [1969] developed a modification of the classical order-by-order analysis of aberrations. Buchdahl's method involves the surface-by-surface computation of contributions to the third-order, fifth-order, etc., geometrical aberrations. This approach helps the designer to determine which surfaces and which parameters are sensitive for the various aberrations. What is more, this surface-by-surface analysis applies also to systems containing one or more gradient-index media, provided these have rotational symmetry about the axis.

Buchdahl's theory calls for tracing two meridional paraxial rays through the system. One ray passes through the axial point of the object plane and has unit height at the entrance pupil. The second ray has unit height at the object plane and zero height at the entrance pupil. A brief outline of Buchdahl's method, as applied in the presence of inhomogeneous media, is presented in an appendix of his book (BUCHDAHL [1969]).

Sands and Moore developed many of the details of Buchdahl theory as

applied to gradient-index media (SANDS [1970, 1971 (a, b, c)], MOORE [1970, 1971], MOORE and SANDS [1971]). In particular, Moore investigated the correction of third-order aberrations in gradient-index singlets. He was able to design a singlet with an axial gradient corrected for third-order spherical aberration and coma and having no third-order distortion. Using a cylindrical gradient he designed a singlet corrected for all third-order monochromatic aberrations except for Petzval curvature of field. His studies indicate that, in a general way, cylindrical gradients appear to be more effective in lens design than axial or spherical gradients.

4.4. COLOR CORRECTION

In the preceding sections it was seen that the paraxial properties of a cylindrical medium are determined by the coefficients N_0 and N_1 of the expansion

$$n(r) = N_0 + N_1 r^2 + N_2 r^4 + \dots . \tag{4.18}$$

In these sections the assumption was tacitly made that the light involved was monochromatic.

The possibility that $N_0, N_1, N_2, \dots$ can depend on wavelength must eventually be faced even though, at the present time, little is known about dispersion properties of the actual inhomogeneous media now being fabricated.

SANDS [1971a] has considered theoretically the question of computing the paraxial chromatic aberration of a general rotation-symmetric system. Thus he assumes an index function of the form

$$n(\xi, z, \lambda) = N_0(z, \lambda) + N_1(z, \lambda)\xi + N_2(z, \lambda)\xi^2 + \cdots , \tag{4.19}$$

where

$$\xi = r^2 = x^2 + y^2 \tag{4.20}$$

and λ is the wavelength.

Following closely the methods of Buchdahl, Sands shows that the chromatic aberration can be computed by summing contributions from all of the surfaces and all of the media in the system even though some of the media may have inhomogeneous index functions of the form shown in eq. (4.19). The definition used for the chromatic aberration is

$$\boldsymbol{\varepsilon} = \boldsymbol{H}' - m\boldsymbol{H}, \tag{4.21}$$

in which $\boldsymbol{H}$ is the two-dimensional vector giving the location of the object point in the object plane, m is the paraxial magnification for light of wavelength λ_0 (the base wavelength) and $\boldsymbol{H}'$ is the vector of position (for wavelength λ) in the ideal image plane for the base color.

We shall not present here the detailed formulas for computing ε. However, it is interesting to note that Sands extends a dispersion theory of Buchdahl's to the inhomogeneous case. Thus each of the coefficients $N_0, N_1, \ldots$ is assumed to be expanded in a power series in a color parameter ω, which is a function of λ and λ_0 of the form

$$\begin{aligned} \omega &= \delta\lambda/(1+\alpha\delta\lambda) \\ \delta\lambda &= \lambda-\lambda_0. \end{aligned} \tag{4.22}$$

Thus one can speak of color aberrations of various chromatic orders. The linear terms in these expressions naturally lead to the primary color aberrations.

4.5. THIRD-ORDER ABERRATIONS OF INHOMOGENEOUS LENSES

Following BUCHDAHL [1969] and SANDS [1970], we consider a medium with refractive index given by

$$n = N_0(z)+N_1(z)\,\xi+N_2(z)\,\xi^2+\ldots, \tag{4.23}$$

where

$$\xi = r^2 = x^2+y^2. \tag{4.24}$$

With n a function of both r and z, rays can be traced by the method of § 3.6.

In a centered optical system, it is known (BUCHDAHL [1970]) that the total vectorial third-order aberration is given by an expression of the form

$$\varepsilon = [\sigma_1\,\xi_1+2\sigma_1\,\eta_1+(\sigma_3+\sigma_4)\zeta_1]\boldsymbol{S}+[\sigma_2\,\xi_1+2\sigma_3\,\eta_1+\sigma_5\,\zeta_1]\boldsymbol{T}. \tag{4.25}$$

Here

$$\boldsymbol{S} = (S_x, S_y), \qquad \boldsymbol{T} = (T_x, T_y) \tag{4.26}$$

are two-dimensional vectors indicating the intersection point of the ray with the object plane and the entrance pupil, respectively, and

$$\begin{aligned} \xi_1 &= S_x^2+S_y^2 \\ \eta_1 &= S_x\,T_x+S_y\,T_y \\ \zeta_1 &= T_x^2+T_y^2. \end{aligned} \tag{4.27}$$

It is not difficult to verify that the five coefficients σ_1 to σ_5 are the familiar Seidel aberration coefficients.

In Buchdahl's theory each of these coefficients is expressed as the sum of contributions from all the surfaces and intersurface transitions in the optical system. In order to describe these contributions we consider two meridional paraxial rays. The a-ray is defined by heights of unity at the object plane and zero at the entrance pupil, the b-ray by heights of zero at the object

plane and unity at the entrance pupil. Tracing these two rays (by paraxial formulas) through the optical system determines heights $y_a(z)$, $y_b(z)$ and slopes $v_a(z) = \dot{y}_a(z)$, $v_b(z) = \dot{y}(z)$ for all points of these rays, the dot indicating differentiation with respect to z.

Consider now a typical refracting surface, assumed to be spherical with curvature c. Sands defines a constant K by the formula

$$K = -c\,\Delta(2N_1 + \tfrac{1}{2}c\,\dot{N}_0), \tag{4.28}$$

where Δ indicates the change in the following quantity in moving from one side of the surface to the other, and N_0, N_1 are the coefficients in eq. (4.23) for either of the two media involved. If both media are homogeneous, $K = 0$. By way of further notation let

$$k_0 = N_0/N_0', \tag{4.29}$$

the ratio of the axial indices before and after the surface, and

$$\begin{aligned} i_a &= v_a + c\,y_a \\ i_b &= v_b + c\,y_b \\ q &= i_b/i_a \\ \lambda &= N_0(z)\,[y_a(z)v_b(z) - y_b(z)v_a(z)] \\ a &= \tfrac{1}{2}N_0(k_0 - 1)\,y_a i_a^{\,2}\,(i_a + v_a'). \end{aligned} \tag{4.30}$$

Evidently, i_a is the angle of incidence, i.e., the angle the a-ray makes with the surface normal before refraction, and i_b correspondingly for the b-ray. The quantity λ is a paraxial invariant of the system and so can be determined at any convenient point in the optical system. The quantity a is the usual spherical surface contribution to the spherical aberration.

With this notation it is found that the contributions of the surface to the five coefficients σ_1 to σ_5 are given by

$$\begin{aligned} a_1 &= a + K y_a^{\,4} \\ a_2 &= aq + K y_a^{\,3} y_b \\ a_3 &= aq^2 + K y_a^{\,2} y_b^{\,2} \\ a_4 &= \tfrac{1}{2}\lambda^2 c\,\Delta(1/N_0) \\ a_5 &= aq^3 + qa_4 + K y_a y_b^{\,3}. \end{aligned} \tag{4.31}$$

The effect of a transfer from any surface to the next one is expressed by five contributions a_1^* to a_5^*, as given by the formulas

$$a_1^* = \tfrac{1}{2}\nabla(N_0 y_a v_a^3) + \int [4N_2 y_a^4 + 2N_1 y_a^2 v_a^2 - \tfrac{1}{2}N_0 v_a^4]\mathrm{d}z$$

$$a_2^* = \tfrac{1}{2}\nabla(N_0 y_a v_a^2 v_b) + \int [4N_2 y_a^3 y_b + N_1 y_a v_a(y_a v_b + y_b v_a) - \tfrac{1}{2}N_0 v_a^3 v_b]\mathrm{d}z$$

$$a_3^* = \tfrac{1}{2}\nabla(N_0 y_a v_a v_b^2) + \int [4N_2 y_a^2 y_b^2 + 2N_1 y_a y_b v_a v_b - \tfrac{1}{2}N_0 v_a^2 v_b^2]\mathrm{d}z \qquad (4.32)$$

$$a_4^* = \lambda^2 \int (N_1/N_0^2)\mathrm{d}z$$

$$a_5^* = \tfrac{1}{2}\nabla(N_0 y_a v_b^3) + \int [4N_2 y_a y_b^3 + N_1 y_b v_b(y_a v_b + y_b v_a) - \tfrac{1}{2}N_0 v_a v_b^3]\mathrm{d}z.$$

Here the symbol ∇ indicates the change in the following quantity in the transfer from one surface to the next. The integrals are to be computed for z-values from the first of the surfaces to the next.

It is thus seen to be relatively straightforward to compute the Seidel (third-order) coefficients for a centered system containing one or more inhomogeneous media of the type indicated by eq. (4.23). The procedure requires the tracing of two paraxial rays through the optical system to give the functions $y_a(z)$, $y_b(z)$, $v_a(z)$, $v_b(z)$ and then performing the integrations indicated in eqs. (4.32).

$$\sigma_i = \mu(\sum_j a_{ij} + \sum_j a_{ij}^*), \qquad (4.33)$$

where

$$\mu = -\{N_0(l')v_a(l')\}^{-1}, \qquad (4.34)$$

l' being the z-coordinate of the paraxial image plane.

The two summation signs in eq. (4.33) indicate that each of the five coefficients σ_i is obtained by summing the surface contributions a_i over all surfaces and the transfer contributions a_i^* over all the intervals in the system.

In the case when n is independent of z (cylindrical medium), the ray tracing can be done by eqs. (4.9) when $N_1 < 0$ or by eqs. (4.10) when $N_1 > 0$. In this case the coefficients $N_0, N_1, N_2, \ldots$ are constants. Then the required integrals have the forms

$$\int y_a^4 \mathrm{d}z, \quad \int y_a^2 v_a^2 \mathrm{d}z, \quad \int v_a^4 \mathrm{d}z, \quad \int y_a v_a^2 v_b \mathrm{d}z, \qquad (4.35)$$

etc.

But, for a cylindrical medium, the paraxial quantities y_a, y_b, v_a, v_b are given by simple functions involving $\sin \bar{z}$, $\cos \bar{z}$ when $N_1 < 0$, or $\sinh \bar{z}$ and

cosh $\bar{z}$ when $N_1 > 0$. Consequently, these integrals can be evaluated once and for all in the indefinite form. Then it only remains to insert the appropriate z-values as limits of integration, i.e., the axial values of z at the surfaces involved.

As an illustration, consider the integral

$$A = \int y_a^2 v_a^2 \mathrm{d}z, \tag{4.36}$$

and rewrite eqs. (4.9), noting that $v = -u$, i.e.,

$$\begin{aligned} y_a &= y_{0a} \cos \bar{z} + (v_{0a}/k) \sin \bar{z} \\ v_a &= v_{0a} \cos \bar{z} - y_{0a} k \sin \bar{z} \end{aligned} \tag{4.37}$$

with

$$\bar{z} = kz = z(2|N_1|/N_0)^{\frac{1}{2}}. \tag{4.38}$$

The quantities y_{0a} and v_{0a} are the values of y_a and v_a at the vertex plane of the first of the two surfaces involved.

Substitution of eqs. (4.37) into (4.36) shows that several elementary indefinite integrals must be evaluated, namely,

$$\int \cos^4 \bar{z}\, \mathrm{d}\bar{z}, \quad \int \cos^3 \bar{z} \sin \bar{z}\, \mathrm{d}\bar{z}, \quad \int \cos^2 \bar{z} \sin^2 \bar{z}\, \mathrm{d}\bar{z},$$

$$\int \cos \bar{z} \sin^3 \bar{z}\, \mathrm{d}\bar{z}, \quad \int \sin^4 \bar{z}\, \mathrm{d}\bar{z}, \tag{4.39}$$

where we have changed the variable of integration to $\bar{z}$. As pointed out, these and the other necessary indefinite integrals can be evaluated once and for all, so that no numerical integrations are required.

§ 5. GRIN Rods

5.1. INTRODUCTION

A GRIN rod consists of a transparent, cylindrical rod having a cylindrical index function with $N_1 < 0$. The negative value of the coefficient produces a focusing effect, so that an image on one end of the rod can be transmitted to the other with unit magnification provided the parameters of the rod are suitably chosen.

Ray tracing in GRIN rods is essentially the same as for any cylindrical medium. Hence, eqs. (3.58) and (3.60), in conjunction with eqs. (3.45), (3.56) and (3.63), with suitable initial conditions, serve as the basis for a computer program to trace the rays.

5.2. MERIDIONAL RAYS

With $M < 0$ meridional paraxial rays can be traced by means of eqs. (4.6) and (4.9), i.e.,

$$\begin{aligned} y &= y_0 \cos \bar{z} - (u_0/k) \sin \bar{z} \\ u &= u_0 \cos \bar{z} + y_0 k \sin \bar{z} \\ \bar{z} &= kz, \qquad k = (2|N_1|/N_0)^{\frac{1}{2}}. \end{aligned} \tag{5.1}$$

It is evident that meridional rays are periodic, the period L being given by

$$L = 2\pi/k. \tag{5.2}$$

Thus, if the length of the rod is a multiple of L, any meridional ray entering one end (at $z = 0$) at height y_0 will have the same height at the other end. Therefore, in the paraxial approximation, all points at one end are sharply imaged, in the limited sense that all paraxial meridional rays through a point P_0 at $z = 0$ will pass through a single point P at the other end.

If the index function $n(r)$ is suitably chosen, it is possible for the same property to hold exactly for *all* meridional rays, except for those reflected at the surface of the cylinder.

Consider a medium given by

$$n(r) = N_0 \operatorname{sech}(\alpha r) = N_0(1 - \tfrac{1}{2}\alpha^2 r^2 + \tfrac{5}{24}\alpha^4 r^4 + \ldots). \tag{5.3}$$

For a meridional ray, from eqs. (3.45), (3.56), and (3.57) with $c = 0$, we find

$$z = \pm l_0 \int_{r_0}^{r} \frac{\mathrm{d}r}{\sqrt{(n^2 - l_0^2)}}, \tag{5.4}$$

taking the $\pm$ sign as $\dot{r} \gtrless 0$. Insertion of eq. (5.3) into (5.4) leads to

$$z = A \int_{r_0}^{r} \frac{\cosh(\alpha r)\mathrm{d}r}{\sqrt{\{1 - A^2 \sinh^2(\alpha r)\}}}, \tag{5.5}$$

with

$$A = \pm l_0 (N_0^2 - l_0^2)^{-\frac{1}{2}}. \tag{5.6}$$

Equation (5.5) is easily evaluated by introducing a variable

$$v = A \sinh(\alpha r). \tag{5.7}$$

The result, with r replaced by y, is

$$\begin{aligned} y &= \frac{1}{\alpha} \sinh^{-1} \left[\frac{1}{A} \sin(\alpha z + \sin^{-1} v_0) \right] \\ v_0 &= A \sinh(\alpha r_0). \end{aligned} \tag{5.8}$$

In this case, evidently, every meridional ray is periodic with period

$$\Delta z = 2\pi/\alpha. \tag{5.9}$$

If the length L of the rod is chosen so that

$$L = 2\pi/\alpha \tag{5.10}$$

(or any multiple of Δz), it is clear that any point on one end of the rod will be sharply imaged (from the viewpoint of meridional rays alone).

Thus the gradient function given by eq. (5.3) can be considered as ideal for meridional rays. Unfortunately, with this type of gradient, skew rays are not imaged sharply, as has been pointed out, for instance, by RAWSON, HERRIOTT and MCKENNA [1970]. In fact, it is not possible for a cylindrical gradient to image sharply both meridional and skew rays within its own medium.

5.3. HELICAL RAYS

A ray is called helical if all of its points are at the same distance r_0 from the axis. For such a ray

$$n \equiv n(r_0) \equiv n_0, \qquad r \equiv r_0, \qquad \dot{r} \equiv 0, \qquad \ddot{r} \equiv 0. \tag{5.11}$$

Equation (3.59) gives then

$$\theta = \theta_0 + (c/l_0)(z/r_0^2), \tag{5.12}$$

showing that the ray is periodic with period

$$\Delta z = 2\pi r_0^2 l_0/|c|. \tag{5.13}$$

We can ask what form of gradient function is ideal for helical rays. If the period Δz is equal to the length L of the rod, we have

$$L = 2\pi r_0^2 l_0/|c|. \tag{5.14}$$

It will be seen that, in order for a ray to be helical, the value of c must be related in a certain way to the index function and the radius of the ray. For, with $\ddot{r} = 0$, we see from eq. (3.60) that $m' = 0$ at each point of the ray. Hence, from eq. (3.45),

$$mm' = nn' + c^2/r^3 = 0, \tag{5.15}$$

at each point of the ray. For P_0 this gives

$$c = \pm(-n_0 n_0' r_0^3)^{\frac{1}{2}}. \tag{5.16}$$

This shows that a cylindrical medium can support helical rays only if n is

a decreasing function of r. Furthermore, if $c \neq 0$, there are evidently two symmetrical helical rays passing through a given starting point, the $\pm$ in eq. (5.16) depending on whether $\dot{\theta} \gtrless 0$.

From eq. (3.54) we have $l_0 = m$ along a helical ray, and therefore eq. (3.45) reduces to

$$l_0^2 = n_0^2 - c^2/r_0^2, \tag{5.17}$$

which determines l_0, with the help of eq. (5.16), if the index function $n(r)$ is given along with the radius r_0.

Elimination of c and l_0 from eq. (5.14) [or eq. (5.13)] by means of eqs. (5.16) and (5.17) shows that the period of a helical ray depends, in general, on the radius r_0 of the ray. However, with a suitable choice of the index function, this is not the case. The elimination of l_0 and c from eq. (5.14) leads to

$$n_0'/n_0 = r_0/\{r_0^2 + (L/2\pi)^2\}. \tag{5.18}$$

Following RAWSON, HERRIOTT and MCKENNA [1970] we assume this to hold for all radial values between 0 and r_0 with L independent of r_0. We integrate the differential equation between these limits, the result being

$$n_0/N_0 = [1 + (2\pi r_0/L)^2]^{-\frac{1}{2}}, \tag{5.19}$$

where N_0 is the axial value of the index. An index function of this form may be regarded as ideal for helical rays.

§ 6. Conclusions

The history of optics contains numerous cases where inhomogeneous media have been studied theoretically. In recent years, improved methods of fabricating optical elements having index gradients were found, and this has stimulated new investigations to determine how such elements can be put to practical use.

Two applications of gradient index materials are of special interest at present. Narrow rods (GRIN rods) with cylindrical index gradients show promise as image-transfer devices, and conventional-type optical elements, modified by an inhomogeneity of the index of the material, provide a new design tool for classical types of optical systems. These elements are potentially useful in improving quality, reducing manufacturing costs, and perhaps helping to meet constraints on space and weight in optical systems.

In connection with recent studies of gradients, the role of ray tracing has been important. Although the basic differential equations that determine the ray paths have been known for many years, it is important to cast these into

forms that are efficient for the particular geometries under consideration. In some cases it is found that the differential equations can be either partially or completely integrated.

A substantial amount of theoretical work has appeared recently in the optical literature on methods of analyzing the image-forming properties of gradient-index lenses. However, not much has been reported, as yet, on the application of such elements in the design of practical lenses.

References

BORN, M. and E. WOLF, 1970, Principles of Optics, 4th ed. (Pergamon Press, London and New York).

BUCHDAHL, H. A., 1969, Optical Aberration Coefficients (Dover, New York).

BUCHDAHL, H. A., 1970, An Introduction to Hamiltonian Optics (Cambridge, Univ. Press, New York) Sec. 19.

HAMBLEN, D. P., 1969, Gradient Refractive Index Optical Lenses, U.S. Patent No. 3, 486, 808.

HERZBERGER, M., 1958, Modern Geometrical Optics (Interscience Publishers, New York and London).

LUNEBURG, R. K., 1964, Mathematical Theory of Optics (University of California Press, Berkeley and Los Angeles).

MARCHAND, E. W., 1970, J. Opt. Soc. Am. **60**, 1.

MARCHAND, E. W., 1972, Appl. Opt. **11**, 1104.

MAXWELL, J. C., 1854, Cambridge and Dublin, Math. J. **8**, 188; also Scientific Papers, I (Cambridge Univ. Press) p. 76.

MONTAGNINO, L., 1968, J. Opt. Soc. Am. **58**, 1667.

MOORE, D. T., 1971, J. Opt. Soc. Am. **61**, 886

MOORE, D. T. and P. J. SANDS, 1971, J. Opt. Soc. Am. **61**. 1195

PAXTON, K. B. and W. STREIFER, 1971, Appl. Opt. **10**, 1164.

PEARSON, A. D., W. G. FRENCH and E. G. RAWSON, 1969, Appl. Phys. Lett. **15**, 76.

RAWSON, E. G., D. R. HERRIOTT and J. MCKENNA, 1970, Appl. Opt. **9**, 753.

SANDS, P. J., 1970, J. Opt. Soc. Am. **60**, 1436.

SANDS, P. J., 1971a, J. Opt. Soc. Am. **61**, 777.

SANDS, P. J., 1971b, J. Opt. Soc. Am. **61**, 879.

SANDS, P. J., 1971c, J. Opt. Soc. Am. **61**, 1086.

SINAI, P., 1971, Appl. Opt. **10**, 99.

STREIFER, W. and K. B. PAXTON, 1971, Appl. Opt. **10**, 769.

WOOD, R. W., 1905, Physical Optics (Macmillan, New York) p. 71.

This Page Intentionally Left Blank

AUTHOR INDEX

SUBJECT INDEX

CUMULATIVE INDEX – VOLUMES I–XI

This Page Intentionally Left Blank