# I. THEORY OF THE LONGITUDINAL FREE ELECTRON LASER

# II. A THEORETICAL MODEL OF THE LINEAR ELECTROOPTIC EFFECT

Thesis by

Chun-Ching Shih

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1980

(Submitted April 4, 1980)

*To*

*Su-Miau*

*and my Parents*

*with love*

# ACKNOWLEDGMENTS

I would like to express my heartfelt appreciation to my thesis advisor, Professor Amnon Yariv, for his excellent guidance and constant support I enjoyed while pursuing this research. His invaluable inspirations and encouragements provided me with a most fruitful and enlightening experience for my future career. I am always grateful for all I have learned from him.

I would also like to express my gratitude to Professors R. Feynman, M. Gell-Mann, T. McGill, D. Smith of Caltech, and to Drs. A. Palmer, A. Gentile of Hughes Research Laboratories for many helpful and stimulating discussions.

Special thanks go to Mrs. Ruth Stratton for her masterful typing of this thesis. The critical proofreading of the manuscript by Mr. I. Ury is also greatly appreciated.

Financial support received from the Air Force Office of Scientific Research and the California Institute of Technology made this research possible and is highly appreciated.

Finally, I owe my greatest debt to my wife, Su-Miau, for the love, encouragement, and understanding she gave me all along. The importance of her contribution to my graduate research is immeasurable.

## ABSTRACT

The first part of this work describes theoretical studies of a proposed longitudinal free electron laser. Stimulated coherent radiation in the infrared region is generated in the proposed laser by a relativistic electron beam passing through a periodically corrugated waveguide. The wavelength of the radiation is widely tunable by changing the electron energy.

Theoretical investigations are based on the single-electron analysis. Both linear and nonlinear treatments of the laser mechanism in a free electron laser are carried out analytically. The phenomena of homogeneous and inhomogeneous interactions, lossy gain, space-charge effect, large-signal behavior, large-gain amplification, and electron dynamics are discussed in detail.

The second part of this work consists of a theoretical study of the linear electrooptic effect. Application of a d.c. or low frequency electric field to a crystal can change its electric susceptibility at optical frequencies. This effect is known as the electrooptic or Pockel's effect. The semiclassical approach used is based on a one-energy gap model, dielectric theory, and the concept of bond-charge. A general expression is obtained for the electrooptic coefficient of a crystal and is applied to the calculation for diatomic and ternary compounds. The results are generally in good agreement with the measured values for nearly all the crystals in which the electrooptic coefficient had been determined.

## TABLE OF CONTENTS

# PART I

# THE LONGITUDINAL FREE ELECTRON LASER

# Chapter 1

# GENERAL INTRODUCTION

## 1.1 Introduction

The mechanism of laser emission can be described by the interaction of radiation and a system of electrons. The emission of radiation results in the transition between electron energy states. The radiation emitted in the process then acts back on the radiating electrons and further stimulates the emission process. According to the nature of the electron energy states, lasers are in general divided into two classes: bound and free electron lasers [1]. Bound electron lasers include most of the conventional lasers so far developed. The electrons involved in the stimulating process are in the atomic or molecular orbits which are discrete energetically. The emissive transition takes place between two well defined electron energy states. In a free electron laser the radiation interacts with a stream of electrons with a continuous energy spectrum. The transition during the stimulating process is between two states which are part of this continuum. The frequency of the radiated electromagnetic wave is determined by the electron energy. The population inversion is achieved by accelerating the electron beam such that the electron energy distribution function is shifted toward the high energies.

Some of the advantages of free-electron lasers are: First, since the frequency of radiation depends on the electron energy, the output wavelength can be tuned over a wide range by changing the accelerating voltage of the electron beam. Second, the laser medium includes the electron beam only, and the interaction region is essentially a vacuum. The problems of radiation reabsorption and material damage at high power levels are thus avoided. Due to these two unique properties, the free electron laser is a promising candidate as a high power tunable laser from the infrared to the soft x-ray region of the spectrum.

Several types of free electron lasers have been proposed. In general, they fall into two groups. In the transverse type of free electron lasers, the electron beam is deflected periodically at right angles to the propagation direction such that electrons can interact coherently with the transverse field component of the radiation field. The deflection of the electron beam is caused by either a periodic magnet or electric fields. In the longitudinal type of free electron lasers the radiation is confined to a waveguide and electrons interact with the longitudinal component of the electric field. To derive the gain equation and laser mechanism of the longitudinal free electron laser, we use a ballistic analysis and classical electrodynamics. The results are compared to those which obtain in a transverse free electron laser.

## 1.2 Previous Work on Free Electron Lasers

Schrödinger [2] was the first to discuss stimulating Compton scattering. This process involves the scattering of photons from electrons. A possible experiment was proposed by Kapitza and Dirac [3] to observe the stimulating Compton scattering of electrons from standing light waves. Almost two decades later, Motz and Nakamura [4,5] analyzed the possibility of generating coherent radiation by passing a relativistic electron beam through a periodic magnetic field. Then, Pantell, Soncini and Puthoff [6] showed how to convert a long wavelength electromagnetic wave into short wavelength radiation by stimulating scattering from a relativistic electron beam. Using the Weizsäcker-Williams approximation, Madey [7] was able to derive quantum mechanically the gain of a free electron laser in which a relativistic electron beam passes through a static helical magnetic field. At the same time, Palmer [8] gave a classical description of the energy transfer between an electron beam and electromagnetic waves inside a static helical magnet. The possibility of using such a device as a laser and a particle accelerator was discussed. Sukhatme and Wolff [9] analyzed the stimulating Compton scattering in a finite length of an interaction region.

Based on this theoretical background, the first free electron laser was demonstrated at Stanford [10] in 1972. A relativistic electron beam passes through the axis of a superconducting coil which generates a static helical magnetic field. Using this device, they were able to demonstrate the amplification of radiation at 10μ [11] and, later, the laser oscillation at 3.4μ [12].

The first experimental demonstration of stimulating amplification was followed by numerous analyses. It became apparent that a classical treatment is more appropriate to describe the free electron laser. Using coupled Maxwell's and Boltzmann equations, Hopf, Meystre,Scully, and Louisell [13,14] derived the gain classically in the small and large signal regions. Colson [15,16] used one-body electron dynamics and related the equation of motion of the electrons in a FEL to that of a pendulum. Kwan, Dawson, and Lin [17] demonstrated electron bunching and gain behavior by computer simulation. Gover and Yariv [18] gave a quantum mechanical view of the interaction between the electron beam and radiation in the single-electron and collective regions. Kroll and McMullin [19] derived the dispersion equation for the stimulated radiation and obtained the exponential gain in the limit of a large cavity. Louisell, Lam, Copeland, and Colson [20] solved the classical pendulum equation and showed the gain saturation and the evolution of the electron distribution. In a separate paper [21] they also discussed the space charge effect to a first order approximation. Baier and Milstein [22] pointed out the importance of the phase of the radiation in the free electron laser. Bernstein and Hirshfield [23] indicated that the gain depends critically on the axial momentum distribution of the beam.

In parallel with the development of the Stanford experiment, some other types of free electron lasers have been proposed utilizing different interaction mechanisms. The electron cyclotron maser was proposed to generate enhanced high power submillimeter waves [24]. Stimulated Compton scattering was observed directly using the up-conversion of microwave radiation "colliding" head-on with an electron beam [25]. The stimulated Smith-Purcell effect was studied in search of a possible means of producing coherent radiation with an electron beam passing close to a grating [26]. Replacing the periodic magnet by a corrugated waveguide in the interaction region is found theoretically to achieve low-power and high-efficiency operation [27].

## 1.3 Outline of Part I

In Chapter 2 the propagation of electrons and radiation in a waveguide is discussed. The physical origin of the electron band structure is also analyzed. The spontaneous emission of an electron passing through a corrugated waveguide is derived using the radiation theory of classical electrodynamics. Finally, the limitations on the applicability of the classical approach is discussed.

In Chapter 3 the classical approach of the single-electron picture is used in the linear theory of the longitudinal free electron laser. The gain behavior in the homogeneous and inhomogeneous interactions is then discussed. The electron dynamics are investigated to find the electron energy and phase distribution. In this analysis the phase diagram is used to describe the evolution of electrons. A two-stage system is analyzed in the discussion of electron bunching.

In Chapter 4 the nonlinear theory of the longitudinal free electron laser is introduced. The space-charge effect due to the high current density is considered. At high radiation energies it is shown that the equation of motion can be solved exactly by use of special functions in the low gain limit. A result of this solution is to show the gain saturation explicitly. In the large gain limit the growth of the field amplitude along the interaction region is described analytically.

In Chapter 5 the theoretical work is summarized and experimental conditions are discussed. The electron circulation and the possibility of electron bunching are studied.

## CHAPTER 1 - REFERENCES

1. D. Marcuse, Engineering Quantum Electrodynamics (Harcourt, New York, 1970).
2. E. Schrödinger, Annalen der Physik IV Folge 82, 257 (1927).
3. P. L. Kapitza and P.A.M. Dirac, Proc. Cambridge Phys. Soc. 29, 297 (1933).
4. H. Motz, J. Appl. Phys. 22, 527 (1951).
5. H. Motz and M. Nakamura, Ann. Phys. 7, 84 (1959).
6. R. H. Pantell, G. Soncini, and H. E. Puthoff, IEEE J. Quantum Electron. 4, 905 (1968).
7. J.M.J. Madey, J. Appl. Phys. 42, 1906 (1971).
8. R. V. Palmer, J. Appl. Phys. 43, 3014 (1972).
9. V. P. Sukhatme and P. A. Wolff, J. Appl. Phys. 44, 2331 (1973).
10. J.M.J. Madey, H. A. Schwettman, and W. M. Fairbank, IEEE Trans. Nucl. Sci. 20, 980 (1973).
11. L. R. Elias, W. M. Fairbank, J.M.J. Madey, H. A. Schwettman, and T. I. Smith, Phys. Rev. Lett. 36, 717 (1976).
12. D.A.G. Deacon, L. R. Elias, J.M.J. Madey, G. R. Ramian, H. A. Schwettman, and T. I. Smith, Phys. Rev. Lett. 38, 892 (1977).
13. F. A. Hopf, P. Meystre, M. O. Scully, and W. H. Louisell, Optics Commun. 18, 413 (1976).
14. F. A. Hopf, P. Meystre, M. O. Scully, and W. H. Louisell, Phys. Rev. Lett. 37, 1342 (1976).
15. W. B. Colson, Phys. Lett. 59A, 187 (1976).
16. W. B. Colson, Phys. Lett. 64A, 190 (1977).

17. T. Kwan, J. M. Dawson, and A. T. Lin, Phys. Fluids 20, 581 (1977).

18. A. Gover and A. Yariv, Appl. Phys. 16, 121 (1978).

19. N. M. Kroll and W. A. McMullin, Phys. Rev. A 17, 300 (1978).

20. W. H. Louisell, J. Lam, D. A. Copeland, and W. B. Colson, Phys. Rev. A 19, 288 (1979).

21. W. H. Louisell, J. Lam, and D. A. Copeland, Phys. Rev. A 18, 655 (1978).

22. V. N. Baier and A. I. Milstein, Phys. Lett. 65A, 319 (1978).

23. I. B. Bernstein and J. L. Hirshfield, Phys. Rev. Lett. 40, 761 (1978).

24. M. Friedman, D. A. Hammer, W. M. Manheimer, and P. Sprangle, Phys. Rev. Lett. 31, 752 (1973).

25. P. Sprangle, V. L. Granatstein, and L. Baker, Phys. Rev. A 12, 1697 (1975).

26. J. M. Wachtel, J. Appl. Phys. 50, 49 (1979).

27. A. Yariv and C. Shih, Optics Commun. 24, 233 (1978).

# Chapter 2

# ELECTRODYNAMICS IN A WAVEGUIDE

## 2.1 Introduction

In this chapter we study the phenomena of electrodynamics in a waveguide which is proposed to be used as an interaction region in a longitudinal free-electron laser. A wave in a corrugated waveguide cannot be described as a state having a definite momentum. Its spectrum is studied in momentum space. An electron passing through the corrugated waveguide generates spontaneous radiation. The method of "image charge" is applied to evaluate the output spontaneous power. The wave spectrum and the spontaneous process are studied classically. The coexistence of radiation and electrons results in the photon-induced electron band structure. The equation of motion using relativistic quantum mechanics is solved exactly in terms of momentum eigenstates. The continuous spectrum of the electron beam is divided into regions of stability and instability. The physical process in the electron stopping band is investigated by the Lorentz transformation and space-time invariant. Finally, the applicability of the classical approach is discussed from the point of view of the uncertainty principle. The limitation on the wave frequency and the radiated power is calculated quantitatively.

## 2.2 Wave Propagation

The prototypical configuration of the proposed longitudinal free electron laser is shown in Figure 2.1a. The main part of the device is

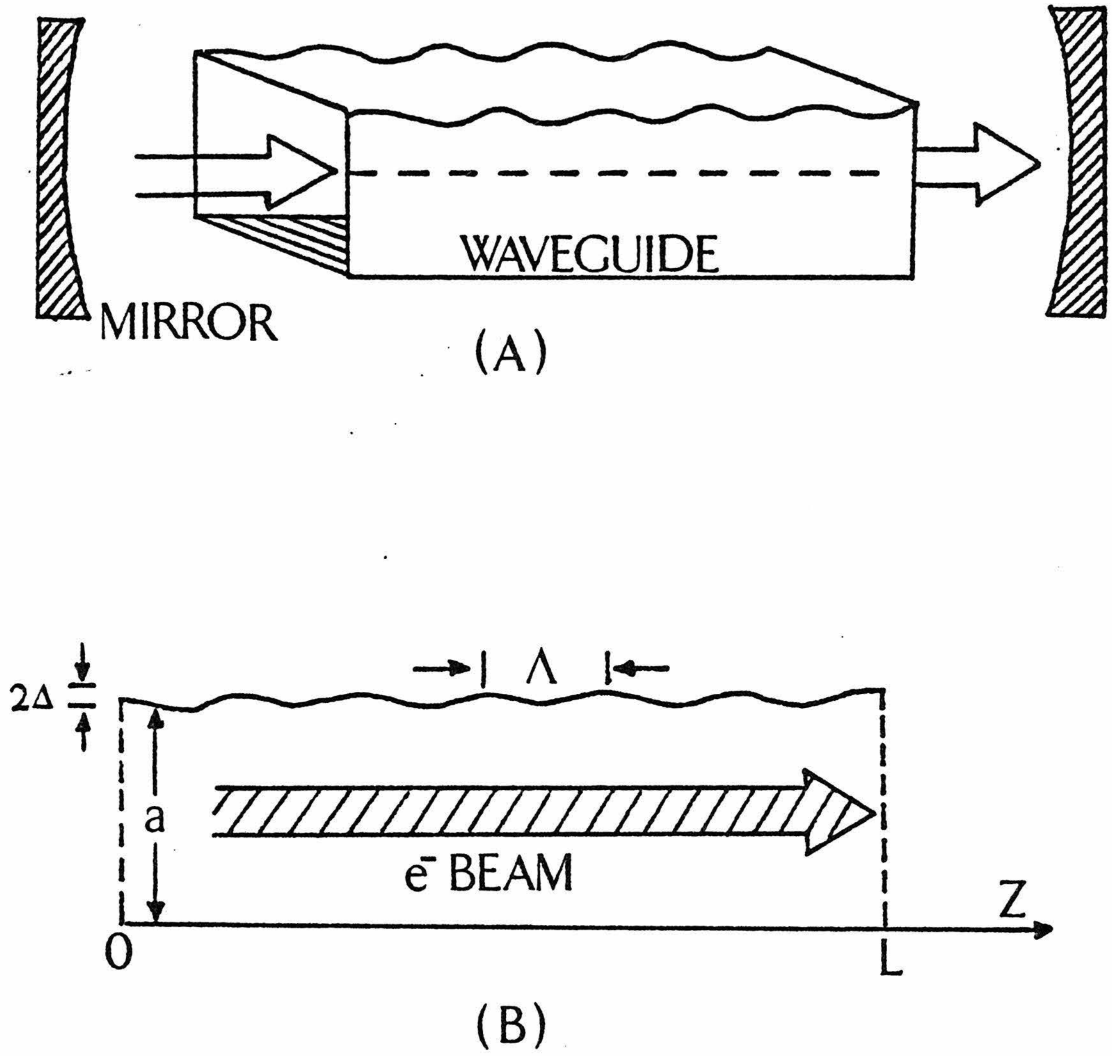

Figure 2-1: (A) Shows the prototypical configuration of a longitudinal free-electron laser; (B) defines the parameters in the device.

a corrugated metallic waveguide with the electron beam passing through its axis. Two mirrors at the two ends are adjusted to form an optical cavity. The dimensions of the waveguide are given in Figure 2.1b. L is the length of the interaction region, a is the width, $\Lambda$ is the period of the corrugation, and $\Delta$ is the depth of the corrugation. The corrugation is necessary for an efficient energy transfer from the electron beam to the radiation. Generally speaking, the periodic structure supplies the extra momentum needed when the electrons emit photons.

Since electrons interact mostly with the longitudinal component of the field, we are only interested in the waveguide TM modes. From Maxwell's equations and the boundary conditions, the TM modes of a rectangular metallic waveguide with cross section (a x b) are readily found [1]

$$\begin{aligned}
E_z &= E_o \sin(k_x x)\ \sin(k_y y)\ e^{i(kz-\omega t)} \\
E_x &= i\,\frac{kk_x}{k_c^2}\,E_o \cos(k_x x)\ \sin(k_y y)\ e^{i(kz-\omega t)} \\
E_y &= i\,\frac{kk_y}{k_c^2}\,E_o \sin(k_x x)\ \cos(k_y y)\ e^{i(kz-\omega t)} \\
B_x &= -\frac{k_o}{k}\,E_y \\
B_y &= \frac{k_o}{k}\,E_x \\
B_z &= 0
\end{aligned} \tag{2.1}$$

where $k_x = m\pi/a$; $k_y = n\pi/b$; $k_o = \omega/c$; $k_c = \sqrt{k_x^2 + k_y^2}$ is the cut-off wave vector and $k = \sqrt{k_o^2 - k_c^2}$ is the propagation wave number. m and n are integers and denote the TM modes. The fundamental mode in a square waveguide has m = n = 1. The phase velocity is

$$v_p = \omega/k = c\Big/\sqrt{1 - \frac{k_c^2}{k_o^2}} = c\Big/\sqrt{1 - \frac{\lambda^2}{2a^2}} \tag{2.2}$$

where $\lambda$ is the wavelength of the radiation in free space.

In a corrugated waveguide such as the one depicted in Figure 2.1b Maxwell's equations cannot be solved exactly. The wave propagating in this structure does not have a definite value of phase velocity and cannot interact, as a single wave, with an electron as in the case of Compton scattering. The best way to proceed is to find its spectrum in momentum space. Each component in the expansion has a definite value of momentum and can interact with electrons as a single photon.

The fractional sinusoidal variation of the waveguide width is defined as

$$\begin{aligned} a(z) &= a + \Delta \cos k'z \\ k' &= 2\pi/\Lambda \end{aligned} \tag{2.3}$$

where k' is the unit of lattice momentum. For convenience, the variation is assumed to be small and slow (i.e., $\Delta << a$ and $a << \Lambda$) and that the adiabatic approximation is valid. Physically, we expect the wave to propagate smoothly along the waveguide. The propagation constant k at z is determined locally by the waveguide width according to (2.2). With this picture the z-dependent wave equation after the separation of variables can be written as

$$\frac{\partial 2_E}{\partial z^2} + k^2(z)E = 0 \tag{2.4}$$

Using the second order WKB approximation, the solution to equation (2.4) is

$$E = E_0 \frac{\sqrt{k}}{\sqrt{k(z)}} e^{i \int k(z)\, dz} \tag{2.5}$$

k(z) can be expanded to first order in Δ

$$k(z) = k + \frac{\partial k}{\partial a} \Delta \cos k'z \tag{2.6}$$

An explicit calculation of (2.5) results in the following expression for E

$$E = E_0[1 - \frac{\Delta}{2k} \frac{\partial k}{\partial a} \cos k'z]\, e^{i[kz + \frac{\Delta}{k'} \frac{\partial k}{\partial a} \sin k'z]} \tag{2.7}$$

The wave is defined only in the interaction region which extends from -L/2 to L/2 with N periods of corrugation. Because both the field amplitude and the phase depend on position z, the wave must be decomposed into the states of definite momentum and constant amplitude. Due to the finite interaction region, the spectrum of the wave is continuous and can be found by Fourier transformation

$$E = E_0\, e^{ikz} \int_{-\infty}^{\infty} a(q)\, e^{-iqk'z}\, dq \tag{2.8}$$

where q is dimensionless. The Fourier coefficient a(q) is obtained by the inverse transform

$$a(q) = \frac{1}{2\pi} \int_{-L/2}^{L/2} [1 - \frac{\Delta}{2k} \frac{\partial k}{\partial a} \cos k'z]\, e^{-i[qk'z - \frac{\Delta}{k'} \frac{\partial k}{\partial a} \sin k'z]} dz$$

$$= H_q(s) - \frac{\Delta}{4k} \frac{\partial k}{\partial a} [H_{q+1}(s) + H_{q-1}(s)] \tag{2.9}$$

where $s = \frac{\Delta}{k'} \frac{\partial k}{\partial a}$

$$H_q(s) = \frac{\sin N q\pi}{N \sin q\pi} \mathcal{J}_q(s) \tag{2.10}$$

$\mathcal{J}_q(s)$ is the Anger function [2], which is defined by the integral representation:

$$\mathcal{J}_q(s) = \frac{1}{\pi} \int_0 \cos(qx - s \sin x)\, dx \tag{2.11}$$

This representation is similar to the one for the Bessel function $J_q(s)$, but q is not necessarily an integer. The dependence of the Auger function on its index and argument is shown in Figure 2.2. The difference between the Auger function and Bessel function is apparent, especially when s = 0. At s = 0, $J_q(0) = 0$ only when q is an integer and not equal to zero, while $\mathcal{J}_q(0) = 0$ except when q = 0. As $\Lambda >> \lambda$, only the first term on the right side of (2.9) dominates. In Figure 2.3 we show the spectrum of the wave in a waveguide having ten corrugation periods (i.e., N = 10). The contribution of the first harmonic to the total wave reaches its maximum at s = 1.84. The spectrum amplitude is enhanced when q is close to an integer. The width of the enhanced peak is proportional to 1/N. Usually, N is very large and the spectrum becomes discrete,

$$a(q) = \mathcal{J}_q(s)\, \delta(q-n), \text{ n is an integer} \tag{2.12}$$

In this limit the electric field becomes

$$E = E_0 \sum_{n=-\infty}^{\infty} J_n(s)\, e^{i(k+nk')z} \tag{2.13}$$

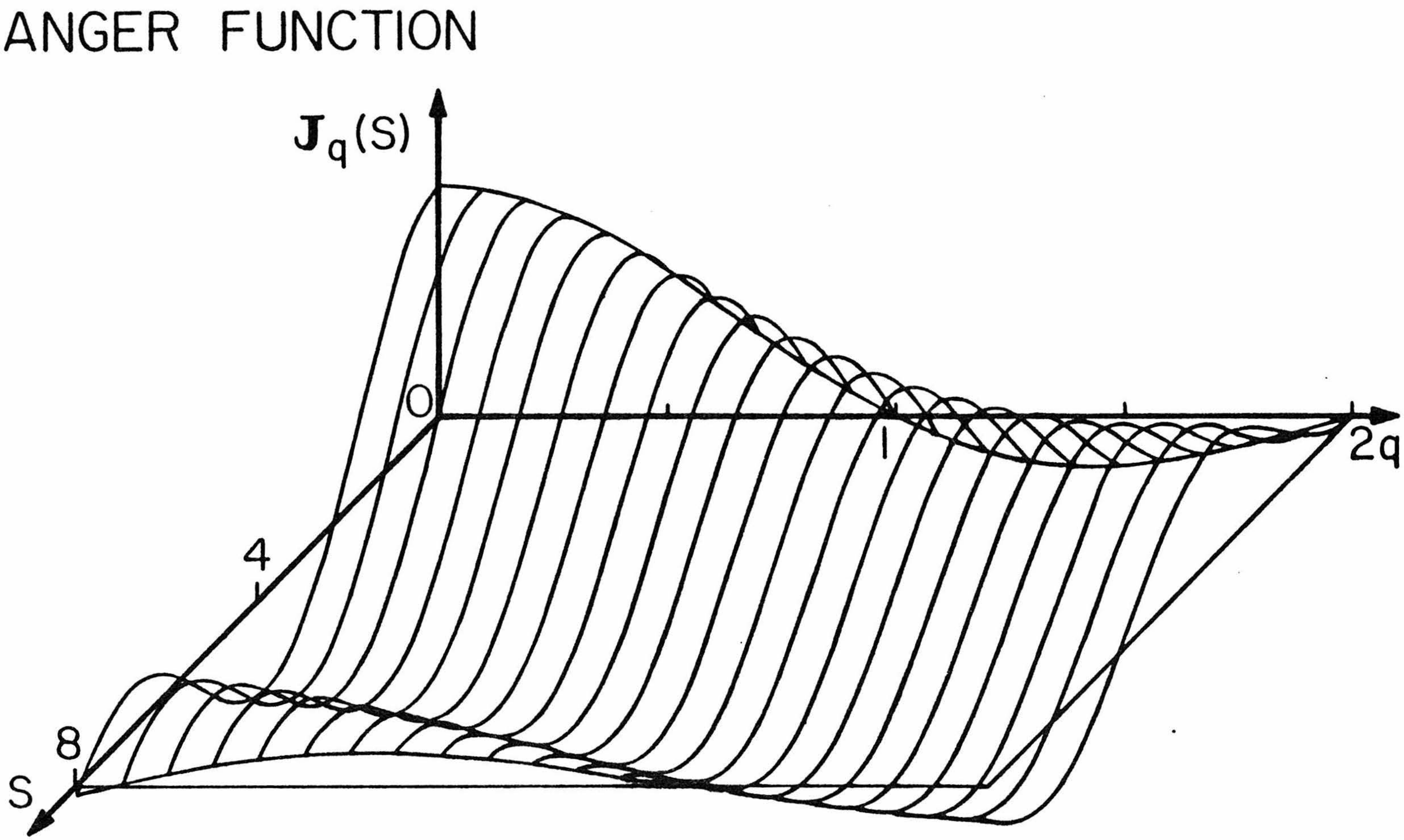

Figure 2-2: Anger function $\mathbf{J}_q(S)$.

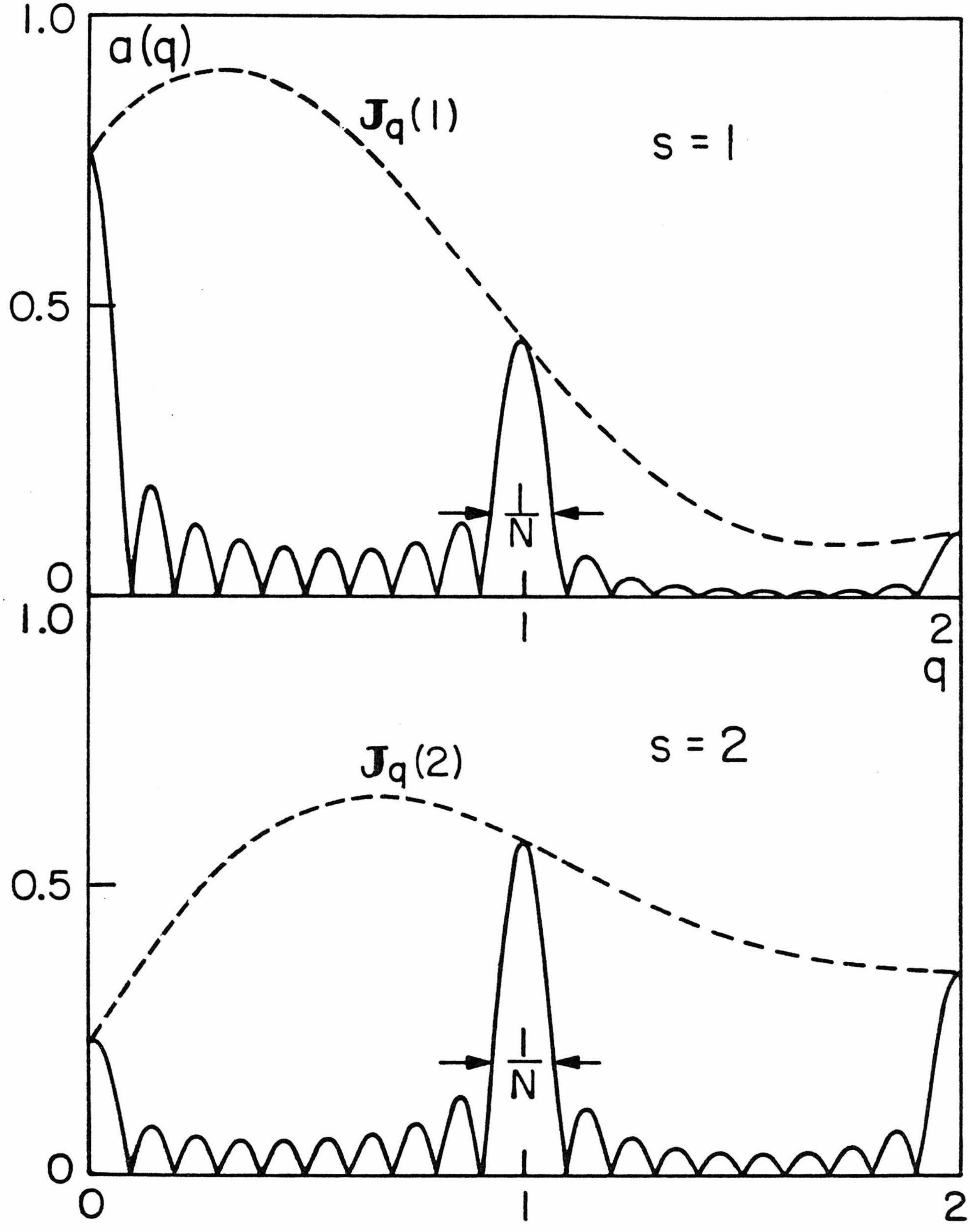

Figure 2-3: The spectrum of a propagating wave in a waveguide

In the longitudinal free electron laser, the stimulating interaction takes place mostly between electrons and the first harmonic. The growth and instability of the first harmonic field distributes its energy among other harmonics via the interaction of the wave with the periodic structure. The fractional intensity of radiation involved in the stimulating process is thus only $J_1^2(s)$ of the total intensity.

The interaction mechanism of a transverse free electron laser is somehow different from that of the longitudinal laser described above. Although the static helical magnet supplies the necessary momentum compensation, the physical process is usually understood as the interaction between transversely deflected electrons and the transverse electric field of the radiation. It seems that the stimulating strength should be much larger because the total wave stimulates the emission process. But the deflection of electrons by the magnetic field is so small that the efficiency of the energy transfer is actually much lower than the efficiency in the longitudinal device.

## 2.3 Spontaneous Radiation

The spontaneous radiation is due to the classical acceleration or deceleration of electrons in the waveguide. The acceleration and deceleration are the result of interactions between the electrons and the corrugated metal wall. In general, the interaction is very small so that the velocity of the electron does not change significantly during the flight. A section of the symmetrically corrugated waveguide is shown in Figure 2.4. The best and simplest way to describe the interaction is to use the method of "image charge." Based on the adiabatic

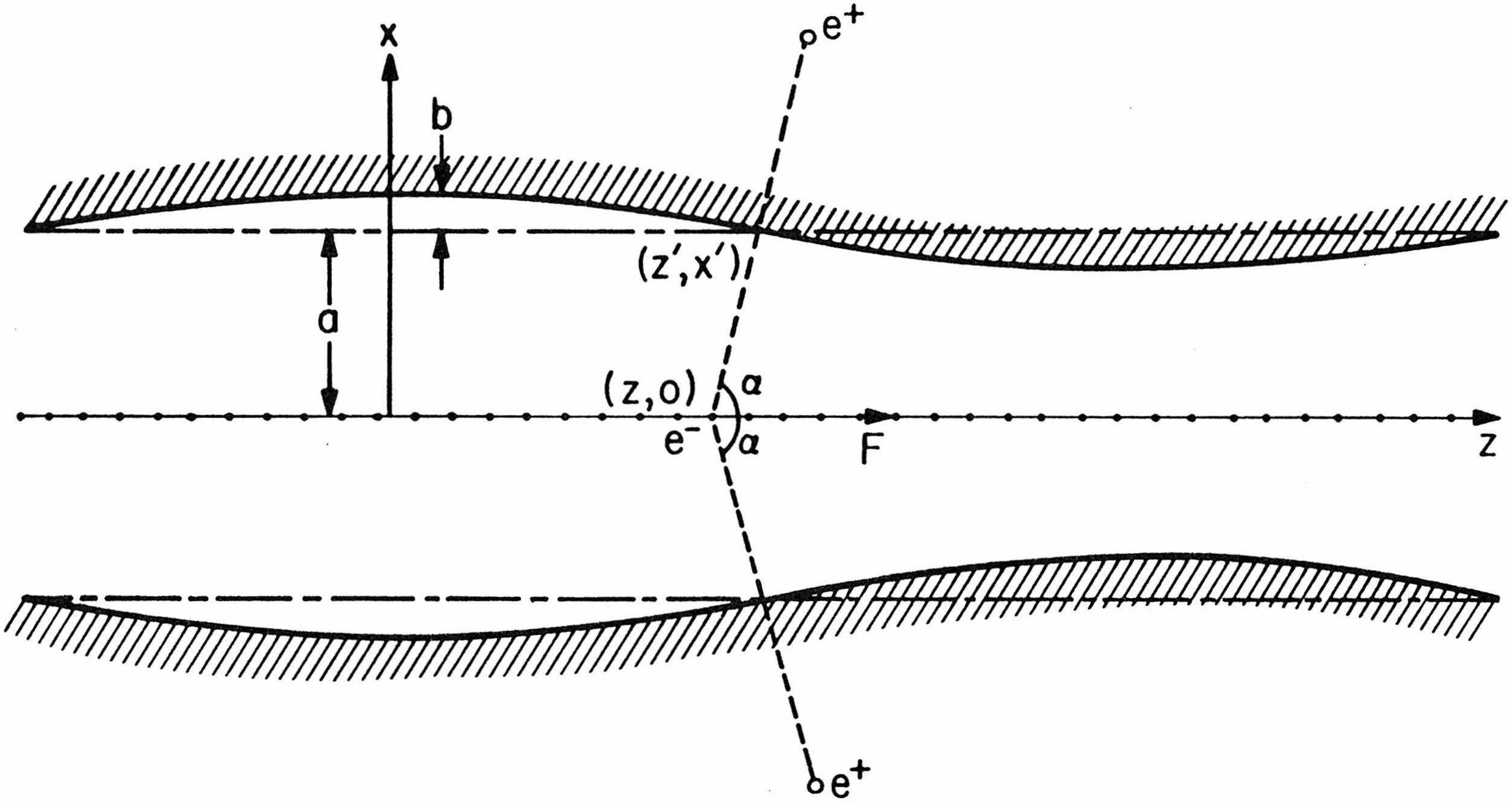

Figure 2-4: A corrugated waveguide. The acceleration and deceleration of the electron is due to its interaction with the metallic wall which can be represented by image charges $e^+$.

approximation, we assume that an electron at any position on the axis has one and only one image charge on each wall. The line connecting the electron and the image charge is perpendicular to the wall. This assumption is true only when the radius of curvature at any point on the wall is larger than the width of waveguide, a. Assuming a wall profile

$$x' = a + b \cos k'z' \tag{2.14}$$

The criterion is equivalent to

$$k'^2 ab \leq 1 \tag{2.15}$$

which is always obeyed in a practical device. The interaction between an electron and the wall can be simulated by the force between the electron and the positive image charge. When the electron travels down the waveguide, the image charge swings along the path parallel to the axis. The force due to the excursive motion causes the acceleration and deceleration of the electron along the periodic structure.

From Figure 2.4, the force in the z-direction is

$$F = \frac{2e^2 \cos \alpha}{4r^2} \tag{2.16}$$

where

$$r = \sqrt{(z - z')^2 + x'^2}$$

$$= (a + b \cos k'z') \sqrt{1 + k'^2 b^2 \sin^2 k'z'} \tag{2.17}$$

$$\cos \alpha = \frac{k'b \sin k'z}{\sqrt{1 + k'^2 b^2 \sin^2 k'z}} \tag{2.18}$$

To the first order in k'b, the force is simplified as

$$F = e^2 k'b \frac{\sin k'z}{2(a + n \cos k'z)^2} \qquad (2.19)$$

z in the right side of (2.19) can be replaced by $v_0 t$. The time-dependence velocity is found to be

$$v = v_0 + \frac{e^2}{2m\gamma^3 v_0} \left[\frac{1}{a+b} - \frac{1}{a + b \cos k'v_0 t}\right] \qquad (2.20)$$

The radiation intensity generated by this interaction per unit solid angle dΩ and per unit frequency interval dω is given [1]

$$\frac{dI}{d\Omega d\omega} = \frac{e^2 \omega^2}{4\pi^2 c^2} \left| \int_{-\infty}^{\infty} \hat{n} \times (\hat{n} \times \vec{v})\, e^{i\omega[t - \hat{n} \frac{z}{c}]}\, dt \right|^2 \qquad (2.21)$$

Practically, it is impossible to observe the angular dependence of the radiation intensity in a waveguide. Even the frequency dependence of the intensity at the output does not follow (2.21). From (2.2), it can be seen that k and ω are not independent of each other for a propagating waveguide mode. Only that part of the radiation which obeys the condition for the guided mode can be detected at the output. The total power which can be detected is less than the total power loss of electrons. To estimate an upper limit on the spontaneous radiation intensity, we integrate (2.21) over angle and frequency to obtain the total power loss of electrons.

If n is defined in the direction of (θ,φ), we have

$$\hat{n} \times (\hat{n} \times \vec{\beta}) = \beta[\sin\theta \cos\theta(\cos\phi\, \hat{x} + \sin\phi\, \hat{y}) - \sin^2\theta\, \hat{z}] \qquad (2.22)$$

and

$$t - \hat{n}\cdot\frac{z}{c} = t - \frac{z}{c}\cos\theta \qquad (2.23)$$

Due to the collective interference of waves, the radiation pattern consists of many spectral lines. The fundamental line is the radiation from the electron acceleration at the fundamental frequency $k'v_o$. From (2.21), the electron velocity can be resolved into different orders by expanding it in power series of b/a. Considering only the fundamental line, we have

$$v = v_o - \frac{e^2b}{2m\,{}^3v_oa^2}(1 - \cos k'v_ot) \qquad (2.24)$$

The interaction region is finite, so the integral in (2.21) extends only from $t = -N\Lambda/2v_o$ to $t = N\Lambda/2v_o$. The radiation intensity is then found as

$$\frac{dI}{d\Omega d\omega} = A\omega^2\sin^2\theta\left[\int \cos k'v_ot\, e^{i(1 - \beta\cos\theta)\omega t}\, dt\right]^2 \qquad (2.25)$$

$$= A\omega^2\sin^2\theta\left[\frac{\frac{N\Lambda\omega}{2v_o}(1 - \beta\cos\theta) + N\pi]}{(1 - \beta\cos\theta) + k'v_o} + \frac{\sin[\frac{N\Lambda\omega}{2v_o}(1 - \beta\cos\theta) - N\pi]}{\omega(1 - \beta\cos\theta) - k'v_o}\right] \qquad (2.26)$$

where

$$A = \frac{1}{c}\left(\frac{e^3b}{4\pi m\gamma^2v_oca^2}\right)^2 \qquad (2.27)$$

The functional form, sin Nx/x becomes sharply peaked at x = 0 when N is very large. Therefore, the radiation frequency of the fundamental line is

$$\omega = \frac{v_0 k'}{1 - \beta \cos \theta} \tag{2.28}$$

and the first term in (2.26) is highly suppressed. The radiation spectrum of the fundamental line is then obtained as

$$\frac{dI}{d\Omega d\omega} = Ak'^2 v_0^2 \sin^2\theta \frac{\sin^2[\frac{N\Omega\omega}{2v_0}(1 - \beta \cos \theta) - N\pi]}{[(1 - \beta \cos \theta)\omega - k' v_0]^2 (1 - \beta \cos \theta)^2} \tag{2.29}$$

By integrating (2.29) over $\omega$, we obtain the angular distribution

$$\frac{dI}{d\Omega} = ANk' \ v_0 \pi^2 \frac{\sin^2\theta}{(1 - \beta \cos \theta)^3} \tag{2.30}$$

Tht total power is found by the integration of (2.30) over $\Omega$ and divided by the flight time $T = N\Lambda/v_0$.

$$P = A\pi^2 k'^2 v_0^2 \frac{1}{\beta^3}(\ell n \ \frac{1-\beta}{1+\beta} + \frac{4\beta - 1}{1-\beta^2}) \tag{2.31}$$

When $\beta \rightarrow 1$, the total power approaches

$$P = A\pi^2 k'^2 v_0^2 \ 3\gamma^2 = \frac{3\pi}{4} \frac{\gamma_0^3 \ b^2}{\gamma^4 \Lambda^2 a^4} mc^3 \tag{2.32}$$

In the last step, (2.27) has been used for the expression of A. $\gamma_0 = e^2/mc^2 = 2.82 \times 10^{-13}$ cm, is the classical electron radius. In the transverse free electron laser, the spontaneous power is proportional to $\gamma^2$, while it is proportional to $\gamma^{-4}$ in (2.32). This dramatic drop in the $\gamma$-dependence comes from two sources: First, the dependence of the radiation spectrum on $\sin^2\theta$ in (2.29) introduces a factor of $\gamma^{-2}$ in the result. Second, it is more difficult to accelerate electrons longitudinally than transversely, which adds a factor of $\gamma^{-2}$ in the expression of v.

From (2.32), the total power of the fundamental line emitted spontaneously by an electron passing through the corrugated waveguide is calculated to be about $10^{-16}$ eV. For a continuous electron beam of 1 A, the emitted power is only about $10^{-16}$ watts which is much lower than the value in the transverse free electron laser.

## 2.4 Electron Band Structure

The energy spectrum of a free electron beam is continuous. However, in the presence of an electromagnetic field, the spectrum is modified and generates electron band structure. It has been known [3] theoretically that the presence of photons can induce band structure in a medium. The physical origin of this phenomenon is based on the fact that $k \neq \omega/c$ in such a medium. This effect should be observable for an em wave propagating in a waveguide, even though the waveguide cannot be represented by a simple index of refraction, n. Therefore, we expect that an electron beam passing through a waveguide containing an electromagnetic field should display a band structure in its energy spectrum. The following analysis follows from the relativistic and quantum mechanical points of view. All physical quantities are written in 4-vector notation and the dimensional choice, $\hbar = c = 1$, for convenience. A four-vector $A^\mu$ represents

$$A^\mu \equiv (A^0, A^1, A^2, A^3) \tag{2.33}$$

For example, the wave-vector $k^\mu$ is

$$k^\mu \equiv (\omega, k^1, k^2, k^3) \tag{2.34}$$

and the momentum-vector $p^\mu$ is

$$p^\mu \equiv (E,p^1,p^2,p^3) \tag{2.35}$$

The usual space vector is denoted as $\vec{A}$, such as $\vec{k}$ and $\vec{p}$. A four-vector $A^\mu$ transforms like a vector in four space. The scalar product of two vectors is defined as

$$\begin{aligned} A\cdot B &\equiv A^\mu B_\mu \equiv g_{\mu\nu}A^\mu B^\nu \\ &\equiv A^o B^o - (A^1B^1 + A^2B^2 + A^3B^3) \end{aligned} \tag{2.36}$$

$$g_{\mu\nu} \equiv \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

and is a scalar in four space. $A^o$ is usually known as a "time component" because

$$x^\mu \equiv (t,x^1,x^2,x^3) \tag{2.37}$$

The band structure can be solved directly from the relativistic equation of motion with a "minimal coupling" to the electromagnetic field [4] (i.e., $p^\mu \rightarrow p^\mu - eA^\mu$ or $\partial^\mu \rightarrow \partial^\mu - ieA^\mu$). For a spin-1/2 particle, such as an electron, the Dirac equation should be used to account for two spin states. However, the spin complication is not essential in obtaining the electron spectrum. For a simpler demonstration, the equation describing a scalar particle is solved. The spin-induced band splitting will be discussed qualitatively at the end of this section. Because the criterion for generating band structure is quite general, we will not limit ourselves to the case of an electron in a waveguide. The analysis follows a general guideline. The physical interpretations are then given for (i) $v_p > c$, and (ii) $v_p < c$, separately.

Consider a spinless particle in a given electromagnetic field. The wavefunction of the particle, $\psi$, is the solution of the Klein-Gordon equation

$$[(P - eA)^2 - m^2]\psi = 0 \tag{2.38}$$

In the x-representation, the equation with the Lorentz condition $\partial_\mu A^\mu = 0$ is rewritten in a covariant form

$$(\partial_\mu \partial^\mu + 2ie\, A^\mu \partial_\mu - e^2 A^\mu A_\mu + m^2)\psi = 0 \tag{2.39}$$

The vector potential in the equation is assumed to be only a function of the single variable $\phi \equiv k\cdot x = \omega t - \vec{k}\cdot\vec{x}$. Without the electromagnetic field coupling, the solution of (2.38) is that of a plane wave, $\psi = e^{ip\cdot x}$ with $p\cdot p = m^2$. It is reasonable to write the solution for (2.39) as

$$\psi = e^{-ip\cdot x}\, F(\phi) \tag{2.40}$$

Inserting (2.40) into the Klein-Gordon equation, we obtain

$$k\cdot kF''(\phi) - 2iP\cdot k\, F'(\phi) + (2eP\cdot A - e^2 A\cdot A)\, F(\phi) = 0 \tag{2.41}$$

The order of the differential equation (2.41) depends on the value of $k\cdot k$. If $k\cdot k = 0$, it is only a first order equation and the field corresponds to a freely propagating electromagnetic wave in vacuum as $\omega = |\vec{k}|$. When $k\cdot k \neq 0$, it becomes a second order equation and two cases can be distinguished

(i) $k\cdot k > 0$: This case indicates $\omega > |\vec{k}|$ and $v_p > c$. This kind of electromagnetic field can be generated either in a plasma medium ($n < 1$) or in a hollow waveguide ($k\cdot k = \omega_c^2$).

(ii) $k\cdot k < 0$: This case indicates $\omega < |\vec{k}|$ and $v_p < c$. Such electromagnetic fields can be found in an ordinary medium ($n > 1$, $k\cdot k = \omega^2(1 - n^2)$).

The vector field $A^\mu \equiv (\phi, A^1, A^2, A^3)$, in general can be written as

$$A^\mu = f^\mu \cos\phi + g^\mu \sin\phi \qquad (2.42)$$

where $f^\mu$ and $g^\mu$ are space-like vectors ($f\cdot f < 0$, $g\cdot g \quad 0$) and orthogonal to each other ($f\cdot g = 0$). The polarization of the field in 4-space is then defined by the relative magnitude of $f^\mu$ and $g^\mu$ . If $f\cdot f = g\cdot g$, $A^\mu$ is circularly polarized. If $f^\mu = 0$ or $g \; = 0$, $A^\mu$ is linearly polarized. Otherwise, $A$ is a field of elliptic polarization.

For a real photon, $k\cdot k = 0$, equation (2.41) is solved to obtain

$$F(\phi) = \exp -i\Big[\frac{e^2}{2p\cdot k}\Big(\frac{f\cdot f + g\cdot g}{2} + \frac{f\cdot f - g\cdot g}{4}\sin 2\phi\Big) - \frac{e}{p\cdot k}(p\cdot f \sin\phi - p\cdot g\cos\phi)\Big] \qquad (2.43)$$

and

$$\psi = e^{ip\cdot x}\, F(\phi) = e^{iP_{eff}\cdot x} \qquad (2.44)$$

Usually, $P_{eff}$ is a $x^\mu$-dependent quantity. Consider a special case: The wave is circularly polarized and propagates in the z direction. The electron also travels along z. Then we have $f\cdot f = g\cdot g$ and $p\cdot f = p\cdot g = 0$

In this case the effective momentum is a constant

$$p^{\mu}_{eff} = p^{\mu} + \frac{e^2 f\cdot f}{2p\cdot k} k^{\mu} \tag{2.45}$$

Now we consider the case when $k\cdot k \neq 0$. In order to eliminate the first derivative term, we choose

$$F(\phi) = G(\phi)\ \exp[i(p\cdot k/k\cdot k)\phi\ ] \tag{2.46}$$

(2.4.1) reduces to an equation for $G(\phi)$

$$G'' + [\frac{(p\cdot k)^2}{(k\cdot k)^2} + \frac{2eA\cdot P}{k\cdot k} - \frac{e^2 A\cdot A}{k\cdot k}]\,G = 0 \tag{2.47}$$

For a wave of circular polarization, (2.47) can be written as

$$\frac{d^2 G}{d\eta^2} + (r + q\cos 2\eta)\,G = 0 \tag{2.48}$$

where

$$\eta = \frac{1}{2}(\phi - \tan^{-1}\frac{p\cdot g}{p\cdot f}) \tag{2.49}$$

$$r = 4[\frac{(p\cdot k)^2}{(k\cdot k)^2} - \frac{e^2 f\cdot f}{k\cdot k}] \tag{2.50}$$

$$q = \frac{8e}{k\cdot k}\sqrt{(p\cdot f)^2 + (p\cdot g)^2} \tag{2.51}$$

Equation (2.48) is a Mathieu's equation [2,5]. It has the general solution

$$G(\eta) = c_1 w(\eta)e^{\nu\eta} + c_2 w(-\eta)e^{-\nu\eta} \tag{2.52}$$

where $c_1$ and $c_2$ are constants, $w(\eta)$ is a periodic function with period $\pi$, and $\nu(r,q)$ is a characteristic root determining the stability of the solution. The characteristic root, $\nu$, can be either pure imaginary or a complex value. If $\nu$ is an imaginary value, the particle can propagate freely through the electromagnetic field with an effective momentum

$$P_{eff} = P^{\mu} - (\frac{p\cdot k}{k\cdot k} \pm |\nu|)\ k^{\mu} \qquad (2.53)$$

If $\nu$ has a nonvanishing real part, the wavefunction would include a factor of $e^{\pm\phi(\mathrm{Re}\ \nu)}$ which is a nonstable function in time and space. According to the nature of the wavefunction, the solution on the r-q plane can be divided into two regions, stable and unstable, separated by characteristic curves. In Figure 2.5, the stability chart of Mathieu's equation is shown. The shaded area is for the region of stability. The features of the chart can be summarized as

(i) It is symmetric upon $q \to -q$;

(ii) For $q = 0$, stability is restricted to $r \geq 0$, instability to $r < 0$;

(iii) For $|q| >> 1$, the zone of stability becomes very narrow and centers about

$$r = -|q| + \sqrt{2}\ (2n+1)\ \sqrt{|q|} + o(q^{\circ})$$

with the bandwidth being

$$w = \sqrt{2/\pi}\ 2^{\frac{7}{2}n + \frac{17}{4}}\ \frac{|q|^{\frac{n}{2}+\frac{3}{4}}\ e^{-\sqrt{8|q|}}}{n!} \qquad (2.55)$$

(iv) The stable region is confined to $r \geq |q|$.

In any practical situation, the solution can be placed anywhere on the stability chart by choosing proper values of $p^{\mu}$, $k^{\mu}$, $f^{\mu}$, and $g^{\mu}$.

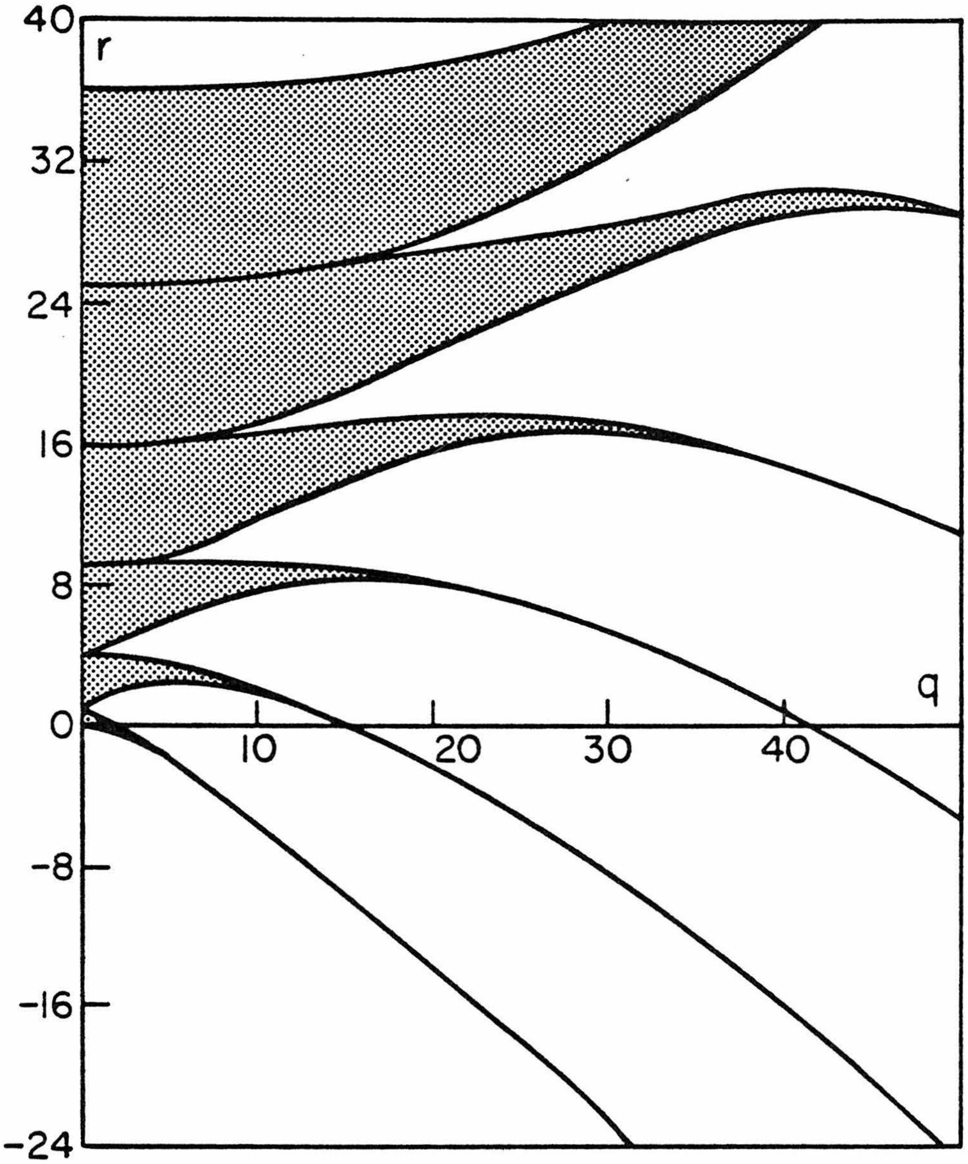

Figure 2-5: The stability chart of the Mathieu's function. The shaded areasare the regions of stable solution.

Next we treat the case of the waveguide. The necessary condition for the applicability of the above analysis is that we must find a four-vector A which can describe the field of a TM mode (2.1). The electric and magnetic fields are related to the vector field as:

$$\vec{B} = \vec{\nabla} \times \vec{A} \quad ; \qquad \vec{E} = -\vec{\nabla}\Phi - \frac{\partial \vec{A}}{\partial t} \tag{2.56}$$

or in tensor form

$$F^{\mu\nu} = \partial^{\mu}A^{\nu} - \partial^{\nu}A^{\mu} \tag{2.57}$$

We can solve $\Phi$ and $\vec{A}$ by inserting (2.1) into (2.56). The established forms for $\Phi$ and $\vec{A}$ can be proven to be a four-vector by showing that they obey the Lorentz condition

$$\partial^{\mu}A_{\mu} = 0 \tag{2.58}$$

The vector field A is then obtained up to an arbitrary constant c:

$$A^{\mu} \equiv (\Phi, A^1, A^2, A^3)$$

with

$$\begin{aligned} A^1 &= k_x c E_o \cos k_x x \sin k_y y \; e^{-i\phi} \\ A^2 &= k_y c E_o \sin k_x x \cos k_y y \; e^{-i\phi} \\ A^3 &= i(kc - \frac{k_o}{k_c^2}) E_o \sin k_x x \sin k_y y \; e^{-i\phi} \\ \Phi &= i(k_o c - \frac{k}{k_c^2}) E_o \sin k_x x \sin k_y y \; e^{-i\phi} \end{aligned} \tag{2.59}$$

Comparing this with the general expression for $A^{\mu}$ in (2.42), we find

$f^0 = f^3 = g^1 = g^2 = 0$. The 4-vectors $f^\mu$ and $g^\mu$ are orthogonal because $f \cdot g = 0$. The sense of polarization is then determined by c. In general, $A^\mu$ is a field of elliptical polarization, but it becomes linearly polarized when $c = 0$ and circularly polarized when $c = k_c^{-2}/\sqrt{2}$. The assumption of a circularly polarized wave is not absolutely necessary for the analysis. It does, however, lead usually to a simpler result. If $A^\mu$ is a wave of linear polarization, the equation for $G(\eta)$ becomes a Hill-type equation which includes both $\cos \eta$ and $\cos^2\eta$ in the coefficient. Obtaining the solution of a Hill's equation is much more complicated. Qualitatively, the solutions can also be divided into regions of stability and instability according to its parameters. The characteristic curves of the Hill's equation have not been well defined. So it is more convenient to treat the problem with a circular wave.

Using (2.50), (2.51), and (2.59) with $c = k_c^{-2}/\sqrt{2}$, we have calculated r and q to be

$$r = \frac{2}{k_c^4} [2(k_o E - pk)^2 - e^2 E_o^2] \tag{2.60}$$

$$q = \frac{8eE_o}{k_c^4} [k_o(\frac{E}{\sqrt{2}} + p) - k(E + \frac{p}{\sqrt{2}})] \tag{2.61}$$

Usually q is a very large number. r can be negative or positive, very small or very large, depending on the values of parameters. At $\lambda = 10\mu$, $a = 50\mu$, and $\gamma \simeq 5$, the value of r changes sign around the field intensity of $E \simeq 10^{10}$ V/m.

For most cases, q is very large and $r < 0$, and the electron spectrum resides mostly in the regions of instability. A physical interpretation of the "instability" is given in the following argument. For simplicity, let us assume that the electron momentum does not change. Therefore, there is a possibility that the electron in the region of instability emits or absorbs photons and changes its momentum and energy until it enters the region of stability. This corresponds to the process of Compton scattering and bremstrahlung.

If the electron stays in the same momentum state, we note that the factor

$$e^{\pm\nu\eta} \sim e^{\pm\nu(\omega t - \vec{k}\cdot\vec{x})} \tag{2.62}$$

is similar to the wave decay in time or wave attenuation in space. However, the simultaneous existence of t and $\vec{x}$ complicates the interpretation of (2.62). The best way to solve this problem is to find a Lorentz transformation such that the system in the new frame can be interpreted easily. We have noticed that $k\cdot x$ is an invariant quantity under transformation. Therefore, for a time-like wave with $k\cdot k > 0$, all scalars are kept positive. A transformation exists which can make $k\cdot x$ equal to $\omega' t'$ where the space part disappears. As a consequence, only the electric field is present in this frame. For a space-like wave with $k\cdot k < 0$, the time part of a scalar product and the electric field can be eliminated by a Lorentz transformation. In summary,

(i) A time-like electromagnetic wave is equivalent to a <u>time varying homogeneous electric field</u>.

(ii) A space-like electromagnetic wave is equivalent to a <u>constant periodic magnetic field</u>.

The propagation of an electron in a time-varying electric field or a periodic magnetic field is demonstrated clearly in Figure 2.6. The

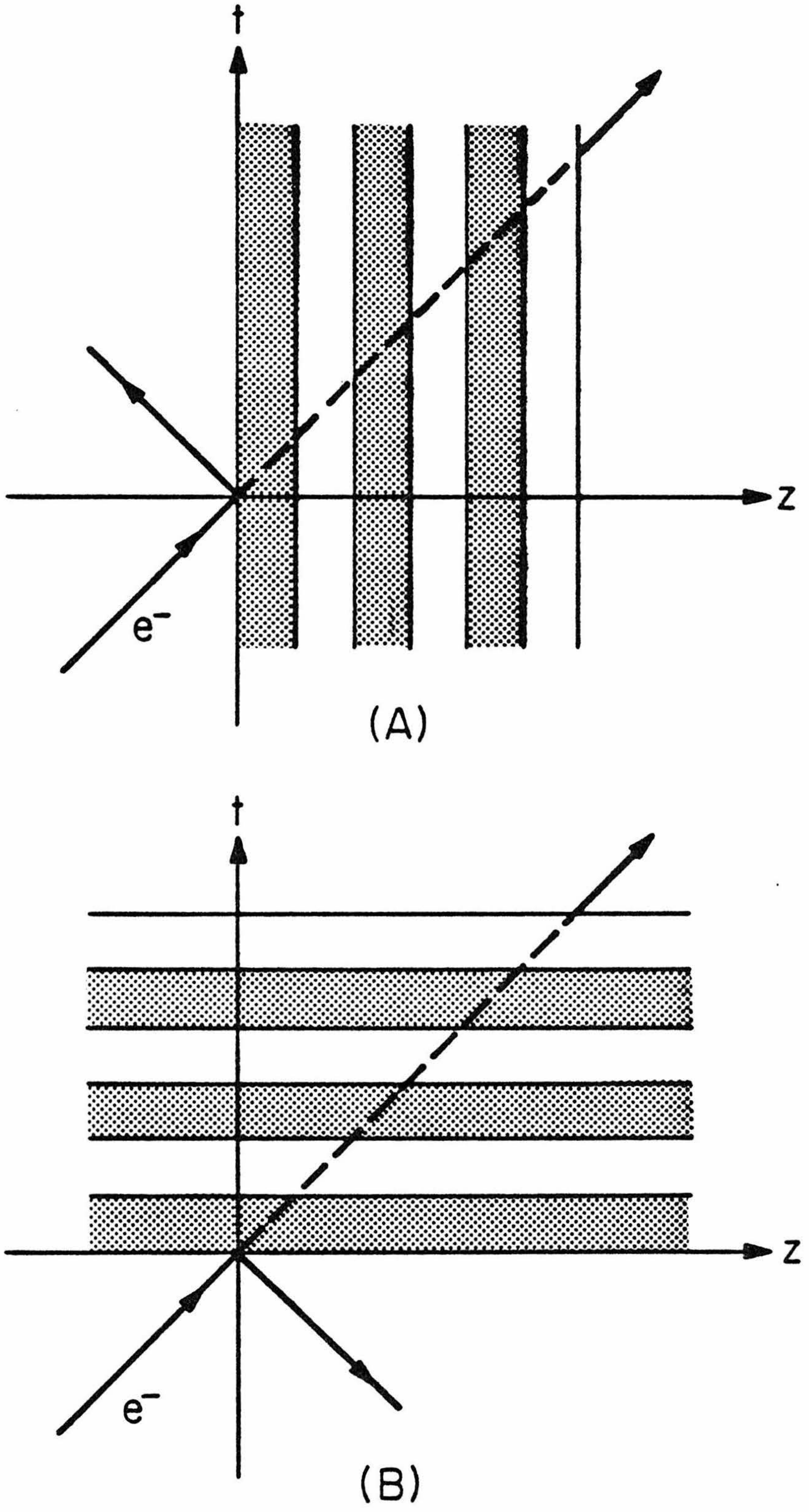

Figure 2-6: The reflection and transmission of electrons by periodic structure in space (A) and in time (B).

path of the electron is shown on a space-time diagram. On the diagram both the electric field and magnetic field can be represented by a periodic structure, but one is in time and the other in space. The transmission and reflection of electrons are easy to understand graphically. In Figure 2.6a, the electron is reflected by a periodic magnetic field and propagates in the opposite direction. The situation is very similar to Bragg reflection. In Figure 2.6b, the electron is reflected by a periodic (in time) electric field and propagates in the direction of "reverse time!" An electron propagating into the past can be interpreted as a positron propagating into the future. Therefore, the interaction between an electron and a time-varying electric field results in the creation of electron-positron pairs. From this picture, we arrive at a very interesting conclusion. The physical process in the unstable region of the electron band structure is the total reflection for a space-like electromagnetic wave and the creation of electron-positron pairs for a time-like electromagnetic wave.

Similar phenomena have been discussed for an optical wave interacting with a dielectric medium. It is well known [6] that an electromagnetic wave has a band structure when interacting with a spatially periodic dielectric medium. The Bragg reflection occurs in the forbidden band. This case is similar to the electron passing through a periodic magnetic field or a space-like electromagnetic wave. When an electromagnetic wave propagates in a time-varying dielectric medium it generates a time-reversal (or conjugated)wave [7]. This is similar to the creation of positrons when electrons pass through a time-varying electric field or a time-like electromagnetic wave.

If the electron spin is considered, the Dirac equation must be solved to find out the wavefunction and define the stable and unstable regions. The solution of the Dirac equation with minimal coupling

$$(\not{P} - e\not{A} - m)\psi = 0 \qquad (2.63)$$

is [6]

$$\psi_p(x) = e^{iP^{\mu}_{eff}x_{\mu}} F_{\pm}(\phi)\phi_p \qquad (2.64)$$

where $\phi_p$ is a constant spinor satisfying $(P - m)\phi_p = 0$, $F_{\pm}(\phi)$ is different for spin-up and spin-down, and depends only on $e^{i\phi}$ and $e^{-i\phi}$ which have no effect on the band structure. The effective momentum is

$$P^{\mu}_{eff} = p^{\mu} - [\frac{p\cdot k}{k\cdot k} \pm \frac{1}{2} + \sqrt{\frac{(p\cdot k)^2}{(k\cdot k)^2} \pm \frac{1}{2}\frac{1}{(k\cdot k)} - \frac{e^2A\cdot A}{k\cdot k}]}\; k^{\mu} \qquad (2.65)$$

The upper and lower signs refer to different spin states. The value under the square-root sign can be positive or negative. When it is negative, the region corresponding to the given $P^{\mu}$, $K^{\mu}$, and $A^{\mu}$ belongs to the region of instability. It can be seen clearly that the band structures for opposite spin orientations are slightly different. This effect has been proposed [3] for selecting the spin states in an electron beam.

## 2.5 Quantum Limitations

In Section 2.3, we have used the classical approach to describe the spontaneous radiation of electrons in a periodic waveguide. In Section 2.4, we solved the electron band structure quantum mechanically.

Before we describe the stimulated process, a discussion about the advantages and the applicability of the two approaches becomes necessary.

Quantum mechanics usually describes physical phenomena more accurately. However, it involves mathematical complications and for their solution only the perturbation method is available. Although we have found the electron band structure exactly, it is still impossible to describe the interaction in detail.

The classical method, including mechanics and electrodynamics, is well developed. Its mathematics is readily understood and the results are easy to interpret. But the classical approach has its limitations. It describes an electron as a particle and a photon as a pure wave. Therefore, several physical quantities have to be specified precisely such as the momentum P and position x of an electron, as well as the amplitude E and phase $\phi$ of the wave.

Consider an electron with momentum P at position x. Because P and x form a conjugate pair of operators, quantum mechanics shows that it is impossible to measure both quantities very accurately. The uncertainties in the measurement follow the uncertainty principle,

$$\Delta p \ \Delta x \geq \hbar/2 \qquad (2.66)$$

or, in terms of the dimensionless electron energy $\gamma$ ,

$$\Delta\gamma \ \Delta x(\mathring{A}) \geq 2\sqrt{1 - \frac{1}{\gamma^2}} \times 10^{-3} \qquad (2.67)$$

For the classical approach to be valid, $\Delta\gamma$ has to be well within the electron energy distribution and $\Delta x$ is much smaller than an optical wavelength. Based on the present day electron beam technology, the energy resolution in an accelerator is about 0.01%. From (2.67), the classical approach is no longer valid when the device is operated in the soft x-ray region, i.e., $\lambda < 20\text{Å}$. For the present stage of experiments this limitation is still of no concern.

Now we consider the electromagnetic wave. The amplitude and phase have to be measured very accurately at the laboratory frame. However, according to quantum mechanics, such a measurement is impossible. The uncertainties in the photon number and the wave phase obey the uncertainty principle [8]

$$(\Delta N)^2 \; \frac{[(\Delta \cos \phi)^2 + (\Delta \sin \phi)^2]}{(<\cos \phi>^2 + <\sin \phi >^2)} \geqslant \frac{1}{4} \tag{2.68}$$

where the uncertainty in $\phi$ has been expressed in terms of the uncertainties in variables $\cos \phi$ and $\sin \phi$. The quantum limitation on the photon number can be obtained by assuming the uncertainty of $\phi$ to be less than $2\pi$. Let's say, $\Delta \cos \phi \simeq \Delta \sin \phi \simeq 1$. Then,

$$\Delta N \geqslant 1/2\sqrt{2} \tag{2.69}$$

$\Delta N$ is the number of photons per mode. The radiation power of that mode corresponding to the uncertainty in the photon number $\Delta N$ is

$$P = \Delta N \cdot \hbar\omega \cdot c/L$$

where $\omega$ is the radiation frequency, L is the length of the cavity. Since $\Delta N \geqslant 1/2\sqrt{2}$, we have

$$P \geqslant \frac{h\omega}{2\sqrt{2T}} \quad ; \; T = \frac{L}{c} \qquad (2.70)$$

For $\lambda = 10\mu$ and $L = 10\text{cm}$, we have

$$P \geqslant 2 \times 10^{-11} \text{ watts} \qquad (2.72)$$

This power is far below the value during the laser oscillation. Therefore, the use of a classical approach is justified. Since P is proportional to the radiation frequency, the quantum limitation of the radiation power might become substantial at very short wavelength.

# CHAPTER 2 - REFERENCES

1. J. D. Jackson, Classical Electrodynamics, 2nd ed. (Wiley, New York, 1975).
2. M. Abramowitz and I. A. Stegun, eds., Handbook of Mathematical Functions (Dover, New York, 1970).
3. C. Cronstrom and M. Noga, Phys. Lett. 60A, 137 (1977).
4. S. Gasiorowicz, Elementary Particle Physics (Wiley, New York, 1966).
5. N. W. McLachlan, Theory and Application of Mathieu Functions (Dover, New York, 1964).
6. P. A. Yeh, California Institute of Technology, Ph.D. Thesis, 1978.
7. J. AuYeung, Private communication.
8. P. Carruthers and M. M. Nieto, Rev. Mod. Phys. 40, 411 (1968).

# Chapter 3

# LINEAR THEORY OF THE LONGITUDINAL FREE ELECTRON LASER

In Section 2.5 the classical approach has been justified for the analysis of a free-electron laser device. Therefore the following analysis will be entirely from the classical point of view.

The linear theory formulation is the simplest method to understand the fundamental properties and laser mechanism of a free electron laser. The "linear" theory is based on three assumptions. First, the current density is so low that the Coulomb repulsion between electrons is completely negligible. The gain is thus proportional to the total current. Second, the field amplitude is sufficiently small so that it can be used as an expansion constant. An iterative method is then suitable for the analysis. The energy increase is proportional to the input intensity. Third, the energy transfer between the electron beam and the radiation is very small in the interaction region. The field amplitude and phase are thus assumed to be constant, which simplifies the analysis.

Based on these three assumptions, we formulate the linear theory of the longitudinal free electron laser. Starting from the force equation, we calculate the homogeneous and inhomogeneous gain constants and demonstrate the tunability of the device. We will also study the electron dynamics on the phase diagram. Finally, we will analyze an interesting device--a free electron laser which utilizes a two-stage system.

## 3.2 The Single Electron Analysis

In the simplest classical model, the stimulating process in a longitudinal free electron laser is described by the energy transfer between electrons and electromagnetic waves. The radiation field loses or gains energy, depending on whether work is done on or by the driven electron, which in turn is determined by the relative phase between the electron and the wave. In the limit of low current density, every electron interacts with the electromagnetic wave independently. The total energy transfer is then the average of the individual energy transfer over the electron phase distribution. This is the single-electron model and the calculated gain should be proportional to the current.

The equation of the electron motion in the presence of an electromagnetic field is given as

$$\frac{d\vec{P}}{dt} = e(\vec{E} + \frac{\vec{v}}{c} \times \vec{B}) \tag{3.1}$$

where $\vec{P}$ and $\vec{v}$ are the momentum and the velocity of the electron, $\vec{E}$ and $\vec{B}$ are the electric and magnetic fields. If the field corresponds to a wave there is a common factor of $e^{i(\omega t - \beta z + \phi)}$ in $\vec{E}$ and $\vec{B}$. $\omega$ and $\beta$ are the frequency and the wave number of the wave. The variables t and z correspond to the time and position of the electron which are defined to be zero when the electron is at the entrance of the interaction region. The phase of the electron upon entering the interaction region is given by $\phi$.

In general, (3.1) can be solved for the variation of the electron velocity in the transverse and longitudinal direction. The energy gain of an electron in transit can be obtained by calculating

$$\Delta\varepsilon = mc^2\Delta\gamma = mc^2[\gamma(L) - \gamma(0)] \tag{3.2}$$

or from the work equation

$$\Delta\varepsilon = \int_0^T e\vec{E} \cdot \vec{v}\, dt \tag{3.3}$$

To obtain a simpler analytical result, we solve equation (3.1) and evaluate (3.2), (3.3) in the region of low field intensity. Every physical quantity in (3.1), (3.2), and (3.3) can be expanded in power series of the field amplitude. Practically, we are interested only in $\langle\Delta\varepsilon\rangle_\phi$ up to second order, where $\langle\ \rangle\phi$ indicates an average over $\phi$. If (3.2) is used, the velocity has to be obtained up to second order. However, in (3.3), we have

$$\begin{aligned}\Delta\varepsilon &= \int_0^{T(\phi)} e\vec{E}\cdot\vec{v}\, dt \\ &= e\vec{E}^{(1)}\cdot\vec{v}^{(0)}\,\Delta T(\phi) + e\int_0^T \vec{E}^{(1)}\cdot\vec{v}^{(1)}\, dt \\ &\quad + e\int_0^T \vec{E}^{(2)}\cdot\vec{v}^{(0)}\, dt\end{aligned} \tag{3.4}$$

where

$$\vec{E} = \vec{E}^{(1)} + \vec{E}^{(2)} + \cdots$$

$$\vec{v} = \vec{v}^{(0)} + v^{(1)} + v^{(2)} + \cdots$$

$$T(\phi) = T + \Delta T(\phi) \quad ; \qquad T \equiv L/v^{(0)}$$

In most cases, the fractional change in the electron velocity which is about $\lambda/L$ is smaller than $10^{-4}$. This value is much smaller than any possible expansion constant in the power series. Therefore, the first and second terms on the right side of (3.4) are negligible compared to the third term. So,

$$\Delta\varepsilon = e \int_0^T \langle \vec{E}^{(2)} \cdot \vec{v}^{(0)} \rangle_\phi \, dt \tag{3.5}$$

In order to obtain $\vec{E}^{(2)}$, only the velocity up to first order is necessary.

In terms of velocity components, equation (3.1) is rewritten as

$$m\gamma \frac{dv_x}{dt} = - \frac{m\gamma^3}{c^2} (v \cdot \frac{dv}{dt}) \, v_x + e[E_x + \frac{v_z}{c} B_y] \tag{3.6}$$

$$m\gamma \frac{dv_y}{dt} = - \frac{m\gamma^3}{c^2} (v \cdot \frac{dv}{dt}) \, v_y + e[E_y - \frac{v_z}{c} B_x] \tag{3.7}$$

$$m\gamma \frac{dv_z}{dt} = - \frac{m\gamma^3}{c^2} (v \cdot \frac{dv}{dt}) \, v_z + e[E_z - \frac{v_x}{c} B_y + \frac{v_y}{c} B_x] \tag{3.8}$$

where $B_z = 0$ in the TM mode. In general, $E_{x,y,z}$ and $B_{x,y}$ are functions of x and y. From (2.1), we know their values are comparable excepting factors of sine and cosine. It is also assumed that the electron has only a z-component of initial velocity.

Let us examine the transverse force equations (3.6) and (3.7). It is obvious that $v_x$ and $v_y$ are to lowest order $v_x^{(1)}$ and $v_y^{(1)}$. According to equation (3.5), the contribution of $E_x v_x$ and $E_y v_y$ is too small to be considered. But $v_x^{(1)}$ and $v_y^{(1)}$ could enter the right side of (3.8), but they generate terms proportional to the square of the field amplitude which can be disregarded because $v_z$ is only up to first order. Therefore, the transverse force equations and the magnetic force in (3.8) can be neglected. The remaining equation is very simple

$$\frac{dv}{dt} = \frac{eE}{m\gamma^3} \cos[\omega t - \beta z + \phi] \tag{3.9}$$

where we have deleted the subindex z without leading to ambiguity, and included the propagation factor in the cosine factor.

Assuming that $\Delta\gamma << \gamma$, we solve equation (3.9) for v and z in an iterative way. The zeroth order solution for the electron position is $z(t) = v_o t$ where $v_o$ is the electron initial velocity. From (3.9) we obtain the velocity to first order as

$$v = v_o + \frac{eE}{m\gamma^3\Omega} [\sin(\Omega t - \phi) + \sin\phi] \tag{3.10}$$

and

$$z = v_o t + \frac{eE}{m\gamma^3\Omega^2} [\cos\phi - \cos(\Omega t - \phi) + \Omega t \sin\phi] \tag{3.11}$$

$$\begin{aligned} \Omega &\equiv \beta v_o - \omega \\ &= \omega[\frac{v_o}{v_p} - 1] \quad , \qquad v_p = \omega/\beta \end{aligned} \tag{3.12}$$

$\Omega$ is thus the wave frequency as "seen" by the electron. Exact synchronism, i.e., wave phase velocity $v_p$ equals electron velocity $v_o$, obtains when $\Omega = 0$.

The electric field "seen" by an electron is no longer a perfect sinusoidal wave due to the variation of the electron position from $v_o t$. Substituting $z = v_o t + \Delta z$ into the field expression,

$$\begin{aligned} E(t,\phi) &= E\cos(\omega t - \beta z + \phi) \\ &= E\cos[\omega t - \beta(v_0 t + \Delta z) + \phi] \\ &= E\cos(\Omega t - \phi) \\ &\quad - \frac{\beta e E^2}{m\gamma^3\Omega^2}\sin(\Omega t - \phi)[\cos\phi - \cos(\Omega t - \phi) + \Omega t\sin\phi] \end{aligned}$$

The integrand in (3.5) can be calculated to be

$$\begin{aligned} <\vec{E}^{(2)}\cdot\vec{v}^{(0)}>_\phi &= v_0<E^{(2)}>_\phi \\ &= -\frac{\beta v_0 e E^2}{2m\gamma^3\Omega^2}\{\sin\Omega t - \Omega t\cos\Omega t\} \end{aligned} \qquad (3.14)$$

The phase-averaged energy loss per electron is obtained by integrating (3.14) over t

$$<\Delta\varepsilon>_\phi = \frac{\beta v_0 e^2 E^2}{2m\gamma^3\Omega^3}\{2 - 2\cos\Omega T - \Omega T\sin\Omega T\} \qquad (3.15)$$

$$= \frac{\omega e^2 E^2 T^3}{2m\gamma^3}\{\frac{1}{\Omega T}[\frac{\sin^2(\Omega T/2)}{(\Omega T/2)^2} - \frac{\sin\Omega T}{\Omega T}]\} \qquad (3.16)$$

The function within the curled brackets contains the dependence of the energy transfer on the electron and wave velocities. The dependence

$$f(\Omega T) = \frac{1}{\Omega T}[\frac{\sin^2(\Omega T/2)}{(\Omega T/2)^2} - \frac{\sin\Omega T}{\Omega T}] \qquad (3.17)$$

is plotted in Figure 3.1. It is a fundamental synchronism function for a single electron-wave interaction. The functional dependence is

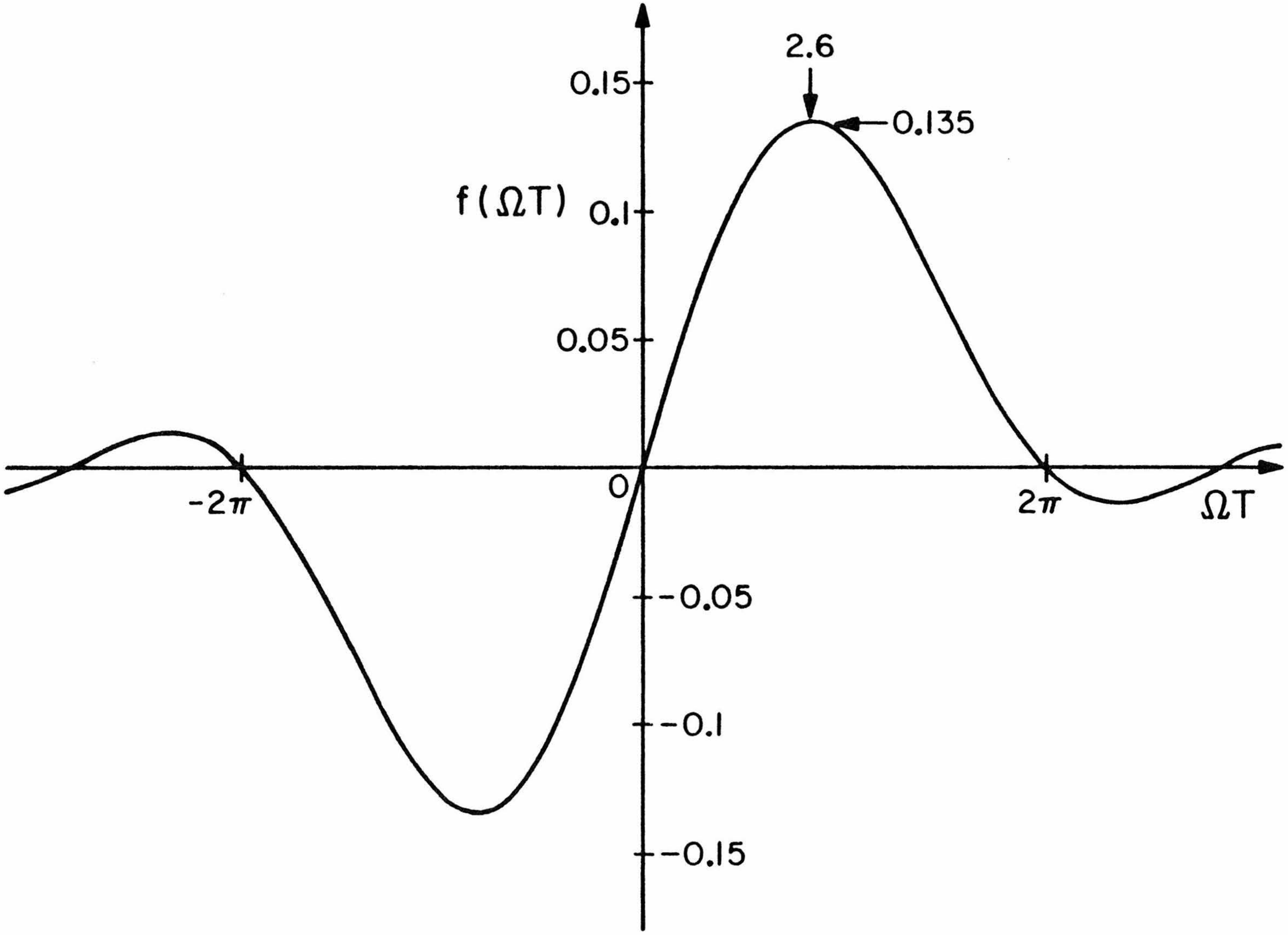

Figure 3-1: The stimulated gain function f(ΩT).

identical to the derivative of the spontaneous radiation spectrum

$$f(x) = -\frac{1}{2}\frac{d}{dx}\left[\frac{\sin^2(x/2)}{(x/2)^2}\right] \tag{3.18}$$

which is in agreement with the quantum mechanical analysis [1] of the free electron laser.

Note that $f(\Omega T)_{max} = 0.135$ at $\Omega T = 2.6$. Physically, $\Omega > 0$ means the electron velocity is larger than the phase velocity of the wave. But it is well known that the phase velocity of a waveguide mode is always larger than c (see (2.2)). This why it is necessary to introduce corrugations on the waveguide. The period of the corrugation is chosen so that the phase velocity of the first harmonic is less than c. In this case it becomes possible for the electron to interact with the wave in positive gain region. It is also noticed that there are other positive gain regions below $\Omega T = -2\pi$. The phase velocity in these regions is larger than the electron velocity. It is not necessary to have a periodic structure to generate slow waves. However, the operation in these regions requires extremely high electron energies and results in only very low gain. This method is rather impractical compared to the operation around $\Delta T \simeq 2.6$.

Once we obtain the average energy loss per electron, the power loss by the electron beam is found as

$$\Delta P = -\langle \Delta\varepsilon \rangle_{\phi} \frac{I}{|e|}$$

$$= \frac{\omega |e| E^2 T^3 I}{2m\gamma^3} f(\Omega T) \tag{3.19}$$

The field amplitude appearing in (3.19) is the amplitude of the first harmonic, which is related to the total field by the wave spectrum discussed in Section 2.2. For a longitudinal free electron laser, N is about 500. The spectrum at q = 1 is very narrow with an amplitude

$$E_1 = E_o J_1(s) \tag{3.20}$$

$$S = \frac{\Delta}{k'}\frac{\partial k}{\partial a}$$

$$= \frac{1}{2}\ \frac{\Lambda\Delta\lambda}{a^3\sqrt{1 - \lambda^2/2a^2}} \tag{3.21}$$

where $E_o$ is the field amplitude of the fundamental TM mode. The expressions of S and k can be found in Section 2.2. The field $E_o$ is related to the total electromagnetic power P by

$$E_o^2 = 2\beta_o^2\ P\ K_o \tag{3.22}$$

where $K_o$ is the waveguide impedance for the fundamental mode

$$k_o = \sqrt{\frac{\mu}{\varepsilon_o}}\ \frac{\lambda^4}{2\pi^2 a^4 (1 - \lambda^2/2a^2)^{3/2}} \tag{3.23}$$

Using (3.20), (3.22), and (3.23) in (3.19), we obtain the following expression for the gain per pass

$$G = \frac{\Delta P}{P} = \frac{\omega|e|J_1^2(S)\ T^3 I\ \beta_o^2 k_o}{m\gamma^3}\ f(\Omega T) \tag{3.24}$$

By choosing the beam velocity $v_o$ so that T = 2.6, we can write the maximum gain as

$$G_{max} = \frac{0.135|e|(\mu/\varepsilon_o)^{1/2}}{\pi^2 mc^2} \times \frac{(\omega T)^3 J_1^2(S)\lambda^4\left(1 - \frac{\lambda^2}{2a^2}\right)^{3/2}}{2\gamma^3 a^4}\ I \tag{3.25}$$

For the practical case, $a >> \lambda$ and $s << 1$, we have

$$G_{max} = 0.78 \times 10^{-4} \frac{\Lambda^2 \Delta^2 \lambda^3 L^3 I}{\gamma^3 a^{10}} \tag{3.26}$$

Equation (3.25) is thus the basic result for the gain in the case of a perfectly monoenergetic electron beam. It applies in practice to a beam in which the velocity spread satisfies $v_o/v_o < (\omega T)^{-1}$ (or, equivalently, $\Delta\Omega < T^{-1}$). In terms of energy resolution, this condition reduces to

$$\frac{\Delta\gamma}{\gamma} = \gamma^2 \frac{\Delta v}{v}$$

$$< \frac{\gamma^2 \lambda}{2\pi L} \sim 2 \times 10^{-4} \tag{3.27}$$

This case is referred to as the homogeneous situation, in analogy with ordinary lasers, and requires an electron beam of high energy resolution.

The synchronism condition determines the relation between the electron energy, radiation frequency and the corrugation period

$$\beta_1 = \beta_o + \frac{2\pi}{\Lambda} = \frac{\omega}{v_o} \tag{3.28}$$

Using the dimensionless electron energy, $\gamma$, we have

$$\gamma = \left\{1 - \left[(1 - \lambda^2/2a^2)^{1/2} + \frac{\lambda}{\Lambda}\right]^{-2}\right\}^{-1/2} \tag{3.29}$$

which gives us the tuning curve for the longitudinal free electron laser. With a fixed period $\Lambda$, the wavelength of the output radiation can be continuously tuned by only changing the electron energy. The relations between $\gamma$ and $\lambda$ are plotted in Figure 3.2 for different periods. For each period, the curve has a minimum which means that there is a lower

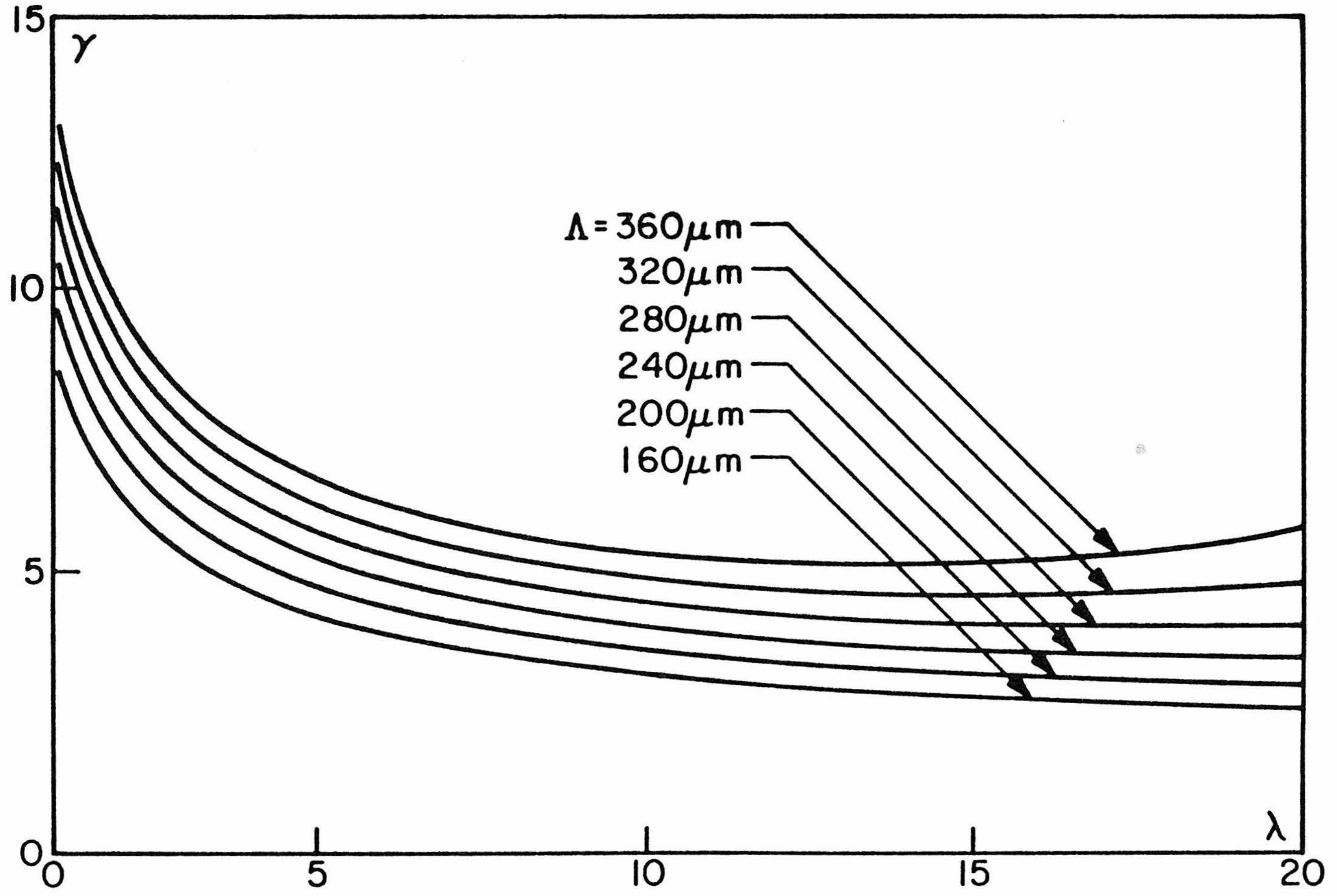

Figure 3-2: The dependence of the wavelength ($\lambda$) on the electron energy ($\gamma$) for different corrugation period $\Lambda$.

limit to the phase velocity for the first harmonic. In practice, $\lambda \ll a$ and $\lambda \ll \Lambda$, in which case (3.29) becomes

$$\gamma^2 = \Lambda/2\lambda \tag{3.30}$$

which is identical to the result of the transverse free electron laser [1]. As expected, in the limit of low current intensity, $G_{max}$ is proportional to I. Other interesting features of $G_{max}$ are:

(i) $G_{max} \propto L^3$: This implies that the gain due to a monoenergetic electron beam is not exponential. The gain can be made higher by using a longer interaction region if $\Omega T$ is kept constant so that in practice as we increase $T(=L/v_0)$ we need to operate closer to synchronism.

(ii) $G_{max} \propto \lambda^{9/2}$: The gain drops dramatically at shorter wavelengths. That is the reason why it is difficult to operate the free electron laser at very high frequencies.

(iii) $G_{max} \propto a^{-10}$: This steep tenth power dependence on "a" reflects the importance of the waveguide dimension. A factor of $a^{-6}$ is due to operation in the TM mode. Smaller waveguides have a larger longitudinal component of the electric field and increases the stimulating strength. A decrease of one-fifth in the dimension leads to a gain of ten times. This advantage could be used to compensate for the troubling waveguide loss which is proportional to a.

To get an appreciation of the level of gain predicted by (3.25), we consider the following example

$$\lambda = 10\ \mu\text{m}\ ; \qquad a = 50\ \mu\text{m}\ ;$$

$$\Delta = 10\ \mu\text{m}\ ; \qquad \Lambda = 200\ \mu\text{m}\ ;$$

$$L = 10\ \text{cm}\ ; \qquad I = 1\ \text{mA}\ .$$

From these independent values, we have $\gamma = 3.64$,$s=0.163$, and $J_1^2(s) = 0.0065$. The calculated gain is

$$G_{max} = 7\% \text{ per pass} \tag{3.31}$$

This gain is sufficient for laser oscillation when mirror feedback is present, provided the losses of the waveguide are negligible. We have to emphasize that the value in (3.31) depends critically on physical parameters, especially $a$ . For example, if $a = 40\ \mu\text{m}$, the gain will rise to a value of 63%.

The gain expression (3.24) applies to the case of a perfectly monoenergetic electron beam and is called the homogeneous gain. However, when the electron velocity distribution is sufficiently broad, electrons with different velocities provide different value of gain. The total gain should be the integral of the gain weighted by the electron distribution. In Figure 3.3 the cases of narrow and broad distributions are shown in terms of velocity. If the distribution is smooth and its width exceeds $\Delta T$ by a large factor, it can be expanded around the wave velocity

$$g(v_e) = g(\Omega)\omega/v_p \tag{3.32}$$

and

$$g(\Omega) = g_\Omega(0) + (\partial g_\Omega(0)/\partial\Omega)\Omega \tag{3.33}$$

The total gain is found by multiplying (3.33) and (3.24), and then integrating over $\Omega$.

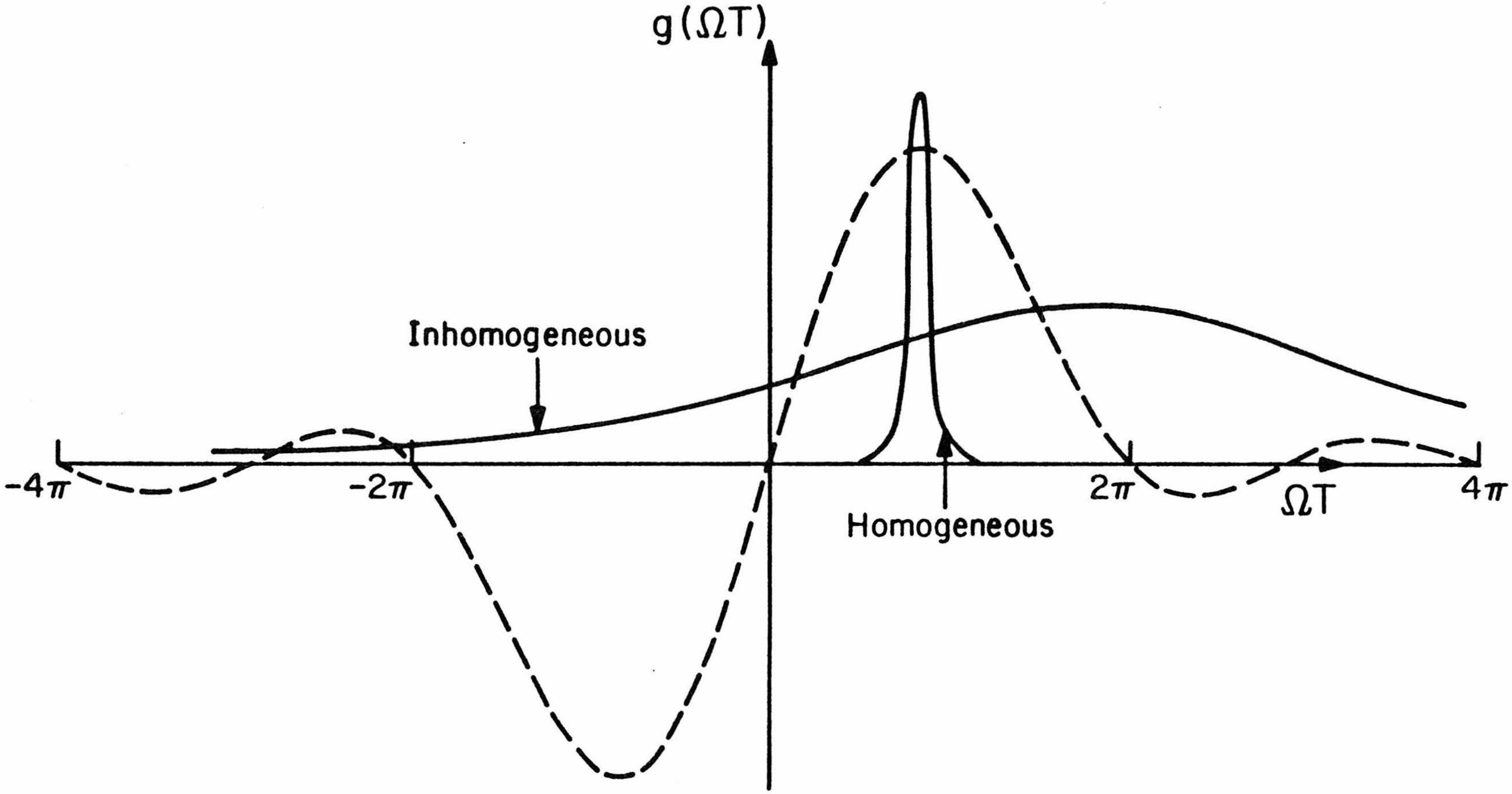

Figure 3-3: The comparison of the width of the electron distribution (solid line) and the gain distribution (dotted line).

$$G = \frac{\partial g_\Omega(0)}{\partial\Omega} \int \Omega G(\Omega)\, d\Omega$$

$$= \pi G'T \frac{\partial g_\Omega(0)}{\partial\Omega} \qquad (3.34)$$

where

$$G' = \omega|e|Ik_0\ J_1^2(s)\ \beta_0^2 / m\gamma^3 \qquad (3.35)$$

The result in (3.34) is called the "inhomogeneous gain." Since it is proportional to T, the gain is exponential. A gain which is proportional to the derivative of the electron distribution at the resonance is well known in the traveling wave tube [2]. It seems that we can get a very large gain if the system is operated at the sharp edge of the electron distribution, but this is in contradiction with the assumption made in expanding $g(\Omega)$. Therefore, the gain in (3.35) is usually a small value. Physically, a broad beam leads to a reduced gain because the net gain is due to the small excess of electrons which lose energy to the wave over those which gain energy. If possible, it is best to avoid the inhomogeneous interaction experimentally.

The gain we calculate in the homogeneous case is the total gain at the output, but not the infinitesimal gain along the interaction region. At the beginning of the device, the gain increases as $t^4$ and reaches its maximum at the output when $\Omega T = 2.6$. For $\Omega T > 2.6$, the gain at the output has already passed its maximum, so that it can be increased by using a shorter interaction region. For $\Omega T < 2.6$, the gain is still increasing at the output. The behavior of this gain is quite different from the loss mechanism of the waveguide which is an exponential function of the distance. At first, the loss is proportional to t which is larger than the gain for small t. It is thus possible to have a situation in which the overall gain exceeds the loss but not be larger than the loss on an incremental basis everywhere. This situation merits an investigation, since it is not even clear a priori if such a device can oscillate.

The analysis proceeds just as in the calculation of the homogeneous gain, except that the force equation (3.9) is replaced by

$$\frac{dv}{dt} = \frac{\dot{e}E}{m\gamma^3} e^{-\alpha z} \cos[\omega t - \beta z + \phi] \tag{3.36}$$

where $\alpha$ is the loss constant of the waveguide. The derivation is tedious but straightforward. The only integral we need to calculate is

$$\int e^{-at} \cos(bt+c)\, dt = \frac{e^{-at}}{a^2+b^2} [b \sin(bt+c) - a \cos(bt+c)] \tag{3.37}$$

Other integrals needed in the derivation can be obtained by taking the derivatives of (3.37) with respect to a, b, or c.

The result is identical to (3.24) except the sychronism function is replaced by f(p,q), where

$$f(p,q) = \frac{qe^{-p}}{(p^2+q^2)^2} \{[2 \cosh p - 2 \cos q - (\frac{p}{q} + \frac{q}{p})p \sin q] + \frac{\alpha}{\beta} [(\frac{p}{q} - \frac{q}{p}) \sinh p + 2 \sin q - (\frac{p}{q} + \frac{q}{p})p \cos q]\} \tag{3.38}$$

$$p \equiv \alpha L \quad , \quad q \equiv \Omega T$$

In general, $\beta >> \alpha$, and the second term in (3.38) can be neglected. When $\alpha \to 0$, (3.38) reduces to the form in (3.17). The total gain is then equal to

$$G_T = G + (e^{-\alpha L} - 1)$$

$$\simeq G - \alpha L \qquad (3.39)$$

where G is given as in (3.14) with f(p,q). The result, (3.39), is important. It shows that for small traveling wave gain and waveguide loss we can calculate each one independent of the other and then subtract the result to get the net gain (or loss).

## 3.3 Electron Dynamics

In a given interaction region, the electron changes its velocity and position from $v_0$ and $v_0t$, respectively, along the path due to the interaction with the electromagnetic wave. The change of position is on the order of the radiation wavelength. The variation in the longitudinal direction essentially changes the relative phase of the electron with respect to the wave. However, the variation in the transverse direction results in the divergence of the electron beam and can be neglected.

The modulation of electrons in real and velocity spaces is best described by plotting its distribution intensity on the v-$\phi$ phase plane. The evolution of the electron distribution is then clearly visualized by the changing of its shape and density. In order to determine the evolution of the distribution, it is necessary to understand how an electron propagates in the phase plane. The path of an electron is determined uniquely by its velocity and phase at time t. Taking t as a parameter, we can trace the electron motion and form a stream line for that electron. The stream lines will not interact with each other unless they coincide exactly all the time. The stream line of a single electron can be found from the force equation (3.9)

$$\frac{dv}{dt} = \frac{eE}{m\gamma^3} \cos \phi \tag{3.9}$$

where $\phi = \beta z - \omega t - \phi_o$ is the phase of the electron with respect to the electromagnetic wave at z and t. $\phi_o$ is the electron phase upon entry. Defining a new variable,

$$w = v - v_p = \frac{1}{\beta} \frac{d\phi}{dt} \tag{3.40a}$$

which is the relative velocity between the electron and the wave, equation (3.9) is rewritten as

$$\frac{dw}{dt} = \frac{eE}{m\gamma^3} \cos \phi \tag{3.40b}$$

Multiplying both sides of (3.40b) by 2w and integrating,

$$d(w^2) = \frac{2eE}{\beta m\gamma^3} d(\sin \phi) \tag{3.41}$$

Assuming the electron stream line passes through a point $(w_o, \phi_o)$, (3.41) is solved to obtain the equation for the stream line

$$w^2 - w_o^2 = \frac{2eE}{\beta m\gamma^3} [\sin \phi - \sin \phi_o] \tag{3.42}$$

A set of stream lines based on (3.42) is plotted in Figure 3.4. This is a reduced phase plane which shows only the stream lines within one optical wavelength. The lines ending at $\phi = 2\pi$ should reappear at $\phi = 0$. The complete phase plane is obtained by placing the reduced phase planes side by side. The stream lines are then continuous in the complete

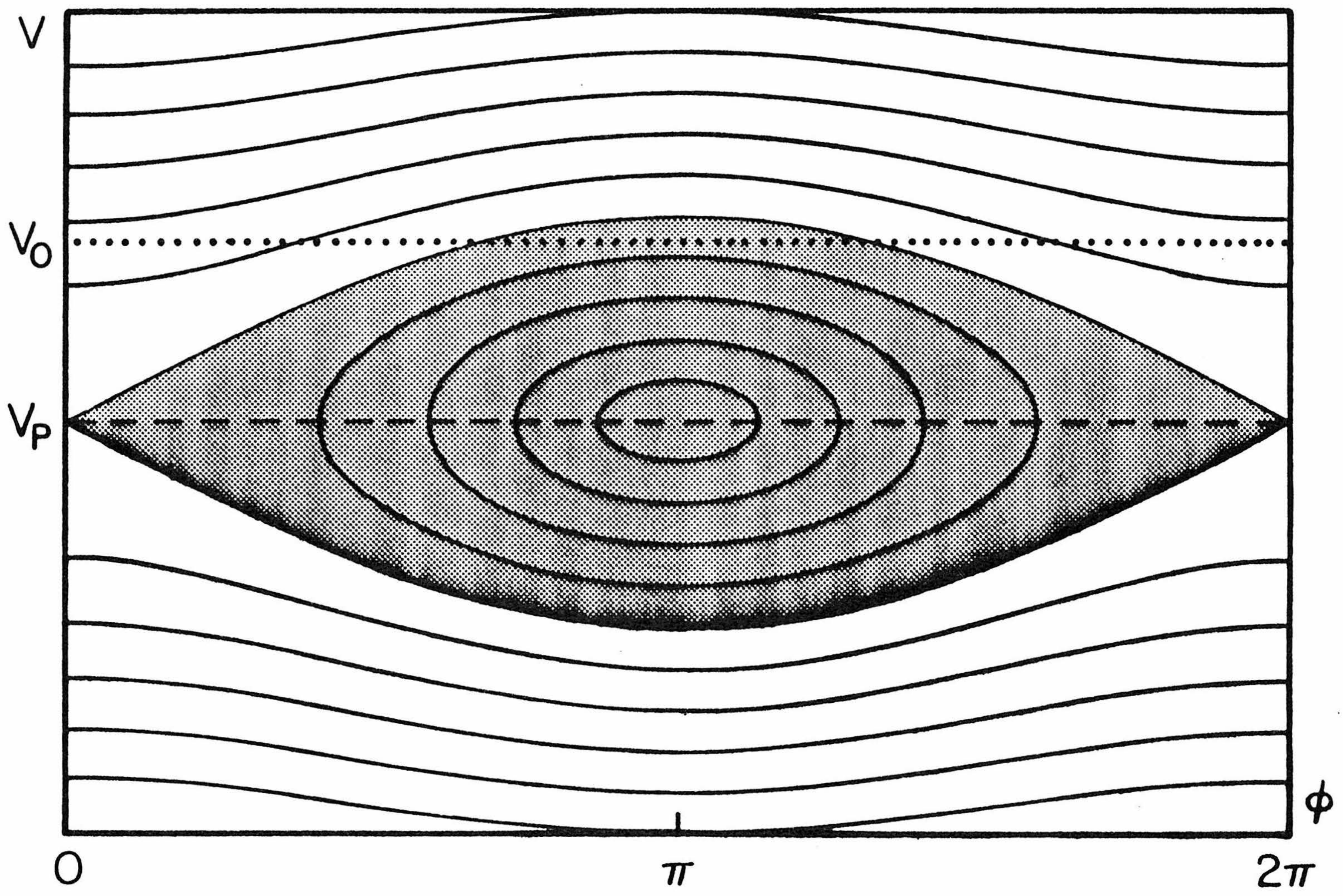

Figure 3-4: The phase diagram of electrons. The shaded area is the region of stability.

phase plane. The phase plane can be divided into two regions of stability and instability. The stable region is shown as the shaded area in Figure 3.4. All the stream lines in this region are closed. Electrons in this area will be trapped within the same wavelength of radiation and keep circling around the center. The region of instability includes all the open stream lines. Electrons in this area will keep overtaking the wave of radiation. It should be noticed that the stream-line picture breaks down when the field amplitude is not a constant. Electrons will jump continuously from one line to the other due to the variation of E.

If $w_o = v_o - v_p$, the corresponding $\phi_o$ is the entry phase. Instead of an electron we consider a monoenergetic electron beam distributed evenly in $\phi$. It is easy to see from (3.42) that if

$$w_o^2 > \frac{4eE}{\beta m\gamma^3} \tag{3.43}$$

w always has a solution for arbitrary given $\phi$ and $\phi_o$. If the electron beam is plotted on the phase plane, the entire beam is in the unstable region. Thus the electron will not be trapped locally inside the wave if the electron velocity is high enough. It is interesting to compare the condition (3.43) with the gain curve. The electron velocity obeying (3.43) with an equality sign will appear on the gain curve at

$$\Omega T = \sqrt{\frac{4}{m}\,\frac{eEL}{c^2\gamma^3}\,\beta L} \tag{3.44}$$

For the previously given example of a longitudinal free electron laser, the point of maximum gain operation will appear in the unstable region if $E \lesssim 6 \times 10^3$ V/m. The electron beam following (3.43) is plotted in Figure 3.5a. The shaded region is the allowed region for electrons during propagation.

If $w_o^2 < \frac{4eE}{\beta m\gamma^3}$, part of the electron beam enters the region of stability and will be trapped during the interaction. In Figure 3.5b, we show the position of the electrons and the allowed region. The shaded area in Figure 3.5a and b also reveals the possible width of the electron distribution at the output. For a monoenergetic electron beam, the maximum range of the distribution in velocity is

$$\sqrt{w_o^2 + \frac{4eE}{\beta m\gamma^3}} > w > \sqrt{w_o^2 - \frac{4eE}{\beta m\gamma^3}} \qquad \text{for} \qquad w_o^2 > \frac{4eE}{\beta m\gamma^3}$$

and (3.45)

$$\sqrt{w_o^2 + \frac{4eE}{\beta m\gamma^3}} > w > -\sqrt{\frac{4eE}{\beta m\gamma^3}} \qquad \text{for} \qquad w_o^2 < \frac{4eE}{\beta m\gamma^3}$$

If the electron distribution has a finite width, it is represented by a strip instead of a line, and each electron still follows the stream line. The evolution of the distribution can be seen qualitatively on the phase plane. Because time t is an implicit parameter in the stream line, the determination of the electron distribution at any instant is not straightforward.

The electron distribution on the phase plane is $N_t(v,\phi)$. The conservation of electron number requires that $N_t(v,\phi)dv\,d\phi$ should be invariant. Given the initial distribution $N_o(v_o,\phi_o)$, the distribution at time t can be found from

$$N_t(v,\phi)dv\,d\phi = N_o(v_o,\phi_o)dv_o d\phi \qquad (3.46)$$

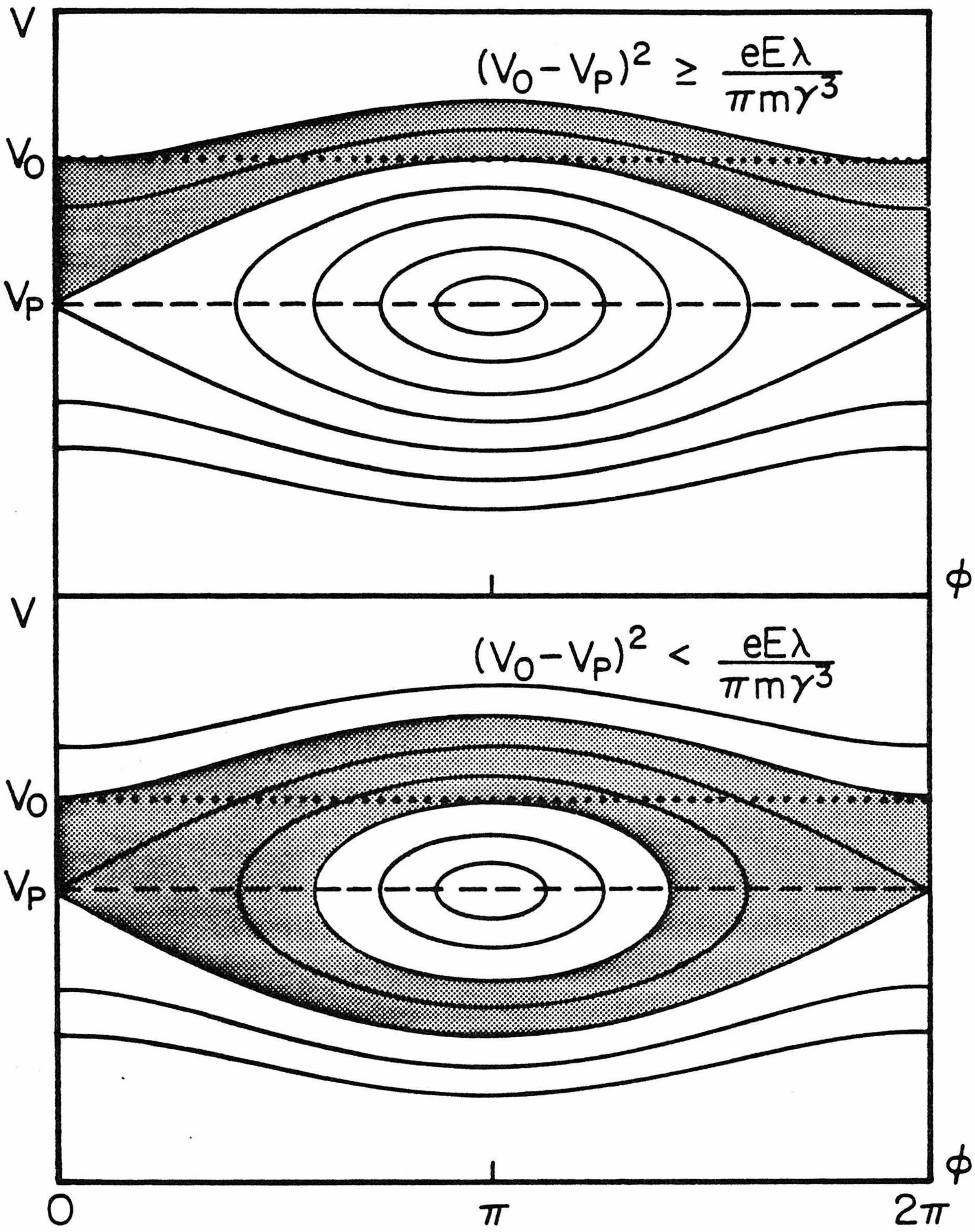
V
$(V_O - V_P)^2 \geq \frac{eE\lambda}{\pi m \gamma^3}$
$V_O$
$V_P$
$\phi$
V
$(V_O - V_P)^2 < \frac{eE\lambda}{\pi m \gamma^3}$
$V_O$
$V_P$
$\phi$
0
$\pi$
$2\pi$

Figure 3-5: Electron dynamics in un-trapped (A) and partly trapped (B) cases. The shaded area is the allowed region for electrons.

The infinitesimal areas dv d$\phi$ and dv$_o$d$\phi_o$ are related by

$$dv_o d\phi_o = \begin{vmatrix} \frac{\partial v_o}{\partial v} & \frac{\partial \phi_o}{\partial v} \\ \frac{\partial v_o}{\partial \phi} & \frac{\partial \phi_o}{\partial \phi} \end{vmatrix} \quad dv\ d\phi \tag{3.47}$$

The distribution at time t is

$$Nt(v,\phi) = N_o(v_o,\phi_o)[\frac{\partial v_o}{\partial v}\frac{\partial \phi_o}{\partial \phi} - \frac{\partial \phi_o}{\partial v}\frac{\partial v_o}{\partial \phi}] \tag{3.48}$$

Generally, v and $\phi$ are functions of $v_o$ and $\phi_o$. In order to evaluate (3.48) it is necessary to invert the functions and obtain $v_o$,$\phi_o$ as functions of v and $\phi$. From the force equation, v and z can be solved iteratively, and $\phi$ is equal to

$$\phi = \phi_o - [\Delta z(v_o,\phi_o) + \Omega t] \tag{3.49}$$

Using (3.48), it is very easy to calculate the shift and spread of the electron distribution in velocity. The shift is calculated as

$$\begin{aligned} \Delta v \equiv \langle v\rangle - v_o &\equiv [\int v N_t(v,\phi) dv\, d\phi] - v_o \\ &= [\int v\ N_o(v_o,\phi_o) dv_o d\phi_o] - v_o \end{aligned} \tag{3.50}$$

If $v(v_o,\phi_o)$ is explicitly known, then $\Delta v$ is obtained by (3.50). The spread is defined as

$$\sigma_v = \sqrt{\langle v^2\rangle - \langle v\rangle^2} \tag{3.51}$$

$\langle v\rangle$ is given in (3.50). $\langle v^2\rangle$ is also obtained in the same way.

A special case is one where the electrons are distributed uniformly and monoenergetically. Then, due to the conservation of probability, we have

$$N(v)\ dv = \frac{N_0}{2\pi}\ d\phi_0 \tag{3.52}$$

where we have integrated over $\phi$ to obtain a distribution which depends on v only. $N_0$ is the total electron number in unit length, and

$$N(v) = \frac{N_0}{2\pi} \frac{1}{\left|\frac{dv}{d\phi_0}\right|} \tag{3.53}$$

In the small field approximation, $dv/d\phi_0$ is evaluated by solving (3.10) up to second order in E. Considering the shift and spread in the lowest order in E we have, in general,

$$v = v_0 + p(v_0) + f(v_0) \cos \phi_0 + g(v_0) \sin \phi_0 \tag{3.54}$$

whereby (3.53) gives

$$N(v) = \frac{N_0}{2\pi} \frac{1}{\sqrt{(f^2 + g^2) - (v - v_0 - p)^2}} \tag{3.55}$$

For given $v_0$, (3.55) is plotted in Figure 3.6a. The shape is symmetric about $v_0 + p$ and the width of the distribution as calculated from (3.50) is found to be equal to $\sqrt{(f^2 + g^2)/2}$. The shift is on the order of $E^2$, while the spread is on the order of E and is much larger than the shift. When the electrons have an initial spread $N_0(v)$ in v, (3.55) can be easily generalized to

$$N(v) = \frac{1}{2\pi} \int \frac{N_0(v_0)}{\sqrt{(f^2 + g^2) - (v - v_0 - p)^2}}\ dv_0 \tag{3.56}$$

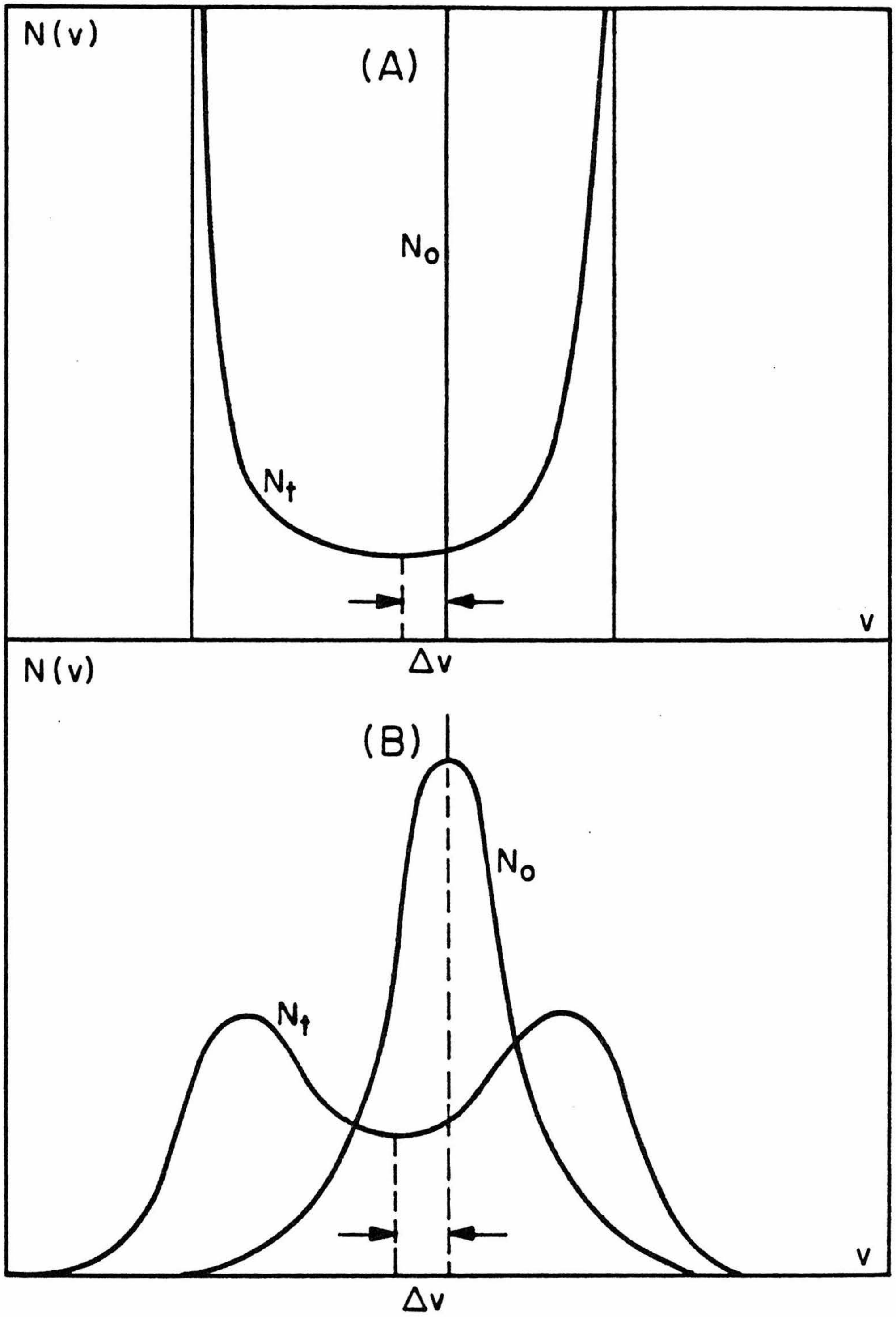

Figure 3-6: Electron distribution N , N before and after the interaction for a monoenergetic (A) or a Gaussian (B) input beam.

An example where $N_0(v_0)$ is a Gaussian distribution is shown in Figure 3.6b. The splitting structure is obvious.

In the small field limit, gain saturation is not well defined, since the neglect of space charge forces allows the predicted gain to increase indefinitely. However, we can get an estimate to the upper limit of energy transfer by using the phase diagram. For example, consider the electron beam in the instability region. The maximum energy transfer for an electron is

$$\Delta\gamma = \gamma^2 \frac{\Delta v}{c}$$

$$= \frac{\gamma^2}{c} [\sqrt{\omega_0^2 + \frac{4eE}{\beta m\gamma^3}} - \sqrt{\omega_0^2 - \frac{4eE}{\beta m\gamma^3}}] \tag{3.57}$$

If the operation is very near the condition (3.44),

$$\Delta\gamma \simeq \frac{\gamma^2}{c} \sqrt{2}\, \omega_0$$

$$\simeq \gamma^2 \sqrt{2} \frac{\Omega}{\omega} \simeq 10^{-3} \tag{3.58}$$

It follows that the upper limit for the efficiency of a longitudinal free electron laser is about 0.1 percent.

## 3.4 Two-Stage System

As practical free electron lasers were considered, it was realized that the electron beam has to be recycled in order to achieve higher overall efficiency. Using a magnetic field, the output electrons can be brought back and made to re-enter the interaction region. The energy loss of an electron in the interaction region can be compensated by supplying to the electron on its return, energy equal to that lost in the interaction region. However, the energy compensation mechanism is of necessity a homogeneous effect. Every electron in the beam is subjected to the same amount of energy change. The whole electron distribution is shifted without changing the shape. Therefore, the recycled beam becomes increasingly broader as it cycles. Although the average beam energy can be brought back to the point of maximum gain, the gain actually decreases due to the increasing spread of the velocity distribution. This broadening mechanism highly limits the efficiency of the free electron laser, even when the electron recirculation is considered.

In order to show how the electron beam shift and spread affect the device efficiency, we study their expressions in a single stage quantitatively and, later, compare them with the results obtained from the new proposed scheme. In the small-signal limit the electron velocity for a single interaction region can be found from the integration of equation (3.9) as

$$\begin{aligned} v &= v_0 + \frac{A}{\Omega}\{(1-\cos\Omega t)\sin\phi + \sin\Omega t\cos\phi\} \\ &\quad - \frac{A^2\beta}{\Omega^3}\{1-\cos\Omega t - \frac{\Omega t}{2}\sin\Omega t\} + \cdots \qquad (3.59) \\ A &= eE/m\gamma^3 \end{aligned}$$

The $\phi$ dependent terms have been kept only up to first order in A, while the $\phi$ independent terms have been kept to second order. In the case of an initially uniform and monoenergetic electron beam, the average velocity shift $\Delta$ and the r.m.s. velocity spread $\sigma$ of electrons are calculated by taking the proper average of v in equation (3.59) over $\phi$.

$$\Delta = - \frac{A^2}{\Omega^3} \{1 - \cos \Omega T - \frac{\Omega T}{2} \sin \Omega T\}$$

$$\sigma = \frac{\sqrt{2}\, A}{\Omega} \sin \frac{\Omega T}{2} \qquad (3.60)$$

where T = L/c is the total flight time of the electron through the interaction region of length L. The maximum gain takes place when T = 2.6, where the shift and the spread are

$$\Delta = -0.0675\beta\ A^2 T^3$$

$$\sigma = 0.525\ AT \qquad (3.61)$$

It is obvious that the velocity shift is a second order effect ($\propto A^2$), while the spread is a first order effect ($\propto A$). In the small signal region the spread actually dominates over the shift.

If we recirculate an electron beam which initially is monoenergetic, and initial distribution in $\phi$ space is uniform, then each time the electron beam re-enters the interaction region, we find that the velocity shift is proportional to the number of circulations N, while the velocity spreads to $\sqrt{N}$

$$\Delta_N = \Delta N$$

$$\sigma_N = \sigma\sqrt{N} \qquad (3.62)$$

The maximum number of circulations is thus determined by the maximum allowable velocity spread. From the theoretical expression for velocity dependence of the gain profile, it can be seen that the gain is reduced by 10% within a range of $\Delta\Omega T = 1$ around the peak gain point. We assume that the electron beam is no longer useful when its distribution is wider than this range and define a maximum allowable spread of electrons $\sigma_{max}$. The maximum number of cycles is found to be

$$N_{max} = (\sigma^2_{max}/\sigma^2) \tag{3.63}$$

During $N_{max}$ cycles of electron circulation, the total velocity shift becomes

$$\Delta_{max} = (\Delta/\sigma^2)\sigma^2_{max} \tag{3.64}$$

Since the extractable energy from the electron beam is proportional to $\Delta_{max}$, the overall efficiency of the device depends on the value of the factor $R \equiv (\Delta/\sigma^2)$. Consequently, an increase in the efficiency can be achieved by either enhancing the single-pass shift $\Delta$ or reducing the single-pass spread $\sigma$.

The velocity spread is due to the different entry phases of electrons which thus see different electric fields during their transit. If we can "invert" the interaction between wave and electrons in a second interaction region, then we may expect a reduction in the velocity spread. Such inversion can be achieved if the entry phase of each electron is shifted by $\pi$ radians with respect to that of the first region. When this happens, each electron in the second region experiences an EM force which is almost the same in magnitude, but opposite in direction to the force it has experienced in the first region. Electrons which were accelerated in the first region will be slowed down and slow electrons will be accelerated in the second region. Thus, two-stage devices have been proposed [3-5] which consist of two identical interaction regions, separated by a drift distance between them. Because of the velocity difference between wave and electrons, the $\pi$ shift of the entry phase can be obtained by adjusting the value of $L_D$ (Fig. 3.7).

As a separate, but related issue, we consider the problem of maximizing the single pass velocity shift and thus the single-pass gain by using two interaction regions. The basic reasoning for this approach is derived from the operation of the klystron.[6] The first interaction region acts as a buncher giving rise to a strong velocity modulation. In the drift space between the two interaction regions the velocity modulation gives rise to bunches (current modulation). These bunches are then made to enter the second interaction region at the optimum phase for deceleration and energy extraction.

The proposed device consists of two identical sections which are separated by a drift space of length $L_D$. The propagation distance of electrons $L_e$ and radiation $L_r$ may not be equal to $L_D$. For example, a bending magnetic field and an accelerating gap changes the length of the electron path, while a system of mirrors can delay the arrival of the radiation at the entrance to the second section. The length of the equivalent drift distance can thus be adjusted independently for the electron and radiation.

Each interaction region is similar to a single-element device. From equation (3.59) we obtain the expressions for the electrons' velocity at the output of the two interaction regions.

$$v_2' = v_2 + \frac{A}{\Omega_2}\{(1 - \cos \Omega_2 T) \sin \phi_2 + \sin \Omega_2 T \cos \phi_2\}$$
$$- \frac{A^2\beta}{\Omega_2^3}\{1 - \cos \Omega_2 T - \frac{\Omega_2 T}{2} \sin \Omega_2 T\} \tag{3.65a}$$

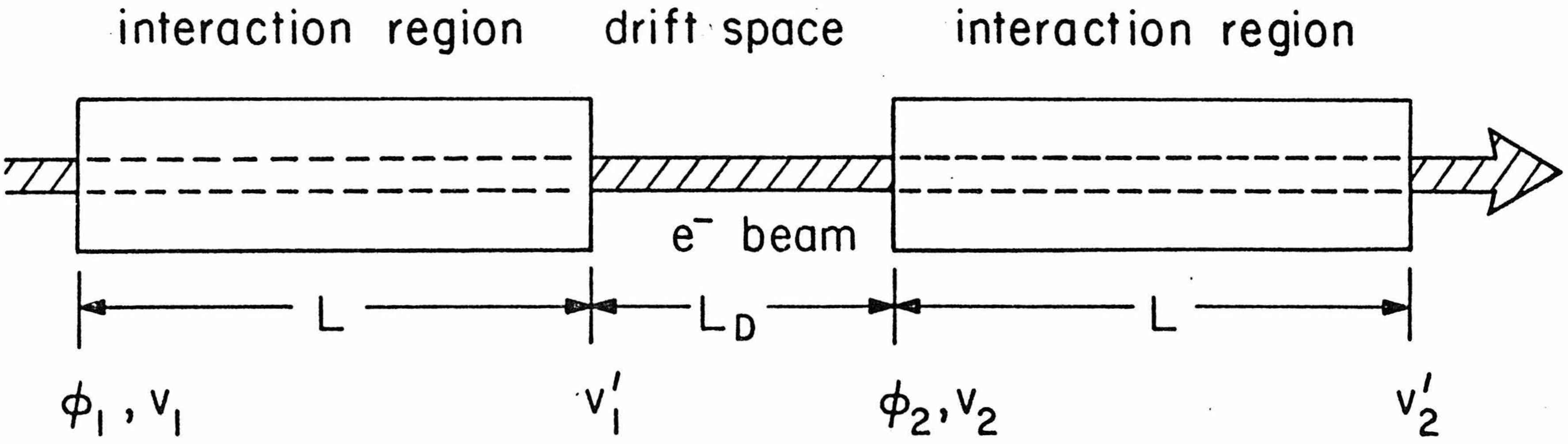

Figure 3-7: A two-stage system with drift space $L_D$.

$$v_1' = v_1 + \frac{A}{\Omega_1} \{(1 - \cos \Omega_1 T) \sin\phi_1 + \sin \Omega_1 T \cos \phi_1\}$$

$$- \frac{A^2\beta}{\Omega_1^3} \{1 - \cos \Omega_1 T - \frac{\Omega_1 T}{2} \sin \Omega_1 T\} \qquad (3.65b)$$

where the subindices, 1 and 2, indicate the quantities in the first and second region. We express the final velocity $v_2'$ in terms of $v_1$ and $\phi_1$ which requires a knowledge of the relations between $v_2$, $\phi_2$, and $v_1$, $\phi_1$. Since $v_{1,2}'$ is considered only up to second order in A, retention of terms in these relations up to first order in E should be sufficient. In a general two-element system we have

$$\Omega_2 = \Omega_1 + \beta\Delta v_1$$

$$\phi_1' = \phi_1 - \Omega_1 T - \beta\Delta z_1$$

$$\phi_2 = \phi_1' - m - n\Delta v_1 \qquad (3.66)$$

where

$$v_1 = \frac{A}{\Omega_1} \{(1 - \cos \Omega_1 T) \sin \phi_1 + \sin \Omega_1 T \cos \phi_1\}$$

$$z_1 = \frac{A}{\Omega_1^2} \{(\Omega_1 T - \sin \Omega_1 T) \sin \phi_1 + (1 - \cos \Omega_1 T)\cos \phi_1\}$$

$$m = \omega(\frac{L_e}{c} - \frac{L_r}{v_1})$$

$$n = \frac{L_e}{v_1^2}$$

From (3.65) and (3.66), the electron velocity at the exit of the second region is expressed in terms of the input conditions (to the first region) as

$$v_2' = v + \frac{A}{\Omega}\{(1 - \cos\Omega T)[\sin\phi + \sin(\phi - \alpha)]$$

$$+ \sin\Omega T[\cos\phi + \cos(\phi - \alpha)]\}$$

$$- \frac{\beta A^2 T^3}{2}\{f(\Omega T)(2 + \frac{n}{\beta T})\sin\alpha + g(\Omega T)(1 + \cos\alpha)\} \qquad (3.67)$$

$$\alpha = m + \Omega T$$

$$f(x) = 2\,\frac{1 - \cos x}{x^2}$$

$$g(x) = 2\,\frac{2 - 2\cos x - x\sin x}{x^3}$$

where we have neglected higher order terms and also the $\phi$-dependent terms of second order. Using (3.67) we can obtain the electron velocity shift and spread in a two-element device. There are three parameters appearing in (3.67): $\Omega$, $\alpha$, and n. The effect of changing n is obvious. It does not alter the electron spread, but does change the magnitude of the shift. Since

$$\frac{n}{\beta T} \simeq \frac{L_e}{L} \qquad (3.68)$$

the velocity shift variation is significant only when the adjustment of the drift distance is comparable to the device length. Although $\alpha$ also depends on the drift distance, we'll show later that the adjustable distance, for which the $\alpha$-dependent factor in the velocity shift reaches its maximum, is about the period of the magnet. Thus, $\alpha$ and n can be seen as two independent parameters.

First we study the problem of maximizing the average velocity shift (the term involving $A^2$ in equation (3.67) and temporarily neglect the problem of velocity spread. Two terms with different dependencies on T are involved. The maximum of $g(\Omega T)$ is 0.27 at $\Omega T = 2.6$, while $f(\Omega T) = 1$ at $\Omega T = 0$. Near resonance ($\Omega = 0$), the first term dominates over the second. Furthermore, the presence of n in the first term makes it possible to increase the shift by using longer drift distances. Neglecting the second term and choosing the operation conditions:

$$\Omega T = 0 \qquad \text{and} \qquad \alpha = \pi/2 \tag{3.69}$$

we have

$$\Delta = -\beta A^2 T^3 (1 + \frac{L_e}{2L}) \qquad \text{and} \qquad \sigma = AT \tag{3.70}$$

for the shift and spread of the two-elements device. We next compare (3.70) with the shift and spread of a single-element device of length 2L which is the total interaction length in (3.70). The drift distance $L_e$ is taken to be equal to the original device length 2L.

|  | 2-element (L - L) | 1-element (2L) |  |
|---|---|---|---|
| $\Delta$ | $-2\beta A^2 T^3$ | $-0.54\beta A^2 T^3$ | (3.71) |
| $\sigma$ | AT | 1.05 AT |  |

Equation (3.71) shows that for a given total interaction distance the velocity shift can be enhanced by a two-element system, especially when the electron drift distance is much larger than the length of each interaction region. Furthermore, it is also shown that the increase in the velocity shift is not accompanied by an increase in the spread of electrons.

The parameter m is determined from the lengths of the optical and electron path. From (3.66) and (3.67) we have

$$\alpha = \frac{2\pi}{\lambda}\left(L_r - \frac{L_e}{1 - \frac{1}{\gamma^2}}\right) \tag{3.72}$$

In the high relativistic limit ($\gamma >> 1$), $\alpha$ becomes

$$\alpha = \frac{2\pi}{\lambda}\left(L_r - L_e - \frac{L_e}{2\gamma^2}\right) \tag{3.73}$$

In the special case where we do not use a bending magnet of mirror system as in Figure3.7,the optical path is equal to the electron path, $L_e = L_r = L_D$.

$$\alpha = -\frac{\pi L_D}{\lambda\gamma^2} \tag{3.74}$$

The optimum operation can be achieved within a change of $2\pi$ in $\alpha$ which means an adjustment of the drift distance within $\Delta L_D$,

$$L_D = 2\lambda\gamma^2 \tag{3.75}$$

$\lambda\gamma^2$ is almost a constant value for a specific device, since the factor of the relativistic up-conversion of the radiation frequency is $4\gamma^2$. For an estimate of a typical value of $\Delta L_D$ we consider the Stanford device and find $\Delta L_D$ is about 5 cm. This value is very reasonable for a practical experimental setup. If $L_r$ and $L_e$ are adjusted independently, the adjustment to an optimum operation is within a distance of radiation wavelength. This may cause power instability in the output radiation, especially at short wavelengths.

The electron energy loss in a two-stage system depends linearly on the electron drift distance $L_e$. With the condition, $\Omega = 0$, it is interesting to point out that the average velocity shift is zero at the exit of the first region. However, the velocity spread of electrons results in electron bunching during free flight in the drift space. To the lowest order approximation, the bunching effect is proportional to the drift distance. The adjustment of the drift distance causes the bunched electron beam to have an optimum entry phase with respect to the radiation which induces the energy loss of the electron beam in the second interaction region.

Equation (3.60) shows that the spread is much larger than the shift in the small signal region. It is found that the first order term in equation (3.67), and hence the first order contribution to the spread, can be made zero by choosing

$$\Omega = 2\pi \qquad \text{or} \qquad \alpha = \pi \tag{3.76}$$

Unfortunately, we also find that the second order (shift) term becomes zero under either of these two conditions. Qualitatively, the spread is now of second order and the shift is fourth order in A. The value of R is still on the order of $A^0$.

This result follows directly from Madey's theorem [7] which can be written in the relativistic approximation as

$$<\Delta v>_\phi = \frac{1}{2} \frac{\partial}{\partial v} [<\Delta v^2>_\phi] \tag{3.77}$$

where triangular parentheses represent the ensemble average over $\phi$. To first order in A we can take $\Delta v$ in general as

$$v = A[M(\Omega t) \cos \phi + N(\Omega t) \sin \phi] \tag{3.78}$$

and

$$< \Delta v >_\phi = \frac{A^2}{2} [M \frac{\partial M}{\partial v} + N \frac{\partial N}{\partial v}] \tag{3.79}$$

In order to eliminate the spread, we have to let M = 0 and N = 0, which leads to the vanishing of the shift $< \Delta v >_\phi$ . We conclude that, up to second order, it is impossible to eliminate the first order spread without sacrificing the second order gain.

We have considered the problem of increasing the shift and eliminating the spread separately. However, the final purpose is to improve the value of R. Here we will investigate the conditions for optimizing R. From (3.69) , we have

$$\sigma^2 = \frac{4A^2}{\Omega^2} (1 - \cos \Omega T)(1 + \cos \alpha)$$

and

$$R = \frac{\beta T}{4} \{(2 + \frac{n}{\beta T}) \tan \frac{\alpha}{2} + (\frac{2}{\Omega T} - \cot \frac{\Omega T}{2})\} \tag{3.80}$$

Although the shift depends on $\alpha$ and $\Omega$ in a complicated manner, the expression for R contains two terms which depend exclusively on $\alpha$ and $\Omega$. So the optimum value for $\alpha$ and $\Omega$ can be found independently. Consider the first term. It becomes infinitely large when $\alpha \to \pi$. The second term which depends only on $\Omega$ approaches infinity as $\Omega T \to 2\pi$. As we have shown before, at these two values the shift is actually zero. However, $\sigma^2$ approaches zero at a faster rate than $\Delta$ which results in an increasing value of R.

In conclusion, we have shown that the efficiency of a free electron laser in beam circulation and the gain of a single-pass device can be greatly improved by using a two-element system. For the gain enhancement the system is operated at the resonance, and $\alpha$ is equal to $\pi/2$. The gain is found linearly related to the drift distance. For the efficiency improvement we choose $\Omega T$ as close as possible to $2\pi$ and $\alpha$ as close as possible to $\pi$. However, the choice of $\Omega T$ and $\alpha$ must be such that single pass gain is higher than the threshold condition.

## CHAPTER 3 - REFERENCES

1. J.M.J. Madey, J. Appl. Phys. 42, 1906 (1971).
2. J. R. Pierce, Traveling-Wave Tubes (New York: Van Nostrand, 1950).
3. C. Shih and A. Yariv, Tenth International Quantum Electronics Conference, Atlanta, Georgia, May 29, 1978.
4. F. A. Hopf, P. Meystre, G. T. Moore, and M. O. Scully, in Physics of Quantum Electronics, S. F. Jacobs, M. Sargent III, and M. O. Scully, eds., Vol. 5 (Addison Wesley, Reading, Mass., 1978).
5. C. Shih and A. Yariv, Optics Letters 5, 76 (1980).
6. D. L. Webster, J. Appl. Phys. 10, 501 (1939).
7. J.M.J. Madey, Nuovo Cimento 50B, 64 (1979).

# Chapter 4

# NONLINEAR THEORY OF LONGITUDINAL FREE ELECTRON LASER

## 4.1 Introduction

In Chapters 1-3 we formulated the linear theory of free electron lasers based on three assumptions: low current density, small field amplitude, and small stimulated gain. In this chapter, the nonlinear theory is introduced by removing the three restrictions. In a dense electron beam, the space-charge fields play an important role in the electron dynamics and cannot be neglected. The Coulomb field is included in the electron modulation and the stimulated process. By incorporating Poisson equation with the force equation, a gain expression which is valid at arbitrary current densities is obtained. For an arbitrary field amplitude the force equation becomes a simple undamped pendulum equation. The equation can be solved in terms of special functions. The distribution of electrons in velocity and real space is described. A gain map is then obtained to show the saturation due to the radiation intensity. The case of large gain is then considered. In the limit of small field, an integral equation is derived to describe the growth of the field amplitude. The equation is solved exactly. The solution demonstrates clearly the regions of bound and exponential gains. In the limit of large fields, the method of harmonic expansion is used. A set of coupled differential equations demonstrates the amplitude growth and the electron dynamics.

## 4.2 Space-Charge Effect

In the linear analysis we have neglected the Coulomb interaction between electrons. Each electron was assumed to interact with the radiation field alone. Obviously this assumption is valid only if the current density is very small. In the Stanford experiment of stimulated amplification it was found that the gain was actually proportional to the current up to 70 mA. With the beam diameter of 1 mm this corresponds to a current density of 7 A/cm$^2$. So the experiment concluded that there is no detectable saturation effect below 7 A/cm$^2$. The following questions can be asked. Is there any saturation phenomenon due to high current density? How high is the current density at which gain saturation begins? It seems apparent that the gain should deviate from its linear relation with the current density. But it must be determined whether the gain increases more slowly or whether it saturates completely.

Qualitative or quantitative answers to these questions are important to the design of electron beam generators for the free electron laser.

Using the concept of plasma resonant waves in the electron beam, the space-charge effect has been considered to the lowest order in the limit of the large cavity [1]. By solving the Maxwell-Boltzmann equations coupled with Poisson's equation, this effect is also evaluated to the lowest order in the limit of the small cavity [2]. However, these solutions show only how the gain is suppressed when the current density is slightly above the linear region. They do not provide any quantitative information as to the value of the saturation current density. Any estimate of the saturation current density requires an exact solution.

Due to the repulsion between electrons, two effects will be observed in the interaction region. For the radiation, it generates a net Coulomb field superimposed upon the electromagnetic field. For the electron beam, the velocity and position of electrons are modified. Because the divergence of the electron beam due to the repulsive force is very small, we neglect the dependence on the transverse variable, x and y. In the following analysis, we do not assume any plasma wave in the electron beam, i.e., no wave-wave interaction is preliminarily taken into account. The dynamic equation, coupled with the Poisson equation, is used to obtain the gain.

In Section 3.2 we have solved the collisionless force equation to obtain the electron position as in (3.11)

$$Z(t) = v_0 t + \frac{eE}{m\gamma^3\Omega^2}\{(1 - \cos \Omega t)\cos\phi + (\Omega t - \sin \Omega t)\sin \phi\} \quad (4.1)$$

Equation (4.1) describes the trajectory of an electron which is assumed to pass the entrance of the interaction region z = 0, when t = 0, with phase $\phi$. Such a periodic dependence on $\phi$ results in a non-uniform beam which generates a space-charge field. Equation (3.9) describes only the situation when this field is negligible compared to the ponderomotive force. In general, the space-charge field $E_c$ should be included in the analysis, since the total field to which an electron is subjected is the sum of this field as well as the external applied field.

The space-charge field $E_c$ is included in the equation as

$$\frac{d^2}{dt^2}\Delta z(t) = \frac{eE}{m\gamma^3}\cos[\Omega t - \beta\Delta z(t) + \phi] + \frac{eE_{cz}}{m\gamma^3} \quad (4.2)$$

The second term on the right side represents the contribution of the space-charge effect. $\gamma^3$ in the denominator comes from the relativistic consideration.

If the change of the transverse space-charge field is assumed to be very small and neglected, the longitudinal space-charge field obeys the Poisson equation

$$\frac{\partial}{\partial z} E_{cz}(z,t) = \frac{e}{\varepsilon_0} [N(z,t) - N_0] \tag{4.3}$$

where $N(z,t)$ is the electron density at position z and time t; $N_0$ is the initial electron density. In order to follow the evolution of the electron density, we consider an infinitesimal section of the beam at $z_0$ when $t = 0$. Its width is $\delta z_0$ and density is $N_0$. After time t, it propagates to position z and develops into a section with width $\delta z$ and density $N(z,t)$. According to the conservation of the electron number, the following relation is correct provided that the electrons retain their orders in space during the propagation (the single-stream assumption)

$$N(z,t)\ \delta z = N_0\ \delta z_0 \tag{4.4}$$

or equivalently,

$$N(z,t) = N_0 \left(\frac{\partial z}{\partial z_0}\right)^{-1} \tag{4.5}$$

In general, the position z is a function of t and $z_0$. it can be written as

$$z(z_0,t) = z_0 + v_0 t + \Delta z(z_0,t) \tag{4.6}$$

Substituting (4.6) into (4.5) and assuming $(\partial \Delta z/\partial z_0)$ is small, we have

$$N(z,t) = N_0 [1 - \frac{\partial}{\partial z_0} \Delta z(z_0,t)] \tag{4.7}$$

Using (4.7) in (4.3), we obtain

$$\frac{\partial}{\partial z} E_{cz}(z,t) = - \frac{eN_0}{\varepsilon_0} \frac{\partial}{\partial z_0} \Delta z(z_0,t) \qquad (4.8)$$

Equation (4.8) involves partial derivatives with respect to different variables: z and $z_0$. However, they are equivalent in the case where only the partial differentiation is concerned, and $\Delta z$ is small compared to $(z_0 + v_0 t)$. This approximation is valid even in the strong signal regime. The integration of (4.8) over $z_0$ leads to

$$E_{cz}(z,t) = - \frac{eN_0}{\varepsilon_0} [\Delta z(z_0,t) + h(t)] \qquad (4.9)$$

The function h(t) does not depend on $z_0$. Since $E_{cz}$ becomes zero when $\Delta z$ is uniform (i.e., independent of $z_0$), it is natural to identify h(t) as the ensemble average of the position deviation $<\Delta z(z_0,t)>_{z_0}$. It is noted that it does not make any difference if we replace $z_0$ by $\phi$ to label electrons. We have thus found a way to relate the space-charge field to the dynamic variable of an electron, $\Delta z$:

$$E_{cz}(z,t) = - \frac{eN_0}{\varepsilon_0} [\Delta z(\phi,t) - <\Delta z(\phi,t)>_{\phi}] \qquad (4.10)$$

Physically, it means that the space-charge field experienced by an electron is proportional to its "net" position deviation.

We have shown that the space-charge term in (4.2) can be related to the single electron position deviation through the key equation (4.10). Since no assumption was made concerning the electron density, the analysis which follows should apply to beams with arbitrary current density provided other conditions are satisfied. By combining equations (4.2) and (4.10) we can write the force equation as

$$\frac{d^2}{dt^2}\Delta z + \omega_p^2[\Delta z - \langle\Delta z\rangle_\phi] = \frac{eE}{m\gamma^3}\cos[\Omega t - \beta\Delta z + \phi]$$
$$\omega_p^2 = \frac{e^2N_o}{\varepsilon_o m\gamma^3} \qquad (4.11)$$

$\omega_p$ is the relativistic plasma frequency at the electron density $N_o$. To solve for $\Delta z$, we consider the perturbation expansion in the limit of small radiation field

$$\Delta z = \Delta z^{(1)} + \Delta z^{(2)} + \cdots \qquad (4.12)$$

where $\Delta z^{(n)}$ is the nth order deviation proportional to $E^n$. Substituting (4.12) into (4.11) and considering the self-consistency in $\phi$, we find $\langle\Delta z\rangle_\phi$ contains only even-order terms. Therefore, $\langle\Delta z\rangle_\phi$ is at most a second order effect. The solution is

$$\Delta z^{(1)} = \frac{eE/m\gamma^3}{\omega_p^2 - \Omega^2}\{(\cos\Omega t - \cos\omega_p t)\cos\phi + (\sin\Omega t - \frac{\Omega}{\omega_p}\sin\omega_p t)\sin\phi\} \qquad (4.13)$$

and

$$\Delta z^{(2)} = \frac{(eE/m\gamma^3)^2\beta}{4\omega_p(\omega_p^2 - \Omega^2)^3}\{(\omega_p+\Omega)^3\sin(\omega_p-\Omega)t - (\omega_p-\Omega)^3\sin(\omega_p+\Omega)t - 4\,\omega_p\Omega(\omega_p^2-\Omega^2)t\} \qquad (4.14)$$

where, in $\Delta z^{(2)}$, only the part independent of $\phi$ is written explicitly.

The modulation of the electron position results in the modulation of the beam density which in turn can drive the radiation field according to Maxwell's equations. Since we are only interested in the energy gain of the radiation within the constant field approximation, it is more convenient and straightforward to consider directly the energy exchange between the electron beam and the radiation. In the case where the Coulomb interaction is neglected, the energy loss of electrons is converted completely into the radiation energy. However, when Coulomb interactions are considered, the energy extracted from the beam must be distributed between the radiation and the space-charge field. Since only the increase in the radiation field is available as useful output, we must be able to calculate the increase in the space charge field energy and subtract it from the total energy lost by the beam.

The energy change of an electron in a single pass can be calculated from (3.4) with the integrand including the longitudinal space-charge field as well as the transverse radiation field

$$\vec{v}\cdot\vec{E} = v_{\perp} E_r + (v_0 + \Delta v_z)\, E_{cz} \tag{4.15}$$

If the ensemble average is taken before the integration is executed, we find immediately $< v_0 E_{cz} >$ disappears

$$< E_{cz} >_{\phi} = - \frac{eN_0}{\varepsilon_0} < \Delta z - <\Delta z>_{\phi} >_{\phi} \equiv 0 \tag{4.16}$$

We also find the energy loss due to $(\Delta v_z E_{cz})$ is

$$\begin{aligned} - \left|\frac{J}{e}\right| e < \int \Delta v_z E_{cz}\, dt >_{\phi} \\ = \left|\frac{J}{e}\right| \frac{e^2 N_0}{\varepsilon_0} < \int \Delta v_z\, \Delta z\, dt >_{\phi} \\ = \frac{c(eN_0)^2}{2\varepsilon_0} < (\Delta z)^2 >_{\phi} \end{aligned}$$

$$= \frac{\varepsilon_o c}{2} < E_{cz}^2 >_\phi \tag{4.17}$$

To obtain the result in (4.17), we have neglected $<\Delta z>_\phi$ in $E_{cz}$ because it is at second order, which results in a third order term in the energy exchange after multiplication by $\Delta v_z$. Physically, (4.17) shows that the energy loss due to $\Delta v_z E_{cz}$ is exactly equal to the space-charge energy. Therefore, the energy increase of the radiation comes exactly from the contribution of $v_\perp E_r$. With the explicit expression of $\Delta z^{(1)}$, we estimate roughly that the energy for the build-up of the space-charge field is only a very small part of the energy loss of the electron beam. Their ratio is $\sim\Omega/\omega$ or $\sim\lambda/\ell$, which is only $10^{-4}$ for the Stanford device.

Following a procedure similar to that used to derive the no-space-charge gain expression (3.24), we find that when we include the space-charge field the gain becomes

$$G(\theta,\theta_p) = G_o' \frac{\theta\,\theta_p^2}{(\theta_p^2 - \theta^2)^2} \{2 - 2\cos\theta_p \cos\theta - (\frac{\theta_p}{\theta} + \frac{\theta}{\theta_p})\sin\theta_p \sin\theta\}$$

$$G_o' = G_o/\theta_p^2 \tag{4.18}$$

$$\theta \equiv \Omega T \qquad , \qquad \theta_p \equiv \omega_p T$$

$G_o$ is a constant independent of $\theta$ and $\theta_p$. It is interesting to note that the gain spectrum is almost the same in terms of either variable, $\theta$ or $\theta_p$, although they have completely different physical meanings. $\theta_p$ indicates the electron density, while $\theta$ represents the velocity detuning from the resonance condition. The condition $\theta = \theta_p$ leads to a well-known phenomenon, "plasma resonance." However, it is approached for the first time from the single-particle point of view. For an appreciation of the value of $\theta_p$, we calculate it using the example of Section 3.2, L = 10 cm, $\gamma = 3.64$, I = 1 mA in a waveguide of 50 μm x 50 μm. The value of $\theta_p$ is found to be about 0.3

It is expected that the new gain expression (4.18) should reduce to (3.24) when the current density is very small. Indeed, if let $\theta_p$ approach zero, we find

$$G^{(1)} = G'_o \, \theta_p^2 \, f(\theta) \tag{4.19}$$

where the superscript (1) indicates that the gain is proportional to the first power of the electron density. To obtain the lowest order correction to the collisionless gain, we expand (4.18) up to the order of $\theta_p^4$ and find it to be

$$G^{(2)} = -G'_o \, \theta_p^4 \, g(\theta) \tag{4.20}$$

$$g(\theta) \equiv \frac{(24-6\theta^2)\cos\theta + (18\theta - \theta^3)\sin\theta - 24}{6\theta^5}$$

The result in (4.20) is identical to that obtained by Louisell et al. [2] using the coupled Maxwell-Boltzmann equations. The fundamental spectrum $f(\theta)$ and the correction function $g(\theta)$ are shown in Figure 4.1. Up to the first-order correction, the gain becomes smaller for $\theta < 4.6$. The non-uniform reduction results in an up-shift of $\theta_{max}$. The up-shift is proportional to the electron density and can be written as

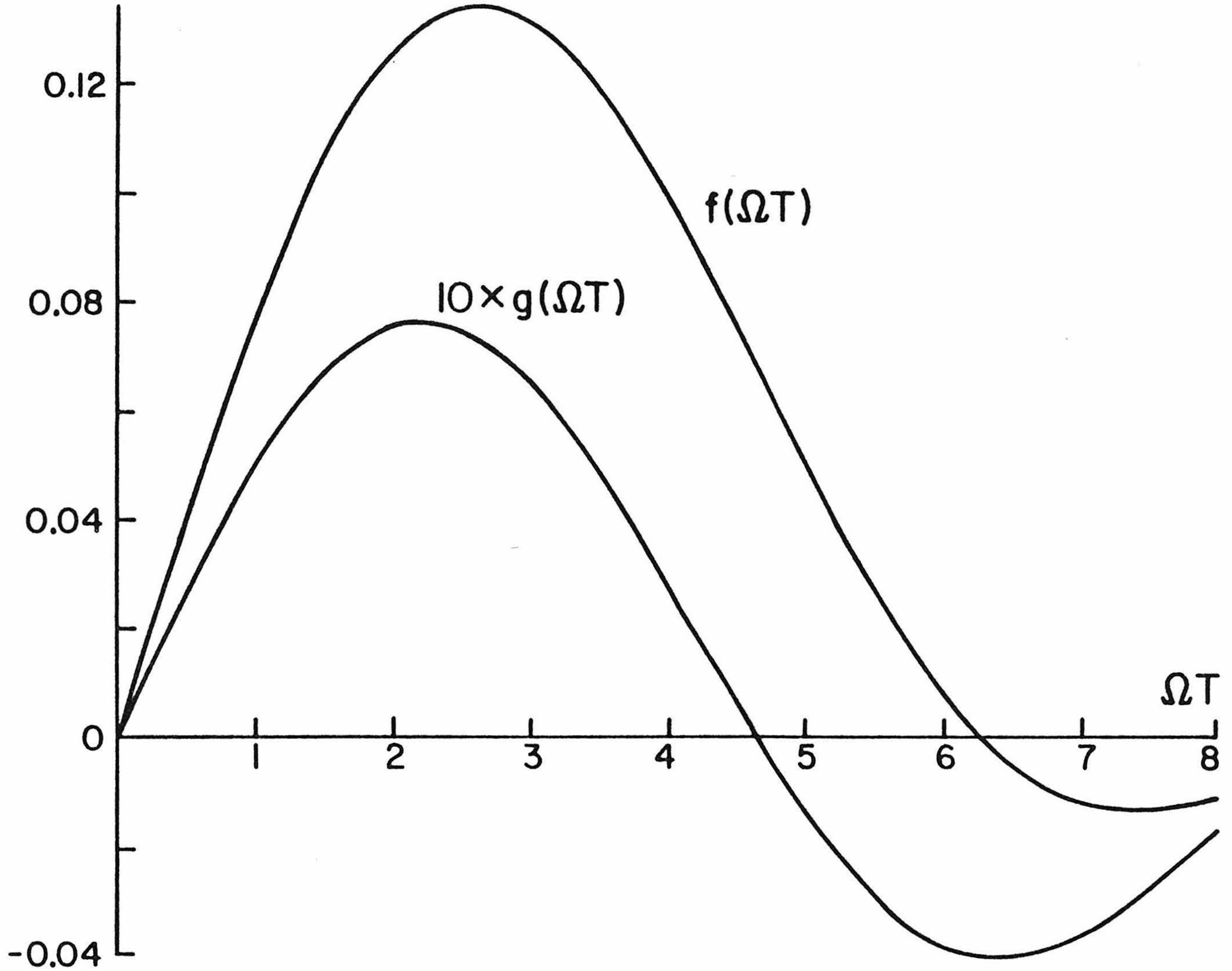

Figure 4-1: f(ΩT) is the gain without the space-charge consideration. g(ΩT) is the first order correction.

$$\Delta\theta_{max} = \frac{g'(\theta_{max})}{f''(\theta_{max})}\,\theta_p^2 \qquad (4.21)$$

To demonstrate the phenomenon of gain saturation, the normalized gain $[G(\theta,\theta_p)/G_o'\ \theta_p^2]$ is plotted for different values of $\theta_p$ (Fig. 4.2). The reason why we normalize the gain with respect to the electron density (through $\theta_p^2$) is to compare it with the gain in the collisionless situation, where it is proportional to the electron density. Therefore, the normalized gain for $\theta_p = 0$ as shown in Figure 4.2 corresponds to the case of collisionless electron beam. In general, it is observed that the peak (normalized) gain decreases and shifts to the right with increasing electron density. Physically, the reduced gain is due to the repulsive force between the electrons which weaken the tendency of the electrons to bunch together. This reduces the beam alternating current which can couple to the electromagnetic field. The increase of $\theta_{max}$ with $\theta_p$ is due to increasing plasma frequency. In a practical device which is used as an amplifier, the radiation frequency is fixed by the input field. If the electron energy does not change (i.e., $\theta$ is a given constant), the normalized gain drops very fast with the current density. If the electron energy is adjustable, we can choose $\theta$ to correspond to the value which yields the maximum gain. This reduces the effect of saturation. If the device is used as a laser oscillator, the radiation frequency adjusts itself automatically until the gain is maximum. It thus makes sense to study the effect of gain saturation by enquiring what happens to the peak gain as a function of the space-charge parameter $\theta_p$.

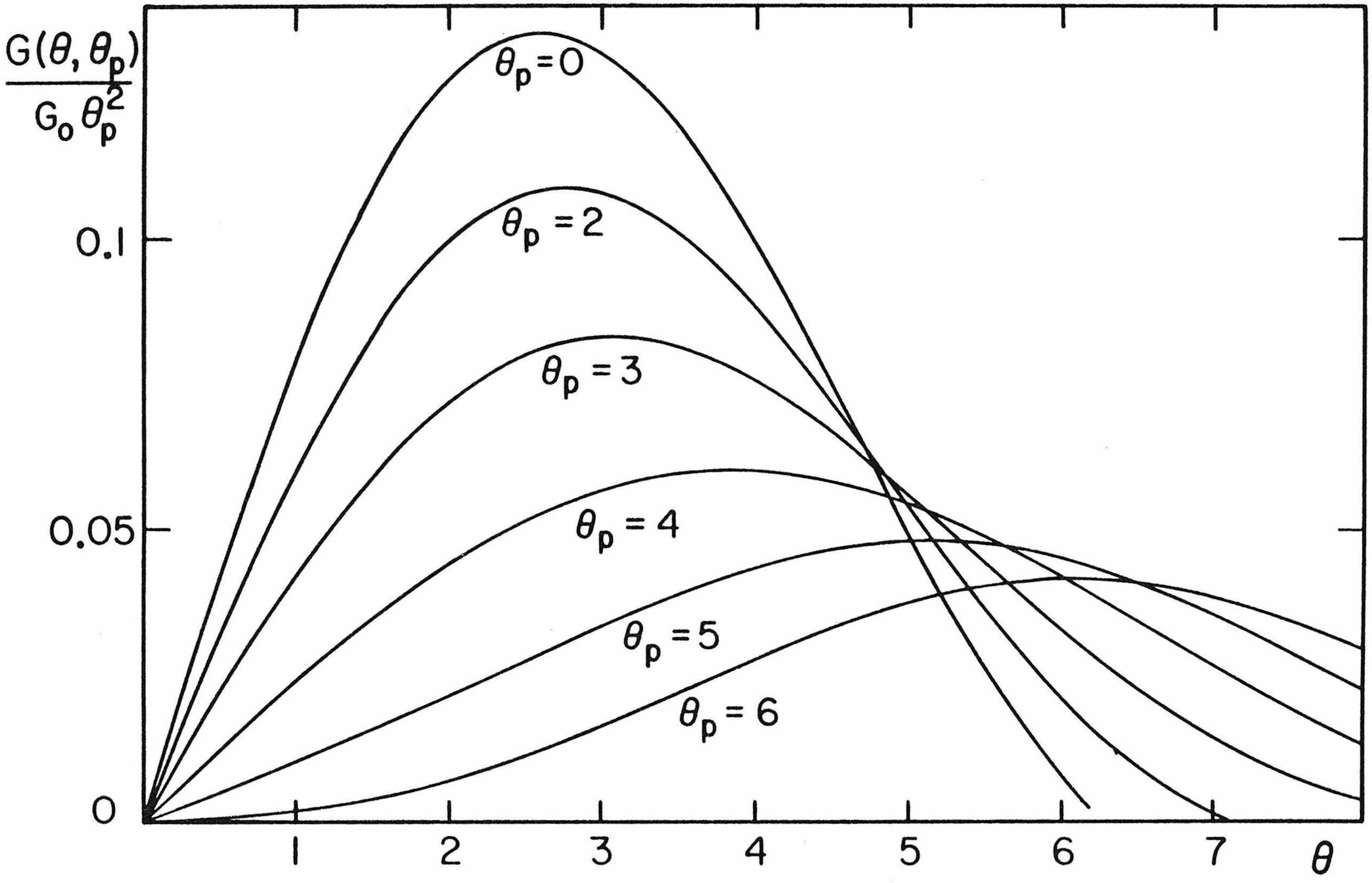

Figure 4-2: Normalized gain $[G(\theta,\theta_p)/\theta_p^2]$ for different electron density $(\theta_p)$.

The behavior of the maximum gain is easier to follow if we use an alternate expression for the gain

$$G(\theta,\theta_p) = \frac{G_o'\,\theta_p}{2}\left\{\frac{1-\cos(\theta-\theta_p)}{(\theta-\theta_p)^2} - \frac{1-\cos(\theta+\theta_p)}{(\theta+\theta_p)^2}\right\} \qquad (4.22)$$

For given $\theta_p$, the value of $\theta_{max}$ can be found from the solution of the equation

$$\frac{\partial}{\partial\theta} G(\theta,\theta_p) = \frac{G_o'\,\theta_p}{2}\;\{f(\theta+\theta_p) - f(\theta-\theta_p)\} = 0 \qquad (4.23)$$

where f is a function identical to the fundamental spectrum appearing in (3.24). It is obvious that $(\theta,\theta_p) \equiv (n\pi, m\pi)$ is always a solution of (4.23) whenever $(n \pm m)$ is an even integer number. The gain becomes a local maximum at these positions. Among those solutions, it can be observed that the solution $\theta = \theta_p$ leads to the overall maximum gain for given $\theta_p = m\pi$. The curve in Figure 4.3 shows the trace of the maximum gain. It crosses the plasma resonance line $(\theta = \theta_p)$ whenever $\theta_p$ is a multiple of $\pi$, or, in general, is a solution of the equation $f(2\theta_p) = 0$.

The radiation frequency is determined in terms of the detuning parameter

$$\omega = 2\gamma^2 c\left(\frac{2\pi}{\ell} + \frac{\theta}{L}\right) \qquad (4.24)$$

Therefore, the curve reveals clearly the transition of the radiation frequency from the single-particle to the plasma region. $\theta$ represents the deviation of the frequency from the "lattice frequency" $(2\pi c/\ell)$. At low electron density $(\theta_p \rightarrow 0)$, $\theta$ approaches the value of 2.6 where the field-interference process dominates. When $\theta_p$ begins to increase, $\theta$ approaches the value of $\theta_p$ very fast and starts to oscillate around the line

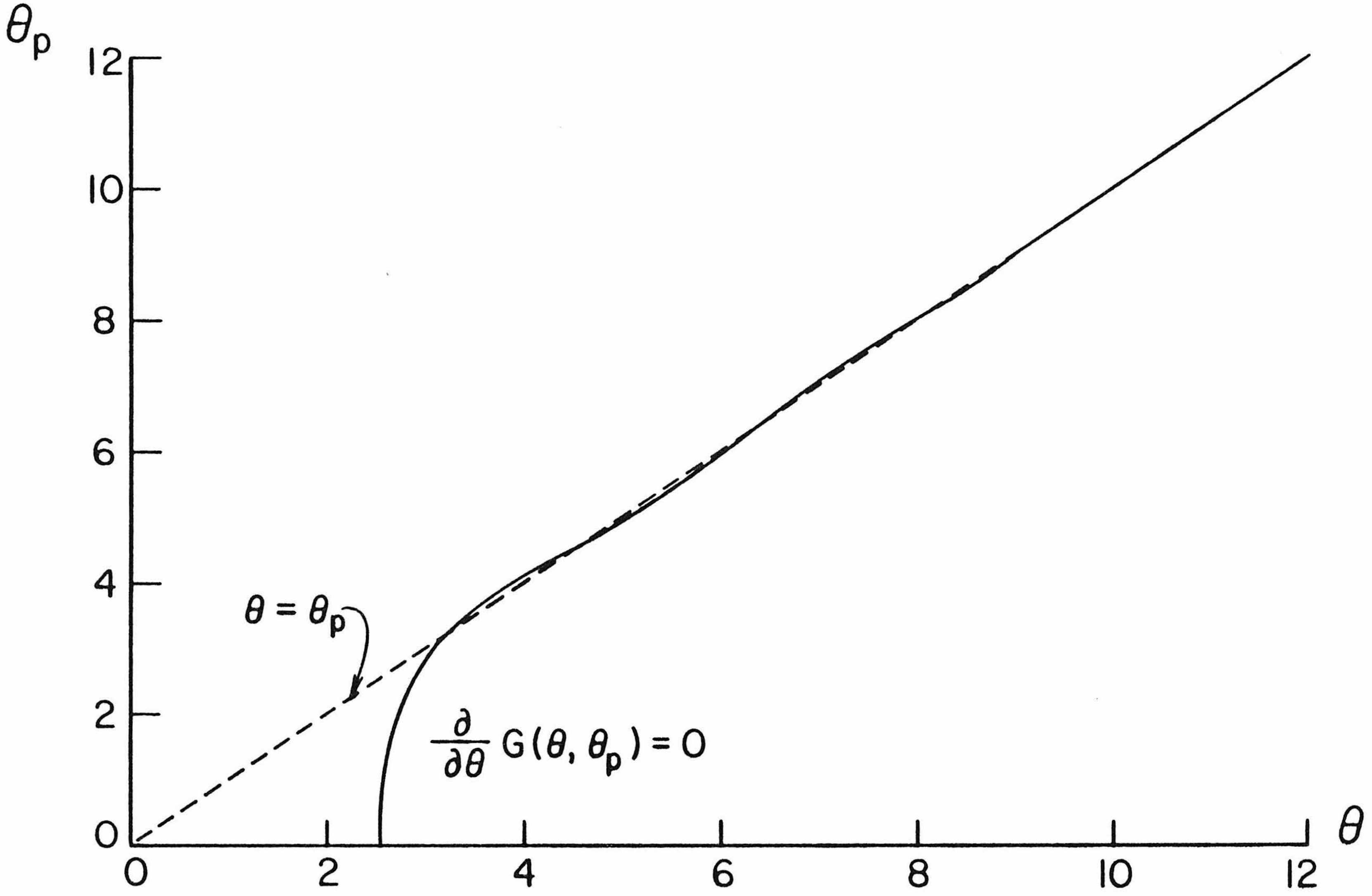

Figure 4-3: The trace of the maximum gain on $(\theta,\theta_p)$.

$\theta = \theta_p$ with a decreasing oscillating amplitude

$$(\theta - \theta_p) \xrightarrow[\theta_p \to \infty]{} - \frac{3 \sin 2\theta_p}{\theta_p^2} \tag{4.25}$$

The point $\theta_p = \pi$ is then observed to serve as a critical boundary where the transition between two regions occurs. Above this point, the effect of plasma resonance dominates. The radiation frequency is then given by the sum of the lattice and plasma frequencies except for the small deviation (4.25).

We have shown the frequency shift due to the space-charge effect. The maximum gain along this curve is shown in Figure 4.4 as a function of $\theta_p^2$. In the limit of small electron density, the maximum gain is proportional to $\theta_p^2$ ($G_{max} = 0.135\ G_o\theta_p^2$). When the beam density increases, the maximum gain begins to saturate with a smaller growth rate. However, there is no upper bound to the gain. In the limit of high electron density the gain is proportional to the square root of the electron density ($G_{max} \to \theta_p/4$).

We have performed all the quantitative analyses in terms of the dimensionless parameter $\theta_p$. In order to get an appreciation of its value in a practical device, we write $\theta_p^2$ as

$$\theta_p^2 = 7.382\ \frac{J[A/cm^2]\ L^2[m]}{\gamma^3} \tag{4.26}$$

where J is the current density in the unit of amperes per centimeter squared. For the typical example of longitudinal free electron laser, we have shown in Section 3.2, $\theta_p^2$ is found to be about 0.3. Thus it is still far from the space-charge saturation.

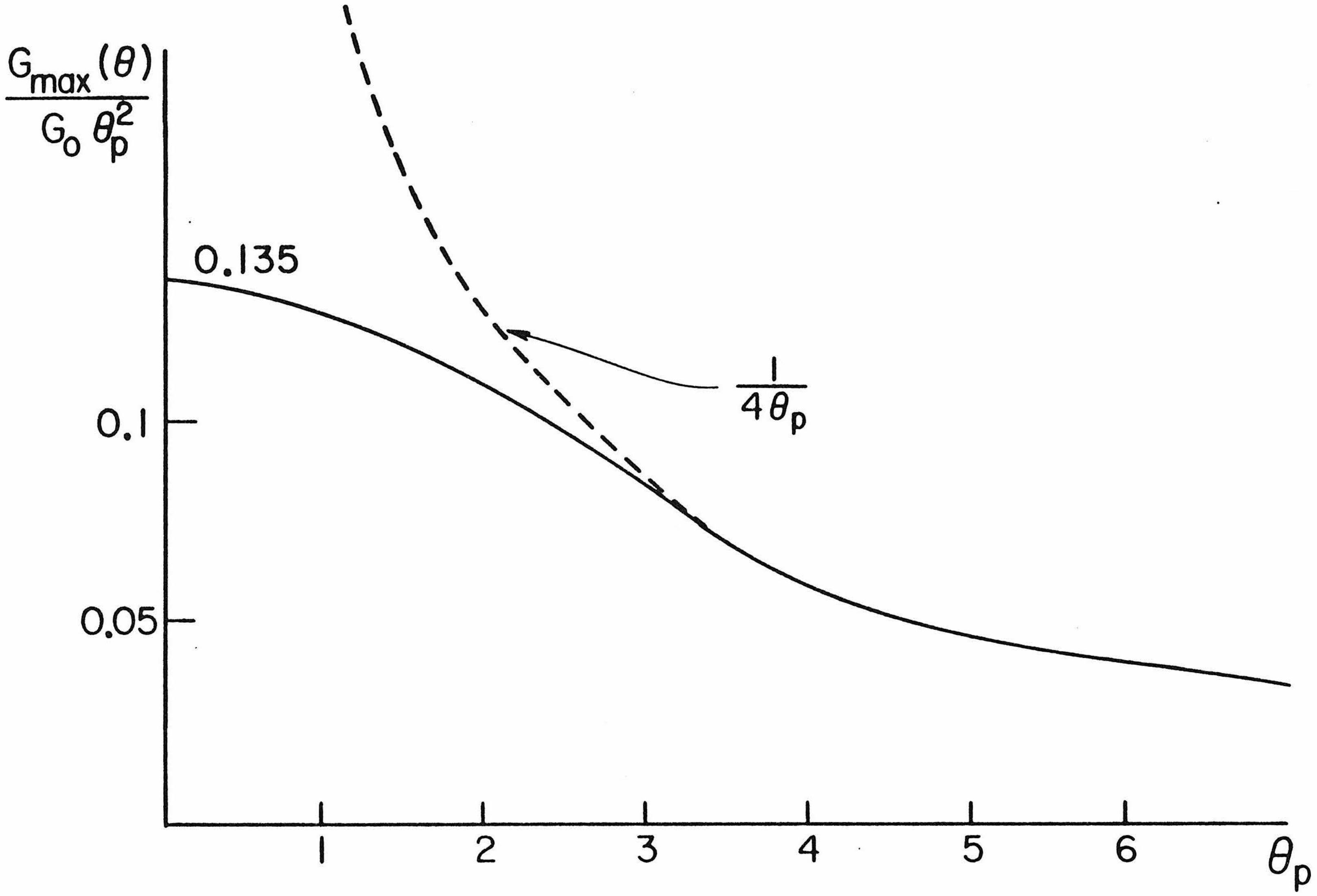

Figure 4-4: The normalized gain along the solid line in Figure 4-3 as function of the electron density ($\theta_p$). The normalized gain reduction shows the saturation.

If the pumping is strong, such that the gain is high enough, the increase of the field amplitude in the interaction region can no longer be neglected. In this case, a self-consistent treatment of the field is required. However, the self-consistent method is important only when it is applied to a high-gain amplifier. For the analysis of a laser device using a resonator with high reflectivity mirrors, the field amplitude increase per pass must equal the small mirror loss per pass, and the constant amplitude approximation is very good.

In summary, we have included the space-charge effect exactly in the single-electron analysis of free-electron lasers. The gain is found in (4.18) through the key equation (4.10). In the small signal and low gain region, the maximum gain is found to saturate at high electron densities and the radiation frequency approaches the condition of the plasma resonance.

## 4.3 The Pendulum Analysis

In Section 4.2 we have removed the small space charge, and in the process found how the gain saturates at high current densities. The saturation current density is then defined quantitatively. Although the value of the current density is beyond the limits of current experiments, it deserves attention for application to future experiments at higher currents. We have to point out that this is the unique feature of the free electron laser. A similar situation, but not quite the same in the conventional gas laser, is the depopulation of the upper level due to molecular collisions. The relaxation time decreases with increasing gas density. Another saturation mechanism in the conventional laser is due to the high radiation intensity. The gain drops as the intensity goes up. The same holds true in the free electron laser.

In order to investigate the saturation due to the radiation intensity, the perturbation method is no longer useful. The applicability of the power expansion in equation (4.11) is limited by the requirement that the expansion constant must be less than 1. This criterion leads to

$$\frac{eE\beta T^2}{m\gamma^3} < 1 \tag{4.53}$$

or, equivalently,

$$eEL \cdot \beta L < mc^2\gamma^3 \tag{4.54}$$

In the given example of the longitudinal free electron laser in Section this corresponds to

$$E < 4 \times 10^3 \text{ V/m} \tag{4.55}$$

and the energy flux is about

$$S = \frac{\varepsilon_o}{2} c\ E^2 \simeq 16 \text{ Watts/cm}^2 \tag{4.56}$$

The equivalent condition for the transverse free electron laser is

$$\frac{2e^2EBT^2}{m^2\gamma^2c^2} < 1 \tag{4.57}$$

In the stimulating amplification, $\gamma = 50$ and $B \simeq 2.4$ kG , the condition is equivalent to

$$E < 1.7 \times 10^5 \text{ V/m} \tag{4.58}$$

and the energy flux is about

$$S = 2.56 \times 10^4 \text{ Watts/cm}^2 \tag{4.59}$$

Therefore, the linear theory breaks down for an energy flux higher than (4.56) or (4.59). In the laser oscillation experiment, the gain per pass is only about 1.5%. The field amplitude can be assumed to be constant during the interaction. The change of $\gamma$ is on the order of less than $10^{-2}$ and is negligible. The original force equation we obtained in Section 3.1 is

$$\frac{dv}{dt} = \frac{eE}{m\gamma^3} \sin(\omega t - \beta z + \phi_0) \tag{4.60}$$

Define $\phi \equiv \phi_0 + \omega t - \beta z$ which is the phase of the wave seen by the electron at z and t. As $d^2\phi/dt^2 = \beta\, dv/dt$, we have

$$\frac{d^2\phi}{dt^2} = - \frac{\beta eE}{m\gamma^3} \sin \phi \tag{4.61}$$

The equation is identical to a pendulum equation with pendulum length $\ell$, acceleration constant g, and $\frac{\beta eE}{m\gamma^3} = \frac{g}{\ell}$ . Multiplying both sides of (4.61) by $2\dot{\phi}$ where the dot means the derivative with respect to t,

$$\frac{d}{dt}[\dot{\phi}^2] = - \frac{2\beta eE}{m\gamma^3} [\sin\phi \cdot \dot{\phi}] \tag{4.62}$$

Integration of (4.62) results in

$$\dot{\phi}^2 - \dot{\phi}_0^2 = \frac{2\beta eE}{m\gamma^3} [\cos\phi - \cos\phi_0] \tag{4.63}$$

where

$$\dot{\phi}_0 = \omega - \beta v_0 = - \Omega \tag{4.64}$$

Equation (4.63) can be written as

$$\dot{\phi}^2 = \frac{4\beta eE}{m\gamma^3 R^2}\{1 - R^2\sin^2\phi/2\} \tag{4.65}$$

where

$$R = \{\frac{m\Omega^2\gamma^3}{4\beta eE} + \sin^2\phi_o / 2\}^{-1} \tag{4.66}$$

With the change of variable,

$$\eta = \phi/2 \tag{4.67}$$

we have

$$\int_{\eta_o}^{\eta} \frac{d\eta}{\sqrt{1 - R^2\sin^2\eta}} = \sqrt{\frac{\beta eE}{m\gamma^3 R^2}} \int_0^t dt \tag{4.68}$$

The integral on the left side is, by definition, the Jacobian elliptic function [3]. The exact solution of the force equation is then obtained as

$$\sin\eta = \mathrm{sn}(u|R^2) \qquad \text{if} \qquad R^2 < 1 \tag{4.69}$$

or

$$\sin\eta = \frac{1}{R}\,\mathrm{sn}(Ru\Big|\frac{1}{R^2}) \;\text{if} \qquad R^2 > 1 \tag{4.70}$$

$$u = \sqrt{\frac{\beta eE}{m\gamma^3 R^2}}\, t + \int_0^{\eta_o} \frac{d\eta}{\sqrt{1 - R^2\sin^2\eta}} \tag{4.71}$$

where sn is the Jacobian elliptic function with parameter $R^2$. Comparing (4.69),(4.70) with the condition of stability (3.43), we find that $R^2 < 1$ and $R^2 > 1$ separate the electron stream lines into regions of instability and stability.

Solution of equation (4.69) or (4.70) represents the evolution of the phase as a function of time. It defines the flow rate of the electron on its own phase diagram stream line. The energy change of an electron is found from

$$\begin{aligned}\Delta\varepsilon &= ev_0 \int E_e \, dt \\ &= ev_0 \, E \int_0^T \sin\phi \, dt \\ &= 2ev_0 \, E \int_0^T \sin\eta \sqrt{1-\sin^2\eta} \, dt \end{aligned} \qquad (4.72)$$

Substituting (4.69) into (4.72) and using the integral formula for the Jacobian elliptic function, we obtain

$$\Delta\varepsilon = 2ev_0 E \int_0^T \mathrm{sn}(u|R^2) \sqrt{1-\mathrm{sn}^2(u|R^2)} \, dt \qquad (4.73)$$

$$= -\frac{2ev_0 E}{R^2}\sqrt{\frac{m\gamma^3 R^2}{\beta eE}} \cdot \sqrt{1-R^2\mathrm{sn}^2(u|R^2)}\,\Big|_0^T \qquad (4.74)$$

$$= -\sqrt{\frac{4mv_0^2\,\gamma^3 eE}{\beta R^2}}\left\{\sqrt{1-R^2\mathrm{sn}^2(u(T)|R^2)} - \sqrt{1-R^2\sin^2\eta_0}\right\} \qquad (4.75)$$

We have obtained the analytic form for $\Delta\varepsilon$ in terms of the special function $\mathrm{sn}(u|R^2)$. However, the dependence of $\Delta\varepsilon$ on $\phi_0$ is very complicated. It is impossible to average $\Delta\varepsilon$ over the entry phase analytically. The final gain is calculated numerically and normalized to the maximum gain of $\Omega T = 2.6$, which obtains when $E \rightarrow 0$.

$$G_{(\mathrm{normalized})} = 30\left[\frac{m\gamma^3}{\beta eET^2}\right]^{3/2} \left\langle \sqrt{\frac{1}{R^2} - \mathrm{sn}^2(u|R^2)} - \sqrt{\frac{1}{R^2} - \sin^2\frac{\phi_0}{2}} \right\rangle_{\phi_0} \qquad (4.76)$$

The result is plotted in Figure 4.5. The contour indicates the level of constant gain. For a given field intensity, the maximum gain shifts to higher values of $\Omega T$. However, the gain does not approach zero as the field intensity monotonically goes to infinity, but oscillates between positive and negative values. This phenomenon is quite different from saturation in conventional lasers. The reason for the difference is that they have different energy transfer mechanisms. In the conventional laser, the transition of an electron is between two well-defined energy levels, so all the electrons are subjected homogeneously to the same energy transfer and also the same saturation condition. In the free electron laser, the energy transfer of electrons is not homogeneous and depends on their entry phase. The net result comes from the ensemble average over all the electrons with different entry phases. Although the gain is zero at certain field levels, it does not mean that there is no energy transfer for individual electrons. It is only the average which goes to zero. Every electron is still active in interacting with the radiation. Thus, when the field intensity increases further, a nonzero gain will appear again, but with different sign.

We are also interested in the electron dynamics in the region of arbitrary field intensity. With a monoenergetic and uniform input electron beam, the distribution in velocity and phase can be formulated in the following way. The results show the velocity spreading and phase bunching effects.

As in the analysis of the electron dynamics, the velocity distribution is found by the conservation of probability

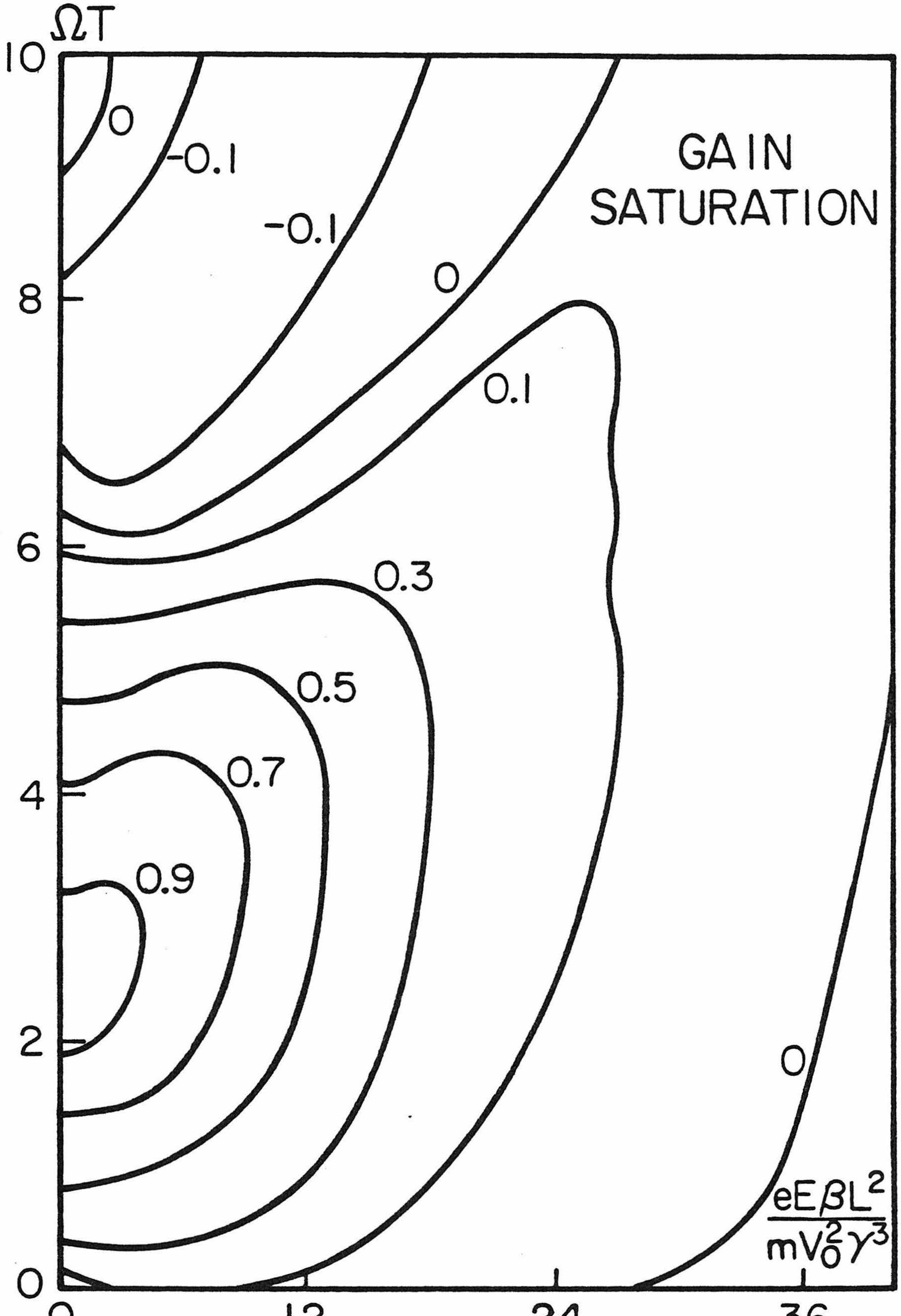

Figure 4-5: Gain diagram of a free-electron laser.

$$N(v) = (N_o/\beta)\left|\frac{d\phi_o}{dv}\right| \tag{4.77}$$

From the definition of $\phi$, we have

$$\dot{\phi} = \omega - \beta v \tag{4.78}$$

So,

$$v = v_p - \sqrt{\frac{4eE}{m\gamma^3 \beta R^2}}\,\{1 - R^2 \sin^2 \tfrac{\phi}{2}\}^{1/2}$$

$$= v_p - \sqrt{\frac{4eE}{m\gamma^3 \beta R^2}}\,\{1 - R^2 \mathrm{sn}^2(u|R^2)\}^{1/2} \tag{4.79}$$

Combining (4.77) and (4.80), we obtain the electron distribution in velocity at the output

$$N(v) = N_o\,\frac{m\gamma^3}{4\beta eE}\,\{\frac{d}{d\phi_o}\,[\sqrt{\frac{1}{R^2} - \mathrm{sn}^2(u(T)|R^2)}]\}^{-1} \tag{4.80}$$

The phase distribution is derived in the same manner

$$N(\phi) = N(\phi_o)\left|\frac{d\phi_o}{d\phi}\right| \tag{4.81}$$

$$= \frac{2N_o}{\beta}\left[\frac{(d/d\phi_o)[\mathrm{sn}(u|R^2)]}{\sqrt{1 - \mathrm{sn}^2(u|R^2)}}\right]^{-1} \tag{4.82}$$

Although the differentiations in (4.80) and (4.82) can be performed, the results have proven to be very complicated. The structures of N(v) and N($\phi$) strongly depend on the physical parameters, especially on E. For a monoenergetic beam, there are discontinuities in N(v) and N($\phi$) which jump from being finite to being infinite. The discontinuity should not disturb us because an energy distribution of a practical beam always has finite width. For a very narrow beam, such a discontinuity means only a sharp edge in the output distribution.

### 4.4 Large Gain Approximation

In an oscillating laser the constant field approximation makes physical sense. The intensity gain per single pass must be equal in steady state to the transmittivity of the mirrors. The transmittivity is usually very small in order to reduce the oscillation pumping threshold. The field amplitude is determined on the gain map (Fig. 4.5) by the given T and gain values. In the case of a free electron laser amplifier, the gain is required to be as large as possible. For example, in the signal amplification, the output field could be many times larger than the input field. Obviously, the constant field approximation is no longer valid. In solving the force equation, we have to consider E as a function of time or position, too.

In the following, we will consider the large gain amplification in the small signal region by solving the force equation directly. We then proceed with the analysis of the large signal regime by using the harmonic expansion.

In the large gain and small signal region, the field amplitude at the input is so small that the power expansion of physical variables in terms of $E_0$ is still applicable. But the variation of amplitude is so large that we must keep it as a function of time, E(t). In general, E should be a function of position. Because the position perturbation of an electron by the field is very small, it is more convenient to write E as a function of time, i.e., E[z(t)]. We have for the force equation

$$\frac{d^2z}{dt^2} = \frac{eE(t)}{m\gamma^3} \cos(\omega t - \beta z + \phi) \qquad (4.83)$$

Once the position is solved, the energy transfer of an electron is calculated as

$$\frac{d\varepsilon}{dt} = ev_o E(t) \cos(\omega t - \beta z + \phi) \qquad (4.84)$$

Because E(t) is the radiation field amplitude experienced by an electron at z(t), it is independent of the phases of the electrons. The averaged energy transfer is

$$\frac{d<\varepsilon>}{dt} = -e\omega\ E(t) <\Delta z \sin(\Omega t - \phi)>_\phi \qquad (4.85)$$

Due to the conservation of energy, the total energy of electrons and radiation is a constant. So

$$\frac{d<\varepsilon>}{dt} = - \frac{|e|}{I} \cdot \frac{dP}{dt} \qquad (4.86)$$

$$= - \frac{|e|}{I} \frac{1}{2\beta^2 k_o J_1^2(s)} \frac{d}{dt} (E^2(t)) \qquad (4.87)$$

where all physical parameters have been defined in Section 3.2. One must be careful when the $\phi$ average and time derivative are exchanged on the left side of (4.85). The exchange is applicable only when the time variation of $\varepsilon$ is negligible within a distance of one wavelength. As we know, physical quantities, such as z, v, E and $\varepsilon$ are slowly varying. So the time derivatives in (4.83) and (4.85) have the same range of applicability.

From (4.85), (4.86), and (4.87), we obtain the equation

$$\frac{dE}{dt} = -\beta^2 \omega k_0 J_1^2(s)\ I\ \langle\Delta z \sin(\Omega t - \phi)\rangle_\phi \tag{4.88}$$

Now z can be expanded to first order in $E_0$,

$$z = v_0 t + E_0\{f(\Omega t)\cos\phi + g(\Omega t)\sin\phi\} \tag{4.89}$$

where $f(\Omega t)$ and $g(\Omega t)$ are functions of $\Omega t$ and to be determined later. Subsituting $\Delta z$ into (4.83) and (4.88), we obtain three coupled equations

$$\begin{aligned} f'' &= a\tilde{E}\cos\Omega t \\ g'' &= a\tilde{E}\sin\Omega t \\ \tilde{E}' &= b(g\cos\Omega t - f\sin\Omega t) \end{aligned} \tag{4.90}$$

where

$$a = e/m\gamma^3\Omega^2$$

$$b = \beta^2\omega k_0 J_1^2(s) I / 2$$

and

$\tilde{E} = E(t) / E_0$ is the normalized field amplitude.

Defining a new variable $x = \Omega t$, and letting $F = f/a$, $G = g/a$, $c = 2ba/\Omega$, equations (4.90) become

$$\begin{aligned} F(x) &= \int_0^x dx' \int_0^{x'} dx''\ \tilde{E}(x'')\cos x'' \\ G(x) &= \int_0^x dx' \int_0^{x'} dx''\ \tilde{E}(x'')\sin x'' \\ \tilde{E}(x) &= 1 + \frac{c}{2}\int_0^x [G(x')\cos x' - F(x')\sin x']\,dx' \end{aligned} \tag{4.91}$$

Combining the three equations in (4.91), we have

$$\tilde{E}(x) = 1 + \frac{c}{2} \int_0^x dx' \int_0^{x'} dx'' \int_0^{x''} dx''' \; E(x''') \sin(x''' - x') \qquad (4.92)$$

The principle of causality, x > x' > x" > x"', is shown explicitly in the upper limits of the integrals. The order of integration can be reversed by proper change of the upper and lower limits of each integration. The result is

$$\tilde{E}(x) = 1 + \frac{c}{2} \int_0^x dx''' \int_{x'''}^{x} dx'' \int_{x''}^{x} dx' \; \tilde{E}(x''') \sin(x''' - x') \qquad (4.93)$$

The integrations over x' and x" can be executed easily

$$\tilde{E}(x) = 1 + \frac{c}{2} \int_0^x dx''' \{\sin(x''' - x) - (x''' - x) \cos(x''' - x)\} \; \tilde{E}(x''') \qquad (4.94)$$

Equation (4.94) is an integral equation for E(x), i.e.,

$$\tilde{E}(x) = 1 + \frac{c}{2} \int_0^x M(x - y) \; E(y) \; dy \qquad (4.95)$$

where $M(y) = (\sin y - y \cos y)$ is the kernel of the integral equation. The type of integral equation (4.95) is best solved by the method of Laplace transform. By taking the Laplace transform on both sides, we have

$$\tilde{E}(s) = \frac{1}{s} + \frac{c}{2} M(s) \; \tilde{E}(s) \qquad (4.96)$$

where

$$M(s) = 2 / (s^2 + 1)^2 \qquad (4.97)$$

The Laplace transform of the normalized field amplitude is found to be

$$\tilde{E}(s) = \frac{1}{s}\,[1 + \frac{c}{(s^2+1)^2 - c}] \qquad (4.98)$$

The inverse Laplace transformation yields the field amplitude as a function of t

$$\tilde{E}(t) = \frac{1}{1-c} + \frac{\sqrt{c}}{2}\,\{\frac{\cos\sqrt{1+\sqrt{c}}\,\Omega t}{1+\sqrt{c}} - \frac{\cos\sqrt{1-\sqrt{c}}\,\Omega t}{1-\sqrt{c}}\} \qquad (4.99)$$

This is an exact solution in the small signal region. The constant c is a measure of the interaction strength. In the small gain region, i.e., $c \ll 1$, we find the gain as

$$G = c[2 - 2\cos\Omega t - \Omega t \sin\Omega t] \qquad (4.100)$$

The result in the small gain region shows that the field amplitude oscillates along the interaction region with its maximum increasing. However, the exact solution shows the correct behavior of the field. Due to the saturation, when $c < 1$, the field is always bound between two extreme values

$$\frac{1}{1-\sqrt{c}} > \tilde{E} > \frac{1}{1+\sqrt{c}} \qquad (4.101)$$

Actually, the field oscillates sinusoidally with two different periods. When c is very small, the interference between those two terms generates beats. The field changes periodically as $\cos\Omega t$ and $\sin\Omega t$, with amplitude envelope $\cos\sqrt{c}\,\Omega t/2$ and $\sin\sqrt{c}\,\Omega t/2$.

When $c > 1$, the expression (4.99) should be written in terms of a hyperbolic cosine

$$\tilde{E}(t) = \frac{1}{1-c} + \frac{\sqrt{c}}{2}\,\{\frac{\cos\ 1+\sqrt{c}\,\Omega t}{1+\sqrt{c}} + \frac{\cosh\sqrt{\sqrt{c}-1}\,\Omega t}{\sqrt{c}-1}\} \qquad (4.102)$$

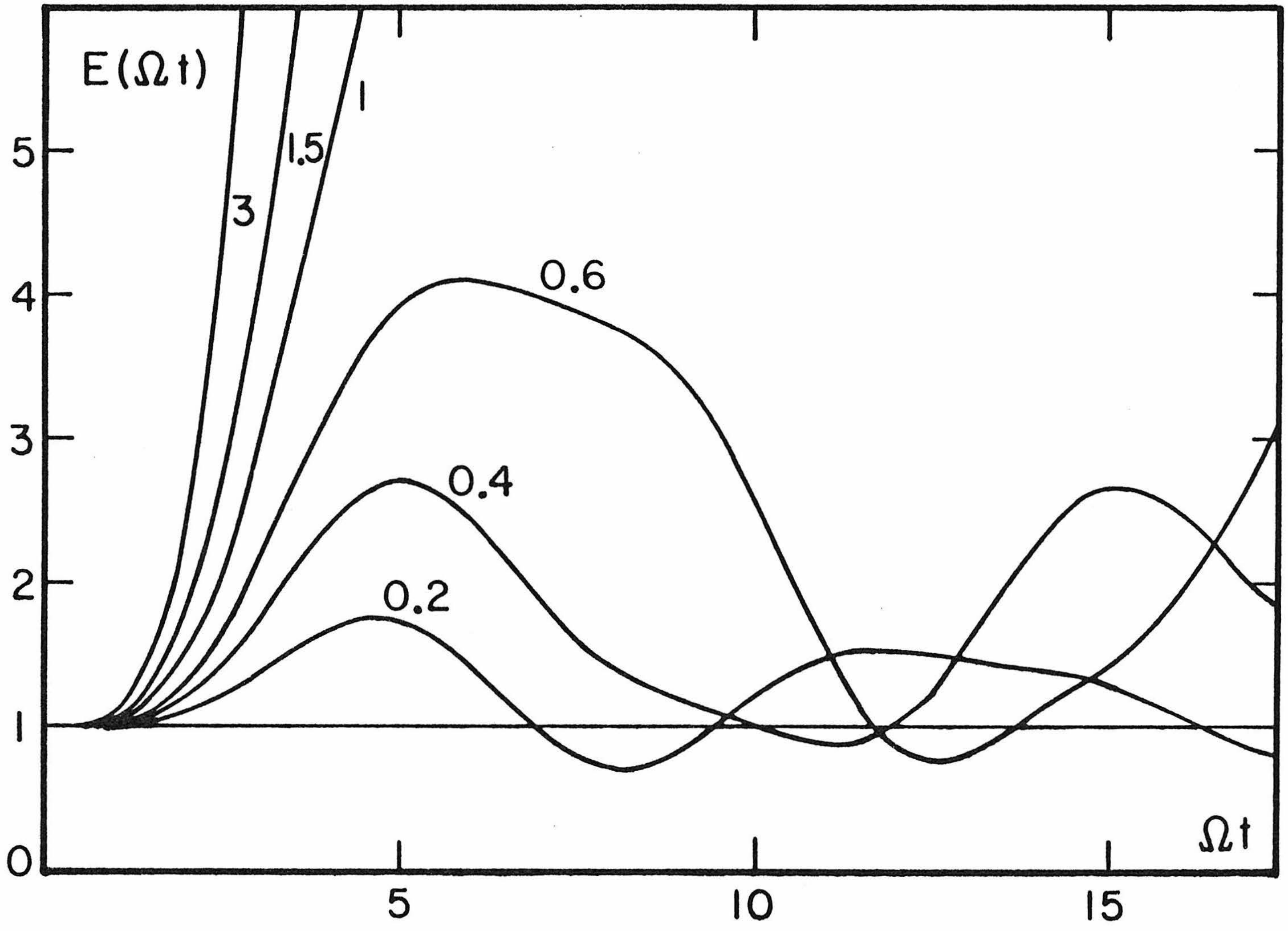

Figure 4-6: The field growth along the interaction region. The curve with c=1 is the boundary between a finite growth and an exponential growth.

and the field is seen to be a combination of oscillating and exponential terms. However, we note that the exponential term is always dominant If c is close to 1, the exponential terms are large due to the small denominator. If c >> 1, the term is large due to its exponential behavior. As a matter of fact, the numerical plots for c > 1 show that it is hard to note the existence of the oscillating term. When c is very large and $\Omega t$ is not a very small value, we have

$$\tilde{E}(t) \longrightarrow \frac{1}{4} e^{c^{1/4}\Omega t} \qquad (4.103)$$

The field amplitude is growing exponentially with distance, so the exact solution in (4.99) or (4.102) explains the difference between oscillating gain [4] and exponential gain [1]. It also defines quantitatively the range of applicability of two different gains. In Figure 4.6, we plot the behavior of E(t) for different values of c. The regions of bound gain and unlimited gain are divided by the curve with c equal to 1

$$E(t) = \frac{1}{4} (3 + \cos\sqrt{2}\, \Omega t + \Omega^2 t^2) \qquad (4.104)$$

If the field amplitude is not small enough, the power expansion of physical variables, like (4.89), is not appropriate due to its divergence. Another quantity in the force equation which could be used as an expansion parameter is the entry phase, $\phi$. It must follow that every physical variable of the electron is a periodic function of $\phi$. Therefore, the method of harmonic expansion in $\phi$ is natural for the analysis. For example, the electron position z can be put in a form

$$z = v_0 t + \Delta z$$
$$= v_0 t + \sum_{n=0}^{\infty} [f_n(t) \cos n\phi + g_n(t) \sin n\phi] \qquad (4.105)$$

where $f_n(t)$ and $g_n(t)$ are the coefficients in the cosine and sine expansions. $f_m$ represents the symmetric position modulation and $g_n$ the asymmetric position modulation. Both coefficients contain the saturation information. If they can be solved exactly, then all the physical processes can be described completely. However, a complete set of differential equations cannot be obtained due to mathematical complications. For simplicity, we truncate the harmonic expansion in (4.100) and leave only terms up to n = 1. The correction from higher harmonic terms has been evaluated by using the computer solution of (4.100) up to n = 2. It is found that for most situations the discrepancy is less than 15%. So the harmonic expansion up to n = 1 is applied to demonstrate, at least qualitatively, the field evolution and electron dynamics. Defining the phase $\phi$ seen by an electron at time t as

$$\phi = \phi_0 + p(t) + f(t) \cos \phi_0 + g(t) \sin \phi_0 \qquad (4.106)$$

the force equation becomes

$$\frac{d^2\phi}{dt^2} = - \frac{eE\beta}{m\gamma^3} \cos \phi$$

$$p'' + f'' \cos \phi_0 + g'' \sin \phi_0$$
$$= - \frac{eE\beta}{m\gamma^3} \cos[\phi_0 + p + f \cos \phi_0 + g \sin \phi_0]$$
$$= - \frac{eE\beta}{m\gamma^3} \cos[\phi_0 + p + h \sin(\phi_0 + \alpha)] \qquad (4.107)$$

where $h^2 = f^2 + g^2$

$\alpha = \tan^{-1}(f/g)$

Equation (4.107) should be true for any value of $\phi$. It can be decoupled into three differential equations for p, f, and g, respectively.

$$p'' = -\frac{eE\beta}{2\pi m\gamma^3}\int_{-\pi}^{\pi}\cos[\phi_0 + p + h\sin(\phi_0 + \alpha)]\,d\phi_0$$

$$f'' = -\frac{eE\beta}{\pi m\gamma^3}\int_{-\pi}^{\pi}\cos\phi_0\cos[\phi_0 + p + h\sin(\phi_0 + \alpha)]\,d\phi_0 \qquad (4.108)$$

$$g'' = -\frac{eE\beta}{\pi m\gamma^3}\int_{-\pi}^{\pi}\sin\phi_0\cos[\phi_0 + p + h\sin(\phi_0 + \alpha)]\,d\phi_0$$

The average energy transfer of an electron is given as

$$\frac{d\langle\varepsilon\rangle}{dt} = ev_0E\langle\cos\phi\rangle$$

$$= \frac{ev_0E}{2\pi}\int_{-\pi}^{\pi}\cos[\phi_0 + p + h\sin(\phi_0 + \alpha)]\,d\phi_0 \qquad (4.109)$$

Using the relation between the transferred energy and the field amplitude in (4.87) we have the differential equation for E

$$E' = \frac{\beta\omega k_0 J_1^2(s) I}{2\pi}\int_{-\pi}^{\pi}\cos[\phi_0 + p + h\sin(\phi_0 + \alpha)]\,d\phi_0 \qquad (4.110)$$

The integrations in (4.108)-(4.110) lead to Bessel functions. The results are

$$p'' = -\frac{eE\beta}{m\gamma^3}\cos(\alpha - p)\,J_1(h)$$

$$f'' = \frac{eE\beta}{m\gamma^3}\{\cos p\,J_0(h) + \cos(2\alpha - p)\,J_2(h)\} \qquad (4.111)$$

$$g'' = -\frac{eE\beta}{m\gamma^3}\{\sin p\, J_0(h) + \sin(2\alpha - p)\, J_2(h)\}$$

$$E' = -\beta\omega k_0 J_1^2(s) I \cos(\alpha - p)\, J_1(h)$$

Using the $\alpha$ given in (4.103), we have

$$p'' = -A\tilde{E}\, QJ_1(h)/h$$

$$f'' = 2A\tilde{E}\{\cos p\ J_1'(h) + gQJ_2(h)/h^2\} \tag{4.112}$$

$$g'' = -2A\tilde{E}\{\sin p\, J_1'(h) + fQ\, J_2(h)/h^2\}$$

$$\tilde{E}' = -BQ\, J_1(h)/h$$

where $A = eE_0\beta/m\gamma^3\Omega^2$

$$B = \beta\omega k_0 J_1^2(s) I/\Omega^2 E_0$$

$$Q(t) = g\cos p + f\sin p$$

and the differentiations are given with respect to a dimensionless quantity $\Omega t$. We found it impossible to solve for E from (4.112) due to complexity. Using a computer calculation, it is straightforward to obtain the behavior of p, f, g, and $\tilde{E}$ along the interaction region for a given set of conditions. The initial conditions for p, f, g, and $\tilde{E}$ are

$$p(0) = f(0) = g(0) = f'(0) = g'(0) = 0 \tag{4.113}$$

$$p'(0) = -1 \quad \text{and} \quad \tilde{E}(0) = 1$$

With these conditions, it is found that $\tilde{E} \sim t^4$ at the beginning of interaction, which result is identical to that obtained in the linear analysis. A typical behavior of the field amplitude is sketched in Figure 4.7. If we plot the amplitude gain ($E_{out}$ - $E_{in}$) as a function of input amplitude ($E_{in}$), we obtain the curve shown in Figure 4.8. The output is shown in Figure 4.7 to oscillate even when $E_{in} < E_s$. In an optical cavity the field amplitude increases every time it reenters the interaction region. Finally, the amplitude clamps at $E_s$ and will not change with further reamplification.

$E_s$ is the saturation field amplitude. Figure 4.8 is similar to the curve in Figure 4.6 if we plot the gain against E with constant $\Omega T$. It is obvious that the contour with gain equal to zero represents the saturation field for a given $\Omega T$ with the presence of electron circulation. However, the gain map should differ slightly from Figure 4.6 in the large gain approximation, although it is similar qualitatively.

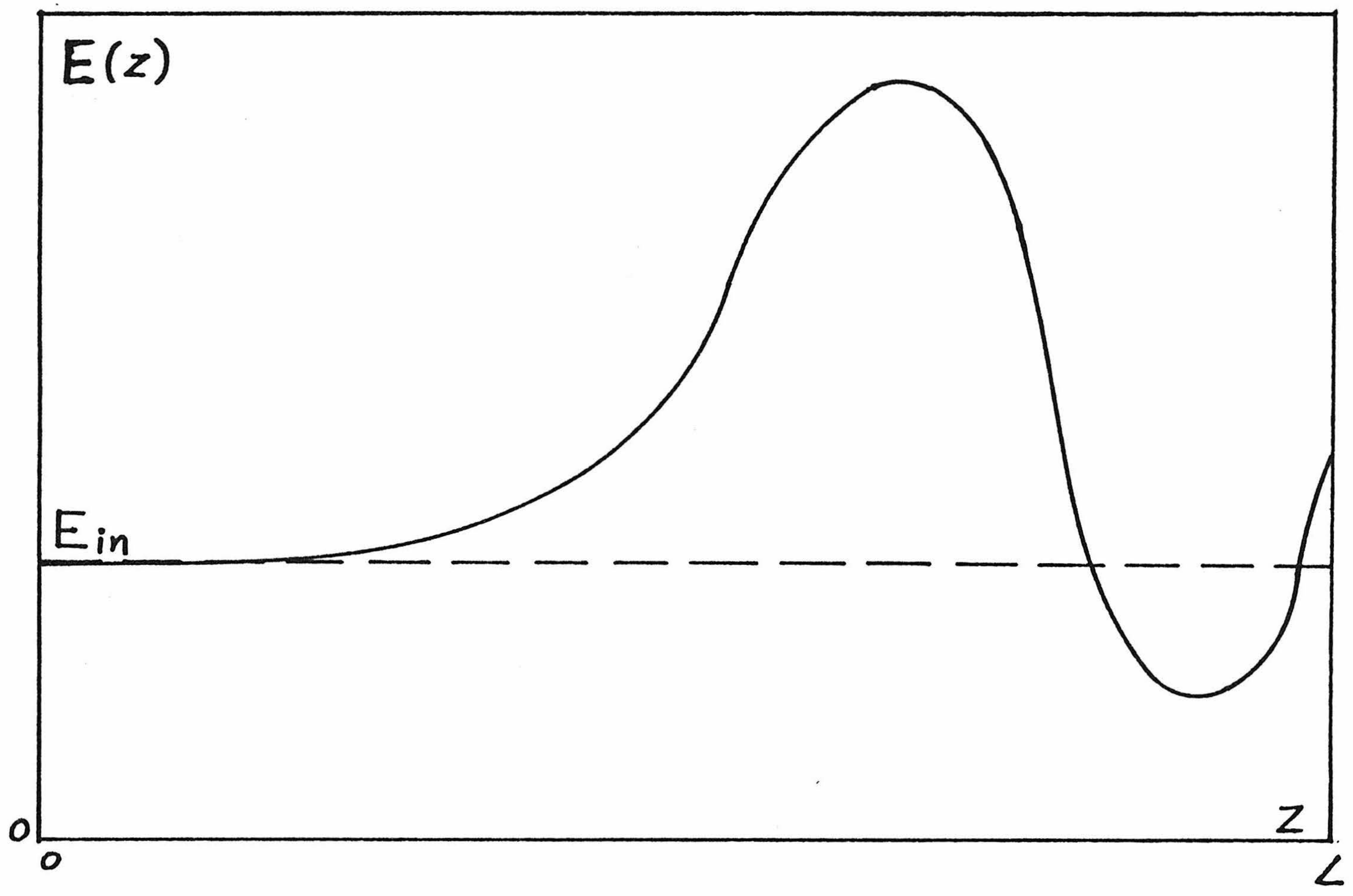

Figure 4-7: The field growth in the large signal region.

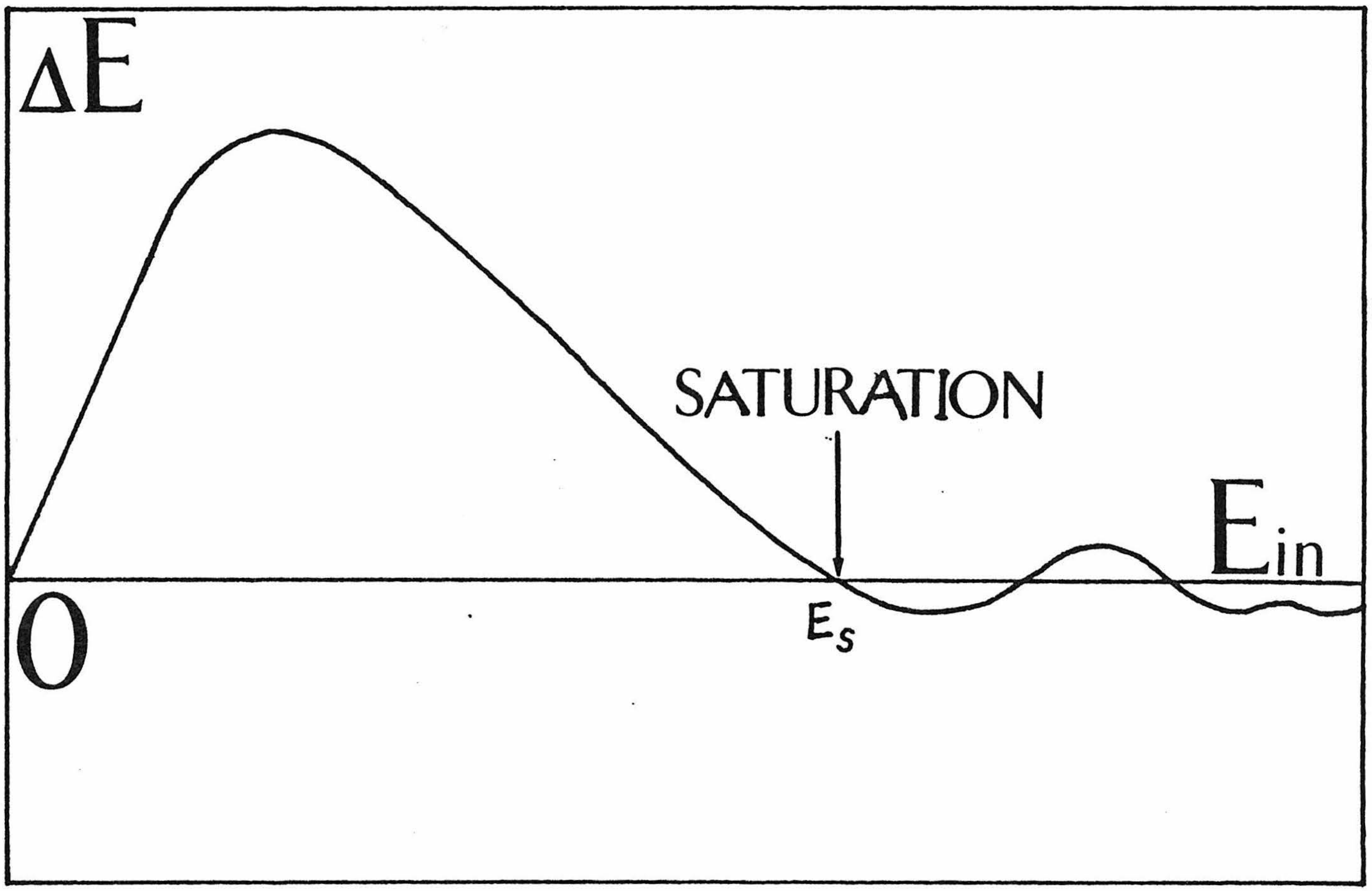

Figure 4-8: The amplitude increase versus the input amplitude to show the saturation field ($E_s$) for beam circulation device.

## CHAPTER 4 - REFERENCES

1. N. M. Kroll and W. A. McMullin, Phys. Rev. A 17, 300 (1978).
2. W. H. Louisell, J. Lam, and D. A. Copeland, Phys. Rev. A 18, 655 (1978).
3. M. Abramowitz and I. A. Stegun, eds., Handbook of Mathematical Functions (Dover, New York, 1970).

# CHAPTER 5
# CONCLUSION

The longitudinal free electron laser has been proposed to produce tunable, low-power coherent radiation efficiently. The device consists of a medium energy electron beam and a corrugated waveguide as the interaction region. We have studied the phenomena of electrodynamics in a waveguide. We described the wave spectrum, spontaneous emission, and discovered the electron band structure, as well as the quantum limitations. From the classical approach, we formulated the linear theory of the longitudinal free electron laser. Starting with the force equation, the homogeneous gain, inhomogeneous gain, and lossy gain were derived. Using a phase diagram, we studied the electron dynamics and the evolution of the electron distribution. We also analyzed a special device--a two-stage system. In order to understand the laser mechanism and phenomena in the nonlinear region, we used different approaches to investigate the effects of a dense beam, high radiation, and large gain. In the regime where space-charge effects are important, we obtained the dependence of the gain on the current density. The saturation current density was pointed out quantitatively. In the case of a large radiation field, we solved the force equation exactly by using special functions. A gain diagram was shown to demonstrate the saturation of the gain due to the high radiation intensity. In the large gain amplification, we found an integral equation describing the field growth in the low field limit and a set of differential equations in the high field limit. Combining the linear and nonlinear theory of the device, we have completely analyzed its laser mechanism and physical processes under all possible situations. However, we also considered the effects in special situations separately to show the influences of those given parameters. The complex situation, such as dense electron beam with high radiation field, can be understood qualitatively, while a quantitative analysis is not absolutely essential.

In a practical set-up of the longitudinal free electron laser, many design and engineering problems have to be considered thoroughly. Due to the need for a high resolution electron beam, the Van de Graaff accelerator and the currently developing Microtron could be the two best candidates for electron generators. Both these devices have electron beams of good resolution and energies in the desired range. The electron beam is then focused to within a diameter of about 30 μm or less by using several stages of quadrupole focusing elements. The fabrication of a 50 μm wide waveguide, 10 cm long, is a very delicate job. The straightness of the waveguide wall must be good within 0.01%. A possible way to obtain the waveguide is to use two separators of thickness 50 μm between two smoothly polished metal surfaces. The corrugation of period 200 μm can be done mechanically on the two separators before the assembly of the waveguide. The entire system should be in a vacuum. The details of the system design and the possible experiments are, however, beyond the scope of this thesis and will not be discussed here.

An important question is how to improve the beam extraction efficiency of a device. For the free electron laser, the solution is to recycle the electron beam. A circulation picture is shown in Figure 5.1. The energy loss of an electron in the interaction region can be restored

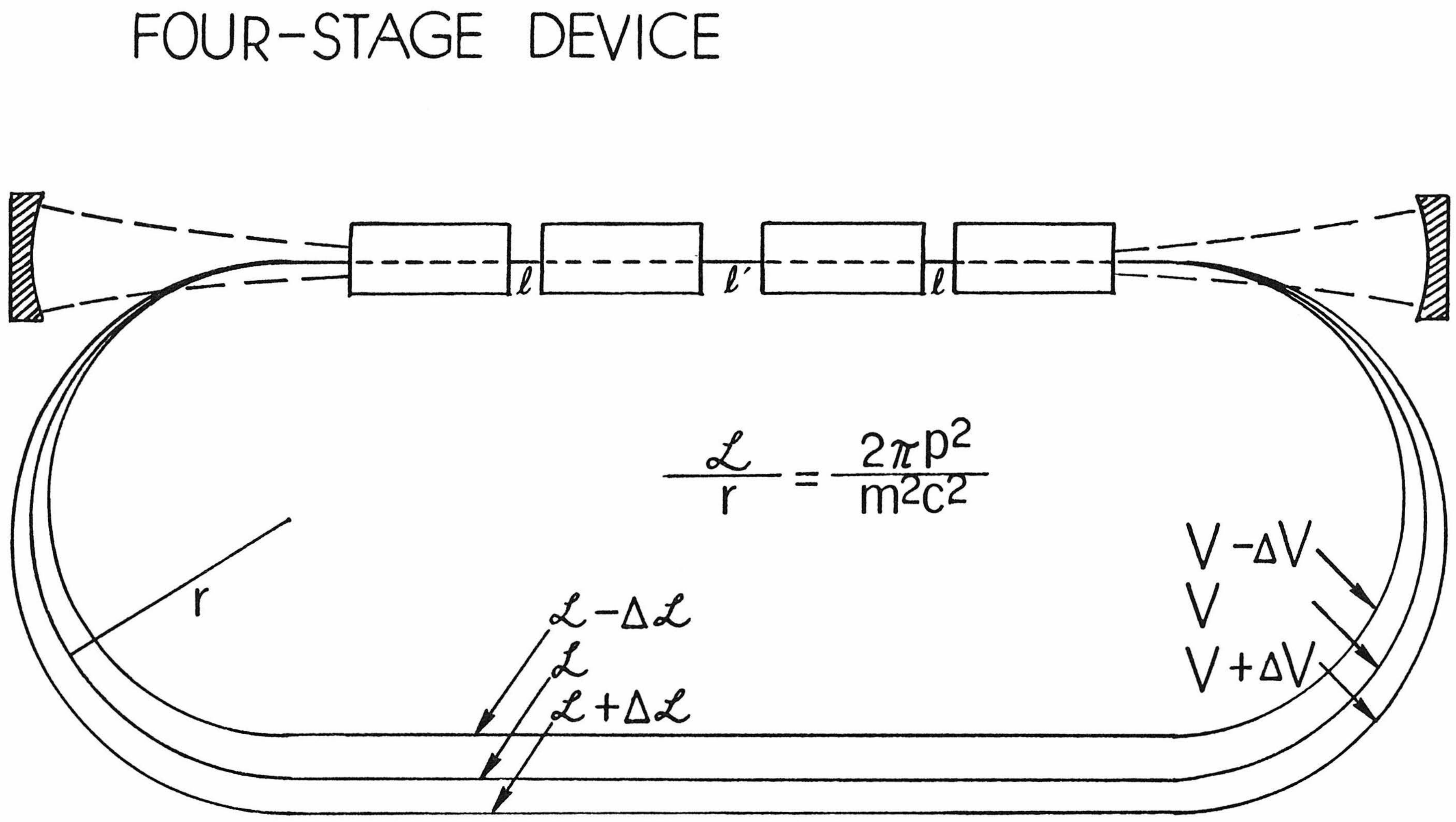
FOUR-STAGE DEVICE
$\ell$
$\ell'$
$\ell$
$\frac{\mathcal{L}}{r} = \frac{2\pi p^2}{m^2c^2}$
r
$\mathcal{L}-\Delta\mathcal{L}$
$\mathcal{L}$
$\mathcal{L}+\Delta\mathcal{L}$
V-ΔV
V
V+ΔV

Figure 5-1: Electron beam circulation. $L$ is the optimum length in space bunching.

by an energy supply device on the return path. The overall efficiency of the system depends on the number of cycles. The number of cycles is limited by the broadening of the electron distribution during the interaction with the radiation. In the case of a prebunched electron beam, it is required that the electron pulse will not spread on the return path due to different velocities. This problem can be solved by properly choosing the length of returning path, $L$. In Figure 5.1 we note that the outer track has a longer path, but also has larger electron velocity. Two factors compensate each other. It is expected that there is an optimum value of $L$ such that the first order effect due to the velocity difference can be eliminated. By using the relativistic formula, it is found that

$$L = r \frac{2\pi p^2}{m^2 c^2} = 2\pi r \, \beta^2 \gamma^2 \tag{5.1}$$

where r is the radius of magnetic bending

$$r = P/Be \tag{5.2}$$

With the designed length, the pulse width can be bunched by a factor of $10^3$ for a very short electron pulse. It is also pointed out that such space-bunching becomes unnecessary when the length of the electron pulse is larger than about 1 mm.

The bunching of the electron velocity distribution is essential in an attempt to increase the number of cycles which result in a significant radiation gain, thus to increase the overall efficiency. The ratio of the velocity shift to its spread is a measure of the effectiveness of the electron energy extraction. We have pointed out that, by the use of a two-element device, the single-pass gain or the device efficiency can be highly improved.

Another possible way to increase the single-pass gain of the device is to introduce a variably corrugated waveguide instead of a constant-period one. This is due to the down-shift of the electron velocity distribution. In order to maintain the system at its high gain condition, we can decrease the phase velocity of the first harmonic simultaneously which can be accomplished by a gradually decreasing corrugation period along the interaction region. The energy extraction from the electron beam will be kept at its optimal condition and the gain can be much higher than the value in the case of constant period. This variable waveguide problem is a challenge for future experiments.

# PART II

# A THEORETICAL MODEL OF THE LINEAR ELECTROOPTIC EFFECT

# CHAPTER 1
# GENERAL INTRODUCTION

## 1.1 Introduction

In the weak field approximation, the induced polarization in matter is linearly related to the applied electric field as

$$\vec{P} = \varepsilon_0 \overline{\overline{\chi}}:\vec{E} \tag{1.1}$$

where $\varepsilon_0$ is the electric permittivity in vacuum; $\chi$ is the susceptibility of the medium. P and E are vectors, so $\chi$ is a tensor of rank two. Due to the complex interaction between the electric field and matter, the susceptibility $\chi$ is in general field-dependent. Such a nonlinearity becomes obvious when the field amplitude is large enough. The polarization can be expressed as an expansion in terms of the total field in the medium

$$P = \varepsilon_0 \chi E + \varepsilon_0 r EE + \varepsilon_0 R EEE + \cdots \tag{1.2}$$

The coefficients, r and R, represent the relation between P and E in second and third orders. r is a tensor of rank three, and R is a tensor of rank four. If the medium has inversion symmetry, it follows that the second order coefficient is identically zero. The lowest nonlinear effect in such a system is in the third order. This happens in gas systems or centrosymmetric crystal, such as rock salt or cesium chloride.

In the absence of inversion symmetry, the nonlinear phenomenon is due to the self-interaction of the total field. In general, the total field can be written as

$$E_T = (E^{\omega}e^{i\omega t} + E^{\Omega}e^{i\Omega t} + (c.c.) \qquad (1.3)$$

where $E^{\omega}$ represents the amplitude of the electric field in the electromagnetic wave of frequency , $E^{\Omega}$ is the low frequency or static electric field amplitude, $\omega >>> \Omega$. The interactions between each term result in several physical phenomena:

(1) Second harmonic generation:

$$P^{2\omega} = r\, E^{\omega}E^{\omega}$$

(2) Parametric frequency mixing:

$$P^{\omega_1+\omega_2} = 2r\, E^{\omega_1}\, E^{\omega_2}$$

or

$$P^{\omega_1-\omega_2} = 2r\, E^{\omega_1}\, E^{\omega_2^*}$$

(3) Optical rectification:

$$P^{o} = r\, E^{\omega}\, E^{\omega*}$$

(4) Electrooptic effect:

$$P^{\omega+\Omega} = 2r\, E^{\omega}\, E^{\Omega}$$

These four nonlinear effects have been discussed in detail [1]. In this thesis we concentrate on the last effect. In terms of the low frequency field $E^{\Omega}$, (1.2) is rewritten as

$$\begin{aligned} P^{\omega} &= \varepsilon_o[\chi E^{\omega} + 2r\, E^{\omega}E^{\omega}] \\ &= \varepsilon_o[\chi + 2r\, E^{\Omega}]\, E^{\omega} \qquad (1.4) \end{aligned}$$

Physically, the applied static field $E^{\Omega}$ changes the optical susceptibility of the medium, and consequently, changes the index of refraction. The large interest in this physical property is due in part to its wide application in light modulators and numerous laser devices.

In many applications it is necessary to modulate the amplitude, phase, frequency, or direction of a laser beam at high speed. The electrooptic effect is found to be an excellent method for performing these tasks due to its fast response and accurate control. The modulating low frequency signal $E^{\Omega}\ e^{i\Omega t}$ is applied to the crystal through which the laser beam passes. With a choice of specific orientations of the crystal, we can modulate the amplitude, phase, frequency, or direction of the incident beam. The techniques of modulation are beyond the scope of this dissertation. An excellent review article discussing these techniques in detail can be found in Ref. 2. Common to all the applications listed above is the need for crystals with high electrooptic constant r such that the modulation power can be lowered. Many efforts have been made in this direction. These, however, were based on trial and error or by empirical means. It is concluded [2] that "Perhaps the development of a theoretical understanding of the electrooptic effect will lead to the discovery of synthesis of the ideal substance for each application in a logical way... ." This is the motivation behind the investigation reported in the second half of this thesis.

## 1.2 Previous Work on the Electrooptic Effect

The electrooptic effect was discovered before the beginning of the twentieth century. Kerr observed the quadratic electrooptic effect in liquids such as carbon disulphide. The linear electrooptic effect was first found in quartz by Rontgen and Kundt. A systematic examination of the linear effect has been performed in the crystals, quartz, tourmaline, potassium chlorate, and Rochelle salt by Pockels [3]. The linear electrooptic effect is also called the Pockels effect. He showed the dependence of the linear electrooptic effect on the point symmetry of the crystal.

Thirty years after Pockels' demonstration, Zwicker and Scherrer reported the electrooptic properties of KDP ($KH_2PO_4$) and KDDP ($KD_2PO_4$) and noted the relation to their ferroelectric behavior [4]. As KDP and KDDP are similar in structure, they found the electrooptic response is proportional to the dielectric constant. Although it is now known that it is not exactly proportional to the dielectric constant, but qualitative, it contributed greatly to the understanding of the effect. The first application of the electrooptic effect was the construction of a high speed light shutter using KDP and ADP ($NH_4H_2PO_4$) in sound recording by Billings and Carpenter [5,6].

The advent of the laser opened the era of optical communication. The guiding and switching of light beams can be achieved by electrooptic materials. In the course of these research projects, the electrooptic coefficients of many crystals with various point group symmetries have been measured [7]. However, the theoretical understanding of the problem did not begin until the late 1960's.

Kaminow derived a simple relation between an electrooptic coefficient measured at radio frequencies and the corresponding Raman-scattering efficiencies [8]. Using the measured efficiencies, he calculated the electrooptic coefficients of $LiNbO_3$ and $LiTaO_3$, which are in good agreement with experiment [9]. With the macroscopic equations between polarization, ion displacement, and electric field, Kelly found an expression for the coefficient of zinc-blende crystals [10]. The result is applied to CuCl and ZnS with satisfactory agreement with measurements. The electrostatic point-charge model and dielectric theory have been used to determine the electrooptic coefficients of III-V compounds [11] and II-VI crystals [12] by Flytzanis. The theoretical treatments are so far limited to diatomic crystals. The measurement of the electrooptic coefficient reveals that the high response materials are actually complex crystals with more than two kinds of atoms. Therefore, a generalized theory to describe the electrooptic behavior of simple crystals as well as complex systems is badly needed.

A complete review of the formal theory of the nonlinear optical effect can be found in Ref. 13.

## 1.3 Outline of Thesis, Part II

In Chapter 2, the theory of the linear electrooptic effect is formulated. Based on the quantum mechanical approach of the one-gap model, the susceptibility of a diatomic crystal is found to depend on the energy gap. The energy gap is then phenomenologically interpreted as the combination of symmetric and asymmetric parts. From the microscopic point of view, these two parts are the result of the motion of bond-charge in the bond region. Its harmonic and anharmonic motions respond to the total field in the crystal inducing the linear and nonlinear susceptibilities. In the low frequency region, the ions are displaced from their normal sites due to the applied electric field. From the derived expression for the dependence of the optical susceptibility on bond-rotation and bond-stretch, we derive an expression for the electrooptic coefficient of crystals. Presumably, the result is applicable to any material with arbitrary complex structure.

In Chapter 3, the theory is applied to the calculation of the coefficient of various crystals. They include zinc-blende and wurtzite crystals, quartz, $LiNbO_3$ and $LiTaO_3$, KDP family, chalcopyrite compounds, and $Ag_3AsS_3$. The characteristics of these crystals are listed in the following:

(1) Zinc-blende and wurtzite: AB type, single bond, tetragonal coordination, point group $\bar{4}3m$ and 6 mm;

(2) $SiO_2$: $AB_2$ type, single bond, distorted tetragonal coordination, point group 32;

(3) $LiNbO_3$ and $LiTaO_3$: $ABC_3$ type, one kind of bonds in two different bond-lengths, distorted octahedral coordination, point group 3m;

(4) KDP family: Single bond, tetragonal coordination, point group $\bar{4}2m$;

(5) Chalcopyrite: $ABC_2$ type, slightly distorted tetragonal coordination, two different bonds, point group $\bar{4}2m$;

(6) Proustite: $A_3BC_3$ type, complex coordination, two different

bonds, point group 3m.

In Chapter 4, the theory and calculation are summarized. A possible direction for seeking out better electrooptic materials is pointed out. The limitation and the possible development of the theory are discussed.

## CHAPTER 1 - REFERENCES

1. A. Yariv, Quantum Electronics, 2nd ed. (Wiley, New York, 1975).
2. I. P. Kaminow and E. H. Turner, Proc. IEEE 54, 1374 (1966).
3. F. Pockels, Lehrbuch der Kristalloptik (Teubner, Leipzig, 1906).
4. F. Jona and G. Shirane, Ferroelectric Crystals (Macmillan, New York, 1962).
5. B. H. Billings, J. Opt. Soc. Am. 39, 797, 802 (1949); 42, 12 (1952).
6. R. O'B. Carpenter, J. Opt. Soc. Am. 40, 225 (1950).
7. I. P. Kaminow and E. H. Turner, Handbook of Lasers (Chemical Rubber Co., Cleveland, Ohio, 1971)., pp. 447-459.
8. I. P. Kaminow, "Ferroelectricity: Proceedings of the Symposium on Ferroelectricity, General Motors Research Laboratories, Warren, Mich., 1966," E. F. Weller, ed. (Elsevier, New York, 1967).
9. I. P. Kaminow and W. D. Johnston, Jr., Phys. Rev. 160, 519 (1967).
10. R. L. Kelly, Phys. Rev. 151, 721 (1966).
11. C. Flytzanis, Phys. Rev. Lett. 23, 1336 (1969).
12. C. Flytzanis, Phys. Lett. 34A, 99 (1971).
13. C. Flytzanis, Quantum Electronics, Vol. 1, Part A, H. Rabin and C. L. Tang, eds. (Academic Press, New York, 1975), p. 9.

# CHAPTER 2

# THEORY OF THE LINEAR ELECTROOPTIC EFFECT

## 2.1 Introduction

In this chapter we introduce the semiclassical model for the linear electrooptic effect. The starting pointsfor our analysis are the dielectric description of diatomic crystals in the Phillips-Van Vechten (PV) theory [1-5] and the bond-charge calculation of the bond-nonlinearity in Levine's theory [6-8]. With the concept of the effective ionic charge and the isotropic displacement, a general expression for the electrooptic coefficient is formulated. Due to the large uncertainty in the measurement of coefficients, a large enough margin is allowed for the accuracy of the theory in order to achieve its generality. However, the uncertainty in the theoretical estimate is usually comparable to or less than that in the experiment.

To meet the purpose of predicting properties of new materials, the theory is required to employ as few physical parameters as possible. In order to serve as a guide to the crystal grower, it is important that the parameters entering the theory can be measured on small crystals or powder. This will obviate the need for expensive and lengthy growth of crystals often to find out that the coefficients are disappointingly small. The dependence of the electrooptic coefficient on the parameters is then studied. The calculated results are compared with experiment in the following chapter.

## 2.2 The Dielectric Theory of Optical Susceptibility of Crystals

The interaction between an external electric field and the electrons in a solid is responsible for its index of refraction. Using the nearly free electron model in the semiconductor, it is found that the dielectric constant at long wavelengths can be calculated from [9]

$$\varepsilon(\infty) = 1 + \left(\frac{\hbar\omega_p}{Eg}\right)^2 \left[1 - \frac{Eg}{4E_F} + \frac{1}{3}\left(\frac{Eg}{4E_F}\right)^2\right] \tag{2.1}$$

where $\omega_p^2 \equiv 4\pi Ne^2/m$ is the plasma frequency. $E_F$ is the Fermi energy level and Eg is the average energy gap of the semiconductor. $Eg/4E_F$ is usually about 0.1. So the susceptibility in this model is expressed in a very simple way

$$\chi(\infty) \simeq (\hbar\omega_p/Eg)^2 \tag{2.2}$$

However, we still have an unknown parameter Eg which varies for different materials.

In the simplest situation with diatomic crystals, the potential acting on the electron can be Fourier transformed into symmetric and antisymmetric parts with the origin chosen at the middle of the bond between the two atoms. If the potential is given as

$$V(x) = V_1(|x_1-x|) + V_2(|x_2-x|)$$

$$= V_1(|x-r_o|) + V_2(|x+r_o|) \tag{2.3}$$

where $x_1$ and $x_2$ are the positions of atom 1 and 2, x is the position of the electron, and $2r_o$ is the bond length. The Fourier transform of V(x) is

$$\begin{aligned} V(x) &= \sum_G V_G\, e^{iGx} \\ V_G &= \cos Gr_o\, V_G^s + i \sin Gr_o V_G^a \\ V_G^s &= \frac{1}{2}\,(V_{G1} + V_{G2}) \\ V_G^a &= \frac{1}{2}\,(V_{G1} - V_{G2}) \end{aligned} \qquad (2.4)$$

and
$$V_{G1,2} = \frac{2}{\ell} \int V_{1,2}(x)\, e^{-iGx}\, dx$$

where $\ell$ is the length of a unit cell. This description is for the one-dimensional situation. In the actual three-dimensional crystal, there are many $V_G$'s due to the coupling of different electron states in many directions. If we consider only the average gap between the valence and conduction bands, the average result of all $V_G$ is represented by one complex expression

$$V = E_h + ic \qquad (2.5)$$

The effective band gap Eg is given by

$$Eg^2 = V \cdot V^* = E_h^2 + c^2 \qquad (2.6)$$

It is clear that $E_h$ is the symmetric part of the energy gap relating to the covalent bonding, while c is the asymmetric part relating to the ionic bonding. In diamond-like crystals, the asymmetric part vanishes and Eg is identically equal to $E_h$. Now the unknown Eg is divided into two quantities, $E_h$ and c, which are to be determined empirically. The behavior of $E_h$ and c are best observed from the expression

$$Eg^2 \equiv E_h^2 + c^2 \simeq (\hbar\omega_p)^2/\chi \qquad (2.7)$$

By plotting $\chi^{-1}$ versus the square of the valence difference z for crystals composed of atoms in the same row, it is found [4] that

$$\chi^{-1} = a_n + b_n(\Delta z)^2 \qquad (2.8)$$

Both constants, $a_n$ and $b_n$, decrease with increasing row number n. In the same row, the lattice length is almost constant, so $\omega_p$ and $E_f$ are also constant. Comparing (2.8) with (2.7), it is obvious that $E_h$ depends on the row number only, and c is approximately proportional to $\Delta z$ with the proportionality constant depending only on n. Since the lattice constant increases as the row number, it is reasonable to assume a relation for $E_h$ as

$$E_h = A\ d_o^{-s} \qquad (2.9)$$

where $d_o$ is the nearest neighbor, and s is a constant to be determined. Using the values $\varepsilon(\infty) = 5.7$ and 12.0 for diamond and silicon, one obtains the indicial value $s = 2.48$ [4] and the proportionality constant (A) 39.74 if $E_h$ is in eV and $d_o$ is in Å.

Because c has the dimension of energy, it may have the form

$$c_{\alpha\beta} \propto (z_\alpha/r_\alpha - z_\beta/r_\beta) \qquad (2.10)$$

between atoms $\alpha$ and $\beta$ belonging to the same row. Physically, c is due to the difference in the Coulomb potentials. However, the electron should be somehow screened by core electrons. It follows from the Thomas-Fermi theory [10] that the screening wave number $k_s$ is

$$k_s^2 = \frac{4}{\pi a_o}(3\pi^2 N)^{1/3} \qquad (2.11)$$

where $a_0$ is the Bohr radius, N is the electron density corresponding to eight electrons per diatomic volume. The complete c is found to be

$$c = be^2(\frac{z_\alpha}{r_\alpha} - \frac{z_\beta}{r_\beta})\ e^{-k_s R} \tag{2.12}$$

R is half the interatomic distance, b is a dimensionless constant depending on row number. The variable c in (2.12) is not exactly the difference between two screened Coulomb potentials because of the common factor $e^{-k_s R}$. We can find Eg from the measurement of $\chi$. Knowing $E_h$, we can obtain c from (2.6). The value of b calculated in this way is found to be between 1 and 2. Actually, most crystals have the values of b between 1.4 and 1.6. The assignment of b = 1.5 has an uncertainty of 26%. The above analysis for the $E_h$ and c is only for nontransition atoms. The effect of d-electrons is considered in the expression of $\chi$ as a constant D [4]

$$\chi = (h\omega_p)^2\ DA/Eg^2 \tag{2.13}$$

where

$$A = 1 - \frac{Eg}{4E_F} + \frac{1}{3}\left(\frac{Eg}{4E_F}\right)^2$$

D is found empirically to be

$$D = \Delta\alpha\Delta\beta - (\delta\alpha\delta\beta - 1)(z_\alpha - z_\beta)^2 \tag{2.14}$$

where the constants $\Delta$ = 1.0, 1.0, 1.12, 1,21, 1.31, and $\delta$ = 1.0, 1.0, 1.0025, 1.005, 1.0075 for the atoms from row 1 to row 5 [4,8].

As c depends on $\Delta z$, it is argued [7] that a better result for b can be obtained by letting $r_\alpha \simeq r_\beta \simeq r_0$, where $r_0$ is half the bond length. With this argument, b is found to be 1.62 ± 14% for the zincblende and wurtzite crystals. Considering different structures, Levine found that b follows a simple relation with the coordination number $N_c$

$$b = (0.089 \pm 10\%)\ N_c^2 \tag{2.15}$$

It must be mentioned that the evaluation of the susceptibility in the PV theory requires only the knowledge of the atomic radius and the structure which determines b. The model is applied to all diatomic crystals. In this theory the covalency and ionicity parameters of the bond are defined as

$$\begin{aligned} f_c &= E_h^2/Eg^2 \\ f_i &= c^2/Eg^2 \end{aligned} \tag{2.16}$$

A complete list of values of c, $E_h$, $f_c$, and $f_i$ for 68 diatomic crystals is in Ref. 4.

## 2.3 Theory of the Electrooptic Effect

In the previous section we have reviewed briefly the PV dielectric theory for diatomic crystals. The reason for the appeal of the theory is that it depends on only two physical quantities: atomic radius and coordination number. We have avoided the complex calculation of the electron wave function by replacing the versatile band structure with a simple energy gap. The simplicity of this semi-empirical result makes it a suitable starting point for the theory of the nonlinear optical effect. However, the division of the potential into symmetric and antisymmetric parts applies only to diatomic crystals. Any extension of the

PV theory should be limited to diatomic crystals. In order to extend our theory to more complex materials, we look into the unit cell from the microscopic point of view. It is assumed that the bulk susceptibility is due to the geometrical composition of the susceptibilities of individual bonds. Or equivalently, that the total dipole moment can be taken as the vector sum over a unit cell of individual dipoles, each associated with a single bond,

$$\chi = \frac{1}{v} \sum_{n} \beta_n \alpha_{ni} \alpha_{nj} \tag{2.17}$$

where $\alpha_{ni}$ is the direction cosine of the bond n in the ith direction, $\beta_n$ is the bond susceptibility along the bond direction, v is the volume of a unit cell, n indicates the individual bond, and the summation runs through the bonds in one unit cell. In the formulation of (2.17), we assume that the susceptibility is isotropically along the bond direction and we neglect the transverse contribution. Usually, the transverse susceptibility is much smaller than the longitudinal one, since it involves promotion of the electron into antibonding orbitals with high energies, so the expression (2.17) is usually a good approximation. For diatomic crystals, there is only one kind of bond in several different directions.

$$\chi = \frac{\beta_n}{v} \sum \alpha_{ni} \alpha_{nj} \tag{2.18}$$

It can be seen that the bond susceptibility is proportional to the bulk susceptibility in crystal of the same structure. We assume that relation (2.2) gives the bond susceptibility except for a proportionality constant depending on the crystal structure. The quantities, $E_h$ and c, are fundamental properties of the bond. With this concept, we can calculate the linear and nonlinear susceptibility in an extended PV theory.

The physical interpretation of $E_h$ and c is based on the bond-charge model [8]. It is thought that the overlap of the electron distribution in two adjacent atoms generates a bond charge in the bond region. It represents only a small amount of charge, but it has high mobility. Its harmonic and anharmonic motion responding to the applied total field is the source of linear and nonlinear dipole moment. The concept of bond charge has been established theoretically and experimentally. The numerical calculation of the electron density in simple crystals reveals the existence of the bond charge. The quantitative estimate of the bond charge is found empirically to be

$$q = en_v(\frac{1}{\varepsilon} + kf_c) \tag{2.19}$$

where $n_v$ is the number of electrons per formula unit divided by the number of bonds. k is a constant found to be 1/3. The relation has been tested for single-atom and diatomic crystals [8] and found to be within an uncertainty of 18%. Since q enters linearly into the calculation of the nonlinear susceptibility, (2.19) is good enough to evaluate the magnitude of the bond charge.

Although the calculation of $E_h$ and c is insensitive to the individual atomic radius and depends on the bond length only, the anharmonic motion of the bond charge should be sensitive to the position of the bond charge. Therefore, the utilization of $E_h$ and c needs a revised interpretation.

The expression for c given in (2.12) should be good for the nonlinear effect because it shows the dependence on $r_\alpha$ and $r_\beta$ . The calculation of $\chi$ is obtained by taking $r_\alpha \simeq r_\beta \simeq r_o$, but such an approximation should be made only after the derivation of the nonlinear effect.

The homopolar energy gap $E_h$ given in (2.9) is only a function of $(r_\alpha + r_\beta)$. This is approximately true for the harmonic motion of the bond charge, but it may not be correct for the anharmonic motion, especially for the highly unequal atomic radii. Considering the small contribution of core electrons to the bond susceptibility, it is proposed that a generalized homopolar part of the energy gap is [8]

$$E_h^{-2} = (E_h^{-2})_o \frac{(r_\alpha - r_c)^{2s} + (r_\beta - r_c)^{2s}}{2(r_o - r_c)^{2s}} \tag{2.20}$$

where $(E_h^{-2})_o$ is the homopolar gap when $r_\alpha = r_\beta = r_o$, $r_c$ is the average core radius. With the expressions given in (2.20) and (2.12), we are ready to calculate the nonlinear susceptibility.

From the assumption of the geometric composition of the susceptibility, the change of the susceptibility is expressed in terms of the changes of the bond susceptibility and direction cosines, (Fig. 2-1),

$$\Delta\chi_{ij} = \frac{1}{v}\{\sum (\Delta\beta_n)\ \alpha_{ni}\ \alpha_{nj}$$

$$+ \sum \beta_n(\Delta\alpha_{ni})\ \alpha_{nj} + \sum \beta_n\alpha_{ni}(\Delta\alpha_{nj})\} \tag{2.21}$$

The changes of direction cosines are due to the relative displacement of atoms. Assuming the electric field is applied in the k direction and induces the isotropic displacement of $\Delta x_k$, we have

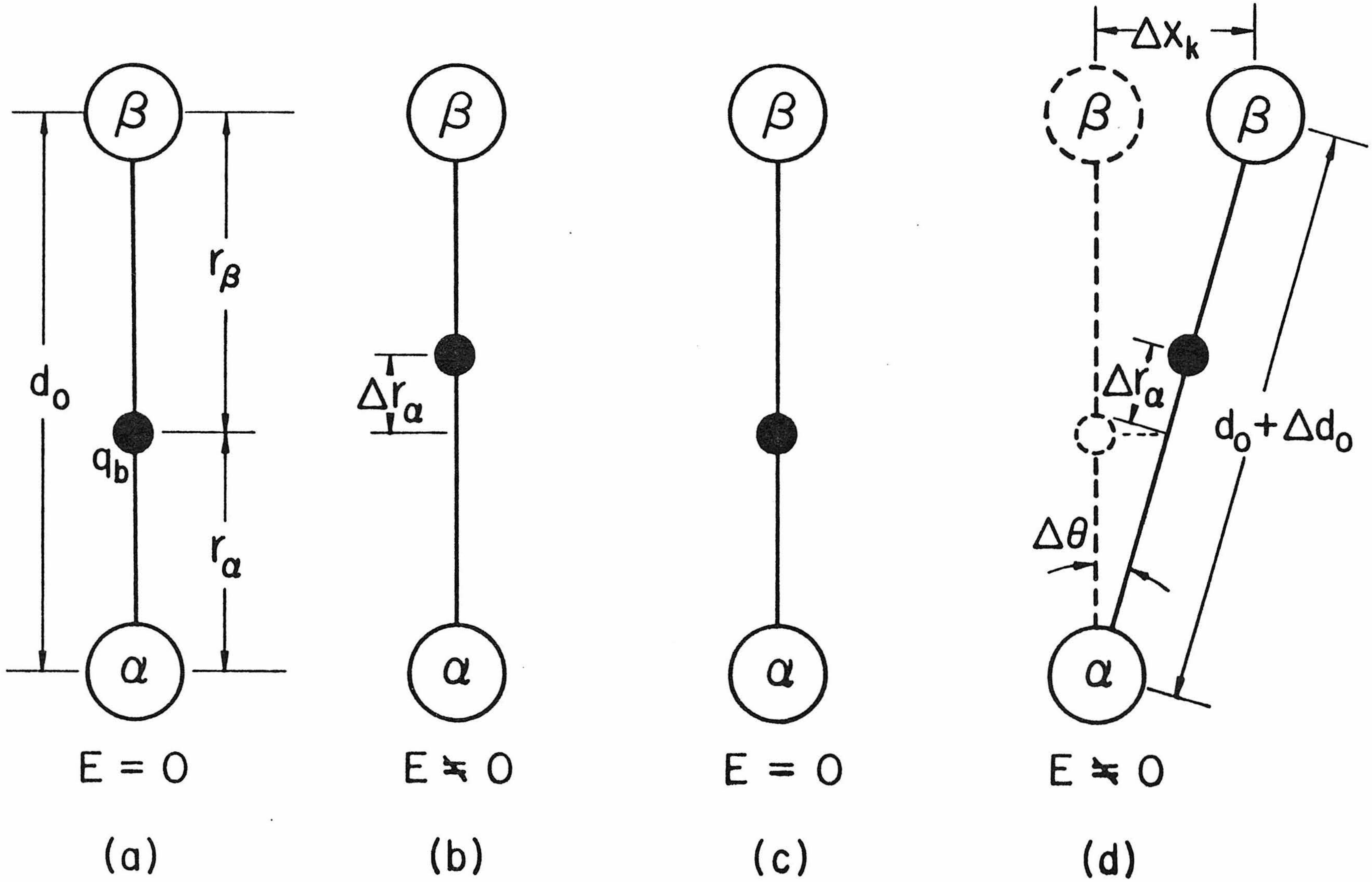

Figure 2-1 : The response of a crystal to the applied electric field. In (b), the atoms stay stationary while the bond charge follows the high frequency field. In (d), both respond to the low frequency field.

$$(\Delta\alpha_i)_k = \frac{\partial}{\partial x_k} (\alpha_i) \, \Delta x_k$$

$$= \frac{\partial}{\partial x_k} (x_i/d_o) \, \Delta x_k$$

$$= \frac{1}{d_o} [\delta_{ik} - \alpha_i\alpha_k] \, \Delta x_k \qquad (2.22)$$

where $d_o$ is the bond length, $d_o = 2r_o$.

The change of the bond susceptibility is due to the displacement of the bond charge and the stretch of the bond. As $\beta \propto (h\omega_p)^2/Eg^2$, where the small contribution from D and A has been neglected, we have

$$\frac{\Delta\beta}{\beta} = \frac{\Delta(\omega_p^2)}{\omega_p^2} + f_c \, E_h^2 \, \Delta(E_h^{-2}) - 2f_i \, \frac{\Delta c}{c} \qquad (2.23)$$

The variations of $\omega_p^2$, $E_h^{-2}$, and c depend on $d_o$, $r_\alpha$ and $r_\beta$, so we have to relate these quantities to the known physical parameters. When the bond length changes, it is assumed that the ratio of atomic radii remains constant. The two independent parameters, $r_\alpha$ and $r_\beta$, can be transformed into two parameters relating directly to the macroscopic properties of the crystal

$$\Delta r_\alpha = \frac{r_\alpha}{d_o} \Delta d_o + \delta$$

$$\Delta r_\beta = \frac{r_\beta}{d_o} \Delta d_o - \delta \qquad (2.24)$$

where $\delta$ is the displacement of the bond charge independent of the ionic motion, and $\delta$ corresponds to the optical susceptibility while $\Delta d_o$ corresponds to the low frequency dielectric constant.

From (2.12), (2.20), (2.23), and (2.24), it is a straightforward procedure to obtain

$$\frac{\Delta\beta}{\beta} = \{[f_i(1 + \frac{k_s r_o}{2}) + sf_c - \frac{3}{2}] \frac{\Delta d_o}{r_o}$$

$$+ [4f_i \frac{z_\alpha + z_\beta}{z_\alpha - z_\beta} + s(2s-1) \frac{f_c \rho d_o^2}{(r_o - r_c)^2}] \frac{\delta}{d_o} \} \tag{2.25}$$

where $\rho = (r_\alpha - r_\beta)/(r_\alpha + r_\beta)$ measures the importance of the unequal atomic radii. In the first square bracket, the term $k_s r_o/2$ is obtained because the screening wave number $k_s$ is proportional to $r_o^{-1/2}$. The number (-3/2) is due to the fact that $\omega_p$ is proportional to $r_o^{-3/2}$. The expression in the second square bracket is the contribution from the bond-charge response. It is exactly identical to the result obtained by Levine in his calculation of the nonlinear optical susceptibility [8]. Thus the expression in the first square bracket is the ionic contribution of a single bond due to the bond stretching. The electronic contribution has been studied in detail elsewhere and will not be discussed in this thesis. Our concern here is only with the ionic contribution to the linear electrooptic effect.

Substituting (2.22) and (2.25) into (2.21), the ionic contribution to $(\Delta\chi_{ij})_k$ is

$$\Delta\chi_{ijk}^{ion} = \{\frac{1}{V} \sum_n \frac{\beta_n}{r_o} [f\ \alpha_{ni}\alpha_{nj}\alpha_{nk}$$

$$+ \frac{1}{2} (\alpha_{ni}\delta_{jk} + \alpha_{nj}\delta_{ik})]\} \ \Delta x_k \tag{2.26}$$

where

$$f = f_i(1 + \frac{k_s r_o}{2}) + sf_c - 2.5$$

$$= (\frac{k_s r_o}{2} - 1.48)\ f_i - 0.02$$

f is the ionicity factor and $\delta_{jk}$ is the Kronecker delta function. The relative displacement of atoms $\Delta x_k$ is related to the dielectric constant of the crystal as

$$Ne_c^* \Delta x_k = \varepsilon_o(\varepsilon'_{dc_k} - \varepsilon'_{\infty_k}) E_k^s \tag{2.27}$$

where N is the number of pairs of atoms per unit cell, $e_c^*$ is the Callen effective ionic charge [11], $\varepsilon'_{dc}$ is the relative dielectric constant, $\varepsilon'_\infty$ is the relative optical permittivity, and $E_k^s$ is the low frequency electric field in the k direction. The effective charge $e_c^*$ is related to the Szigette effective charge [11] $e_s^*$ as

$$e_c^* = e_s^*(e_\infty + 2)/3\varepsilon_\infty \tag{2.28}$$

Furthermore, the Szigette effective charge has been found empirically equal to $(c/\hbar\omega_p)$ in diatomic crystals [12]. If we adopt this relation for even more complex crystals, it becomes possible to calculate the effective charge from the knowledge of the atomic radius and structure. The only physical quantity which we cannot calculate is the static dielectric constant $\varepsilon_{dc}$. In this thesis, we use the experimental data of $\varepsilon'_{dc}$ to calculate the electrooptic coefficient.

Next we will relate the change of the susceptibility to the electrooptic coefficient. It is convenient to define the coefficient in terms of the change of $1/n^2$, i.e.,

$$rE = \Delta(1/\varepsilon') \tag{2.29}$$

The advantage of this definition can be seen from the "index ellipsoid" describing the optical properties of a crystal. The general equation of this surface is [13]

$$\left(\frac{1}{n^2}\right)_1 x^2 + \left(\frac{1}{n^2}\right)_2 y^2 + \left(\frac{1}{n^2}\right)_3 z^2 + 2\left(\frac{1}{n^2}\right)_4 yz$$

$$+ 2\left(\frac{1}{n^2}\right)_5 xz + 2\left(\frac{1}{n^2}\right)_6 xy = 1 \tag{2.30}$$

in an arbitrary coordinate system x, y, z. If we choose the coordinates to be parallel to the principal dielectric axes of the crystal, then with zero applied field,

$$\begin{aligned} \left(\frac{1}{n^2}\right)_1 &= \frac{1}{n_x^2} & \left(\frac{1}{n^2}\right)_2 &= \frac{1}{n_y^2} \\ \left(\frac{1}{n^2}\right)_3 &= \frac{1}{n_z^2} & \left(\frac{1}{n^2}\right)_4 &= \left(\frac{1}{n^2}\right)_5 = \left(\frac{1}{n^2}\right)_6 = 0 \end{aligned} \tag{2.31}$$

According to the definition (2.29), we have the following matrix representing the change of constant due to an arbitrary low frequency electric field

$$\begin{bmatrix} \Delta\left(\frac{1}{n^2}\right)_1 \\ \Delta\left(\frac{1}{n^2}\right)_2 \\ \Delta\left(\frac{1}{n^2}\right)_3 \\ \Delta\left(\frac{1}{n^2}\right)_4 \\ \Delta\left(\frac{1}{n^2}\right)_5 \\ \Delta\left(\frac{1}{n^2}\right)_6 \end{bmatrix} \quad \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \\ r_{41} & r_{42} & r_{43} \\ r_{51} & r_{52} & r_{53} \\ r_{61} & r_{62} & r_{63} \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix} \tag{2.32}$$

Because the index of refraction of a crystal depends on the optical polarization relative to the crystal axes, the matrix form for r must reflect the crystal symmetry. Highly symmetric crystal should have fewer independent nonzero elements in its matrix. A complete list of the matrix form for various point groups can be found in Ref. 13.

The coefficient $r_{ijk}$ is related to the change in optical susceptibility by

$$r_{ijk}E^{s}_{k} = -\frac{\Delta\chi_{ijk}}{\varepsilon'_i\varepsilon'_j} \tag{2.33}$$

Combining (2.33), (2.26), and (2.27), the final expression of the electrooptic coefficient is

$$r_{ijk} = \frac{\varepsilon_o(\varepsilon'_{dck} - \varepsilon'_{\infty k})}{vNe^*_c\ \varepsilon'_i\ \varepsilon'_j}\{\sum_n \frac{\beta_n}{r_o}[f\alpha_{ni}\alpha_{nj}\alpha_{nk} + \frac{1}{2}(\alpha_{ni}\delta_{jk} + \alpha_{nj}\delta_{ik})]\} \tag{2.34}$$

As the direction cosines appear only as $\alpha_i$ or $\alpha_i\alpha_j\alpha_k$, it is apparent that $r_{ijk} = 0$ if the crystal has inversion symmetry. Usually, $\alpha_i\alpha_j\alpha_k$ should be an order of magnitude smaller than $\alpha_i$ if $\Sigma\alpha_i \neq 0$. The ionicity factor has been calculated for many crystals. Its value is about 0.1 to 0.2 for diatomic crystals and never exceeds 0.3 for most complex crystals.

## CHAPTER 2 - REFERENCES

1. J. C. Phillips, Phys. Rev. Lett. 20, 550 (1968).
2. J. C. Phillips, Phys. Rev. 166, 832 (1968); 168, 905 (1968).
3. J. C. Phillips and J. A. Van Vechten, Phys. Rev. Lett. 22, 705 (1969).
4. J. A. Van Vechten, Phys. Rev. 182, 891 (1969); 187, 1007 (1969).
5. J. C. Phillips, Rev. Mod. Phys. 42, 317 (1970).
6. B. F. Levine, Phys. Rev. Lett. 22, 787 (1969); 25, 440 (1970).
7. B. F. Levine, J. Chem. Phys. 59, 1463 (1973).
8. B. F. Levine, Phys. Rev. B 7, 2591 (1973); 7, 2600 (1973).
9. D. R. Penn, Phys. Rev. 128, 2093 (1962).
10. C. Kittel, Introduction to Solid State Physics, 5th ed. (Wiley, New York, 1976).
11. H. B. Callen, Phys. Rev. 76, 1394 (1949).
12. P. Lawaetz, Phys. Rev. Lett. 26, 697 (1971).
13. A. Yariv, Quantum Electronics, 2nd ed. (Wiley, New York, 1975).

# CHAPTER 3
# THE CALCULATION OF THE ELECTROOPTIC COEFFICIENTS OF DIATOMIC AND TERNARY COMPOUNDS

## 3.1 Introduction

We have established a theoretical model and derived an expression to calculate the electrooptic coefficients of crystals. In this chapter the theory will be applied to the crystals with two or three atoms per formula. At first, we calculate the coefficients of zinc-blende and wurtzite crystals which are the basic lattice structures of the PV dielectric theory. Then, quartz with two atoms per formula, is chosen to carry out the extension of the theory to the range beyond the AB type.

In some of the ternary compounds, not all the bonds have to be taken into account in the calculation. Lithium niobate and tantalate are used to demonstrate the negligible contribution of the highly ionic bond like Li—O. It is also shown that the distorted octahedron which results in two different bond lengths for Nb—O is responsible for the nonlinear effect. The $KH_2PO_4$ (KDP) family is interesting in its nonlinear properties. The calculation shows the dominant role played by P—O bonds. In order to investigate the materials with wider range of transparency in the infrared region, we calculate the coefficient of chalcopyrite crystals which has the unit cell structure evolved from the zinc-blende crystal.

The last ternary crystal we consider is proustite whose complex structure contains 54 bonds per unit cell.

We calculate the ionic part of the electrooptic coefficient. The electronic contribution is taken numerically from the SHG coefficients. The sum of these two parts is compared with the measured value. In general, our calculated values are in good agreement with those obtained experimentally.

## 3.2 Diatomic Crystals--Zinc-Blende and Wurtzite

The diatomic crystals can be divided into four groups according to their structures: zinc-blende, wurtzite, rock salt and CsCl type. The structures in a unit cell of these four groups are shown in Figure 3.1. The crystals with rock salt and CsCl structures have inversion symmetry and, consequently, do not have a linear electrooptic effect. Zinc-blende crystals are cubic, while wurtzite crystals are hexagonal, but both are in tetragonal coordination. There is no difference between the two structures if one looks at only the nearest atoms. They can only be distinguished by comparing the position of the next nearest neighbors. The definition of the ionicity in terms of the symmetric and antisymmetric energy gaps shows its advantage in the classification of those groups with different coordination numbers. It has been noticed that $f_i = 0.785$ is the dividing line between tetragonal coordination and rock salt . The lowest ionicity in the rock salt crystals is 0.785 for CdO. The crystals, MgS ($f_i = 0.786$) and MgSe ($f_i = 790$) have an ionicity close to 0.785 and exist in both structures of rock salt and wurtzite. For the CsCl type crystals, $N_c = 8$, the lowest ionicity is 0.929. Therefore, the concept of the average homopolar and heteropolar energy gaps seems appropriate in describing bond properties.

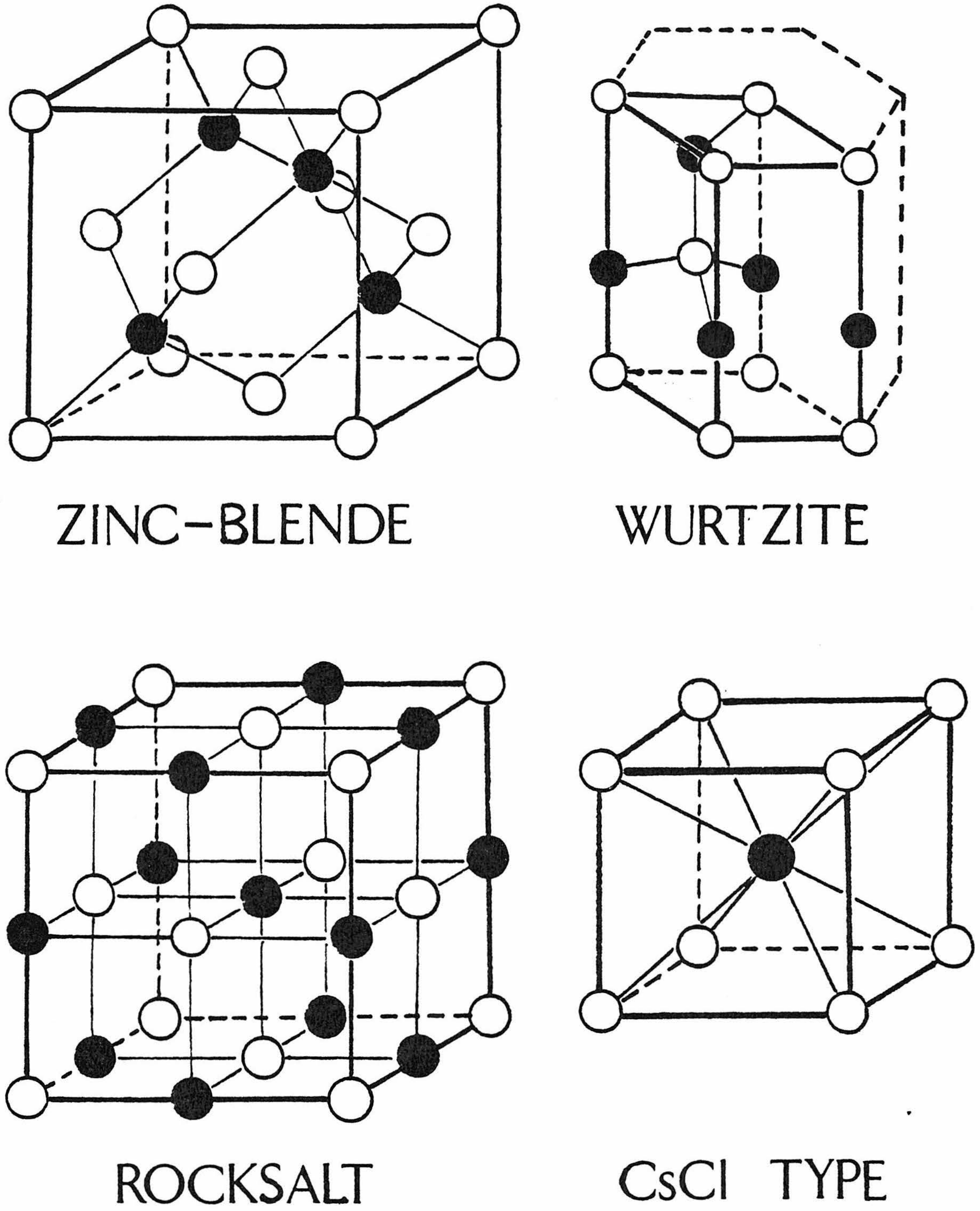

Figure 3-1 : Four types of diatomic crystal structures.

If the slight distortion from the perfect tetragonal structure in wurtzite [1] is neglected, the coefficients of both types of crystals can be calculated in a similar way. In Table I, the parameters used in (2.34) are presented. For the calculation of $\Sigma\alpha_i\alpha_j\alpha_k$ we have chosen the coordinates such that one of the four bonds points in the direction of (111) for zinc-blende and in the +z direction for wurtzite. The sense of the positive polarization of the bond is defined as the direction from the positive ion to the negative ion. Those definitions are in compliance with the conventional choices used in experimental measurements. With the quantities in Table I, the electrooptic coefficient of diatomic crystals is found to be

Zinc-blende:

$$r_{14} = 0.3689 \frac{a_o^2 w f}{e_c^*/e} \tag{3.1}$$

Wurtzite:

$$r_{33} = -2r_{13} = 0.4260 \frac{a_{eff}^2 w f}{e_c^*/e} \tag{3.2}$$

and other coefficients are identically zero due to the symmetric property of the crystals. In (3.1) and (3.2), $r$'s are in units of $10^{-12}$ m/V while $a_o$ and $a_{eff}$ are in units of Å.

We have calculated the electrooptic coefficients of nine diatomic crystals using formulas (3.1) and (3.2). The results for the zinc-blende (GaAs, GaP, ZnSe, ZnS , ZnTe, CuCl) and wurtzite (ZnS , CdS, CdSe) are compared with experiment as shown in Table II. Values of parameters $a_o$, $a_{eff}$, w, $f_i$, f, $e_c^*$ are also listed in the table and discussed in the following.

Table I. Several parameters for zinc-blende and wurtzite crystals

$$a^3_{eff} = \sqrt{3}\, a^2_o\, C_o$$

| | zinc-blende | wurtzite |
|---|---|---|
| # of atoms in a unit cell | 4 | 2 |
| Volume of a unit cell | $a^3_o$ | $\sqrt{3}\, a^2_o C_o/2$ |
| # of atoms per unit volume | $4/a^3_o$ | $4/a^3_{eff}$ |
| $\Sigma\alpha_i$ | 0 | 0 |
| $\Sigma\alpha^2_i$ | 16/3 | 16/3 |
| $\Sigma\alpha_1\alpha_2\alpha_3$ | $16/3\sqrt{3}$ | 0 |
| $\Sigma\alpha^2_1\alpha_3 = \Sigma\alpha^2_2\alpha_3$ | 0 | -16/9 |
| $\Sigma\alpha^3_3$ | 0 | 32/9 |

Table II. Parameters and results of equations (3.1 ) and (3.2 ). $a = a_0$ or $a_{eff}$. r's represent $r_{14}$ (zinc-blende) and $r_{33}$ (wurtzite) and are in units of $10^{-12}$m/v. $r_{exptl}$ are measurements with clamped crystals. Their signs are not yet determined, unless so specified.

| | Zinc-blende | | | | | | Wurtzite | | |
|---|---|---|---|---|---|---|---|---|---|
| AB | GaAs | GaP | ZnSe | ZnS | ZnTe | CuCl | ZnS | CdS | CdSe |
| a [a] | 5.65 | 5.45 | 5.67 | 5.41 | 6.09 | 5.41 | 5.39 | 5.85 | 6.08 |
| $\varepsilon'_{dc}$ | 13.2[b] | 12.0[c] | 9.1[d] | 8.3[d] | 10.1[d] | 7.5[e] | 8.7[f] | 9.4[d] | 10.2[d] |
| w | 0.192 | 0.284 | 0.450 | 0.528 | 0.331 | 0.656 | 0.567 | 0.652 | 0.562 |
| $f_i$ [g] | 0.310 | 0.370 | 0.630 | 0.623 | 0.546 | 0.749 | 0.623 | 0.683 | 0.699 |
| f | -0.091 | -0.113 | -0.163 | -0.179 | -0.119 | -0.212 | -0.181 | -0.162 | -0.147 |
| $e_c^*/e$ | 0.20 | 0.23 | 0.33 | 0.35 | 0.26 | 0.27 | 0.35 | 0.41 | 0.36 |
| $r_{ionic}$ | +1.03 | +1.53 | +2.64 | +2.93 | +2.07 | -5.56 | +3.63 | +3.75 | +3.61 |
| $r_{elec}$ | -2.73[h] | -3.20[i] | -4.68[j] | -4.77[k] | -6.41[l] | +2.66[m] | -5.63[k] | -6.71[k] | -7.40[j] |
| $r_{sum}^{theo}$ | -1.7 | -1.7 | -2.0 | -1.8 | -4.3 | -2.9 | -2.0 | -3.0 | -3.8 |
| $r_{exptl}$ | -1.6[h] | -1.1[n] | 2.0[p] | 1.6[p] | 4.3[p] | -2.4[p] | 1.8[p] | 3.0[p] | 4.3[p] |

[a] R.W.G. Wyckoff, Crystal Structures, 2nd ed., Vol. 1 (1963).

[b] S. Jones and S. Mao, J. Appl. Phys. 39, 4038 (1968).

[c] I. P. Kaminow and E. H. Turner, Proc. IEEE 54, 1374 (1966).

[d] D. Berlincourt, H. Jaffe, and L. R. Shiozawa, Phys. Rev. 129, 1009 (1963).

[e] P. Alomas, G. Sherman, C. Wittig, and P.D. Coleman, Appl. Optics 8, 2557, (1968).

[f] I. B. Kobyakov, Soviet Phys.-Cryst. 11, 369 (1966).

[g] J. A. Van Vechten, Phys. Rev. 182, 891 (1969); 187, 1007 (1969).

[h] A. Mooradian and A. L. McWhorter, Light scattering spectra of solids, G. B. Wright, ed. (Springer, New York, 1969).

[i] J. J. Wynne and N. Bloembergen, Phys. Rev. 188, 1211 (1969).

[j] R. A. Soref and H. W. Moos, J. Appl. Phys. 35, 2152 (1964).

[k] C.K.N. Patel, Phys. Rev. Lett. 16, 613 (1966).

[l] R. K. Chang, J. Ducuing, and N. Bloembergen, Phys. Rev. Lett. 15, 415 (1965).

[m] D. Chemla, P. Kupecek, C. Schwartz, C. Schwab, and A. Goltzene, IEEE J. Quant. Electron. 7, 126 (1971).

[n] D. F. Nelson and E. H. Turner, J. Appl. Phys. 39, 3337 (1968).

[p] I. P. Kaminow and E. H. Turner, Handbook of Lasers, R. J. Pressley, ed. pp. 453, Chemical Rubber Co., Cleveland, Ohio 1971.

(1) a: a represents $a_o$ for zinc-blende and $a_{eff} = (\sqrt{3}\, a_o^2 c_o)^{1/3}$ for wurtzite. Because both crystals possess tetragonal structure, the density of atoms is expected to be almost the same. This fact is reflected by the values of a in ZnS of different structures.

(2) $\varepsilon'_{dc}$: The values of dielectric constants are taken from the experimental data.They show an interesting relation to the number of valence electrons, $n_v$. The crystals with higher $n_v$ have larger $\varepsilon'_{dc}$. This behavior is probably not related to the value of $n_v$ directly, because it can be seen that the effective ionic charge is actually smaller for higher $n_v$ crystals. It should be the result of the bond flexibility. The crystal with lower $n_v$ is more ionic and has stronger bonds. The distortion of the crystal responding to the applied field becomes smaller. Therefore, it has a smaller relative displacement between atoms which results in smaller $\varepsilon'_{dc}$. This argument is also applied to the crystals with the same $n_v$. For example, ZnTe > ZnSe > ZnS because the atomic radius and the covalency follow the relation Te > Se > S. So the relative displacement is larger for ZnTe, which means higher $\varepsilon'_{dc}$.

(3) $w = (\varepsilon' - 1)(\varepsilon'_{dc} - \varepsilon')/\varepsilon'^2$: Because the higher mobility of the bond charge for the crystal with higher $n_v$, $\varepsilon'$ has a similar behavior as $\varepsilon'_{dc}$. The value of w also falls into three distinct groups, according to the values of $n_v$. However, for $n_v = 2$, w for wurtzite is in general larger than the value for zinc-blende.

(4) $f_i$: The value of $f_i$ is calculated from $E_h$ and c [2]. The ionicity of crystals has been discussed in detail elsewhere [3].

(5) f: The ionicity factor is a very important parameter which is calculated from the screening wave number and the bond ionicity. It

characterizes the magnitude of the contribution from the summation of the triple product of direction cosines (see (2.34)). The value of f in diatomic crystals ranges from 0.1 to about 0.2. Because it is such a small value, the accuracy of the estimate of s becomes very important. The uncertainty in s limits the uncertainty in f to be higher than 10%. However, f is no doubt a negative quantity.

(6) $e_c^*/e$: The effective ionic charge is calculated from the Szigette effective charge. It is a very small value due to the local field correction factor.

(7) $r_{ionic}$: The ionic contribution of the electrooptic coefficient is obtained by using (3.1) and (3.2). It is obvious that the crystal with lower $n_v$ has higher $r_{ionic}$. This is due to its higher value of W and f. The uncertainty of $r_{ionic}$ is due mostly to two sources. One is the uncertainty in the measured values of the physical parameters. The other is in the assumptions and approximations of the theory. Over all, the uncertainty of $r_{ionic}$ should be about 15-20%, which is usually also the standard deviation of the measurement.

(8) $r_{elec}$: The purely electronic contribution is obtained from the coefficient of the second harmonic generation (SHG) using the relation

$$r_{ijk} = -4d_{ijk}/\varepsilon_i\varepsilon_j \qquad (3.3)$$

We have assumed that the coefficient $d_{ijk}$ has no dispersion in the frequency. As $r_{ijk}$ is obtained in the limit of long wavelength, $d_{ijk}$ should be the measurement at long wavelengths where the nonlinear effect and linear effect are less dispersive. The value of $r_{elec}$ seems to have no obvious relation to the number of valence electrons.

(9) $r_{sum}^{theo}$: The predicted coefficient is the sum of the theoretical calculation $r_{ionic}$ and the experimental data $r_{elec}$. We intend to show the coefficient r, although the modulation strength depends on $n^3r$.

(10) $r_{exptl}$: The predicted value is compared with the measurement. $r_{exptl}$ is obtained with the clamped crystals where the strain induced effect can be neglected.

The comparison between $r_{sum}^{theo}$ and $r_{exptl}$ shows that the prediction is in good agreement with experiment including the determination of signs. The worst discrepancy is in GaP, but it is interesting to note that the electronic contribution is about double the ionic contribution. This is in good agreement with the observation from Raman scattering [4]. Another interesting example is CuCl. It is found that $r_{elec}$ of CuCl is positive and $r_{exptl}$ is negative, which implies that $r_{ionic}$ must be negative and a large value. If we assign a negative sign to our calculated result, the value of $r_{sum}^{theo}$ is close to what we expect it should be. The question arises as to why CuCl should possess different sign for $r_{ionic}$ compared to the other crystals. This novel behavior has been explained by considering the d-electron contribution [5]. Due to their high mobility around the molecule, the valence electrons do not contribute to the molecular polarization. The polarization thus is determined by the charges of the nucleus and core electrons. For example, in GaAs, Ga has a total "core" charge of +3 and As has +5. In CuCl, Cl has a total core charge of +7 while, excluding the d-electrons, Cu has +11.

So it seems that the bond polarization of CuCl has the opposite sense from that of GaAs. This explanation is supported by the fact that CuCl also has a different sign for $r_{elec}$ from other crystals. However, the sense of polarization cannot be known by the low frequency electric field. The displacement is also too small ($\sim 0.01\mathring{A}$) in the case of a static field to be detected by crystallography. Therefore, the polarization problem in CuCl is still an open question. As expected, this should happen also in the crystal CuI.

Recently, the electrooptic coefficient of InP was measured for the first time at Hughes Research Laboratories [6]. Before knowing the result, we used our theory to calculate the electrooptic content. The measurement was performed at a wavelength of 3.39μ, but the second harmonic generation coefficient which is used as part of our input has only been observed in the visible region where the material is very dispersive. Using the Miller rule [7], we estimated the electronic contribution at 3.39μ and predicted the total electrooptic coefficient to be $(1.3 \pm 15\%) \times 10^{-12}$ m/V. The experimental measurement following our calculation yielded a value of 1.3 to 1.6. So, again, the predicted value is in good agreement with experiment.

It should not be surprising that agreement of the theory with experiment is satisfactory, since it is based on the theory describing the bond properties of diatomic crystals. However, the dc dielectric constant is not much larger than $\varepsilon'_\infty$ and the value of $\varepsilon'_{dc}$ varies with the experimental conditions. A small amount of error in $\varepsilon'_{dc}$ could lead to a large discrepancy in the value of w. In other complex crystals with very large values of $\varepsilon_{dc}$ such critical dependence disappears. The prediction becomes more reliable with the given experimental value of $\varepsilon_{dc}$.

## 3.3 Diatomic Crystal--Quartz

Quartz is the first crystal discovered to have the linear electro-optic property. The interest in quartz is due mostly to its availability in nature and its excellent optical properties. The transparency region extends from UV (∿ 1800Å) to the infrared (∿ 7 μm).

At room temperature, quartz ($SiO_2$) has the α-structure. Its point group is 32. The positions of silicon and oxygen are shown in Figure 3.2, where all the atoms have been projected onto the x-y plane and the number in the circle shows the position of the atom as a fraction of the lattice constant on the c-axis. There are three molecules in a unit cell.

Although quartz has only two kinds of atoms, it is completely different from the zinc-blende and wurtzite structures. For the diatomic crystals of $AB_n$ type, it has been suggested that the heteropolar part of the energy gap is replaced by the expression [8]

$$c = be^2\left(\frac{z_\alpha}{r_o} - \frac{nz_\beta}{r_o}\right) e^{-k_s R} \tag{3.4}$$

where n is the number of atoms B per formula. The reason n is included is that the valence electron spends about n times more of the period around atom β than around atom α. The effective screened Coulomb potential is thus multiplied by n for atom β.

From the crystallographic data, we find the bond length is 1.61Å and the volume of a unit cell is 113Å$^3$. Every silicon atom bonds to its four nearest oxygen atoms, while every oxygen atom bonds to two silicon

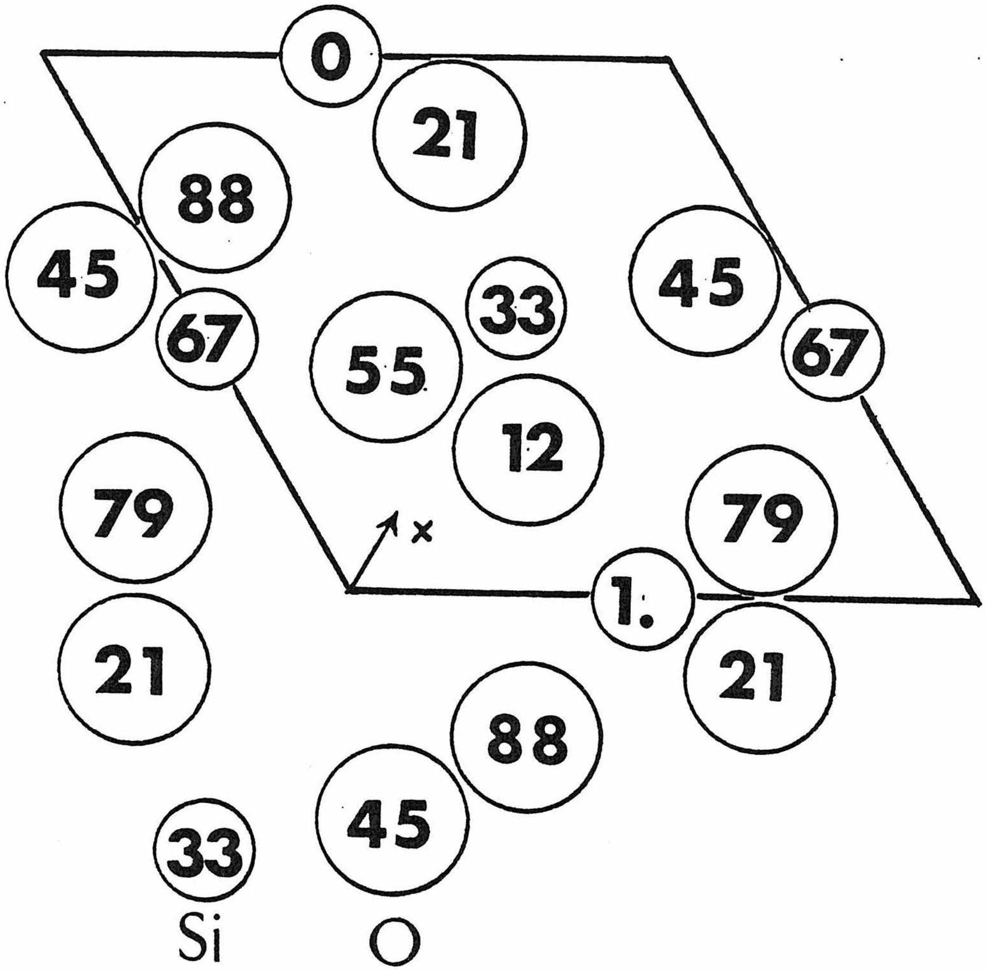

Figure 3-2 : Quartz ( $SiO_2$ ) structure. The number in the circle indicates the position of atom as fraction of a unit cell.

atoms. There are twelve bonds in a unit cell which are all identical except in direction.

As in the conventional coordinates of quartz crystals, the x axis is chosen as in Figure 3.2, and the z axis coincides with the c axis. The summations of direction cosines over the twelve bonds are found to be

$$\begin{aligned}
&\sum \alpha_i = 0 \\
&\sum \alpha_1 = \sum \alpha_2 = 3.97 \;; \quad \sum \alpha_3 = 4.06 \\
&\sum \alpha_1^3 = 0.967 \quad ; \quad \sum \alpha_2^3 = \sum \alpha_3^3 = 0 \qquad (3.5) \\
&\sum \alpha_1\alpha_2^2 = -0.967 \\
&\sum \alpha_1\alpha_3^2 = \sum \alpha_2\alpha_1^2 = \sum \alpha_2\alpha_3^2 = \sum \alpha_3\alpha_1^2 = \sum \alpha_3\alpha_2^2 = 0 \\
&\sum \alpha_1\alpha_2\alpha_3 \simeq 10^{-5}
\end{aligned}$$

Although quartz possesses birefringence, the difference in the index of refraction is only about 1%. For convenience, we use the average value without bringing significant uncertainty into the result, $\sum \alpha^2 = 4.0$.

From the summations of the direction cosines, the coefficient $r_{11}$ which has been measured experimentally is derived from (2.34)

$$r_{11} = \frac{\varepsilon_0}{Ne_c^*} \frac{\chi(\varepsilon'_{dc} - \varepsilon')}{\varepsilon'^2} \frac{\sum \alpha_1^3}{\sum \alpha^2} \frac{f}{r_0} \qquad (3.6)$$

For demonstration and clarity, all physical quantities appearing in the calculation are presented in Table III. The prediction is com-

Table III. Calculation of $r_{11}$ for quartz

| Symbol | Value | Remarks |
|---|---|---|
| $v_b$ | $9.4\text{Å}^3$ | bond volume |
| $n_b$ | 2 | number of valence electrons per bond |
| $n$ | $0.21\text{Å}^{-3}$ | $n = n_b/v_b$ |
| $\hbar\omega_p$ | 17.1 eV | $\hbar\omega_p(\text{eV}) = 37.16 \times \sqrt{n(\text{Å}^{-3})}$ |
| $k_F$ | $1.85\ \text{Å}^{-1}$ | $k_F = (3\pi^2 n)^{1/3}$ |
| $k_S$ | $2.10\text{Å}^{-1}$ | $k_S = (4k_F/\pi a_B)^{1/2}$ |
| $E_h$ | 12.2 eV | homopolar energy gap[a] |
| $c$ | 14.1 eV | heteropolar energy gap[a] |
| $E_g$ | 18.6 eV | $Eg^2 = E_h^2 + c^2$ |
| $f_i$ | 0.57 | $f_i = c^2/Eg^2$ |
| $f$ | -0.39 | $f = (k_s r_o/2 - 1.48)f_i - 0.02$ |
| $e_s^*/e$ | 0.82 | $e_s^*/e = c/\hbar\omega_p$ |
| $e_c^*/e$ | 0.50 | $e_c^* = e_s^*(\varepsilon' + 2)/3\varepsilon'$ |
| $N$ | $0.106\text{Å}^{-3}$ | number of ionic changes per unit volume |
| $\varepsilon'$ | 2.4 | optical permittivity[b] |
| $\varepsilon'_{dc}$ | 4.5 | dielectric constant[c] |
| $r_o$ | $0.805\text{Å}$ | half of bond length |
| $r_{11}^{ion}$ | $0.61 \times 10^{-12}\text{m/V}$ | use (2.34) |
| $d_{11}$ | $0.4 \times 10^{-12}\text{m/V}$ | SHG coefficient at 1.06 μm[d] |
| $d_{11}$ | $0.44 \times 10^{-12}\text{m/V}$ | assume Miller index is constant at 0.633 μm |
| $r_{11}^{elec}$ | $-0.32 \times 10^{-12}\text{m/V}$ | $r_{11}^{elec} = -4d_{11}/\varepsilon^2$ |
| $r_{11}$ | $0.29 \times 10^{-12}\text{m/V}$ | $r_{11} = r_{11}^{ion} + r_{11}^{elec}$ |
| $r_{11}^{exptl}$ | $0.29 \times 10^{-12}\text{m/V}$ | $r_{11}^{exptl}$ at 0.633 μm[e] |

Table III (continued)

[a] B. F. Levine, J. Chem. Phys. 59, 1463 (1973).

[b] I. P. Kaminow and E. H. Turner, *Handbook of Lasers*, R. J. Pressley, ed., Chemical Rubber Co., Cleveland, Ohio, 1971.

[c] V. G. Zubov, M. M. Firsova, and T. M. Molokova, Soviet Phys. Cryst. 8, 85 (1963).

[d] R. C. Miller, Appl. Phys. Lett. 5, 17 (1964).

[e] R. D. Rosner, E. H. Turner, and I. P. Kaminow, Appl. Opt. 6, 779 (1967).

pared with the measurement. The excellent agreement should not be overemphasized. The uncertainty in $r_{11}^{expt}$ is 10% and in $r_{11}^{ion}$ is even higher. In Table III two quantities having abnormal values should be discussed. The value f is much higher than the value in zinc-blende and wurtzite, and $e_c^*$ is also much larger. The higher values of f and $e_c^*$ are actually the result of a lower valence electron density. In a comparable size of the unit cell, quartz has twelve bonds while zinc-blende crystals have sixteen bonds which affects the value of the plasma frequency.

The only experimental value used in this calculation is the dielectric constant $\varepsilon'_{dc}$. This value has been measured over a wide range of temperatures. It is found that for a purified crystal $\varepsilon'_{dc}$ is constant up to about 600°C where the phase change to β-modification takes place [9].

## 3.4 Lithium Niobate and Tantalate

We have calculated the electrooptic coefficients of crystals which contain only two kinds of atoms and only one type of bond. However, the principle of the geometrical superposition of susceptibilities does not limit its applicability to the diatomic crystals. In the following we will extend this principle and the theory to complex crystals.

The first materials we shall consider are lithium niobate ($LiNbO_3$) and lithium tantalate ($LiTaO_3$). These two materials are now used extensively in integrated optics. Large crystals (> 1 cm) with good optical and electrical qualities are available. Due to their high transition

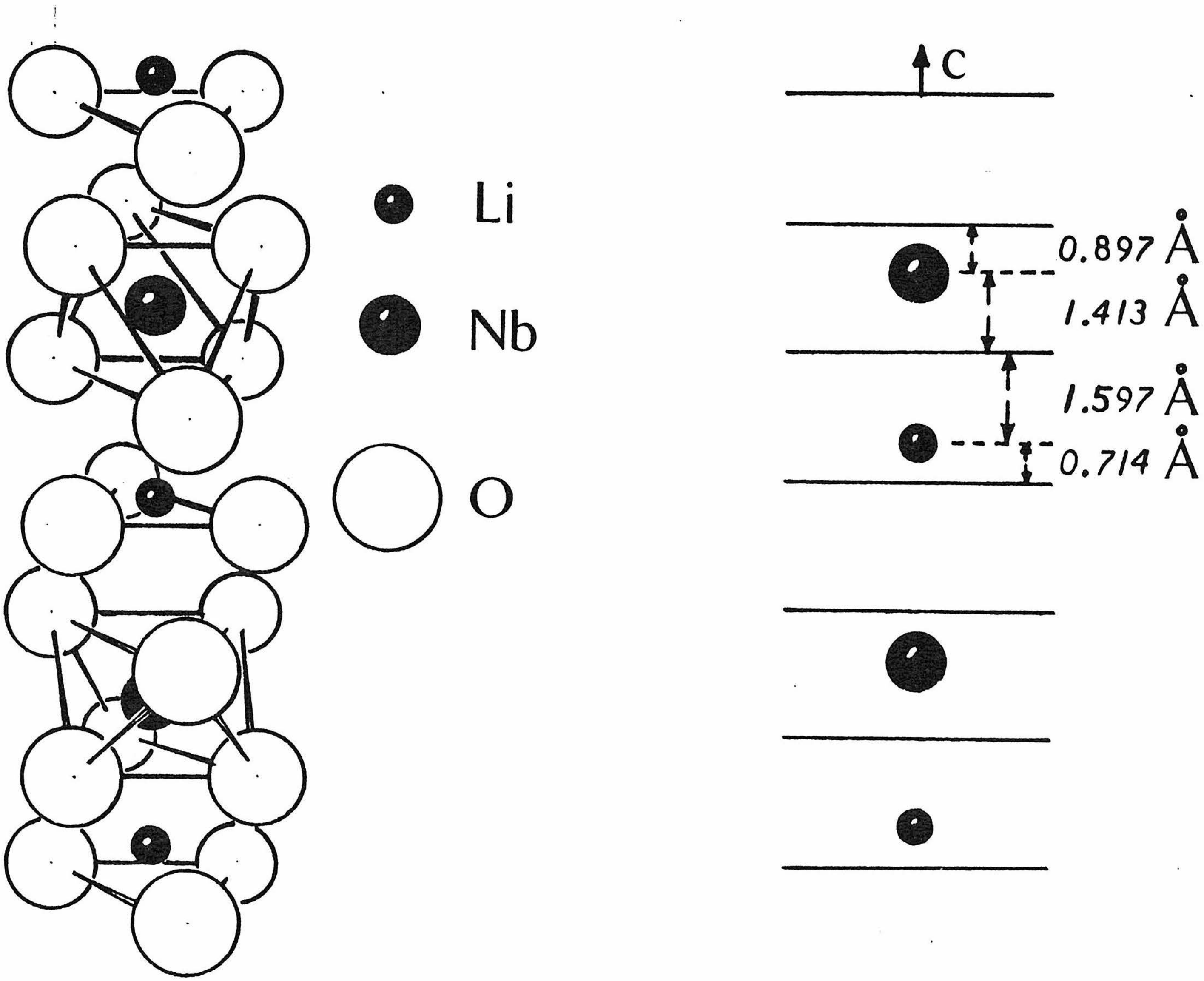

Figure 3-3 : Lithium Niobate ( $LiNbO_3$ ) structure. The diagram on the right side showsthe relative position of the Li ion, Nb ion and the oxygen plane ( horizontal line ).

temperature, both crystals are easy to handle, for example, cut, polish, press, without creating additional domains.

At room temperature, the structures of $LiNbO_3$ [10] and $LiTaO_3$ [11] are rhombohedral with point group 3m. Oxygen atoms form octahedral coordinations with the three-fold axis in the z direction. There are two formulas per unit cell. Two lithium ions and two niobium ions occupy four out of six octahedral sites in a unit cell. In Figure 3.3a we show the rhombohedral $LiNbO_3$-type unit cell in three dimensions. In Figure 3.3b, the positions of oxygen layers, lithium and niobium on c-axis are shown quantitatively. It is obvious that the positions of the lithium and niobium ions are distorted from the center of octahedron due to the occurrence of an empty site for every three octahedral structures. The distortion is actually responsible for the second harmonic generation and linear electrooptic effect. The electrooptic tensor has the non-vanishing components: $r_{33}$, $r_{13} = r_{23}$, $r_{22} = -r_{12} = -r_{16}$, and $r_{42} = r_{51}$.

The crystallographic data show the following results for the lithium niobate [10]:

In a hexagonal unit cell (6 formula)

$a_H = 5.14829\text{Å}$

$c_H = 13.8631\text{Å}$

In a rhombohedral unit cell (2fw)

$a_R = 5.4944\text{Å}$

$\alpha = 55°52'$

Volume per formula: $53.0\text{Å}^3$

Nb—O: 1.889Å and 2.112Å

Li—O: 2.068Å and 2.238Å

and for lithium tantalate [11]:

In a hexagonal unit cell (6fw)

$a_H = 5.15359\text{Å}$

$c_H = 13.78070\text{Å}$

In a rhombohedral unit cell (2fw)

$a_R = 5.4740\text{Å}$

$\alpha = 56°10.5'$

Volume per formula: $52.9\text{Å}^3$

Ta—O: 1.891Å and 2.071Å

Li—O: 2.076Å and 2.293Å

We can see that for each kind of bond there are two different bond lengths. Naturally, different bond lengths result in different bond susceptibilities. Rigorously speaking, there should be four different bonds in a unit cell. However, it has been found that the contribution of Li—O bond to the linear and nonlinear susceptibility is so small that we can almost neglect it [12]. The "almost" means that we can neglect its susceptibility but cannot neglect its existence. Li—O bonds still have to be taken into account when calculating the bond volume plasma frequency and screening factor. So we have only to be concerned with the structured information of the Nb—O bonds and the Ta—O bonds. The results are presented in Table IV. In the table we use the subindex s to indicate quantities for shorter bonds and L for longer bonds. The structure information is only applied to obtain $(f \sum\alpha_3\alpha_1^2 + \frac{1}{2}\sum\alpha_3)$ and

Table IV. The properties of bonds, Nb–O and Ta–O in $LiNbO_3$ and $LiTaO_3$

| $LiMO_3$ | $LiNbO_3$ | | $LiTaO_3$ | |
|---|---|---|---|---|
| M–O | $(Nb–O)_S$ | $(Nb–O)_L$ | $(Ta–O)_S$ | $(Ta–O)_L$ |
| $d_o$ | 1.889Å | 2.112Å | 1.891Å | 2.070Å |
| $E_h$ | 8.20 eV | 6.22 eV | 8.18 eV | 6.53 eV |
| c | 17.54 eV | 13.75 eV | 19.24 eV | 15.75 eV |
| Eg | 19.36 eV | 15.09 eV | 20.91 eV | 17.05 eV |
| $f_i$ | 0.821 | 0.830 | 0.847 | 0.853 |
| f | -0.292 | -0.241 | -0.282 | -0.238 |
| $\sum \alpha_3$ | 2.84913 | -4.01420 | 3.0438 | -3.8760 |
| $\sum \alpha_3\alpha_1^2$ | 1.10334 | -1.10871 | 1.1302 | -1.1293 |
| $\sum \alpha_3$ | 0.396 | -0.195 | 0.410 | -0.261 |
| $\sum \alpha_3^3$ | 0.64244 | -1.79678 | 0.7833 | -1.6175 |
| $\chi^\mu/\chi^\mu_{avg}$ | 0.921 | 1.085 | 0.9350 | 1.070 |
| $f\sum \alpha_3\alpha_1^2 + \frac{1}{2}\sum \alpha_3$ | 1.10236 | -1.7405 | 1.2032 | -1.6698 |
| $f\sum \alpha_3^3 + \sum \alpha_3$ | 2.66154 | -3.5821 | 2.8229 | -3.4918 |

$(f \sum\alpha_3^3 + \sum\alpha_3)$ which correspond to the coefficients $r_{51}$ and $r_{33}$, respectively, since these two coefficients are large and of most interest for a 3 m crystal. The reason why $r_{51}$ and $r_{33}$ are large becomes obvious in Table IV. In these two crystals, $\sum\alpha_3 \neq 0$. Furthermore, $\sum \alpha_3$ has about four times the value of $\sum\alpha_3^3$ and nearly three times the value of $\sum \alpha_1^2\alpha_3$. As f is a small value ($f < 0.3$), the term $\sum \alpha_3$ actually dominates over the term containing f which involves the information of the electronic structure.

The results of calculations are shown in Table V. The ionic charge of niobium is found from the average c and $\hbar\omega_p$. However, the ionic charge of lithium is set equal to 1 due to the high ionicity of the Li–O bond. If it is assumed that under the applied electric field, the displacements of the niobium and lithium ions have the same magnitude and are in the same direction, the effective ionic charge per formula is obtained to be 1.8e for $LiNbO_3$ and 2.0e for $LiTaO_3$. The low frequency dielectric constants which are different in z and x or y directions are listed to calculate the factor $(\varepsilon'_{dc} - \varepsilon')/\varepsilon'\varepsilon'$. The value found is the ionic contribution to the electrooptic coefficient and enters as $r^{ionic}$. The electronic contribution is obtained from the coefficient of SHG [13]. The theoretical prediction $r^{sum}$ is taken as the sum of those two values and compared with the experimental measurement [14]. It is found that the prediction is in good agreement with experiment. The difference is less than 10% which is well within the uncertainty of the theory and the measurement.

Table V. Results of $r_{15}$ and $r_{33}$ for $LiNbO_3$ and $LiTaO_3$

| | $LiNbO_3$ | $LiTaO_3$ |
|---|---|---|
| c | 15.5 eV | 17.4 eV |
| $\hbar\omega_p$ | 28.6 eV | 29.3 eV |
| $e_s^*/e$ | 3.7 | 4.0 |
| $e_c^*/e$ | 1.8 | 2.0 |
| $\varepsilon'_{dc,3}$ | 28 | 43 |
| $\varepsilon'_{dc,1,2}$ | 43 | 41 |
| $(\varepsilon'_{dc1}-\varepsilon'_{\infty 1})\ /\ \varepsilon'_1\varepsilon'_3$ | 1.5419 | 1.6393 |
| $(\varepsilon'_{dc3}-\varepsilon'_3)/\varepsilon'^2_3$ | 1.0245 | 1.7210 |
| $r_{51}^{ionic}$ | $+19.7 \times 10^{-12}$m/V | $+16.5 \times 10^{-12}$m/V |
| $r_{51}^{elec}$ | $+0.8 \times 10^{-12}$m/V | $+0.2 \times 10^{-12}$ m/V |
| $r_{51}^{sum}$ | $+20.5 \times 10^{-12m/V}$ | $+16.7 \times 10^{-12}$m/V |
| $r_{51}^{exptl}$ | $+23 \times 10^{-12}$m/V | $+15 \times 10^{-12}$m/V |
| $r_{33}^{ionic}$ | $+19.9 \times 10^{-12}$m/V | $+27.8 \times 10^{-12}$m/V |
| $r_{33}^{elec}$ | $+6.0 \times 10^{-12}$m/V | $+3.7 \times 10^{-12}$ m/V |
| $r_{33}^{sum}$ | $+25.9 \times 10^{-12}$m/V | $+31.5 \times 10^{-12}$m/V |
| $r_{33}^{exptl}$ | $+28 \times 10^{-12}$m/V | $+30 \times 10^{-12}$ m/V |

## 3.5 The KDP Family

Potassium dihydrogen phosphate (KDP) and ammonium dihydrogen phosphate (ADP) are the best known nonlinear materials. They can be grown easily from a water solution with dimensions as large as several centimeters. The crystals are usually of good optical quality and can be cut or polished without difficulty. At room temperature, KDP and ADP are piezoelectric, and belong to the point group $\overline{4}2m$. Although the SHG has been observed [15] below the Curie temperature, no electrooptic measurements have been made at low temperatures where ADP is antiferroelectric and KDP is ferroelectric. Above $T_c$ the only nonvanishing electrooptic coefficients are $r_{41} = r_{52}$ and $r_{63}$. The transparent region for the crystal is from $0.2\mu$ to about $1 \sim 2\mu$. Both the electrooptic coefficients and the index of refraction are almost constant in this range.

Without changing the crystal structure, the KDP family is obtained by replacing K, H, P with some atoms from the corresponding columns in the periodic table or with some equivalent clusters, e.g., K can be replaced by $NH_4$. So far, only five members of the family have had their dielectric constants determined. They are $KH_2PO_4$ (KDP), $KD_2PO_4$ (KDDP), $KH_2AsO_4$ (KDA), $RbH_2AsO_2$ (RDA), and $NH_4H_2PO_4$ (ADP). Therefore, we will apply the theory on these crystals and compare the results with experiment.

The crystal structure of KDP is shown in Figure 3.4. Both K and P are in tetragonal coordination with oxygen atoms. Hydrogen is about 0.21 $\overset{\circ}{A}$ from the midpoint of the line joining the oxygens. There exist

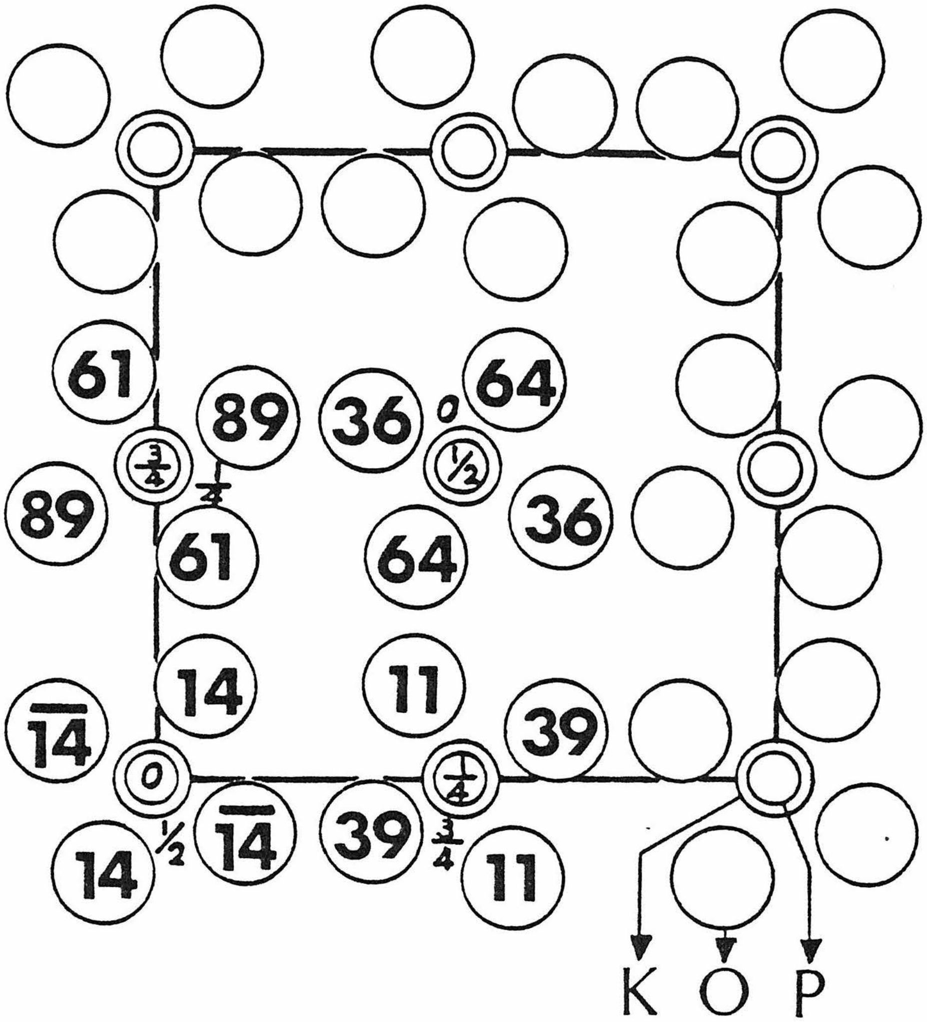

Figure 3-4 : KDP ($KH_2PO_4$) structure.

three different bonds: K—O, H—O, and P—O. Using an argument similar to that used when discussing the Li—O bond in $LiNbO_3$, the contribution of the K—O bond to the linear and nonlinear susceptibility can be neglected due to its high ionicity. The H—O bond is covalent and contributes to the linear susceptibility. However, the almost random distribution of H—O bonds makes the contributions to the nonlinear susceptibility cancel each other, so the only contribution to the electrooptic coefficient comes from P—O bonds. Usually, from the principle of the geometrical decomposition of susceptibility, the bond polarizability can be obtained easily. In KDP, the participation of H—O bonds in $\chi$ prevents us from calculating the polarizability of P—O. In order to avoid such difficulty, we can assume that the polarizability of a bond is almost the same in two crystals if the environments of the bond are similar. In the crystals without hydrogen atoms, $AlPO_4$ is the best candidate to find the polarizability of the P—O bond because the P atom in KDP and $AlPO_4$ has the same coordination structure. The bond properties of P—O in $AlPO_4$, including C, $E_h$, $f_i$, have been obtained by decomposition [8]. Since all members in the KDP family have the same structure, the properties and direction cosines of P—O bonds should not deviate much from crystal to crystal. With this argument, the electrooptic coefficient for any crystals of the KDP type is represented by the formula

$$r = \frac{\varepsilon_0}{Ne_c^*} \frac{\chi(P\text{-}O)}{r_0} \frac{\Sigma\alpha_1\alpha_2\alpha_3}{\Sigma\alpha^2} \frac{\varepsilon_{dc} - \varepsilon}{\varepsilon^2} f \qquad (3.7)$$

The values of parameters appearing in the formula are given in the following:

$c(\text{P—O}) = 12.7$ eV

$E_h(\text{P—O}) = 13.2$ eV

$f_i(\text{P—O}) = 0.481$

$h\omega_p(\text{P—O}) = 25.1$

$e_s^*/e(\text{P}) = 0.506 \times 5$

$e_s^*/e$ per formula $= 5.8$

$e_c^*/e$ per formula $= 3.2$

$N = 0.010$ formula/Å$^3$

$\chi(\text{P—O}) = 0.85$

$r_o(\text{P—O}) = 0.78$ Å

$\Sigma\alpha_1\alpha_2\alpha_3/\Sigma\alpha^2 = 0.5$ [16]

$k_s r_o = 1.92$

$f = 0.271$

Using these values, the coefficient is found to be

$$r = \frac{\varepsilon'_{dc} - \varepsilon'}{\varepsilon^2} 2.47 \times 10^{-12} \text{ m/V} \tag{3.8}$$

The electronic contribution is calculated from the SHG coefficient and ranges from 0.2 to 0.4. It is negligible compared with the total value which is about 10. Thus the electrooptic effect in KDP is essentially ionic: Since the coefficient $r_{41}$ in most crystals has not been measured at high frequencies, we present only the calculation of $r_{63}$ in Table VI. The predictions are in good agreement with experiment. In the calculation, $\varepsilon'$ is taken to be 2.3 because it is almost a constant for various crystals. We also assume that all positive ions, K, H, and P, are displaced by the same distance under the applied electric field. Without

Table VI. The electrooptic coefficient $r_{63}$ of KDP, KDDP, and ADP. The values corresponding to constant stress are noted by (T) and constant strain (clamped) by (S)

| | $\varepsilon_{dc3}$ | $\frac{\varepsilon_{dc3} - \varepsilon}{\varepsilon^2}$ | $r_{63}^{theor}$ $(10^{-12}$ m/V) | $r_{63}^{exptl}$ $(10^{-12}$m/V) |
|---|---|---|---|---|
| KDP | (T) 21 [a] | 3.535 | 8.7 | 9.4 [e] |
| | (S) 21 [b] | 3.535 | 8.7 | 8.8 [f] |
| KDDP | (T) 50 [c] | 9.017 | 22.3 | 26.4 [c] |
| | (S) 48 [d] | 8.639 | 21.3 | 24.0 [g] |
| ADP | (T) 15 [a] | 2.401 | 6.5 | 8.5 [h] |
| | (S) 14 [d] | 2.212 | 5.5 | 5.5 [h] |

[a] D. A. Berlincourt, D. R. Curran and H. Jaffe, Physical Acoustics, Vol. 1, pt. A, W. P. Mason, ed. (Academic Press, New York, 1964).

[b] I. P. Kaminow and G. O. Harding, Phys. Rev. 129, 1562 (1963).

[c] T. R. Sliker and S. R. Burlage, J. Appl. Phys. 34, 1837 (1963).

[d] I. P. Kaminow, Phys. Rev. 138 A, 1539 (1965).

[e] O. G. Blokh, Sov. Phys.-Cryst. 7, 509 (1962).

[f] R. D. Rosner, E. H. Turner, and I. P. Kaminow, Appl. Optics 6, 778 (1967).

[g] T. M. Christmas and C. G. Wildey, Electr. Lett. 6, 152 (1970).

[h] R. O'B. Carpenter, J. Opt. Soc. Am. 40, 225 (1950); 25, 1145 (1953).

assumption, the calculation of the displacement of P atoms from the knowledge of $\varepsilon'_{dc}$ is impossible.

## 3.6 Ternary Chalcopyrite Compounds

Ternary compounds are interesting because of their large electrooptic coefficients. As we have shown previously for $LiNbO_3$, $LiTaO_3$, and the KDP family, the coefficient is as high as $30 \times 10^{-12}$m/V which is an order of magnitude higher than the coefficient for diatomic crystals. However, the use of such materials is limited at long wavelengths. The infrared absorption of KDP begins at 1.5μ due to the vibration of H ions. The upper limit of the transparency range of oxides is at about 5μ. In order to extend the applicability of electrooptic materials to longer wavelengths, we have to choose compounds with heavier atoms such that the resonant energy of vibration is lower. For example, oxygen can be replaced by other atoms in the same column, such as S, Se or Te. The simplest compound has the structure of chalcopyrite ($CuFeS_2$) with point group $\bar{4}2m$. As shown in Figure 3.5, a unit cell of the chalcopyrite structure consists of two unit cells of the zinc-blende structure, such as GaAs. Sulfur atoms occupy the positions of As, while Cu and Fe share evenly the positions of Ga. With a little distortion from the perfect tetragonal coordination, the ratio of the lattice constants, c/a, is usually less than 2 [16].

In general, the compounds with chalcopyrite structure are written as $ABC_2$ which can be the composition of II-IV-$V_2$ or I-III-$VI_2$. Both A—C and B—C bonds are not extremely ionic and contribute comparably to the linear and nonlinear susceptibility. Therefore, these compounds

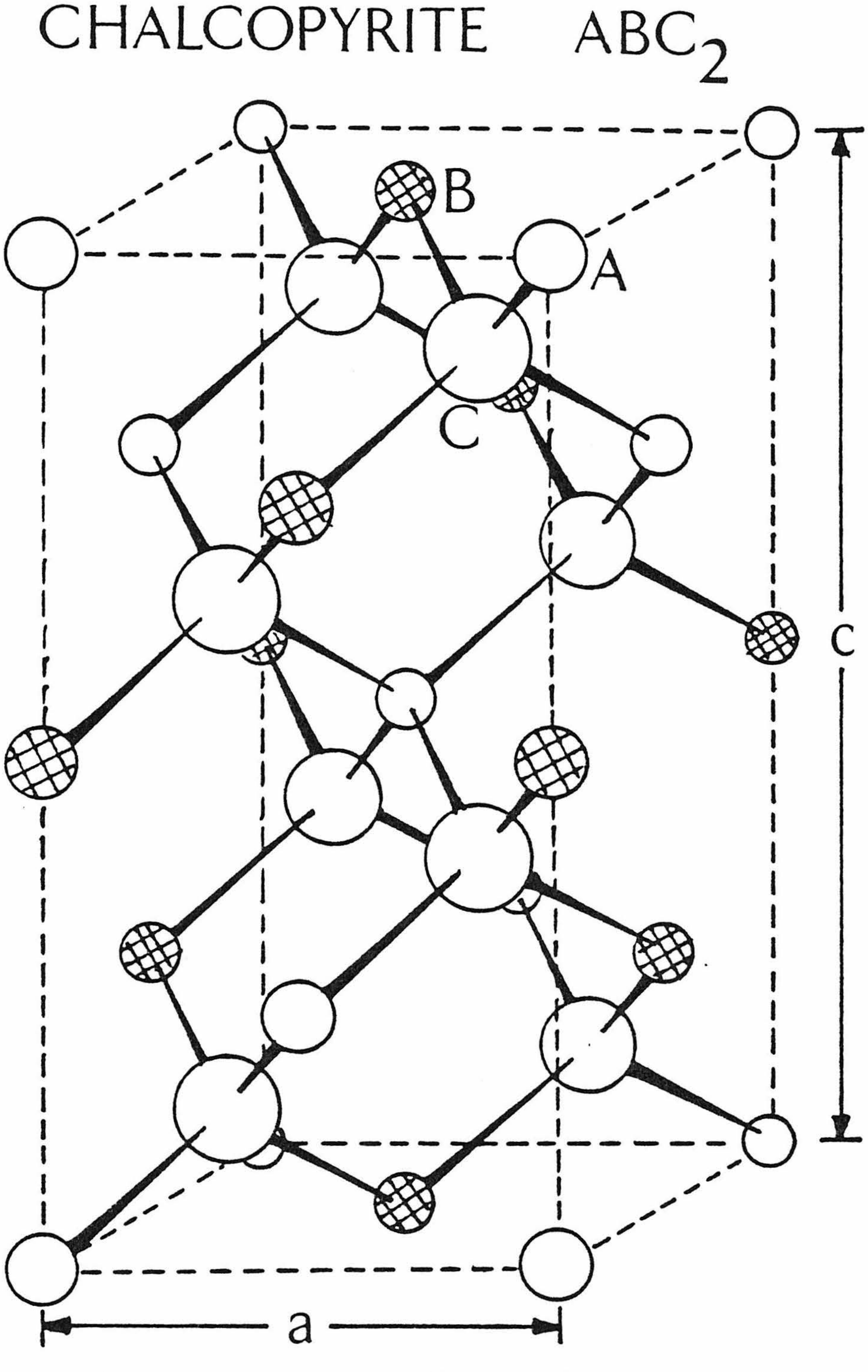

Figure 3-5 : Chalcopyrite structure.

could test the theory in the case of two kinds of completely different bonds.

Up to now, both the dielectric constant and linear electrooptic coefficient have been measured for only three crystals, i.e., $AgGaS_2$ [17], $CuGaS_2$, and $ZnGeP_2$ [18]. Our calculation will concentrate on these three compounds. The results will be compared with experimental measurements.

For the chalcopyrite crystals, only the nonzero summation $\Sigma\alpha_1\alpha_2\alpha_3$ of both bonds contributes to the electrooptic coefficient. Since these crystals have birefringence, we have $r_{41} = r_{52} \neq r_{63}$ with different dielectric constants $\varepsilon_1 = \varepsilon_2 \neq \varepsilon_3$ and, in general, $\Sigma\alpha_1^2 = \Sigma\alpha_2^2 \neq \Sigma\alpha_3^2$. However, we can assume the crystal is in perfect tetragonal coordination in the calculation of the summation of direction cosines without introducing significant error. Therefore, the value $\Sigma\alpha_1\alpha_2\alpha_3/\Sigma\alpha^2 = 0.57735$ will be used for both A—C and B—C bonds in all chalcopyrite crystals.

In the calculation of the electrooptic coefficient, the way to obtain the values of the physical parameters in (2.34) is explained in the following. N, $r_0$, $\Sigma\alpha_s'$, $k_s$, and $k_F$ (to calculate $\hbar\omega_p$) are obtained from the crystallographic data. $E_h$ is also easy to obtain from a knowledge of the bond length. We then determine $f_i$ and $e_c^*/e$ using the value of c. Finally, the bond susceptibility has to be determined. In the case of only one kind of bond, there is no problem in obtaining the bond susceptibility if we have the value of the crystal susceptibility and the knowledge of its geometrical factor. However, it is impossible to obtain the bond susceptibility of each bond separately by way of decomposition if the crystal has two kinds of bonds. A very novel method to obtain the bond susceptibility has been suggested to investigate the bond ionicity [8].

Let us consider the crystal $ZnGeP_2$. It can be seen as the superposition of two fictitious zinc-blende crystals: ZnP and GeP. The susceptibility is then the average of that of ZnP and GeP. We also note that Zn, Ga, and Ge are in the same row, and P and S are in the same row. Among the crystals composed of the atoms from these two rows, GaP and ZnS are well known. We assume that the empirical formula (2.8) is also applicable to the crystals composed of atoms from different rows. Then, with $\Delta Z(ZnS) = 4$ and $\Delta Z(GaP) = 2$, we can construct a straight line on the plot of $\chi^{-1}$ versus $(\Delta Z)^2$. By knowing $\Delta Z(ZnP) = 3$ and $\Delta Z(GeP) = 1$, the susceptibility of ZnP and GeP can be found easily on the line. Using this procedure, we can find the susceptibility of both bonds without knowing the susceptibility of the original crystal, $ZnGeP_2$. This is true only if all atoms are not bonding via the d-electrons. In the case of $AgGaS_2$ and $CuGaS_2$, the definition of $\Delta Z$ for Ag-S and Cu-S becomes ambiguous. Only the susceptibility of GaS can be obtained by the empirical formula (2.8). $\chi$(Ag–S) is then found by considering $\chi(AgGaS_2)$ is the average of $\chi$(Ag-S) and $\chi$(Ga–S).

The values of parameters and the calculation results are shown in Table VII for the three crystals $AgGaS_2$, $CuGaS_2$ and $ZnGeP_2$. The results are compared with experiment. The measurement of the coefficient is at 0.633μ for $AgGaS_2$, while at 3.39μ for $CuGaS_2$ and $ZnGeP_2$. Due to the dispersion of the optical permittivity, we use the value of $\varepsilon_3' = 6.50$

Table VII. The calculation of electrooptic coefficients $r_{63}$ and $r_{41}$ for the ternary chalcopyrite crystals. r's are in units of $10^{-12}$ m/V.

| Characteristics | | $AgGaS_2$ | $CuGaS_2$ | $ZnGeP_2$ |
|---|---|---|---|---|
| A—C[a] | $r_o$ | $1.28\mathring{A}$ | $1.1945\mathring{A}$ | $1.1945\mathring{A}$ |
| | $\chi$ | 3.68 | 4.43 | 5.84 |
| | $v_b$ | $12.45\mathring{A}^3$ | $9.763\mathring{A}^3$ | $10.414\mathring{A}^3$ |
| | $n_v$ | $0.1606\mathring{A}^{-3}$ | $0.2049\mathring{A}^{-3}$ | $0.1680\mathring{A}^{-3}$ |
| | $k_s$ | $2.012\mathring{A}^{-1}$ | $2.095\mathring{A}^{-1}$ | $2.027\mathring{A}^{-1}$ |
| | $E_h$ | 3.87 eV | 4.58 eV | 4.58 eV |
| | C | 9.87 eV | 9.95 eV | 4.08 eV |
| | $f_i$ | 0.867 | 0.825 | 0.442 |
| | f | -0.1868 | -0.2086 | -0.1391 |
| | $\hbar\omega_p$ | 21.71 eV | 24.52 eV | 15.23 eV |
| | $e_s^*/e$ | 0.91 | 0.81 | 0.67 |
| B—C[a] | $r_o$ | $1.14\mathring{A}$ | $1.162\mathring{A}$ | $1.162\mathring{A}$ |
| | $\chi$ | 5.84 | 6.07 | 10.55 |
| | $v_b$ | $8.80\mathring{A}^3$ | $8.987\mathring{A}^3$ | $9.586\mathring{A}^3$ |
| | $n_v$ | $0.2557\mathring{A}^{-3}$ | $0.2504\mathring{A}$ | $0.2347\mathring{A}^{-3}$ |
| | $k_s$ | $2.174\mathring{A}^{-1}$ | $2.166\mathring{A}^{-1}$ | $2.143\mathring{A}^{-1}$ |
| | $E_h$ | 5.147 eV | 4.91 eV | 4.91 eV |
| | C | 5.416 eV | 5.60 eV | 2.60 eV |
| | $f_i$ | 0.525 | 0.565 | 0.219 |
| | f | -0.1465 | -0.1451 | -0.0714 |
| | $\hbar\omega_p$ | 18.79 eV | 18.59 eV | 18.00 eV |
| | $e_s^*/e$ | 0.72 | 0.75 | 0.51 |

Table VII (continued)

| Characteristics | | | | ZnGeP$_2$ |
|---|---|---|---|---|
| | $e_c^*/e$ | 0.732 | 0.688 | 0.477 |
| | $\frac{\chi_A}{r_A} f_A + \frac{\chi_B}{r_B} f_B$ | $0.6438\text{Å}^{-1}$ | $0.7658\text{Å}^{-1}$ | $0.6643\text{Å}^{-1}$ |
| | $\Sigma\alpha_1\alpha_2\alpha_3/\Sigma\alpha^2$ | 0.57735 | 0.57735 | 0.57735 |
| | $\frac{\varepsilon_o}{Ne}$ [$10^{-12}$ $\frac{m\text{Å}}{V}$] | 47.0 | 41.5 | 44.3 |
| $r_{63}$ | $\varepsilon_{dc3}$ | $14^b$ | $10^c$ | $12^c$ |
| | $\frac{\varepsilon_{dc} - \varepsilon}{\varepsilon^2}$ | 0.192 | 0.096 | 0.037 |
| | $r^{ion}$ | +4.58 | +2.56 | +1.32 |
| | $r^{elec}$ | $-7.85\ ^d$ | $-1.55\ ^e$ | $-4.84\ ^f$ |
| | $r^{sum}$ | -3.27 | +1.01 | -3.52 |
| | $r^{exptl}$ | $3.0\ ^b$ | $+1.05\ ^c$ | $-0.97\ ^c$ |
| $r_{41}$ | $\varepsilon_{dc1}$ | $10^b$ | $9.3^c$ | $15^c$ |
| | $\frac{\varepsilon_{dc} - \varepsilon}{\varepsilon^2}$ | 0.096 | 0.078 | 0.074 |
| | $r^{ion}$ | +2.30 | +2.08 | +2.64 |
| | $r^{elec}$ | $-7.58^d$ | $-1.55^e$ | $-4.84^f$ |
| | $r^{sum}$ | -5.28 | +0.53 | -2.20 |
| | $r^{exptl}$ | $4.0^b$ | $+1.1^c$ | ? |

[a] R.W.G. Wyckoff, *Crystal Structures* (Interscience, New York, 1964), Vol. 2.

[b] V. M. Cound, P. H. Davies, K. F. Hulme, and P. Robertson, J. Phys. C **3**, L83 (1970).

Table VII (continued)

[c] E. H. Turner, E. Buehler, and H. Kasper, Phys. Rev. B9, 558 (1974).

[d] D. A. Kleinman, Phys. Rev. 128, 1761 (1962).

[e] G. D. Boyd, H. Kasper, and J. H. McFee, IEEE J. Quantum Electron. QE-7, 563 (1971).

[f] S. Bhagvantam, Crystal Symmetry and Physical Properties (Academic Press, New York, N.Y., 1966).

and $\varepsilon_1' = 6.25$ at $0.633\mu$ for $AgGaS_2$. The values of $r_{63}$ are in good agreement with experiment, while the results of $r_{41}$ are satisfactory. Although the sign of the coefficient is correct for $ZnGeP_2$, the magnitude is in poor agreement. Such a large discrepancy cannot be due to the uncertainty of the theory and the measurement. Since there is only one indirect measurement of $\varepsilon_{dc3}'$ for $ZnGeP_2$, we suggest rechecking this value experimentally. With respect to the theory, the assumption of the uniform displacement may not be suitable in the chalcopyrite compound, especially for $ZnGeP_2$. It is quite possible that the relative movements of Zn and Ge with respect to P have different magnitudes. However, we need more experimental data to check this assumption. At the same time, it is a challenge to find a method to account for the different displacements theoretically.

## 3.7 Other Ternary Compounds

The ternary compounds are of continuous interest due to the properties of simple structure, large nonlinearity, and wide range of transparency. There are many different chemical compositions in the ternary system. However, only those of $A^{II}B_2^{III}C_4^{VI}$ and $A_3^IB^{III}C_3^{VI}$ type have been investigated in some detail.

In the $A^{II}B_2^{III}C_4^{VI}$ type, A could be Zn, Cd, or Hg, B could be Al, Ga, or In. C could be S, Se, or Te. Most of these compounds are in the point group $\overline{4}$ or $\overline{4}2m$ (Figure 3.6), except $ZnIn_2S_4$, which is 3m [19]. It is obvious that the crystals of $\overline{4}$ or $\overline{4}2m$ (defect chalcopyrite) have similar structures to the zinc-blende or chalcopyrite compounds, except

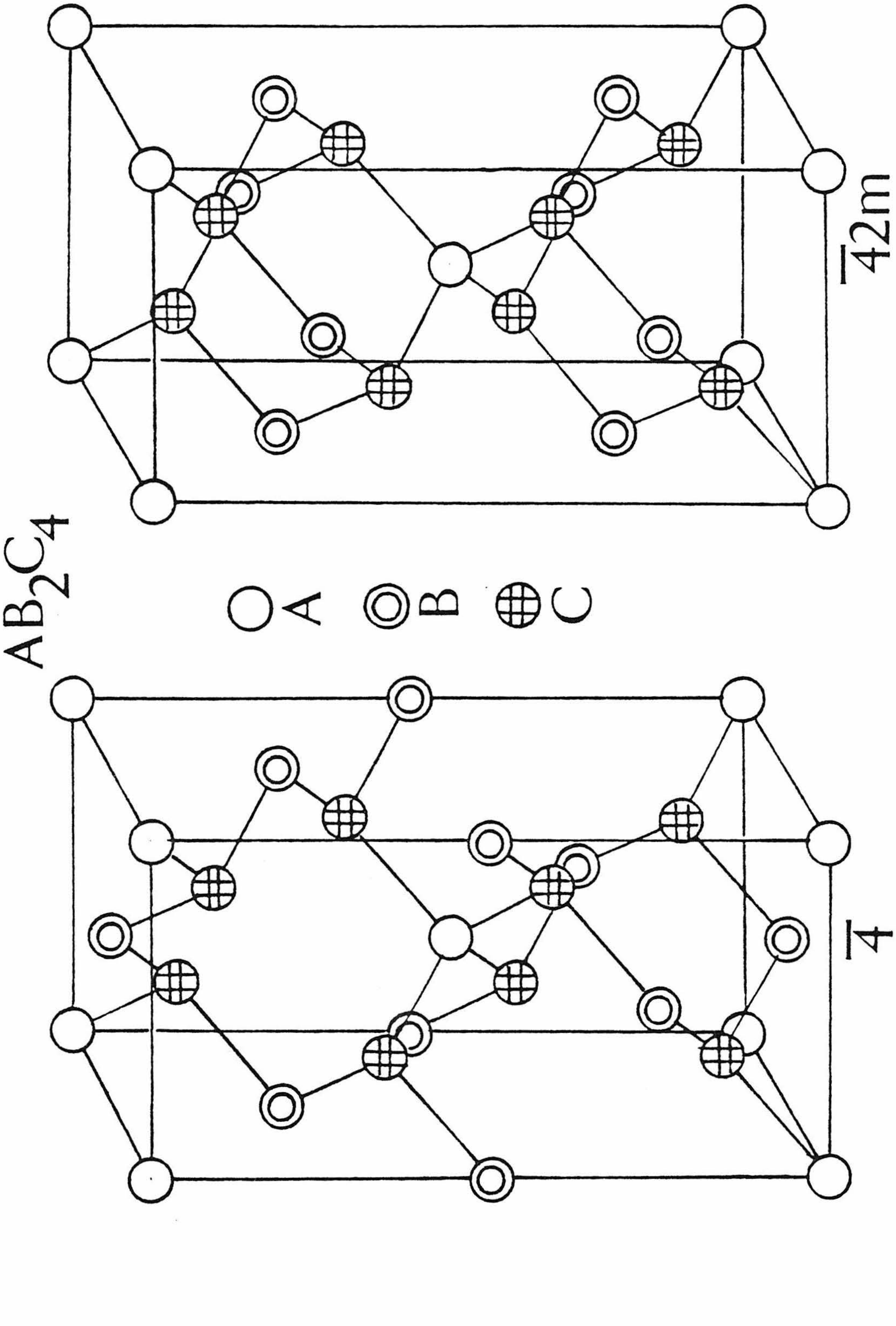

Figure 3-6 : Two typical structures of the ternary compound ($AB_2C_4$).

that there are 25 percent vacancies of atom positions. It is shown in Figure 3.6 that the defect chalcopyrite has only 24 bonds per unit cell while chalcopyrite has 32. From the principle of the susceptibility composition, $\chi$ of the defect chalcopyrite should be in general smaller than $\chi$ of chalcopyrite. This fact is seen clearly by comparing the susceptibility of $CdGa_2S_4$ and $AgGaS_2$, $\chi[CdGa_2S_4] = 4.25$, while $\chi[AgGaS_2] = 4.76$. A detailed crystallographic study of the defect chalcopyrite structure has been done more than twenty years ago [20]. Although the structures of $\bar{4}$ and $\bar{4}2m$ are different, the summations of direction cosines $\Sigma\alpha_i$, $\Sigma\alpha_i\alpha_j$, and $\Sigma\alpha_i\alpha_j\alpha_k$ are the same for both structures if considered in perfect tetragonal coordination.

The compound $ZnIn_2S_4$ has a layer structure of the point group 3m with three formulas per unit cell [21]. The lattice constants are a = 3.85 and c = 37.0Å. The two indium ions have different coordination numbers. $In_O$ occupies slightly compressed octahedral void, where $In_T$ and Zn occupy enlarged tetrahedral spaces.

Although the structure information for $A^{II}B_2^{III}C_4^{VI}$ crystals is complete, no measurement of the dielectric constant has been carried out. The calculation of the ionic part of the electrooptic coefficient is still impossible.

The other type of interesting ternary compound is $A_3^IB^{III}C_3^{VI}$. Three crystals of this type, $Ag_3AsS_3$, $Ag_3SbS_3$, and $Tl_3AsSe_3$ have been studied somewhat in detail for their structures, optical properties and nonlinearities. However, only the dielectric constant and electrooptic coefficients of $Ag_3AsS_3$ have been measured.

$Ag_3AsS_3$ belongs to the point group 3m with six formulas per unit cell. The lattice constants of its hexagonal unit cell are a = 10.80Å and c = 8.69Å. The projection of atoms on the x-y plane is shown in Figure 3.7. The bond distance of an As atom to the nearest three S atoms is 2.293Å. Every S atom is bonded to one As atom (2.293Å) and two Ag atoms (2.44Å). In all, there are 54 bonds in a unit cell, where As—S has 18 and Ag—S has 36. The summations of direction cosines for both bonds are listed in Table VIII. From the nonzero summations of $\Sigma\alpha_i\alpha_j\alpha_k$, we know the nonzero coefficients of $Ag_3AsS_3$ (proustite) are $r_{33}$, $r_{22} = -r_{12} = -r_{61}$, $r_{13} = r_{23}$, and $r_{51} = r_{42}$.

As in the chalcopyrite compounds, we can obtain the values of most physical parameters for both bonds from the structural information. However, we cannot find the bond susceptibility in the same way, because there is no equivalent structure in 3m which has only two atoms in the formula. So we try to decompose the susceptibility into the individual bonds by using the bond-additive principle:

$$\begin{aligned} \chi_{Ag-S}(\Sigma\alpha_{1,2}^2)_{Ag-S} + \chi_{As-S}(\Sigma\alpha_{1,2}^2)_{As-S} &= \chi_o \\ \chi_{Ag-S}(\Sigma\alpha_3^2)_{Ag-S} + \chi_{As-S}(\Sigma\alpha_3^2)_{As-S} &= \chi_e \end{aligned} \qquad (3.9)$$

From the measured ordinary and extraordinary susceptibilities $\chi_o$ and $\chi_e$, we can solve for the contribution of the individual bonds $\chi_{Ag-S}$ and $\chi_{As-S}$. The susceptibility in Table VIII is taken as if all the bonds in the crystal are occupied by the same kind of bonds. The averaged susceptibility is calculated from $\chi_{avg} = (2\chi_o + \chi_e)/3$. The values of all physical parameters are shown in Table VIII, including the effective ionic charge.

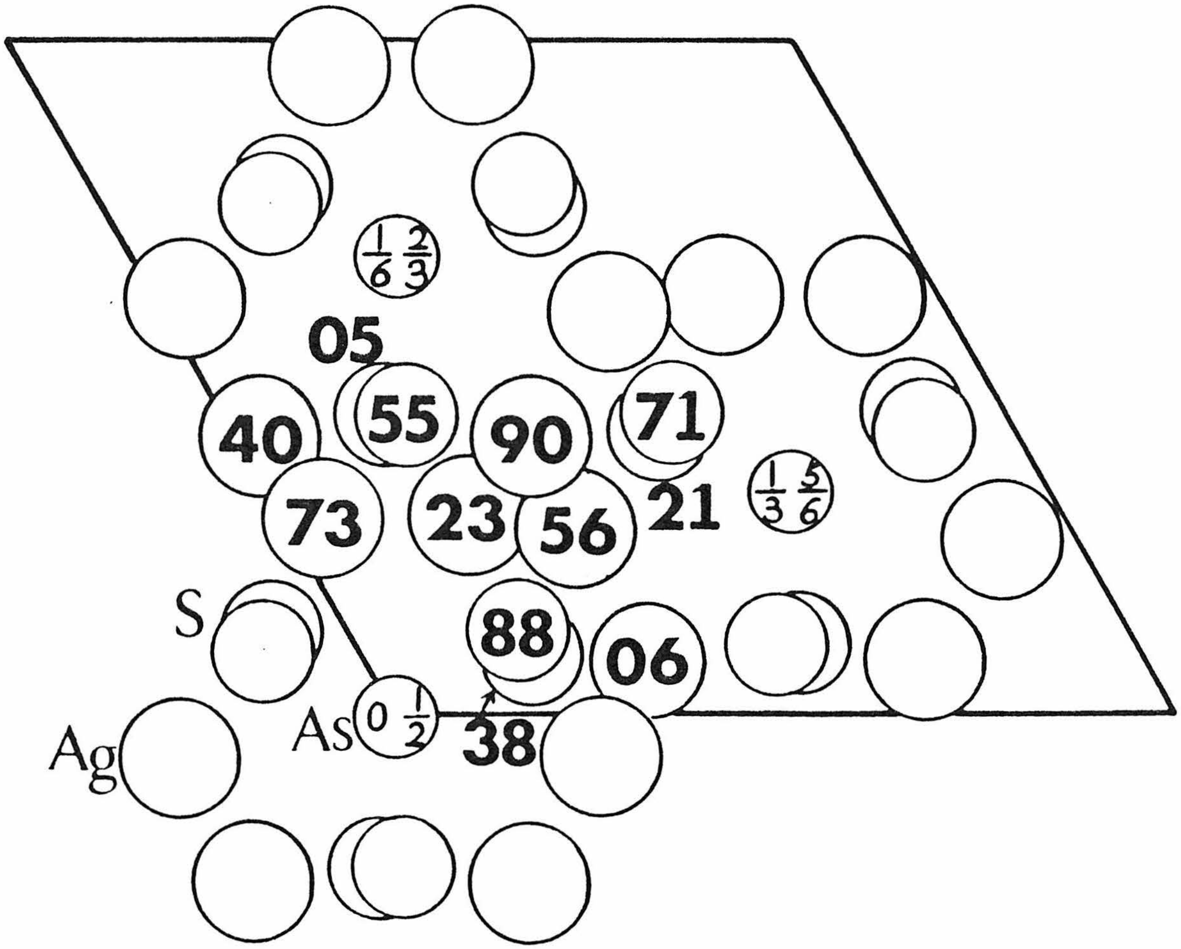

Figure 3-7 : Proustite ($Ag_3AsS_3$) structure.

Table VIII. Values of physical parameters used in the calculation of the electrooptic coefficient of proustite ($Ag_3AsS_3$)

| Characteristics | As—S | Ag—S |
|---|---|---|
| $\Sigma\alpha_3$ | 7.844 | 2.137 |
| $\Sigma\alpha_1^2 = \Sigma\alpha_2^2$ | 7.291 | 11.589 |
| $\Sigma\alpha_3$ | 3.418 | 12.811 |
| $\Sigma\alpha_3^3$ | -1.490 | -2.266 |
| $\Sigma\alpha_2^3 = -\Sigma\alpha_2\alpha_1^2$ | 3.190 | 2.005 |
| $\Sigma\alpha_3\alpha_1^2 = \Sigma\alpha_3\alpha_2^2$ | -3.177 | 0.089 |
| $\chi_{avg}$ | 6.9832 | 5.6383 |
| $r_o$ | 2.293Å | 2.440Å |
| $v_b$ | 14.305Å$^3$ | 17.231Å$^3$ |
| $n_v$ | 0.2097Å$^{-3}$ | 0.1741Å$^{-3}$ |
| $k_s$ | 2.4117Å$^{-1}$ | 2.4877Å$^{-1}$ |
| $E_h$ | 5.0735 eV | 4.3501 eV |
| $C$ | 3.9107 eV | 5.2620 eV |
| $f_i$ | 0.3727 | 0.5940 |
| $f$ | -0.1222 | -0.1603 |
| $\hbar\omega_p$ | 17.0175 eV | 15.5051 eV |
| $e_s^*/e$ | 0.5745 | 2.0362 |
| $e_c^*/e$ | 1.1158 | |
| $\varepsilon_o/Ne$ | 80.96 mÅ/V | |

The calculated values of the electrooptic coefficients are shown in Table IX. The sign of coefficient depends on the definition of the crystal polarity. Since all measurements of SHG coefficients and electrooptic coefficients are in absolute value, we compare the values of $r^{ion}$, $r^{elec}$, and r to deduce the possible sign for each quantity. The only thing we know presumably about the sign is the relative sign between the ionic parts of the coefficients. The estimate of signs is entered in parentheses. The prediction is in good agreement with experiment. The comparison is on the smaller coefficients where both contributions are comparable. We also predict the value of $r_{33}$ in spite of the invalidity of $d_{33}$, since the ionic effect dominates in the electrooptic response.

Table IX. Comparison of the theoretical prediction for the electrooptic coefficients of $Ag_3AsS_3$ with experiment. All coefficients are in units of $10^{-12}$m/V. For the coefficient $r_{ijk}$, the first two rows are calculated as $\sum \frac{\chi}{r} [f\alpha_i\alpha_j\alpha_k + \frac{1}{2} (\alpha_i\delta_{jk} + \alpha_j\delta_{ik})]$ and $\frac{\varepsilon'_{dck} - \varepsilon'_k}{\varepsilon'_i\varepsilon'_j}$.

| $r'_s$ | $r_{22}$ | $r_{13}$ | $r_{51}$ | $r_{33}$ |
|---|---|---|---|---|
| $\frac{\chi}{r} [f\alpha^3 + \alpha]$ | -0.2144 | 0.1277 | 1.7295 | 2.1375 |
| $\frac{\varepsilon'_{dc} - \varepsilon'}{\varepsilon\cdot\varepsilon'}$ | 0.1310 | 0.1504 | 0.1592 | 0.2220 |
| $r^{ionic}$ | (+) 2.04 | (-) 1.39 | (-)19.98 | (-)34.43 |
| $r^{elec}$ [a] | (-) 1.10 | (-) 0.97 | (-) 0.97 | ? |
| $r^{sum}$ | (+) 0.94 | (-) 2.36 | (-)20.95 | ? |
| $r^{exptl}$ [b] | (+) 1.05 | (-) 2.54 | ? | ? |

[a] K. F. Hulme, O'Jones, P. H. Davis, and M. V. Hobden, Appl. Phys. Lett. 10, 133 (1967);
D. M. Boggett and A. F. Gibson, Phys. Lett. 28A, 33 (1968).

[b] J. Warner, Brit. J. Appl. Phys. (J. Phys. D), Ser. 2, 1, 949 (1968).

## CHAPTER 3 - REFERENCES

1. P. Lawaetz, Phys. Rev. B 5, 4039 (1972).
2. J. A. Van Vechten, Phys. Rev. 182, 891 (1969); 187, 1007 (1969).
3. J. C. Phillips, Rev. Mod. Phys. 42, 317 (1970).
4. W. L. Faust and C. H. Henry, Phys. Rev. Lett. 17, 1265 (1966).
5. I. P. Kaminow and E. H. Turner, Phys. Rev. B 5, 1564 (1972).
6. A. Gentile, private communication.
7. R. C. Miller, Appl. Phys. Lett. 15, 17 (1964).
8. B. F. Levine, J. Chem. Phys. 59, 1463 (1973).
9. V. G. Zubov, M. M. Firsova, and T. M. Glushkova, Soviet Phys. Cryst. 9, 729 (1965).
10. S. C. Abrahams, J. M. Reddy, and J. L. Bernstein, J. Phys. Chem Solids 27, 997 (1966).
11. S. C. Abrahams, W. C. Hamilton, and A. Segueira, J. Phys. Chem. Solids 28, 1693 (1967).
12. C. R. Jeggo and G. D. Boyd, J. Appl. Phys. 41, 2741 (1970).
13. R. C. Miller and A. Savage, Appl. Phys. Lett. 9, 169 (1966).
14. E. H. Turner, Th A13, Opt. Soc. Amer., San Francisco, Oct. 20, 1966.
15. J. P. Van der Ziel and N. Bloembergen, Phys. Rev. 135, A1662 (1964).
16. J. L. Shay and Wernick, Ternary Chalcopyrite Semiconductors (Pergamon Press, New York, 1975).
17. V. M. Cound, P. H. Davies, K. F. Hulme, and D. Robertson, J. Phys. C 3, L83 (1970).
18. E. H. Turner, E. Buehler, and H. Kasper, Phys. Rev.B 9, 558 (1974).

19. L. I. Berger and F. C. Prochukhan, Ternary Diamond-like Semiconductors (Consultants Bureau, New York, 1969).

20. H. Hahn, G. Frank, W. Klingler, A. D. Störger, and G. Störger, Z. Anorg Allgem. Chem. 279, 241 (1955).

21. F. Lappe, A. Niggli, R. Nitsche, and J. G. White, Zeit. Krist. 117, 146 (1962).

# Chapter 4
# CONCLUSION

The electrooptic effect plays an important role in light modulation and laser control systems. Seeking out new materials with higher electrooptic response has been a major effort joined by engineers and scientists. The theory just given makes it possible to predict the electrooptic coefficient of crystals.

Starting from the one-gap model for the semiconductor, we relate the optical permittivity to the energy gap. The energy gap is then modeled by the decomposition of the interaction potential of electrons in the periodic lattice into symmetric and antisymmetric parts. The empirical expressions for the symmetric and antisymmetric contributions are obtained by observing the dependence of the susceptibility on the row number of atoms and the difference of the numbers of valence electrons. Instead of an average effect over the whole crystal, the microscopic interpretation of the two contributions to the energy gap is carried out using the principle of bond additivity. The idea of the geometrical composition of the bond susceptibility is the key point in generalizing the theory to more complex crystals.

Assuming the ionic part of the dielectric constant is due to the relative displacement of positive and negative ions, we relate the change of the susceptibility to the distortion of the crystal lattice. Such relative displacement results in bond rotation and bond stretch. The effect due to the bond stretch is found from a knowledge of the dependence of the bond susceptibility on atomic radii. The final expression for the ionic part of the electrooptic coefficient is obtained in (2.34).

The theory has been applied to the calculation of the electrooptic coefficient of many crystals with different structures. The crystals used in the comparison possessed tetragonal and octahedral coordinations, single-bond, one kind of bond with different bond lengths, two different kinds of bonds. In general, we find that the theoretical calculations are in good agreement with experiment. This proves the versatile applicability of the theory to various crystals.

Referring to the key result (2.34), we point out that the structural information is more important than the detailed electronic information in seeking better electrooptic materials. The ionicity factor, f, is a new factor in the theory and is very small for any bond, so the term $\Sigma\alpha_i$ usually dominates over the term f $\Sigma\alpha_i\alpha_j\alpha_k$ if $\Sigma\alpha_i$ is not equal to zero. As a consequence, the coefficient of a crystal with nonzero $\Sigma\alpha_i$ should have a higher value. This is fully demonstrated by the fact that $r_{33}$ and $r_{51}$ in the 3m crystals like $LiNbO_3$, $LiTaO_3$, and $Ag_3AsS_3$ is an order of magnitude larger than in most other crystals. However, not all 3m crystals have higher coefficients. For example, $ZnIn_2S_4$ has zero value of $\Sigma\alpha_i$ and is not expected to have a large electrooptic coefficient. But a distorted octahedral structure is surely a promising mechanism for a larger electrooptic response. Such octahedral structure usually occurs in the point group with 3-fold or 6-fold axis. Therefore, the crystals belonging to those point groups are promising candidates for electrooptic applications.

Although the theory has proved successful in predicting the electrooptic constants of diatomic and ternary compounds, it remains on somewhat shaky ground in the case of quarternary crystals, where more than two kinds of bonds have to be taken into account. It is very important to point out some basic assumptions and crucial criticisms inherent in the theory.

(1) The relation between the dielectric constant and the ion-displacement depends on the bond-strength and is not completely known. In the theory, we accept it as a parameter and use its measured value in the calculation. It will be very important to know the dependence of $\varepsilon_{dc}'$ on the structure and atom information like the crystal susceptibility, $\chi$.

(2) The displacement of ions is assumed to be uniform. This is not true in the practical situation. The understanding of the displacements for different ions is still beyond the scope of the theory, but becomes important in more complex crystals.

(3) The crystal susceptibility is considered as the geometrical composition of only the bond susceptibility along the bond direction. The transverse polarization has been completely neglected. This is justifiable for most crystals, but not for the highly anisotropic bond.

(4) In the ideal situation, it is expected to calculate the bond susceptibility just from knowledge of the crystal structure. However, we obtain the bond susceptibility for chalcopyrite and 3m crystals from the measured crystal susceptibility. The reason is that we find the "b" value used in the calculation of c cannot be determined theoretically with the same accuracy as in the case of diatomic crystals. In diatomic

crystals, b is related to a high degree of accuracy to the coordination number. Qualitatively, b in the ternary compounds still follows the relation but deviates highly from the predicted value. So we would rather rely on the principle of bond-addivity to find the bond susceptibility. A generalized method to obtain b becomes very important to the further improvement of the theory.

In conclusion, the theory has been applied successfully to the understanding of the electrooptic effect in the diatomic and ternary compounds. However, considerable further improvement is needed to meet the challenge of complex crystals and the task of seeking out new electrooptic crystals.