Paper C2 lecture notes: Errors. Updated 19/1/2022

Optics 2

1. The first line of eqn (2.53) should read,

$$U(X,u+v)=\frac{1}{\mathrm{i}v\lambda}\exp\left[\mathrm{i}k(u+v)\right]\exp\left(\mathrm{i}k\frac{X^2}{2v}\right)\int U(x,u)\exp\left[\left(\mathrm{i}k\frac{x^2}{2v}\right)\right]\exp\left(-\mathrm{i}k\frac{Xx}{v}\right)\mathrm{d}x$$

Optics 4

1. Equation (4.7) should read:

$$\exp\left(\mathrm{i}kL_\mathrm{c}\right)K_{\mathrm{free-space}}(X,Y,x,y)=\frac{1}{\mathrm{i}\lambda L_\mathrm{c}}\exp\left(\mathrm{i}kr_{01}\right),$$

   i.e. the argument of the complex exponential should be kL_c not $2kL_\mathrm{c}$

2. In the text immediately before eqn (4.9) the adjoint functions should be written as $u^\dagger_{nm}(x,y)$ (not $E^\dagger_{nm}(x,y)$).

Optics 7

1. Equation (7.51) should read:

$$\frac{I_\mathrm{trans}}{I_\mathrm{in}}=\frac{1}{2}\left\{1-\cos\left[\alpha V(t)+\pi/2\right]\right\}=\frac{1}{2}\left\{1+\sin\left[\alpha V(t)\right]\right\}\approx\frac{1}{2}\left[1+\alpha V(t)\right]$$

2. In eqns (8.19) to (8.23) χ_{kij} should be replaced by χ_{ijk}.

3. In Fig. 8.3 the field oscillating at 2ω is the e-wave, and so should be labelled as $E_\mathrm{e}(2\omega)$.

Paper C2: Laser Science and Quantum Information Processing. Laser Physics

Prof. Simon Hooker

Academic year 2021–2022

Contents

1 Basic Laser Physics **1**

1.1 Introduction . . . 1
1.2 Recommended texts for Lasers and Optics parts of the course . . . 1
1.3 Lineshapes . . . 2
1.3.1 Homogeneous broadening . . . 2
1.3.2 Inhomogeneous broadening . . . 3
1.3.3 Comparison of homogeneous and inhomogeneous transitions . . . 5
1.4 The optical gain cross-section . . . 7
1.4.1 Laser rate equations . . . 8
1.5 Gain saturation . . . 8
1.5.1 homogeneous broadening . . . 9
1.5.2 Inhomogeneously broadened transitions . . . 11
1.5.3 Summary . . . 13
1.6 Threshold behaviour . . . 13
1.6.1 Homogeneously broadened laser transitions . . . 13
1.6.2 Inhomogeneously broadened systems . . . 15

2 Solid State Laser Materials **16**

2.1 General considerations . . . 16
2.1.1 Radiative transitions . . . 18
2.1.2 Non-radiative transitions . . . 19
2.1.3 Line broadening . . . 19
2.2 Trivalent rare earths, $4f^n - 4f^n$ transitions . . . 20
2.2.1 Energy level structure . . . 20
2.2.2 Transition linewidth . . . 21
2.2.3 Nd:YAG laser . . . 21
2.2.4 Other crystalline hosts . . . 26
2.2.5 Nd:Glass laser . . . 26
2.2.6 Erbium lasers . . . 27
2.3 Trivalent iron group, $3d^n - 3d^n$ transitions . . . 29
2.3.1 Energy level structure . . . 29
2.3.2 The ruby laser . . . 35
2.3.3 Alexandrite laser . . . 38

2.3.4 Ti:sapphire . . . 41
2.3.5 Host materials . . . 43

3 Q-Switching **46**
3.1 Laser Spiking and relaxation oscillations . . . 46
3.1.1 Rate equations . . . 48
3.1.2 Cavity lifetime . . . 49
3.1.3 Numerical analysis of laser spiking . . . 49
3.1.4 Analysis of relaxation oscillations . . . 50
3.2 Q-switching . . . 52
3.2.1 Techniques for Q-switching . . . 53
3.2.2 Analysis of Q-switching . . . 57

4 Modelocking **62**
4.1 Introduction . . . 62
4.2 General ideas . . . 62
4.2.1 Simple analysis . . . 63
4.2.2 Further general comments . . . 66
4.3 Active modelocking techniques . . . 66
4.3.1 AM modelocking . . . 66
4.3.2 Synchronous pumping . . . 72
4.3.3 Pulse duration of actively modelocked, homogeneously broadened lasers . . . 73
4.4 Passive modelocking techniques . . . 73
4.4.1 Saturable absorbers . . . 73
4.4.2 Kerr lens modelocking . . . 75
4.5 Examples of modelocked lasers . . . 77
4.5.1 CW modelocking . . . 77
4.5.2 Pulsed modelocking . . . 77

5 Ultrafast Lasers **79**
5.1 Propagation in dispersive media . . . 79
5.1.1 Propagation through an arbitrary system . . . 80
5.1.2 Propagation of Gaussian pulses . . . 82
5.1.3 Nonlinear effects: self-phase-modulation and the B-integral . . . 85
5.2 Dispersion control . . . 86
5.2.1 The grating pair . . . 88
5.2.2 The prism pair . . . 89
5.2.3 Introduction of positive GDD . . . 90
5.2.4 Chirped mirrors . . . 91
5.3 TW and PW laser systems . . . 92
5.3.1 Regenerative and multipass amplifiers . . . 93
5.3.2 Multipass amplifiers . . . 95
5.3.3 Low-energy, TW laser systems . . . 95
5.3.4 High-energy, PW laser system . . . 96

5.4 Applications of ultrafast lasers . . . 97

6 Semiconductor & Fibre lasers 99
6.1 Review of semiconductor physics . . . 99
6.1.1 Distribution over energy levels . . . 101
6.1.2 Doped semiconductors . . . 102
6.2 Radiative transitions in semiconductors . . . 104
6.2.1 Joint density of states . . . 105
6.2.2 Density of possible transitions . . . 106
6.2.3 The gain coefficient . . . 107
6.2.4 Condition for gain . . . 107
6.2.5 Generating gain: the injection density . . . 108
6.3 Homojunction diode lasers . . . 109
6.3.1 Structure of homojunction diode lasers . . . 111
6.4 Double heterostructure lasers . . . 112
6.4.1 Forming heterostructures: Material properties . . . 113
6.5 Geometries of heterostructure diode lasers . . . 114
6.6 Output characteristics of diode lasers . . . 115
6.7 Fibre lasers . . . 117
6.7.1 Pumping configurations and fibre materials . . . 118
6.7.2 An example . . . 118
6.7.3 Erbium-doped fibre lasers . . . 119

7 Precision frequency control 121
7.1 Single-mode operation . . . 121
7.1.1 Short cavity . . . 121
7.1.2 Intra-cavity etalon . . . 122
7.1.3 Ring resonators . . . 123
7.2 Output linewidth . . . 123
7.2.1 Theoretical limit . . . 123
7.2.2 Practical limitations . . . 127
7.3 Frequency locking . . . 127
7.3.1 Locking to atomic or molecular transitions . . . 130
7.3.2 Locking to an external cavity . . . 131

Lecture 1

Basic Laser Physics

1.1 Introduction

In these lectures we build on the basic laser physics covered last year in Paper B3. All of the material covered in those lectures will be required for this course, and hence you should revise all of that work.

I would be very grateful if you were to bring to my attention any errors or unclear passages (email: simon.hooker@physics.ox.ac.uk).

1.2 Recommended texts for Lasers and Optics parts of the course

You should read beyond these lectures notes. The following books are at about the right level, and contain sections on almost everything that we will cover:

Books primarily on lasers

- "Laser Physics," Simon Hooker & Colin Webb, Oxford University Press.
- "Principles of Lasers," Orazio Svelto, fourth edition, Plenum Press.
- "Lasers and Electro-Optics: Fundamentals and Engineering," Christopher Davies Cambridge University Press.
- "Laser Fundamentals," William Silfvast, Cambridge University Press.

Books primarily on optics

- "Modern Classical Optics," Geoffrey Brooker, Oxford University Press.
- "Optics," Eugene Hect, Pearson

More advanced texts

You may want to consult the following, more advanced texts:

- "Lasers," Anthony Siegman, University Science Books. This book approaches laser physics from more of an engineering viewpoint, but contains a wealth of information. It is particularly strong on cavity modes and Gaussian beams.

- "Principles of Optics," Max Born & Emil Wolf, Cambridge University Press. The bible of classical optics.
- "Introduction to Fourier Optics," Joseph Goodman, Roberts & Company. A more advanced text on the use of Fourier techniques for calculating beam propagation.

1.3 Lineshapes

As we saw on the B3 course, a variety of mechanisms can lead to broadening of the frequency spectrum of an optical transition. These mechanisms can be classified as being **homogeneous** or **inhomogeneous**.

The key property of homogeneous broadening mechanisms is that the broadening mechanism affects all atoms in the sample in the same way. As a consequence all atoms in the same energy level interact with radiation of frequency ω with the same strength. In contrast, inhomogeneous broadening arises from processes which alter the frequency of an atomic transition by an amount that depends on some property of the atom, such as its velocity or position. In this case, the atoms may be considered to belong to different **classes**, depending on the frequency the frequency of the transition. The strength with which an atom will interact with radiation of frequency ω then depends on the class to which it belongs.

1.3.1 Homogeneous broadening

Homogeneous line broadening was discussed in detail as part of the B3 course. Examples include: lifetime, or natural, broadening; pressure broadening; and phonon broadening. In each case it was found that the spontaneous emission from a sample of atoms in some energy level E_2 on a transition to a lower level E_1 gave rise to a spectrum proportional to $g_{\rm H}(\omega-\omega_0)$, where ω_0 is the central frequency of the transition, $\hbar\omega_0 = E_2 - E_1$. The function $g_{\rm H}(\omega-\omega_0)$ is normalized[1] such that $\int_0^\infty g_{\rm H}(\omega-\omega_0)d\omega = 1$. For example, in the case of lifetime broadening we find that $g_{\rm H}(\omega-\omega_0)$ is a Lorentzian function:

$$g_{\rm H}(\omega-\omega_0) = \frac{1}{\pi}\frac{\gamma/2}{(\omega-\omega_0)^2+(\gamma/2)^2}, \tag{1.1}$$

where γ is the full-width at half maximum of the lineshape. This is equal to the sum of the contributions from the upper and lower levels:

$$\gamma = \frac{1}{\tau_2^{\rm rad}} + \frac{1}{\tau_1^{\rm rad}} \tag{1.2}$$

where $\tau_i^{\rm rad}$ is the radiative lifetime of level i.

[1] Actually the function $g_{\rm H}(x)$ is normalized such that $\int_{-\infty}^{\infty} g_{\rm H}(x)dx = 1$. In eqn (1.1) the argument of $g_{\rm H}(x)$ extends only to $-\omega_0$. However, since the linewidth is always very small compared to ω_0, extending the lower limit to $-\infty$ will have negligible effect on the integral.

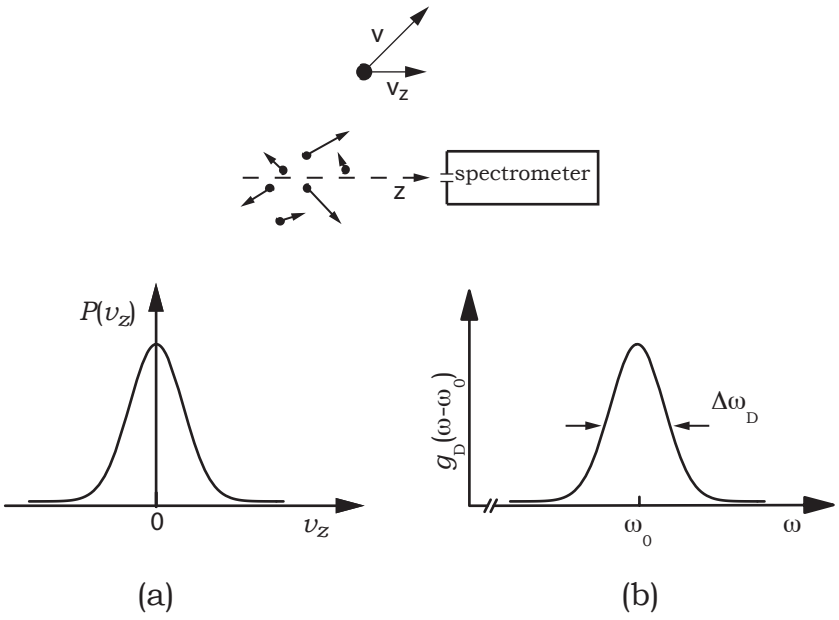


Figure 1.1: The formation of an inhomogeneously broadened spectral line by Doppler broadened gas atoms. Radiation emitted by the atoms in the z-direction is detected by the spectrometer. The distribution of atomic velocities in the z-direction (a) leads, via the Doppler effect, to a distribution of frequencies measured by the spectrometer, and hence broadening of the spectral line (b).

1.3.2 Inhomogeneous broadening

Inhomogeneous broadening mechanisms cause the transition frequency of different atoms to be shifted by different amounts. Such broadening mechanisms are said to be **inhomogeneous**.

Inhomogeneous broadening will always be associated with some homogeneous broadening since natural broadening is always present, and the atoms may also be subject to other homogeneous broadening mechanisms such as phonon broadening. The transition lineshape observed from a sample subjected to both homogeneous and inhomogeneous broadening processes is given by the convolution of the homogeneous and inhomogeneous lineshapes. It should be clear, however, that when the frequency width of the inhomogeneous distribution is much greater than that of the homogeneous distribution the observed transition lineshape will simply be the inhomogeneous lineshape.

Doppler broadening

For gas lasers the most important line broadening mechanism is usually **Doppler broadening**, which arises from a combination of the Doppler effect and the thermal motion of the atoms: the observed frequency ω emitted by atoms moving towards the observer with velocity v_z will be $\omega - \omega_0 = (v_z/\mathrm{c})\omega_0$, and hence any distribution of velocities will give rise to a distribution of observed frequencies. If the distribution is Maxwellian then the proportion $P(v_z)dv_z$ of atoms with velocities in the range v_z to

$v_z + dv_z$ is given by :

$$P(v_z)dv_z = \sqrt{\frac{M}{2\pi k_B T}} \exp\left[-\frac{Mv_z^2}{2k_{\rm B}T}\right] \, dv_z, \tag{1.3}$$

where M is the mass of each atom, and $k_{\rm B}$ the Boltzmann constant. The observed lineshape $g_D(\omega-\omega_0)$ is then found by making the substitution $v_z/c = (\omega - \omega_0)/\omega_0$, to yield

$$g_{\rm D}(\omega - \omega_0)d\omega = \sqrt{\frac{M}{2\pi k_{\rm B}T}} \exp\left[\frac{-Mc^2}{2k_{\rm B}T}\left(\frac{\omega-\omega_0}{\omega_0}\right)^2\right] \, \frac{c}{\omega_0}d\omega. \tag{1.4}$$

This is a Gaussian distribution with a full-width at half maximum, in this context known as the **Doppler width** $\Delta\omega_D$, of

$$\Delta\omega_D = \sqrt{8\ln 2}\sqrt{\frac{k_B T}{Mc^2}}\omega_0. \tag{1.5}$$

It is often convenient to re-write the lineshape in terms of $\Delta\omega_D$, *i.e.*

$$g_{\rm D}(\omega - \omega_0) = \frac{2}{\Delta\omega_D}\sqrt{\frac{\ln 2}{\pi}} \exp\left\{-\left[\frac{\omega-\omega_0}{(\Delta\omega_D/2)}\right]^2 \ln 2\right\}. \tag{1.6}$$

This lineshape is normalized so that $\int_{-\infty}^{\infty} g_{\rm D}(x)dx = 1$.

Broadening in amorphous solids

In a number of important laser systems the lasing species are ions doped into a solid. If the solid is a good quality crystal, the laser transition will be broadened homogeneously by natural and phonon broadening. However if the solid is non-uniform, ions in different regions will experience different local environments. Of particular importance is the local value of the strain of the crystal lattice since this affects the electric field experienced by the active ion, which in turn affects the energy levels of the ion through the Stark effect. Other aspects of the local environment which can affect the transition frequencies of an active ion are the presence of impurity ions, or variations in the orientation of the crystal lattice. Since such effects change the centre frequency of the active ions according to their location in the medium, this contribution to the broadening of the frequency response of a macroscopic sample is inhomogeneous.

These effects are particularly important for ions doped in glasses, such as the Nd^{3+} ions in the Nd:glass laser, since in a glassy material the local environment varies significantly with position, leading to substantial inhomogeneous broadening. Very often the distribution of centre frequencies is found to follow a Normal, *i.e.* Gaussian, distribution, in which case the transition lineshape is also Gaussian.

Treatment of inhomogeneous broadening

We can develop a mathematical treatment of inhomogeneous broadening by considering that atoms with different centre frequencies behave as if they were different 'types' of atom. The corollary of this is that atoms with similar centre frequencies are of the same type, or **class**. The frequency range over which each class can interact appreciably is determined by the (smaller) degree of homogeneous broadening experienced by the class. For example, the strength with which the class of atoms with centre frequency ω_c interacts with radiation of frequency ω is determined by the homogeneous optical cross-section $\sigma_{21}(\omega - \omega_c)$. Note that the strength with which each class of atoms interacts with radiation of frequency ω depends on the detuning of ω from ω_c, not on the detuning from ω_0, the average centre frequency of the entire inhomogeneous lineshape.

In treating the interaction of inhomogeneously broadened atoms with radiation we can treat each class separately. We can define the number density of atoms in the upper and lower levels with centre frequencies between ω_c and $\omega_c + \delta\omega_c$ as $\Delta N_2(\omega_c)\delta\omega_c$ and $\Delta N_1(\omega_c)\delta\omega_c$ respectively. For example, suppose inhomogeneously broadened atoms interact with radiation of total intensity I in a narrow band near ω_L. The rate equation for atoms in the upper level would then be of the form:

$$\frac{d\Delta N_2(\omega_c)}{dt} = R_2(\omega_c) - \Delta N^*(\omega_c)\sigma_{21}(\omega_L - \omega_c)\frac{I}{\hbar\omega_L} - \frac{\Delta N_2(\omega_c)}{\tau_2} + \ldots, \tag{1.7}$$

where $R_2(\omega_c)$ is the pump rate of atoms of class ω_c into the upper level, $\Delta N^*(\omega_c) = N_2(\omega_c) - (g_2/g_1)N_1(\omega_c)$ is the population inversion density for atoms with a centre frequency ω_c, and τ_2 is the fluorescence lifetime of the upper level (assumed to be the same for all classes). In principle the rate equations for each class can then be solved separately, and the results integrated over ω_c to give the total population densities in the upper and lower levels.

1.3.3 Comparison of homogeneously and inhomogeneously broadened laser transitions

Imagine passing a probe beam of narrow-band, tunable radiation through a sample of atoms, as shown schematically in Figure 1.2. If we tune the frequency of the probe beam in the region around one of the absorption transitions of the atoms, we can record the fluorescence spectrum — that is, the spectrum of radiation emitted spontaneously from the excited level[2].

The behaviour of homogeneously and inhomogeneously broadened transitions is compared in Table 1.1.

[2]Note that it is important to ensure that we record the spectrum of spontaneous emission rather than scattered radiation. In principle this can be achieved by using a pulsed probe beam and turning the detector on — 'gating' it — a short time after the passage of the probe pulse. Scattered radiation, which is prompt and (in the absence of nonlinear effects) has the *same* frequency as the incident radiation, is not then detected.

Homogeneous transition	**Inhomogeneous transition**
For a given frequency ω of the probe laser, all the atoms in the sample will interact with the probe beam with the same strength	For a given frequency ω of the probe beam, only those atoms with centre frequencies ω_c close to ω will be excited.
The strength of the interaction increases as ω is tuned close to the centre frequency ω_0 of the transition since the homogeneous cross-section increases	The strength of the interaction increases as ω is tuned close to ω_0 since more atoms have this centre frequency
Atoms excited to the upper level will emit radiation with the same spectrum — described by $g_H(\omega - \omega_0)$	Atoms excited to the upper level will emit a homogeneously broadened spectrum centred at $\omega_c = \omega$
The *shape* and centre frequency of the fluorescence spectrum will be unchanged as the frequency of the probe laser is tuned across the transition	The shape of the fluoresence spectrum is unchanged, but the centre frequency tunes with ω.
The *intensity* of the fluorescence changes as the probe frequency is tuned; the fluorescence will be brightest when ω is tuned close to ω_0, and be negligibly weak when tuned many homogeneous linewidths away from ω_0	The intensity of the fluorescence changes as the probe frequency is tuned; the intensity of the spectrum will be greatest for probe frequencies close to ω_0 since more atoms can interact with the probe beam.

Table 1.1: Comparison of classes of line broadening

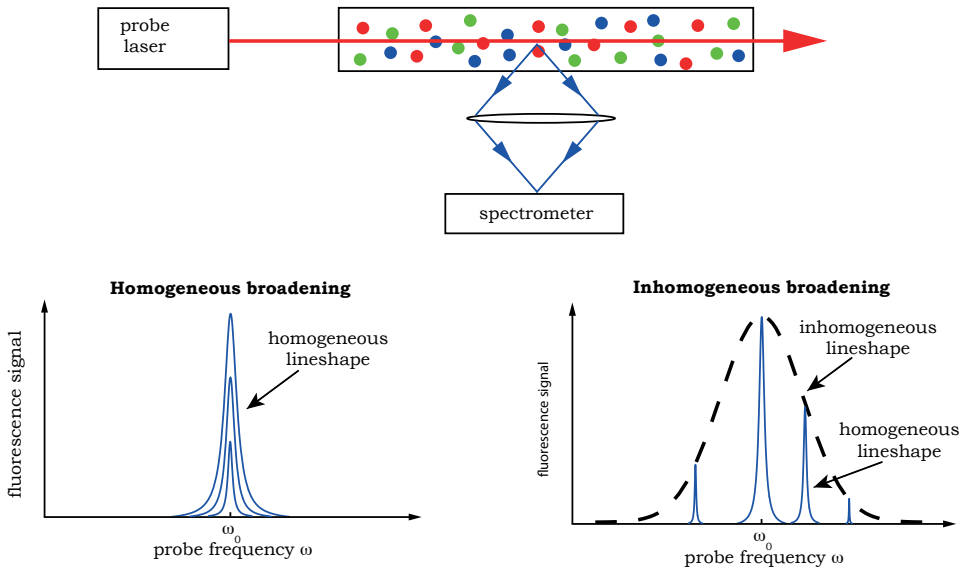


Figure 1.2: Fluorescence spectra recorded as a function of the frequency ω of a probe laser for the cases of homogeneous and inhomogeneous broadening.

1.4 The optical gain cross-section

In the B3 course it was shown that the growth of a beam of radiation of spectral intensity $I(\omega, z)$ interacting with atoms via a homogeneously broadened transition is given by,

$$\frac{\partial I}{\partial z} = N^* \sigma_{21}(\omega - \omega_0) I(\omega, z), \tag{1.8}$$

where the **population inversion density** is defined as,

$$N^* = N_2 - \frac{g_2}{g_1} N_1, \tag{1.9}$$

and the **optical gain cross-section** is given by,

$$\sigma_{21}(\omega - \omega_0) = \frac{\hbar \omega_0}{c} B_{21} g_{\mathrm{H}}(\omega - \omega_0) = \frac{\pi^2 c^2}{\omega_0^2} A_{21} g_{\mathrm{H}}(\omega - \omega_0). \tag{1.10}$$

Note that it is often useful to define the **gain coefficient** $\alpha_{21}(\omega - \omega_0)$:

$$\alpha_{21}(\omega - \omega_0) = \frac{1}{I} \frac{\partial I}{\partial z} = N^* \sigma_{21}(\omega - \omega_0). \tag{1.11}$$

1.4.1 Laser rate equations in terms of the optical cross-section

We will often need to write rate equations to describe the transfer of population between levels of an atom. In terms of the Einstein coefficients, the rate equation for (say) the upper laser level can be written:

$$\frac{dN_2}{dt} = R_2 - (N_2 B_{21} - N_1 B_{12}) \int_0^\infty g_{\rm H}(\omega - \omega_0)\rho(\omega) d\omega + \ldots \tag{1.12}$$

$$= R_2 - N^* \int_0^\infty \sigma_{21}(\omega - \omega_0) \frac{I(\omega)}{\hbar\omega} d\omega + \ldots \tag{1.13}$$

where R_2 is the rate of pumping of the upper laser level, $\rho(\omega)$ is the spectral energy density of the beam, and the $+\ldots$ indicate that in general there may be other processes to consider. The second line follows from the relations between the Einstein B-coefficients and the definition of the optical gain cross-section.

For narrow-band radiation the gain cross-section varies slowly over the spectral width of the radiation and so $I(\omega)$ acts like a Dirac delta function: $I(\omega) = I_T \delta(\omega - \omega_L)$ where I_T is the total intensity and ω_L the centre frequency of the beam. We then have:

$$\frac{dN_2}{dt} = R_2 - N^* \sigma_{21}(\omega_L - \omega_0) \frac{I_T}{\hbar\omega_L} + \ldots, \tag{1.14}$$

In a similar way, we can integrate eqn (1.8) over the bandwidth of the beam to find the growth equation for the total intensity.

$$\frac{dI_T}{dz} = N^* \sigma_{21}(\omega_L - \omega_0) I_T = \alpha_{21}(\omega_L - \omega_0) I_T. \tag{1.15}$$

Equations (1.14) and (1.15) will be used frequently in our analysis of laser systems.

1.5 Gain saturation

Gain saturation refers to the phenomenon whereby the optical gain on a laser transition depends on the intensity of the radiation interacting with it. Figure 1.3 shows schematically the processes which affect the level populations of a laser operating under steady-state conditions in the presence of an intense beam of radiation. In the absence of radiation on the laser transition, pumping of the upper and lower levels at rates R_2 and R_1, respectively, will form a population inversion $N^*(0)$. However the populations in the upper and lower laser levels will be altered by the presence of radiation on the laser transition owing to stimulated transitions from $2 \to 1$ and absorption from $2 \leftarrow 1$. Hence the population inversion in general will be a function of the intensity[3] I of the (assumed narrow-band) radiation and the frequency detuning of this radiation from the centre frequency of the transition.

[3] We will usually drop the subscript T, and distinguish spectral intensity by writing it in the form $I(\omega)$.

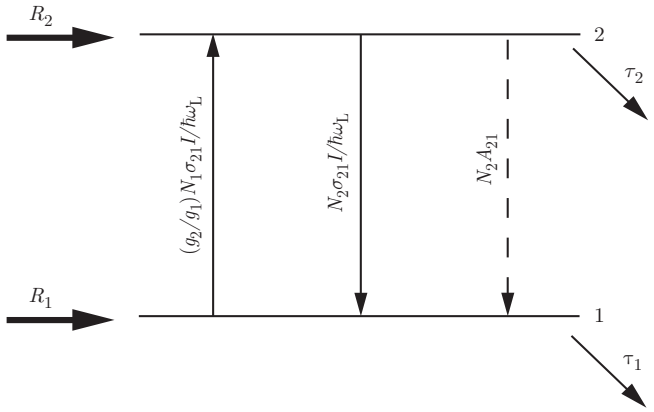


Figure 1.3: Processes affecting the upper and lower laser levels in a laser operating under steady-state conditions in the presence of an intense radiation beam.

1.5.1 homogeneous broadening

For a homogeneously broadened laser transition the rate equations for the laser levels may be written as,

$$\frac{dN_2}{dt} = R_2 - N^*\sigma_{21}(\omega_L - \omega_0)\frac{I}{\hbar\omega_L} - \frac{N_2}{\tau_2} \tag{1.16}$$

$$\frac{dN_1}{dt} = R_1 + N^*\sigma_{21}(\omega_L - \omega_0)\frac{I}{\hbar\omega_L} + N_2A_{21} - \frac{N_1}{\tau_1}. \tag{1.17}$$

In last year's lectures we showed how these may be solved straightforwardly to give the steady-state population inversion:

$$N^*(I) = \frac{N^*(0)}{1 + I/I_s(\omega_L - \omega_0)}. \tag{1.18}$$

where I_s is the **saturation intensity** (Units $\mathrm{W\,m^{-2}}$), and is given by

$$I_s(\omega_L - \omega_0) = \frac{\hbar\omega_L}{\sigma_{21}(\omega_L - \omega_0)\tau_R}, \qquad \textbf{Saturation intensity} \tag{1.19}$$

where we have emphasized that the saturation intensity depends (through the optical cross-section) on the detuning of the intense radiation from the centre-frequency of the transition[4]. The *recovery time* τ_R is given by

[4]The saturation intensity is *not* a spectral quantity; it has units of intensity! However, just like the optical gain cross-section itself, it *does* depend on the frequency detuning. In order to avoid clutter, we will often drop the explicit frequency dependence of the saturation intensity and just write I_s.

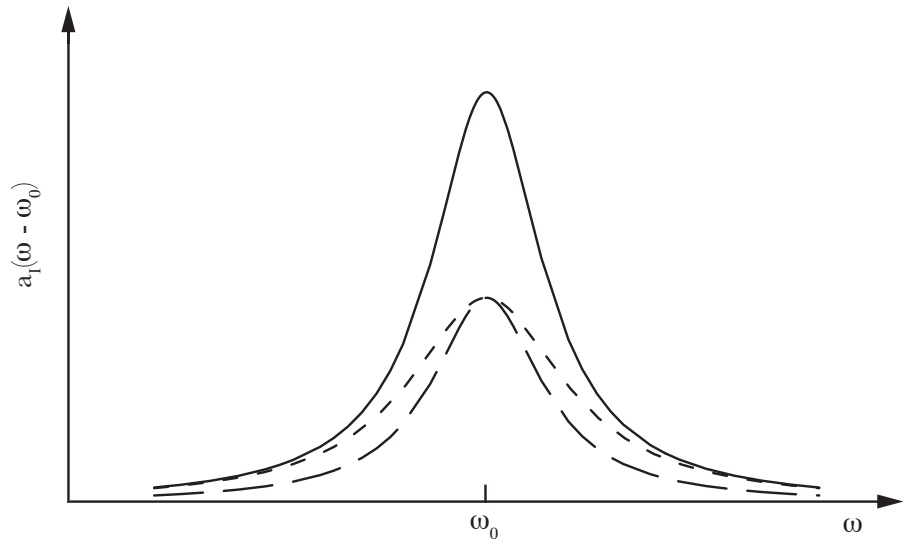


Figure 1.4: The gain coefficient measured for a homogeneously broadened transition by: a weak probe beam (solid line); a weak probe beam in the presence of an intense beam of intensity $I_s(\omega_L - \omega_0)$ at frequency ω_L (long-dashed line); an intense probe beam of intensity $I_s(0)$ (short-dashed line).

$$\tau_R = \tau_2 + \frac{g_2}{g_1}\tau_1[1 - A_{21}\tau_2]. \tag{1.20}$$

The recovery time depends on the lifetimes of both the upper and lower laser levels. It can usefully be thought of as giving an approximate measure of the time taken for the population inversion to re-establish itself after a perturbation, such as the passage of an intense laser pulse.

The **saturated gain coefficient** can be found directly from the saturated population inversion,

$$\alpha_I(\omega - \omega_0) = N^*(I)\sigma_{21}(\omega - \omega_0) = \frac{\alpha_0(\omega - \omega_0)}{1 + I/Is}. \tag{1.21}$$

Note that we can calculate the saturation intensity for absorption in a similar way. The result is of the same form as eqn (1.19), but with a different expression for the recovery time τ_R.

To illustrate how the gain coefficient saturates in a homogeneously broadened laser medium, we imagine measuring the gain coefficient under the following conditions:

1. Using a weak, tunable probe beam in the absence of any other radiation;

2. Using a weak, tunable probe beam in the presence of a second (intense) beam of radiation at a frequency ω_L of intensity $I_s(\omega_L - \omega_0)$.

3. An intense, tunable probe beam of constant intensity $I_s(0)$.

Clearly in the first case the gain coefficient measured as a function of the frequency ω of the probe beam will simply be the small-signal gain coefficient $\alpha_0(\omega - \omega_0)$. In the second case, however, the presence of the intense beam will decrease the population inversion density. Since the intense beam has an intensity equal to the saturation intensity (by definition, whatever the frequency ω_L of the intense beam), the population inversion will burnt down to *half* that generated by the pumping in the absence of the intense beam. As the frequency of the probe beam is tuned across the laser line, the gain measured will therefore always be half of the small-signal gain. Finally, for the case of the intense probe beam, the degree of saturation of the population inversion depends on the detuning of the probe from the centre frequency of the transition. In this case the intensity of the probe is equal to the saturation intensity when tuned to the line-centre. Hence when $\omega = \omega_0$ the measured gain coefficient will be half the small-signal gain coefficient. At large detuning, however, the intensity of the probe beam will be small compared to the saturation intensity; here the beam interacts only weakly with the atoms and hence perturbs their populations only slightly. The measured gain coefficient must, then, tend to the small-signal value at large detuning. The three measured gain profiles are shown in Figure 1.4.

1.5.2 Inhomogeneously broadened transitions

The phenomenon of gain saturation is considerably more complicated when the laser transition is inhomogeneously broadened since we must now consider separately the extent to which the population inversion of each frequency class is saturated, and the contribution by each such class to the overall gain.

We will restrict ourselves to a qualitative discussion.[5] Consider an inhomogeneously broadened gain medium subjected to a beam of intense, narrow-band radiation with frequencies close to ω_L. The extent to which each class of atoms is saturated — that is, the extent to which, for each class of atoms, the population density of the upper level is reduced — depends on the detuning of the intense beam from the centre frequency of the class. It should be clear that significant saturation will only occur for those classes of atoms with centre frequencies such that $|\omega_c - \omega_L| \ll \Delta\omega_H$; for these atoms the population inversion will be strongly burnt down. However, those classes with centre frequencies significantly different from that of the radiation (on the scale of $\Delta\omega_H$) will not be saturated to any appreciable extent. As such, a **spectral hole** will be burnt into the population inversion, as illustrated in Fig. 1.5(a). The width of this spectral hole will be approximately $\Delta\omega_H$.

The gain coefficient measured by a weak probe is given by integrating the contribution to the gain from each class. Hence we expect the measured gain coefficient in this case to exhibit a spectral hole, reflecting that in the population inversion. This phenomenon is shown in Figure 1.5(b).[6]

[5] A mathematical treatment may be found in Chapter 5 of H&W.

[6] It can be shown that in the limit of extreme inhomogeneous broadening — i.e. the inhomogeneous linewidth is much greater than the homogeneous linewidth — and if the homogeneous lineshape is Lorentzian, the spectral hole burnt in the gain profile is also Lorentzian, with a full-width at half maximum of $\Delta\omega_{hole} = \Delta\omega_H \left[1 + \sqrt{1 + I/I_s(0)}\right]$. For this case we see that if the intensity of the saturating beam is small compared to the saturation intensity $\Delta\omega_{hole} \approx 2\Delta\omega_H$. This result is consistent with our expectation that the spectral hole burnt in the population inversion is approximately $\Delta\omega_H$ wide, and the range of frequency classes with which the probe interacts is also approximately equal to $\Delta\omega_H$.

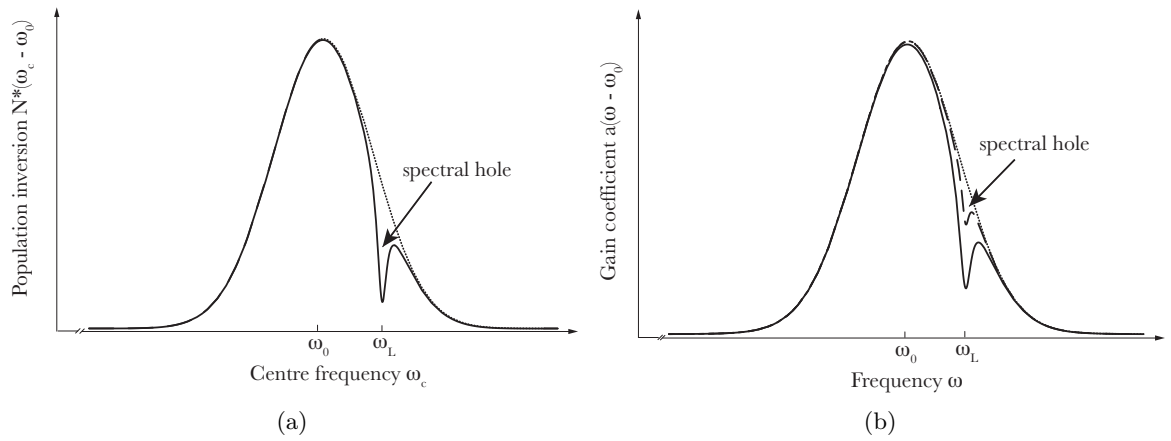


Figure 1.5: Saturation effects on inhomogeneously broadened transitions. (a) The effect of a beam of intense, narrow-bandwidth radiation of total intensity I with frequencies close to $\omega_{\rm L}$ on the distribution of the population inversion over centre frequencies. The dotted curve shows the unperturbed distribution for $I = 0$, and the solid curve the distribution for $I \approx I_{\rm s}(0)$. For this plot $\Delta\omega_{\rm H}/\Delta\omega_{\rm D} = 0.1$. (b) Spectral hole burning in the gain profile of an inhomogeneously broadened laser transition by a narrow-bandwidth beam of intensity I and with frequencies close to $\omega_{\rm L}$. The gain profile measured by a weak probe beam is shown for $I = 0$ (dotted), $I = I_{\rm s}(0)$ (dashed), $I = 10I_{\rm s}(0)$ (solid). For this plot $\Delta\omega_{\rm H}/\Delta\omega_{\rm D} = 0.03$.

Intense probe The saturation of the gain as measured by an intense probe (with no other beam of radiation present) is very different from that shown in Figure 1.5(b) since as the probe is tuned across the transition the class of atoms which provide gain are also the class which experiences saturation. Hence in this situation we do not expect a spectral hole to be burnt into the gain profile.

It can be shown that for the case of extreme inhomogeneous broadening (i.e. $\Delta\omega_{\rm D} \gg \Delta\omega_{\rm H}$), the gain coefficient measured by an intense probe is equal to a constant fraction of that measured by a weak probe. Further, if the homogeneous lineshape is Lorentzian, it may be shown that the saturated gain coefficient is,

$$\alpha_I^{\rm D}(\omega - \omega_0) = \frac{\alpha_0^{\rm D}(\omega - \omega_0)}{\sqrt{1 + I/I_{\rm s}(0)}}. \qquad \text{(Special case)} \tag{1.22}$$

Notice that this looks rather similar to eqn (1.21), which was obtained for homogeneous broadening. However, this similarity is very mis-leading since eqn (1.21) refers to the gain measured for a homogeneously- broadened transition by a weak probe in the presence of a saturating beam at $\omega = \omega_{\rm L}$; eqn (1.22) refers to the gain measured for an extreme inhomogeneously broadened transition by an intense probe in the absence of any other saturating radiation. The situations couldn't be more different!

1.5.3 Summary

To summarize: for homogeneously broadened gain media an intense beam of radiation at $\omega = \omega_{\rm L}$ causes the entire population inversion to be burnt down by a factor $[1 + I/I_{\rm s}(\omega_{\rm L} - \omega_0)]$ such that the gain profile measured by a weak probe beam has the same shape as that measured in the absence of the saturating beam, but with the gain being reduced by the factor $[1+I/I_{\rm s}(\omega_{\rm L}-\omega_0)]$; for inhomogeneously broadened gain media a spectral hole is burnt into the population inversion, and the gain measured by a weak probe beam is only reduced for frequencies close to $\omega_{\rm L}$. For homogeneously broadened systems the gain measured by an intense probe beam is reduced by a greater extent at frequencies close to the line centre, leading to a flattening and broadening of the measured gain profile; for strongly inhomogeneously broadened media the gain is everywhere reduced by the same factor $\sqrt{1 + I/I_{\rm s}(0)}$, and the measured gain profile has the same shape as measured by a weak probe beam.

1.6 Threshold behaviour

We now examine how lasers operating on homogeneously and inhomogeneously broadended transitions behave as they are brought above threshold.

Let us suppose that the gain medium is inserted into a optical cavity of length L. In general the modes of an optical cavity are described by the set of integers (l, m, p). The integers l and m determine the transverse profile of the right- and left-going beams in the cavity, and are said to determine the **transverse modes** of the cavity; the integer p determines the number of half-wave loops of the radiation field along the axis of the cavity, and so describes the **longitudinal modes**. The frequencies of the cavity modes are determined primarily by p; for example, for the case of a plane-plane cavity the longitudinal modes are separated in frequency by an amount,

$$\omega_{p,p-1} = \frac{\pi {\rm c}}{L}. \tag{1.23}$$

As discussed in the B3 course, in order for laser oscillation to occur on a cavity mode the round-trip gain must equal the round-trip loss.

1.6.1 Homogeneously broadened laser transitions

As the pumping is increased from zero the population inversion N^*, and hence the small-signal gain coefficient $\alpha_0(\omega)$, will increase. For low levels of pumping, spontaneous emission will increase the energy in all of the cavity modes with frequencies lying within the linewidth of the laser transition. The intensity will be low, and the number of photons per mode will typically be much less than unity.

With further increases in pumping the small-signal gain coefficient will increase until the threshold condition is met for the cavity mode closest to the line centre, causing the onset of oscillation on that cavity mode. The onset of oscillation will be accompanied by a very large (by a factor of order 10^{15}) increase in the energy density and number of photons in the oscillating mode.

Further increases in the pumping do not increase either the gain coefficient or the population inversion. In fact this must be the case since in the steady-state the round-trip gain must always be balanced by the round-trip loss. Instead, as the pumping is increased the intensity of the oscillating

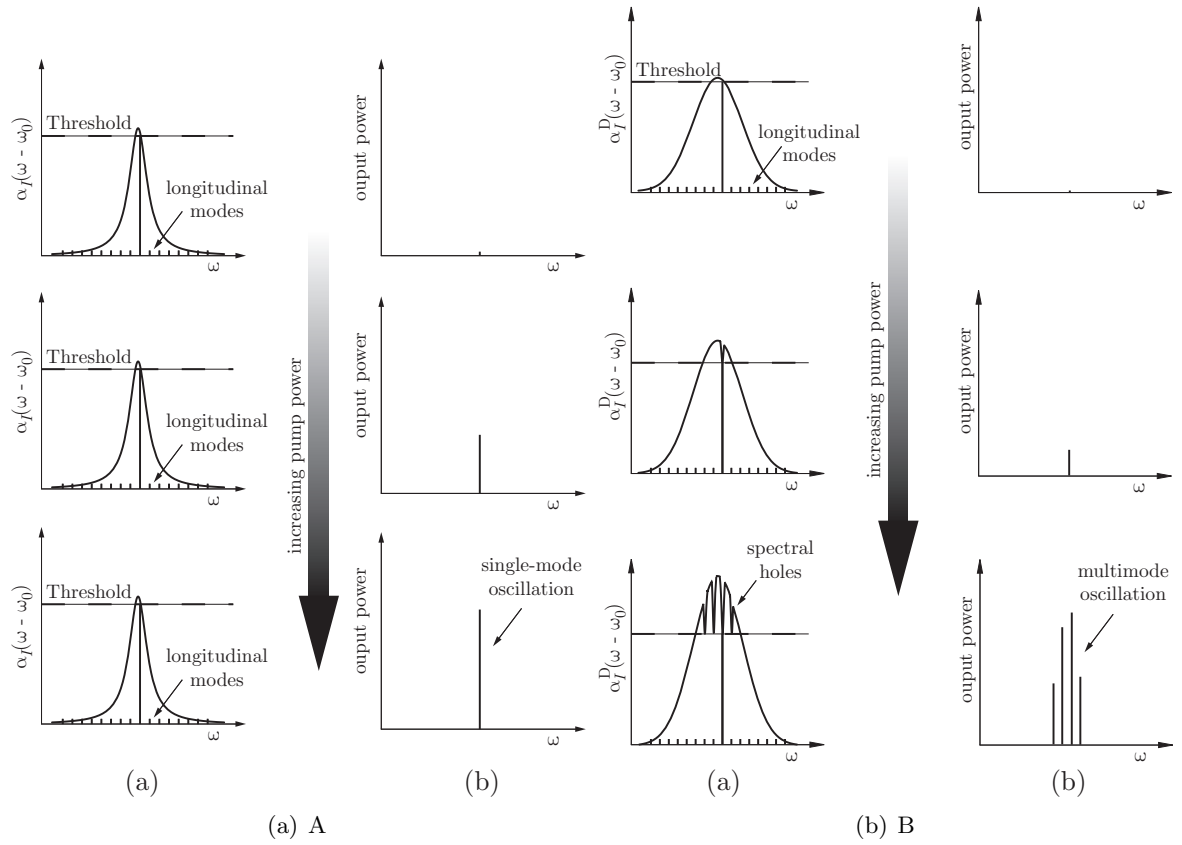


Figure 1.6: Behaviour of the small-signal gain (a) and output power (b) as the pumping power is increased above the threshold value for lasers operating on : (A) homogeneously; and (B) inhomogeneously broadened transitions

mode increases — this, after all, is what we are trying to achieve! The increased intensity of radiation circulating within the cavity reduces, or burns down, the population inversion to the threshold value. In other words, the laser transition is saturated. For a given level of pumping the intensity of the oscillating mode is determined by the condition that the saturated round-trip gain is equal to the round-trip loss, or, equivalently, that the population inversion is burnt down from the value, N_0^*, that would be achieved at that level of pumping, to the threshold value N_{thresh}^*.

The behaviour of the saturated gain coefficient $\alpha_I(\omega - \omega_0)$ and the output power of the laser as the level of pumping is increased is illustrated schematically in Figure 1.6(a).

Spatial hole burning

In principle, for a homogeneously broadened laser system only the cavity mode with the largest gain can ever oscillate; since the population inversion is clamped to the threshold value, and all atoms interact with radiation of a given frequency with the same strength, the other modes will always be below the threshold for oscillation. However, in practice in some circumstances more than one mode can oscillate.

This behaviour arises from the fact that the spatial distribution of the intensity of the oscillating mode is not uniform. In particular the oscillating mode will form a standing wave within the cavity. Near the anti-nodes of the standing wave the population inversion will be burnt down to threshold value; near the nodes the intensity will be low, and the population inversion essentially unsaturated. The non-uniform burning down of the population inversion in a homogeneously broadened laser system is known as **spatial hole-burning**, by analogy with the spectral hole-burning discussed in the next section. Spatial hole-burning allows other modes, with slightly different frequencies, to feed off regions of unsaturated population inversion and reach the threshold for oscillation. In such circumstances two or more cavity modes can oscillate, leading to **multimode oscillation** at several frequencies within the linewidth of the transition.

1.6.2 Inhomogeneously broadened systems

The above-threshold behaviour of inhomogeneously broadened lasers is quite different since different classes of atoms interact with different frequencies, and hence different cavity modes. Once again, as the pumping is increased to the threshold value the mode with the largest gain will start to oscillate. Now, however, as the pumping is increased the population inversion is only clamped to the threshold value for those classes of atoms which can interact with the frequency of the oscillating mode; other classes will be essentially unaffected, and for these classes the population inversion will continue to increase as the pumping increases. As a consequence the gain coefficient $\alpha_I^{\mathrm{D}}(\omega-\omega_0)$ develops a spectral hole at the frequency of the oscillating mode.

Further increases in the pumping level will allow other modes, of different frequencies, to reach the threshold for oscillation. Thus, in general as the pumping level is increased in an inhomogeneously broadened laser oscillator, the number of oscillating modes increases as well as the output power of each oscillating mode. The process is illustrated schematically in Figure 1.6(b).

Lecture 2

Solid State Laser Materials

A large number of technologically and scientifically important lasers operate by optically pumping transitions within ions doped as an impurity species into a variety of crystalline or glass hosts. The term **solid state lasers** is usually reserved for systems of this type, lasers operating between levels of the electron band structure found in semiconductors being referred to as **semiconductor** or simply **diode** lasers.

Solid state laser materials offer advantages over liquid and gas lasers in that they are robust, chemically inert, require no special handling, and do not degrade or become contaminated with use.

The spectroscopy of ions doped into solid hosts is a complex subject. Fortunately for our purposes we need only understand a few key concepts in order to appreciate how different laser systems work, and why they behave as they do. In particular we would like to know the energy level structure of the ions, the lifetimes of the levels, and the linewidth of the transitions between them.

2.1 General considerations

Ions doped into a solid host differ from free ions in that they are subjected to the electric field — the **crystal field** — arising from the ions of the solid host. The interaction with the crystal field leads to an additional term in the Hamiltonian of the form,

$$H_{\mathrm{c}} = -\mathrm{e}V_{\mathrm{c}}, \tag{2.1}$$

where V_{c} is an electrostatic potential describing the crystal field. In general the effect of this interaction is to split and shift the energy levels from their positions in the absence of the crystal field, an effect known as the Stark effect[1].

The crystal field has a symmetry reflecting that of the crystal lattice. Consequently a proper treatment of the effect of the crystal field involves detailed consideration of this symmetry; indeed in some circumstances these symmetry properties are used to label the energy levels arising from the interaction with the crystal field.

The nature of the influence of the crystal field on the energy level structure depends critically on its strength relative to the other terms in the Hamiltonian. Three limiting cases may be identified:

[1] Named after the German physicist Johannes Stark (1874-1957).

- **Weak field**: $|H_{\rm c}| \ll |H_{\rm s-o}| \ll |H_{\rm re}|$. In this case the interaction of the crystal field is weak compared to the spin-orbit and residual electrostatic interactions (as well the interaction with the central field) so that the energy levels are changed only slightly from those in the free ion. The effect of the crystal field in this case is to cause a small splitting and shifting of the levels of the free ion to form a so-called **manifold** of closely-spaced levels. Since the energy shifts are small, the labelling of the manifolds is simply that of the energy levels of the free ion. Further, in the weak field case the energy level structure of an ion will be almost independent of the crystal host in which it is embedded. A good example of the weak field case are the energy levels formed by the $4f^n$ configurations in trivalent rare earth ions discussed in Section 2.2.

- **Intermediate field**: $|H_{\rm s-o}| \ll |H_{\rm c}| \ll |H_{\rm re}|$. In this situation the crystal field interaction must be considered as a perturbation acting on the terms formed by the residual electrostatic interaction before the spin-orbit interaction is taken into account; the crystal field acts on the terms formed by the residual electrostatic interaction and splits and shifts them. The spin-orbit interaction leads to further splitting of the energy levels. For intermediate crystal fields it is no longer possible to label the energy levels with the quantum numbers L, S, and J. The first-row transition metal ions, such as Cr^{3+} frequently exhibit crystal fields of intermediate strength. It is more difficult to calculate the energy levels for intermediate crystal fields than for weak- or strong-fields. Frequently, therefore, the energy levels of ions experiencing intermediate crystal fields are labelled with either a weak- or strong-field notation.

- **Strong field**: $|H_{\rm s-o}|, |H_{\rm re}| \ll |H_{\rm c}|$. Here the spin-orbit or residual electrostatic interactions may be ignored (in the first approximation) and the crystal field acts on the single-electron orbitals of the central potential to give single-electron 'crystal-field orbitals' which reflect the symmetry of the crystal field. As discussed below, the energies of the orbitals depends on the spatial distribution of the orbital wave function with respect to the crystal lattice. The energy levels are labelled by the symmetry properties of the electron wave function using a notation derived from group theory. Strong crystal fields can arise in the second- and third-row transition-metal ions.

For each electron the relative strengths of the residual electrostatic, spin-orbit, and crystal field interactions depends very strongly on the radial distance r_i of the electron from its nucleus, and the distance d from the active ion to the neighbouring ions. For example, for neighbouring ions distributed with octahedral symmetry the energy of interaction is proportional to,

$$\frac{\langle r_i^4 \rangle}{d^5}$$

where $\langle r_i^4 \rangle$ is the expectation value of r_i. In contrast, the energy shift arising from the spin-orbit interaction is proportional to,

$$\left\langle \frac{1}{r_i^3} \right\rangle$$

Hence the ratio of the interaction with the crystal field to the spin-orbit interaction varies approximately as,

$$\frac{\Delta E_{\rm c}}{\Delta E_{\rm s-o}} \propto \frac{\langle r_i^7 \rangle}{d^5}. \tag{2.2}$$

As a consequence very small changes in the relative size of the mean electron radius or the nearest-neighbour distance can change the relative strength of the interaction with the crystal field enormously.[2]

2.1.1 Radiative transitions

For all the solid-state laser transitions discussed in this lecture, the laser transition occurs between two levels *within the same electron configuration.* The parity of the electron wave function does not change in such transitions, and hence they are forbidden by the selection rules of electric dipole radiation. Whilst this would indeed strictly be the case for transitions between the equivalent levels in a free ion, for ions doped into a solid, these electric dipole transitions may occur for two reasons. The first is simply that the presence of the crystal field means that there is no longer inversion symmetry at the site of the active ion: in other words the electron wave function no longer has a definite parity and under the transformation $\boldsymbol{r}_i \rightarrow -\boldsymbol{r}_i$ the electron wave function can change magnitude (as well as possibly change sign). A *pure* electron configuration *must* have a well-defined parity since it is a solution of the Schrödinger equation in a spherically-symmetric central field. However, the crystal field interaction is not spherically symmetric and therefore introduces small admixtures of configurations of opposite parity to form a wave function which is no longer spherically symmetric, and no longer has a well-defined parity. These admixtures of configurations of opposite parity can allow transitions to occur between two levels which nominally have the same electron configuration.

A second mechanism by which electric dipole transitions may occur between two levels of the same configuration is a dynamically-induced transition strength through vibrations of the crystal lattice which destroy the inversion symmetry at the site of the active ion. This second mechanism becomes important when the active ion is located at a crystal site exhibiting a high degree of symmetry. As might be expected, in such cases the radiative lifetime of the level depend strongly on the temperature of the crystal.

Notwithstanding the above, the dipole-forbidden nature of transitions within a single configuration means that the transition rates are significantly slower than fully dipole-allowed transitions. The *radiative* lifetimes of the upper levels of visible transitions within ions doped into a solid are 3 to 6 orders of magnitude longer than fully dipole-allowed transitions in free atoms and ions.

Finally we note that the strength of a radiative transition can depend on the orientation of the polarization of the radiation with respect to the axes of the crystal. This is true for both absorption and emission. In such cases the orientation of the crystal axes with respect to the axis of the laser

[2]Frequently it is said that the relative strength of an electron's interaction with the crystal field depends on the extent to which it is shielded by other electrons in the ion. For example, the 4f electrons in trivalent rare earth ions are often said to be shielded by the 'outer' 5s and 5p electrons. Shielding effects of this type do occur, but calculation shows them to be approximately an order of magnitude smaller than the effects of changing mean electron radius and nearest-neighbour distance. For a discussion of this point see G. Burns, *Solid State Physics*, Academic Press.

cavity and the direction and polarization of the pump radiation can be important, and the laser output can be partially or totally polarized.

2.1.2 Non-radiative transitions

Non-radiative transitions play a very important role in the operation of solid-state lasers: rapid non-radiative decay can provide an efficient route for feeding population from one or more excited levels into the upper laser level, and similarly can help to keep the population of the lower laser level low.

Phonon de-excitation

The energy levels of an active ion are coupled to the vibrations of the crystal lattice via the crystal field interaction. The lattice vibrations have a spectrum of normal modes,[3] and their excitation is quantized; the unit of excitation being known as a **phonon**. Transitions between energy levels may therefore occur not by emission of radiation, but by the emission or absorption of one or more lattice phonons. The rate of such **phonon de-excitation** is found to decrease very rapidly with the number of phonons involved. For example, for the case of weak coupling it may be shown that the rate of de-excitation on a transition of energy E_{21} by phonons of energy $E_{\rm p}$ varies as $\exp[-\beta(E_{21}/E_{\rm p})]$, where β is a positive constant which depends on the transition. In practice the de-excitation rate is found to be small if the number of phonons required is greater than approximately 5, and consequently the dominant contribution to the decay rate is from the phonons with the largest possible energy.

The rate of phonon de-excitation depends strongly on the strength of coupling between the electrons of the ions and the crystal field, and on the phonon spectrum of the lattice — and in particular on the maximum phonon energy. The de-excitation rates therefore vary widely for different combinations of impurity ions and hosts; the rate of phonon de-excitation also increases rapidly with temperature which reflects the increasing density of phonons in a given mode with temperature (determined by the Bose-Einstein distribution). For small energy differences the rate of non-radiative transitions can be extremely high: at room temperature the rate is of order 10^{11} - 10^{12} $\rm s^{-1}$.

The strong dependence of the rate of phonon de-excitation on the energy gap between the initial and final levels of the ion means that if a level is to be suitable as an upper laser level it should lie significantly above the nearest lower-lying level. This is necessary if the level is to have a long fluorescence lifetime, which allows a large population to build up in the level; and ensures that it decays predominantly by emission of radiation, which enables the population in the upper laser level to be used efficiently.

2.1.3 Line broadening

The natural linewidths of transitions between the levels of an impurity ion are typically very small, owing to the long radiative lifetimes of the levels. In addition a temperature-dependent lifetime broadening arises from the increased rate of decay of the levels caused by non-radiative transitions. In practice, however, the measured linewidths of transitions are significantly broader than can be

[3] There are two classes of these modes, known as acoustic and optical modes. A discussion of normal modes of crystals may be found in textbooks on solid-state physics.

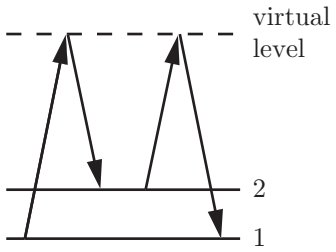


Figure 2.1: Schematic diagram of the two-phonon Raman process leading to line broadening. In this process phonons of two different energies transfer ions from level 1 to level 2 via a virtual level, followed by the reverse process which returns the ion to level 1. The net result is no change in either the level occupied by the ion or the phonon spectrum but a broadening of the energy level by an amount proportional to the rate at which this process occurs.

accounted for by the measured lifetimes of the upper and lower levels, and hence other processes must be responsible for the additional broadening.

A variety of mechanisms can cause this additional broadening. One of these, two-phonon Raman scattering is illustrated in Fig. 2.1. In this process two phonons of different energy are successively absorbed and emitted so as to remove, and then return an ion from its energy level. Since the ion is returned to its initial level there is no change in the lifetime of the level, but a broadening occurs proportional to the rate of scattering. As might be expected the rate of scattering, and hence its contribution to the linewidth, depends strongly on temperature.

Phonons may also play a more direct role in determining the linewidth of a radiative transition. For some ion-host combinations radiative transitions can be accompanied by the emission or absorption of one or more phonons. These so-called vibronic transitions (discussed in Section 2.3.1) have very broad, temperature-dependent linewidths, as discussed below.

2.2 Trivalent rare earths, $4f^n - 4f^n$ transitions

2.2.1 Energy level structure

A large number of solid-state lasers operate on $4f^n - 4f^n$ transitions in trivalent rare earth ions, and in particular the lanthanide series — the actinides being radioactive, and hence difficult to work with. Important trivalent rare-earth laser ions include: Ce^{3+}, Nd^{3+}, Ho^{3+}, and Er^{3+}.

Neutral atoms of these rare earths have a ground state electronic configuration of the form $[Xe]\,4f^{n+1}\,6s^2$, or $[Xe]\,4f^n\,5d\,6s^2$ where $n \geq 0$ and [Xe] represents the ground state configuration of xenon:

$$1s^2\,2s^2 2p^6\,3s^2 3p^6 3d^{10}\,4s^2 4p^6 4d^{10}\,5s^2 5p^6.$$

In forming the trivalent ion, the three loosest-bound electrons are lost:[4] the two 6s electrons and

[4]In the free ion the electrons are removed from the atom, in a solid they form bonds with the atoms of the host lattice.

either one of the 4f electrons or the 5d electron, to give a ground-state configuration which is always of the form [Xe] $4f^n$. The absorption spectrum of the ions corresponds to transitions within the 4f configuration ($4f^n \leftarrow 4f^n$ transitions), or $4f^{n-1}\,5d \leftarrow 4f^n$ transitions to the lowest-lying empty orbital, 5d.[5]

Of key importance to the spectroscopy of the trivalent rare earth ions in crystalline media is the fact that the wave function of the 4f orbital is more compact than those of the 5s and 5p orbitals. As such the mean radius of the 4f electrons are relatively small compared to the size of the ions, and hence to the nearest-neighbour distance. The interaction with the crystal field is therefore weak, producing a series of manifolds which are only slightly perturbed from those of the levels of the free ion. As a consequence the energy level structure of the 4f configurations in the trivalent rare earths is approximately independent of the crystal host, although the optical cross-sections of transitions may vary significantly.

The manifolds are labelled by the levels of the free ion from which they arise, i.e. $^{2S+1}L_J$. If the total spin quantum number S is an integer, the maximum number of levels within the manifold is given by $(2J+1)$; if S is half-odd-integer the levels of the crystal field are all doubly degenerate,[6] and consequently the maximum number of levels within the manifold is reduced to $(2J+1)/2$. It should be emphasized that these are the maximum number of non-degenerate levels; the actual number formed depends on the symmetry of the crystal field.

2.2.2 Transition linewidth

The natural linewidth of the $4f^n - 4f^n$ transitions is very narrow owing to the long radiative lifetimes of the levels. This linewidth is increased substantially by phonon collisions to give a homogeneously-broadened transition with a linewidth which depends on temperature. At room temperature the linewidth is typically in the range 0.1 - 3 THz (10 - 100 cm^{-1}).

Inhomogeneous broadening can arise if the crystal field varies with position in the crystal. For example, the presence of crystal defects, strain, or impurity ions can cause variation in the crystal field and hence local shifts of the ion energy levels. For good quality laser crystals these effects should be small, in which case the line broadening will be predominantly homogeneous. In contrast, ions doped into glasses experience a very wide range of local environments. The energies, broadening, and even the number of levels varies significantly from site to site and consequently all transitions will be strongly inhomogeneously broadened. The inhomogeneous linewidth of the $4f^n - 4f^n$ transitions in ions doped into a glass host are of order 30 THz (1000 cm^{-1}).

2.2.3 Nd:YAG laser

Perhaps the most important example of a laser based on a $4f^n - 4f^n$ transition in a trivalent rare earth ion is the Nd:YAG laser. Neodymium-based lasers are very widely used in science and industry. They are frequently used as pump lasers for dye lasers, or other solid-state lasers such as Ti:sapphire. In medicine they find applications in removing secondary cataracts or tissue removal. High-power Nd:YAG lasers can be used in laser drilling and welding.

[5] The 5s and 5p orbitals already being occupied.

[6] This is known as Kramers' degeneracy.

Crystal properties

In Nd:YAG the Nd^{3+} ion replaces the Y^{3+} ion. Since the size of the Nd ion, which has a radius of 98 pm, is greater than that of the Y ion, of radius 90 pm, it is not possible to introduce Nd ions with an atomic concentration much above 1.5% without straining the crystal lattice unduly. It is worth noting that an atomic concentration of Nd ions equal to 1% corresponds to a density of Nd ions of $1.386 \times 10^{20}\,\mathrm{cm}^{-3}$.

Energy levels

Figure 2.2 shows the energy levels of the Nd^{3+} ion and how these are split into manifolds by the crystal field of YAG. Broad pump bands are provided by several closely-spaced manifolds with energies between approximately 12 000 and 33 000cm^{-1} (300 - 800 nm). Particularly strong pumping occurs at 810 nm (12 300 cm^{-1}) and 750 nm (13 300 cm^{-1}), which can be accessed by diode laser pumping.

The manifolds excited by optical pumping on these bands are relatively closely spaced compared to the maximum phonon energy in YAG of $850\,\mathrm{cm}^{-1}$. Consequently excited ions cascade down through these manifolds by very rapid non-radiative transitions until the two levels (denoted R_1 and R_2) of the $^4\mathrm{F}_{3/2}$ manifold is reached. The levels of the $^4\mathrm{F}_{3/2}$ manifold are metastable because the nearest lower-lying level — the top of the $^4\mathrm{I}_{15/2}$ manifold — is separated by some $4\,698\,\mathrm{cm}^{-1}$, corresponding to more than 5 phonons. The rate of non-radiative decay of the $^4\mathrm{F}_{3/2}$ manifold is therefore slow, and it decays instead almost entirely by radiative transitions to levels of the ^{4}I manifolds, with a fluorescence lifetime of 230 μs. In contrast, the levels of the ^{4}I manifolds, are separated by less than $1\,500\,\mathrm{cm}^{-1}$, and consequently decay non-radiatively with lifetimes of order 100 ps.

It can be seen, therefore, that the energy levels of Nd:YAG are very well suited to achieving laser oscillation: the levels of the $^4\mathrm{F}_{3/2}$ manifold may be populated efficiently by cascade from broad pump bands, and they are metastable — ideal properties of an upper laser level. Suitable lower laser levels exist within the ^{4}I manifolds; these lie significantly above the ground state and decay very rapidly to it. The system is therefore an almost ideal example of a four-level laser system. Figure 2.3 shows a simplified energy level scheme for the strongest Nd:YAG laser transition; transitions to the levels of other ^{4}I manifolds are analogous, with the exception of those to the lowest-lying manifold, $\mathrm{I}_{9/2}$. Lasers operating on transitions to this manifold behave quite differently than those operating on transitions to the higher-lying manifolds, since the levels of the $\mathrm{I}_{9/2}$ manifold can have a significant thermal population. As such, these lasers behave more like three-level systems.

Lasing has been achieved from the $^4\mathrm{F}_{3/2}$ manifold on more than 20 transitions to the $\mathrm{I}_{13/2}$, $\mathrm{I}_{11/2}$, and $\mathrm{I}_{9/2}$ manifolds, with wavelengths near 1319 nm, 1064 nm, and 946 nm respectively. Of these, the strongest — and certainly the most commonly used — transition is the $^4\mathrm{F}_{3/2} \rightarrow {}^4\mathrm{I}_{11/2}$ transition.[7]

The R_1 and R_2 levels of the $^4\mathrm{F}_{3/2}$ manifold are separated by $84\,\mathrm{cm}^{-1}$. Rapid non-radiative transitions ensure that the populations of these two levels is maintained in thermal equilibrium and, importantly, the rate at which the population in these two levels is mixed is sufficiently fast that this remains the case even during laser oscillation. Hence at room temperature the relative populations of these levels are approximately 60% and 40% respectively. Of the possible $^4\mathrm{F}_{3/2} \rightarrow {}^4\mathrm{I}_{11/2}$ transitions, the strongest is the ℓ_2 transition between the R_2 and Y_3 levels with a vacuum wavelength of

[7] Indeed, in order to obtain lasing to the other manifolds it is necessary to employ frequency-selective cavities which suppress lasing to the $^4\mathrm{I}_{11/2}$ manifold and enhance the feedback on the desired transition.

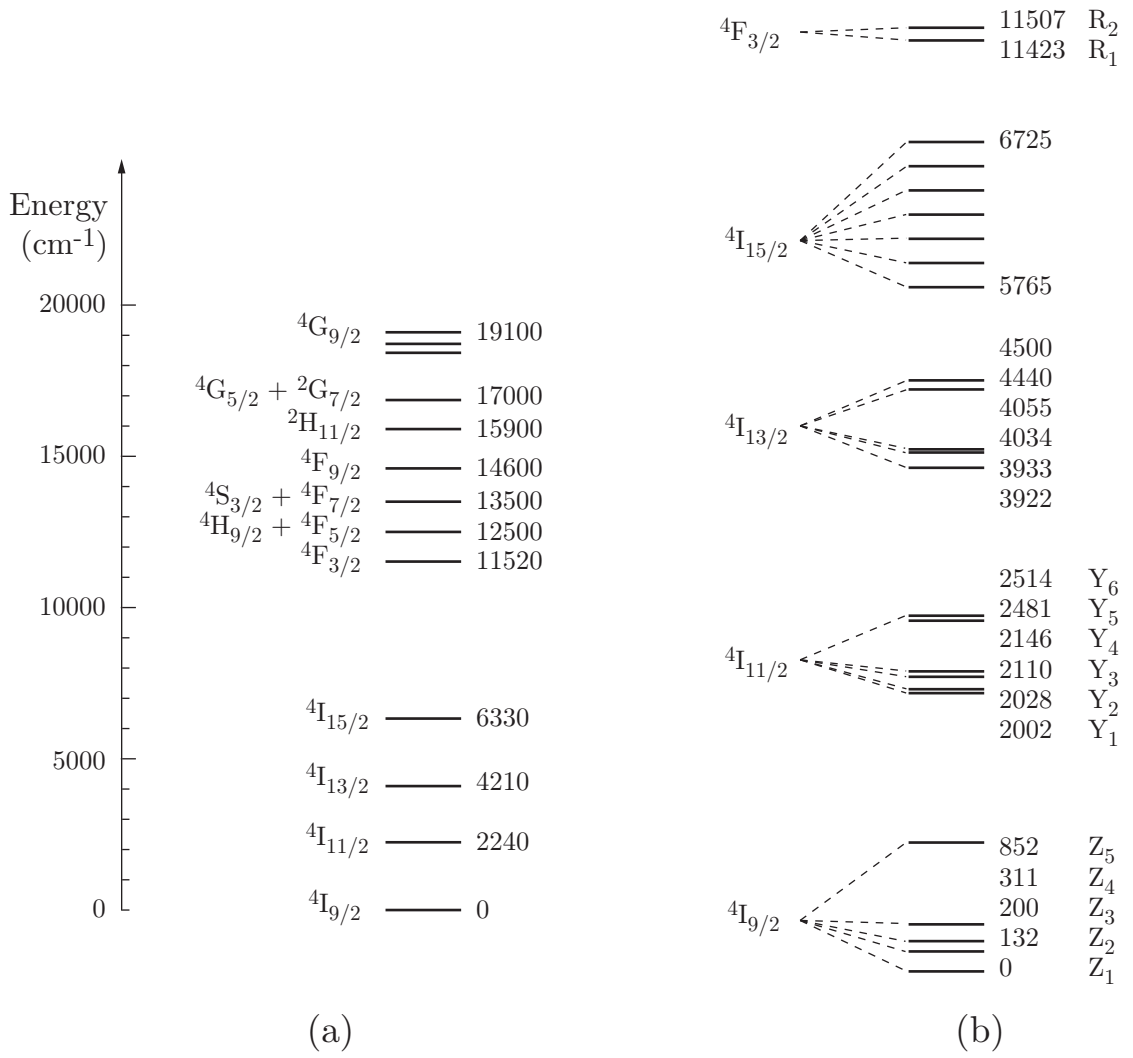


Figure 2.2: Energy level diagrams for the Nd^{3+} ion in YAG. (a) shows the lowest-lying levels of the free ion formed from the [Xe] $4f^3$ configuration; (b) illustrates how these levels are split by the crystal field in YAG to form manifolds of closely-spaced levels. Note that in (b) the levels of each manifold are shown on an expanded energy scale, and the spacing between manifolds is not drawn to scale. The energies of the levels is given in cm^{-1}.

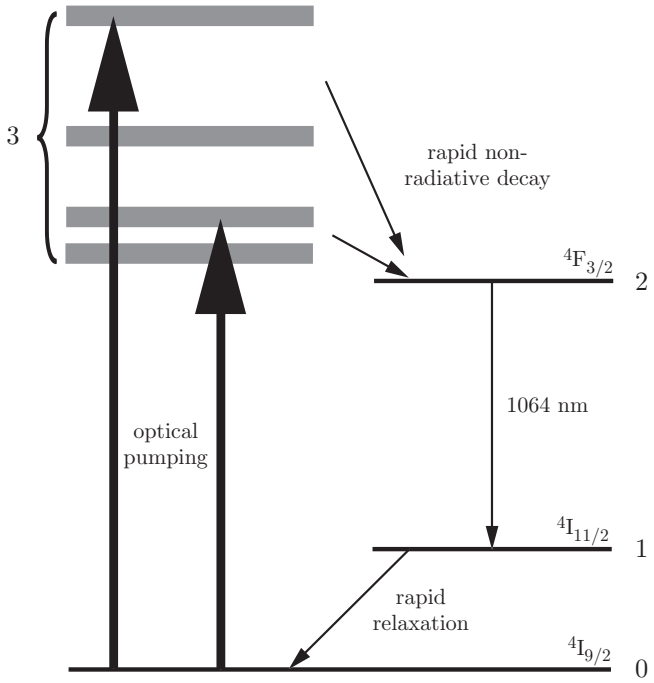


Figure 2.3: Simplified energy level diagram of the $^4F_{3/2} \rightarrow ^4 I_{11/2}$ laser transition in Nd:YAG showing the four-level nature of the laser scheme.

1064.15 nm. At room temperature this transition dominates the laser output, since the population of the R_2 levels is sufficiently large that the gain on this transition is largest. In contrast, if the crystal is cooled to low temperatures almost all the population of the $^4F_{3/2}$ manifold resides in the R_1 level and lasing occurs from this level on the ℓ_1 transition from R_1 to the Y_2 level of $^4I_{11/2}$ at a vacuum wavelength of 1064.40 nm.

Broadening

In Nd:YAG the $^4F_{3/2} \rightarrow {}^4I$ transitions are homogeneously broadened by phonon collisions, with a full-width at half-maximum of approximately 190 GHz ($6.5\,cm^{-1}$). Note that this is much larger than the natural broadening — dominated by the short lifetime of the lower laser level — which is of order 1 GHz.

Effective gain cross-section

Calculation of the optical gain is complicated by the fact that both the ℓ_1 and ℓ_2 transitions in Nd:YAG can contribute to the gain. In such cases it is useful to use a single **effective cross-section** which takes into account the contribution to the gain from all the transitions involved. The effective cross-section is defined in such a way that the gain coefficient is given by multiplying it by the *total* population of the manifold of levels, rather than the population of any single level.

The peak optical gain cross-sections of the ℓ_1 and ℓ_2 transitions are $1.9 \times 10^{-19}\,cm^2$ and $7.1 \times 10^{-19}\,cm^2$ respectively. Hence the dominant contribution comes from the ℓ_2 transition; if we neglect the contribution of the gain from the ℓ_1 transition, the peak value of the effective gain cross-section is given by multiplying the peak cross-section of the ℓ_2 transition by the proportion of population in the R_2 level (39%): $\sigma_{\text{eff}}(0) \approx f_{R_2}\sigma_{R_2}(0) = 2.8 \times 10^{-19}\,cm^2$. In fact, contribution from the spectral wings of the ℓ_1 transition increases the effective cross-section to $3.7 \times 10^{-19}\,cm^2$.

Practical implementation

Nd:YAG lasers are one of the most common types of laser in use, and can be operated in a wide variety of different configurations: they can run continuously or in a pulsed mode; they can be pumped by flashlamps or laser diodes; and they can provide mean output powers from a few milliwatts up to several kilowatts. Very often Nd:YAG lasers are frequency doubled, tripled, or quadrupled to generate radiation of wavelength 532, 355, or 266 nm respectively.

Flashlamp-pumped systems employ pump chambers with single or multiple flashlamps. In such systems the laser rod is typically of about 5 mm diameter and 30 - 150 mm long, and the optical cavity is of order 0.5 m in length. With pulsed pumping the laser is often Q-switched at a pulse repetition of 10 - 50 Hz to yield output pulses of 5 - 10 ns duration and energies of a few tens to several hundred millijoules. The addition of one or more Nd:YAG amplifier stages can increase the output pulse energy to several Joules.

Flashlamp-pumped systems can generate pulsed output with a much higher pulse repetition frequency by employing continuous pumping and employing acousto-optic modulators to Q-switch the cavity at a repetition rate of 10 - 20 kHz. The mean output power of this type of system can be as high as 15 W.

Table 2.1: Important parameters of Nd:YAG, Nd:YLF, Nd:$GdVO_4$, and Nd:Glass (Hoya LHG-5 phosphate) lasers.

	Nd:YAG	Nd:YLF	Nd:$GdVO_4$	Nd:Glass
λ (nm)	1064	1053	1063	1054
τ_2 (μs)	230	450	90	290
τ_1 (ps)	100s	100s	100s	100s
$\Delta\nu$ (GHz)	160	380	180	5 000
σ_{21} (10^{-20} cm^{-2})	37	18.7	76	4.1
κ (W m^{-1} K^{-1})	13.0	6.0	12.3	1.19

ND:YAG lasers may be pumped by GaAs diode lasers operating at 808 nm which closely matches the peak absorption corresponding to excitation of levels of the $^2H_{9/2}$ and $^4F_{5/2}$ manifolds. The excited levels lie only 900 cm^{-1} above the upper laser levels, and consequently much less energy is deposited in the crystal in the form of heat (phonons) than is the case for flashlamp pumping which excites all manifolds up to some 10 000 cm^{-1} above the upper laser levels.

Diode-pumped Nd:YAG lasers can generate continuous-wave output with powers of: 10 W or more with end-pumping; above 100 W with side-pumping; and of order 1 kW with slab-pumping. Q-switching and modelocking is also possible. The slope efficiencies of diode-pumped Nd:YAG can be as high as 50 - 60% for end-pumped configurations, and 25 - 40% for side-pumping. These figures are much higher than for flashlamp-pumped lasers, for which the slope efficiency is typically 3%.

2.2.4 Other crystalline hosts

Neodymium ions have exhibited laser oscillation in several crystalline hosts including lithium yttrium fluoride ($LiYF_4$ — usually written as YLF) several vanadates such as YVO_4 and $GdVO_4$. Vanadates are particularly useful in conjunction with diode-pumping since in these hosts the pump bands are some 80% broader, thereby reducing the tolerance required on the wavelength of the diode laser.

2.2.5 Nd:Glass laser

A wide variety of oxide-, fluoride-, and sulphide-based glasses have been developed as hosts for Nd ions. The wavelength of the $^4F_{3/2} \rightarrow {}^4I_{11/2}$ transition varies from approximately 1054 to 1062 nm depending on the glass host. The lifetime of the upper laser level is similar to that in YAG.

In glass hosts the laser transition is strongly inhomogeneously broadened, the linewidth increasing by a factor of approximately 50 to typically 6.5 THz ($\Delta\lambda \approx 25$ nm) compared to that found in Nd:YAG. Largely as a result of the greater linewidth, the gain cross-section of Nd:Glass is an order of magnitude smaller than that in Nd:YAG.

The small optical gain cross-section of Nd:Glass allows a large population inversion density to be generated without the onset of amplified spontaneous emission. The material is therefore frequently employed to amplify laser pulses to large energies. The low thermal conductivity of Nd:Glass restricts

the pulse repetition rate of amplifiers based on this material to no more than a few pulses per second; and for very high energy amplifiers, the maximum pulse repetition rate is much lower. Finally we note that the large linewidth supports amplification of pulses as short as 100 fs.

2.2.6 Erbium lasers

Laser action has been achieved in erbium ions doped into a variety of garnet and fluoride crystalline hosts, as well as in several types of glass. There are two transitions of interest, with wavelengths of approximately 2.9 μm and 1.5 μm. Radiation at 2.9 μm is absorbed very strongly by water, and consequently erbium lasers operating on this transition have found applications in medicine. The 1.5 μm transition matches the third transparency window of optical fibres, and consequently is important for optical communications, particulary in the form of the **erbium-doped fibre amplifier (EDFA)**, the construction of which is discussed in Lecture 6. Further, this wavelength falls into the so-called 'eye-safe' window,[8] and consequently erbium lasers may also be used in telemetry and laser-ranging applications.

Figure 2.4 shows the energy levels of Er^{3+} ions in a glass host, and the important $^4I_{13/2} \rightarrow ^4 I_{15/2}$ transition at 1.54 μm. The upper and lower manifolds of levels are split by the crystal field into 7 and 8 degenerate levels respectively. In a glass host the 56 possible transitions between these levels are broadened homogeneously by lifetime and phonon broadening, and inhomogeneously by spatial variations in the local environment. The broadening characteristics of the laser transition are therefore complex, and to some extent the relative importance of homogeneous and inhomogeneous broadening depends on the properties of the glass host employed. However, the linewidth of the laser transition is approximately an order of magnitude greater than in a crystalline environment and hence the transition may be considered to be predominantly inhomogeneously broadened.

The most important pump band occurs at 980 nm (10 200 cm^{-1}) to the $^4I_{11/2}$ manifold since this can be accessed by InGaAs/GaAs diode lasers. In-band pumping can also occur at 1480 nm (6 800 cm^{-1}) to the higher levels of the $^4I_{13/2}$ manifold. For either pump band, excitation is followed by rapid non-radiative decay to the lowest level of the $^4I_{13/2}$ manifold, which acts as the upper laser level.

As for the case of Er:YAG, since the $^4I_{13/2}$ level lies well above the next lower level, those of the ground manifold, the rate of multiphonon decay of this level is low and the level decays predominantly radiatively with a fluorescence lifetime of approximately 8 ms. Lasing occurs from the bottom of the $^4I_{13/2}$ manifold to the levels of the ground-state manifold, $^4I_{15/2}$. Note that for the $^4I_{13/2} \rightarrow ^4 I_{15/2}$ transition, up-conversion[9] depletes the upper laser level; as a consequence the concentration of Er^{3+} ions must be kept low, generally to below $1 \times 10^{20}\,cm^{-3}$.

The Er:Glass laser is interesting in that it exhibits both three- and four-level characteristics. Figure 2.5 shows the measured gain coefficient as a function of wavelength for different pump intensities.

[8] Lasers operating between 1.45 and 1.70 μm are sometimes known as 'eye-safe' since radiation in this region is strongly absorbed by the cornea of the eye, and so cannot reach the retina. The term, however, is something of a misnomer since a sufficiently high-power laser operating in the 'eye-safe' region could still cause damage to the eye.

[9] Up-conversion proceeds through the same ion-ion interactions that cause concentration quenching. In up-conversion ion A decays from an intermediate level to the ground state, causing ion B to be excited from an intermediate level to a higher-lying level. If the two ions are of the same species this process is known as up-conversion, and can proceed very rapidly if the energies of the two transitions are closely-matched. As for concentration quenching, up-conversion transitions are non-radiative.

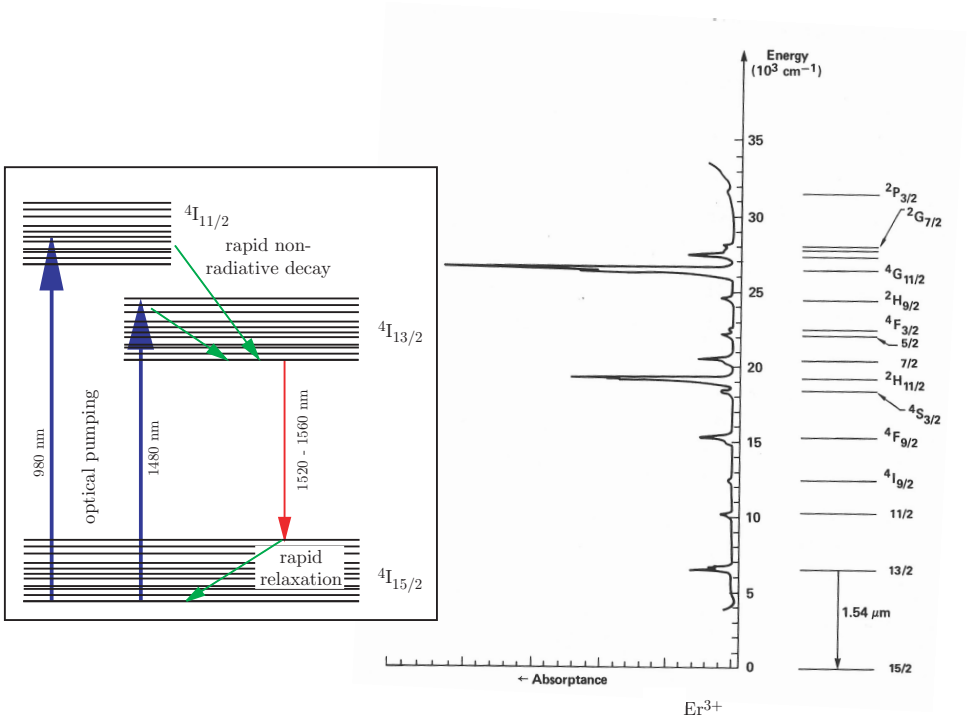


Figure 2.4: Energy level structure of the Er^{3+} ion in a glass host showing some possible laser transitions. Inset: simplified energy level diagram for the important transition at $\approx$ 1550 nm used in the EDFA.

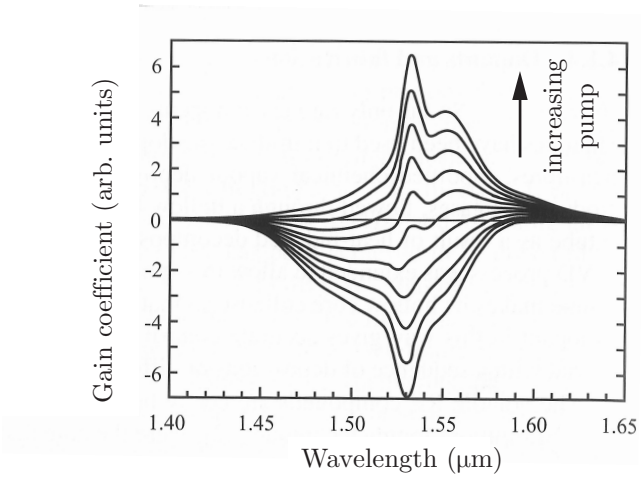


Figure 2.5: Measured gain coefficient as a function of wavelength for the $^4I_{13/2} \rightarrow ^4 I_{15/2}$ Er:Glass laser transition for different pump laser intensities.

Table 2.2: Important parameters of Er:YAG and Er:Glass lasers.

	Er:YAG		Er:Glass	
	$^4I_{11/2} \rightarrow ^4 I_{13/2}$	$^4I_{13/2} \rightarrow ^4 I_{15/2}$	$^4I_{11/2} \rightarrow ^4 I_{13/2}$	$^4I_{13/2} \rightarrow ^4 I_{15/2}$
λ (nm)	2940	1646		1540
τ_2 (μs)	100	7700	100	8000
τ_1 (μs)	7700	∞	8000	∞
$\Delta\nu$ (GHz)				3500
σ_{21} (10^{-20} cm^{-2})	2.6	0.5		0.7
κ (W m^{-1} K^{-1})	13	13	1	1

It is seen that at long wavelengths positive gain is achieved at low levels of pumping, i.e. four-level laser behaviour. In contrast, at shorter wavelengths gain only occurs for high pump intensities, which is characteristic of a three-level laser. This complex behaviour reflects the fact that the lower manifold of states extends over approximately 850 cm^{-1}, compared to $k_B T \approx 210$ cm^{-1}. Lasing at long wavelengths occurs to the top of the $^4I_{15/2}$ manifold which has only a low thermal population and which undergoes rapid relaxation. However, lasing at shorter wavelengths occurs to lower-lying Stark levels which have a large thermal population; this population must be significantly reduced by the pumping before a population inversion is achieved.

The properties of some Er lasers are listed in Table 2.2.

Er:Glass lasers have been operated in a variety of formats and in both continuous and pulsed mode. Q-switched operation is able to generate pulses of nanosecond duration, which is of interest in eye-safe laser-ranging. Alternatively, modelocked microlasers can generate pulses of approximately 20 ps duration with a pulse repetition rate as high as several GHz.

2.3 Trivalent iron group, $3d^n - 3d^n$ transitions

2.3.1 Energy level structure

The ground-state configuration of neutral atoms in the iron group of the transition metals is of the form [Ar] $3d^{n+1}\,4s^2$ or [Ar] $3d^{n+2}\,4s$, and hence that of the trivalent ions is [Ar] $3d^n$. The spectroscopy of these ions is very different from that of the $4f^n$ configuration in the trivalent rare earths owing to the fact that the size of the 3d orbitals are larger than those of the 4f orbitals in the trivalent rare earths, and the ionic size (and hence d) is smaller for the iron group than the rare earths. As a consequence the interaction with the crystal field is more than an order of magnitude bigger for the 3d electrons in the iron group than the 4f electrons in the rare earths, to the extent that the interaction with the crystal field is stronger than the spin-orbit interaction; the LS coupling scheme no longer applies.

It is highly instructive to consider in outline the form of the wave functions and energy levels in the strong crystal field limit. In this case the effect of the crystal field must be considered before the residual electrostatic and spin-orbit interactions, and consequently we determine the effect of the crystal field acting on the single-electron orbitals formed in the central potential of the active ion.

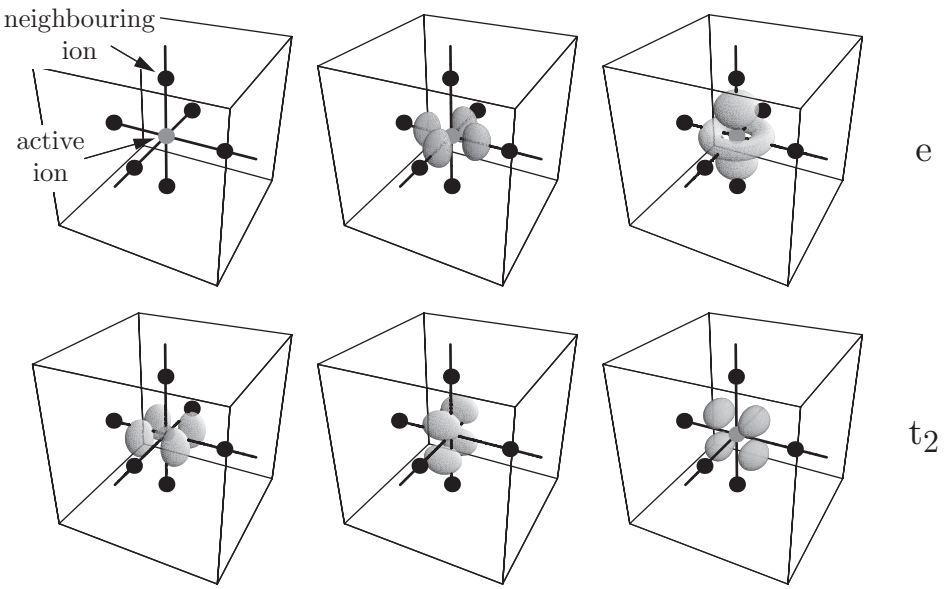


Figure 2.6: Crystal field orbitals of a d-electron in a field with octahedral symmetry. The top row shows the relative positions of the active ion and the neighbouring ions, and a surface of constant probability density for the two degenerate wave functions corresponding to the e_g orbitals. The bottom row shows equivalent plots for the three degenerate wave functions corresponding to the t_2 orbital. Notice that the wave functions of the e orbitals have lobes pointing towards the neighbouring ions, and hence have higher energy than the t_2 orbitals for which the lobes in the wave function point between the neighbouring ions.

The crystal field removes the spherical symmetry of the free ion, and hence ℓ is no longer a good quantum number. Instead the modified electron states are labelled by its symmetry properties, and by the electron spin, which is unchanged by the interaction with the crystal field.

As an example, we consider the case of a d-electron moving in a crystal field with octahedral symmetry.[10] Octahedral symmetry of the crystal field arises when the active ion is located at the centre of a cube, and identical neighbouring ions are located on each face of the cube, as illustrated in Fig. 2.6. In the absence of the crystal field the d-electron is five-fold degenerate, corresponding to $m = -2, -1, 0, 1, 2$. A crystal field with octahedral symmetry partially lifts this degeneracy to form a triply-degenerate crystal field orbital labelled t_2, and a higher-lying doubly-degenerate orbital e.[11] The spatial distributions of the t_2 and e orbitals are very different, as illustrated schematically in Fig. 2.6. The e wave functions have regions of high probability density centred on the axes connecting the

[10] We note that other, generally lower, symmetries of the crystal field are often treated by considering the difference between the true symmetry and octahedral symmetry as a perturbation which acts on the energy levels found for the case of octahedral symmetry.

[11] Just as in molecular spectroscopy, a subscript 'g' (standing for 'gerade') may also be included in the label to indicate that the orbitals have even parity, i.e. e_g and t_{2g}. Orbitals with odd parity can be labelled with a subscript 'u', standing for 'ungerade'.

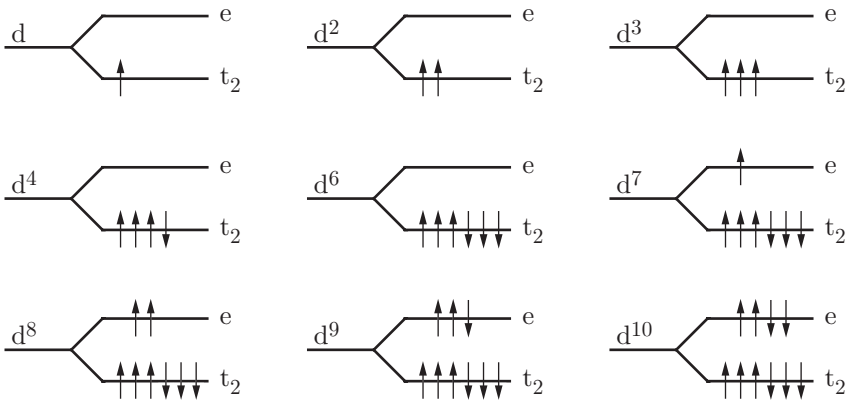


Figure 2.7: Schematic diagram showing the ground-state configurations arising from splitting of a d^n configuration a strong crystal field with octahedral symmetry. The arrows indicate electron spin.

active ion with the neighbouring ions, whereas for the t_2 orbitals these regions are centred between the axes. Consequently electrons in the e orbitals are on average closer to the neighbouring ions than those in the t_2 orbitals and therefore their interaction with the neighbouring ions — that is, with the crystal field — causes their energy to be increased from the unperturbed energy by a greater amount than electrons in the t_2 orbitals.

We may now consider the configurations formed in the crystal field as the number of d-electrons is increased. The ground state configuration of a single d-electron is simply t_2. Adding a second and third d-electron to the system gives rise to ground-state configurations comprised of two or three t_2 electrons respectively, since the t_2 orbital is three-fold degenerate. Since the spin-orbit interaction is minimized if the number of aligned spins is maximized,[12] the spins of the electrons will all point in the same direction.[13] A fourth d-electron added to the system may also occupy the t_2 orbital, but its spin must point in the opposite direction to satisfy the Pauli exclusion principle.[14] Considerations of this kind lead straightforwardly to the ground-state configurations illustrated schematically in Fig. 2.7.

The possible configurations of excited levels may be identified in a similar way. For example, for the case of three d-electrons two different excited configurations may be formed: one of the spins of the t_2 electrons may be reversed to give a configuration lying above the ground-state owing to the larger spin-orbit interaction; alternatively, one of the electrons can be promoted to the e orbital. The two types of excited configurations give rise to two types of transition, as illustrated in Fig. 2.8. In a **spin-flip** transition all the electrons remain in the same orbital, but the spin of one of the electrons is reversed, as illustrated in 2.8(a); in a **configuration transition** one electron makes a transition

[12]This is 'Hund's rule'.

[13]Hence the ground-state configurations of t_2^2 will have S = 1 (a triplet), and that of t_2^3 will have S = 3/2 (a quartet).

[14]We note that if the crystal field were weak, the ground-state configuration would be t_2^3e, with all spins parallel since this minimizes the spin-orbit interaction; in other words the increase in energy associated with promoting an electron to the e orbital would be smaller than the increase in energy caused by reversing the spin of the electron.

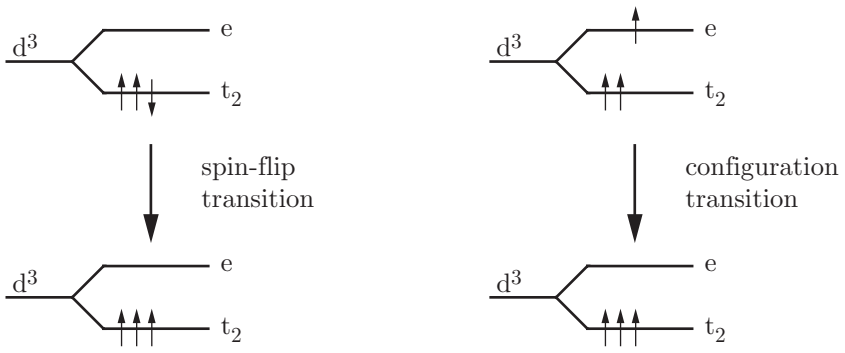


Figure 2.8: Spin-flip and configuration transitions of a d-electron in a crystal field with octahedral symmetry.

from an e orbital to a t_2 orbital.[15]

As we shall explore further below, these two types of transition have very different characteristics. In a spin-flip transition the spatial distribution of the total electronic wave function is unchanged. As a consequence the strength of the interaction of the active ion with the neighbouring ion is unchanged and, as we will see, this results in narrow-band transitions in which there is no change in the vibrational energy of the crystal lattice. Transitions of this type are known as **zero-phonon** transitions.[16]

In contrast, in a configuration transition the spatial distribution of the total electron wave function changes significantly. Since the interaction of the active ion with the neighbouring ions is very different for the upper and lower levels, the neighbouring ions move during the transition. Configuration transitions are therefore associated with changes in the vibrational motion of the lattice, and are characterized by broad-band, temperature-dependent lineshapes — the broad bandwidth arising from the emission or absorption of varying numbers and energies of lattice phonons. Transitions of this type are therefore known as **phonon-assisted** or **vibronic transitions** and are the basis of an important class of widely-tunable solid-state laser.

Configuration co-ordinate diagrams

Further insight into the properties of energy levels in the strong-coupling limit is provided by the **configuration co-ordinate diagram**. In this model the active ion and the surrounding ions are treated as a single system. To simplify matters this multi-dimensional problem[17] is reduced to one-dimension, and the distances between the active ion and the surrounding ions is parameterized by a 'configuration co-ordinate' Q. We can think of Q as representing some sort of average value of the

[15] The spin of the electron may, or may not, also change.

[16] The $4f^n - 4f^n$ transitions considered above are zero-phonon transitions since the interaction with the crystal field is weak for both the upper and lower levels.

[17] If there are N neighbouring ions there are in principle N different values of the spacing between the active ion and the nearest neighbours. Whether all N values are different depends on the nature of the lattice vibrations, as discussed in footnote 20.

distance between active and neighbouring ions.[18]

Within the configuration co-ordinate picture the Schrödinger equation for the active ion may be written in the form,

$$[T_{\rm n} + T_{\rm e} + V(Q, \boldsymbol{r})]\, \Psi(Q, \boldsymbol{r}) = E\Psi(Q, \boldsymbol{r}), \tag{2.3}$$

where: $\boldsymbol{r}$ represents the co-ordinates of the electrons; $T_{\rm n}$ is the kinetic energy operator of the nuclei of the active and surrounding ions; $T_{\rm e}$ is the sum of the kinetic energy operators of the electrons in the active ion[19]; and $V(Q, \boldsymbol{r})$ is the electrostatic potential energy, comprised of terms arising from the attraction of the electrons to the nuclei of the active and neighbouring ions, mutual repulsion between the electrons of the active and neighbouring ions, and interactions between the nucleus of the active ion and the electrons and nuclei of the neighbouring ions.

Just as for the case of diatomic molecules discussed last year, we may make the **Born-Oppenheimer approximation** and assume that the total wave function may be written as a product of an electronic wave function $\phi_{\rm e}(\boldsymbol{r})$ and a nuclear (or 'vibrational') wave function $\phi_{\rm n}(Q)$. We assume also that the electronic wave function depends only weakly on Q. It is straightforward to show that the Schrödinger equation may then be separated:

$$[T_{\rm e} + V(Q, \boldsymbol{r})]\, \phi_{\rm e}(\boldsymbol{r}) = E(Q)\phi_{\rm e}(\boldsymbol{r}) \tag{2.4}$$

$$[T_{\rm n} + E(Q)]\, \phi_{\rm n}(Q) = E\phi_{\rm n}(Q). \tag{2.5}$$

In principle, these equations may be solved as follows. The Schrödinger equation for the electronic wave function can be solved *for fixed* Q to yield an energy eigenvalue, $E(Q)$, that is a function of the configuration co-ordinate. We see from eqn (2.5) that this energy eigenvalue then acts as a potential well within which the nuclei of the ions move. The energy of the nucleus moving in this potential is the *total energy of the system* (i.e. the energy of the active ion and the vibrational motion of the lattice).

The potential well in which the ions move resembles that in which the nuclei of a diatomic molecule move, and hence in a similar way the ion will exhibit a ladder of vibrational energy levels as shown.

The strength of an optical transition between two electronic levels is proportional to the square modulus of the dipole (or higher order) matrix element. Since the dipole operator does not operate on the vibrational wave function $\phi_{\rm n}(Q)$, the strength of optical transitions between two electronic levels is proportional to the square of the overlap integral of the vibrational wave functions of the upper and lower level — just as in the case of transitions in a diatomic molecule.

An important parameter of the potential well is the value of the configuration co-ordinate, Q_0, for which the potential is a minimum; this is the equilibrium value of the configuration co-ordinate.[20] If

[18] Although see Note 20.

[19] We will ignore the electrons in the neighbouring ions since they will not be involved in transitions.

[20] We see now that Q need not be thought of as the average value of the distance between active and neighbouring ions. Rather, when $Q = Q_0$ all the neighbouring ions are in their equilibrium positions; when $Q = Q_0 + \delta Q$ all the neighbouring ions have increased their separation from the active ion by a distance δQ. The configuration co-ordinate model is therefore effectively considers the lattice vibrations to be such that all the neighbouring ions change their separation from the active ion by the same amount, and consequently can be represented by a single value of Q. Vibrational modes of this type are known as 'breathing modes.'

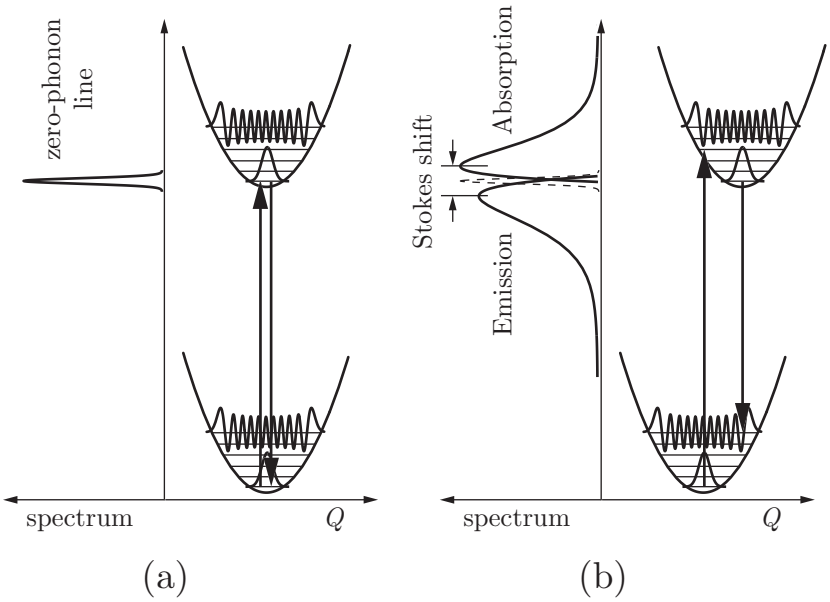


Figure 2.9: Transitions between electronic levels: (a) when the equilibrium configuration co-ordinate is essentially equal for the upper and lower electronic levels; (b) when the equilibrium configuration coordinates are different for the two levels, causing the potential curves to be displaced. Vibrational wave functions are illustrated schematically for the lowest vibrational level and an excited vibrational level. In (b) the position of the zero-phonon line is shown by the dotted curve.

the spatial distribution of the electronic wave functions is similar for the upper and lower electronic levels then the interaction of the ion with the neighbouring ions will be similar for the two levels, and hence the upper and lower potential wells will have similar values of the equilibrium values of the configuration co-ordinate and so lie vertically above one another, as shown in 2.9(a). In contrast, if the spatial distributions of the wave functions for the upper and lower levels is very different, the interaction of the electrons of the active ion will be different, and consequently Q_0 will be different for the two levels. This situation is shown in 2.9(b).

We may now use the configuration co-ordinate picture to deduce the form of the absorption and emission spectra of active ions for the two types of transition, as illustrated in Fig. 2.9. Suppose that the lower electronic level corresponds to the electronic ground-state, and let us consider optical excitation to the upper electronic level. The vibrational wave functions for the lowest vibrational level generally resemble a Gaussian curve, whilst those of high-lying levels are strongly peaked near the classical turning points.[21] Absorption from the lower electronic level will be dominated by transitions from the lowest vibrational level, since this will have by far the largest thermal population. The

[21] For a given vibrational energy level the classical turning points are the values of Q for which the potential curve has an energy equal to the total vibrational energy. In classical mechanics the kinetic energy of a particle is zero at such points, and hence they represent the extrema of its oscillatory motion. Classically, the probability of finding a particle is greatest at the turning points, since there it is stationary. The Correspondence Principle therefore tells us to expect the wave functions of high-lying (i.e. with large energy) vibrational levels to be strongly peaked at these points.

strongest absorption will therefore by from here to the lowest vibrational level of the upper electronic level, as shown in 2.9(a), since the overlap integral for these two vibrational wave functions will be large. The overlap integral will be comparatively small for transitions to high-lying vibrational levels since the vibrational wave function of such levels show rapid oscillations and are only large near the classical turning points, where the amplitude of the wave function of the lowest vibrational level of the ground state is small. In any case, excitation to excited vibrational levels of the upper electronic level will be followed by extremely rapid de-excitation to the lowest vibrational level by phonon de-excitation. The emission spectrum will therefore be dominated by emission from the lowest vibrational level of the upper electronic level. Just as for the absorption spectrum, and for the same reason, the emission will be dominated by transitions to the lowest vibrational level of the electronic ground-state. In other words, when the equilibrium co-ordinates of the upper and lower electronic levels are similar — or, equivalently, when the spatial distribution of the wave functions of the upper and lower levels are similar, such as in spin-flip transitions — the absorption and emission spectra are dominated by a single, zero-phonon transition as we would expect from our earlier considerations.

In contrast, consider the case shown in 2.9(b) in which the spatial distribution of the electronic wave functions are different for the upper and lower levels, and consequently the equilibrium configuration co-ordinates are different. Now absorption from the lowest vibrational level of the ground electronic state is distributed over many vibrational levels of the upper electronic level, leading to absorption extending to higher frequencies than that of the zero-phonon line. Absorption at frequencies above that of the zero-phonon line corresponds to electronic excitation accompanied by excitation of lattice vibrations. Following absorption, ions in excited vibrational levels of the upper electronic level will undergo rapid phonon de-excitation to the lowest-lying vibrational level. Just as for the absorption, emission will then occur to a wide range of vibrational levels of the ground electronic state, producing a broad emission spectrum shifted to frequencies below that of the zero-phonon line. The absorption and emission spectra are widely separated, the separation in frequency between the peaks of the two spectra being known as the Stokes shift. Hence, as expected from our earlier discussion, transitions between electronic levels with different spatial distributions (and hence different values of Q_0), such as configuration transitions, are broad-band, vibronic transitions. Note that the width of the absorption and emission bands, and the magnitude of the Stoke's shift, both increase with the magnitude of the difference in the equilibrium configuration co-ordinates of the two levels.

2.3.2 The ruby laser

The ruby laser is of enormous significance since it was the first optical maser. It is much less commonly used today than it once was, but it still finds applications in holography (where the relatively short wavelength for a solid-state laser is better suited to the response of photographic emulsions), pulsed interferometry, and some medical applications. It is also the best known example of a three-level laser.

Crystal properties

Ruby is formed by doping Cr^{3+} ions into a sapphire (Al_2O_3) host with a doping level[22] of approximately 0.05 wt%, corresponding to an active ion density of $1.58 \times 10^{19}\,\mathrm{cm}^{-3}$. The chromium ions, of radius

[22] In ruby gemstones the doping level of chromium ions is much higher, of order 1%, which gives them a rich red colour. Ruby laser rods typically appear pink.

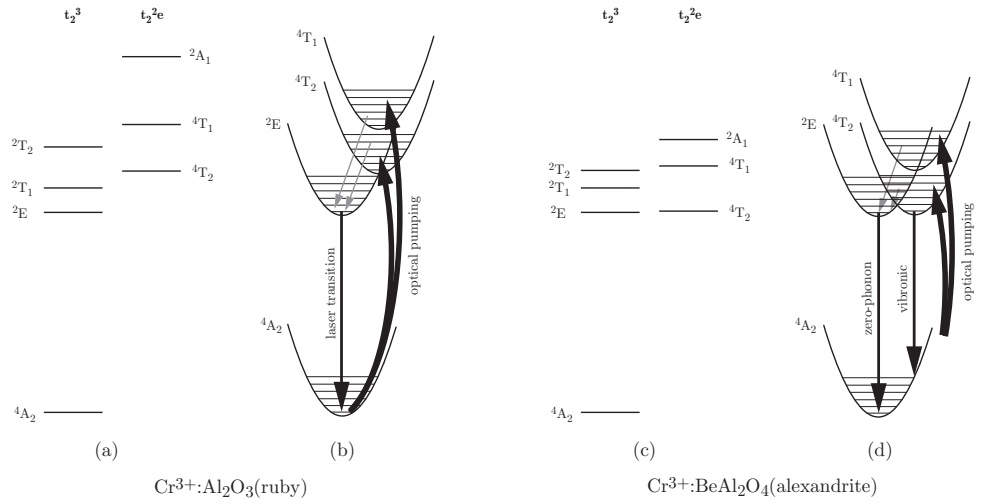


Figure 2.10: The crystal field configurations and energy levels (a and c) and simplified configuration co-ordinate diagrams (b and d) for ruby and alexandrite. Note that the diagrams are not to scale, and for clarity not all the energy levels shown in (a) and (c) are shown in the corresponding configuration co-ordinate diagram.

62 pm, substitute for the aluminium ions, of radius 53 pm, with little distortion of the host lattice. Ruby may be grown by the Czochralski method, and has excellent mechanical, thermal, and chemical properties. Ruby laser rods typically have a diameter of between 3 and 25 mm and a lengths as long as 200 mm.

Energy levels

Figure 2.10 shows the energy levels of Cr^{3+} ions doped in Al_2O_3 together with a simplified configuration co-ordinate diagram. In ruby the crystal field is strong, and hence, as discussed in Section 2.3.1 the crystal field configuration of the ground state is t_2^3. This configuration forms crystal field terms labelled[23] as $^4A_2$ (the ground-state), ^2E, $^2T_1$, and $^2T_2$ where the superscripts are equal to $2S + 1$. The lowest-lying excited configuration is t_2^2 e, which forms the levels $^4T_2$, $^4T_1$ and $^2A_1$. Of particular importance is the fact that since the crystal field is strong the $^4T_2$ and $^4T_1$ terms (which arise from the t_2^2e configuration) lie above the ^2E term (which arises from the t_2^3 configuration.[24]

Following the argument above, the wave functions of the levels of the t_2^2 e configuration have a very different spatial distribution than those of the t_2^3 configuration, and consequently, as shown in Figure 2.10 the equilibrium configuration co-ordinate of the $^4T_2$ and $^4T_1$ levels is quite different from that of the $^4A_2$ and ^2E levels.

The absorption spectrum of ruby is dominated by the $^4T_2 \leftarrow {}^4A_2$ and $^4T_1 \leftarrow {}^4A_2$ transitions since

[23] Just as for levels in a free atom or ion, single-electron orbitals are labelled with lower-case letters, crystal field terms are labelled with upper-case letters.

[24] The $^4T_2$ and $^4T_1$ levels may also be denoted in an alternative notation as $^4F_2$ and $^4F_1$ respectively.

these are fully spin-allowed. Since these are configuration transitions, as discussed in Section 2.3.1, they form broad, vibronic pump bands some 50 nm wide and centred at approximately 554 nm and 404 nm respectively.[25] These may be pumped conveniently by flashlamp radiation. Absorption to the doublet levels are much weaker since they are spin-forbidden.

Absorption on the $^4T_{2,1} \leftarrow {}^4A_2$ pump bands is followed by rapid non-radiative relaxation to the lowest-lying excited level: the ^2E level. This level is metastable since it can only decay on the spin-forbidden $^2E \rightarrow {}^4A_2$ transition; in fact it has one of the longest fluorescence lifetimes of any level in the solid state, approximately 3 ms. Since the $^2E \rightarrow {}^4A_2$ transition involves no change in the crystal field orbitals, it is a narrow-band, zero-phonon transition.

The ^2E level is actually split[26] into two levels separated by only $29\,cm^{-1}$. Population is rapidly transferred between these two levels by phonon collisions, so that the levels may be considered to be in thermal equilibrium, with a slightly larger population in the lower of the two ^2E levels. The transitions from the two levels of ^2E to $^4A_2$ are denoted R_1 and R_2.

Laser parameters

The R_1 and R_2 transitions are homogeneously broadened by phonon interactions to give a temperature-dependent linewidth, which at room temperature is approximately equal to 450 GHz. The optical gain cross-section of the R_1 transition in ruby is approximately $2.5 \times 10^{-20}\,cm^{-2}$.

The energy levels of ruby therefore have many desirable features for a laser transition: broad pump bands allow for efficient absorption of broad-band pump radiation; almost all the ions excited on these transitions are transferred by rapid non-radiative transitions to the upper laser level; the upper laser level has a long fluorescence lifetime, and therefore acts as a storage level for excitation. The main difficulty, of course, is that the laser transition occurs to the ground-state and consequently the laser is a three-level laser system as illustrated schematically in Fig. 2.11.

In order to achieve a population inversion the optical pumping must be sufficiently strong to deplete the population of the ground-state by approximately one-half, as discussed in the course last year. The splitting of the ^2E level causes the lasing transition to be split into two, closely-spaced lines denoted R_1 and R_2 with wavelengths of 694.3 nm and 692.8 nm respectively. Of these two transitions, the R_1 line at 694.3 nm usually dominates owing to the fact that it arises from the lower of the ^2E levels, and consequently has a higher thermal population.

Practical implementation

Since ruby is a three-level laser, it is almost invariably operated in a pulsed mode. Optical pumping is usually provided by a medium-pressure xenon flashlamp.

The long lifetime of the upper laser level means that it is straightforward to Q-switch the laser, and indeed this is necessary if strong laser spiking is to be avoided. Q-switched ruby lasers typically

[25] It is these strong absorption bands in the green and violet part of the spectrum which are responsible for the striking colour of ruby gemstones.

[26] The splitting arises from a combination of the spin-orbit interaction and lowering of the symmetry of the crystal field from octagonal symmetry. The $^4A_2$ ground state is also split by the same interactions, but its splitting is even smaller. The split levels of the ^2E level are denoted $^2E_{3/2}$ and $^2E_{1/2}$, the latter being the lower. In the alternative notation mentioned in Note 2.3.2 these levels are denoted $\bar{E}$ and $2\bar{A}$.

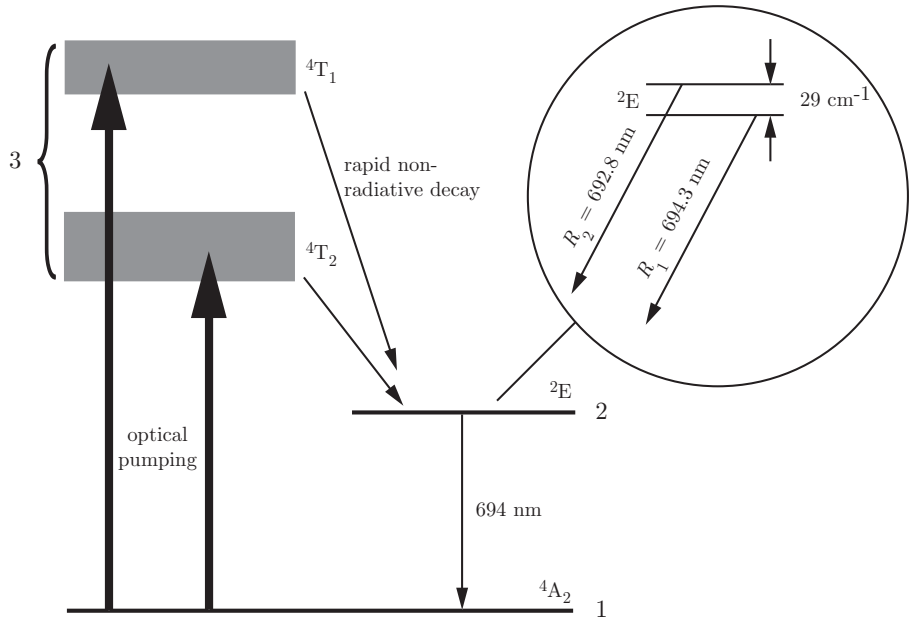


Figure 2.11: Simplified energy level scheme for the ruby laser. Inset: Splitting of the upper laser level.

generate pulses with an energy of several Joules, and a duration between 10 and 50 ns. Ruby lasers may also be actively or passively modelocked to give output pulses of 5 to 10 ps duration.

2.3.3 Alexandrite laser

Alexandrite, or chromium-doped chrysoberyl (Cr^{3+}:$BeAl_2O_4$), was the first commercially important tunable solid-state laser. Alexandrite lasers have proved useful in applications where their wide tunability is useful, such as detection of pollutant species in the atmosphere and in several medical applications.

Crystal properties

The Aluminium ions of the host material occupy two types of sites within the lattice: one in which the crystal field has a mirror symmetry, and one for which the field exhibits inversion symmetry. Typical doping levels are 0.04 - 0.1 at.%, the chromium ions substituting for aluminium ions, with a significant majority (78%) going into the sites with mirror symmetry.

The host is optically biaxial, leading to a strong polarization dependence on its optical properties which dominates thermal or stress-induced birefringence. The host has a very high thermal conductivity, a high optical damage threshold, is chemically stable, and has good mechanical strength and hardness. Laser rods of high optical quality may be produced with diameters of about 5 mm diameter and lengths up to 100 mm.

Table 2.3: Important parameters of the ruby, alexandrite, and Ti:sapphire laser transitions.

	Ruby	Alexandrite	Ti:sapphire
λ (nm)	694.3	680.4, 700 - 820	660 - 11 80
τ_2 (μs)	3000	260 (vibronic)	3.8
τ_1	∞	$\approx$ 100 ps (vibronic)	$\approx$ 100 ps
σ_{21} (10^{-20}cm^{-2})	2.5	1	0.34
$\Delta\nu$ (GHz)	330	330,26 000	100 000
Doping level (%)	0.05	0.1	0.1
Pump bands (nm)	375 - 425, 525 - 575	380 - 630	380 - 620

Energy levels

Figure 2.10(b) and (d) shows the energy levels of alexandrite. In this material the crystal field is of intermediate strength and consequently the $^4T_1$ and $^4T_2$ levels are closer to the ^2E level than is the case for Cr^{3+} ions in Al_2O_3 (i.e ruby). For example, in alexandrite the $^4T_2$ level lies only 800 cm^{-1} above the ^2E level, compared to approximately 1500 cm^{-1} in ruby.

As in the case of ruby, broad pump bands are provided by the vibronic $^4T_2 \leftarrow {}^4A_2$ and $^4T_1 \leftarrow {}^4A_2$ transitions lying at 550 - 650 nm and 350 - 450 nm respectively. These may be pumped by xenon or mercury flashlamps, or by argon and krypton ion lasers.

Absorption on the $^4T_2 \leftarrow {}^4A_2$ pump bands is followed by rapid non-radiative relaxation to the lowest vibrational level of the $^4T_2$ level and to the ^2E level. Since the $^4T_2$ level lies only 800 cm^{-1} above the ^2E level, population is transferred between these levels by rapid non-radiative transitions to form a thermal distribution over the two levels. The radiative lifetime of the ^2E level is similar to that of the same level in ruby: 2.3 ms; that of the $^4T_2$ level is somewhat shorter, 7.1 μs, since it may decay to the ground state via a spin-allowed transition. Since the population in these levels is shared, the effective lifetime of ions in either level depends on the relative population in each level, and is therefore strongly temperature dependent. At room temperature the effective lifetime of either level is approximately 300 μs.

Laser parameters

Lasing can occur in alexandrite on the analogous transition to the ruby laser transition: the zero-phonon $^2E \rightarrow {}^4A_2$ transition at 680.4 nm, which operates as a three-level scheme. However, of more importance is the $^4T_2 \rightarrow {}^4A_2$ since this is a vibronic transition and hence is widely tunable between approximately 700 nm and 818 nm.

The vibronic laser transition operates as a four-level laser: Pumping on the broad-band $^4T_{2,1} \leftarrow {}^4A_2$ transitions is followed by rapid non-radiative relaxation to the bottom of the $^4T_2$ level and to the ^2E level; the ^2E level acts as a population storage level, increasing the effective lifetime of the upper laser level considerably; lasing can then occur to excited vibrational levels of the $^4A_2$ ground electronic level, followed by rapid non-radiative decay to the lowest vibrational level of the ground electronic state. A simplified energy level diagram of the operation of the vibronic laser transition in

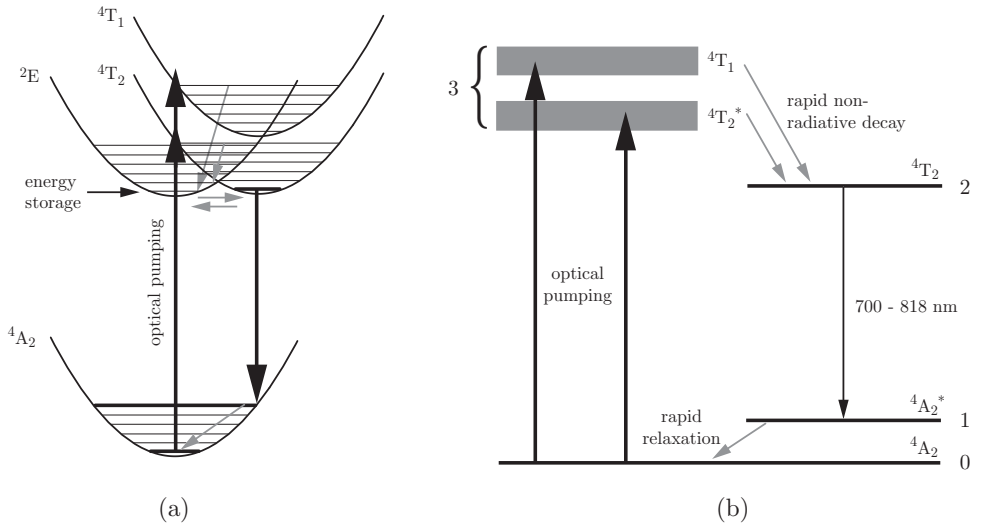


Figure 2.12: Simplified energy level scheme for the vibronic laser transition in alexandrite. (a) Simplified configurational co-ordinate diagram illustrating how the lowest vibrational level of the ^{2}E level acts as a storage level and is in thermal equilibrium with the lowest vibrational level of $^4T_2$; (b) Energy level scheme showing the four-level nature of the laser transition. Note in (b) an asterisk indicates that the level is in an excited vibrational level.

alexandrite is shown in Fig. 2.12.

Since the populations of the $^4T_2$ and ^{2}E levels are closely-coupled, it is appropriate to use an effective optical gain cross-section together with the total population in the two levels, as discussed on page 25 for the case of the Nd:YAG laser. The effective cross-section for the vibronic transition in alexandrite is approximately $1 \times 10^{-20}\,\mathrm{cm}^2$, although this is strongly dependent on temperature.

The main parameters of the laser transitions in ruby and alexandrite are given in Table 2.3

Practical implementation

Alexandrite lasers operate in a similar configuration to Nd:YAG lasers. The main difference is that the laser rod may be heated to between 50 and 70 °C in order to increase the relative population in the $^4T_2$ level.[27]

Alexandrite lasers are usually flashlamp-pumped. Although they may in principle be operated continuously, the relatively low gain cross-section means that the threshold population inversion density is high, and so pulsed operation is more usual. Pulse repetition rates between 10 and 100 Hz are typical, with output pulse energies of up to about 1 J.

Alexandrite lasers naturally produce output pulses with a duration close to the effective lifetime of the upper laser level, i.e. of 200 - 300 μs. Q-switching reduces the duration of the output pulse to

[27]Heating to higher temperatures would increase the population of the $^4T_2$ level still further. However, raising the temperature of the laser rod also increases the thermal population of the lower laser level, and hence the optimum temperature is below that which would maximize the thermal population of the upper laser level.

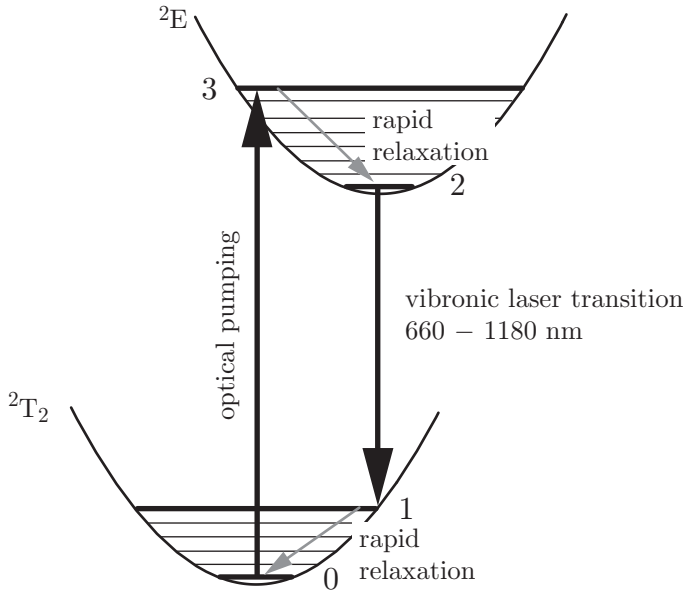


Figure 2.13: Simplified energy level structure and laser transitions of the Ti^{3+} ion in Al_2O_3 (Ti:sapphire).

typically 50 ns.

2.3.4 Ti:sapphire

Crystal properties

In titanium-doped sapphire, Ti:Al_2O_3, titanium ions (ionic radius 74.5 pm) substitute for aluminium ions (ionic radius 53 pm) with a typical concentration of 0.1 wt%. High-quality crystals may be grown by the Czochralski technique.

Energy levels

The Ti^{3+} ion has only a single d-electron, and consequently its energy level structure is rather simple: the ground crystal field configuration is t_2, which forms a $^2T_2$ term; the only other possible crystal field orbital for a d-electron is e, which gives rise to a ^{2}E term, as shown in Figure 2.13.

Radiative transitions between these two electronic levels obviously involve a change in the crystal field configuration, and are therefore broad, vibronic transitions. It is worth noting that the sparse energy level structure of Ti:sapphire means that losses arising from excited-state absorption are negligible.

Figure 2.14 shows the absorption and emission spectra of Ti:sapphire. The very broad bands, and their wide separation is a consequence of the large difference in equilibrium configuration co-ordinates for the upper and lower levels. Absorption from the ground electronic level occurs on the strong vi-

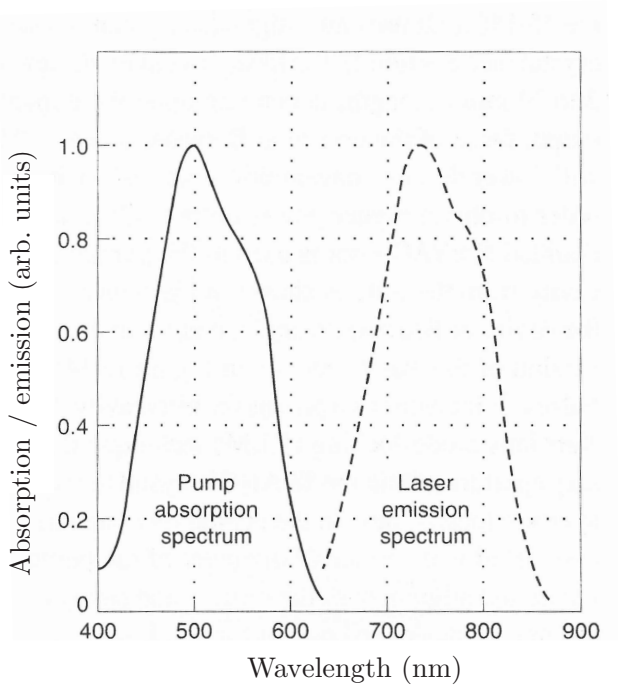


Figure 2.14: Absorption and emission bands of the Ti:sapphire laser.

bronic $^2\mathrm{E} \leftarrow {}^2\mathrm{T}_2$ transitions to the excited vibrational levels of the ^{2}E level, with wavelengths between 630 nm and 400 nm. Ions in excited vibrational levels decay rapidly by non-radiative transitions to the lowest vibrational level of the ^{2}E level, which forms the upper level of the laser transition. At room temperature ions in this level decay predominantly radiatively to the vibrational levels of the $^2\mathrm{T}_2$ level, with a fluorescence lifetime of approximately 3.2 μs, the relatively short lifetime reflecting the fact that the transition to the ground state is spin-allowed.

Laser parameters

The Ti:sapphire laser is a four-level laser: lasing may occur from the lowest vibrational level of the ^{2}E level to excited vibrational levels of the $^2\mathrm{T}_2$ level, at wavelengths tunable between 660 nm and 1180 nm. Indeed Ti:sapphire has the largest bandwidth of any laser.

Although the bandwidth of the laser transition is very large, the large radiative decay rate of the laser transition ensures that the optical gain cross-section remains relatively high for a solid-state laser: approximately $3 \times 10^{-19}\,\mathrm{cm}^2$, which is comparable to that in Nd:YAG

Practical implementation

A major difference between Ti:sapphire and all the lasers discussed so far in this chapter is the short fluorescence lifetime of the upper laser level. Flashlamp pumping is therefore not usually very efficient owing to the much longer duration of the flashlamp pulse. Instead, laser pumping is usually

employed, the most commonly employed pump lasers being frequency-doubled Nd lasers (YAG, YLF or vanadate).

Figure 2.15 shows the construction of two common types of Ti:sapphire laser oscillator. In both designs the Ti:sapphire laser rod is 2 - 10 mm long, with end faces cut at Brewster's angle so as to eliminate reflection losses for one polarization. The z-folded cavity compensates for astigmatism. This arises because it is necessary to focus the pump radiation to a small cross-sectional area in order to achieve sufficient population inversion. The design of the optical cavity design must therefore be such that the transverse modes are also brought to a focus at the Ti:sapphire crystal. However, the Brewster-cut faces of the crystal introduce astigmatism to the intra-cavity beam — i.e. the divergence of the beam becomes different in the horizontal and vertical planes. In the cavity shown in Fig. 2.15 this is corrected by the two concave mirrors used away from normal incidence, which introduce astigmatism of the opposite sign to that caused by refraction at the laser rod.

Continuously-tunable c.w. Ti:sapphire lasers may be tuned between approximately[28] 700 nm and 1000 nm, with output powers of 1 - 2 Watts. Wavelength tuning may be accomplished by a variety of techniques. In the design illustrated in Figure 2.15(a), tuning is achieved by adjusting a birefringent filter.[29] Other techniques, such as prisms or gratings, may be employed for selecting the output wavelength. For the oscillator shown in Fig. 2.15(a) the linewidth of the laser output is typically about 40 GHz; this may be reduced to better than 1 GHz by employing an intra-cavity etalon. Further reduction of the linewidth to around 10 MHz is possible by employing a ring cavity, and by locking the operating frequency to an external cavity the linewidth can be reduced to as little as 50 kHz.

Ti:sapphire lasers are widely used to generate very short optical pulses by the Kerr modelocking technique discussed in Lecture 4. The cavity design shown[30] in 2.15(b) is typical of modelocked Ti:sapphire lasers, and incorporates a pair of prisms for dispersion control as discussed in Lecture 5. The train of modelocked laser pulses from such a system typically has a pulse repetition rate of 80 MHz, a mean power of approximately 0.5 W (i.e. $\approx$ 5 nJ per pulse), with pulse durations of 35 - 50 fs. Modelocked oscillators of this type can also be wavelength tuned, although the tuning range is more restricted: 780 - 820 nm is common.

2.3.5 Host materials

The host into which the active ion of a solid-state laser is doped plays an important role in the operation of the laser through its physical properties and its influence on the energy level structure of the ion. Desirable physical properties include: hardness, so that the material surface may be polished

[28] The upper and lower limits to the tuning curve may be extended to approximately 100 nm and 675 nm respectively by employing cavity mirrors optimized for these wavelengths. However, it is not possible continuously to tune from 675 nm to 1100 nm without changing the cavity mirrors.

[29] At its simplest a birefringent filter is simply a plate of birefringent material with its optical axis in the plane of the plate. The retardance introduced by the plate depends on the angle between the optic axis and the electric field of the incident light, and on the wavelength of light. As the plate is rotated the wavelength(s) for which the retardance is equal to an integer times 2π varies. The polarization of these wavelength(s) is unchanged, unlike other wavelengths which in general become elliptically polarized, and consequently experience lower losses in passing through a linear polarizer (or any other polarization-sensitive element, such as the laser rod). In order to reduce losses, the filter can be mounted at Brewster's angle, and tuning achieved by rotating the filter about an axis parallel to its normal.

[30] In designing suitable cavities it is desirable to maximize the ratio of the round-trip losses under c.w. operation to those when Kerr lens modelocking occurs.

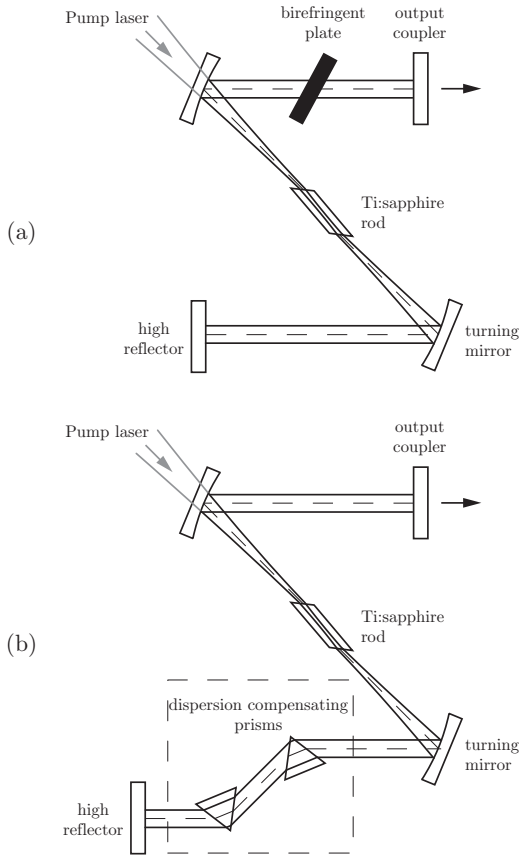


Figure 2.15: Two types of Ti:sapphire laser oscillator: (a) a wavelength-tunable oscillator; (b) a modelocked oscillator incorporating intra-cavity dispersion control.

to an optical-quality finish; high thermal conductivity, so that waste heat may be conducted away rapidly; and ease of growth, so that the material may be grown to the required size whilst maintaining high optical quality throughout the sample. It is also desirable that the crystal exhibits low thermal lensing[31] and low thermally-induced birefringence.[32] At the microscopic level, the host material must be capable of accepting the active ion into its structure.

Solid state hosts fall into three classes: crystals, glasses, and ceramics, as discussed below.

[31]Thermal lensing refers to a spatial variation of the refractive index of the laser gain medium arising from spatial variations in temperature and thermal stresses within the material. Frequently the induced variation in refractive index – and, for a solid material, possible changes in the surface figure of the laser rod or disk – causes the gain medium to act as a thin lens of focal length f. For solid-state lasers f may be positive or negative, with a magnitude that can be as small as a few 10s of centimetres. Further details may be found in Koechner and Bass *Solid-State Lasers: A Graduate Text*, Springer (2003).

[32]The change in refractive index resulting from thermally-induced strain within the crystal can be different for different polarizations. For example, in a laser rod the change in refractive index will in general be different for the radial and azimuthal polarization components.

Crystalline hosts

Crystalline hosts have thermal conductivities which are typically an order of magnitude greater than in glass hosts, and can therefore sustain higher mean pumping powers before the thermal lensing of the laser rod becomes problematic. Laser transitions in crystal hosts are usually homogeneously broadened, which makes single mode operation easier to achieve, and gives narrower linewidths for non-vibronic transitions (see Section 2.3.1). Disadvantages of crystal hosts are that they are often birefringent, making them polarization sensitive, and it can be difficult to grow large crystals of high optical quality.

Glass hosts

Glass hosts have several advantages over crystals: glass can be cast into many forms — from optical fibres to metre-sized discs — with good homogeneity and high optical quality; glass is inexpensive; and it has low, or zero, birefringence. Further, to some extent, the physical properties of the glass and the properties of the laser transition can be controlled by adjusting the composition of the glass. The biggest drawback of glass is the relatively low thermal conductivity, which makes thermal management more difficult. Optical transitions within ions doped into a glass host are always inhomogeneously broadened, with linewidths several orders of magnitude larger than non-vibronic transitions in crystalline hosts. This offers advantages in providing broad pump bands, but means that the laser transition will also be broad. As a consequence, the optical gain cross-section will be low, making it more difficult to achieve laser oscillation in an oscillator owing to the lower round-trip gain (although this restriction does not apply to fibre lasers). The low gain cross-section has significant advantages, though, in laser *amplifiers*. In an amplifier an upper limit on the single-pass gain is set by the onset of parasitic laser oscillation caused by small Fresnel reflections from the ends or edges of the laser rod, or depletion of the gain by strong amplified spontaneous emission. The low gain cross-section obtained with glass laser materials allows a larger population inversion to be generated before this is a problem, and this allows more energy to be extracted by an injected laser pulse.

To summarize: the higher gain and good thermal conductivity of crystalline hosts means that they are often employed in laser oscillators and lower-power amplifiers; glass hosts are frequently used in high-energy pulse amplifiers (with low mean power) owing to the low optical gain cross-section and ease with which large rods and discs may be fabricated.

Lecture 3

Q-Switching

3.1 Laser Spiking and relaxation oscillations

Two types of transient behaviour are frequently observed in the output of laser systems:

1. **Laser spiking:** This can occur when a laser is first turned on, and comprises large amplitude, somewhat irregularly-spaced, pulses of laser radiation.

2. **Relaxation oscillations:** These are more regular oscillations in the output power of the laser, typically with an exponentially-decaying amplitude. Relaxation oscillations can occur in the tail of a series of laser spikes, or can follow a small perturbation to the laser system such as disturbance of a cavity mirror, or a change in the pumping power.

The two phenomena are related; both arise from the fact that the population inversion density and the intra-cavity laser intensity are coupled together via stimulated emission and feedback from the cavity mirrors, and that the inversion density and intra-cavity intensity respond to changes in the operating conditions of the laser on different time scales.

Consider first the case of laser spiking. Immediately after the pumping is switched on, the population in the upper laser level will grow from zero in a time of order the upper level lifetime τ_2. At this point the density of photons in the cavity will be very low since the photons arise from spontaneous emission only. At some point the population inversion density will reach the threshold for laser oscillation, $N_{\rm th}^*$, but laser oscillation will not commence since the density of photons in the relevant cavity mode will still be low. As pumping continues, the population inversion grows beyond the threshold value, and the density of photons in the laser mode will build up by stimulated emission. Eventually the density of photons in the laser mode will be sufficiently high to burn the population inversion down towards $N_{\rm th}^*$, further increasing as it does the rate of stimulated emission and the photon density. Inevitably, the population inversion will be driven well below the threshold value, leading to reduction in the photon density and the cessation of lasing. The dramatic growth and rapid decline of the intra-cavity photon density correspond to a spike in the output power of the laser. As pumping continues, the population inversion will start to grow again and the cycle can repeat. Laser spiking of this type will not usually continue indefinitely, even in continuously-pumped lasers, owing to the fact that the population inversion and photon densities do not decay to zero after the spike. Consequently,

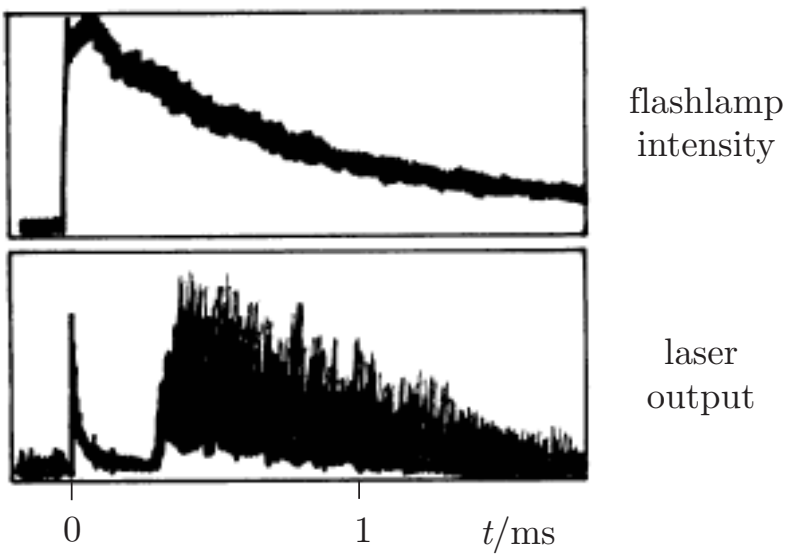


Figure 3.1: Laser spiking in the first ruby laser. The top trace shows the flashlamp pulse, the lower the output from the ruby laser as recorded by a photodiode. Note that the sharp feature at $t = 0$ is caused by flashlamp light reaching the photodiode.

after each spike these quantities tend to end up closer to their steady-state values, in which case the spiking will decay. Figure 3.1 shows a typical example of laser spiking from the first ever ruby laser.

It should be noted that laser spiking is often associated with rapid jumps in the longitudinal and/or transverse mode of the laser. This occurs for essentially the same reasons as spatial hole-burning in lasers operating on homogeneously broadened transitions: immediately after a laser spike, significant population inversion may remain in regions near nodes of the longitudinal nodes, or in regions where the transverse profile of the mode of the previous pulse was low in intensity. This gain can be exploited by other modes which, as pumping continues, may therefore reach threshold sooner than the mode (or modes) of the previous spike.

Laser spiking tends to occur when the time scale on which the population inversion responds to the pumping, typically of order τ_2, is slow compared to that in which the photon density changes in the cavity, which is of order the cavity lifetime τ_c. Hence we expect laser spiking to occur if,

$\tau_2 \gg \tau_c$ Approximate condition for laser spiking. (3.1)

3.1.1 Rate equations

For a homogeneously broadened laser the rate equations for the upper and lower laser level may be written simply as[1],

$$\frac{dN_2}{dt} = R_2(t) - N^*(t)\sigma_{21}\frac{I(t)}{\hbar\omega_{\rm L}} - \frac{N_2(t)}{\tau_2} \tag{3.2}$$

$$\frac{dN_1}{dt} = R_1(t) + N^*(t)\sigma_{21}\frac{I(t)}{\hbar\omega_{\rm L}} - \frac{N_1(t)}{\tau_1} + N_2(t)A_{21}, \tag{3.3}$$

where R_i is the pump rate of level i, and $\omega_{\rm L}$ and $I(t)$ are the angular frequency and total intensity of the oscillating laser mode.

Notice that in the absence of stimulated emission ($I(t) = 0$) the population of the upper laser level grows in a time of order τ_2.

The rate per unit volume at which photons are stimulated into the oscillating mode is simply $N^*\sigma_{21}I/\hbar\omega_{\rm L}$. Hence if we suppose that the fraction of the volume of the oscillating mode lying within the gain medium is equal to $f_{\rm c}$, the rate equation for the density of photons is simply,

$$\frac{dn}{dt} = f_{\rm c}N^*(t)\sigma_{21}\frac{I(t)}{\hbar\omega_{\rm L}} - \frac{n(t)}{\tau_{\rm c}}. \tag{3.4}$$

The last term gives, in terms of the **cavity lifetime** $\tau_{\rm c}$,the rate of loss of photon density by coupling through the output mirror (plus any other losses). We derive an expression for $\tau_{\rm c}$ in Section 3.1.2.

This equation may be simplified by noting that $I(t) = n(t)\hbar\omega_{\rm L}c$. Then,

$$\frac{dn}{dt} = \left(\frac{N^*}{N^*_{\rm th}} - 1\right)\frac{n}{\tau_{\rm c}}, \tag{3.5}$$

where,

$$N^*_{\rm th} = \frac{1}{f_{\rm c}\sigma_{21}c\tau_{\rm c}} \tag{3.6}$$

and we have dropped the time dependence of $N^*(t)$ and $n(t)$ to avoid clutter. We see that $N^*_{\rm th}$ is the threshold population inversion density since we must have $N^* > N^*_{\rm th}$ for the photon density to grow (by stimulated emission). Equation (3.5) also shows that the time scale in which the photon density changes is $\tau_{\rm c}$ divided by a factor which depends on the extent to which the population inversion exceeds $N^*_{\rm th}$. As such we expect the photon density to change in a time of order $\tau_{\rm c}$, as stated earlier.

[1]In writing the rate equations in this form we have made several simplifying assumptions: we assume that the laser oscillates on a single longitudinal and transverse mode; and we ignore transverse variations in the intra-cavity laser intensity and in the population inversion.

3.1.2 Cavity lifetime

To calculate the cavity lifetime we consider some photons circulating within a laser cavity. In the absence of further photons being supplied by any source (an injected beam, or by spontaneous or stimulated emission), the number of photons within the cavity will decrease with time. Suppose that at some moment the cavity contains n photons per unit volume. The flux — i.e. the number of photons per unit area, per unit time — of photons striking each mirror is $\frac{1}{2}nc$, the factor of one-half arising since half the photons travel to the left, and half to the right. For each mirror i the proportion of these photons which are transmitted is equal to T_i. Hence, ignoring any other losses, the density of photons within the cavity will decay according to

$$\frac{\mathrm{d}n}{\mathrm{d}t} = -\frac{\frac{1}{2}ncA_{\mathrm{mode}}\left(T_1 + T_2\right)}{V_{\mathrm{mode}}},$$

where A_{mode} and V_{mode} are the area and volume of the oscillating mode. Hence we may write,

$$\frac{\mathrm{d}n}{\mathrm{d}t} = -\frac{n}{\tau_{\mathrm{c}}},$$

where the cavity lifetime is given by

$$\tau_{\mathrm{c}} = \frac{2V_{\mathrm{mode}}}{A_{\mathrm{mode}}c}\frac{1}{T_1 + T_2}.$$

Now $V_{\mathrm{mode}} \approx A_{\mathrm{mode}}L_{\mathrm{c}}$, and hence

$$\tau_{\mathrm{c}} = \frac{2L_{\mathrm{c}}}{c}\frac{1}{T_1 + T_2} \approx \frac{L_{\mathrm{c}}}{c}\frac{1}{1-R} \qquad \text{Cavity lifetime}$$

where $R = \sqrt{R_1 R_2}$ and the approximation holds if $R_1 \approx R_2 \approx 1$.

3.1.3 Numerical analysis of laser spiking

Fig. 3.2 shows numerical solutions of these rate equations for a system with a filling factor f_{c} of unity and constant pumping at a rate corresponding to an over-pumping ratio (see Section 3.1.4) $r = 4$. In each of these simulations the pumping would, in the absence of stimulated emission, produce a population inversion density which was a factor of r greater than the threshold value. However, once the population inversion rises above N_{th} the rate of stimulated emission exceeds the rate of loss of photons from the cavity, increasing the density of photons within the cavity, and causing in turn the population inversion to burn down towards the threshold value. In all the cases shown the population inversion and photon densities tend to their steady-state values N_{th} and n_{th} respectively at long times after the pumping is switched on at $t = 0$. Substantial laser spiking is observed when $\tau_2/\tau_{\mathrm{c}} \gg 1$, consistent with our discussion above, and the number and amplitude of the spikes in the laser output

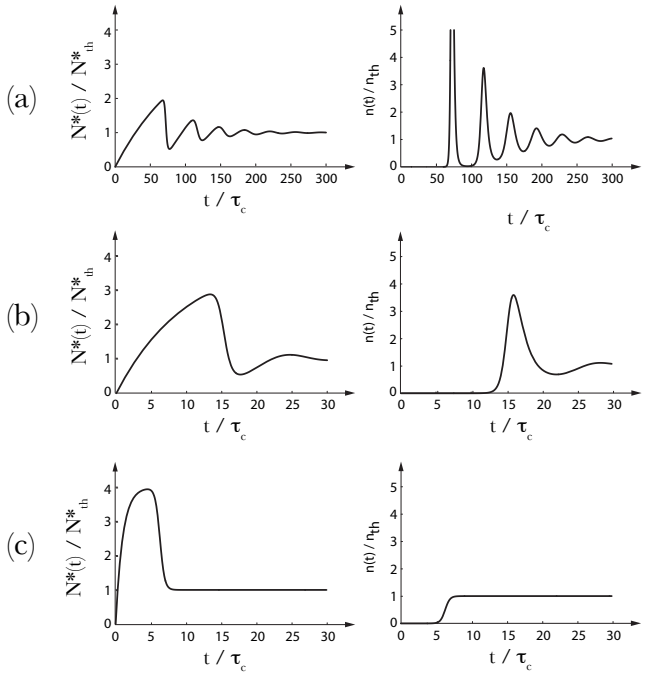


Figure 3.2: Numerical analysis of laser spiking in an ideal four-level laser with a filling factor f_c of unity pumped with an over-pumping ratio $r = 4$ and: (a) $\tau_2/\tau_c = 100$; (b) $\tau_2/\tau_c = 10$; (c) $\tau_2/\tau_c = 1$. In each case the plots show the temporal evolution of the population inversion density and the photon density normalized to the steady-state value, n_{th}.

is seen to decrease as the ratio τ_2/τ_c is reduced. In particular, for the case $\tau_2/\tau_c = 1$ there is essentially no spike in the photon density. Instead the population inversion density photon density reaches rN_{th} before the smoothly-rising photon density burns the population inversion down to the steady-state value.

3.1.4 Analysis of relaxation oscillations

Relaxation oscillations, being somewhat gentler than laser spiking, are amenable to algebraic analysis. We consider the simple example of an ideal homogeneously broadened, four-level laser system in which the population of the lower laser level is essentially zero at all times. In this case eqn (3.2) is also the rate equation for the population inversion density:

$$\frac{dN^*}{dt} = \frac{dN_2}{dt} = R_2 - N^*\sigma_{21}\frac{I}{\hbar\omega_L} - \frac{N_2}{\tau_2} \tag{3.7}$$

$$= R_2 - \frac{1}{f_c}\frac{N^*}{N^*_{th}}\frac{n}{\tau_c} - \frac{N^*}{\tau_2} \tag{3.8}$$

Notice that in the absence of any stimulated emission ($I = 0$) the population inversion would reach a steady-state value of $R_2\tau_2$.

Under steady-state conditions $N^*(t) = N^*_{\rm th}$. Substituting this into eqn (3.8) then yields an expression for the steady-state photon density,

$$n_{\rm th} = \left(R_2 - \frac{N^*_{\rm th}}{\tau_2}\right) f_{\rm c}\tau_{\rm c} \tag{3.9}$$

$$= (r-1)\frac{N^*_{\rm th} f_{\rm c}\tau_{\rm c}}{\tau_2}, \tag{3.10}$$

where the parameter $r = R_2\tau_2/N^*_{\rm th}$ is the ratio of the population inversion that would be produced in the absence of any stimulated emission to $N^*_{\rm th}$. It is also equal to the over-pumping ratio that will appear in our discussion of Q-switching.

We now suppose that having established steady-state conditions, the laser is subject to a small perturbation that shifts the population inversion and photon densities to:

$$N^*(t) = N^*_{\rm th} + \Delta N^*(t) \tag{3.11}$$

$$n(t) = n_{\rm th} + \Delta n(t) \tag{3.12}$$

We may then linearize the differential equations for $N^*(t)$ and $n(t)$ by substituting eqns (3.11) and (3.12) into (3.5) and (3.8), and ignoring the small product terms in $\Delta N^*(t)\Delta n(t)$. We find:

$$\frac{d\Delta N^*}{dt} = -r\frac{\Delta N^*}{\tau_2} - \frac{1}{f_{\rm c}}\frac{\Delta n}{\tau_{\rm c}} \tag{3.13}$$

$$\frac{d\Delta n}{dt} = (r-1)f_{\rm c}\frac{\Delta N^*}{\tau_2} \tag{3.14}$$

We now assume solutions of the form,

$$\Delta N^*(t) = a\exp(mt) \tag{3.15}$$

$$\Delta n(t) = b\exp(mt). \tag{3.16}$$

Substitution into eqns (3.13) and (3.14) then gives,

$$\left(m + \frac{r}{\tau_2}\right)a + \frac{1}{f_{\rm c}\tau_{\rm c}}b = 0 \tag{3.17}$$

$$(r-1)\frac{f_{\rm c}}{\tau_2}a - mb = 0. \tag{3.18}$$

These two simultaneous equations can be solved by the usual matrix method: non-trivial solutions require the determinant to be zero, yielding,

$$m = -\frac{r}{2\tau_2} \pm \sqrt{\left(\frac{r}{2\tau_2}\right)^2 - \frac{r-1}{\tau_2\tau_c}}. \tag{3.19}$$

Clearly, relaxation oscillations will occur if m is complex, i.e. if

$$\left(\frac{r}{2\tau_2}\right)^2 < \frac{r-1}{\tau_2\tau_c} \tag{3.20}$$

$$\Rightarrow \tau_2 > \frac{r^2}{r-1}\frac{\tau_c}{4}. \tag{3.21}$$

In other words if the upper level lifetime is sufficiently short, relaxation oscillations do not occur and perturbations to $N^*(t)$ and $n(t)$ decay exponentially back to their steady-state solutions. Note that for $r \gg 1$ eqn (3.21) is in agreement with our earlier estimate (3.1).

If the above condition is not met relaxation oscillations will occur with an angular frequency

$$\omega_{ro} = \sqrt{\frac{r-1}{\tau_2\tau_c} - \left(\frac{r}{2\tau_2}\right)^2} \approx \sqrt{\frac{r-1}{\tau_2\tau_c}}, \tag{3.22}$$

where the approximation holds if $\tau_2 \gg \tau_c$. Notice that very often $r - 1 \approx 1$, in which case ω_{ro} is close to the geometric mean of the upper level and cavity decay rates.

The condition (3.21) explains why relaxation oscillations (and also laser spiking) are unusual in gas lasers. Such systems typically operate on allowed transitions with upper level lifetimes of 10s to 100s of nanoseconds[2]. This is reasonably short compared to a typical cavity lifetime of 10 ns. In contrast, the upper level lifetime for many solid state laser transitions is of the order of microseconds or even milliseconds. In such cases laser spiking and relaxation oscillations are to be expected.

3.2 Q-switching

Q-switching is a technique that generates pulses of laser light with a peak power very much greater than the mean power that can be produced in steady-state operation. The high peak power is particularly useful in applications that require a high peak intensity, such as: laser cutting or drilling; or in pumping other laser systems such as Ti:sapphire amplifiers.

In a Q-switched laser the cavity is deliberately spoiled — imagine, for the moment, that one of the cavity mirrors is blocked — whilst the pumping builds up the population inversion. Since the losses of the modified cavity are so high, the population inversion can build up to much greater levels than the threshold inversion for lasing in the normal cavity without any lasing occurring. Once the population

[2]A notable exception is the CO_2 laser which has an upper level lifetime of 4 ms, and hence can exhibit relaxation oscillations.

inversion has grown to a large value the mirror is unblocked, thereby switching the 'Q' of the cavity from a low value to a high value. High rates of stimulated emission then cause the density of photons in the cavity to grow very rapidly, burning the population inversion density well below the threshold value. The output from the laser takes the form of a giant pulse of radiation — essentially a single laser spike of the type discussed above.

Note that the pumping in Q-switched lasers may be pulsed or continuous. With pulsed pumping the ratio of the peak population inversion to the threshold value can be larger than can be achieved with continuous pumping, leading to a higher peak output power (although the mean output power may be similar).

3.2.1 Techniques for Q-switching

A wide variety of techniques have been used to implement Q-switching.

Rotating mirror

Historically one of the earliest techniques was to rotate one of the cavity mirrors with a motor. The cavity would then be completed briefly whenever the mirror was rotated into alignment. It is best to use a roof prism as shown schematically in Figure 3.3, rather than a plane mirror, since any ray incident in a plane perpendicular to the axis of the roof is reflected in the anti-parallel direction to the incident ray. Hence for the system illustrated alignment in the vertical plane is automatic — this is useful given the inevitable vibration in a mechanical system. Alignment in the horizontal plane is achieved once per rotation.

The disadvantages of this approach are obvious. The switching is rather slow and the timing of the output pulse has a relatively high uncertainty. In order to estimate the switching time, we suppose that the mirror has to be aligned to an angular tolerance about equal to the divergence of the oscillating cavity mode. This divergence will typically be of the order of 1 mrad. Hence, for a motor running at 24 000 r.p.m. the switching time is about 400 ns. The disadvantages are such that this technique is now used only very rarely.

Electro-optic switching

A more modern approach is the use of an electro-optic switch, as illustrated schematically in Figure 3.4. The refractive index of an electro-optic material changes when it is subjected to an external electric field. In the **Pockels effect**[3], observed in non-centro-symmetric materials, the induced refractive index varies linearly with the applied electric field. Suitable materials are KD*P or lithium niobate for visible to near infrared wavelengths, or cadmium telluride for the mid-infrared. In the **Kerr effect**[4] the change in refractive index is proportional to the square of the electric field. We note that all materials exhibit the Kerr effect to some extent. Since the Kerr effect arises from higher-order processes than in the Pockels effect (via the third-order susceptibility rather than the second-order), larger electric fields are usually required. Since these must be switched quickly for Q-switching to be effective, it is more usual to use the Pockels effect in this application.

[3]Named after the German physicist F. C. A. Pockels (1865 - 1913).

[4]Named after the Scottish physicist John Kerr (1824 - 1907).

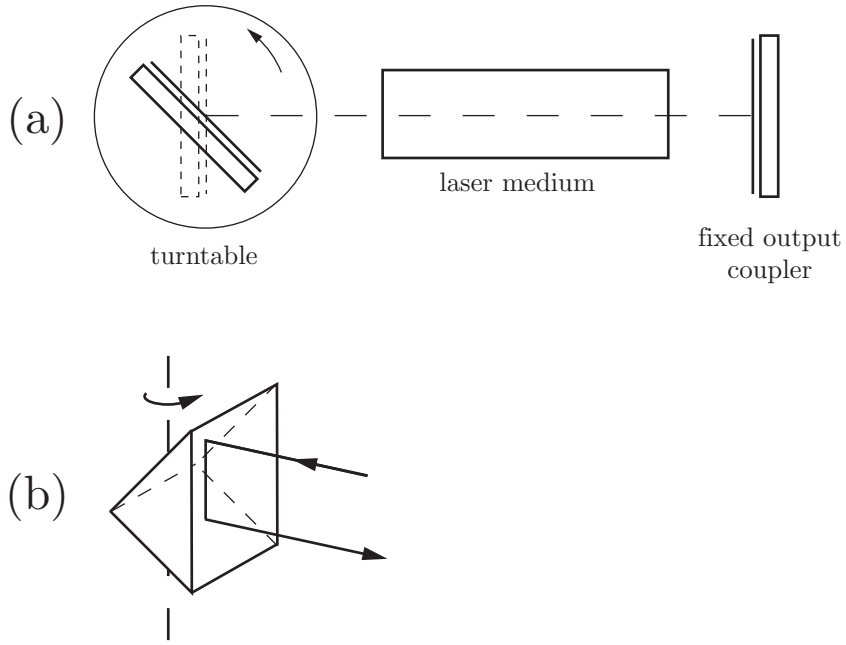


Figure 3.3: Q-switching with a rotating mirror: (a) schematic diagram of the layout of the optical cavity; (b) use of a roof prism as the rotating element. A roof prism has the advantage of automatically ensuring alignment of the reflected beam in one plane (in this case the vertical plane).

We can treat electro-optic devices of this type as wave plates with a birefringence which depends on the strength of the applied electric field. In the arrangement shown in Figure 3.4 a Pockels cell is oriented with its axis at 45^o to the axis of a polarizer. The cavity is held in a low-Q state by applying a certain voltage (the 'quarter-wave' voltage) to the Pockels cell, which then behaves as a quarter-wave plate. Consequently vertically-polarized light transmitted by the polarizer becomes circularly polarized after it transmission through the Pockels cell. Upon reflection from the end mirror of the cavity the handedness of the circular polarization changes, and hence on passing through the Pockels cell again it is converted to horizontally-polarized light and so reflected out of the cavity by the polarizer. Hence, with the quarter-wave voltage applied to the Pockels cell the end mirror of the cavity is effectively blocked.

The mirror can be unblocked by switching the voltage across the Pockels cell to zero, which reduces the birefringence of the Pockels cell to zero[5]. Vertically-polarized light can now pass through the polarizer and Pockels cell with essentially zero loss, and the Q-switched pulse can build up. A Kerr cell would be employed in essentially the same way.

Electro-optic switching has the advantages of fast switching times (of the order of 10 ns) and a high hold-off ratio, which allows the population inversion to be built up to many times the threshold value

[5]It would also be possible to devise an arrangement in which the low-loss condition was reached by switching *on* the Pockels cell. However, it is usually technically easier to switch off voltages rapidly — a technique known as 'crowbaring' — than to switch them on quickly.

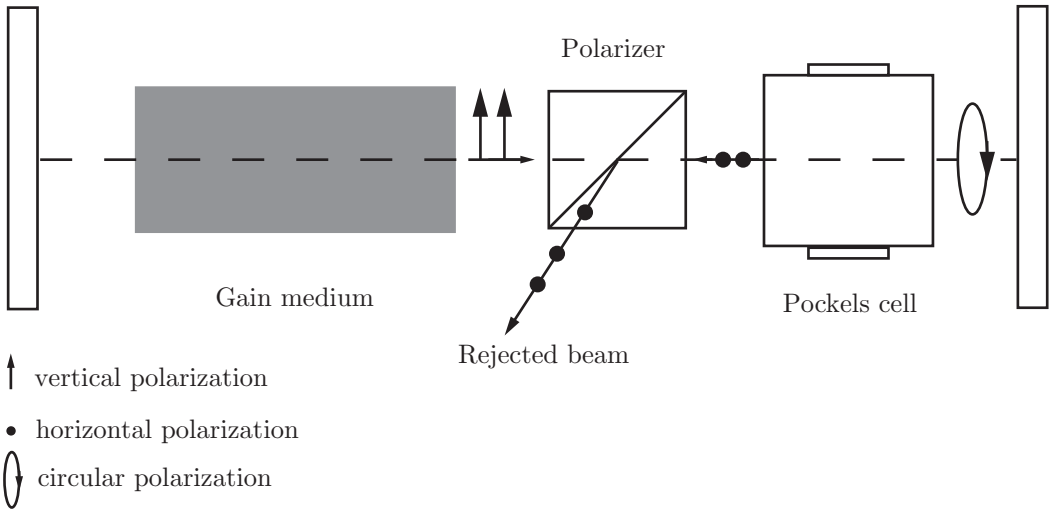


Figure 3.4: Electro-optic Q-switching. The system is shown with the Pockels cell voltage set so as to reject the beam reflected from the right-hand mirror.

for the high-Q cavity. The disadvantages are that the technique requires several optics to be placed in the cavity (each with some insertion loss and subject to possible damage), and the devices and associated electronics are relatively expensive. However, the advantages are such that the technique is very widely used.

Acousto-optic switching

If an acoustic wave is launched through an acousto-optic crystal the refractive index is changed slightly at the peaks and troughs as a result of the local expansion or contraction of the crystal lattice. The resulting periodic variation in the refractive index will cause incident radiation to be diffracted. We can think of this as Bragg reflection from the high- and low-index planes established in the crystal.

A wide variety of acousto-optic materials are available, including: fused silica, tellurium dioxide and lithium niobate for visible wavelengths; gallium arsenide for use in the infrared; and silicon for radiation in the 10 μm region.

Figure 3.5 shows an acousto-optic switch employed in a Q-switched laser. With the radio-frequency (RF) voltage applied to the crystal an acoustic wave propagates along the length of the crystal and is damped at the opposite end, the damping often being achieved by cutting the end of the crystal to form a wedge. Some fraction of the laser radiation incident on the crystal will be Bragg reflected away from the cavity axis and hence lost from the cavity mode. With the RF voltage applied, then, the acousto-optic crystal increases the losses of the laser cavity. Turning off the RF voltage removes this additional loss, and sets the cavity to a high-Q condition.

The advantages of acousto-optic switching are that it is relatively cheap, and insertion losses may be made low by orienting the crystal at Brewster's angle. However the hold-off ratio is low, and hence the technique is usually employed in continuously-pumped Q-switched lasers for which the over-

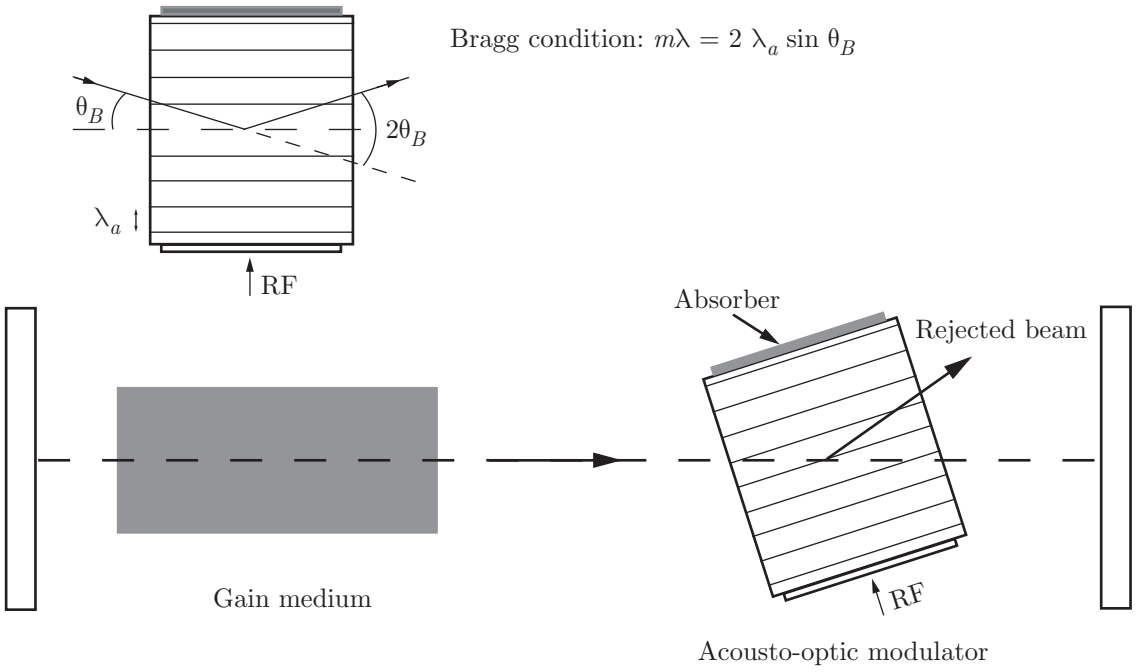


Figure 3.5: Acousto-optic Q-switching. Note that the Bragg angle θ_B has been greatly exaggerated. A value closer to 1° would be more typical. The insertion loss of the acousto-optic crystal can be eliminated (for one polarization) by rotating the crystal out of the plane of the paper through Brewster's angle.

pumping ratio is relatively modest. The switching times achieved are also relatively slow, the switching time corresponding to the time taken for the acoustic wave to propagate out of the region of the cavity mode. Taking the speed of sound in the crystal to be 5 $\mathrm{km\,s^{-1}}$, we can estimate the switching time to be 200 ns for a 1 mm diameter beam. The switching time can be decreased by focusing the cavity mode through the acousto-optic crystal.

Saturable absorbers

A very simple, and quite widely used, method for achieving Q-switching is to insert a saturable absorber into the laser cavity. During the pump pulse the absorber will cause the laser mode to experience a high loss and lasing is prevented. However once the population inversion reaches a high value the absorber will be subjected to an ever increasing flux of spontaneous and weak stimulated emission. As population is transferred to the upper level of the absorbing species the absorption starts to saturate, increasing the feedback from the cavity mirror and so increasing the rate of stimulated emission. This positive feedback leads to rapid saturation (or 'bleaching') of the absorber until, at full saturation, it represents essentially zero loss. The cavity is now in a high-Q state, and a giant pulse develops. It is clear that the absorber must have a strong absorption transition at the laser wavelength and also that the lifetime of the upper level must be reasonably long so that the bleached state is maintained

for the duration of the Q-switched pulse. On the other hand, the upper level lifetime should be short compared to the desired time between Q-switched pulses. Many types of saturable absorbers have been used, such as solutions of dyes, and doped crystals such as Cr^{4+}:YAG.

Saturable absorbers are a simple (there is no external drive circuit) and cheap way of achieving Q-switching. The relatively slow turn on corresponds to many cavity round-trips, which allows good mode discrimination. As a consequence it is quite straightforward to achieve Q-switched lasing on a single transverse mode of the cavity. The disadvantages are that the shot-to-shot fluctuation in the output energy and in the timing of the output pulse can be relatively high, and that the absorber can degrade over time.

3.2.2 Analysis of Q-switching

An analysis of Q-switching is relatively straightforward, providing we make a few simplifying assumptions: we neglect pumping during the development of the Q-switched pulse, and ignore spontaneous transitions. With these assumptions, the rate equations (3.2) and (3.3) become:

$$\frac{dN_2}{dt} = -N^*(t)\sigma_{21}\frac{I}{\hbar\omega_{\mathrm{L}}} \tag{3.23}$$

$$\frac{dN_1}{dt} = +N^*(t)\sigma_{21}\frac{I}{\hbar\omega_{\mathrm{L}}}. \tag{3.24}$$

The equations above correspond to the case of **severe bottlenecking** in which the rate of spontaneous decay of the lower laser level is negligible. As such, stimulated emission causes the population to build up in the lower laser level. The rate equation for the evolution of the population inversion during the Q-switched pulse can be found simply by subtraction:

$$\frac{dN^*}{dt} = \frac{dN_2}{dt} - \frac{g_2}{g_1}\frac{dN_1}{dt} \tag{3.25}$$

$$= -\beta N^*(t)\sigma_{21}\frac{I(t)}{\hbar\omega_{\mathrm{L}}}, \tag{3.26}$$

where $\beta = 1 + g_2/g_1$.

A second case is also of interest, however: that of an ideal four-level laser. In these systems the lower level decays so quickly that throughout the Q-switched pulse $N_1(t) \approx 0$. In this case $N^*(t) \approx N_2(t)$ and hence the rate equation for the population inversion is identical to that for $N_2(t)$.

It is convenient, therefore, to combine these two cases into one set of equations for the population inversion and the photon density:

$$\frac{dN^*}{dt} = -\frac{\beta}{f_{\mathrm{c}}}\frac{N^*}{N^*_{\mathrm{th}}}\frac{n}{\tau_{\mathrm{c}}} \tag{3.27}$$

$$\frac{dn}{dt} = \left(\frac{N^*}{N^*_{\mathrm{th}}} - 1\right)\frac{n}{\tau_{\mathrm{c}}}, \tag{3.28}$$

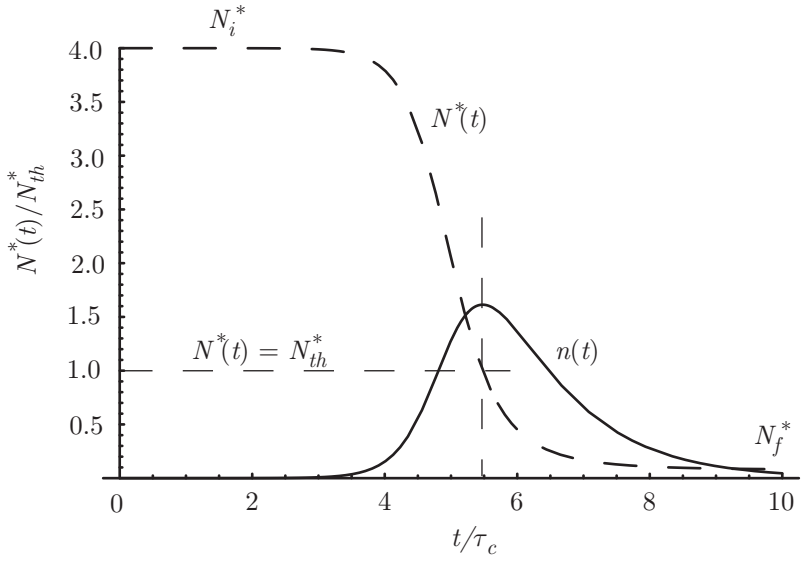


Figure 3.6: General temporal evolution of the population inversion (dashed) and photon (solid) densities during a Q-switched laser pulse for the case $N_i^* = 4N_{\rm th}^*$. The Q-switch is opened at $t = 0$. Notice that it takes several cavity lifetimes for the Q-switch pulse to develop.

where again we have used $I(t) = n(t)c\hbar\omega_{\rm L}$ and eqn (3.6), and where

$$\beta = \begin{cases} 1 & \text{ideal 4-level} \\ 1 + \frac{g_2}{g_1} & \text{severe bottlenecking} \end{cases} \tag{3.29}$$

The general form of the solution to these equations is shown in Fig. 3.6. Notice that the population inversion density is exactly equal to the threshold value, $N_{\rm th}^*$, when the photon density reaches its peak. Since stimulated emission will continue to occur whilst the photon density falls, the population inversion can be burnt down to a level well below the threshold value $N_{\rm th}^*$.

If we integrate eqn (3.28) over the complete duration of the pulse we find,

$$\int_{n(t=-\infty)}^{n(t=\infty)} dn = \int_{-\infty}^{\infty} \frac{N^*}{N_{\rm th}^*} \frac{n}{\tau_{\rm c}} dt - \int_{-\infty}^{\infty} \frac{n}{\tau_{\rm c}} dt = 0 \tag{3.30}$$

where we have used the fact that the first integral is zero since the initial and final densities of photons in the cavity are both zero.

Now, the rate at which photons are coupled out of the laser is equal to $nV_{\rm c}/\tau_{\rm c}$, where $V_{\rm c}$ is the volume taken up by the oscillating mode in the cavity. Hence the power output from the laser is,

$$P = \hbar\omega_{\rm L} V_{\rm c} \frac{n}{\tau_{\rm c}}. \tag{3.31}$$

Applying this result to eqn (3.30), the energy of the output pulse is given by,:

$$E = \hbar\omega_{\rm L} V_{\rm c} \int_{-\infty}^{\infty} \frac{n}{\tau_{\rm c}} dt \tag{3.32}$$

$$= \hbar\omega_{\rm L} V_{\rm c} \int_{-\infty}^{\infty} \frac{N^*}{N^*_{\rm th}} \frac{n}{\tau_{\rm c}} dt \tag{3.33}$$

$$= -\hbar\omega_{\rm L} V_{\rm c} \frac{f_{\rm c}}{\beta} \int_{N_i^*}^{N_f^*} dN^* \qquad \text{(from eqn (3.27))} \tag{3.34}$$

$$= \frac{\hbar\omega_{\rm L}}{\beta} V_g (N_i^* - N_f^*), \tag{3.35}$$

where N_i^* and N_f^* are the initial and final population inversion densities, and $V_g = f_{\rm c} V_{\rm c}$ is the volume of the oscillating mode in the gain medium. We therefore can write,

$$E = \eta N_i^* V_g \hbar\omega_{\rm L}, \tag{3.36}$$

where the **energy utilization factor** is defined as,

$$\eta = \frac{N_i^* - N_f^*}{\beta N_i^*}. \tag{3.37}$$

The interpretation of this result is straightforward: The initial number of inverted atoms is equal to $N_i^* V_g$; for an ideal four-level laser transition the number of photons emitted in the pulse is simply $(N_i^* - N_f^*)V_g$; whereas in the case of severe bottlenecking this number is reduced by a factor of β since for these systems each stimulated emission event reduces the population inversion by β. Eqns (3.36) and (3.37) then follow from the fact that each photon carries an energy $\hbar\omega_{\rm L}$.

We may gain further information by dividing eqn (3.28) by (3.27) to give,

$$\frac{dn}{dN^*} = \frac{f_{\rm c}}{\beta}\left(\frac{N^*_{\rm th}}{N^*} - 1\right) \tag{3.38}$$

$$\Rightarrow \int_{-\infty}^{t} dn = \frac{f_{\rm c}}{\beta} \int_{-\infty}^{t} \left(\frac{N^*_{\rm th}}{N^*} - 1\right) dN^*, \tag{3.39}$$

which gives a useful relation between the photon and inversion densities at all times during the Q-switched pulse:

$$n(t) = \frac{f_{\rm c}}{\beta}\left\{ N^*_{\rm th} \ln\left[\frac{N^*(t)}{N_i^*}\right] + N_i^* - N^*(t)\right\}. \tag{3.40}$$

After the pulse the photon density is zero, and hence setting $t \to \infty$ in the above yields,

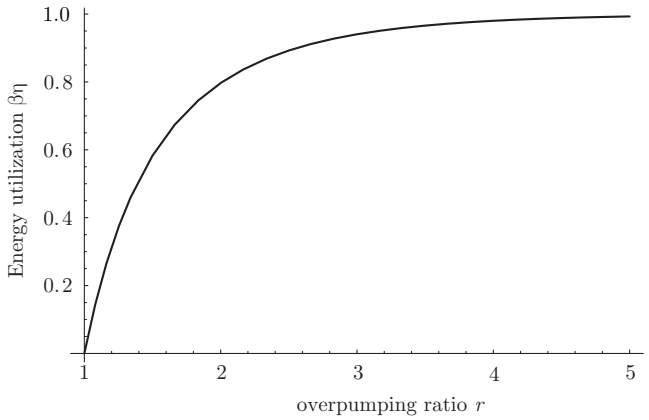


Figure 3.7: The product of the energy utilization factor η and β plotted as a function of the over-pumping ratio r.

$$N_i^* - N_f^* = -N_{\text{th}}^* \ln\left(\frac{N_f^*}{N_i^*}\right) \tag{3.41}$$

$$\Rightarrow \frac{N_i^* - N_f^*}{\beta N_i^*} = -\frac{N_{\text{th}}^*}{\beta N_i^*} \ln\left(\frac{N_f^*}{N_i^*}\right). \tag{3.42}$$

We define the **over-pumping ratio** as the ratio of the initial population inversion density to the threshold value:

$$r = \frac{N_i^*}{N_{\text{th}}^*}. \tag{3.43}$$

Equation (3.42) may then be rearranged to give a relation between the over-pumping ratio and the energy utilization factor:

$$r = -\frac{\ln(1-\beta\eta)}{\beta\eta}. \tag{3.44}$$

This tells us, for example, that to extract 80% of the energy stored in the initial population inversion in an ideal four-level laser requires pumping the initial population inversion to twice the threshold value. Figure 3.7 plots $\beta\eta$ as a function of the over-pumping ratio. We see that once r reaches approximately 5, virtually all the available energy is extracted from the population inversion.

Peak output power and pulse duration

The peak density of photons in the cavity occurs when $dn/dt = 0$, or, $N^* = N^*_{\rm th}$. Equation (3.40) then gives the peak photon density as,

$$n_p = \frac{f_{\rm c}}{\beta} N^*_{\rm th} \left(r - 1 - \ln r\right). \tag{3.45}$$

From (3.31) the peak output power is then,

$$P_p = \frac{n_p V_{\rm c} \hbar \omega_{\rm L}}{\tau_{\rm c}} = \frac{N_{\rm th} \hbar \omega_{\rm L} {V_g}^*}{\beta \tau_{\rm c}} \left(r - 1 - \ln r\right). \tag{3.46}$$

We may then estimate the duration $\tau_{\rm Qs}$ of the Q-switched pulse by dividing the pulse energy E by the peak power, to give:

$$\tau_{\rm Qs} = \beta\eta \frac{r}{r - 1 - \ln r} \tau_{\rm c}. \tag{3.47}$$

We note that as $r \to \infty$, $\beta\eta \to 1$ and hence $\tau_{\rm Qs} \to \tau_{\rm c}$. In this limit the Q-switched pulse rises very rapidly owing to the sudden onset of a high rate of stimulated — the population inversion is extracted almost instantaneously — followed by an exponential decrease with a time constant of $\tau_{\rm c}$ as the photons leak out of the cavity.

Lecture 4

Modelocking

4.1 Introduction

Modelocking is an extremely important technique that is used in many laser systems. Its value lies in the ability to generate pulses of laser radiation which have a short duration — typically picoseconds (1 ps $\equiv 10^{-12}$ s) to femtoseconds (1 fs $\equiv 10^{-15}$ s) — and a peak power many orders of magnitude greater than the mean power that can be extracted from the laser medium. Pulses from modelocked lasers have found applications in a wide variety of scientific and technical applications including: imaging techniques such as two-photon laser-scanning fluorescence imaging, ballistic imaging, and optical coherence tomography; ultrafast chemistry, including coherent control of chemical reactions; pump-probe measurements of solid-state materials; and short-pulse laser machining and processing of materials.

4.2 General ideas

We recall that the longitudinal modes of a laser cavity of length L are spaced in angular frequency by an amount,

$$\Delta\omega_{p,p-1} = 2\pi\frac{c}{2L} = \frac{2\pi}{T_\mathrm{c}}, \tag{4.1}$$

where T_c is the time taken for light to travel once round the cavity. In order to avoid clutter, it will be convenient to let $\Delta\omega \equiv \Delta\omega_{p,p-1}$ throughout this lecture.

The right- and left-going waves of a cavity mode may be written in the simplified form,

$$E(z,t) = a_p \exp\left(\mathrm{i}[k_p z \mp \omega_p t + \phi_p]\right), \tag{4.2}$$

where a_p is the amplitude of the wave, ϕ_p is the phase of the wave, the upper (lower) sign refers to waves propagating towards positive (negative) z, and we have ignored the details of the transverse profile of the waves.

Suppose that the laser oscillates simultaneously on N such modes. For simplicity, we will take there to be an odd number of oscillating modes labelled by $p = -P/2 \ldots P/2$. The amplitude of the laser radiation emerging from the output coupler would have the form,

$$E(z,t) = \sum_{p=p_0-P/2}^{p=p_0+P/2} a_p \exp\left(\mathrm{i}[k_p z - \omega_p t + \phi_p]\right), \tag{4.3}$$

where p_0 is the number of the central longitudinal mode.

Figure 4.1(a) shows a calculation of the output intensity of a laser oscillating simultaneously on 61 modes with equal amplitudes, but with random — though constant — relative phases ϕ_p. The output has two key features,

1. During an interval equal to T_c the output contains sharp fluctuations in intensity which are randomly distributed;

2. The sharpest of these peaks has a duration of order $2\pi/\Delta\omega_\mathrm{osc}$, where $\Delta\omega_\mathrm{osc} = (N-1)\Delta\omega$ is total bandwidth occupied by the oscillating modes;

3. The pattern is periodic with a period equal to the cavity round-trip time T_c.

In contrast, Figure 4.1(b) shows the intensity calculated for the case when the modes all have the same phase ($\phi_p = 0$). Now the output consists of a train of short pulses separated by the cavity round-trip time. This situation is known as **modelocking**.

4.2.1 Simple analysis

To gain further insight, let us calculate the output for the case where all the modes have the same amplitude, E_0. The frequencies of the oscillating modes may be written,

$$\omega_p = \omega_{p_0} + p\Delta\omega, \tag{4.4}$$

where ω_{p_0} is the frequency of the centre mode. Similarly, we consider the case in which the phases of adjacent modes differ by a constant amount, $\Delta\phi$, so that we have,

$$\phi_p = \phi_{p_0} + p\Delta\phi. \tag{4.5}$$

From eqn (4.3) the amplitude of the radiation emitted by the laser will then be:

$$E(\boldsymbol{r},t) = E_0 \sum_{p=-P/2}^{p=+P/2} \exp\left[\mathrm{i}(k_p z - \omega_{p_0} t - p\Delta\omega t + \phi_{p_0} + p\Delta\phi)\right]. \tag{4.6}$$

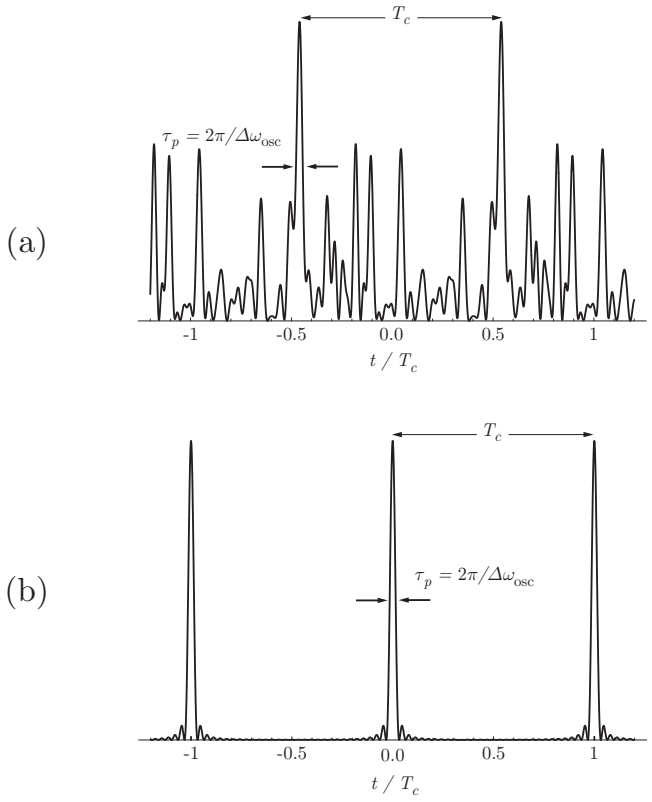


Figure 4.1: Calculated intensity as a function of time emerging from a laser with $N = 61$ modes oscillating simultaneously with equal amplitudes and (a) random (but fixed) phases; (b) identical phases. Notice that for both plots the pattern repeats every cavity round-trip time $T_{\rm c}$, and that the sharpest features have the same width. It should also be emphasized that the vertical scales of the two plots are quite different: the peak intensity of the modelocked pulses are approximately N times greater than the mean intensity in (a).

For convenience we consider the amplitude at $z = 0$ and also take the special case[1] in which $\Delta\phi = 0$. Then,

$$E(0,t) = E_0 \exp\left[-\mathrm{i}(\omega_{p_0} t - \phi_{p_0})\right] \sum_{p=-P/2}^{p=+P/2} \exp\left(-\mathrm{i}p[\Delta\omega t]\right). \tag{4.7}$$

The sum appearing in the above is a geometric progression, which may be summed in the usual way to give,

$$E(0,t) = E_0 \exp\left[-\mathrm{i}(\omega_{p_0} t - \phi_{p_0})\right] \frac{\sin\left(\frac{N}{2}\Delta\omega t\right)}{\sin\left(\frac{1}{2}\Delta\omega t\right)}. \tag{4.8}$$

Hence the output takes the form of a (travelling) wave of mean frequency ω_{p_0} multiplied by an envelope function $\frac{\sin\left(\frac{N}{2}\Delta\omega t\right)}{\sin\left(\frac{1}{2}\Delta\omega t\right)}$.

Taking the intensity to be given by $|E(0,t)|^2$ (i.e. ignoring constants of proportionality):

$$I(t) = E_0^2 \frac{\sin^2\left(\frac{N}{2}\Delta\omega t\right)}{\sin^2\left(\frac{1}{2}\Delta\omega t\right)}. \tag{4.9}$$

This is just the function plotted in Fig. 4.1(b).

The temporal behaviour calculated above corresponds to a train of intense modelocked pulses. Notice that the peak intensity of each pulse is equal to $N^2 E_0^2$, and hence the peak intensity is a factor of N greater than the mean intensity of N modes oscillating with random phases[2].

Pulses occur at times given by $\Delta\omega t_m = 2\pi m$, and hence the separation of pulses is simply $2\pi/\Delta\omega = T_\mathrm{c}$. The width, τ_p, of the pulses may be taken to be the interval from the peak of a modelocked pulse to the first zero. It is straightforward to show that this is equal to:

$$\tau_p = \frac{2\pi}{N\Delta\omega} = \frac{T_\mathrm{c}}{N} \approx \frac{2\pi}{\Delta\omega_\mathrm{osc}}, \tag{4.10}$$

where $\Delta\omega_\mathrm{osc} = (N-1)\Delta\omega$ is the (angular) frequency bandwidth covered by the oscillating modes.

It should be clear from the above that if we were able to lock the phase of N longitudinal modes of a laser we could generate a stream of pulses of high peak intensity and short duration. In practice the number of modes that can be locked ranges from several hundred to several thousand, leading to a huge increase in peak power. The duration of the pulses is determined by the frequency width of the laser transition; transitions with broad linewidths can (in principle) generate shorter pulses. In practice pulses with durations from a few picoseconds to a few femtoseconds can be generated; a very active area of current research is the development of techniques related to modelocking to generate pulses in the attosecond range (1 as $\equiv 10^{-18}$ s).

[1] It can be shown quite simply that if $\Delta\phi \neq 0$ the peaks of all the modelocked pulses shift in time by a constant value of $\Delta t = \Delta\phi/\Delta\omega$, but that otherwise the train is unchanged.

[2] Note that the mean intensity is the same whether or not the N modes are modelocked or not. Modelocking is an interference phenomenon, and interference neither destroys or creates energy!

4.2.2 Further general comments

Modelocking techniques can be divided into two classes: **active** and **passive**, as described below.

In the general discussion above we did not consider in any detail the properties of the laser gain medium. As we might expect, the modelocking behaviour of a laser depends on whether the laser transition is homogeneously or inhomogeneously broadened. We have already seen that an inhomogeneously broadened laser transition can oscillate simultaneously on all longitudinal modes for which the round-trip gain exceeds the round-trip loss. The discussion above applies quite well to such lasers, and in such cases the modelocking technique only needs to lock the phases of modes that are already oscillating simultaneously. As a first approximation, the duration of modelocked pulses output from a laser operating on an inhomogeneously broadened transition will be $\tau_p \approx 2\pi/\Delta\omega_{\mathrm{D}}$, where $\Delta\omega_{\mathrm{D}}$ is the inhomogeneous linewidth of the transition.

Homogeneously broadened laser transitions, however, usually only oscillate on a single longitudinal mode (unless spatial holeburning occurs). The modelocking must therefore play a stronger role, and encourage laser oscillation on many modes. As such, the dependence of the pulse duration on the 'strength' of the modelocking will be different from the case of inhomogeneous broadening, and the duration of the modelocked pulses will generally be longer than $2\pi/\Delta\omega_{\mathrm{H}}$, where $\Delta\omega_{\mathrm{H}}$ is the homogeneous linewidth.

4.3 Active modelocking techniques

Active modelocking techniques fall into two classes: **amplitude modulation (AM)** and **frequency modulation (FM)**. We note that we usually require a single modelocked pulse to circulate within the laser cavity. For a linear laser cavity this is achieved by locating the active component near one of the end mirrors, and driving the component at a frequency equal to the longitudinal mode spacing $\Delta\omega$. However, it is possible to generate N such modelocked pulses within the laser cavity by locating the modulator suitably and driving it at $N\Delta\omega$. **Harmonic modelocking** of this type is occasionally used to increase the pulse repetition frequency. In principle it can also be used to generate shorter pulses in homogeneously broadened lasers (see eqn (4.19)).

4.3.1 AM modelocking

Once modelocking is achieved, the intra-cavity radiation takes the form of a short pulse which circulates round the laser cavity. An obvious way, therefore, to achieve modelocking is to introduce some sort of 'shutter' into the cavity which lets through short pulses, but which blocks — or increases the loss of — longer pulses. Radiation which happens to reach the shutter when it is open will pass through with low loss, will experience net round-trip gain, and hence will grow. In contrast, the radiation arriving at the shutter outside this interval will experience a round-trip loss which is greater than the round-trip gain, and so will be absorbed.

After many round-trips a pulse will form which propagates round the cavity in synchronization with the shutter, as illustrated schematically in Figure 4.2. The duration of the modelocked pulse so formed is determined by an equilibrium between two competing effects. The shutter tends to decrease the duration of the pulse (and hence increase the bandwidth) since the losses of the leading and trailing edges of the pulse are greater than those experienced by the centre of the pulse. However, the gain

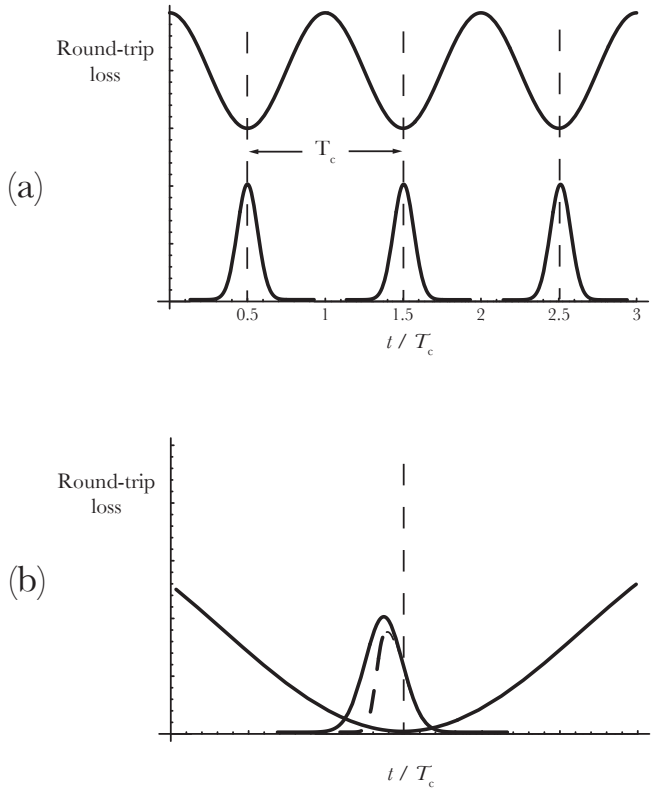


Figure 4.2: Schematic diagram of AM modelocking. In (a) the pulses arrive perfectly synchronized with the shutter and hence experiences minimum loss. The pulse shown in (b) arrives slightly too early, and hence the leading edge is reduced in intensity. The effect is to move the peak of the pulse to later times, closer to synchronization with the shutter.

medium can only amplify a finite frequency bandwidth which limits the minimum possible duration of the modelocked pulse.

The above description is said to be in the time domain. It is also possible to consider the problem from the perspective of the frequency domain. To do this we suppose that the amplitude transmission of the shutter varies harmonically as,

$$T(t) = T_0 + \frac{\delta}{2}\cos\Delta\omega t. \tag{4.11}$$

Suppose that a wave of angular frequency ω_p is incident on the shutter, such that we may write its amplitude as the real part of,

$$E(t) = E_0 \exp(-\mathrm{i}\omega_p t). \tag{4.12}$$

The amplitude of the transmitted wave will be:

$$E(t) = E_0 \exp(-\mathrm{i}\omega_p t)\left[T_0 + \frac{\delta}{4}\exp(\mathrm{i}\Delta\omega t) + \frac{\delta}{4}\exp(-\mathrm{i}\Delta\omega t\right] \quad (4.13)$$

$$= E_0 T_0 \exp(-\mathrm{i}\omega_p t) + \frac{E_0\delta}{4}\exp(-\mathrm{i}[\omega_p - \Delta\omega]t) + \frac{E_0\delta}{4}\exp(-\mathrm{i}[\omega_p + \Delta\omega]t. \quad (4.14)$$

In other words, the amplitude modulation adds **frequency sidebands** at $\omega_p \pm \Delta\omega$. Since the modulation is driven at the frequency difference between longitudinal modes, the frequencies of the sidebands of a longitudinal mode p lie exactly at the frequencies of the adjacent longitudinal modes $p \pm 1$.

This frequency domain picture provides further insight into how it is possible to achieve modelocking in lasers operating on *homogeneously* broadened transitions. For each longitudinal mode, the amplitude modulation couples energy into the adjacent modes; and it does so with a definite phase. As such, we can consider the sidebands as 'injection seeding' the adjacent longitudinal modes. The system is complex, since each of many modes is coupled to the adjacent modes, but it is clear that the amplitude modulation causes energy to flow from the modes closest to the line centre out to the wings. The amplitude modulation therefore tends to force the laser out of single-mode oscillation, and broaden the frequency spectrum. A steady state is reached when this broadening is balanced by the frequency narrowing arising from the finite frequency response of the gain medium.

We note that for a homogeneously broadened laser transition the round-trip gain of the modes near the line centre is actually greater than the round-trip losses; in the wings of the modelocked pulse the round-trip gain is less than the round-trip loss, as shown schematically in Fig. 4.3. In both cases the difference in energy is made up by energy transfer to or from adjacent modes so that a steady-state is maintained.

Finally we note that for both homogeneously and inhomogeneously laser transitions the system adopts the lowest loss configuration, which corresponds to simultaneous oscillation on a large number of longitudinal modes, phase-locked so as to form a short pulse circulating round the cavity.

Experimental implementation

Amplitude modulation may be achieved by using a Pockels cell in much the same configuration as is used to generate Q-switched pulses. The Pockels cell is driven with a sinusoidal voltage of frequency $\Delta\omega$. Since the amplitude modulation does not need to be 100%, the driving voltage need only be some fraction of the quarter-wave voltage.

It is also possible to use an acousto-optic modulator to implement AM modelocking, using a similar configuration to that used for Q-switching, as shown in Fig. 4.4. However, for modelocking the faces of the acousto-optic block are cut parallel, and the length of the block is cut to be an integer number of half (acoustic) wavelengths long. A standing wave can then be driven in the acousto-optic block, rather than the travelling acoustic wave used to Q-switch. The amplitude of the standing wave passes through zero twice per period, corresponding to windows of minimum loss. Hence, if the block is driven with an angular frequency ω_a, AM modelocking will be achieved if $2\omega_\mathrm{a} = \Delta\omega$.

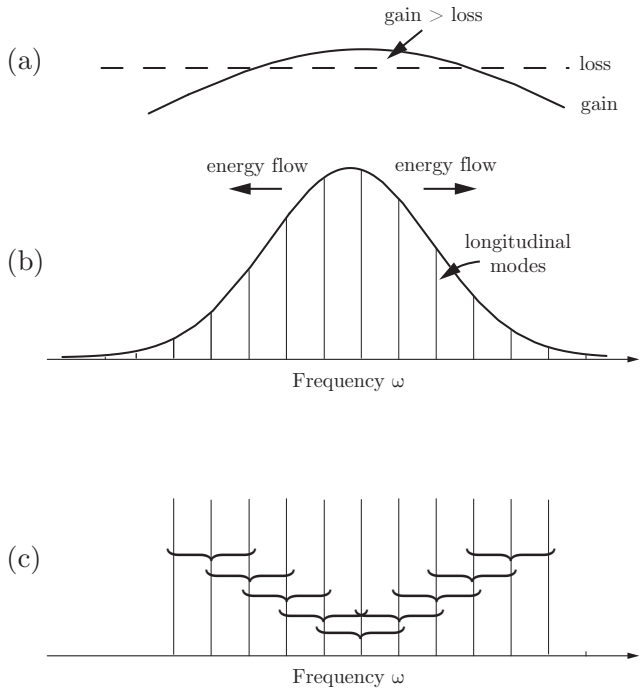


Figure 4.3: Energy flow and coupling of modes in an AM modelocked laser operating on a homogeneously broadened transition. The frequency dependence of the optical gain and loss *in the absence of the amplitude modulation* is shown in (a). As illustrated in (b), the amplitude modulation causes energy to flow from those modes for which the round-trip gain exceeds the round-trip loss to modes in the wings of the gain profile for which the gain is lower than the loss. This energy transfer allows steady-state laser oscillation to be maintained simultaneously on many modes. The amplitude modulation causes each mode to develop sidebands at frequencies equal to those of the adjacent longitudinal modes, and hence couples adjacent modes as illustrated schematically in (c).

FM modelocking

Frequency modulation modelocking is achieved by modulating the refractive index of an intra-cavity component at the mode spacing frequency $\Delta\omega$, as illustrated schematically in Fig. 4.5. As we will see, this affects the frequency spectrum of an incident wave and hence the active device is known as a **frequency modulator**.

We consider first the effect on a wave of angular frequency ω of a modulator of length $L_{\rm mod}$ and time-dependent refractive index $n(t)$. The incident wave will emerge with an amplitude equal to the real part of,

$$E(L_{\rm mod}, t) = E_0 \exp\left\{ \mathrm{i}\omega \left[n(t)\frac{L_{\rm mod}}{c} - t \right] \right\}. \tag{4.15}$$

Let us now suppose that the refractive index is modulated harmonically, such that,

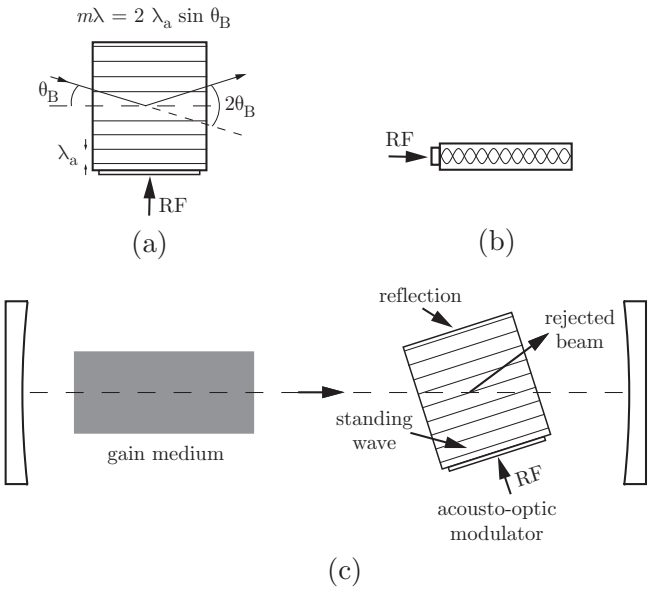


Figure 4.4: Amplitude modulation modelocking with an acousto-optic modulator. In (a) the Bragg condition for reflection of a beam of radiation by incident at an angle θ_B on an acoustic wave of wavelength λ_a. In contrast to the use of acousto-optic crystals in Q-switching, discussed in Lecture 3, in order to achieve modelocking the cavity losses must be modulated at $\Delta\omega$. This is achieved by cutting the acousto-optic crystal so that its two ends are parallel — allowing reflection of the acoustic wave — and separated by an integer of half wavelengths when the modulator is driven at $\Delta\omega/2$. This configuration allows a standing wave to be set-up with an amplitude that passes through zero with a frequency $\Delta\omega$. A typical layout of the optical cavity is shown in (c). Note that the Bragg angle θ_B has been greatly exaggerated. Note also that the insertion loss of the acousto-optic crystal can be eliminated (for one polarization) by rotating the crystal out of the plane of the paper through Brewster's angle.

$$n(t) = n_0 + \frac{\delta}{2}\sin\Omega t.$$

The transmitted wave will be

$$E(L_{\rm mod}, t) = E_0 \exp(-{\rm i}\omega t)\exp({\rm i}\phi_0)\exp({\rm i}\beta\sin\Omega t), \tag{4.16}$$

where $\phi_0 = \omega n_0 L_{\rm mod}/c$ is the (constant) phase shift experienced by light of angular frequency ω, and $\beta = \omega\delta L_{\rm mod}/2c$ is the amplitude of the modulation in phase introduced by the modulator.

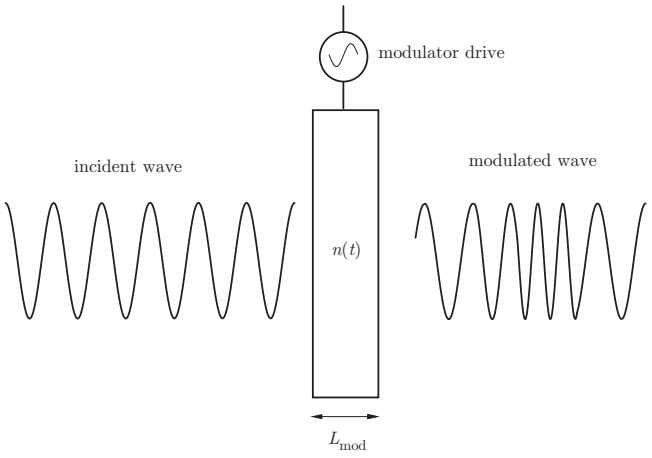


Figure 4.5: Schematic diagram of FM modelocking.

Now we may use the identity,[3]

$$\exp(\mathrm{i}\beta \sin \Omega t) \equiv \sum_{m=-\infty}^{\infty} J_m(\beta) \exp(\mathrm{i}m\Omega t) \tag{4.17}$$

where $J_m(x)$ is the Bessel function of the first kind of order m.

Thus we may write the amplitude of the transmitted wave as,

$$E(L_{\mathrm{mod}}, t) = E_0 \exp(-\mathrm{i}\omega t) \exp(\mathrm{i}\phi_0) \sum_{m=-\infty}^{\infty} J_m(\beta) \exp(\mathrm{i}m\Omega t), \tag{4.18}$$

and we see that the modulator introduces a series of sidebands at frequencies of $\omega \pm m\Omega$, where m is an integer.

Applying these ideas to an intra-cavity modulator driven at the mode spacing frequency $\Delta\omega$, we see that for each mode of frequency ω_p, the modulator generates frequency sidebands at frequencies of $\omega_p \pm m\Delta\omega$ thereby coupling the longitudinal modes in a similar way to amplitude modulation. The width of the pulse generated in this way is determined by the balance between the bandwidth broadening caused by the modulator and the frequency narrowing produced by amplification in the gain medium.

As shown schematically in Fig 4.6, the modelocked pulses are found to be located at the maxima and minima of $n(t)$, when the modulator produces no shift in frequency. This may be understood as follows. The effect of the phase modulation introduced by the frequency modulator is equivalent to modulating the cavity length by driving one of the cavity mirrors sinusoidally. In general a pulse

[3]Consult any text on mathematical methods, for example Riley, Hobson & Bence "Mathematical methods for physics and engineering: a comprehensive guide".

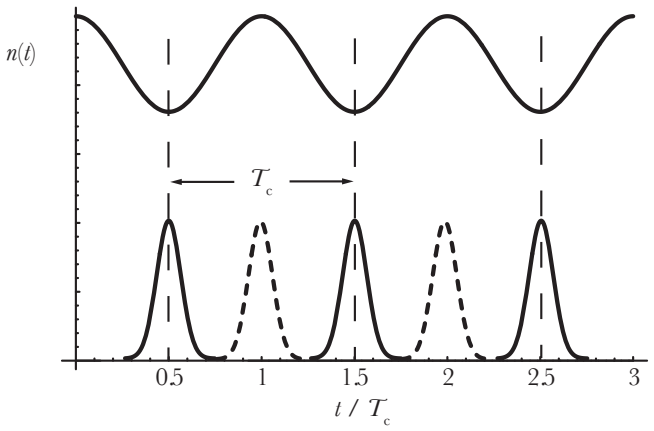


Figure 4.6: Location of the modelocked pulses with FM modelocking.

reflected from the moving cavity mirror will be Doppler-shifted in frequency and, since the mirror is driven with a frequency of $\Delta\omega$, when it returns one round-trip later it will be further shifted in frequency by the same amount. Eventually the pulse will be shifted outside the bandwidth of the amplifying transition and will no longer be amplified. However, this progressive frequency shift is minimized for radiation which strikes the mirror at one of the extrema of its motion or, equivalently, if it passes through the modulator near the maxima or minima in $n(t)$.

Notice that there are *two* stable positions for the modelocked pulse. An undesirable consequence of this is that the inevitable disturbances present in a real laser can cause the modelocking to flip between these two states, or even that two modelocked pulses circulate within the cavity.

Experimental implementation

Frequency modulation may be achieved by orienting a Pockels cell with its optical axis parallel with the polarization of the intra-cavity laser radiation. Applying a voltage to the Pockels cell will then change the refractive index experienced by the radiation, but will not alter its polarization state.

4.3.2 Synchronous pumping

Modulating the gain at the longitudinal mode spacing frequency $\Delta\omega$ can result in modelocking in some laser systems by processes analagous to the AM modelocking discussed above. It should be clear that in order for the gain of the laser transition to respond to the modulation, the population inversion must decay in a time which is short compared to the cavity round-trip time.

For example, dye lasers can be modelocked by optically pumping the dye with pulses from a laser which is itself modelocked[4]. Synchronous pumping can also be achieved in semiconductor lasers by modulating the driving current.

[4]Obviously care must be taken to ensure that the repetition frequency of the modelocked pump pulses corresponds to the longitudinal mode spacing of the dye laser: in other words the cavity round-trip times must be the same for the two lasers

4.3.3 Pulse duration of actively modelocked, homogeneously broadened lasers

For completeness, we note that it can be shown[5] that the duration of the pulses generated by AM modelocking a CW laser operating on a homogeneously broadened laser transition is given by,

$$\tau_p \approx 0.45 \left(\frac{2\alpha_I L_{\rm g}}{\delta} \right)^{1/4} \frac{2\pi}{\sqrt{\Delta\omega_{\rm mod}\Delta\omega_{\rm H}}}, \tag{4.19}$$

where α_I is the (saturated) gain coefficient, $L_{\rm g}$ the length of the gain medium, $\Delta\omega_{\rm mod}$ is the driving frequency of the modulator (usually set equal to $\Delta\omega$, but it may be a harmonic of this), and $\Delta\omega_{\rm H}$ is the homogeneous, linewidth. Notice that the pulse duration decreases with increasing modulation, δ, but only rather slowly. In fact, for most systems the first bracketed term will be close to unity, and hence,

$$\tau_p \approx 0.45 \frac{2\pi}{\sqrt{\Delta\omega_{\rm mod}\Delta\omega_{\rm H}}}. \tag{4.20}$$

A similar result can be obtained for FM modelocking. Notice that since $\Delta\omega_{\rm H} \gg \Delta\omega$ the pulses output by an actively modelocked laser operating on a homogeneously broadened laser will be significantly longer than an inhomogeneously broadened transition of the same linewidth.

4.4 Passive modelocking techniques

Modelocking can also be achieved by inserting into the laser cavity an element with transmission properties which depend on the intensity of the incident radiation. Passive modelocking avoids the need to drive modulators at high frequencies, and is therefore relatively simple to implement practically. Passive modelocking techniques are perhaps of more practical importance than active techniques, and are able to generate the shortest possible laser pulses.

From a theoretical point of view, however, it is more difficult to calculate the pulse duration of the modelocked pulse than in active modelocking, and treatments tend to be specific to the particular case under consideration. In any case, the lower limit of the duration will always be given by approximately by $2\pi/\Delta\omega_{\rm H,D}$, where $\Delta\omega_{\rm H,D}$ is the linewidth of the laser transition.

4.4.1 Saturable absorbers

Just as there is for a transition exhibiting optical gain, there are several key parameters for an absorption transition. In particular, the saturation intensity, $I_{\rm s}$, and the **saturation fluence** (units of $\mathrm{W\,m^{-2}}$), $\Gamma_{\rm s}$:

$$I_{\rm s} = \frac{\hbar\omega}{\sigma_{21}\tau_R} \tag{4.21}$$

$$\Gamma_{\rm s} = \frac{\hbar\omega}{\sigma_{21}}, \tag{4.22}$$

[5] See, for example, Siegman.

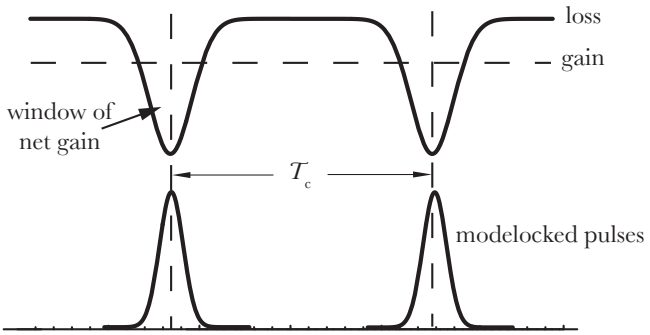


Figure 4.7: Passive modelocking with a fast saturable absorber.

where τ_R is the recovery time of the transition. We note that the time taken for the absorption to return to its high, unsaturated value once an intense, saturating beam of incident radiation is switched off is of the order of the recovery time.

The saturation intensity is relevant to steady-state conditions, or situations in which the duration, τ, of an incident pulse is much longer than the recovery time. In this case, I_s divides two regions: low intensities for which the incident radiation experiences high, unsaturated absorption; high intensities, in which the absorption is strongly reduced by saturation of the transition.

The situation is somewhat different in pulsed conditions. If the duration of an incident laser pulses is short compared to the recovery time the absorption experienced at any time t during the pulse depends only on the intensity integrated over time to that point, i.e. on $\int_{-\infty}^{t} I(t')dt'$. The reason for this is simply that atoms transferred to the upper level of the absorption transition will not decay during the incident laser pulse, and hence the density of atoms in the upper level will depend only on the number of incident photons per unit area, not on the details of the pulse shape. As such, the absorption will start to saturate after a certain number of incident photons per unit area, or, equivalently after a certain incident fluence (energy per unit area). In this case, then, the behaviour is determined by the fluence of the incident pulse relative to the saturation fluence.

Fast saturable absorbers

In the context of modelocking, a **fast saturable absorber** is a medium for which the recovery time is very much shorter than the cavity round-trip time, and in fact is short compared to the typical spacing of the 'noise' spikes observed in the non-modelocked, multimode laser (See Fig. 4.3(b)). Typically, then, the recovery time of a fast saturable absorber will be a few tens of picoseconds, or less.

If a fast saturable absorber is placed within a laser cavity, near one of the laser mirrors, the round-trip loss will be lower for intense noise spikes than for low-intensity spikes. Consequently the most intense spike will tend to grow at the expense of the others, and will start to saturate the available gain. The result is a single, intense pulse propagating round the laser cavity: modelocking. We can think of the saturable absorber as acting as a shutter which is activated by the pulse itself, as illustrated in Figure 4.7.

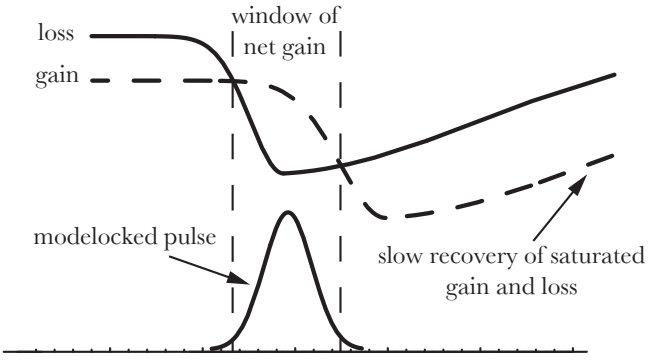


Figure 4.8: Schematic diagram of modelocking with a slow absorber. The interplay of saturation of the gain and loss ensures that net gain only exists in a short window. The saturated gain and loss must recover within a cavity round-trip time T_c.

Slow saturable absorbers

Under some circumstances it is also possible to achieve modelocking using a **slow saturable absorber**, for which the recovery time is comparable to — although it must still be less than — the cavity round-trip time.

Figure 4.8 illustrates schematically how modelocking is achieved with a slow absorber. For simplicity we imagine the saturable absorber and the gain medium lie in the same plane. The leading edge of the modelocked pulse will experience a net round-trip loss, owing to the fact that the (unsaturated) absorption is higher than the available gain. As such the front of the pulse will be eroded. As the pulse passes through the absorber, the absorption will gradually decrease as population is transferred to the upper level of the absorption transition. At some point the round-trip loss introduced by the saturable absorber will fall below the round-trip gain, and the pulse will experience amplification. In order for the generated pulse to be short, however, it is essential that shortly after this moment the growing pulse saturates the round-trip gain below the (saturated) round-trip loss so that the trailing edge of the pulse is also attenuated. Further, in order for the system to settle down to a steady-state in which a single, modelocked pulse circulates within the cavity both the gain and the absorption must recover to their initial values in less than a cavity round-trip time.

It should be appreciated, therefore, that modelocking with a slow saturable absorber requires a delicate balance between the energy of the modelocked pulse, and the saturation fluences and recovery times of both the absorber and the gain medium. The requirement that the gain can be returned to its initial value by the pumping mechanism in less than one round-trip time means that the technique is limited to lasers with a relatively short upper level lifetime, such as dye lasers or semiconductor lasers.

4.4.2 Kerr lens modelocking

Kerr lens modelocking (KLM) is one of the most important types of modelocking used today, and is able to generate the shortest optical pulses obtainable directly from a laser oscillator.

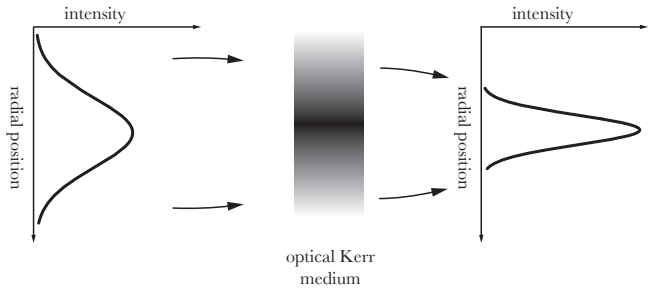


Figure 4.9: Schematic diagram of self-focusing arising from the optical Kerr effect.

As we have already seen, all materials exhibit the Kerr effect, in which the refractive index depends on the square of any applied electric field. If an intense optical pulse is incident on a medium, the **optical Kerr effect** will cause the refractive index to be modified by the electric field of the pulse itself. Since the square of the electric field is proportional to the intensity, I, of the pulse, we may write the refractive index of a material as,

$$n(I) = n_0 + n_2 I. \tag{4.23}$$

It is found that n_2 is positive for nearly all optical materials, with values typically of order 10^{-16} cm^2 W^{-1}.

The optical Kerr effect is responsible for a variety of phenomena, such as self-focusing and self-phase-modulation. The former lies at the heart of Kerr lens modelocking. Figure 4.9 shows the propagation of a beam with a Gaussian transverse profile through an optical Kerr medium. The intensity of the beam will be greater on the propagation axis than in the transverse wings, and hence the refractive index of the medium in the axial region will be greater than in the wings. The optical Kerr medium therefore looks like a thin lens which focuses the beam slightly as it propagates through it.

Figure 4.10 shows schematically how the optical Kerr effect can cause modelocking. The optical Kerr medium is usually the (solid) laser gain medium itself. The laser cavity is designed so that in the absence of any self-focusing the transverse dimensions of the lowest-order cavity mode are comparable to, or larger than, the size of an aperture placed within the cavity. As such, low-intensity radiation experiences a relatively high round-trip loss. Higher intensity radiation, however, will be self-focused by the Kerr effect and so experience lower round-trip losses. Modelocking then arises by essentially the same processes as with a fast saturable absorber: the optical Kerr medium acts as a fast, self-activated shutter.

In order for KLM to be effective, the laser cavity must be operated close to being 'unstable' in order that the weak Kerr lens causes a large change in the transverse extent of the mode.

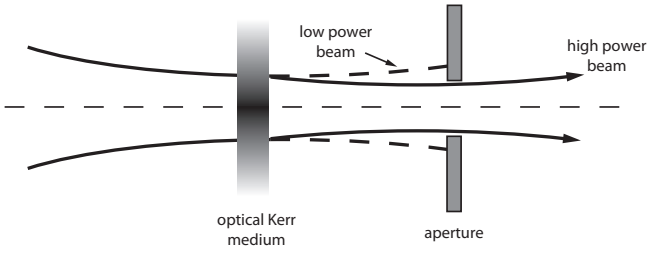


Figure 4.10: Schematic diagram of Kerr lens modelocking.

4.5 Examples of modelocked lasers

4.5.1 CW modelocking

In CW lasers, active modelocking is usually implemented with acousto-optic modulators owing to their lower insertion loss compared to a Pockels cell. The modulator should be placed near one of the end mirrors of the cavity if only a single pulse circulating within the cavity is to be generated.

Some examples are given in Table 4.1.

4.5.2 Pulsed modelocking

The behaviour of pulsed modelocked lasers is determined by the dynamics of the population inversion during the pump pulse. If the lifetime of the upper laser level is short, the duration of the burst of modelocked pulses is determined by the duration of the pumping. However if the lifetime of the upper laser level is long, pulsed modelocking tends to be accompanied by Q-switching. In this case the duration of the burst of modelocked pulses will be approximately equal to the duration of the Q-switched pulse τ_{QS}.

The higher gain found in pulsed lasers means that it is possible to use a Pockels cell to achieve active modelocking, in either AM or FM mode. Passive modelocking of pulsed lasers is achieved with fast saturable absorbers.

Some examples are given in Table 4.1.

Table 4.1: Parameters of some modelocked lasers

Laser	pulsed / c.w.	Method	Pulse duration
He-Ne	c.w.	acousto-optic	1 ns
CO_2	c.w.	acousto-optic	10-20 ns
Ar^+	c.w.	acousto-optic	150 ps
Nd:YAG	c.w.	acousto-optic	100 ps
He-Ne	c.w.	slow sat. abs.	300 ps
GaAs	c.w.	slow sat. abs.	5 ps
Dye	c.w.	slow sat. abs.	25 fs
Dye	c.w.	synch. pumping	1 - 30 ps
Ti:sapphire	c.w.	Kerr lens	10 - 30 fs
Nd:YAG	pulsed	Pockels cell	40 ps
CO_2	pulsed	fast sat. abs.	1 ns
Dye	pulsed	fast sat. abs.	1 ps
Nd:YAG	pulsed	fast sat. abs.	40 ps
Nd:Glass	pulsed	fast sat. abs.	5 ps
Ruby	pulsed	fast sat. abs.	10 ps

Lecture 5

Ultrafast Lasers

5.1 Propagation of short laser pulses in dispersive media

It may be shown that the root-mean-square (RMS) angular frequency bandwidth, $\Delta\omega_{\mathrm{RMS}}$, of a pulse is related to the RMS duration of the pulse, Δt_{RMS}, by,

$$\Delta\omega_{\mathrm{RMS}}\Delta t_{\mathrm{RMS}} \geq \pi. \tag{5.1}$$

In general the value of the **time-bandwidth product** $\Delta\omega_{\mathrm{RMS}}\Delta t_{\mathrm{RMS}}$ depends on:

- the exact shape of the pulse envelope (i.e. Gaussian, square, sech^2 etc.);
- the amount of frequency chirp or phase structure within the pulse.

In addition, of course, different definitions of $\Delta\omega$ and Δt — such as the full-width at half maximum — will yield different time-bandwidth products, although any reasonable definition of these quantities should give a result similar to eqn (5.1).

Short laser pulses, therefore, are associated with a large bandwidth. For propagation *in vacuo*, this presents no difficulties, and the pulse will propagate over arbitrary distances without distortion. However, optical materials are generally dispersive — that is their refractive index varies with frequency — which can lead to distortion of the pulse as it propagates. It is no use generating a short laser pulse if it is stretched and distorted beyond recognition when it reaches the interaction region in an experiment. An understanding of the effect of dispersion, and how it may be controlled, is crucial for both applications and generation of short laser pulses.

In general we may represent an optical pulse by a superposition of harmonic waves. For example, a pulse propagating through a medium is given by the real part of,

$$E(z,t) = \int_0^\infty a(\omega)\exp\left\{\mathrm{i}[k(\omega)z - \omega t]\right\}\mathrm{d}\omega, \tag{5.2}$$

where we take the axis of propagation to be the z axis, and $a(\omega)$ is the amplitude of each angular frequency ω. The wave vector $k(\omega)$ will in general be a function of frequency of the wave, as determined

by the **dispersion relation** for the medium. For example, in terms of the refractive index $n(\omega)$ the wave vector is given by,

$$k(\omega) = \frac{\omega}{c} n(\omega). \tag{5.3}$$

In general we can expand the wave vector about its value at some representative central frequency ω_0:

$$k(\omega) = k(\omega_0) + \left.\frac{\partial k}{\partial \omega}\right|_{\omega_0} (\omega - \omega_0) + \frac{1}{2} \left.\frac{\partial^2 k}{\partial \omega^2}\right|_{\omega_0} (\omega - \omega_0)^2 + \ldots \tag{5.4}$$

If we retain just the first two terms in this expansion, substitution into eqn (5.2) yields,

$$E(z,t) = \exp\left\{\mathrm{i}[k(\omega_0)z - \omega_0 t]\right\} \int_0^\infty a(\Delta\omega) \exp\left[\mathrm{i}(\Delta\omega)\left(\frac{z}{v\mathrm{g}} - t\right)\right] d(\Delta\omega), \tag{5.5}$$

where $\Delta\omega = \omega - \omega_0$, and the **group velocity** of the pulse is given by,

$$v_\mathrm{g} = \left.\frac{\partial \omega}{\partial k}\right|_{\omega_0}. \tag{5.6}$$

It is easy to see that in this approximation the pulse comprises a **carrier wave** of frequency ω_0, the wavefronts of which propagate at the **phase velocity**

$$v_p = \frac{\omega_0}{k(\omega_0)}, \tag{5.7}$$

multiplied by a **pulse envelope** determined by the integral. Within the approximation we have made here, the pulse envelope has a constant shape and moves at the group velocity. If we had included higher-order terms in the expansion of $k(\omega)$, we would find that the envelope of the pulse would change with propagation distance.

5.1.1 Propagation through an arbitrary system

In determining the behaviour of a short pulse as it propagates between two planes $z = z_1$ and $z = z_2$, it is useful to calculate the total phase accumulated:

$$\phi(\omega) = \int_{z_1}^{z_2} k(\omega) dz \tag{5.8}$$

$$= \phi(\omega_0) + \left.\frac{\partial \phi}{\partial \omega}\right|_{\omega_0} (\omega - \omega_0) + \frac{1}{2} \left.\frac{\partial^2 \phi}{\partial \omega^2}\right|_{\omega_0} (\omega - \omega_0)^2 + \ldots \tag{5.9}$$

For a single medium of length L, this expansion is simply equal to eqn (5.4) multiplied by L.

We consider each of the terms above:

- $\phi(\omega_0)$. This is the total phase accumulated at ω_0. Together with the group delay, it determines where the peaks of the carrier wave are located with respect to the envelope of the pulse. As such it becomes important for pulses which comprising only a few optical cycles; techniques for measuring and controlling $\phi(\omega_0)$ are now an active area of research.
- $\left.\frac{\partial\phi}{\partial\omega}\right|_{\omega_0} \equiv \phi^{(1)}(\omega_0)$. This term is known as the **group delay**. To see this, we note that for propagation through a single medium of length L, $\phi^{(1)}(\omega_0) = L/v_{\rm g}$, and gives therefore the time taken for the envelope to propagate between the two planes.
- $\left.\frac{\partial^2\phi}{\partial\omega^2}\right|_{\omega_0} \equiv \phi^{(2)}(\omega_0)$. This term is known as the **group delay dispersion (GDD)** or the **quadratic phase**. It is the lowest term in the expansion responsible for distortion of the pulse envelope with propagation. **Positive dispersion** is defined as $\phi^{(2)}(\omega_0) > 0$. The GDD is generally expressed in units of fs^2.
- $\left.\frac{\partial^n\phi}{\partial\omega^n}\right|_{\omega_0} \equiv \phi^{(n)}(\omega_0)$. There are of course higher terms in the expansion: the cubic, quartic, quintic... phase. For the production and propagation of sufficiently short pulses, these will also be important.

Group delay dispersion (GDD)

Let us consider first the frequency dependence of the group delay. Denoting the transit time through the system for quasi-monochromatic waves with frequencies near ω as $\tau(\omega)$, we have from our discussion above,

$$\tau(\omega) = \frac{\partial\phi}{\partial\omega}. \tag{5.10}$$

Expanding about the mean frequency yields,

$$\begin{aligned} \tau(\omega) &= \tau(\omega_0) + \left.\frac{\partial\tau}{\partial\omega}\right|_{\omega_0}(\omega-\omega_0) + \ldots & (5.11)\\ &= \phi^{(1)}(\omega_0) + \phi^{(2)}(\omega_0)(\omega-\omega_0) + \ldots. & (5.12)\end{aligned}$$

We see that the term in $\phi^{(2)}(\omega_0)$ causes the transit time to increase linearly with frequency. As such, for positive values of the GDD, the high frequency components of the pulse will move to the rear of the pulse; the low frequency components will lead. This will cause the pulse to develop a **positive frequency chirp**, that is for a stationary observer the frequency of the radiation increases as the pulse passes. Further, the difference in group delays will cause the pulse to increase in duration by an amount approximately equal to

$$\Delta\tau_{\rm GDD} = \phi^{(2)}(\omega_0)\Delta\omega, \tag{5.13}$$

where $\Delta\omega$ is the bandwidth of the pulse. These processes are illustrated schematically in Fig. 5.1

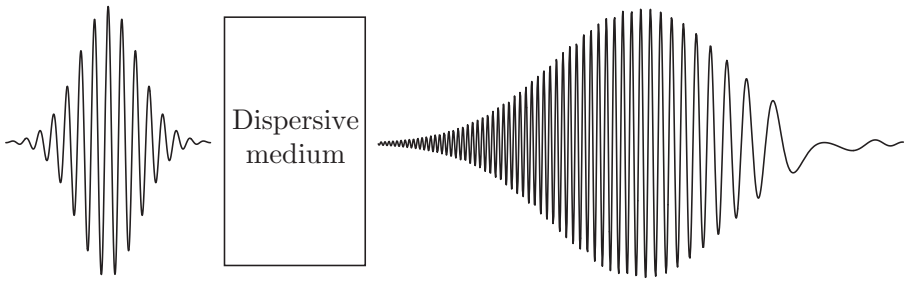


Figure 5.1: Distortion of an initially un-chirped optical pulse by propagation through a dispersive medium with positive GDD. The transmitted pulse is stretched in time, and develops a positive frequency chirp.

5.1.2 Propagation of Gaussian pulses

Considerable insight may be obtained by applying the general methods discussed above to the special case of an optical pulse with a Gaussian temporal profile. The general conclusions we draw will apply to pulses with arbitrary temporal profile.

Properties of a Gaussian pulse

We consider a generalized Gaussian pulse of the following form,[1]

$$E_{\rm in}(t) = {\rm e}^{-\Gamma t^2}{\rm e}^{-{\rm i}\omega_0 t} \tag{5.14}$$

where the complex Gaussian parameter describing the pulse is given by,

$$\Gamma \equiv a + {\rm i}b. \tag{5.15}$$

Let us first establish the main properties of this pulse. The full-width at half maximum duration of the temporal profile of the pulse *intensity* is easily shown to be,

$$\tau_{\rm p} = \sqrt{\frac{2\ln 2}{a}} = \sqrt{\frac{2\ln 2}{\Re(\Gamma)}}, \tag{5.16}$$

from which we note that the duration of the pulse depends only on $\Re(\Gamma)$.

[1]Note that Siegman writes the pulse in the form $\exp(-\Gamma t^2)\exp({\rm j}\omega_0 t)$ with $\Gamma \equiv a - {\rm j}b$. Consequently equations in Siegman's book may be converted to our notation by making the transformation ${\rm j} \to -{\rm i}$, and vice versa.

If $b \neq 0$ the instantaneous frequency changes within the pulse. To see this we note that the phase of the electric field is given by $\psi(t) = -\omega_0 t - bt^2$, and the instantaneous frequency given by,[2]

$$\omega(t) \equiv -\frac{\mathrm{d}\psi(t)}{\mathrm{d}t} = \omega_0 + 2bt. \tag{5.17}$$

Thus in general the frequency of the pulse changes linearly with time about the mean value of ω_0; this frequency chirp will be positive if b is positive.

To investigate how such a pulse propagates through an optical system it is necessary to find the frequency spectrum of the pulse by taking the Fourier transform:

$$a(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} E_{\mathrm{in}}(t) \mathrm{e}^{\mathrm{i}\omega t} \mathrm{d}t \tag{5.18}$$

$$\Rightarrow a(\omega) = \frac{1}{\sqrt{2\Gamma}} \mathrm{e}^{-(\omega-\omega_0)^2/4\Gamma} \tag{5.19}$$

$$= \frac{1}{\sqrt{2\Gamma}} \exp\left[-\Re\left(\frac{1}{4\Gamma}\right)(\omega-\omega_0)^2\right] \exp\left[-\mathrm{i}\Im\left(\frac{1}{4\Gamma}\right)(\omega-\omega_0)^2\right]. \tag{5.20}$$

It is clear that in general the phase $\psi(\omega)$ of the amplitudes of the frequency components varies quadratically with $(\omega - \omega_0)$ with a curvature determined by $\Im(1/\Gamma)$. The *power* spectrum, which is proportional to $|a(\omega)|^2$, has a full-width at half maximum of,

$$\Delta\omega_{\mathrm{p}} = 2\sqrt{\frac{2\ln 2}{\Re(1/\Gamma)}} = 2\sqrt{2\ln 2}\sqrt{a[1+(b/a)^2]}. \tag{5.21}$$

In terms of the full-widths at half maxima, the time-bandwidth product of the Gaussian pulse is found from eqns (5.16) and (5.21) to be

$$\Delta\omega_{\mathrm{p}}\tau_{\mathrm{p}} = 4\ln 2\sqrt{1+(b/a)^2}. \tag{5.22}$$

The time-bandwidth product clearly takes the smallest possible value if $b = 0$, i.e. if the pulse has no frequency chirp or, equivalently, if the phases of the frequency components are all the same.[3] Figure 5.2 summarizes the properties of Gaussian optical pulses of this type.

Propagation of a Gaussian pulse

We now consider the propagation of a Gaussian pulse through an optical system for which the linear and quadratic dispersion are finite, but the higher-order dispersion may be ignored:

$$\phi(\omega) = \phi^{(0)} + \phi^{(1)}(\omega-\omega_0) + \frac{1}{2}\phi^{(2)}(\omega-\omega_0)^2. \tag{5.23}$$

[2]Equation (5.14) may be written as $E_{\mathrm{in}}(t) = \exp(-at^2)\exp[-\mathrm{i}(\omega_0 + bt)t]$. However, it is *not* correct to deduce from this that the effective frequency is $\omega_0 + bt$.

[3]If the widths in eqn (5.22) are replaced with root-mean-square values the factor of $4\ln 2$ is replaced by π. We then see that the case $b = 0$ reduces to the minimum possible value according to eqn (5.1).

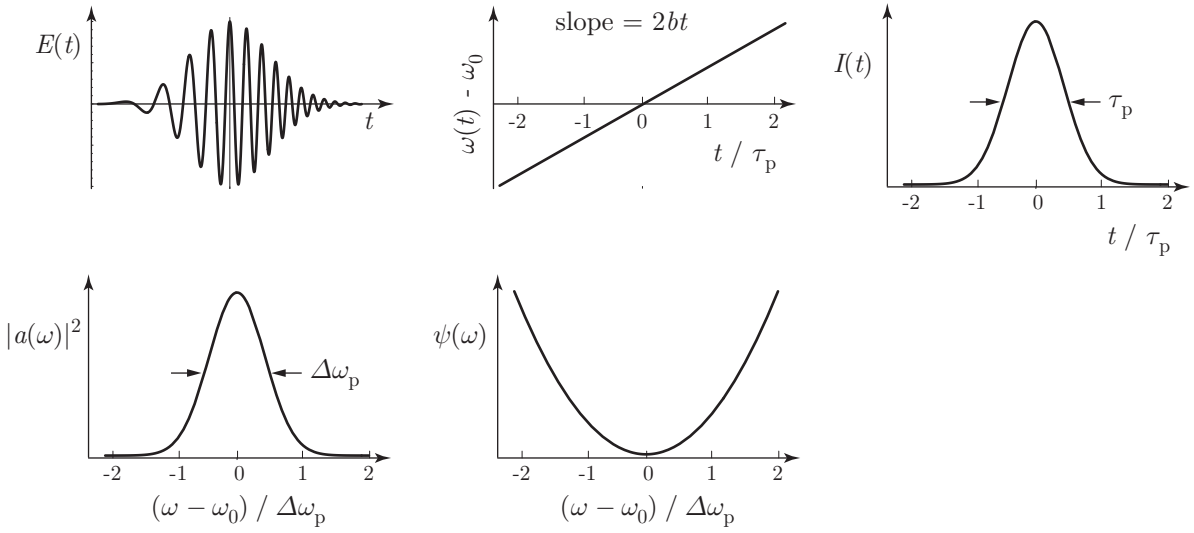


Figure 5.2: Properties of a Gaussian optical pulse with a pulse parameter $\Gamma = a + \mathrm{i}b$ for the case $b > 0$. Shown are the electric field $E(t)$, the instantaneous frequency $\omega(t)$, the temporal profile of the intensity $I(t)$ the power spectrum $|a(\omega)|^2$ and phase $\psi(\omega)$.

To calculate the form of the output pulse we multiply the frequency spectrum of the input pulse by $\exp[\mathrm{i}\phi(\omega)]$ and take the inverse Fourier transform. As explored in the problems, this turns out to be relatively straightforward, and we find

$$E_{\mathrm{out}}(t) = \sqrt{\frac{\Gamma_{\mathrm{out}}}{\Gamma_{\mathrm{in}}}}\exp[\mathrm{i}(\phi^{(0)} - \omega_0 t)]\exp\left\{-\Gamma_{\mathrm{out}}\left[t - \phi^{(1)}\right]^2\right\}, \tag{5.24}$$

where

$$\frac{1}{\Gamma_{\mathrm{out}}} = \frac{1}{\Gamma_{\mathrm{in}}} - 2\mathrm{i}\phi^{(2)}. \tag{5.25}$$

We see that the transmitted pulse still has a Gaussian temporal profile, but that the duration of the pulse envelope and the frequency chirp will in general be different from the input pulse since the parameter Γ will be changed. There will also be a relative shift between the peak of the Gaussian pulse envelope — which occurs at $t = \phi^{(1)}$, i.e. it is shifted in time by the group delay — and the carrier wave which is shifted in phase by $\phi^{(0)}$. Since in general $\phi^{(0)}/\omega_0 \neq \phi^{(1)}$ the carrier-envelope offset of the pulse will be altered by the optical system.

Note that there are strong parallels between eqn (5.24) describing the propagation of a Gaussian optical pulse and the equations describing the propagation of a lowest-order Gaussian beam. Comparison of these two situations reveals that the parameter $1/\Gamma$ plays an equivalent role to that of the complex radius of curvature q, and the dimension of time (shifted by $-\phi^{(1)}$) is equivalent to the transverse co-ordinate r. Thus compression (stretching) of a Gaussian pulse are analogous to focusing

(defocusing) of a Gaussian beam. Notice too that compression of a chirped Gaussian pulse requires the introduction of a medium with quadratic phase variation in frequency; whereas as focusing a Gaussian beam requires a lens which imposes a phase shift which varies quadratically with the transverse co-ordinates of the beam.

These points are made more obvious by considering the case of a uniform medium. Equation (5.24) may then be written as

$$E_{\rm out}(t) = \underbrace{\exp[-{\rm i}\omega_0(t - z/v_{\rm p})]}_{\text{carrier wave}} \overbrace{\sqrt{\frac{\Gamma_{\rm out}}{\Gamma_{\rm in}}}\exp[-\Gamma_{\rm out}(t - z/v_{\rm g})^2]}^{\text{pulse envelope}} \tag{5.26}$$

since $\phi^{(0)} = k(\omega_0)z$ and the group delay $\phi^{(1)} = z/v_{\rm g}$.

In general the pulse duration and frequency chirp of the optical pulse are altered by transmission through the optical system since the Gaussian parameter Γ is changed. Note, however, that the frequency width of the pulse is *not* altered by the propagation, as must be the case for a linear system.[4] Whether the optical pulse is compressed or stretched in time, and whether the pulse becomes more or less chirped, depends on the relative sign of the initial frequency chirp and the GDD of the optical system.

5.1.3 Nonlinear effects: self-phase-modulation and the B-integral

As we saw in Lecture 4, at high intensities the refractive index of a material becomes nonlinear with the intensity:

$$n = n_0 + n_2 I. \tag{5.27}$$

In propagating a distance L a wave of constant intensity I therefore accumulates a nonlinear phase

$$\phi_{\rm NL} \equiv B = \frac{2\pi}{\lambda_0}\int_0^L n_2 I(z)dz \tag{5.28}$$

in addition to that which would be accumulated by a beam of low intensity. Here λ_0 is the vacuum wavelength of the radiation. The integral above is universally known as the **B-integral** for the propagation.

The intensity of a beam of radiation will generally vary with transverse position within the beam, and, as we have seen, this usually leads to self-focusing of the beam. For a laser *pulse* the nonlinear phase will also vary with time during the pulse. Ignoring dispersion, the total accumulated phase of a wave after propagating a distance L into a medium will be given by,

$$\phi = \frac{2\pi}{\lambda_0}\left[n_0 + n_2 I(t)\right] L - \omega t. \tag{5.29}$$

[4] This may be seen from eqn (5.21); the width of the spectrum is determined by $\Re(1/\Gamma)$ and from eqn (5.25) $\Re(1/\Gamma_{\rm out}) = \Re(1/\Gamma_{\rm in})$.

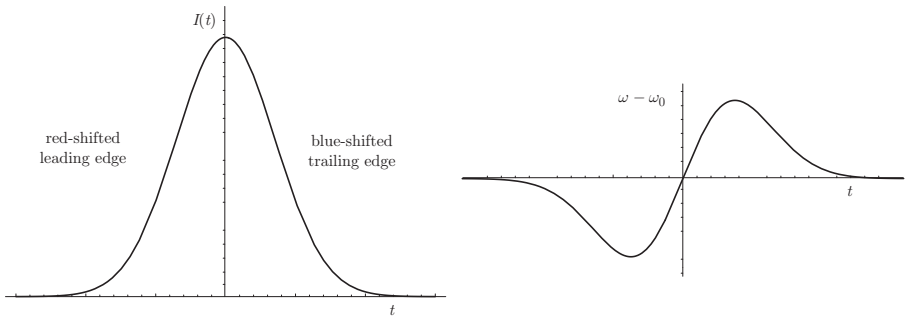


Figure 5.3: Generation of a frequency chirp by self-phase-modulation of an intense optical pulse.

At any point in the pulse, the effective frequency, ω', is given by the (negative) total rate of change of the phase with time:

$$\omega' = -\frac{d\phi}{dt} = \omega - \frac{2\pi}{\lambda_0} n_2 L \frac{\partial I}{\partial t}. \tag{5.30}$$

The nonlinear refractive index therefore causes the phase of the carrier wave in the pulse to vary nonlinearly with time, corresponding to a variation of frequency within the pulse. This process is known as **self-phase-modulation**.

Figure 5.3 shows this process schematically. The leading edge a pulse must have $\partial I/\partial t > 0$ corresponding to a decrease in frequency, or red shift, for materials with $n_2 > 0$. Similarly, the trailing edge of the pulse will be blue-shifted, and as a consequence the pulse will develop a positive chirp. The rate of change of frequency with time is approximately linear over the central part of the pulse, as shown in Fig 5.3.

Self-phase-modulation and self-focusing are nearly always undesirable in a laser system, in which case they must be kept to a minimum. A rule of thumb often employed by system designers is that the total B-integral should be kept below 1 radian.

5.2 Dispersion control

The linear and nonlinear dispersion discussed above causes distortion of short laser pulses as they propagate within a system. In order to overcome this — or at least alleviate it — it is necessary to employ techniques in which additional, controllable dispersion can be introduced. A simple way of achieving this is to pass the pulse through a series of optical components for which the physical (and hence optical) path is different for different wavelengths. In this way controllable, wavelength-dependent phases can be introduced. The dispersion introduced in this way is known as **geometric dispersion**.

Dispersion control is particularly important for modelocked lasers producing very short pulses. In order to generate the shortest possible pulses by modelocking it is necessary for the longitudinal modes must be evenly spaced in frequency, but the presence of dispersion in the cavity will prevent

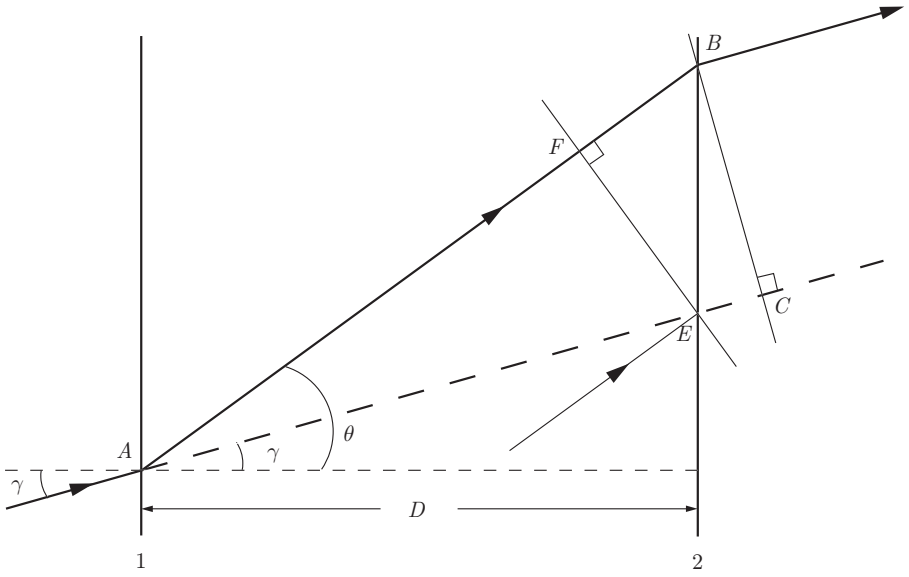


Figure 5.4: Schematic diagram of a general optical system exhibiting geometric dispersion.

this. Clearly the problem will become more acute as the duration of the modelocked pulses is reduced, and the associated bandwidth is increased. In practice it is found that dispersion control is necessary to generate pulses shorter than approximately 150 fs. Most materials exhibit positive dispersion, and hence geometric dispersion introduced into a laser cavity needs to be negative (and controllable).

Figure 5.4 shows a general system exhibiting geometric dispersion. Upon reaching the plane 1, light is refracted or diffracted through a wavelength-dependent angle θ. After propagating some distance it is refracted/diffracted again so as to render the direction of propagation parallel to the initial direction. Since θ depends on the wavelength, the optical path, and hence the group delay, will depend on wavelength; the system therefore introduces dispersion.

In calculating the phase accumulated by a given wavelength, we need to define two planes (an input plane and an output plane) through which all waves pass. There is no difficulty in identifying the input plane: it is the plane passing through A, perpendicular to the *incident* ray. However, the output plane needs to be defined more carefully. For example, the plane BC is not a suitable candidate since this plane is different for each wavelength. There is still a lot of flexibility: the output plane could be any fixed plane perpendicular to the rays *leaving* the system. One candidate would be such a plane passing through E.

However, it is more convenient to move the reference output plane back along the undeflected ray to A – i.e. the input plane! Using these reference planes the additional optical path travelled by the deflected ray is then equal to the distance $AF - AE$.[5]

Now,

[5]The calculated accumulated phase depends on where the input and output planes are defined to be. However, the calculated GDD (and higher orders) do not depend on these locations; all that will change is the total accumulated phase and the group delay for propagating between the two reference planes, just as we would expect.

$$AE = \frac{D}{\cos\gamma} \tag{5.31}$$

$$AF = AE\cos(\theta-\gamma). \tag{5.32}$$

Hence the optical path difference is given by,

$$OP = -\frac{D}{\cos\gamma}\left[1-\cos(\theta-\gamma)\right], \tag{5.33}$$

and hence the phase introduced is simply,

$$\phi(\omega) = -\frac{\omega}{c}\frac{D}{\cos\gamma}\left[1-\cos(\theta-\gamma)\right]. \tag{5.34}$$

It is a simple matter (in principle) to calculate the GDD etc. by differentiating eqn (5.34), providing the relation between θ and ω is known.

5.2.1 The grating pair

Figure 5.5 shows a ray propagating via a pair of parallel reflection grating. Light of wavelength λ will be diffracted at an angle θ from the first grating according to,

$$m\lambda = d[\sin\theta - \sin\gamma] \quad m = 0, \pm1, \pm2, \pm3, \ldots \tag{5.35}$$

where γ is the incident angle and d the grating spacing. At the second grating the incident angle will now be θ and consequently the second grating will diffract the light at an angle γ if the order of diffraction is m for the first grating and $-m$ for the second. The diffracted ray therefore leaves parallel to the incident ray.

The optical phase introduced by the grating pair is given by eqn (5.34), and this, together with eqn (5.35) allows one to calculate the GDD and other phase terms. For this simple case, however, it is easy to deduce the sign of the GDD. Consider the time taken for light to propagate from A to (say) the position of the mirror or roof prism, which we may take as a convenient reference plane. We know that the time taken for the undeflected light (dashed line) to reach this plane is a minimum, since the path it takes is the same as if the gratings were replaced by mirrors. All other paths must take a longer time, and since the diffraction angle θ increases with wavelength, we conclude that the group delay is smaller for shorter wavelengths (higher frequencies). Hence the group delay decreases with frequency, and the GDD must be negative[6]. It should also be clear that the GDD will increase linearly with

[6]Note that the group delay is NOT equal to the optical path difference given by eqn (5.33) divided by the speed of light. To use the optical path we must find the phase introduced by the system and then see how it varies with frequency. Unfortunately this is not as straightforward as the method used in the text to deduce the sign of the GDD: as ω increases the optical path difference decreases, but $\phi(\omega)$ is equal to this path difference multiplied by ω/c and hence it is not obvious whether overall $\phi(\omega)$ increases or decreases with ω. The method above works because we can use our insight to deduce that the group delay, $\phi^{(1)}(\omega)$, decreases with frequency.

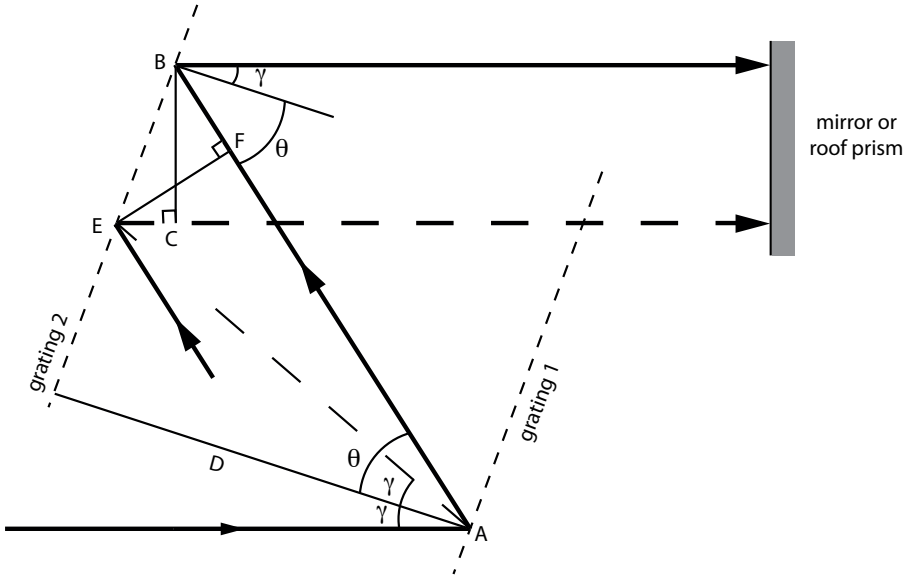


Figure 5.5: A grating pair used to introduce negative GDD.

the grating separation D, and consequently the grating pair can provide controllable negative GDD. Clearly the different wavelengths leave the grating pair at different heights above the incident ray. Consequently in the exiting beam there is a correlation between position and wavelength, a situation known as **spatial chirp**. In general this is undesirable, and the spatial chirp is undone by reflecting the beam back through the grating pair to form a beam counter-propagating along the axis of the incident. This counter-propagating beam may be selected by allowing the incident beam to enter the system at a small angle to the plane of the page; the exiting beam therefore leaves above the incident beam and may be picked off by a small mirror. Alternatively, and somewhat better, the beam may be returned by reflection from a roof mirror which returns the light parallel to the incident rays, but at a different height. Obviously, double-passing the beam through the grating pair will introduce twice the GDD generated from a single pass.

5.2.2 The prism pair

Dispersion control may also be achieved with the arrangement of prisms shown in Figure 5.6[7]. It is clear that the optical paths are different for different wavelengths, and consequently the arrangement will introduce geometric dispersion. Note that the plane AB is a plane of symmetry and consequently the total GDD is simply twice that introduced by the first prism pair. The advantage, however, of the arrangement of four prisms is that the emerging beam propagates along the same axis as the incident beam. This means that the optical system (in particular a laser cavity) can be aligned *without* the prisms in place, and once this is done the prisms can be introduced without changing the alignment

[7]Note that the diagram given in Svelto appears to be incorrect in that the higher frequency light is refracted through a smaller angle.

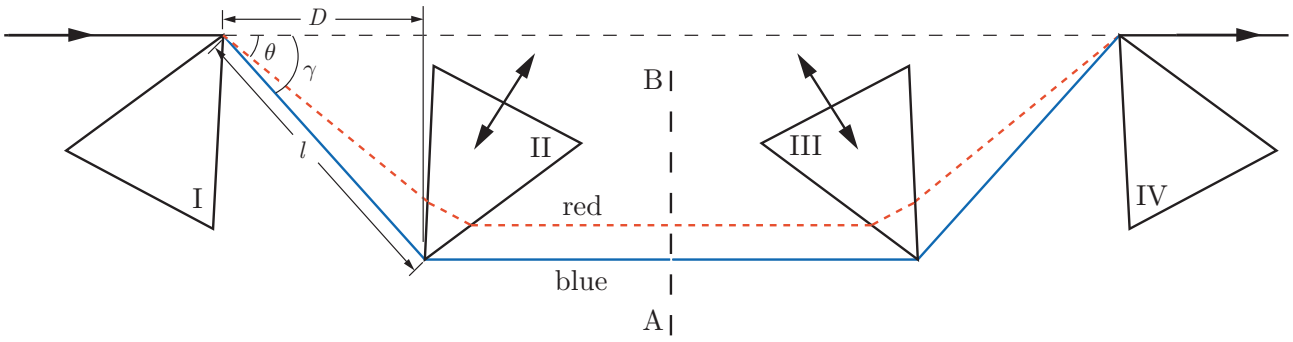


Figure 5.6: A prism pair used to introduce negative GDD. Note that the plane AB is a plane of symmetry; it would be possible to place a mirror here, reflecting rays back through the first two prisms to yield a counter-propagating beam with the same GDD as produced by the four prisms shown.

of other components.

It may be shown that for the first two prisms, the optical path and phase difference between a ray refracted at an angle θ and the ray which passes between the apexes of the prisms are also given by eqns (5.33) and (5.34). There is no simple argument for deducing the sign of the GDD in this case: indeed it may be positive or negative. To evaluate the GDD, and other terms, one must use Snell's Law to calculate θ as a function of frequency. This procedure is not difficult, but it is somewhat messy and gives us no particular insight. In fact it is found[8] that the GDD introduced by a prism pair is also negative, provided that l is sufficiently large. The GDD is proportional to $l = D/\cos\gamma$, and hence may be controlled by adjusting l. In practice, however, it is usual to introduce precise control of the GDD by moving prisms II and III perpendicular to their bases, as indicated in Fig. 5.6. This reduces l and introduces more prism material into the path, and therefore increases the GDD.

The prisms are usually oriented at Brewster's angle so as to minimize the insertion loss; and the apex angles of the prisms are chosen so that they operate at minimum deviation (i.e. the incident and exit angles are equal). The GDD introduced by prism pairs is typically much smaller than that provided by grating pairs, but have the advantage that the insertion loss can be much smaller. As such prism pairs are often used to provide dispersion control in modelocked laser oscillators.

5.2.3 Introduction of positive GDD

The two systems above both introduce negative GDD. Clearly we would also like to be able to introduce positive GDD (especially for CPA - see later), but this would appear to be difficult owing to the fact that for both these systems the GDD is negative for a positive separation D.

A clever solution to this problem is to introduce an imaging system between the two elements which can project the first element in front, behind, or at the same position as the second. In this way the effective separation of the elements may be made to be positive, zero, or negative which allows

[8]This was first shown by R. L. Fork *et al.*, Optics Letters vol. 9, p. 150 (1984).

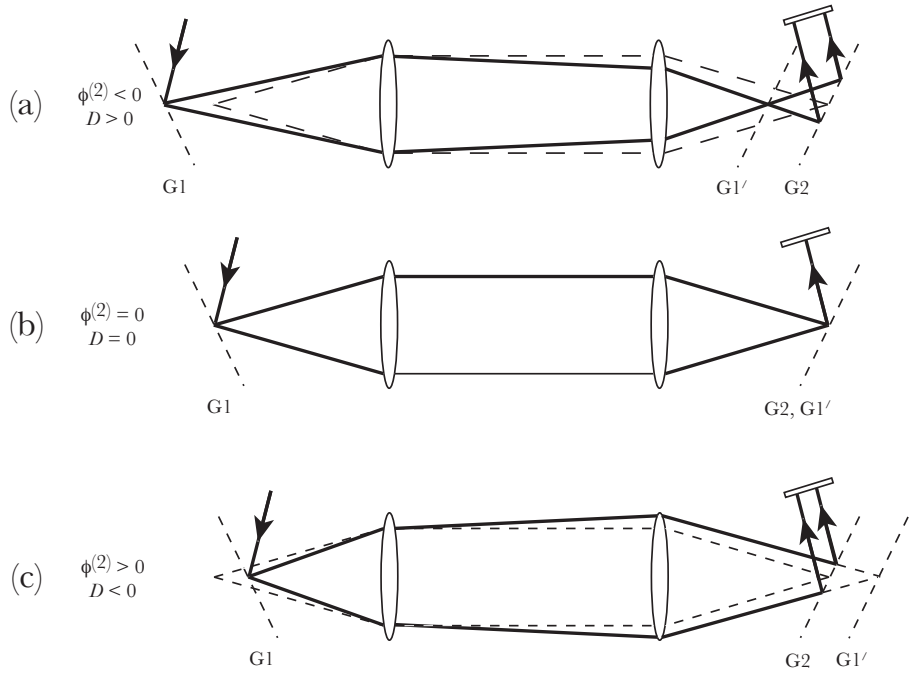


Figure 5.7: Grating configurations giving (a) negative, (b) zero, and (c) positive GDD.

the GDD to be negative, zero, or positive.

In practice this system is usually implemented with diffraction gratings, as shown in Figure 5.7. The two lenses form an image, with unit magnification, of grating $G1$. By adjusting the position of $G1$, its image, $G1'$, may be positioned either side of the second grating $G2$. Notice that the longitudinal magnification is equal to -1, and hence the imaging system changes the orientation of $G1'$ with respect to $G1$.

Many ingenious improvements have been made to this simple idea. More advanced systems replace the lenses with curved mirrors, avoiding problems caused by chromatic aberration. It is also possible to use a folded, single-grating design which automatically ensures alignment of the 'two' gratings as the other components are adjusted.

5.2.4 Chirped mirrors

A recently developed way of introducing controlled GDD is reflection from a **chirped mirror**. A standard 'dielectric mirror' comprises a stack of alternating layers of high and low refractive index material, the thickness of each mirror being equal to a quarter of the wavelength in the medium. In a chirped mirror the thickness of the layers is varied within the stack so that different wavelengths are reflected most strongly at different depths in the stack, as illustrated schematically in Figure 5.8. With careful design and manufacture, it is possible to produce any desired GDD – as well as controlled higher-order phase terms. If the GDD produced by one mirror proves insufficient, it may be increased by multiple reflections.

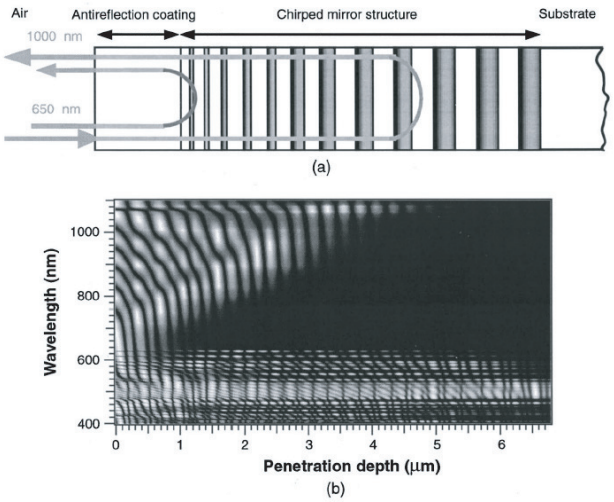


Figure 5.8: Schematic diagram showing: (a) the principle of operation of a chirped mirror; (b) the calculated standing wave pattern in a chirped mirror as a function of wavelength. Note how for this mirror the penetration depth increases with wavelength, giving negative GDD.

Chirped mirrors have the advantage that they are simple to align and set up, and introduce a minimum of optical elements into the system. A disadvantage is that they must be designed specifically for each application, and are therefore less flexible.

5.3 TW and PW laser systems

Perhaps the supreme example of the application of dispersion control is the generation of laser pulses with peak powers in the terawatt ($1\,\mathrm{TW} \equiv 10^{12}\,\mathrm{W}$) or even the petawatt ($1\,\mathrm{PW} \equiv 10^{15}\,\mathrm{W}$) range.[9]

Imagine trying to generate laser pulses with a peak power of 1 TW. If the pulse had a duration of $1\,\mu\mathrm{s}$, our 1 TW laser pulse would require an energy of 1 MJ! The first trick, then, in generating high peak powers is to keep the duration of the pulses short; modelocked pulses in the femtosecond to picosecond ranges are used for ultra-high power laser systems.

The laser pulses generated by modelocked lasers typically have very low energies — a few nanojoules is quite common. In order to reach the TW or PW level the pulse energy must be amplified by factors of 10^6 to 10^{10}. If we were to attempt to do this in a simple chain of amplifiers, we would soon reach peak intensities that would cause strong self-focusing, leading to damage of the optical components in the amplifier chain. In practice the intensity to which the pulses can be amplified is often even lower than the limit set by optical damage; at lower intensities self-phase-modulation leads to the development of a complicated phase structure to the pulse which cannot be compensated easily by controlled dispersion. An obvious way to keep the intensity reasonable is to increase the diameter of

[9] An excellent review of TW and PW laser systems may be found in S. Backus, Review of Scientific Instruments vol. 69 p. 1207 (1998). Further information on PW lasers systems can be found at the NIF website: www.llnl.gov/nif/.

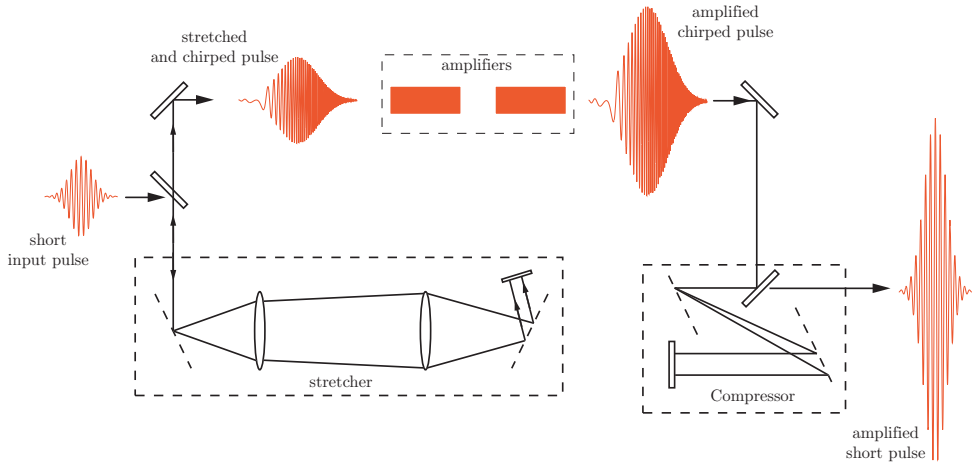


Figure 5.9: Schematic diagram of chirped pulse amplification.

the laser beam. However, this approach quickly requires large diameter laser rods and optics, which are both expensive (if they are available) and unwieldy.

The solution to this problem is **chirped pulse amplification (CPA)**, the technique for which Donna Strickland and Gérard Mourou won (half of) the 2018 Nobel Prize in Physics 'for their method of generating high-intensity, ultra-short optical pulses'.

In this approach the low energy pulses to be amplified to high energy, high peak power pulses in three stages:

1. The pulses are stretched in time by passing them through a pulse stretcher with a large GDD. Stretch factors of order 10^3 to 10^5 are common.

2. The stretched pulses are then amplified in a chain of one or more amplifiers. Since the duration of the stretched pulses is many orders of magnitude longer than the original pulses, they may be amplified to higher energies than in the absence of stretching by a factor approximately equal to the stretch factor.

3. The amplified pulse is then re-compressed to (approximately) the duration of the original pulse.

The CPA process is illustrated schematically in Figure 5.9

The pulse stretcher usually comprises the grating pair illustrated in Figure 5.7(c), set to produce positive dispersion. One reason for this choice is that the material of the following amplifier chain will tend to introduce more positive dispersion, and therefore there is no danger of partially re-compressing the laser pulses as they are amplified, which could lead to higher than expected peak pulse intensities in the laser amplifiers.

5.3.1 Regenerative and multipass amplifiers

In a CPA laser there will typically be several stages of amplification. Immediately after the pulse has been stretched, there will be one or more pre-amplifiers in which most of the gain occurs (perhaps a

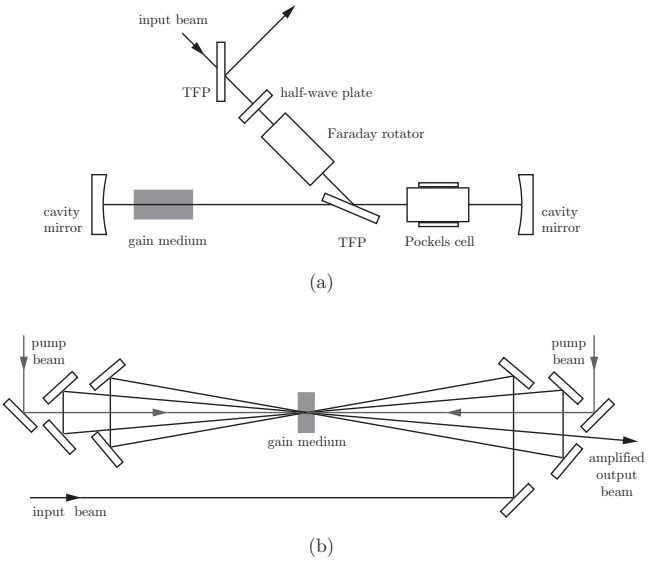


Figure 5.10: Schematic diagram of (a) a regenerative and (b) a multipass amplifier.

factor of 10^7). After the pre-amplifiers the energy of the stretched pulses will be sufficiently high that they can extract energy efficiently from one or more power amplifiers. Most of the energy in the final pulse is added by the power amplifiers.

There are two main classes of amplifier used, as illustrated schematically in Fig. 5.10.

Regenerative amplifiers

In a regenerative amplifier the amplification takes place in an optical cavity. Laser pulses are injected into the cavity by applying a voltage pulse to a Pockels cell as shown schematically in Fig. 5.10(a). In the absence of any voltage on the Pockels cell, the incident laser pulses will, after reflection from the cavity end mirror be reflected out of the cavity by the polarizer. However, if a quarter-wave voltage is applied to the Pockels cell, the pulses will return to the intra-cavity polarizer with the orthogonal polarization and hence will be trapped in the cavity and be amplified. After amplification the pulses are switched out by applying a second quarter-wave pulse to the Pockels cell[10].

[10]The Faraday rotator is included in the arrangement in order to prevent pulses being reflected back into earlier parts of the laser chain (such as the laser oscillator), since such reflections can cause damage or instabilities. It's inclusion complicates a description of the operation of the regenerative amplifier somewhat, although the essential features are included in the discussion above. Let us suppose that the input pulses are vertically polarized. The half-wave plate is oriented at 45° to the vertical so as to rotate the plane of polarization to the horizontal plane. On on passing through the Faraday rotator the plane of polarization is rotated by (say) +45°. In the absence of any voltage on the Pockels cell, the pulses will be rejected by the cavity and their polarization further rotated by the Faraday rotator in the *same sense*, i.e. another +45°, to give vertically-polarized light. The half-wave plate then flips the polarization to the horizontal plane, and consequently the pulses are reflected by the external polarizer. Trapping laser pulses within the cavity operates as discussed above, i.e. applying a voltage to the Pockels cell rotates the polarization of the laser pulses so that they are transmitted by the intra-cavity polarizer. The pulses are rejected from the cavity by applying a second pulse to the Pockels cell, whereupon they are reflected by the external polarizer in the manner discussed in this note.

Since the gain medium is located within an optical cavity, the pumping must be at a level which prevents the single-pass gain from becoming too high, or lasing — or at least strong amplification of spontaneous emission — will occur even in the absence of an injected pulse. As such, in order to extract energy from the gain medium it is necessary for the pulse to undergo many cavity round trips, perhaps of order 20.

The advantages of regenerative amplifiers are that the optical cavity imposes a good transverse mode structure on the amplified pulses. As such, the beam quality should be good. There are several disadvantages, however. Leakage through the polarizers and Pockels cell can give rise to a series of lower-energy pulses propagating ahead of the main pulse. Prepulses of this type can be problematic in many applications. Further, the many round-trips required to extract the energy stored in the population inversion introduces a long length of material into the path of the pulse (100s of millimetres). The consequent linear and non-linear phase structure introduced into the pulse can be difficult to compensate for. Notwithstanding these difficulties, regenerative amplifiers have been used to generate TW pulses with pulse durations as short as 30 fs.

5.3.2 Multipass amplifiers

In a multipass amplifier the pulse is passed through the gain medium several times by reflection from a series of mirrors – sometimes referred to as **angular multiplexing**. Since the gain medium is not located within a cavity the single-pass gain may be much higher without causing too much **amplified spontaneous emission (ASE)**. Multipass amplifiers therefore have the advantage of introducing less material into the path of the pulse, as well as not suffering from prepulses. The disadvantages are that the overlap between the pump laser beam and the amplified laser pulse is not as good as in a regenerative amplifier, and consequently the efficiency will not be as high as in a regenerative amplifier. A typical efficiency for a multipass pre-amplifier is 15%; although multipass power amplifiers may reach 30%, more or less equal to that achieved in a regenerative amplifier.

5.3.3 Low-energy, TW laser systems

A good example of a low-energy TW laser is the system used in the Clarendon Laboratory, which has the following key components:

1. A Ti:sapphire oscillator pumped by a frequency-doubled CW Nd:YLF laser. The oscillator generates a stream of modelocked pulses with a pulse repetition rate of 86 MHz. The duration and energy of each pulse are approximately 35 fs and 10 nJ respectively.

2. A pulse stretcher which stretches the pulses to a duration of approximately 300 ps.

3. A subset of the stretched pulses are selected by a Pockels cell and injected into a **regenerative amplifier** operated at a repetition rate of 1 kHz. The regenerative amplifier comprises a 5 mm diameter Ti:sapphire rod pumped by pulses from a diode-pumped, frequency-doubled Q-switched Nd:YAG laser operating at a pulse repetition rate of 1 kHz The regenerative amplifier increases the energy of the pulses to 0.7 mJ.

4. A subset of the pulses from the regenerative amplifier are injected into a **multipass amplifier** comprising a 20 mm diameter Ti:sapphire rod pumped by a flashlamp-pumped, frequency-doubled, Q-switched Nd:YAG laser which delivers approximately 1 J of green pump light per pulse. This amplifier increases the energy of the pulses to 200 mJ.

5. Finally the amplified pulses are compressed by a double-passed grating pair to a duration of approximately 40 fs. Losses in the compressor are dominated by the four reflections from the diffraction gratings, and are such that the compressed laser pulses have an energy of approximately 120 mJ.

6. The compressed pulses therefore have a peak power of 3 TW and a duration quite close to those from the oscillator.

5.3.4 High-energy, PW laser system

In order to reach very high pulse energies (100s of Joules) it is necessary to increase the beam diameter to 100 mm or more. At present it is not possible to grow large-diameter crystals of Ti:sapphire with sufficient optical quality for this purpose. As a consequence, high-energy systems use as the active medium Nd:Glass since the technology for producing optical-quality rods and disks of this material is well developed.

The laser transition in Nd:Glass is inhomogeneously broadened with a large linewidth of approximately 50 THz, and hence in principle can produce modelocked laser pulses with durations as short as approximately 100 fs. In practice, pulses of duration 0.5 - 1 ps can be generated. Modern high-energy 'glass' laser systems amplify pulses from a Ti:sapphire oscillator and pre-amplifier system, the Ti:sapphire being tuned in wavelength to match the peak of the gain of the Nd:Glass laser. Typically a series of multipass rod amplifiers is used for amplification to pulse energies of several tens of Joules. Amplification to higher energies requires the use of larger beam diameters, and hence large-area discs of Nd:Glass are used.

A good example of a PW class laser is the Vulcan laser at the Rutherford Appleton Laboratory, just down the road near Didcot. This was the first laser in Europe to reach the PW level. The laser has the following features:

- Pulses with an energy of 5 nJ, a duration of 120 fs, and a wavelength of 1055 nm are generated by a modelocked Ti:Sapphire oscillator.

- These pulses are stretched by a factor of 4×10^4 to a duration of 5 ns, and then amplified to an energy of order 10 mJ using a new technique known as optical parametric chirped pulse amplification (OPCPA). In this approach the gain medium is replaced by a non-linear, optical parametric amplifier and the pulses are amplified by conversion of energy from the pump laser beam to the wavelength of the pulses. This method avoids the difficulties of gain narrowing and reduces the pedestal level normally associated with regenerative amplifiers.

- The pre-amplified pulses are then introduced into a chain of flashlamp-pumped rod and disk Nd:Glass amplifiers, which amplifies the pulse energy to 670 J.

- Two gold-coated holographic diffraction gratings (940 mm diameter!) with a groove spacing of 480 lines mm^{-1} re-compress the pulse to 650 fs; somewhat longer than the original pulses, but not bad considering that they have been amplified by a factor of 10^{11}!

- After losses caused by reflection at the compressor gratings and beam-steering optics, the energy of the pulses available to the target chamber is approximately 500 J, i.e. a peak power of 0.8 PW. This can be focused to a spot size of order 7 μm diameter, corresponding to a peak focused intensity of order $10^{21}\,\mathrm{W\,cm^{-2}}$. Note that at this intensity the radiation pressure is 11 orders of magnitude greater than atmospheric pressure!

5.4 Applications of ultrafast lasers

At low pulse energies (fraction of a millijoule), ultrafast lasers find a myriad of applications in 'pump-probe' experiments in which an event, such as a chemical reaction, is triggered by a short laser pulse, and the subsequent evolution is studied using a second, short probe laser pulse. The ability to control the phase structure of short laser pulses is opening up the field of 'coherent control' in which, for example, the products of photo-initiated chemical reactions are selected by controlling the precise variation of the electric field within the laser pulse.

At somewhat higher pulse energies (several millijoules), focusing the laser pulse to a beam diameter of a few tens of microns results in a peak laser intensity of $10^{15}\,\mathrm{W\,cm^{-2}}$ and higher. At such intensities the electric field of the electromagnetic wave is comparable, or stronger, than the Coulomb field binding the valence electrons in an atom. A number of applications arise:

- High-harmonic generation. Generation of very high-order (of order 100) harmonics of the driving laser wavelength to produce beams of coherent radiation with wavelengths in the vacuum ultraviolet to soft x-ray spectral region. High-harmonic generation requires relatively low driving intensities of order $10^{15}\,\mathrm{W\,cm^{-2}}$.

- Driving 'table-top' soft x-ray lasers. Typical intensities required for this are 10^{16} - 10^{17} $\mathrm{W\,cm^{-2}}$.

- Plasma accelerators. At intensities of order $10^{18}\,\mathrm{W\,cm^{-2}}$, a laser pulse propagating through a plasma will drive a longitudinal plasma wave that trails the laser pulse in much the same way that a wake follows a boat travelling across a lake. The longitudinal electric fields within the plasma wave can be at least three orders of magnitude greater than those used in conventional particle accelerators, such as those at CERN. For example, recently experiments by groups in Oxford and Lawrence Berkeley National Laboratory demonstrated acceleration of electrons to 1 GeV energy (c.f. the electron rest mass of 0.5 MeV) in an laser-driven plasma accelerator only 33 mm long.

At somewhat higher pulse energies (several Joules), a variety of exotic phenomena become possible:

- Laser-driven fusion. Of particular interest is so-called fast-ignitor inertial confinement fusion, in which a pre-heated core of material is rapidly heated by an intense ultrashort laser pulse to initiate fusion.

- Generation of ultra-intense magnetic fields (up to 10^5 T) during the interaction of intense laser pulses with a solid target.
- Generation of intense, short-pulse proton or heavy-ion beams.
- 'Laboratory Astrophysics'. The extreme temperatures, pressures, and magnetic fields that can be created when a ultra-high intensity laser interacts with matter allows us to recreate (briefly) the conditions found in several types of astrophysical objects.

Lecture 6

Semiconductor & Fibre lasers

6.1 Review of semiconductor physics

As the constituent atoms of a semiconductor crystal are brought together from infinity, the wave functions of the valence electrons begin to overlap. At inter-atomic separations characteristic of the solid state, the wave functions are sufficiently overlapped that the valence electrons are delocalized and can no longer be associated with a particular atom.

The electron wave functions can then be written as **Bloch wave functions**:

$$\psi(\boldsymbol{r}) = u_k(\boldsymbol{r}) \exp(\mathrm{i}\boldsymbol{k} \cdot \boldsymbol{r}), \tag{6.1}$$

where $u_k(\boldsymbol{r})$ has the periodicity of the crystal lattice.

The energy of the electron is a function of the wave vector $\boldsymbol{k}$. For a *free* electron the energy and momentum of the electron would be given by,

$$\boldsymbol{p} = \hbar\boldsymbol{k} \tag{6.2}$$

$$E = \frac{\hbar^2 k^2}{2m_\mathrm{e}}, \tag{6.3}$$

where m_e is the rest mass of a free electron. However, for electrons moving in the periodic potential of the crystal lattice this relation no longer applies. Deviation from the parabolic dependence of energy on k becomes particularly strong when the wave vector k approaches one of the Brillouin zones[1], which is to be expected since this is just the condition for strong Bragg reflection from the crystal planes. The distortion of the electron energy from the free space value leads to some energies being forbidden and the formation of allowed energy bands with energies $E_n(\boldsymbol{k})$ which are functions of $\boldsymbol{k}$ and an index $n = 1, 2, 3, \ldots$. The formation of the band structure is illustrated schematically in Figure 6.1.

We note that the Bloch wave functions are not momentum eigenstatates[2], and hence $\hbar\boldsymbol{k}$ is not the momentum of the electron. However, the quantity $\hbar\boldsymbol{k}$ does play a role akin to momentum in

[1] The first Brillouin zone is defined as the Wigner-Seitz primitive cell of the reciprocal lattice; the Wigner-Seitz primitive cell about any lattice point is the region of space which is closer to that lattice point than to any other.

[2] Fundamentally this relates to the fact that the Hamiltonian is not invariant under arbitrary translations.

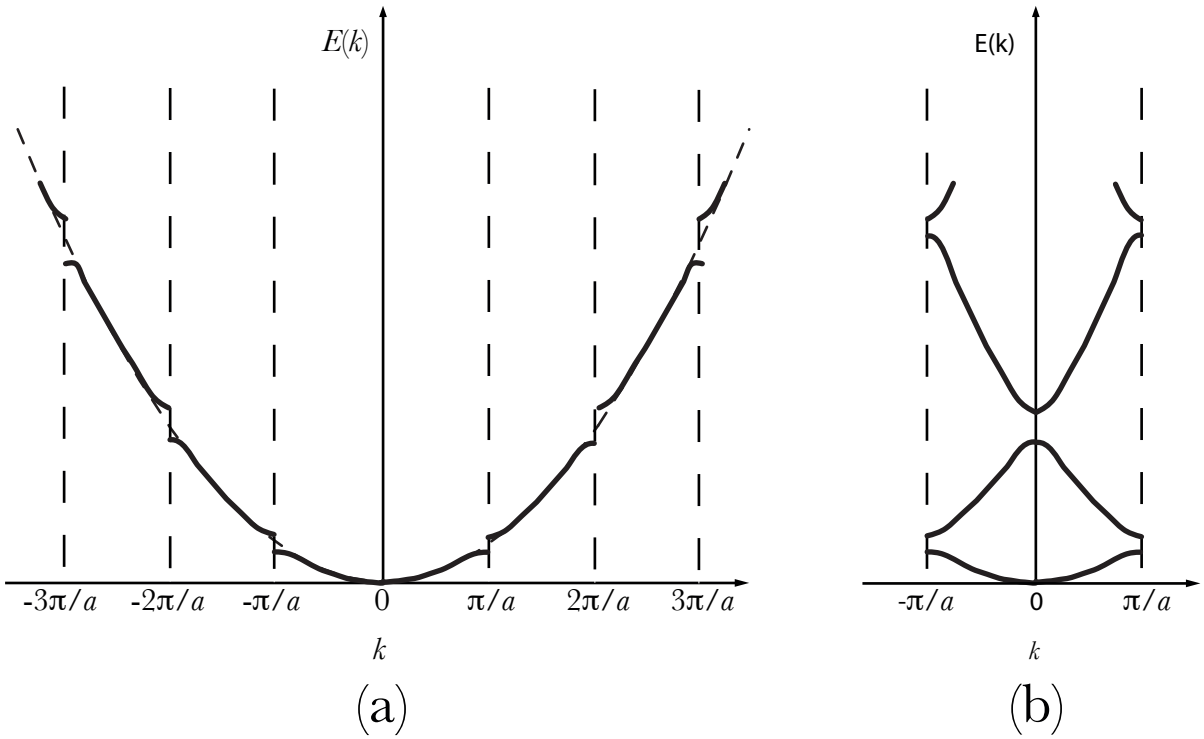


Figure 6.1: Schematic diagram of the formation of band structure in a one-dimensional crystal lattice of lattice spacing a. In (a) is shown the parabolic dependence of the energy $E(k)$ on the wave vector k for a free electron (dashed curve). This dependence is strongly distorted for values of k close to a Brillouin zone, leading to the formation of bands of allowed energies separated by energy gaps. In (b) the bands are shown in the **reduced zone scheme** in which reciprocal lattice vectors are added to (or subtracted from) k so as to bring everything into the first Brillouin zone. This last picture is usually the one used when discussing semiconductors.

conservation laws describing the interaction of electrons with fields or phonons. For this reason it is known as the **crystal momentum**.

The number of electron states per band is equal to twice the number of unit cells in the crystal. A semiconductor contains an even number of electrons per unit cell, and consequently at absolute zero all bands are either completely full or completely empty. As such, at absolute zero a semiconductor has zero conductivity. What differentiates a semiconductor from an insulator is that the energy gap between the top of the highest full energy band, the **valence band**, and the next highest (empty) band, the **conduction band**, is not too large compared to $k_{\mathrm{B}}T$. Typical values of this **band gap** energy are 0.5 - 3.0 eV, compared to $k_{\mathrm{B}}T \approx 0.025$ eV. Thermal excitation of electrons from the valence band to the conduction band leads to a finite electrical conductivity which is between that of insulators and conductors and increases strongly with temperature.

Effective mass

It may be shown that under a force $\boldsymbol{F}$ an electron in one of the energy bands moves according to,

$$\boldsymbol{F} = \hbar^2 \left(\frac{\partial^2 E_n}{\partial k^2} \right)^{-1} \frac{d\boldsymbol{v}_g}{dt}, \tag{6.4}$$

where v_g is the group velocity of the electron. As such, it is useful to define the **effective mass**:

$$m^* = \hbar^2 \left(\frac{\partial^2 E_n}{\partial k^2} \right)^{-1}. \tag{6.5}$$

Hence the effective mass of the electron depends on the curvature of the band energy $E(k)$. Note that for a free electron, the definition above gives $m^* = m_\mathrm{e}$.

Holes

Figure 6.2 shows a semiconductor in which some electrons have been excited from the valence band to the conduction band to give two partially-filled bands. In describing the partially-filled valence band, it is often easier to focus attention on the relatively few empty electron states than on the many filled states. Empty electron states are known as **holes**, and have the following properties:

- $\boldsymbol{k}_\mathrm{h} = -\boldsymbol{k}_\mathrm{e}$, where $\boldsymbol{k}_\mathrm{h}$ is the wave vector of the hole, and $\boldsymbol{k}_\mathrm{e}$ that of the missing electron.
- $E_\mathrm{h} = -E_\mathrm{e}$, where E_h is the energy of the hole, and E_e is the energy that an electron would have if it occupied the empty state.
- $\boldsymbol{v}_\mathrm{h} = \boldsymbol{v}_\mathrm{e}$
- $m^*_\mathrm{h} = -m^*_\mathrm{e}$. This follows from the first and second points above.
- The hole moves under the influence of external electromagnetic fields as if it had a charge of $+e$.

Hence the holes behave as particles of positive charge, with an energy and mass equal to minus that of the missing electron. Notice that for the partially-filled valence band, the curvature of the electron energy in the region near the top of the band is negative, and hence their effective mass would be negative. In contrast, the holes appear to have a positive effective mass which is another reason for dealing with the hole states rather than the electron states.

6.1.1 Distribution over energy levels

The probability of a level with an energy between E and $E + dE$ being occupied by an electron is given by the Fermi-Dirac distribution:

$$f(E)dE = \frac{1}{\exp([E - E_\mathrm{F}]/k_\mathrm{B}T) + 1}, \tag{6.6}$$

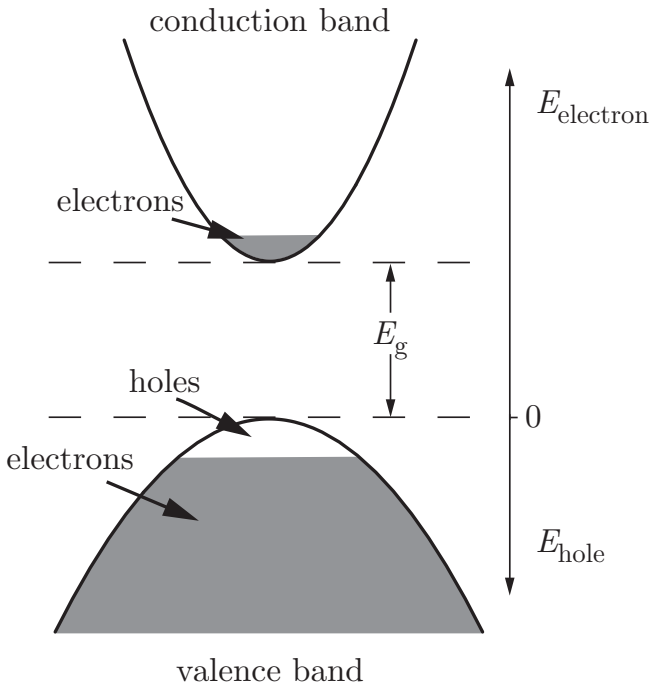


Figure 6.2: Partially-filled valence bands following promotion of electrons from the conduction to the valence band.

where E_F is the **Fermi energy**. From eqn (6.6) we see that the Fermi energy is that energy for which a level has a probability of 1/2 of being occupied. Alternatively, as shown in Figure 6.3(a), the Fermi energy gives the energy to which the band is filled at $T = 0$.

Imagine that we transfer N_e electrons per unit volume from the valence band into the conduction band. After a few picoseconds the electrons will equilibrate within the conduction band to form a local thermal equilibrium at temperature described by T and a **quasi-Fermi level** E_{Fc}. Similarly the electrons in the valence band will equilibrate to a thermal distribution described by a quasi-Fermi level E_{Fv}. Hence the probabilities of a level of energy E being occupied are :

$$f_c(E) = \frac{1}{\exp([E - E_{Fc}]/k_B T) + 1} \tag{6.7}$$

$$f_v(E) = \frac{1}{\exp([E - E_{Fv}]/k_B T) + 1}, \tag{6.8}$$

for the conduction and valence band respectively.

6.1.2 Doped semiconductors

In an ultra-pure sample of a semiconductor compound such as GaAs the number of electrons excited across the energy gap up into the conduction band at room temperature is very small. As indicated

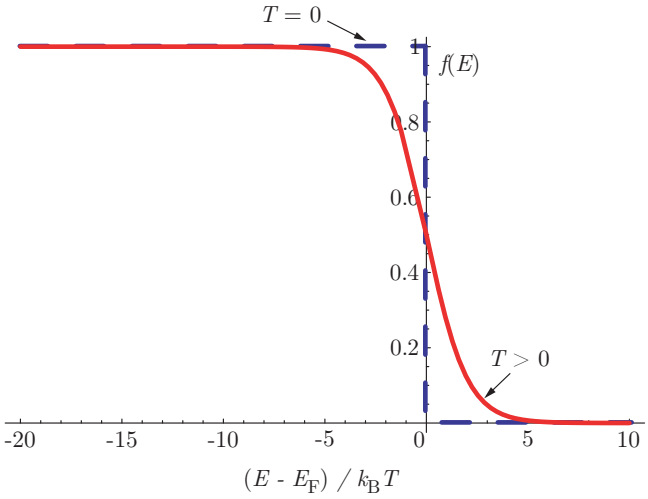


Figure 6.3: The Fermi-Dirac distribution for $T = 0$ (dashed line) and $T > 0$ (solid line).

in Fig. 6.4(a) for such an ultra-pure semiconductor the Fermi energy would lie halfway up the energy gap between the top of the valence band and the bottom of the conduction band, corresponding to the energy for which the probability of occupation is $\frac{1}{2}$.

Higher concentrations of electrons in the conduction band are required for most practical applications. This is achieved by adding carefully controlled concentrations of particular doping species to the bulk semiconductor compound. For example, for the case of GaAs, adding small proportions of Si atoms or Te atoms[3] we introduce a new energy level corresponding to the impurity species which lies just below the conduction band of the host material, as indicated schematically in Figure 6.4. Atoms of these particular added species (Si or Te) have one more electrons in their outer shells than the atoms they replace and can therefore act as donor species by providing a source of electrons which go into the conduction band, because the gap between the impurity level and the conduction band of the host material is small and the impurity atoms are easily ionized at room temperature. This type of material, with an excess of electrons in the conduction band is called **n-type**, and at light doping levels provides a number density of electrons in the conduction band, which although much larger than in the case of the intrinsic undoped material, nevertheless depends strongly on temperature and tends to zero at low temperatures. Under conditions of light doping the Fermi energy lies between the impurity level and the bottom of the conduction band.

In contrast if the same n-type material is **heavily** doped the impurity level merges into the conduction band, and the Fermi level moves up to sit close to the top of the occupied states in the conduction band as indicated in Figure 6.4(c).

In an analogous way, by doping GaAs with electron acceptor atoms such as Be, Zn, or C, **p-type** materials with excess holes in the valence band can be produced and at heavy doping levels the Fermi

[3]Si belongs to group IV and substitutes for atoms originally on Ga sites. Te belongs to group VI and substitutes for atoms originally on As sites. In the case of n-type GaAs, it is Si which is used most commonly.

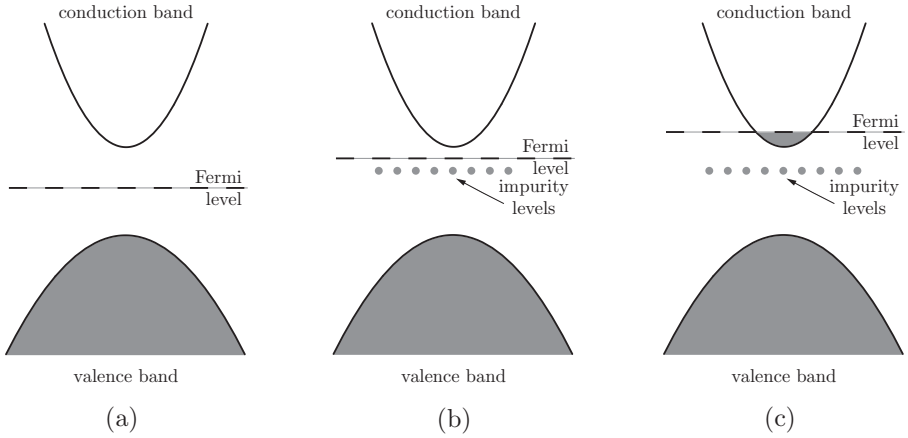


Figure 6.4: Occupancy of conduction and valence bands and position of the Fermi level for: (a) ultrapure semiconductor; (b) after the introduction of a small density of an electron donor impurity species; (c) after the introduction of a large density of electron donor species.

level of such materials moves down into the valence band[4].

Heavily doped semiconductor materials are known as **degenerate** semiconductors. Because they have partially filled bands (like metals) they remain electrically conducting even at very low temperatures.

6.2 Radiative transitions in semiconductors

Suppose that an incident harmonic wave of angular frequency ω causes an electron to make a transition from an upper level of energy E_2 in the conduction band to a level in the valence band of lower energy E_1. If we write the electric field of the incident wave in the form,

$$\boldsymbol{E}(\boldsymbol{r},t) = \boldsymbol{E_0} \exp\left[\mathrm{i}(\boldsymbol{k}_{\mathrm{opt}} \cdot \boldsymbol{r} - \omega t)\right], \tag{6.9}$$

the perturbation takes the form,

$$H' = e\boldsymbol{E} \cdot \boldsymbol{r}. \tag{6.10}$$

The rate of transitions induced by the perturbation is given by Fermi's golden rule:

$$W_{21} = \frac{\pi^2}{\hbar^2} \left|H_{21}\right|^2 \delta(\omega - \omega_{21}), \tag{6.11}$$

[4]Be belongs to group IIA and Zn to group IIB, and these atoms substitute for Ga to give rise to excess holes. C belongs to group IVB and substitutes for As.

where,

$$\hbar\omega_{21} = E_2 - E_1, \tag{6.12}$$

and,

$$H_{21} = \int \psi_2^*(\boldsymbol{r}) \left[e\boldsymbol{r} \cdot \boldsymbol{E_0} \exp(\mathrm{i}\boldsymbol{k}_{\mathrm{opt}} \cdot \boldsymbol{r})\right] \psi_1(\boldsymbol{r}) d\tau \tag{6.13}$$

$$= \int u_2^*(\boldsymbol{r}) \left\{e\boldsymbol{r} \cdot \boldsymbol{E_0} \exp\left[\mathrm{i}(\boldsymbol{k}_{\mathrm{opt}} + \boldsymbol{k}_1 - \boldsymbol{k}_2) \cdot \boldsymbol{r}\right]\right\} u_1(\boldsymbol{r}) d\tau. \tag{6.14}$$

We see from eqn (6.14) that the integrand will oscillate rapidly during the integration, and hence be zero, unless

$$\hbar\boldsymbol{k}_2 = \hbar\boldsymbol{k}_1 + \hbar\boldsymbol{k}_{\mathrm{opt}}. \tag{6.15}$$

This, then, is a selection rule for optical transitions between electron levels in a semiconductor; it resembles a statement of conservation of momentum, except that the quantities $\hbar\boldsymbol{k}_{2,1}$ are *crystal* momenta. In the optical region $|k_{\mathrm{opt}}| \ll |k_{1,2}|$ and hence the selection rule becomes,

$$\hbar\boldsymbol{k}_2 \approx \hbar\boldsymbol{k}_1, \tag{6.16}$$

and hence $\boldsymbol{k}$ is unchanged in the transition — the transitions are 'vertical' in our diagrams.

We also note that since the perturbation does not operate on the electron spin s, we also have the selection rule,

$$\Delta s = 0. \tag{6.17}$$

6.2.1 Joint density of states

We now consider a transition between two levels of energy $E_2 = E_{\mathrm{c}}(k_2)$ and $E_1 = E_{\mathrm{v}}(k_1)$ of a given value of $k = k_1 \approx k_2$. In any particular band the density of states with wavevectors between k and $k + \mathrm{d}k$ is given by the well known result:

$$g(k)\mathrm{d}k = \frac{k^2}{\pi^2}\mathrm{d}k. \tag{6.18}$$

We wish to find the density of transitions with a given energy difference $\hbar\omega$. To this end we note that within the parabolic approximation we may write the energies of the upper and lower electron states as,

$$E_2 = E_g + \frac{\hbar^2 k^2}{2m_c^*} \tag{6.19}$$

$$E_1 = -\frac{\hbar^2 k^2}{2m_v^*}, \tag{6.20}$$

where m_v^* is the mass of an electron in the valence band, i.e. $m_v^* = -m_h^*$ such that $m_v^* > 0$.

Hence the energy of the transition is written in terms of k as,

$$\hbar\omega = E_2 - E_1 = E_g + \frac{\hbar^2 k^2}{2}\left[\frac{1}{m_c^*} + \frac{1}{m_v^*}\right] = E_g + \frac{\hbar^2 k^2}{2\mu}, \tag{6.21}$$

where the **reduced mass** is given by,

$$\frac{1}{\mu} = \frac{1}{m_c^*} + \frac{1}{m_v^*}. \tag{6.22}$$

Now that we have established a relationship between the energy of the transition and the wavevector of the electrons involved in the transition we may now use eqn (6.21) to re-write the density of states in terms of the angular frequency of the transition:

$$g(k)\mathrm{d}k = \frac{k^2}{\pi^2}\mathrm{d}k = g_j(\omega)\mathrm{d}\omega \tag{6.23}$$

$$\Rightarrow g_j(\omega)d\omega = \frac{\hbar}{2\pi^2}\left(\frac{2\mu}{\hbar^2}\right)^{3/2}(\hbar\omega - E_g)^{1/2}d\omega, \tag{6.24}$$

where $g_j(\omega)$ gives the density of transitions of angular frequency ω, and is known as the **joint density of states**.

6.2.2 Density of possible transitions

The joint density of states gives the density of transitions of angular frequency ω, but *not all of these transitions can occur* since that requires that the initial state is occupied by an electron and the final state is empty.

For example, the density of possible stimulated emission transitions with frequencies in the interval ω to $\omega + \mathrm{d}\omega$ is given by,

$$g_j(\omega) f_c(E_2)[1 - f_v(E_1)]\mathrm{d}\omega,$$

i.e. the density of transitions in this frequency interval multiplied by the probability that the upper level is occupied and the lower level is unoccupied. Similarly the density of possible absorption transitions with frequencies in the interval ω to $\omega + \mathrm{d}\omega$ is given by,

$g_{\rm j}(\omega)f_{\rm v}(E_1)[1-f_{\rm c}(E_2)]{\rm d}\omega.$

The density of net stimulated transitions with frequencies in the interval ω to $\omega+{\rm d}\omega$ is given by the difference of these two expressions:

$$g(\omega)[f_{\rm c}(E_2)-f_{\rm v}(E_1)]{\rm d}\omega. \tag{6.25}$$

6.2.3 The gain coefficient

We are now in a position to write down the net rate of stimulated emission per unit volume with frequencies between ω and $\omega+d\omega$. From eqns (6.11) and (6.25) we have:

$$R_{SE}=W_{21}g_j(\omega)d\omega[f_{\rm c}(E_2)-f_{\rm v}(E_1)]. \tag{6.26}$$

Note that for atoms we would write the above in terms of the population inversion density N^* and optical gain cross-section $\sigma_{21}(\omega)$ as,

$$R_{SE}=N^*\sigma_{21}(\omega)\frac{I}{\hbar\omega}=\alpha(\omega)\frac{I}{\hbar\omega}, \tag{6.27}$$

where $\alpha(\omega)$ is the gain coefficient of the transition. Hence we deduce that the gain coefficient is given by $\alpha(\omega)=\hbar\omega R_{SE}/I$. From eqns (6.26) and (6.14) we find, after tidying and eliminating the intensity using $I=(1/2)\epsilon_0{\rm c}nE_0^2$, where n is the refractive index of the medium,

$$\alpha(\omega)=\frac{\omega\left|D_{21}\right|^2}{n\epsilon_0{\rm c}\hbar^3}(2\mu)^{3/2}(\hbar\omega-E_{\rm g})^{1/2}[f_{\rm c}(E_2)-f_{\rm v}(E_1)]. \tag{6.28}$$

6.2.4 Condition for gain

Notice that the term $[f_{\rm c}(E_2)-f_{\rm v}(E_1)]$ plays the role of the population inversion density which appears in the expression for the gain coefficient in atomic or molecular laser transitions. Clearly, in order for there to be positive gain we must have,

$$f_{\rm c}(E_2) > f_{\rm v}(E_1) \tag{6.29}$$

$$\Rightarrow \frac{1}{\exp([E_2-E_{\rm Fc}])/k_{\rm B}T+1} > \frac{1}{\exp([E_1-E_{\rm Fv}])/k_{\rm B}T+1} \tag{6.30}$$

$$\Rightarrow E_2-E_1 < E_{\rm Fc}-E_{\rm Fv}. \tag{6.31}$$

The left-hand side is simply the photon energy. The photon energy must also exceed the gap energy, and hence we can write the condition for gain as,

$$E_{\rm g}<\hbar\omega<E_{\rm Fc}-E_{\rm Fv}. \tag{6.32}$$

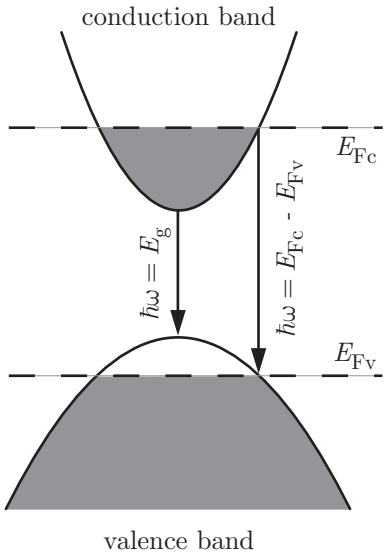


Figure 6.5: Illustrating the condition for gain for the case when $T = 0$.

In order for this condition to be met at any frequency, we must therefore have,

$$E_{\mathrm{Fc}} - E_{\mathrm{Fv}} > E_{\mathrm{g}} \quad \text{Condition for gain.} \tag{6.33}$$

This condition — originally derived in 1961 by Bernard and Duraffourg — holds for all temperatures T, but is particularly easy to understand for the case $T = 0$ as illustrated schematically in Figure 6.5. At absolute zero the gain corresponds to the upper level being occupied whilst the lower level is empty. At finite temperatures the condition for gain is that the probability of a level of energy E_2 being occupied is greater than the probability of a level of energy E_1 being occupied.

6.2.5 Generating gain: the injection density

Gain on optical transitions in semiconductor lasers is generated by raising N electrons per unit volume from the valence band to the conduction band. The density of electrons moved is known as the **injection density**. Since the density of electrons in the conduction band is equal to the density of holes in the valence band we may write,

$$N = \int_0^\infty g_{\mathrm{c}}(E_{\mathrm{c}}) f_{\mathrm{c}}(E_{\mathrm{c}}) dE_{\mathrm{c}} = \int_0^\infty g_{\mathrm{v}}(E_{\mathrm{v}})[1 - f_{\mathrm{v}}(E_{\mathrm{v}})]\mathrm{d}E_{\mathrm{v}}, \tag{6.34}$$

where $g_{\mathrm{c}}(E)$ and $g_{\mathrm{v}}(E)$ are the density of states of energy E in the conduction band and valence band respectively. The density of states may be written in terms of energy by using eqns (6.18), (6.19) and (6.20).

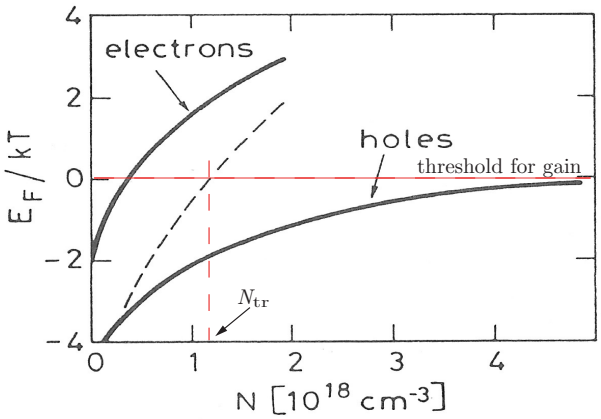


Figure 6.6: Calculation of the variation of the quasi-Fermi levels in the conduction and valence bands of GaAs as a function of the injected electron density N. For the conduction band the energy of the quasi-Fermi level above the bottom of the band is plotted; for the valence band the energy of the holes (which have minus the energy of the corresponding electron state) is plotted. The transparency condition therefore corresponds to the sum of these terms equalling zero, which occurs at an injection density $N_{\rm tr}$ as shown.

The relations above then determine the positions of the quasi Fermi levels in the conduction and valence bands, and hence the frequency width over which gain can be achieved.

For a given density N, the transcendental eqns (6.34) may be solved to find the quasi-Fermi levels. Figure 6.6 shows the result of one such calculation for GaAs. Notice that for low injection densities the quasi-Fermi levels lie outside their respective bands. As the injection density is increased the Fermi level in the conduction band rises more quickly than the Fermi level in the valence band drops. Hence the gain becomes positive.

Once the quasi-Fermi levels have been calculated as a function of the injection density, the gain coefficient may be found from eqn (6.28). Figure 6.7 shows the results of such a calculation. Two points are worth noting. The gain coefficient is typically large, some tens of cm^{-1} and hence laser oscillation is possible with very short gain lengths. Second, the bandwidth of the gain is broad and increases as the injection density of electrons is increased. For the example shown, the gain bandwidth is of order 10 THz, corresponding to a 30 nm bandwidth at the operating wavelength of 840 nm.

The gain first becomes positive when eqn (6.33) is met, corresponding to the injection density exceeding a threshold density, known as the **transparency density**, $N_{\rm tr}$. The condition for gain can therefore be written as $N > N_{\rm tr}$, which is analagous to $N^* > 0$ for gain on an atomic transition.

6.3 Homojunction diode lasers

In a homojunction laser, gain is achieved in an intrinsic region of semiconductor located between layers of degenerate p-type and n-type semiconductor, a so-called **p-i-n** junction.

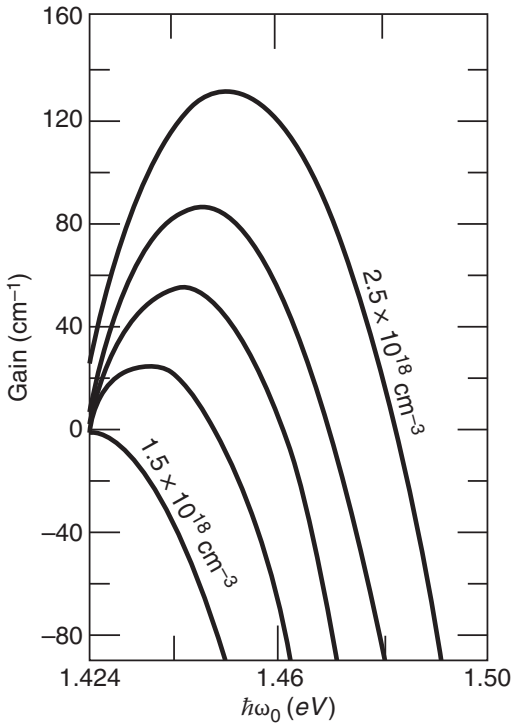


Figure 6.7: Calculated gain coefficient as a function of photon energy for various values of N (in units of 10^{18} cm^{-3}) for GaAs at $T = 300k$.

Figure 6.8(a) shows schematically the positions of the conduction bands and Fermi levels at the (rather theoretical!) moment when the three layers are first contacted. Electrons from the n-type material can drift across the intrinsic layer and recombine with holes in the p-type layer; in a similar way, holes drifting across the intrinsic layer can recombine with electrons in the n-type layer. This process sets up an electric field from the n-type to the p-type layer, and equilibrium is reached when the rate of drift of electrons across the intrinsic layer is balanced by the rate at which they are returned by the electric field. The region in which the electric field exists is virtually devoid of charge carriers since they are quickly removed by the field; it is known as the **depletion layer**.

The electric field raises the potential energy of electrons in the p-type layer, and lowers it in the n-type layer, to give the equilibrium distribution shown in 6.8(b). Notice that under these equilibrium conditions the Fermi energy lies at the same level in all three layers.

If a forward-bias voltage is now applied to the structure, the electric field across the device will be reduced, leading to a distribution of energy levels as shown in 6.8(c). Now there exists a region in the intrinsic layer in which there are electrons in the conduction band and holes in the valence band; gain is possible provided the quasi-Fermi levels obey eqn (6.33), or, equivalently, the density of electrons in the conduction band exceeds the transparency density.

To relate the density of electrons in the conduction band to the diode current, we note that most of the current is carried by the electrons (owing to their greater mobility). If the time for electron-hole re-

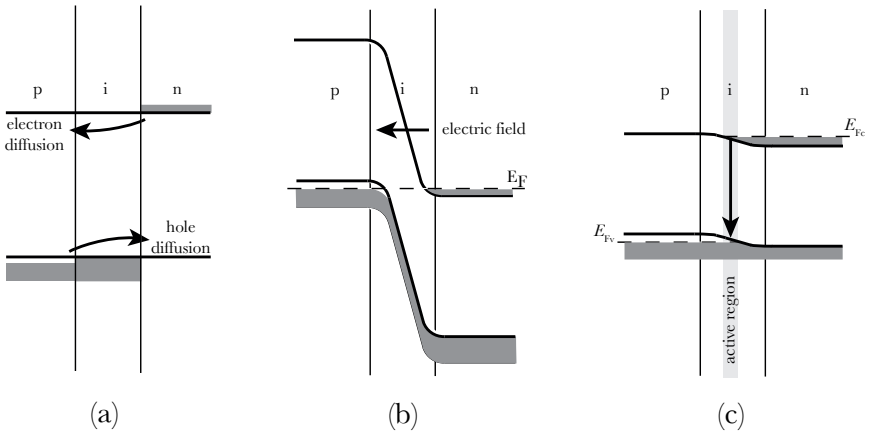


Figure 6.8: Energy levels and Fermi levels in a pin diode: (a) at the (theoretical) moment when the three layers are brought into contact; (b) after the depletion layer has been established; (c) under the influence of a forward-bias voltage.

combination in the intrinsic layer is $\tau_{\rm rec}$, then in the steady-state, and in the absence of stimulated emission, the rate of flow of electrons into the intrinsic layer must be balanced by the rate of recombination. Hence,

$$\eta_{\rm i}\frac{I}{{\rm e}V_{\rm g}} = \frac{N}{\tau_{\rm rec}}, \tag{6.35}$$

where $V_{\rm g}$ is the volume of the active region, and the **internal quantum efficiency** $\eta_{\rm i}$ gives the fraction of carriers that recombine radiatively in the active layer — other carriers undergo non-radiative recombination at the junction boundary, or flow round the active region. Hence the threshold current may be written as,

$$I_{\rm thresh} = \frac{1}{\eta}\frac{N_{\rm tr}{\rm e}V_{\rm g}}{\tau_{\rm rec}}. \tag{6.36}$$

We see that the threshold current increases with N_{tr} and the volume of the active region.

6.3.1 Structure of homojunction diode lasers

Figure 6.9 shows the structure of a homojunction diode laser. Notice that the gain length is very short, only 200 - 500 μm, and perhaps 100 - 200 μm wide. Optical feedback is usually provided by the uncoated surfaces of the cleaved crystal planes at each end of the device. The high refractive index of GaAs ($n = 3.6$) ensures a Fresnel reflectivity of approximately 32 % which is sufficient for lasing to be achieved.

One of the limitations of homojunction lasers is that the potential barrier preventing electrons from entering the p-type material is very low when the forward-bias voltage is applied. Consequently

tb]

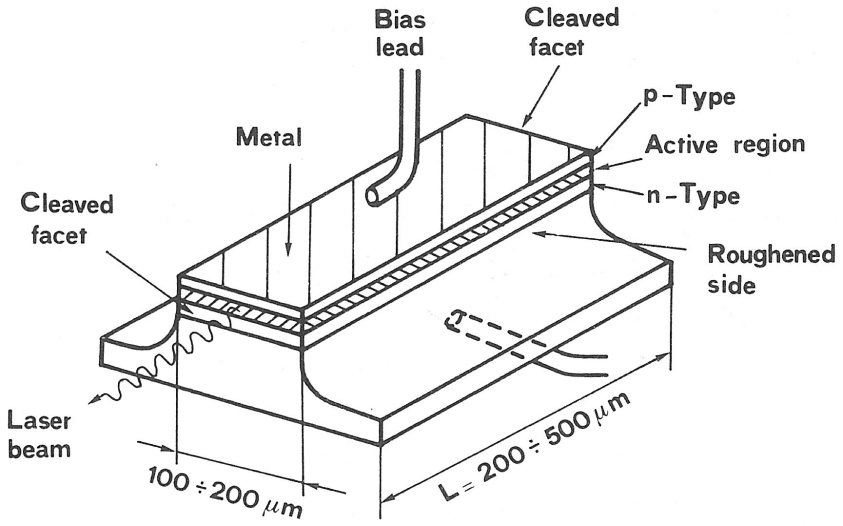


Figure 6.9: Physical structure of a homojunction diode laser.

electrons tend to diffuse into the p-type material, increasing the thickness of the active region beyond that of the depletion layer. In practice diffusion increases the thickness to around 1 μm, much thicker than the thickness of the depletion layer. This increases the threshold current to tens of amps, corresponding to a current density of order 10^5 A cm^{-2}. This high value means that homojunction diode lasers cannot be operated continuously without damaging the structure unless the whole device is cooled to cryogenic temperatures. For this reason, homojunction diode lasers are of historical interest only and all modern diode lasers use heterostructures.

6.4 Double heterostructure lasers

All diode lasers currently used employ a **double heterostructure**, rather than the simple p-i-n geometry discussed above. This structure can reduce the threshold current density by two orders of magnitude (to approximately 10^3 A cm^{-2}) owing to:

- **Photon confinement:** The active layer is surrounded by layers of a higher refractive index. As such the structure acts as a waveguide, and prevents the laser radiation from extending into regions of high loss.
- **Carrier confinement:** The surrounding layers also have large band gaps which confines the current carriers to the active region. The resulting increase in carrier density increases the gain.
- **Reduced absorption:** The large energy gap of the surrounding layers ensures that any laser radiation which does extend beyond the active region experiences much lower absorption than

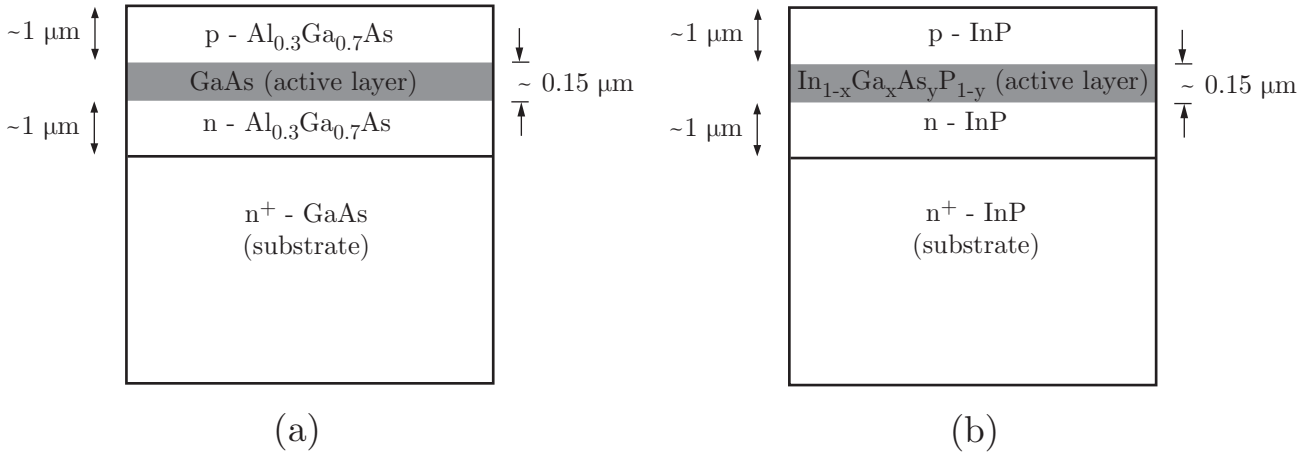


Figure 6.10: Schematic diagram of double heterostructure diode lasers made from: (a) AlGaAs; (b) InGaAsP.

it would if the band gap were the same as that in the active layer.

Figure 6.10 shows schematically the structure of heterostructure diode lasers made from (a) AlGaAs and (b) InGaAsP. In each case the active layer (≈ 0.15 μm) is surrounded by thicker layers of n- and p-doped material with a higher band gap. Figure 6.11 shows the variation of the refractive index, laser intensity, and band gap along a line perpendicular to the p-i-n junction under conditions of forward bias. It is clear that the higher band gap of the surrounding material prevents current carriers from moving beyond the active layer. Further, the diode laser radiation is surrounded by layers of higher refractive index and hence the mode is guided.

6.4.1 Forming heterostructures: Material properties

The selection of materials from which to make the cladding layers is a crucial one; not only must their band gap be larger than that of the intrinsic layer, but they must also be capable of being deposited via epitaxial growth techniques on the active region to form crystals free from defects without grain boundaries or regions of dislocation. This means that the alloy used for cladding must have a lattice constant very closely matched to that of the intrinsic material.

As shown in Figure 6.12, if we start with pure GaAs, by adding increasing atomic proportions x of aluminium to form the ternary alloy $Ga_{1-x}Al_xAs$ the lattice spacing is virtually the same whatever the proportion of added Al, even to the extent of replacing all of the Ga by Al to form AlAs since the lattice spacing of GaAs and AlAs are almost identical: 0.564 and 0.566 nm respectively.

This is however a happy circumstance unique to the GaAs system. In general **quarternary compounds** must be employed. For example, in $In_{1-x}Ga_xAs_yP_{1-y}$ the lattice spacing can be matched to the III-V semiconductor InP if $y = 2.2x$. With this ratio fixed, adjusting x then varies the band gap energy, allowing the output of the diode laser to be adjusted between 1150 nm and 1670 nm. This range is particularly useful since it covers the second (1300 nm) and third (1550 nm) transparency windows of silica fibres. InGaAsP diode lasers therefore find widespread application in telecommunications.

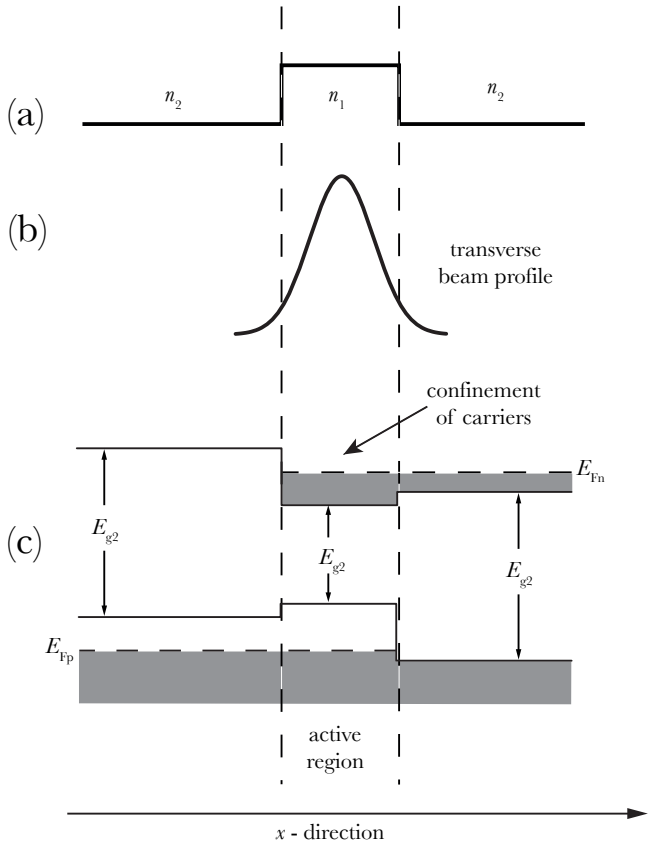


Figure 6.11: Schematic diagram of the variation of: (a) the refractive index; (b) the laser intensity; and (c) the band gap along the x-axis of a double heterostructure diode laser.

Table 6.1 lists the properties of some common types of diode laser.

6.5 Geometries of heterostructure diode lasers

Figure 6.13 shows the design of two types of heterostructure diode lasers. In both designs an oxide layer is added immediately under one of the electrode layers to restrict the extent of the current flow, and hence the gain region, in the y-direction. In the x-direction the region of gain is limited to the thickness of the active layer. By controlling the volume of the gain region in this way, the threshold current for achieving laser oscillation is reduced; although, of course, the laser power that can be extracted is also reduced.

For the design shown in Figure 6.13(a), the radiation mode is confined in the x-direction by the lower refractive index cladding layer. In the y-direction the extent of the beam is determined by the width of the gain region, and hence is said to be **gain-guided**. However, the confinement is weak in this direction, and the oscillating mode extends into regions which are not pumped, and so experiences

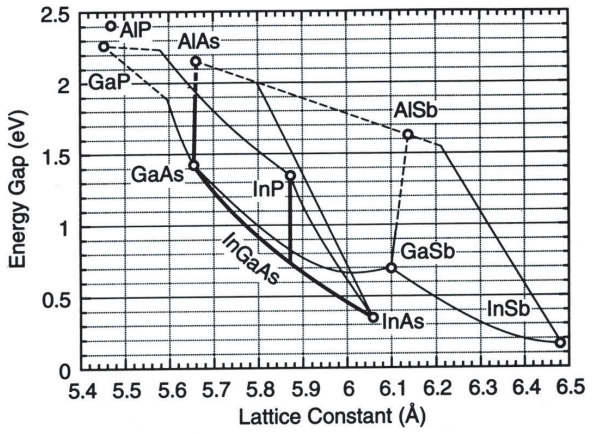


Figure 6.12: The band gap energy versus lattice spacing for some semiconductor compounds.

Active region	Material of cladding layers	Wavelength range	Substrate material
GaAs	(AlGa)As	800 - 900 nm	GaAs
GaInP	(AlGa)InP	630 - 650 nm	GaAs
InGaAs	GaAs	900 - 1000 nm	GaAs
InGaAs	InGaAsP	1550 nm	InP
InGaAsP	InGaAsP	1300 - 1550 nm	InP
InAsP	InGaAsP	1060 -1400 nm	InP
InGaAs	InP	1550 nm	InP
InGaN	AlGaN, GaN	370- 460nm	Al_2O_3,GaN

Table 6.1: Diode laser materials and wavelengths of operation.

high loss. This problem is overcome in the **buried heterostructure** shown in Figure 6.13(b). Here, the active region is also surrounded in the y-direction by cladding, or 'burying', layers which confine the radiation in that direction as well.

6.6 Output characteristics of diode lasers

Diode lasers have very high **slope efficiencies** $dP_{\text{out}}/dP_{\text{in}} = dP_{\text{out}}/(VdI)$, where V is the diode voltage. This can reach approximately 50% may be obtained in commercial devices.

The output power of a single double heterostructure diode laser is limited to powers less than a few hundred mW by over-heating of the diode structure. Much large powers can be generated by manufacturing linear arrays of diode lasers, known as **diode bars**. In turn, diode bars can be stacked to form diode arrays. A wide variety of geometries are available. Diode arrays can give output powers of 1 kW or even higher, and are increasingly used as replacements in flashlamps in solid state lasers such as Nd:YAG, and as the pump laser in fibre lasers.

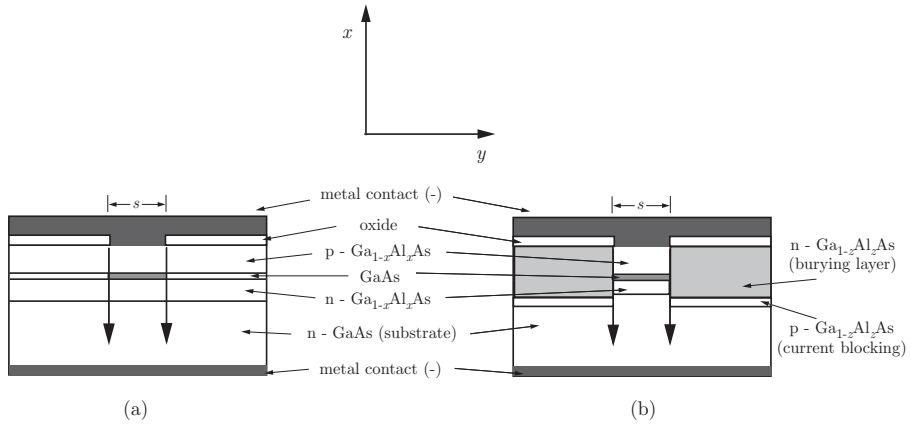


Figure 6.13: Geometries of heterostructure diode lasers: (a) gain-guided laser; (b) buried heterostructure, index-guided laser.

The frequency spectrum of diode lasers often depends on the driving current. Near threshold the laser mode spectrum has the wide bandwidth characteristic of an LED, but as the excitation current is increased a dominant axial mode establishes itself. This competes more and more effectively for the available inversion as the current increases so that the bandwidth spanned by the oscillating modes decreases as the current increases. Diode lasers can be operated in a single longitudinal mode by providing frequency-selective coupling. For example, a **fibre Bragg grating (FBG)** can be fixed to the laser output; in a FBG the refractive index of the fibre core varies periodically, and hence only those wavelengths satisfying the Bragg condition for reflection are efficiently coupled back into the gain region. Alternatively, wavelength-selective coupling can be provided by modulating the thickness of the layers surrounding the diode laser gain region with a "washboard" pattern; for the selected wavelength this provides continuous coupling between the left- and right-going waves in the gain region.

The biggest disadvantage of diode lasers is that they have a high divergence, which reflects the small dimensions of the gain region (typically $L_x = 1 \times L_y = 5$ μm). Further, the divergence is different in the two orthogonal directions ($\theta_x \approx 25^\circ \times \theta_y \approx 5^\circ$) which is often undesirable in applications. Techniques have been developed for making the beam more nearly circular, such as the use of prisms to expand the beam in one direction, or anamorphic lenses which have different curvatures in the x- and y-directions. Nevertheless the output beam quality of diode lasers remains poor in comparison to that from other types of laser.

Figure 6.14 shows the typical variation of output power with the diode current. Notice the low threshold current, a few milliamperes, and the sensitivity of the threshold current with the device temperature. This sensitivity reflects the strong dependence on the quasi-Fermi energies on temperature.

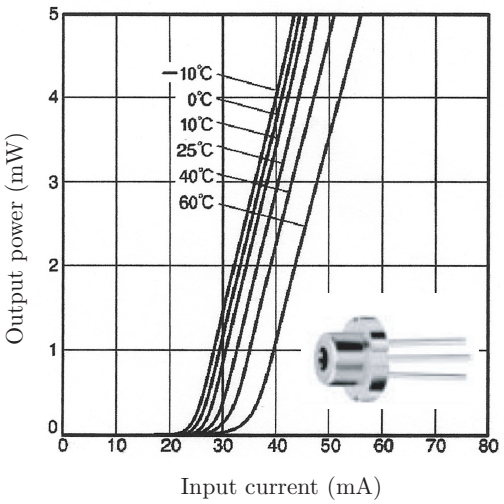


Figure 6.14: Output power of Rohm RLD-78MA double heterostructure diode laser versus current for various operating temperatures. Inset: photo of 5.6 mm diameter packaged diode laser. (From www.rohm.com)

6.7 Fibre lasers

Fibre lasers are particularly efficient at converting the often poor beam quality obtained from a diode laser into diffraction-limited laser radiation. In a fibre laser the active ion is doped into the central core of an optical fibre, the ions being pumped by radiation propagating along the fibre axis.

Fibre lasers have several advantages over the traditional rod or disc geometries:

- **Low pump power threshold:** Achievement of the threshold population inversion density requires the pump laser to exceed a threshold intensity. Since the cross-sectional area of an optical fibre is so low, the power required to reach threshold is orders of magnitude smaller than in a rod or disc where the area of the gain medium is at least an order of magnitude larger. In practice threshold pump powers significantly below 1 mW can be achieved. Using the fibre laser geometry it is possible to achieve lasing in lasers with: low gain coefficients; with high threshold pump intensities; and even in three-level laser systems.
- **Guiding of pump beam:** The pump laser radiation is naturally guided by the fibre, allowing very long gain regions to be employed.
- **High efficiency:** The good overlap between the pump and generated laser radiation along the whole length allows efficient conversion of the pump laser energy into laser output.
- **Good thermal management:** For any solid state laser system it is important to conduct away pump energy deposited into the gain medium in the form of heat, in order to avoid thermal lensing. The high surface area to volume ratio of the fibre geometry allows heat to be conducted away more effectively.

- **Compact gain medium:** Optical fibres may be coiled, keeping the volume of the laser system small.

- **Robustness:** Optical fibres are robust and much less liable to cracking or other physical damage.

The main disadvantages of fibre lasers are that the volume of the gain region is small. Consequently the energy stored in the gain region will also be small, and hence fibre lasers are not useful for amplifying laser pulses to high energy. Further, the long propagation length through the fibre core can be associated with the development of non-linear loss processes such as stimulated Raman scattering or stimulated Brillouin scattering.

6.7.1 Pumping configurations and fibre materials

In early fibre lasers both the pump radiation and the generated laser radiation were guided in the inner core of an optical fibre (Fig. 6.15(a)). As such the pump laser must be close to diffraction-limited for efficient coupling into the inner core of the fibre, which restricts the use of diode lasers to single-stripe devices.

Figure 6.15(b) shows a solution to this problem: **cladding pumping**. For this geometry the core of refractive index is surrounded by *two* further layers: an outer core of lower refractive index, and a cladding layer of still lower refractive index. The cladding layer will usually be protected by a final protective layer. The pump laser radiation is therefore guided in a multi-mode fashion by reflections at the boundary between the outer core and the cladding layer; with that part of the pump radiation passing through the core contributing to the pumping of the fibre laser. The laser radiation generated in the core can be guided mono- or multi-mode.

The use of cladding pumping allows much greater pump powers to be employed, increasing the output of the fibre laser from less than 1 W to greater than 100 W in a monomode beam and up to 1 kW in multi-mode fibre lasers.

6.7.2 An example

Figure 6.16 shows schematically the design of a high-power Nd:Glass fibre laser. For this system the doped core has a diameter of 10 μm, with a numerical aperture of 0.17. The outer core has a diameter of 400 μm and numerical aperture of 0.38. Pump radiation, provided by a diode array operating at 808 nm, is coupled into one end of the fibre through a mirror which has a high transmission at 808 nm, but is highly reflective at the Nd:Glass wavelength. The output coupling mirror is also highly reflective at the pump laser wavelength, but has a transmission of approximately 90% for the Nd:Glass laser radiation. With this design, output powers of 32 W have been achieved for a pump power of 72 W, corresponding to a conversion efficiency of 44%. The output is in the form of a single transverse mode, since although the core supports several low-order modes, the lowest mode has the lowest losses and hence grows at the expense of higher-order transverse modes. Laser oscillation occurs at approximately 1065 nm, with a large spectral width of 15 nm reflecting the large inhomogeneous width introduced by the glass host.

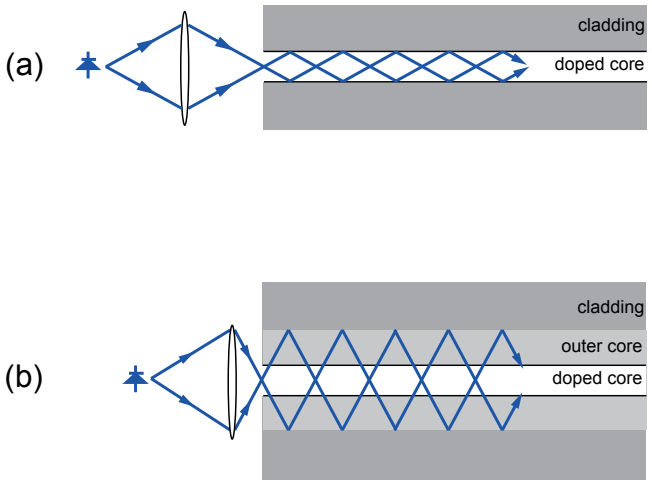


Figure 6.15: Fibre laser pumping configurations: (a) pumping of a conventional fibre; (b) cladding pumping.

6.7.3 Erbium-doped fibre lasers

Silica optical fibres have transmission windows at 1.3 μm and 1.5 μm. Single-mode fibres used at 1.3 μm have low dispersion and so the distance that signals can propagate before they must be amplified or regenerated is determined primarily by the fibre loss of 0.4 dB km^{-1}. Amplifiers are not available at 1.3 μm, and so signals at this wavelength are regenerated in a regenerative repeater, which: (i) detects and interprets the signal; (ii) corrects errors; (iii) regenerates and retransmits the bitstream. The optimum repeater separation is typically 40 km.

At a wavelength of 1.5 μm conventional fibre has a lower loss (approximately 0.3 dB km^{-1}), but higher dispersion than at 1.3 μm. However in dispersion-shifted fibre, in which material and waveguide dispersion are balanced, the dispersion at 1.5 μm can be reduced to 0.1 ps km^{-1} nm^{-1} compared to

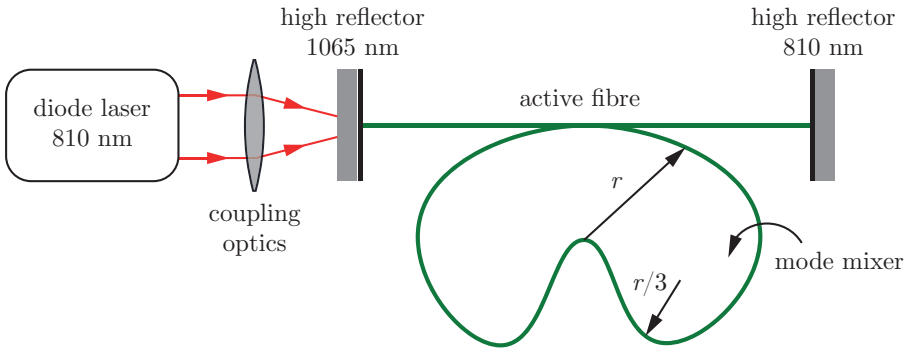


Figure 6.16: Schematic diagram of a high-power Nd:Glass fibre laser.

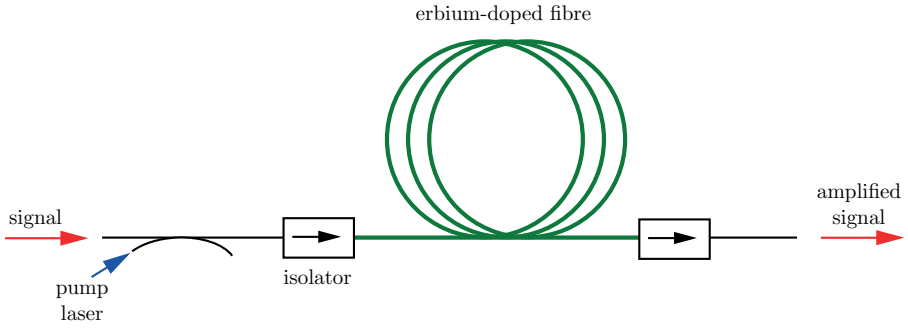


Figure 6.17: Schematic diagram of an EDFA.

17 ps km^{-1} nm^{-1} for conventional fibre at 1.3 μm.[5] Hence dispersion-shifted fibre can transmit signals some two orders of magnitude further — corresponding to distances of tens of thousands of kilometres — before regeneration of the signal is required. Of course, *amplification* of the signal is required more often to overcome the fibre losses; typically amplification is required every few tens of kilometres.

Amplification can be provided by an **erbium-doped fibre amplifier (EDFA)**. The EDFA operates on the $^4I_{13/2} \rightarrow ^4 I_{15/2}$ transition at approximately 1550 nm. As discussed previously, pump radiation can conveniently be provided by diode lasers operating at 800, 980, or 1480 nm.

Figure 6.17 shows schematically the configuration of an EDFA. The pump laser radiation is coupled into the amplifier by splicing together a fibre carrying the pump radiation and a fibre coaxial with the amplifier. The amplifying section is a monomode silica fibre doped with around 0.1% erbium, and with a length ranging from metres to hundreds of metres. The single-pass gain is typically of order 10^3. Notice that optical isolators — based on the Faraday effect — are inserted around the EDFA to prevent parasitic laser oscillation arising from the small Fresnel reflections at junctions in the fibre.

EDFAs have been used in the latest 5 Gbit s^{-1} transatlantic fibres which employ amplifiers every 35 km, but only regenerate the signal at the two land stations!

Finally we note that the very wide bandwidth (some 30 nm or 3.74 THz) of the erbium transition allows amplification of multiple channels each transmitted at a slightly different wavelength. This technique, **wavelength division multiplexing (WDM)**, could allow, for example, 375 channels of 2.5 Gbit s^{-1} and spaced by 10 GHz to be transmitted within the bandwidth of an erbium amplifier. The implementation of WDM, however, requires the development of suitable laser sources and techniques for de-multiplexing the transmitted signals; a great deal of work is being done in this area.

[5] In any optical fibre the dispersion experienced by *each mode* — the intra-modal dispersion — has two contributions: dispersion arising from the dispersion of the materials employed (material dispersion); and dispersion arising from the frequency dependence of the wave vector of each mode (waveguide dispersion), which depends on the geometry of the fibre. In dispersion-shifted fibres the dimensions of the fibre are adjusted so that the total dispersion is low at the desired operating wavelength. In addition, for any fibre different modes have different phase velocities leading to so-called inter-modal dispersion; this can be avoided if only one mode is excited.

Lecture 7

Precision frequency control

Here we discuss some of the issues affecting the frequency linewidth of an oscillating laser. We will often give results in terms of the frequency ν (units of Hertz) rather than the angular frequency ω (units of radians per second) since this allows easier comparison with real laser systems.

7.1 Single-mode operation

In principle a laser operating on a homogeneously broadened transition should oscillate on only a single longitudinal mode, but in practice spatial hole-burning means that oscillation can occur on more than one mode. Lasers operating on homogeously broadened transitions will typically operate on many longitudinal modes. In addition, laser oscillation can occur on higher-order transverse modes, which in general have slightly different frequencies.

It is clear that for a laser to achieve the narrowest possible linewidth it should operate on a single longitudinal and transverse mode. Ensuring the latter is straightforward: by introducing a suitable aperture it is possible to ensure that higher-order transverse modes experience sufficiently high losses that they cannot reach the threshold for oscillation.[1]

Several techniques can be used to force a laser to operate on only a single longitudinal mode.

7.1.1 Short cavity

The simplest approach is to ensure that the longitudinal mode spacing $\Delta\nu_{p,p-1} \approx c/2L_\mathrm{c}$ is larger than the range of frequencies, $\Delta\nu_\mathrm{osc}$, for which the laser gain exceeds the threshold value. Assuming that one of the longitudinal modes coincides with the centre of the laser transition, it is easy to show that single-mode operation will occur if,

$$L_\mathrm{c} < \frac{c}{\Delta\nu_\mathrm{osc}}. \tag{7.1}$$

[1] See Fig. 8.7 of "Modern Classical Optics" by Geoff Brooker, also reproduced as Fig. 6.12 in H&W.

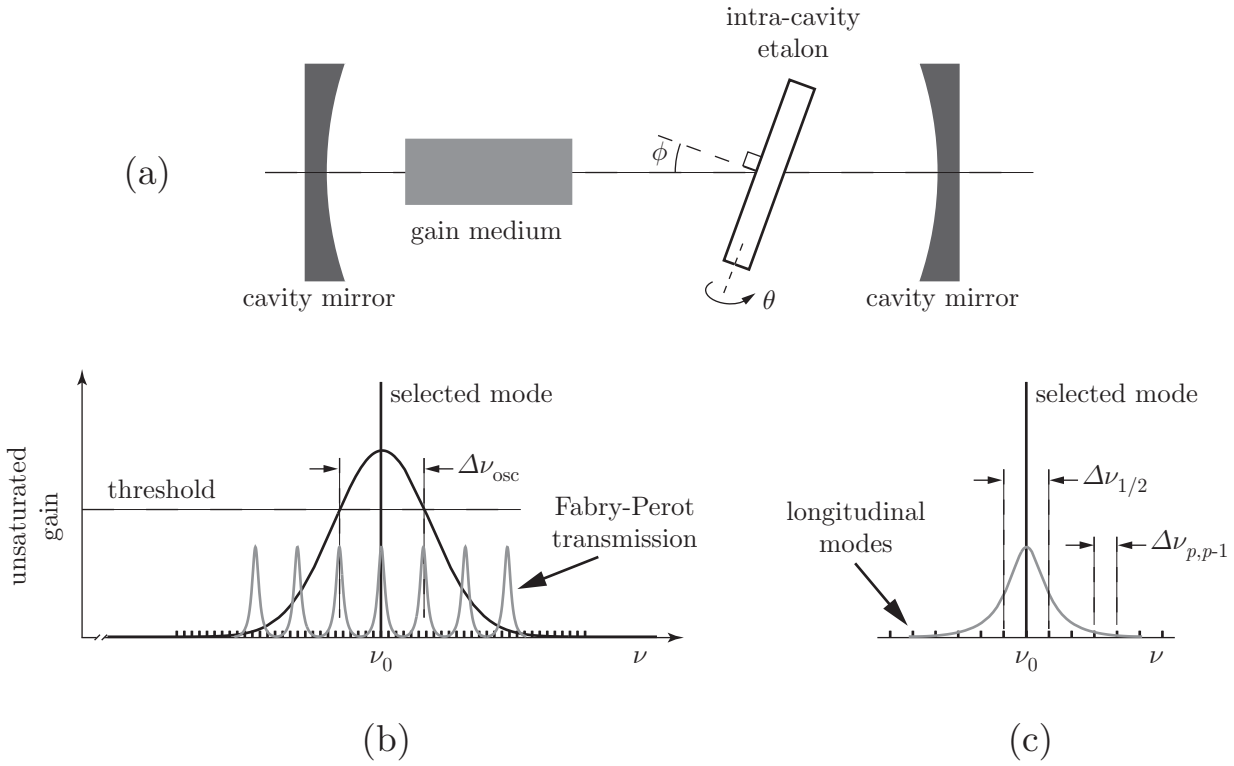


Figure 7.1: SLM operation by using an intra-cavity etalon. (b) Relation between the peaks of the etalon transmission and the laser lineshape when Condition (2) is satisfied. (c) Relation between the FWHM of the etalon and the longitudinal mode spacing when Condition (1) is satisfied.

7.1.2 Intra-cavity etalon

An intra-cavity etalon may be used to increase the losses experienced by longitudinal modes adjacent to the mode nearest line centre. The etalon must satisfy two conditions:

1. The width of the etalon peaks must be sufficiently narrow to discriminate between adjacent longitudinal modes.

2. The free-spectral range of the etalon must be sufficiently large that adjacent etalon peaks lie outside the range of frequencies for which laser oscillation is possible.

Figure 7.1 shows an idealized arrangement in which the orientation of the etalon has been rotated so that one of its transmission peaks coincides with the longitudinal mode nearest the centre of the laser line.

We may express conditions (1) and (2) as:

$$\frac{\Delta\nu_{1/2}}{2} < \Delta\nu_{p,p-1} \tag{7.2}$$

$$\Delta\nu_{\mathrm{FSR}} > \frac{\Delta\nu_{\mathrm{osc}}}{2}, \tag{7.3}$$

where $\Delta\nu_{1/2} = \Delta\nu_{\mathrm{FSR}}/\mathcal{F}$ is the FWHM of one of the etalon peaks. Using the standard relations for a Fabry-Perot etalon it is easy to show that the thickness d of the etalon must satisfy,

$$\frac{\Delta\nu_{\mathrm{osc}}}{2} < \frac{c}{2nd} < \mathcal{F}\frac{c}{L_{\mathrm{c}}}, \tag{7.4}$$

where n is the refractive index of the material between the etalon plates (it is common to use a solid etalon in this application). Hence we find that,

$$L_{\mathrm{c}} < (2\mathcal{F})\,\frac{c}{\Delta\nu_{\mathrm{osc}}}. \tag{7.5}$$

Thus the use of an intra-cavity etalon allows SLM operation to be achieved in a cavity with a length increased by a factor $2\mathcal{F}$ than with no etalon. It may be shown that for N etalons the maximum length of the cavity can be increased by a factor of $(2\mathcal{F}_1)\,(2\mathcal{F}_2)\,(2\mathcal{F}_3)\ldots(2\mathcal{F}_N)$ above that required for no etalon.

7.1.3 Ring resonators

SLM operation on *homogeneously* broadened laser transitions can be achieved by employing a ring resonator and ensuring that oscillation only occurs in one direction around the ring; this prevents spatial hole-burning. Uni-directional oscillation can be achieved by inserting a Faraday isolator within the ring as shown in Fig. 7.2.

7.2 Output linewidth

7.2.1 Theoretical limit

Suppose that SLM oscillation is achieved. What would be the linewidth of the laser output?

Since the laser cavity can be considered to be a Fabry-Perot etalon, we might expect that the linewidth would be given by the FWHM of an etalon peak, i.e. $\Delta\nu_{1/2} = \Delta\nu_{\mathrm{FSR}}/\mathcal{F}$. However, this is *not* the case since the finite linewidth of the etalon arises from the fact that the amplitude of a wave launched within the etalon decays as it bounces between the mirrors; as such the wave is only 'stored' within the cavity/etalon for a time of order the cavity lifetime τ_{c}, leading to a linewidth of order $1/\tau_{\mathrm{c}}$. For an oscillating laser, however, the losses are balanced by the gain so, in effect, the cavity lifetime becomes infinite. This would suggest that the laser linewidth should be zero.

In fact the fundamental limit to the laser linewidth arises from the role of spontaneous emission within the laser cavity. As we will see, the effect of spontaneous emission is a slow drift of the phase of the laser output, which leads to a finite frequency width.

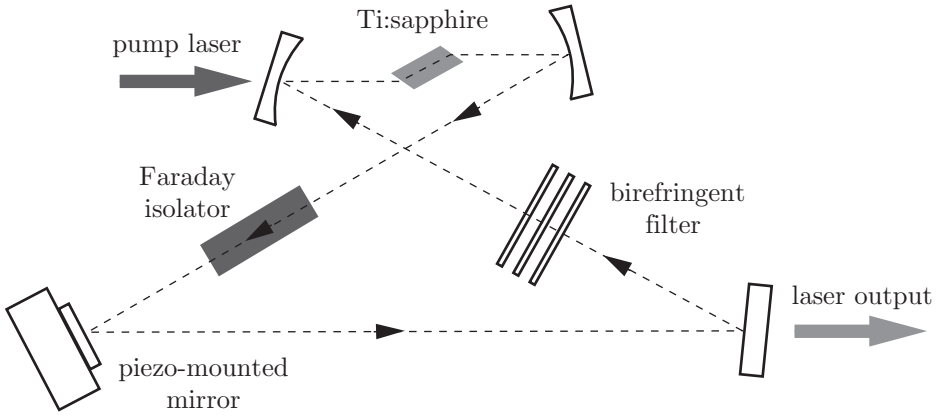


Figure 7.2: Single longitudinal mode tunable Ti:sapphire laser employing a uni-directional ring cavity. Wavelength tuning is achieved by rotating the birefringent filter. One of the cavity mirrors is mounted on a piezoelectric crystal so that the length of the cavity can be adjusted as the wavelength is tuned, ensuring that the same longitudinal mode oscillates as the wavelength is varied.

Supose that an oscillating mode within the cavity contains $\bar{n}$ photons. The energy of the mode is proportional to $\bar{n}$, or, equivalently, to E_0^2 where E_0 is the amplitude of the electric field of the mode. Hence we have the relation $E_0 \propto \sqrt{\bar{n}}$.

If the cavity lifetime is τ_c then photons are lost from the cavity at a rate $\bar{n}/\tau_c$; these form the output beam. Since photons produced by stimulated emission are perfectly coherent with each other, in the absence of spontaneous emission the amplitude and phase of the electric field at any point in the cavity would be constant and could be written in the form $E(t) = E_0 \exp(-\mathrm{i}[\omega_L t - \phi])$. The amplitude of this field could be represented by a constant vector in the complex plane, as shown in Fig. 7.3.

Occasionally a photon will be emitted into the cavity spontaneously. This will have a random phase relative to that of the stimulated photons, leading to a small shift in the phase (and also the amplitude, but this is small) of the oscillating mode. Spontaneous emission therefore causes the phase of the mode to change randomly by small amounts so that the phase of the mode diffuses over time with a characteristic time τ_{ST}, corresponding to a frequency width $\Delta\omega_{ST} = 1/\tau_{ST}$.

We now estimate τ_{ST} and hence the **Schawlow-Townes linewidth**. Since $E_0 \propto \sqrt{\bar{n}}$ we can write,

$$\frac{E_{sp}}{E_0} = \frac{\sqrt{1}}{\sqrt{\bar{n}}}. \tag{7.6}$$

If N photons are emitted spontaneously into the mode the change θ of the phase of the mode is

$$\theta = \frac{\Delta E_\theta}{E_0}, \tag{7.7}$$

where ΔE_θ is the is the component of the vector sum of the electric fields of the N photons along the $\hat{\boldsymbol{\theta}}$ direction.

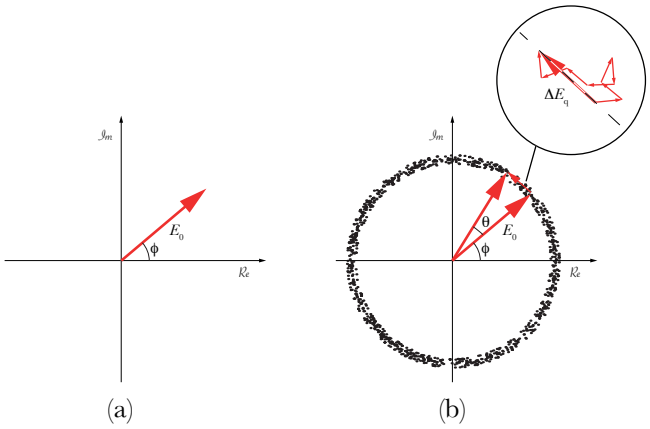


Figure 7.3: Origin of the Schawlow-Townes linewidth. (a) In the absence of spontaneous emission, if a laser oscillates on a single cavity mode the electric field may be written as $E(t) = E_0 \exp(-\mathrm{i}[\omega_\mathrm{L} t - \phi])$ where ϕ is an arbitrary constant phase. This may be represented by a constant vector in an Argand diagram, as shown, where the axes correspond to $\Re\left[E(t)\exp(\mathrm{i}\omega_\mathrm{L} t)\right]$ and $\Im\left[E(t)\exp(\mathrm{i}\omega_\mathrm{L} t)\right]$. The effect of spontaneous emission is shown in (b). After each spontaneously emitted photon is emitted into the oscillating mode, the tip of the electric field vector moves in a random direction by an amount E_sp. The dots shown in (b) show the positions occupied by the tip of the electric field vector after a very large number of photons have been emitted spontaneously. The inset in (b) shows the effect on the electric field vector of the emission of 8 spontaneously emitted photons, and in particular the net displacement, ΔE_θ, along a direction, $\hat{\boldsymbol{\theta}}$, orthogonal to the original electric field vector. Notice that the electric field vector of each spontaneously emitted photon has a different component along $\hat{\boldsymbol{\theta}}$, and consequently on average the displacement ΔE_θ is given by the result of a one-dimensional random walk (along $\hat{\boldsymbol{\theta}}$) of varying step size. The resultant displacement causes the orientation of the electric field vector to rotate through an angle θ, as shown.

The value of ΔE_θ depends on the phases of the spontaneously emitted photons. Since these are uncorrelated, we may use the results of the theory of a one-dimensional random walk of variable step size. These state that for N steps in the random walk — i.e. after the emission of N spontaneous photons — the mean amplitude and mean square amplitude along $\hat{\boldsymbol{\theta}}$ are given by:

$$\overline{\Delta E_\theta} = N\overline{E_{\theta\mathrm{sp}}} \tag{7.8}$$

$$\overline{\Delta E_\theta^2} = N\overline{E_{\theta\mathrm{sp}}^2}, \tag{7.9}$$

where $\overline{E_{\theta\mathrm{sp}}}$ and $\overline{E_{\theta\mathrm{sp}}^2}$ are the mean amplitude and mean square amplitude along $\hat{\boldsymbol{\theta}}$ of a single spontaneously emitted photon. Since for spontaneously emitted photons E_sp is the same, but their phase can take any value with equal probability, it is easy to show that

$$\overline{E_{\theta \mathrm{sp}}} = 0 \tag{7.10}$$

$$\overline{E_{\theta \mathrm{sp}}^2} = \frac{E_{\mathrm{sp}}^2}{2}, \tag{7.11}$$

and hence that after N photons have been emitted into the mode:

$$\overline{\theta} = 0 \tag{7.12}$$

$$\overline{\theta^2} = \frac{\overline{\Delta E_\theta^2}}{E_0^2} = \frac{N}{2}\left(\frac{E_{\mathrm{sp}}}{E_0}\right)^2. \tag{7.13}$$

If photons are spontaneously emitted into the mode at a rate R_{sp}, then after a time τ the root mean square value of θ will be

$$\theta_{\mathrm{rms}} = \sqrt{\overline{\theta^2}} = \sqrt{\frac{R_{\mathrm{sp}}\tau}{2\bar{n}}}, \tag{7.14}$$

where we have used eqn (7.6).

We can say that the uncertainty in the phase of the mode is significant once $\theta_{\mathrm{rms}} \approx \pi/2$, and hence we may write,

$$\sqrt{\frac{R_{\mathrm{sp}}\tau_{\mathrm{ST}}}{2\bar{n}}} \approx \frac{\pi}{2},$$

from which we deduce,

$$\Delta\omega_{\mathrm{ST}} = \frac{1}{\tau_{\mathrm{ST}}} \approx \left(\frac{2}{\pi}\right)^2 \frac{R_{\mathrm{sp}}}{2\bar{n}}. \tag{7.15}$$

To find the rate R_{sp} at which photons are emitted spontaneously into the mode we note that the rate of spontaneous emission into a mode is equal to $1/\bar{n}$ of the rate of stimulated emission from the upper laser level. Further, under steady-state conditions the rate of stimulated emission from the upper level must equal $\bar{n}/\tau_{\mathrm{c}}$, i.e. the rate at which photons leave the cavity. Hence $R_{\mathrm{sp}} = 1/\tau_c$ and so,

$$\Delta\omega_{\mathrm{ST}} \approx \left(\frac{2}{\pi}\right)^2 \frac{1}{2\bar{n}\tau_{\mathrm{c}}}. \tag{7.16}$$

It is convenient to write $\bar{n}$ in terms of the output power P of the laser. Since the cavity loses photons at a rate $\bar{n}/\tau_{\mathrm{c}}$, we may write $P = (\bar{n}/\tau_{\mathrm{c}})\hbar\omega_{\mathrm{L}}$ and hence,

$$\Delta\omega_{\rm ST} \approx \frac{2}{\pi^2}\frac{\hbar\omega_{\rm L}}{P}\frac{1}{\tau_{\rm c}^2} = 8\frac{\hbar\omega_{\rm L}}{P}\Delta\nu_{\rm c}^2 \tag{7.17}$$

$$\Rightarrow \Delta\nu_{\rm ST} \approx \frac{4}{\pi}\frac{h\nu_{\rm L}\Delta\nu_{\rm c}^2}{P}. \tag{7.18}$$

A more complete derivation shows that,

$$\Delta\nu_{\rm ST} = \left(\frac{N_2}{N^*}\right)_{\rm th}\frac{\pi h\nu_{\rm L}\Delta\nu_{\rm c}^2}{P}, \tag{7.19}$$

where $(N_2/N^*)_{\rm th}$ is the ratio of the population density in the upper laser level to the population inversion density under the operating conditions of the laser. Typically this ratio is close to unity.

7.2.2 Practical limitations

The Schawlow-Townes linewidth is usually extremely small (very often much less than 1 Hz!). In the absence of active frequency-locking techniques the linewidth of even a single-mode laser will be much greater than $\Delta\nu_{\rm ST}$.

A change in cavity length will change the frequency by an amount,

$$\frac{\Delta\nu_p}{\nu_p} = -\frac{\Delta L_{\rm c}}{L_{\rm c}}. \tag{7.20}$$

A wide variety of processes (vibration, thermal effects, changes in ambient pressure etc.) can cause the cavity length to change, and hence the output frequency to drift or change.

To give just one example, if the coefficient of thermal expansion of the material on which the cavity mirrors is mounted is $\alpha_{\rm T}$, then a change in temperature of ΔT gives,

$$\frac{\Delta\nu_p}{\nu_p} = -\alpha_{\rm T}\Delta T. \tag{7.21}$$

For invar (a steel alloy with very low coefficient of thermal expansion of $\alpha_{\rm T} \approx 10^{-6}$) then to keep $\Delta\nu_p$ below, say, 1 MHz requires temperature stabilization to better than 2 mK for a laser operating in the visible region of the spectrum.

7.3 Frequency locking

Under practical conditions in order to achieve the narrowest linewidths, and to prevent drift, it is necessary to lock the frequency of a laser to a stable reference frequency. In principle this is straightforward:

1. The frequency ν of the laser to be locked is compared with some fixed frequency $\nu_{\rm ref}$.

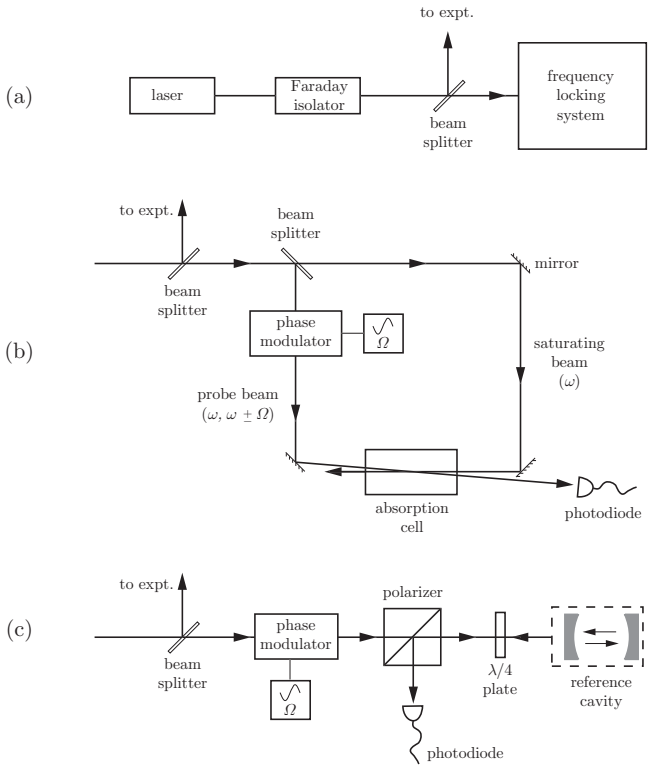


Figure 7.4: (a) Schematic diagram of the arrangement used to frequency lock a laser. (b) Frequency-locking to a Doppler-free absorption line. (c) Frequency-locking to a reference cavity.

2. Any difference in frequency is converted to an error signal, which increases with the magnitude of the frequency error and differs in sign according to the sign of $(\nu - \nu_{\rm ref})$.

3. The frequency of the oscillating mode is adjusted — usually by changing the length of the laser cavity — so as to minimize the magnitude of the error signal, thereby minimizing the frequency error.

This is usually achieved by picking off a small portion of the laser beam and modulating its frequency so that phase-sensitive detection (PSD) techniques can be employed. Figure 7.4 shows how this is achieved in the general case, and for the cases of locking to an absorption line or an external cavity.

Form of error signal

We now consider the form of the error signal for the general case of a frequency-modulated wave passing through (or reflecting from) a reference system with a frequency-dependent *amplitude* response $a(\omega)$ which is strongly peaked in the region of the reference frequency $\omega_{\rm ref}$. The amplitude of the wave leaving a frequency modulator with a time-dependent amplitude transmission $\exp(\mathrm{i}\beta \sin \Omega t)$ may be

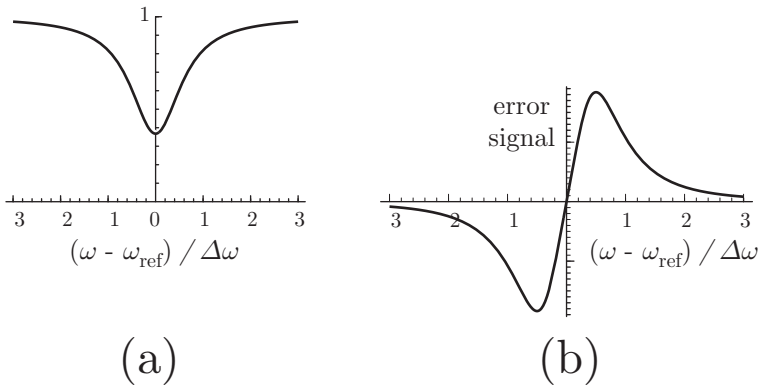


Figure 7.5: Frequency-locking to a Lorentzian response function centred at $\omega_{\rm ref}$ with a full-width at half maximum $\Delta\omega$. (a) the amplitude response function $a(\omega)$; (b) the amplitude of the error signal $S_\Omega(t)$ as a function of the output frequency ω of the laser.

written as,

$$E(t) = E_0 \exp(-\mathrm{i}\omega t) \exp(\mathrm{i}\phi_0) \sum_{m=-\infty}^{\infty} J_m(\beta) \exp(\mathrm{i}m\Omega t), \tag{7.22}$$

where β is proportional to the phase modulation introduced by the modulator. If the amplitudes of the sidebands other than $m = \pm 1$ may be ignored, we have

$$E(t) = E_0 \exp(\mathrm{i}\phi_0) \left\{ -J_1(\beta) \exp\left[-\mathrm{i}(\omega+\Omega)t\right] + J_0(\beta) \exp(-\mathrm{i}\omega t) + J_1(\beta) \exp\left[-\mathrm{i}(\omega-\Omega)t\right] \right\}, \tag{7.23}$$

where we have used the fact that $J_{-1}(\beta) = -J_1(\beta)$. After passing through the reference system the amplitude of the radiation is

$$\begin{aligned} E(t) = E_0 \exp(\mathrm{i}\phi_0) \{ & -J_1(\beta) a(\omega+\Omega) \exp\left[-\mathrm{i}(\omega+\Omega)t\right] + J_0(\beta) a(\omega) \exp(-\mathrm{i}\omega t) \\ & + J_1(\beta) a(\omega-\Omega) \exp\left[-\mathrm{i}(\omega-\Omega)t\right] \}, \end{aligned} \tag{7.24}$$

It may be shown that for a square-law detector, i.e. one which detects a signal $S(t) \propto E(t)E*(t)$, the signal at the modulation frequency Ω is given by,

$$S_\Omega(t) \propto \Omega \left|a(\omega)\right| \left.\frac{\partial \left|a(\omega)\right|}{\partial \omega}\right|_\omega \cos \Omega t. \tag{7.25}$$

As shown in Fig. 7.5 the **error signal** is proportional to the derivative of $|a(\omega)|$, and hence the sign of the error signal depends on whether ω is greater or less than the reference frequency $\omega_{\rm ref}$.

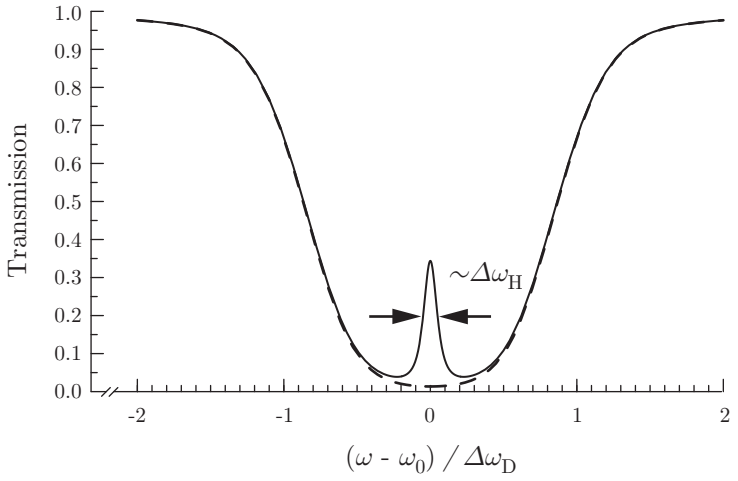


Figure 7.6: Calculated frequency dependence of the transmission of a weak probe through a sample of Doppler broadened atoms with (solid line) and without (dashed line) a counter-propagating saturating beam with the same frequency as that of the probe beam. Note that the main absorption has a frequency width of order the Doppler width $\Delta\omega_{\rm D}$, but the spike of reduced absorption has a width approximately equal to the homogeneous width of the atoms, $\Delta\omega_{\rm H}$. For the purposes of illustration, in this calculation $\Delta\omega_{\rm H}$ was set equal to $0.1\Delta\omega_{\rm D}$; in practice the ratio of the homogeneous and to inhomogeneous linewidths would usually be very much smaller.

7.3.1 Locking to atomic or molecular transitions

Locking to atomic or molecular transitions provides the best long-term stability and gives highly reproducible results: the frequency reproducibility is typically found to be 1 part in 10^{11}.

The most convenient transitions suitable as reference frequencies occur in atoms or molecules in the gaseous state, and consequently it is necessary to use **Doppler-free** techniques to eliminate Doppler broadening, as illustrated in Fig. 7.4(b). In this approach the laser beam is divided into a strong 'saturating' beam and a weak 'probe' beam. The saturating and probe beams pass through the absorption cell in an exact, or near, counter-propagating geometry, and the intensity of the transmitted probe beam is measured by a photodiode. This is the same arrangement used in saturation absorption spectroscopy.

In order to determine the form of the error signal we must find the frequency response $a(\omega)$ of the reference system, which in this case is the frequency dependence of the amplitude transmission of the probe beam. Now, *in the absence of the saturating beam* the transmitted amplitude will be large for frequencies ω such that $|\omega - \omega_0| \gg \Delta\omega_{\rm D}$, and it will be small for detunings $|\omega - \omega_0|$ comparable to or small compared to $\Delta\omega_{\rm D}$. This behaviour is shown in Fig. 7.6.

Since they propagate in opposite directions, the saturating beam and the probe beam usually interact with different velocity classes. As such, for frequencies such that $|\omega - \omega_0| \gg \Delta\omega_{\rm H}$ the absorption of the probe beam is unaffected by the presence of the saturating beam. However, for frequencies $\omega \approx \omega_0 \pm \Delta\omega_{\rm H}$, both beams interact with the same ($v_z \approx 0$) class of atoms and hence the absorption experienced by the probe beam is reduced. This leads to a very narrow feature in the transmission of

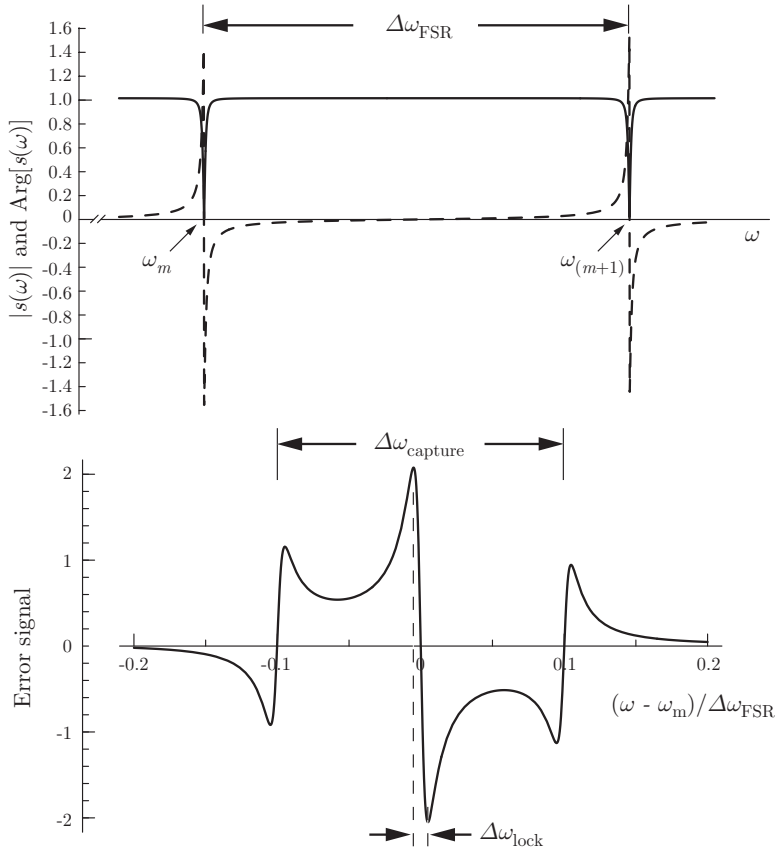


Figure 7.7: The Pound-Drever-Hall technique. (a) shows the amplitude (solid line) and phase (dashed line) of the beam *reflected* from a cavity as a function of frequency, where ω_m and $\omega_{(m+1)}$ are two adjacent resonant frequencies of the cavity. (b) shows the error signal as a function of frequency near one of the resonant frequencies. Notice that the 'capture' range of the error signal is a significant fraction of the free spectral range, $\Delta\omega_{\rm FSR}$, of the cavity, but that the error signal has a slope of very large magnitude near the resonant frequency.

the probe beam, as shown in Fig. 7.6 which can be used to lock the frequency to ω_0.

7.3.2 Locking to an external cavity

Locking to the resonant frequency of an external cavity provides better short-term stability than locking to an atomic transition since the error signals are large and hence have a high signal-to-noise ratio. There are many ways in which this can be done, but the most widely used is the **Pound-Drever-Hall technique** illustrated schematically in Fig. 7.4(c).

The Pound-Drever-Hall technique utilizes the frequency-dependent *reflection* from a high-finesse cavity.[2] The advantage of using the signal reflected from the cavity is that the error signal has a much

[2]In this application cavity finesses up to 2×10^5 may be obtained by using 'super-polished' mirrors.

wider 'capture' range than that which would be obtained by monitoring the transmitted beam.

To see this we note that the cavity acts as a Fabry-Perot etalon with a transmission which is sharply peaked. As shown in Fig. 7.5(b) when used in transmission the resulting error signal would only extend over a frequency range of order the cavity linewidth $\Delta\omega_c$, and this must be narrow for accurate locking. This behaviour may also be considered in the time domain; a narrow linewidth cavity corresponds to a long cavity lifetime τ_c and consequently the transmitted beam does not respond to fluctuations in the beam which occur in a time significantly shorter than τ_c, or, equivalently, to frequency shifts greater than of order $1/\tau_c \approx \Delta\omega_c$.

The beam reflected from a reference cavity behaves differently since it is comprised of a component reflected directly from the first mirror plus leakage of radiation circulating within the cavity. The response function is more complex, and leads to an error signal which has a very wide capture range, yet a narrow central feature which can be used to lock the frequency to a tight tolerance. Fig. 7.7(a) shows the amplitude and phase of the signal reflected from a reference cavity. It may be seen that the reflected amplitude is approximately unity apart from sharp dips in the region of the resonance frequencies of the cavity; this behaviour is to be expected from our understanding of the *transmission* of a Fabry-Perot etalon. Fig. 7.7(b) shows the error signal derived from a beam modulated at a frequency Ω. The error signal is large, and has the correct sign, for frequencies within approximately $\pm\Omega$ of the resonance frequency. At the same time the central feature has a much smaller width, of order $\Delta\omega_{1/2} = \Delta\omega_{\rm FSR}/\mathcal{F}$, where $\mathcal{F}$ is the finesse of the reference cavity, allowing narrow frequency-locked linewidths to be obtained. This approach can be used to achieve extremely narrow linewidths if the reference cavity is isolated against mechanical and thermal fluctuations and drift.

Paper C2: Laser Science and Quantum Information Processing. Optics

Prof. Simon Hooker

Academic year 2021 – 2022

Contents

1 The wave equation and the paraxial wave equation **1**
1.1 The wave equation . . . 1
1.1.1 The scalar wave equation . . . 2
1.1.2 The analytic signal . . . 3
1.1.3 Harmonic waves and the Helmholtz equation . . . 4
1.2 The angular spectrum method . . . 4
1.2.1 Transmission through an aperture . . . 6
1.2.2 Summary . . . 6
1.3 Paraxial solutions of the wave equation . . . 6
1.4 Gaussian beams . . . 7
1.4.1 Derivation of Gaussian beams by the ASM method . . . 7
1.4.2 Derivation of Gaussian beams from the paraxial wave equation . . . 9
1.4.3 Properties of Gaussian beams . . . 10

2 The Theory of Diffraction **12**
2.1 Green's theorem . . . 12
2.2 The integral theorem of Helmholtz and Kirchoff . . . 13
2.3 The Fresnel-Kirchoff diffraction formula . . . 14
2.3.1 The Kirchoff boundary conditions . . . 15
2.4 Fresnel and Fraunhofer diffraction . . . 17
2.4.1 Fresnel diffraction . . . 17
2.4.2 Fraunhofer diffraction . . . 18
2.4.3 Observation of Fraunhofer diffraction . . . 18

3 Ray transfer matrices **24**
3.1 Ray propagation . . . 24
3.1.1 Describing chains of optical elements . . . 25
3.1.2 Sign conventions and reduced slopes . . . 26
3.2 Application to spherical waves . . . 26
3.3 Application to Gaussian beams . . . 27
3.4 Application to diffraction integrals . . . 28

4 Optical cavities and resonators **31**
4.1 Gauss-Hermite beams . . . 31
4.2 Cavity modes . . . 32
4.2.1 General considerations . . . 32
4.2.2 Condition for a low-loss cavity . . . 34
4.2.3 Alternative derivation of the condition for a low-loss cavity . . . 35
4.3 Longitudinal and transverse modes . . . 37
4.3.1 Transverse modes . . . 37
4.3.2 Longitudinal modes . . . 39

5 Waveguides **40**
5.1 Geometric optics considerations . . . 40
5.2 Scalar wave theory of fibre modes . . . 41
5.2.1 General observations . . . 41
5.2.2 Step-index fibre modes . . . 42
5.3 Dispersion in waveguides . . . 47

6 Interference and coherence **49**
6.1 Introduction . . . 49
6.2 General interference . . . 49
6.2.1 Fringe visibility . . . 51
6.3 Temporal coherence . . . 52
6.3.1 Fourier transform spectroscopy . . . 53
6.3.2 Optical coherence tomography . . . 54
6.4 Transverse coherence . . . 55
6.4.1 The van Cittert-Zernike theorem . . . 55
6.4.2 Michelson stellar interferometer . . . 58

7 Electro-optic modulation **60**
7.1 Introduction . . . 60
7.2 Propagation of light in non-isotropic media . . . 60
7.2.1 Relative orientation of fields . . . 61
7.2.2 The phase velocity . . . 62
7.2.3 Propagation in uniaxial crystals . . . 63
7.2.4 The index ellipsoid (a.k.a. the “indicatrix” or the “ellipsoid of wave normals”) 63
7.3 Nonlinear optics . . . 65
7.3.1 The electro-optic effect . . . 66
7.3.2 An example: ADP . . . 67

8 Frequency generation **71**
8.1 Preliminary considerations . . . 71
8.2 The nonlinear wave equation . . . 72
8.2.1 Phase-matching . . . 74
8.3 Second-harmonic generation (SHG) . . . 76

8.3.1 Phase-matching SHG 76
8.3.2 The effective nonlinear coefficient 77
8.3.3 Further comments 80
8.4 Three-wave mixing 80
8.4.1 Sum-frequency generation 80
8.4.2 Difference-frequency generation 80
8.5 Appendix: Missing steps in derivation of the nonlinear wave equation 82

Lecture 1

The wave equation and the paraxial wave equation

Recommended reading: Brooker, Ch 2; Goodman, Ch 3.

1.1 The wave equation

Our starting point is Maxwell's equations:

$$\nabla \cdot \boldsymbol{D} = \rho \tag{1.1}$$
$$\nabla \cdot \boldsymbol{B} = 0 \tag{1.2}$$
$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} \tag{1.3}$$
$$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}. \tag{1.4}$$

We will restrict ourselves to *linear, isotropic, and homogeneous* media for which the following *constitutive relations* apply:

$$\boldsymbol{D} = \epsilon_0 \epsilon_\mathrm{r} \boldsymbol{E} = \epsilon \boldsymbol{E} \tag{1.5}$$
$$\boldsymbol{B} = \mu_0 \mu_\mathrm{r} \boldsymbol{E} = \mu \boldsymbol{H}. \tag{1.6}$$

If we also assume that the region of interest contains no free charges or currents (i.e. $\rho = 0$ and $\boldsymbol{J} = 0$), then we can write,

$$\nabla \times \nabla \times \boldsymbol{E} = -\nabla \times \frac{\partial \boldsymbol{B}}{\partial t} \tag{1.7}$$

$$\Rightarrow \nabla\nabla \cdot \boldsymbol{E} - \nabla^2 \boldsymbol{E} = -\frac{\partial}{\partial t}\nabla \times \boldsymbol{B} = -\frac{\partial}{\partial t}\left(\epsilon\mu \frac{\partial \boldsymbol{E}}{\partial t}\right) \tag{1.8}$$

$$\Rightarrow \nabla^2 \boldsymbol{E} = \frac{1}{v^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2}. \tag{1.9}$$

This is a wave equation for waves of *phase velocity*,

$$v = (\epsilon\mu)^{-1/2} = \frac{c}{n}, \tag{1.10}$$

where the refractive index is given by,

$$n = \sqrt{\epsilon_{\mathrm{r}}\mu_{\mathrm{r}}} \approx \sqrt{\epsilon_{\mathrm{r}}}, \tag{1.11}$$

and in which the approximation is exact for non-magnetic materials. For propagation in vacuo we have $n = 1$ and $v = c = (\epsilon_0\mu_0)^{-1/2}$.

It is straightforward to show that the magnetic field obeys the same wave equation, i.e.

$$\nabla^2 \boldsymbol{H} = \frac{1}{v^2}\frac{\partial^2 \boldsymbol{H}}{\partial t^2}. \tag{1.12}$$

1.1.1 The scalar wave equation

Since the $\boldsymbol{E}$ and $\boldsymbol{H}$ fields obey the same wave equation, each component obeys a scalar wave equation of the form,

$$\nabla^2 E = \frac{1}{v^2}\frac{\partial^2 E}{\partial t^2}, \tag{1.13}$$

where now E represents one of the vector components. For many problems in optics it is sensible, and perfectly ok, to neglect the vector nature of the fields. However, this is not always so. Examples include, non-uniform media, coupling of fields through boundary conditions, waveguides, and interference of beams at large angles.

Comment: A simple example of a situation in which the scalar wave equation does not hold is that of a non-uniform medium. Suppose, for example, that ϵ depends on the position coordinate $\boldsymbol{r}$. Our derivation of the wave equation would then lead to,

$$\nabla\nabla\cdot\boldsymbol{E}-\nabla^2\boldsymbol{E}=-\frac{\partial}{\partial t}\nabla\times\boldsymbol{B}=-\frac{\partial}{\partial t}\left(\epsilon\mu\frac{\partial\boldsymbol{E}}{\partial t}\right). \tag{1.14}$$

Now,

$$\nabla\cdot\boldsymbol{D}=\nabla\cdot(\epsilon(\boldsymbol{r})\boldsymbol{E})=\epsilon(\boldsymbol{r})\nabla\cdot\boldsymbol{E}+\nabla\epsilon(\boldsymbol{r})\cdot\boldsymbol{E}=0 \tag{1.15}$$

$$\Rightarrow\nabla\cdot\boldsymbol{E}=-\frac{1}{\epsilon(\boldsymbol{r})}\nabla\epsilon(\boldsymbol{r})\cdot\boldsymbol{E}. \tag{1.16}$$

Our wave equation then becomes, from (1.8),

$$\nabla^2\boldsymbol{E}+\nabla\left(\frac{1}{\epsilon(\boldsymbol{r})}\nabla\epsilon(\boldsymbol{r})\cdot\boldsymbol{E}\right)=\frac{1}{v^2}\frac{\partial^2\boldsymbol{E}}{\partial t^2}. \tag{1.17}$$

The effect of the additional term is that the vector components of $\boldsymbol{E}$ are coupled: for example, if $\nabla\epsilon(\boldsymbol{r})$ has a component along the y-direction, then the wave equation for E_x will contain terms in E_y.

1.1.2 The analytic signal

To emphasize the fact that we have simplified to the scaler wave equation we will usually denote the wave amplitude by the *real signal* $V(t)$. This signal forms a Fourier transform[1] pair with the *spectrum* $\tilde{V}(\omega)$ of the signal:

$$V(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\tilde{V}(\omega)\exp\left(-\mathrm{i}\omega t\right)\mathrm{d}\omega \tag{1.18}$$

$$\tilde{V}(\omega)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}V(t)\exp\left(+\mathrm{i}\omega t\right)\mathrm{d}t. \tag{1.19}$$

It is easy to show that, since $V(t)$ is real,

$$\tilde{V}(-\omega)=\tilde{V}^*(\omega). \tag{1.20}$$

Note that we can write eqn. (1.18) in the form,

[1]The sign convention for these Fourier transforms between time and frequency has been chosen so that the spectrum of a monochromatic wave written in the form $\exp[\mathrm{i}(kz-\omega_0 t)]$ is proportional to a delta function centred at $\omega=\omega_0$. For the same reason, the opposite sign convention will be used when we take Fourier transforms in the spatial domain.

$$\begin{aligned} V(t) &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \tilde{V}(\omega) \exp\left(-\mathrm{i}\omega t\right) \mathrm{d}\omega + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 \tilde{V}(\omega) \exp\left(-\mathrm{i}\omega t\right) \mathrm{d}\omega \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \tilde{V}(\omega) \exp\left(-\mathrm{i}\omega t\right) \mathrm{d}\omega - \frac{1}{\sqrt{2\pi}} \int_\infty^0 \tilde{V}(-\omega') \exp\left(+\mathrm{i}\omega' t\right) \mathrm{d}\omega' \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \tilde{V}(\omega) \exp\left(-\mathrm{i}\omega t\right) \mathrm{d}\omega + \frac{1}{\sqrt{2\pi}} \int_0^\infty \tilde{V}^*(\omega') \exp\left(+\mathrm{i}\omega' t\right) \mathrm{d}\omega' \\ &= 2\Re\left\{ \frac{1}{\sqrt{2\pi}} \int_0^\infty \tilde{V}(\omega) \exp\left(-\mathrm{i}\omega t\right) \mathrm{d}\omega \right\}, \end{aligned} \tag{1.21}$$

where we have used the substitution $\omega' = -\omega$. We can therefore write the real signal as the real part of a complex **analytic signal** $\mathcal{U}(t)$, where

$$\mathcal{U}(t) = 2\frac{1}{\sqrt{2\pi}} \int_0^\infty \tilde{V}(\omega) \exp\left(-\mathrm{i}\omega t\right) \mathrm{d}\omega. \tag{1.22}$$

We will use the analytic signal extensively in what follows. Note that the Fourier transform of the analytic signal is $\tilde{\mathcal{U}}(\omega) = 2\tilde{V}(\omega)$, i.e. it is just twice the transform of the real signal.

1.1.3 Harmonic waves and the Helmholtz equation

We will very often want to consider the propagation of harmonic waves (i.e. waves with a well defined frequency). Such waves can be described by the analytic function,

$$\mathcal{U}(\boldsymbol{r}, t) = U(\boldsymbol{r}) \exp\left(-\mathrm{i}\omega t\right), \tag{1.23}$$

where, in general, $U(\boldsymbol{r})$ will be complex. Inserting this into the scalar wave equation yields,

$$\nabla^2 U(\boldsymbol{r}) = -\frac{\omega^2}{v^2} U(\boldsymbol{r}) \tag{1.24}$$

$$\Rightarrow \left[\nabla^2 + k^2\right] U(\boldsymbol{r}) = 0 \qquad \text{Helmholtz equation,} \tag{1.25}$$

where the wave vector $k = 2\pi/\lambda$, λ is the wavelength, and $\omega = kv$. Equation (1.25) is known as the **Helmholtz equation**; it describes the propagation of monochromatic components of the field in a uniform medium.

1.2 The angular spectrum method

The *angular spectrum method* (ASM) is an extremely powerful way to solve the *full* scalar wave equation; it is especially useful for numerical solution of the wave equation, but in some circumstances it can be used to find analytical solutions. The essential idea is to find the *spatial* Fourier transform

of the field distribution in some plane ("the angular spectrum"), propagate this through the optical system using a certain set of rules, and to take the inverse transform to find the field at the output of the system.

We consider a harmonic field wave propagating along the z-axis, and described by the analytic function $\mathcal{U}(\boldsymbol{r}) = U(\boldsymbol{r}) \exp(-\mathrm{i}\omega t)$. The angular spectrum of $U(x, y, z)$ is given by the Fourier transform,[2] and its inverse:

$$\tilde{U}(k_x, k_y; z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U(x, y, z) \exp\left[-\mathrm{i}\left(k_x x + k_y y\right)\right] \mathrm{d}x\mathrm{d}y \tag{1.26}$$

$$U(x, y, z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{U}(k_x, k_y; z) \exp\left[+\mathrm{i}\left(k_x x + k_y y\right)\right] \mathrm{d}k_x\mathrm{d}k_y. \tag{1.27}$$

Note that $\tilde{U}(k_x, k_y; z)$ is the amplitude of the spatial frequencies (k_x, k_y) which make up the "object" $U(x, y, z)$.

In order to find how $\tilde{U}(k_x, k_y; z)$ propagates we substitute (1.27) into the Helmholtz equation:

$$\left[\nabla^2 + k^2\right] \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{U}(k_x, k_y; z) \exp\left[\left(\mathrm{i}k_x x + k_y y\right)\right] \mathrm{d}k_x\mathrm{d}k_y = 0 \tag{1.28}$$

$$\Rightarrow \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left\{ \left[-k_x^2 - k_y^2 + k^2\right] \tilde{U}(k_x, k_y; z) + \frac{\mathrm{d}^2\tilde{U}}{\mathrm{d}z^2} \right\} \exp\left[\left(\mathrm{i}k_x x + k_y y\right)\right] \mathrm{d}k_x\mathrm{d}k_y = 0 \tag{1.29}$$

$$\Rightarrow \frac{\mathrm{d}^2\tilde{U}}{\mathrm{d}z^2} + \left[k^2 - k_x^2 - k_y^2\right] \tilde{U}(k_x, k_y; z) = 0. \tag{1.30}$$

Equation (1.30) has the trivial solution,

$$\tilde{U}(k_x, k_y; z) = \tilde{U}(k_x, k_y; 0) \exp\left(\mathrm{i}\sqrt{k^2 - k_x^2 - k_y^2}z\right) = \tilde{U}(k_x, k_y; 0)\mathcal{H}(k_x, k_y, z). \tag{1.31}$$

We see that propagation through a uniform medium corresponds to mulitplying $\tilde{U}(k_x, k_y; z)$ by the propagating function,

$$\mathcal{H}(k_x, k_y, z) = \exp\left(\mathrm{i}\sqrt{k^2 - k_x^2 - k_y^2}\ z\right). \tag{1.32}$$

This makes sense: we can think of the field in some input plane $z = z_1$ as being formed by a superposition of plane waves of the form $\tilde{U}(k_x, k_y; z_1) \exp\left[\mathrm{i}\left(k_x x + k_y y + k_z z_1 - \omega t\right)\right]$, where $k_x^2 + k_y^2 + k_z^2 = k^2$. In propagating to the plane $z = z_2$ these plane waves will accumulate an additional phase $k_z(z_2 - z_1)$, where $k_z = \sqrt{k^2 - k_x^2 - k_y^2}$.

[2] As advertised, here we use the opposite sign convention to those used in eqns (1.18) and (1.19).

1.2.1 Transmission through an aperture

Suppose that the beam is passed through an aperture with an amplitude transmission function $t(x, y)$. The transmitted amplitude will be $U_{\text{out}}(x, y) = U_{\text{in}}t(x, y)$. Hence the angular spectrum of the transmitted beam will be,

$$\tilde{U}_{\text{out}}(k_x, k_y; z) = \text{FT}_{k_x,k_y}\left[U_{\text{in}}(x, y)t(x, y)\right] = 2\pi\tilde{U}_{\text{in}}(k_x, k_y; z) \circledast \tilde{t}(k_x, k_y), \tag{1.33}$$

where we have used the Convolution Theorem. **Hence to propagate an angular spectrum through an aperture one should convolve the angular spectrum with the (spatial) Fourier transform of the transmission function of the object**.

1.2.2 Summary

The ASM method therefore comprises the following steps:

1. Find the angular spectrum $\tilde{U}(k_x, k_y; z_1)$ of the field $U(x, y, z_1)$ in the input plane $z = z_1$.
2. Propagate the angular spectrum through the optical system by multiplying by the propagator $\mathcal{H}(k_x, k_y, z)$ and/or convolving with the Fourier transform $\tilde{t}(k_x, k_y)$ of any apertures in the system
3. Take the inverse Fourier transform of the propagated angular spectrum $\tilde{U}(k_x, k_y; z_2)$ to find the field $U(x, y, z_2)$ in the output plane $z = z_2$.

Optional exercise:

1. Write a computer code (in one-dimension if you prefer) for numerical propagation of a beam by the ASM method.
2. Use your code to explore diffraction of a simple beam (e.g. a plane wave or a Gaussian beam) through simple apertures, and check that the results are consistent with the analytical results we obtain in later sections.
3. Use your code to explore propagation of beams through a thin lens, and check that the results are consistent with the analytical results we obtain in later sections.

1.3 Paraxial solutions of the wave equation

We will often want to investigate the propagation of a beam propagating along an axis, which we take to be the z-axis. As we will see, this beam may diffract, or be focused, but the wave normals will still be predominantly oriented along the z-axis. This will allow us to make a powerful approximation, the paraxial approximation, which we will employ for much of this part of the course.

To emphasize the special nature of the z-component of the propagation we will look for solutions of the form,

$$U(\boldsymbol{r}) = u(x,y,z)\exp(\mathrm{i}kz)\,. \tag{1.34}$$

Substituting this solution into the Helmholtz equation we find,

$$\left[\nabla^2 + k^2\right] u(x,y,z)\exp(\mathrm{i}kz) = 0 \tag{1.35}$$

$$\Rightarrow \exp(\mathrm{i}kz)\,\nabla_\perp^2 u + \frac{\partial}{\partial z}\left[\mathrm{i}ku\exp(\mathrm{i}kz) + \frac{\partial u}{\partial z}\exp(\mathrm{i}kz)\right] + k^2 u\exp(\mathrm{i}kz) = 0 \tag{1.36}$$

$$\Rightarrow \nabla_\perp^2 u + \left[-k^2 u + ik\frac{\partial u}{\partial z} + ik\frac{\partial u}{\partial z} + \frac{\partial^2 u}{\partial z^2}\right] = 0 \tag{1.37}$$

$$\Rightarrow \nabla_\perp^2 u + \frac{\partial^2 u}{\partial z^2} + 2ik\frac{\partial u}{\partial z} = 0, \tag{1.38}$$

where $\nabla_\perp^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$. The above equation is still exact. However, we now recognize that the amplitude u will vary negligibly for changes in z of the order of the wavelength; in fact, as we will see later, u will only vary appreciably over a length scale of the order of the Rayleigh range. We will also assume that variations of u in the z-direction are slow compared to those in the transverse coordinates. These assumptions constitute the *paraxial approximation*:

$$\left|\frac{\partial^2 u}{\partial z^2}\right| = \left|\frac{\partial}{\partial z}\left(\frac{\partial u}{\partial z}\right)\right| \ll k\frac{\partial u}{\partial z} \tag{1.39}$$

$$\left|\frac{\partial^2 u}{\partial z^2}\right| \ll \left|\frac{\partial^2 u}{\partial x^2}\right|, \left|\frac{\partial^2 u}{\partial y^2}\right|. \tag{1.40}$$

Making the paraxial approximation we arrive at the **paraxial wave equation**:

$$\nabla_\perp^2 u + 2ik\frac{\partial u}{\partial z} = 0. \qquad \text{The paraxial wave equation} \tag{1.41}$$

1.4 Gaussian beams

As we shall see when we consider the beams output by laser cavities, an important example of a paraxial solution to the wave equation is the Gaussian beam.

1.4.1 Derivation of Gaussian beams by the ASM method

As an example of the use of the ASM in an *analytical* calculation, let us consider the propagation of a Gaussian beam of the form,

$$U(x,y,0) = \left(\frac{1}{-\mathrm{i}\pi w_0^2/\lambda}\right) U_0 \exp\left(-\frac{x^2+y^2}{w_0^2}\right). \tag{1.42}$$

The reason for the odd pre-factor in brackets will become clear. The angular spectrum is given by,

$$\tilde{U}(k_x,k_y;0) = \left(\frac{1}{-\mathrm{i}\pi w_0^2/\lambda}\right)\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U_0 \exp\left(-\frac{x^2+y^2}{w_0^2}\right)\exp\left[-\mathrm{i}\left(k_x x + k_y y\right)\right]\mathrm{d}x\mathrm{d}y \tag{1.43}$$

$$= \left(\frac{1}{-\mathrm{i}\pi w_0^2/\lambda}\right)\frac{U_0}{2\pi}\int_{-\infty}^{\infty}\exp\left(-\frac{x^2}{w_0^2}\right)\exp\left(-\mathrm{i}k_x x\right)\mathrm{d}x\int_{-\infty}^{\infty}\exp\left(-\frac{y^2}{w_0^2}\right)\exp\left(-\mathrm{i}k_y y\right)\mathrm{d}y \tag{1.44}$$

$$= \left(\frac{1}{-\mathrm{i}\pi w_0^2/\lambda}\right)\frac{U_0}{2\pi}\left\{\sqrt{\pi}w_0\exp\left[-\left(\frac{k_x w_0}{2}\right)^2\right]\right\}\left\{\sqrt{\pi}w_0\exp\left[-\left(\frac{k_y w_0}{2}\right)^2\right]\right\} \tag{1.45}$$

$$= \frac{\mathrm{i}\lambda}{2\pi}U_0\exp\left[-\left(k_x^2+k_y^2\right)\frac{w_0^2}{4}\right], \tag{1.46}$$

where we have used a standard integral.

To propagate this beam to the plane $z = z$ we need to multiply by the propagator $\mathcal{H}(k_x,k_y,z)$. In doing this *we will make the paraxial approximation*, so that,

$$\mathcal{H}(k_x,k_y,z) = \exp\left(\mathrm{i}\sqrt{k^2-k_x^2-k_y^2}z\right) \approx \exp\left[\mathrm{i}\left(k - \frac{k_x^2}{2k} - \frac{k_y^2}{2k}\right)z\right]. \tag{1.47}$$

We can now find the field in the plane $z = z$ by taking the inverse Fourier transform:

$$U(x,y,z) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\tilde{U}(k_x,k_y;z)\exp\left[+\mathrm{i}\left(k_x x + k_y y\right)\right]\mathrm{d}k_x\mathrm{d}k_y \tag{1.48}$$

$$= \frac{\mathrm{i}\lambda}{2\pi}U_0\exp\left(\mathrm{i}kz\right)\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left[-\left(k_x^2+k_y^2\right)\frac{w_0^2}{4}\right]\exp\left[-\mathrm{i}\left(\frac{k_x^2}{2k}+\frac{k_y^2}{2k}\right)z\right] \times \exp\left[+\mathrm{i}\left(k_x x + k_y y\right)\right]\mathrm{d}k_x\mathrm{d}k_y \tag{1.49}$$

$$= \exp\left(\mathrm{i}kz\right)\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left\{\frac{\mathrm{i}\lambda}{2\pi}U_0\exp\left[-\left(\frac{w_0^2}{4}+\mathrm{i}\frac{z}{2k}\right)\left(k_x^2+k_y^2\right)\right]\right\} \times \exp\left[+\mathrm{i}\left(k_x x + k_y y\right)\right]\mathrm{d}k_x\mathrm{d}k_y \tag{1.50}$$

We note that the term in curly brackets is the same as the right-hand side of (1.46), but with

$$\frac{w_0^2}{4} \to \frac{w_0^2}{4} + \mathrm{i}\frac{z}{2k} \tag{1.51}$$

$$\Rightarrow w_0^2 \to w_0^2 + 2\mathrm{i}\frac{z}{k} \tag{1.52}$$

Hence we do not need to evaluate the inverse Fourier transform, but can write immediately,

$$U(x,y,z) = \exp(\mathrm{i}kz)\left[\frac{1}{-\mathrm{i}\pi\left(w_0^2 + 2\mathrm{i}\frac{z}{k}\right)/\lambda}\right] U_0 \exp\left(-\frac{x^2+y^2}{w_0^2 + 2\mathrm{i}\frac{z}{k}}\right) \tag{1.53}$$

$$= U_0 \exp(\mathrm{i}kz)\frac{1}{q}\exp\left(\mathrm{i}k\frac{x^2+y^2}{2q}\right) \tag{1.54}$$

$$= U_0 \exp(\mathrm{i}kz)\frac{1}{q}\exp\left(\mathrm{i}k\frac{\rho^2}{2q}\right), \tag{1.55}$$

where $\rho = \sqrt{x^2+y^2}$ is the transverse distance from the z-axis and the **complex radius** is defined by,

$$q = z - \mathrm{i}\frac{\pi w_0^2}{\lambda}. \tag{1.56}$$

We will see that (1.55) describes a lowest-order Gaussian beam at all points along the z-axis.

1.4.2 Derivation of Gaussian beams from the paraxial wave equation

We can also derive the form of a Gaussian beam from the paraxial wave equation. A spherical wave launched from the point (x_0, y_0, z_0) is described by,

$$U(x,y,z) = \frac{U_0}{r}\exp(\mathrm{i}kr) \tag{1.57}$$

where,

$$r = \sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2} \approx (z-z_0) + \frac{1}{2}\frac{(x-x_0)^2}{(z-z_0)} + \frac{1}{2}\frac{(y-y_0)^2}{(z-z_0)} \tag{1.58}$$

and the approximation holds close to the z-axis.

In the paraxial region the spherical wave can be written as,

$$U(x,y,z) = U_0 \exp[\mathrm{i}(z-z_0)]\frac{1}{z-z_0}\exp\left(\mathrm{i}k\frac{\rho^2}{2(z-z_0)}\right), \tag{1.59}$$

where $\rho = \sqrt{(x-x_0)^2 + (y-y_0)^2}$ is the distance from the propagation axis. Equation (1.59) is a solution to the paraxial wave equation. Now, since z_0 is just a constant, the solution above remains a solution to the paraxial wave equation if we replace z_0 by a complex number. In particular we can let,

$$z - z_0 = q = z - z_0 - \mathrm{i}z_\mathrm{R}. \tag{1.60}$$

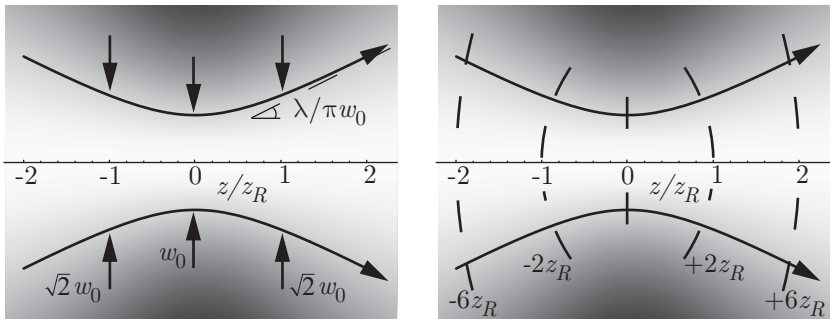


Figure 1.1: The main features of a lowest-order Gaussian beam of focal spot size w_0 and Rayleigh range z_R, with its waist located at $z_0 = 0$: (a) the variation of the spot size of the beam; (b) the variation of the radius of curvature of the wave fronts. For both plots, white represents the highest intensity, black the lowest. The solid curves show the locus of points at a radial distance of one spot size from the propagation axis. Notice that for $z \gg z_R$ these points diverge from the propagation axis at an angle of $\theta = \lambda/\pi w_0$

Then, (1.59) takes the form of the Gaussian beam.

$$U(x, y, z) = U_0 \exp(\mathrm{i}kz) \frac{1}{q} \exp\left(\mathrm{i}k\frac{\rho^2}{2q}\right), \tag{1.61}$$

1.4.3 Properties of Gaussian beams

Figure 1.1 illustrates some of the key properties of a Gaussian beam. To understand these we write,

$$\frac{\mathrm{i}k\rho^2}{2q} = \frac{\mathrm{i}k\rho^2}{2} \frac{(z - z_0) + \mathrm{i}z_\mathrm{R}}{(z - z_0)^2 + z_\mathrm{R}^2} \equiv \frac{\mathrm{i}k\rho^2}{2R(z)} - \frac{\rho^2}{w(z)^2}. \tag{1.62}$$

where,

$$R(z) = \frac{(z - z_0)^2 + z_\mathrm{R}^2}{z - z_0} \tag{1.63}$$

$$w(z) = w_0 \sqrt{1 + \left(\frac{z - z_0}{z_\mathrm{R}}\right)^2} \tag{1.64}$$

$$w_0 = \sqrt{\frac{\lambda z_\mathrm{R}}{\pi}} \tag{1.65}$$

By comparison with (1.59), $R(z)$ is the radius of curvature of the wavefront. The parameter $w(z)$ is known as the spot size of the beam, and it describes the transverse extent of the beam. The parameter

$z_{\mathrm{R}} = \pi w_0^2/\lambda$ is known as the **Rayleigh range** of the beam. It is the distance over which the spot size of the beam increases by a fact of $\sqrt{2}$ over that at the focus.

Similarly,

$$\frac{1}{q} = \frac{1}{(z - z_0) - \mathrm{i}z_{\mathrm{R}}} = \frac{\mathrm{i}\mathrm{e}^{-\mathrm{i}\alpha}}{\sqrt{(z - z_0)^2 + z_{\mathrm{R}}^2}}, \tag{1.66}$$

where,

$$\tan\alpha = \frac{z - z_0}{z_{\mathrm{R}}}. \tag{1.67}$$

In summary the Gaussian beam in the plane $z = z$ can be written in the form,

$$\begin{aligned}
\mathcal{U}(x, y, z, t) &= U(x, y, z)\exp(-\mathrm{i}\omega t) \\
&= \mathrm{i}\frac{U_0}{z_{\mathrm{R}}} && \text{Constants; amplitude at waist} \\
&\times \frac{w_0}{w(z)} && \text{Conserves energy; amplitude gets smaller as beam diffracts} \\
&\times \exp\left[\mathrm{i}\left(kz - \omega t\right)\right] && \text{Wave travels towards sign(k)} \\
&\times \exp\left[-\mathrm{i}\alpha(z)\right] && \text{Additional phase change with propagation} \\
&\times \exp\left[-\frac{\rho^2}{w(z)^2}\right] && \text{Gaussian transverse profile, with spot size } w(z) \\
&\times \exp\left[\mathrm{i}\frac{k\rho^2}{2R(z)}\right] && \text{Wavefronts are spherical with radius of curvature } R(z)
\end{aligned} \tag{1.68}$$

Lecture 2

The Theory of Diffraction

Recommended reading: Brooker, Ch 3; Goodman, Ch 3.

The theory of diffraction attempts to answer the question "If we know the optical field over some surface (often a plane), what is the field on some other surface?" This requires us to find solutions to the wave equation.

2.1 Green's theorem

We shall restrict ourselves to solution of the scalar wave equation. A useful starting point is Green's theorem, which states that if $U(\boldsymbol{r})$ and $G(\boldsymbol{r})$ are single-valued and continuous functions on, and within, a surface S, then,

$$\int_{\mathrm{V}} \left(U\nabla^2 G - G\nabla^2 U\right) \mathrm{d}V = \int_{\mathrm{S}} \left(U\nabla G - G\nabla U\right) \cdot \mathbf{d}\boldsymbol{S}, \tag{2.1}$$

where V is the volume contained within S.

One of these fields, $U(\boldsymbol{r})$, will be the optical field we wish to find. The other will be a well-chosen auxiliary field which will allow solution of the wave equation.

Aside: Note that Green's theorem follows from the Divergence Theorem (i.e. Gauss's Theorem) applied to the vector fields $U\nabla G$ and $G\nabla U$:

$$\int_{\mathrm{S}} (U\nabla G) \cdot \mathbf{d}\boldsymbol{S} = \int_{\mathrm{V}} \left[\nabla \cdot (U\nabla G)\right] \mathrm{d}V = \int_{\mathrm{V}} \left[\nabla U \cdot \nabla G + U\nabla^2 G\right] \mathrm{d}V \tag{2.2}$$

$$\int_{\mathrm{S}} (G\nabla U) \cdot \mathbf{d}\boldsymbol{S} = \int_{\mathrm{V}} \left[\nabla \cdot (G\nabla U)\right] \mathrm{d}V = \int_{\mathrm{V}} \left[\nabla G \cdot \nabla U + G\nabla^2 U\right] \mathrm{d}V. \tag{2.3}$$

Subtraction of these two equations yields Green's theorem.

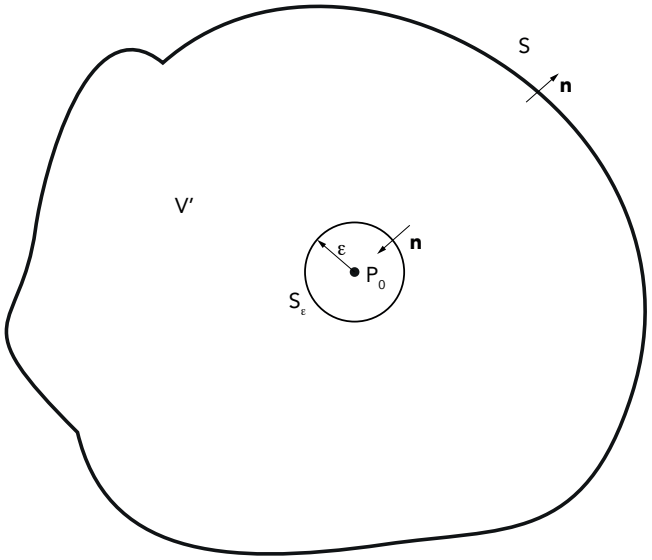


Figure 2.1: Geometry used in the derivation of the theorem of Helmholtz and Kirchoff.

2.2 The integral theorem of Helmholtz and Kirchoff

Let us suppose that we wish to find the optical field at a point P_0, as illustrated in Fig. 2.1. We will first find this field in terms of the fields over a surface S which surrounds P_0. To this end we will try as an auxiliary function a spherical wave leaving P_0, so that the auxiliary field at $\boldsymbol{r_1}$ is:

$$G(\boldsymbol{r_1}) = \frac{1}{r_{01}} \exp\left(\mathrm{i}kr_{01}\right), \tag{2.4}$$

where $\boldsymbol{r_{01}} = \boldsymbol{r_1} - \boldsymbol{r_0}$.

To avoid mathematical complications in the region close to P_0 we surround that point by a spherical surface S_ϵ of small radius ϵ. We will then apply Green's theorem to the volume V' which lies between S and S_ϵ. Remembering that both U and G satisfy the Helmholtz equation, we find:

$$\int_{\mathrm{S}'} (U\nabla G - G\nabla U)\cdot \mathbf{d}\boldsymbol{S} = \int_{\mathrm{V}'} \left(U\nabla^2 G - G\nabla^2 U\right) \mathrm{d}V \tag{2.5}$$

$$\Rightarrow \int_{\mathrm{S}'} (U\nabla G - G\nabla U)\cdot \mathbf{d}\boldsymbol{S} = \int_{\mathrm{V}'} \left(-Uk^2 G + Gk^2 U\right) \mathrm{d}V \tag{2.6}$$

$$\Rightarrow \int_{\mathrm{S}'} (U\nabla G - G\nabla U)\cdot \mathbf{d}\boldsymbol{S} = 0 \tag{2.7}$$

$$\Rightarrow -\int_{\mathrm{S}_\epsilon} (U\nabla G - G\nabla U)\cdot \mathbf{d}\boldsymbol{S} = \int_{\mathrm{S}} (U\nabla G - G\nabla U)\cdot \mathbf{d}\boldsymbol{S}. \tag{2.8}$$

This result allows us to relate the field on S_ϵ to that on S.

To make further progress we note that (remembering that the vector operators operate on the coordinates of $\boldsymbol{r_1}$, not $\boldsymbol{r_0}$)

$$\nabla G = \left(\mathrm{i}k - \frac{1}{r_{01}}\right)\frac{\exp\left(\mathrm{i}kr_{01}\right)}{r_{01}}\hat{\boldsymbol{r}}_{\mathbf{01}}. \tag{2.9}$$

Applying this to (2.8) we find,

$$-\int_{\mathrm{S}_\epsilon}\left[U\left(\mathrm{i}k - \frac{1}{r_{01}}\right)\frac{\exp\left(\mathrm{i}kr_{01}\right)}{r_{01}}\hat{\boldsymbol{r}}_{\mathbf{01}} - G\nabla U\right]\cdot\mathbf{d}\boldsymbol{S} = \int_{\mathrm{S}}\left[U\left(\mathrm{i}k - \frac{1}{r_{01}}\right)\frac{\exp\left(\mathrm{i}kr_{01}\right)}{r_{01}}\hat{\boldsymbol{r}}_{\mathbf{01}} - G\nabla U\right]\cdot\mathbf{d}\boldsymbol{S}. \tag{2.10}$$

We need to evaluate the LHS as $\epsilon \to 0$:

$$\begin{aligned}
&\lim_{\epsilon\to 0}\int_{\mathrm{S}_\epsilon}\left[U(P_0)\left(\mathrm{i}k - \frac{1}{r_{01}}\right)\frac{\exp\left(\mathrm{i}kr_{01}\right)}{r_{01}}\hat{\boldsymbol{r}}_{\mathbf{01}} - G\nabla U(P_0)\right]\cdot\mathbf{d}\boldsymbol{S}\\
&= -\lim_{\epsilon\to 0} 4\pi\epsilon^2\left[U(P_0)\left(\mathrm{i}k - \frac{1}{\epsilon}\right)\frac{\exp\left(\mathrm{i}k\epsilon\right)}{\epsilon} - \frac{1}{\epsilon}\exp\left(\mathrm{i}k\epsilon\right)\frac{\partial U(P_0)}{\partial r}\right]\\
&= 4\pi U(P_0),
\end{aligned} \tag{2.11}$$

where we have remembered that the normal to S_ϵ points towards P_0 (i.e. antiparallel to $\hat{\boldsymbol{r}}_{\mathbf{01}}$).

Substituting this into (2.8) we find,

$$U(P_0) = -\frac{1}{4\pi}\int_{\mathrm{S}}\left(U\nabla G - G\nabla U\right)\cdot\mathbf{d}\boldsymbol{S} \tag{2.12}$$

$$= \frac{1}{4\pi}\int_{\mathrm{S}}\left[\frac{1}{r_{01}}\exp\left(\mathrm{i}kr_{01}\right)\nabla U - U\left(\mathrm{i}k - \frac{1}{r_{01}}\right)\frac{\exp\left(\mathrm{i}kr_{01}\right)}{r_{01}}\hat{\boldsymbol{r}}_{\mathbf{01}}\right]\cdot\mathbf{d}\boldsymbol{S}. \tag{2.13}$$

This is the **integral theorem of Helmholtz and Kirchoff**. It allows the field at any point (P_0) to be determined from the fields on a surface surrounding that point.

2.3 The Fresnel-Kirchoff diffraction formula

We now apply these results to the case of a planar screen, with its normal parallel to the z-axis, as shown in Fig. 2.2. We consider a surface S comprised of: (i) a plane, S_1 in contact with the rear surface of the screen; and (ii) a spherical cap, S_2, of radius R and which connects with S_2. We will consider first the RHS of (2.13) evaluated over S_2:

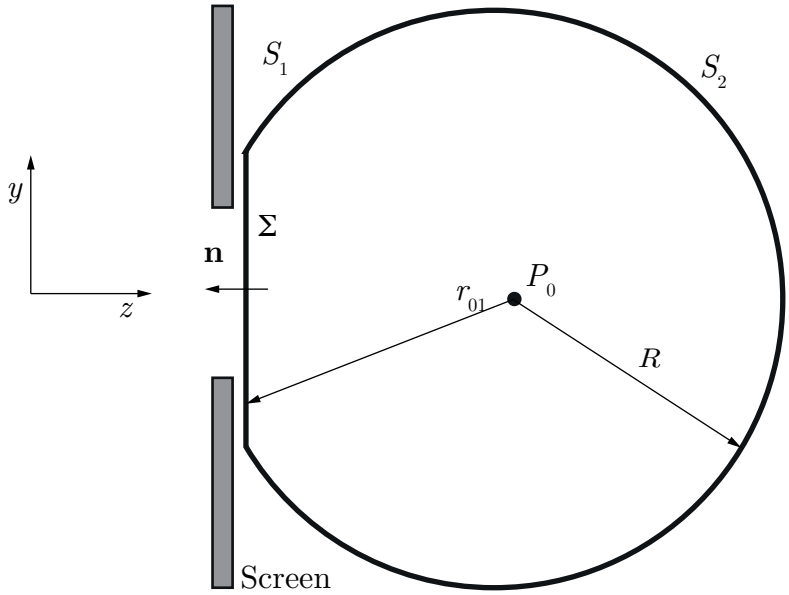


Figure 2.2: Geometry used in the derivation of the Fresnel-Kirchoff diffraction formula.

$$\frac{1}{4\pi}\int_{S_2}\left[\frac{1}{r_{01}}\exp\left(\mathrm{i}kr_{01}\right)\nabla U - U\left(\mathrm{i}k-\frac{1}{r_{01}}\right)\frac{\exp\left(\mathrm{i}kr_{01}\right)}{r_{01}}\hat{\boldsymbol{r}}_{\mathbf{01}}\right]\cdot\mathbf{d}\boldsymbol{S} \tag{2.14}$$

$$=\int_{S_2}\left[\frac{1}{R}\exp\left(\mathrm{i}kR\right)\frac{\partial U}{\partial r} - U\left(\mathrm{i}k-\frac{1}{R}\right)\frac{\exp\left(\mathrm{i}kR\right)}{R}\right]\mathrm{d}S \tag{2.15}$$

$$\approx\int_{S_2}\left[G\frac{\partial U}{\partial r} - \mathrm{i}kGU\right]\mathrm{d}S \tag{2.16}$$

$$\approx\int_{\Omega} G\left[\frac{\partial U}{\partial r} - \mathrm{i}kU\right]R^2\mathrm{d}\Omega. \tag{2.17}$$

Since, on S_2, $G \propto 1/R$ the above integral will tend to zero as R becomes large provided,

$$\lim_{R\to\infty}\left[R\left(\frac{\partial U}{\partial r} - \mathrm{i}kU\right)\right] = 0. \tag{2.18}$$

This condition, the Sommerfeld radiation condition, will hold provided that U vanishes at least as fast as a spherical wave. This will be the case if there are only outgoing waves on S_2.

2.3.1 The Kirchoff boundary conditions

Having got rid of the contribution from S_2 we now need to evaluate that from S_1. To make progress we will assume the **Kirchoff boundary conditions**:[1]

[1] At first glance these boundary conditions seem perfectly reasonable, but they cannot be true since: (i) the boundary conditions lead to discontinuities in the fields; (ii) in detail the behaviour of the fields must depend on the properties of

1. For transparent parts of the screen the field U and its derivatives are exactly the same as those of the incident wave;

2. Over the remainder of S_1 the field U and its derivatives are zero.

We then have,

$$U(P_0) = \frac{1}{4\pi}\int_\Sigma \left(\frac{1}{r_{01}}\exp\left(\mathrm{i}kr_{01}\right)\nabla U - U\left(\mathrm{i}k - \frac{1}{r_{01}}\right)\frac{\exp\left(\mathrm{i}kr_{01}\right)}{r_{01}}\hat{\boldsymbol{r}}_{\mathbf{01}}\right)\cdot \mathbf{d}\boldsymbol{S}. \tag{2.19}$$

We suppose that the aperture is illuminated by a spherical wave originating from the point S_0, so that on Σ:

$$U = \frac{U_0}{r_{s1}}\exp\left(\mathrm{i}kr_{s1}\right) \tag{2.20}$$

$$\nabla U = \left(\mathrm{i}k - \frac{1}{r_{s1}}\right)U_0\frac{\exp\left(\mathrm{i}kr_{s1}\right)}{r_{s1}}\hat{\boldsymbol{r}}_{\boldsymbol{s1}}. \tag{2.21}$$

Provided the distances to the screen from the source and observation point are large, we may neglect r_{01}^{-1} and r_{s1}^{-1} compared to k. We then have,

$$U(P_0) \approx \frac{\mathrm{i}k}{4\pi}\int_\Sigma U_0\frac{\exp\left(\mathrm{i}kr_{s1}\right)}{r_{s1}}\frac{\exp\left(\mathrm{i}kr_{01}\right)}{r_{01}}\left(\hat{\boldsymbol{r}}_{\boldsymbol{s1}} - \hat{\boldsymbol{r}}_{\mathbf{01}}\right)\cdot\mathbf{d}\boldsymbol{S} \tag{2.22}$$

$$\approx \frac{1}{\mathrm{i}\lambda}\int_\Sigma U_0\frac{\exp\left(\mathrm{i}kr_{s1}\right)}{r_{s1}}\frac{\exp\left(\mathrm{i}kr_{01}\right)}{r_{01}}\left\{\frac{\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{r}}_{\boldsymbol{s1}} - \hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{r}}_{\mathbf{01}}}{2}\right\}\mathrm{d}S \tag{2.23}$$

where $\boldsymbol{n}$ is the unit normal to Σ. This is the **Fresnel-Kirchoff** diffraction integral. Notice that the field at P_0 can be interpreted as arising from a superposition of spherical waves $(\exp(\mathrm{i}kr_{01})/r_{01})$ originating from the transparent parts of the diffraction screen with:

1. An amplitude which is proportional to that of the incident wave (the term in square brackets)

2. An amplitude which decreases with angle away from the forward direction, according to the **obliquity factor** (the term in curly brackets)

3. A phase of $-\pi/2$ relative to that of the incident wave.

We seen, then, that the Fresnel-Kirchoff theory reproduces the rather ad hoc assumptions of Huygens' theory, and yields quantitative expressions for the obliquity factor and phase of the "secondary wavelets".

---

the material of the screen (e.g. whether or not it is conducting); and (iii) in reality the fields will extend a short distance into the shadow regions behind the opaque parts of the screen. See §3.18 of Brooker for further discussion of these points.

2.4 Fresnel and Fraunhofer diffraction

Finding solutions to (2.23) requires some further approximations to be made. We will assume that the point P_0 lies sufficiently far from Σ that:

- The obliquity factor is approximately unity over the range of integration
- Variations in the factor $1/r_{s1}$ are small, and hence this can be treated as a constant.

We derived the Fresnel-Kirchoff diffraction integral by considering illumination of an aperture by a spherical wave. We could have considered a more general case in which the incident field takes the form $E_{\text{in}}(x, y)$. In this case, and with the assumptions above, we have,[2]

$$U(X, Y, Z) \approx \frac{1}{\mathrm{i}\lambda}\frac{1}{Z}\int_{\Sigma} E_{\text{in}}(x, y)\exp\left(\mathrm{i}kr_{01}\right)\mathrm{d}x\mathrm{d}y, \tag{2.24}$$

where we have taken the coordinates of P_0 to be (X, Y, Z).

To proceed we write r_{01} as,

$$\begin{aligned} r_{01} &= \left[(X-x)^2 + (Y-y)^2 + (Z-0)^2\right]^{1/2} && (2.25)\\ &= \left[X^2+Y^2+Z^2 - 2Xx - 2Yy + x^2 + y^2\right]^{1/2} && (2.26)\\ &= Z\left[1 - 2\frac{Xx}{Z^2} - 2\frac{Yy}{Z^2} + \frac{X^2+Y^2}{Z^2} + \frac{x^2+y^2}{Z^2}\right]^{1/2} && (2.27)\\ &\approx \underbrace{Z - \frac{Xx}{Z} - \frac{Yy}{Z} + \frac{X^2+Y^2}{2Z}}_{\text{Fraunhofer}} + \underbrace{\frac{x^2+y^2}{2Z}}_{\text{Fresnel}} + \ldots. && (2.28) \end{aligned}$$

2.4.1 Fresnel diffraction

If we ignore terms which are higher than quadratic in the coordinates of the screen then we have **Fresnel diffraction**. In this case the amplitude of the diffracted field is given by,

$$\begin{aligned} U(X, Y, Z) \approx \frac{1}{\mathrm{i}\lambda}\frac{\exp\left(\mathrm{i}kZ\right)}{Z}\exp\left(\mathrm{i}k\frac{X^2+Y^2}{2Z}\right)\\ \times\int_{\Sigma} E_{\text{in}}(x, y)\exp\left(\mathrm{i}k\frac{x^2+y^2}{2Z}\right)\exp\left[-\mathrm{i}k\left(\frac{X}{Z}x + \frac{Y}{Z}y\right)\right]\mathrm{d}x\mathrm{d}y. \end{aligned} \tag{2.29}$$

It is usually not possible to obtain analytic solutions to this integral, in which case it must be solved numerically. One way to do this is to notice that the integral is a Fourier transform of the input field multiplied by the function $\exp\left[\mathrm{i}\left(x^2+y^2\right)/(2Z)\right]$. However, it is often computationally more efficient to employ the angular spectrum method described in §1.2.

[2] In writing the integral in this form we have assumed $\partial E_{\text{in}}/\partial z \approx \mathrm{i}kE_{\text{in}}$. This approximation is consistent with our assumption that variations in the obliquity factor may be ignored.

2.4.2 Fraunhofer diffraction

We can ignore the terms which are of quadratic (or higher) powers of the screen coordinates (x, y) if they are *small compared to the wavelength*, i.e. if

$$\frac{x^2+y^2}{2Z} \approx \frac{a^2}{2Z} \ll \lambda \tag{2.30}$$

$$\Rightarrow Z \gg \frac{a^2}{\lambda}, \tag{2.31}$$

where a is the approximate linear dimension of the screen. This condition is sometimes written as $\mathcal{F} \ll 1$, where the **Fresnel number** $\mathcal{F} = \pi a^2/(\lambda Z)$. Note that for an aperture of 'size' a the width of the diffraction pattern is of order $Z(\lambda/a)$. Hence the Fresnel number is approximately the ratio of the size of the aperture to the width of the diffraction pattern.

In this limit we have,

$$U(X,Y,Z) \approx \frac{1}{\mathrm{i}\lambda}\frac{\exp(\mathrm{i}kZ)}{Z}\exp\left(\mathrm{i}k\frac{X^2+Y^2}{2Z}\right)\int_\Sigma E_{\mathrm{in}}(x,y)\exp\left[-\mathrm{i}k\left(\frac{X}{Z}x+\frac{Y}{Z}xy\right)\right]\mathrm{d}x\mathrm{d}y \tag{2.32}$$

$$= \frac{\sqrt{2\pi}}{\mathrm{i}\lambda}\frac{\exp(\mathrm{i}kZ)}{Z}\exp\left(\mathrm{i}k\frac{X^2+Y^2}{2Z}\right)\tilde{U}_{\mathrm{in}}\left(k\frac{X}{Z}, k\frac{Y}{Z}\right). \tag{2.33}$$

Hence the field in the observation plane is *proportional* to the Fourier transform of the field in the aperture, with spatial frequencies,

$$k_x = k\frac{X}{Z} \tag{2.34}$$

$$k_y = k\frac{Y}{Z}. \tag{2.35}$$

It is important to note that (in general) the diffracted field is not *equal* to the Fourier transform, owing to the **field-dependent phase term** $\exp\left[\mathrm{i}k\left(X^2+Y^2\right)/(2Z)\right]$.

2.4.3 Observation of Fraunhofer diffraction

We will often be interested in observing the diffraction pattern of a screen with an amplitude transmission function $t(x, y)$. The diffraction pattern will be in the Fraunhofer regime if:

1. The screen is illuminated by plane waves and the diffraction is observed in a plane sufficiently far from the diffracting aperture
2. The screen is illuminated by plane waves and the diffraction is observed in the back focal plane of a lens located after the screen
3. The diffraction is observed in the conjugate plane to the source.

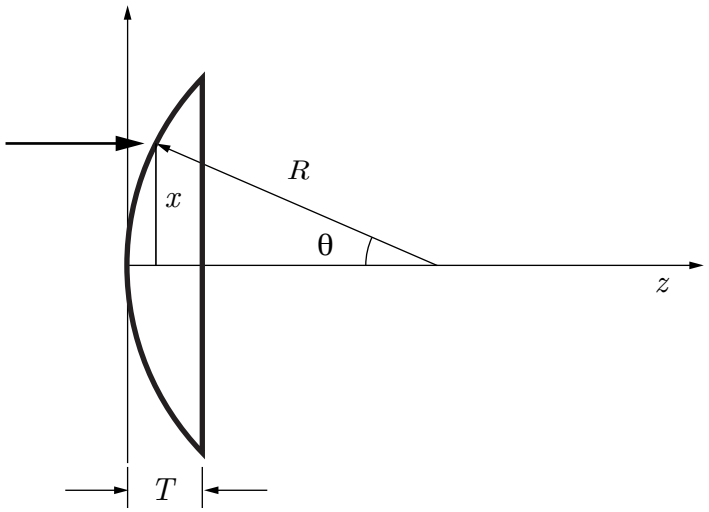


Figure 2.3: Geometry for calculating the phase shift caused by an ideal thin lens.

We have already discussed (1). In order to understand (2) and (3) we first derive the transmission function for an ideal thin lens, as illustrated in Fig. 2.3. Here, and for most of what follows, we will work on one dimension.

The phase shift imposed by the lens is,

$$\phi_{\text{lens}}(x) = k(n-1)t(x), \tag{2.36}$$

where n is the refractive index, and

$$t(x) = T - R(1-\cos\theta) \approx T - R\left[1 - \left(1 - \frac{\theta^2}{2} + \ldots\right)\right] \tag{2.37}$$

$$\approx T - \frac{x^2}{2R}, \tag{2.38}$$

where R is the radius of curvature of the lens, and we have used $\sin\theta \approx \theta \approx x/R$. We will usually be able to ignore the constant term T, in which case

$$\phi_{\text{lens}}(x) \approx -k(n-1)\frac{x^2}{2R} = -k\frac{x^2}{2f}, \tag{2.39}$$

where f is the focal length of the lens.

Aside: Does this make sense? Consider a spherical wave originating a distance u from the lens. The phase of the wave incident on the lens is then,

$$\phi_{\text{in}}(x) = k\sqrt{u^2 + x^2} \approx ku + k\frac{x^2}{2u}. \tag{2.40}$$

After propagating to the lens the phase will be,

$$\phi_{\text{out}}(x) = \phi_{\text{in}}(x) + \phi_{\text{lens}}(x) = ku + k\frac{x^2}{2u} - k(n-1)\frac{x^2}{2R}. \tag{2.41}$$

This should equal the phase of a spherical wave converging on a point a distance v after the lens:

$$\phi_{\text{out}}(x) = \text{const} - k\sqrt{v^2 + x^2} \approx -kv - k\frac{x^2}{2v} = ku + k\frac{x^2}{2u} - k(n-1)\frac{x^2}{2R} \tag{2.42}$$

$$\Rightarrow -\frac{1}{2v} = \frac{1}{2u} - \frac{n-1}{2R} \tag{2.43}$$

$$\Rightarrow \frac{1}{u} + \frac{1}{v} = \frac{n-1}{R} = \frac{1}{f}. \tag{2.44}$$

This is just the well-known lens maker's formula.

Observation of diffraction in the back focal plane of a lens

We can now consider case (2) above. Suppose that a plane wave, of amplitude U_0 illuminates a screen of amplitude transmission $t(x)$. Immediately behind this screen we place a thin lens of focal length f. From (2.29) the amplitude in the back focal plane of the lens will be,

$$U(X,f) = \frac{U_0}{\mathrm{i}\lambda}\frac{\exp(\mathrm{i}kf)}{f}\exp\left(\mathrm{i}k\frac{X^2}{2f}\right)\int\left[t(x)\exp\left(-\mathrm{i}k\frac{x^2}{2f}\right)\right]\exp\left(\mathrm{i}k\frac{x^2}{2f}\right)\exp\left[-\mathrm{i}k\left(\frac{X}{f}x\right)\right]\mathrm{d}x, \tag{2.45}$$

where the term in square brackets is the amplitude immediately after the lens. Clearly this simplifies to,

$$U(X,f) = \frac{U_0}{\mathrm{i}\lambda}\frac{\exp(\mathrm{i}kf)}{f}\exp\left(\mathrm{i}k\frac{X^2}{2f}\right)\int t(x)\exp\left[-\mathrm{i}k\left(\frac{X}{f}x\right)\right]\mathrm{d}x \tag{2.46}$$

$$= \sqrt{2\pi}\frac{U_0}{\mathrm{i}\lambda}\frac{\exp(\mathrm{i}kf)}{f}\exp\left(\mathrm{i}k\frac{X^2}{2f}\right)\tilde{t}\left(k\frac{X}{f}\right). \tag{2.47}$$

Hence the amplitude of the diffraction pattern is proportional to the Fourier transform of the transmission function of the screen.

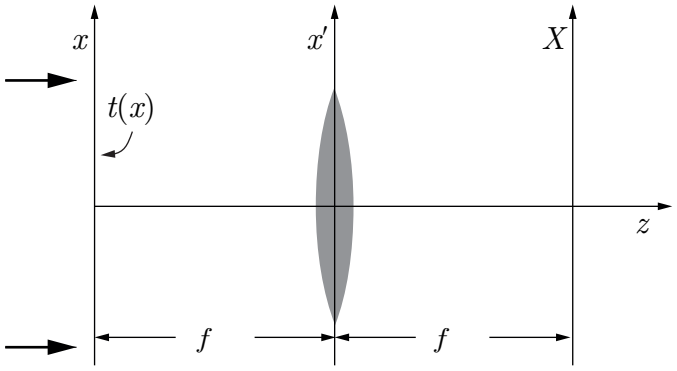


Figure 2.4: $2f$ geometry for eliminating the field-dependent phase.

Eliminating the field-dependent phase ($2f$ geometry)

Nevertheless, the diffraction pattern does still contain a field-dependent phase. Are there geometries in which this disappears? The answer is the geometry shown in Fig. 2.4. To find the amplitude in the observation screen we have to evaluate two Fresnel diffraction integrals. First we find the amplitude of the field in the plane immediately before the lens:

$$U(x',f) = \frac{U_0}{\mathrm{i}\lambda}\frac{\exp(\mathrm{i}kf)}{f}\exp\left(\mathrm{i}k\frac{x'^2}{2f}\right)\int t(x)\exp\left(\mathrm{i}k\frac{x^2}{2f}\right)\exp\left[-\mathrm{i}k\left(\frac{x'}{f}x\right)\right]\mathrm{d}x. \tag{2.48}$$

We then multiply this amplitude by the transmission function of the lens, and perform a second Fresnel diffraction integral to the observation plane:

$$\begin{aligned}
U(X,2f) &= \frac{1}{\mathrm{i}\lambda}\frac{\exp(\mathrm{i}kf)}{f}\exp\left(\mathrm{i}k\frac{X^2}{2f}\right)\int U(x',f)\exp\left(-\mathrm{i}k\frac{x'^2}{2f}\right)\exp\left(\mathrm{i}k\frac{x'^2}{2f}\right)\exp\left[-\mathrm{i}k\left(\frac{X}{f}x'\right)\right]\mathrm{d}x' \\
&= \frac{1}{\mathrm{i}\lambda}\frac{\exp(\mathrm{i}kf)}{f}\exp\left(\mathrm{i}k\frac{X^2}{2f}\right)\int U(x',f)\exp\left[-\mathrm{i}k\left(\frac{X}{f}x'\right)\right]\mathrm{d}x'' \\
&= \frac{U_0}{\lambda^2}\frac{\exp(\mathrm{i}2kf)}{f^2}\exp\left(\mathrm{i}k\frac{X^2}{2f}\right) \\
&\times\int\left\{\exp\left(\mathrm{i}k\frac{x'^2}{2f}\right)\int t(x)\exp\left(\mathrm{i}k\frac{x^2}{2f}\right)\exp\left(-\mathrm{i}k\frac{x'x}{f}\right)\mathrm{d}x\right\}\exp\left[-\mathrm{i}k\left(\frac{Xx'}{f}\right)\right]\mathrm{d}x' \\
&= \frac{U_0}{\lambda^2}\frac{\exp(\mathrm{i}2kf)}{f^2}\exp\left(\mathrm{i}k\frac{X^2}{2f}\right) \\
&\times\int\left\{\int\exp\left(\mathrm{i}k\frac{x'^2}{2f}\right)\exp\left(-\mathrm{i}k\frac{x'x}{f}\right)\exp\left[-\mathrm{i}k\left(\frac{Xx'}{f}\right)\right]\mathrm{d}x'\right\}t(x)\exp\left(\mathrm{i}k\frac{x^2}{2f}\right)\mathrm{d}x.
\end{aligned} \tag{2.49}$$

The integrals may be written as,

$$\begin{aligned}\mathcal{I} &= \int \left\{ \int \exp\left[\mathrm{i}\frac{k}{2f}\left(x'^2 - 2x'x - 2Xx'\right)\right] \mathrm{d}x' \right\} t(x) \exp\left(\mathrm{i}k\frac{x^2}{2f}\right) \mathrm{d}x \\ &= \int \left\{ \int \exp\left[\mathrm{i}\frac{k}{2f}\left(x' - (X+x)\right)^2\right] \mathrm{d}x' \right\} t(x) \exp\left(\mathrm{i}k\frac{x^2}{2f}\right) \exp\left[-\mathrm{i}\frac{k}{2f}\left(X+x\right)^2\right] \mathrm{d}x.\end{aligned}$$

Now,

$$\int_{-\infty}^{\infty} \exp\left(\mathrm{i}\beta x^2\right) \mathrm{d}x = \sqrt{\frac{\pi}{-\mathrm{i}\beta}}. \tag{2.50}$$

Hence,

$$\begin{aligned}\mathcal{I} &= \sqrt{\frac{\lambda f}{-\mathrm{i}}} \int t(x) \exp\left(\mathrm{i}k\frac{x^2}{2f}\right) \exp\left[-\mathrm{i}\frac{k}{2f}\left(X+x\right)^2\right] \mathrm{d}x \\ &= \sqrt{\frac{\lambda f}{-\mathrm{i}}} \exp\left(-\mathrm{i}\frac{kX^2}{2f}\right) \int t(x) \exp\left(-\mathrm{i}k\frac{X}{f}x\right) \mathrm{d}x.\end{aligned}$$

Thus we find, at last,

$$\begin{aligned}U(X,2f) &= \frac{U_0}{\lambda^2}\frac{\exp\left(\mathrm{i}2kf\right)}{f^2} \exp\left(\mathrm{i}k\frac{X^2}{2f}\right) \sqrt{\frac{\lambda f}{-\mathrm{i}}} \exp\left(-\mathrm{i}\frac{kX^2}{2f}\right) \int t(x) \exp\left(-\mathrm{i}k\frac{X}{f}x\right) \mathrm{d}x \\ &= \sqrt{\frac{2\pi f}{-\mathrm{i}\lambda^3}} U_0 \frac{\exp\left(\mathrm{i}2kf\right)}{f^2} \tilde{t}\left(k\frac{X}{f}\right).\end{aligned} \tag{2.51}$$

We see that with this geometry the diffracted amplitude is a true Fourier transform (multiplied by some constants), *with no field-dependent phase.*

Example of observation in the conjugate plane to the source

Finally we consider an *example* of diffraction observed in the conjugate plane to the source. Let a point source be located a distance u from a screen, immediately behind which is placed a lens of focal length f. We will view the diffraction pattern in the plane $z = u + v$. If we write the amplitude of the spherical wave as $U_0 \exp(\mathrm{i}kr)/r$, then immediately behind the lens we have, from (1.59),

$$U(x,u) = U_0 \frac{\exp(\mathrm{i}ku)}{u} \exp\left(\mathrm{i}k\frac{x^2}{2u}\right) t(x) \exp\left(-\mathrm{i}k\frac{x^2}{2f}\right). \tag{2.52}$$

We can then use a Fresnel diffraction integral to find the field in the observation plane:

$$
\begin{aligned}
U(X, u+v) &= \frac{1}{\mathrm{i}uv\lambda} \exp\left[\mathrm{i}k(u+v)\right] \exp\left(\mathrm{i}k\frac{X^2}{2v}\right) \int U(x,u) \exp\left[\left(\mathrm{i}k\frac{x^2}{2v}\right)\right] \exp\left(-\mathrm{i}k\frac{Xx}{v}\right) \mathrm{d}x \\
&= \frac{U_0}{\mathrm{i}uv\lambda} \exp\left[\mathrm{i}k(u+v)\right] \exp\left(\mathrm{i}k\frac{X^2}{2v}\right) \int t(x) \exp\left[\mathrm{i}\frac{kx^2}{2}\left(\frac{1}{u}+\frac{1}{v}-\frac{1}{f}\right)\right] \exp\left(-\mathrm{i}k\frac{Xx}{v}\right) \mathrm{d}x \\
&= \sqrt{2\pi}\frac{U_0}{\mathrm{i}uv\lambda} \exp\left[\mathrm{i}k(u+v)\right] \exp\left(\mathrm{i}k\frac{X^2}{2v}\right) \quad \tilde{t}\left(k\frac{X}{v}\right). \qquad (2.53)
\end{aligned}
$$

We see that if we observe in the conjugate (i.e. image) plane of the source, such that where $1/u+1/v = 1/f$, the quadratic phase terms in the integrand vanish, and once again the observed amplitude is proportional to the Fourier transform of the transmission function (albeit with a field-dependent phase).

Lecture 3

Ray transfer matrices

Recommended reading: Brooker, Ch 7; Siegman Ch 15.

Ray transfer matrices provide a simple method for calculating the propagation of a ray through a sequence of optical elements. More importantly, however, the same matrices can be used to:

- Calculate how spherical waves propagate through an optical system
- Calculate how Gaussian beams propagate through an optical system
- Find the modes of a laser cavity.

3.1 Ray propagation

Consider the propagation of a ray between the planes $z = z_1$ and $z = z_2$. In order to define the ray (in one transverse dimension) we need only specify the transverse position of the ray and its propagation angle, θ, with respect to the z-axis. Referring to Fig. 3.1(a) we see that,

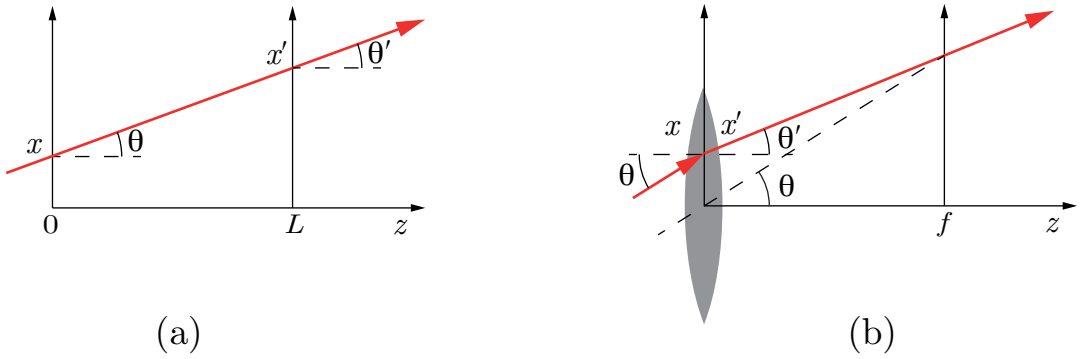


Figure 3.1: Geometries used to find the ray transfer matrix for: (a) propagation over distance L; (b) propagation through a thin lens of focal length f.

$$x' = x + L\theta \tag{3.1}$$
$$\theta' = \theta. \tag{3.2}$$

We can write these equations in matrix form as,

$$\begin{pmatrix} x' \\ \theta' \end{pmatrix} = \begin{pmatrix} 1 & L \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ \theta \end{pmatrix} = \boldsymbol{M}_{\text{trans}} \begin{pmatrix} x \\ \theta' \end{pmatrix}, \tag{3.3}$$

where the ray transfer matrix for translation over a distance L is:

$$\boldsymbol{M}_{\text{trans}} = \begin{pmatrix} 1 & L \\ 0 & 1 \end{pmatrix}. \tag{3.4}$$

Similarly, the ray coordinates either side of a thin lens are related by,

$$x' = x \tag{3.5}$$
$$\theta' = \frac{f\theta - x}{f}, \tag{3.6}$$

and hence the transfer matrix for a thin lens is,

$$\boldsymbol{M}_{\text{lens}} = \begin{pmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{pmatrix}. \tag{3.7}$$

3.1.1 Describing chains of optical elements

We will frequently want to describe chains of optical elements. For example, an imaging system is comprised of a section of propagation over a distance u, propagation through a thin lens of focal length f, and propagation to the image plane a distance v beyond the lens. It is straightforward to show that in the coordinates of the rays in the image plane are related to those in the object plane by,

$$\begin{pmatrix} x \\ \theta \end{pmatrix}_{\text{image}} = \boldsymbol{M}_{\text{trans,v}} \boldsymbol{M}_{\text{lens}} \boldsymbol{M}_{\text{trans,u}} \begin{pmatrix} x \\ \theta \end{pmatrix}_{\text{object}}. \tag{3.8}$$

In other words, the transfer matrix for a chain of optical elements is obtained by multiplying the matrices for each element in the reverse order in which the light encounters them.

Optional exercise: Find the transfer matrix for the imaging system described above. Show that the position, x_{image}, in the image plane of a ray leaving a point x_{obj} in the object plane is independent of the angle at which it leaves the object plane. Use the transfer matrix for the system to show that the transverse magnification is given by,

$$\frac{x_{\text{image}}}{x_{\text{obj}}} = -\frac{v}{u}. \tag{3.9}$$

3.1.2 Sign conventions and reduced slopes

The two matrices we have just derived are both **unitary**, i.e. their inverse is equal to their conjugate transpose. An important consequence of this is that their determinants are equal to 1. As we will see, this is a valuable property of the transfer matrices. For example, the transfer matrix describing a chain of optical elements is equal to the product of the matrices describing each element. Since the determinant of a product of matrices is equal to the product of the determinants, the transfer matrix for the complete optical system will also be equal to unity.

However, there are situations in which the approach adopted above would lead to non-unitary matrices. These include refraction at a boundary between media of two refractive indices n_1 and n_2, and reflections from mirrors. In these cases the unitary nature transfer matrices can be maintained if θ is replaced by $n\theta$, where n is the refractive index at that point, and if the refractive index of a ray propagating towards negative z is taken to be negative.

We will not adopt this approach since the it will not be needed for the simple cases we will consider. However, it is helpful to be aware that you may find in textbooks transfer matrices which include the refractive index and/or which contain terms with different signs than presented here.

3.2 Application to spherical waves

Consider the propagation of rays through an optical system with transfer matrix $\boldsymbol{M}_{\text{syst}}$,

$$\begin{pmatrix} x' \\ \theta' \end{pmatrix} = \boldsymbol{M}_{\text{syst}} \begin{pmatrix} x \\ \theta \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ \theta \end{pmatrix}, \tag{3.10}$$

where the elements of the transfer matrix are denoted A, B, C, D. We therefore have,

$$x' = Ax + B\theta \tag{3.11}$$
$$\theta' = Cx + D\theta. \tag{3.12}$$

Now, suppose that the rays describe a spherical wave. In this case, in any plane the radius of curvature of the wave is given by $R = x/\theta$. Hence we may write,

$$R' = \frac{x'}{\theta'} = \frac{Ax + B\theta}{Cx + D\theta} = \frac{AR + B}{CR + D}. \tag{3.13}$$

This is known as the **ABCD rule**.

An easier way to remember the ABCD rule is to re-write this last result as:

$$R' = \mathcal{UF}\left[AR + B\right] \tag{3.14}$$
$$1 = \mathcal{UF}\left[CR + D\right], \tag{3.15}$$

where $\mathcal{UF}$ stands for "uninteresting factor" (it is equal to $(CR + D)^{-1}$). This last pair of equations can then be written in the suggestive form,

$$\begin{pmatrix} R' \\ 1 \end{pmatrix} = \mathcal{UF} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} R \\ 1 \end{pmatrix}. \tag{3.16}$$

We see that we can use *the same ray transfer matrices* to describe the propagation of a spherical wave of radius of curvature R. Further, eqn (3.16) leads to the ABCD rule in natural, and easily remembered way.

3.3 Application to Gaussian beams

We have just seen how to use the ray transfer matrices can be used to describe the propagation of a spherical wave through an optical system. We will see that we can use the same matrices to describe the propagation of a Gaussian beam through the system

To see this, let's compare the form of a (paraxial) spherical wave and Gaussian beam:

$$U_{\text{spherical}}(x, y, z) = U_0 \exp\left[\mathrm{i}k(z - z_0)\right] \frac{1}{R} \exp\left(\mathrm{i}k\frac{\rho^2}{2R}\right) \tag{3.17}$$
$$U_{\text{Gaussian}}(x, y, z) = U_0 \exp\left[\mathrm{i}k(z - z_0)\right] \frac{1}{q} \exp\left(\mathrm{i}k\frac{\rho^2}{2q}\right), \tag{3.18}$$

where for the Gaussian beam $q = z - z_0 - \mathrm{i}z_{\mathrm{R}}$. We see that for the Gaussian beam the complex radius q plays the same role as the radius of curvature R does for a spherical wave. *It is important to note that the Gaussian beam is entirely described by* q; once we know the value of this parameter we know the amplitude of the beam in all planes, and in particular the location and size of the beam waist.

Given the similar forms of spherical waves and Gaussian beams we expect that can find the propagation of a Gaussian beam through an optical system by using an analogous rule to eqn. (3.16):

$$\begin{pmatrix} q' \\ 1 \end{pmatrix} = \mathcal{UF} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} q \\ 1 \end{pmatrix}. \tag{3.19}$$

This is indeed the case! A very simple is the case of propagation. If in some plane a Gaussian beam has a complex radius q, then a distance L downstream this parameter will be $q' = q + L$. It is straightforward to show that the ABCD rule and the transfer matrix for propagation give the same result.

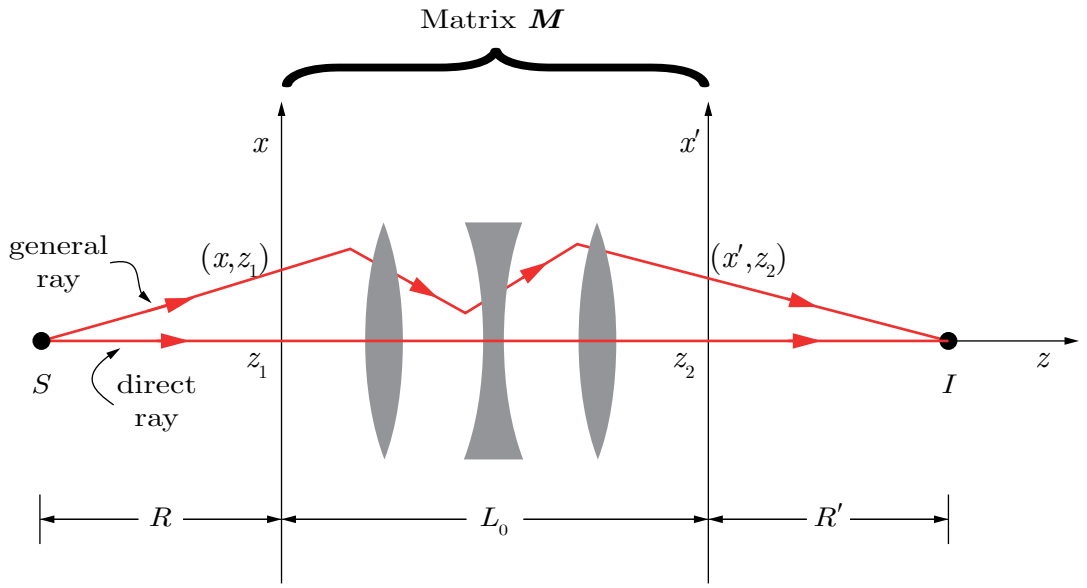


Figure 3.2: Use of the ray transfer matrices in calculating diffraction in a generalized imaging system.

Optional exercise: Consider a Gaussian beam of complex radius q incident on a thin lens. Using the phase shift introduced by a thin lens, i.e. eqn. (2.39), write down the amplitude $U(x, y, z)$ of the beam on the output surface of the lens. Show that the complex radius of the transmitted beam is given by,

$$\frac{1}{q'} = \frac{1}{q} - \frac{1}{f}. \tag{3.20}$$

Show that the ABCD rule and the transfer matrix for a thin lens give the same result.

3.4 Application to diffraction integrals

Suppose we wish to consider the propagation of waves between two planes $z = z_1$ and $z = z_2$ which lie within an optical imaging system, as illustrated in Fig. 3.2. Let the source S be located a distance R in front of z_1 and let this be imaged to the point I, which lies a distance R' beyond z_2. Suppose that the ray transfer matrix describing propagation between the planes z_1 and z_2 be $\boldsymbol{M} = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$.

From eqn. (2.24) we have,

$$U(x', z_2) = \frac{1}{\mathrm{i}\lambda} \frac{1}{|z_2 - z_1|} \int_\Sigma U(x, z_1) \exp\left[\mathrm{i}k\ell(x, x')\right] \mathrm{d}x, \tag{3.21}$$

where $\ell(x, x')$ is the optical distance taken by rays propagating between x and x'.

We consider two rays: a 'direct' ray which propagates along the axis from s to I; and a general ray which also propagates between the source and image points, but via the points (x, z_1) and (x', z_2).

The optical paths for these two rays are:

$$D_{\text{direct}} = R + L_0 + R' \tag{3.22}$$
$$D_{\text{ray}} = \sqrt{R^2 + x^2} + \ell(x, x') + \sqrt{R'^2 + x^2} \tag{3.23}$$

By Fermat's principle these two optical paths must be equal, and so we can write,

$$\sqrt{R^2 + x^2} + \ell(x, x') + \sqrt{R'^2 + x'^2} = R + L_0 + R' \tag{3.24}$$
$$\Rightarrow \ell(x, x') = R + L_0 + R' - \sqrt{R^2 + x^2} - \sqrt{R'^2 + x'^2} \tag{3.25}$$
$$= R + L_0 + R' - \left(R + \frac{x^2}{2R} + \ldots\right) - \left(R' + \frac{x'^2}{2R'}\right) \tag{3.26}$$
$$\approx L_0 - \frac{x^2}{2R} - \frac{x'^2}{2R'}. \tag{3.27}$$

Now, the rays passing through the points (x, z_1) and (x', z_2) are related by,

$$\begin{pmatrix} x' \\ \theta' \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ \theta \end{pmatrix} \tag{3.28}$$
$$\Rightarrow \begin{pmatrix} x' \\ \theta' \end{pmatrix} = \begin{pmatrix} Ax + B\theta \\ Cx + D\theta \end{pmatrix}. \tag{3.29}$$

Solving these two equations for the angles, and using the fact that $AD - BC = 1$, we find,

$$\theta = \frac{x' - Ax}{B} = \frac{x}{R} \tag{3.30}$$
$$\theta' = \frac{Dx' - x}{B} = \frac{x'}{-R'}, \tag{3.31}$$

where the far right-hand sides follow from the facts that the point (x, z_1) is illuminated by a spherical wave of radius R, and the wave leaving (x', z_2) is a spherical wave converging on I (and hence it has a radius of curvature equal to $-R'$).

Returning to eqn. (3.27) we can now write,

$$\ell(x, x') \approx L_0 - \frac{x}{2}\frac{x' - Ax}{B} + \frac{x'}{2}\frac{Dx' - x}{B} \tag{3.32}$$
$$\approx L_0 + \frac{1}{2B}\left[Ax^2 + Dx'^2 - 2xx'\right]. \tag{3.33}$$

This is an extremely powerful way of finding the optical path $\ell(x, x')$ needed in a diffraction integral.

Optional exercise: Apply this method to the geometry of Fig. 2.4. Show that in this case the transfer matrix for propagation between (x, z_1) and (x', z_2) is,

$$\boldsymbol{M} = \begin{pmatrix} 0 & f \\ -\frac{1}{f} & 0 \end{pmatrix}. \tag{3.34}$$

Hence show that the observed diffraction will occur in the Fraunhofer regime, and that there will be no field-dependent phase.

Lecture 4

Optical cavities and resonators

Recommended reading: Brooker, Ch 8; Hooker & Webb Ch 6; Siegman Ch 14.

Here we explore the properties of optical cavities. We first outline the derivation of higher-order Gaussian beams, then see how these higher-order beams can be used to find the properties of optical cavities.

4.1 Gauss-Hermite beams

We will seek solutions to the paraxial wave equation (1.41):

$$\nabla_\perp^2 u' + 2ik\frac{\partial u'}{\partial z} = 0, \tag{4.1}$$

and in particular solutions which can be written as a product of two independent functions of (x, z) only and (y, z) only, i.e. in the form $u'(x, y, z) = u_m(x, z)u_n(y, z)$. In this case the paraxial wave equation becomes,

$$\frac{1}{u_m}\left[\frac{\partial^2 u_m}{\partial x^2} + 2\mathrm{i}k\frac{\partial u_m}{\partial z}\right] + \frac{1}{u_n}\left[\frac{\partial^2 u_n}{\partial y^2} + 2\mathrm{i}k\frac{\partial u_n}{\partial z}\right] = 0. \tag{4.2}$$

Since we want $u_m(x, z)$ and $u_n(y, z)$ to be independent of each other, each must satisfy an equation of the form

$$\frac{\partial^2 u_m}{\partial x^2} + 2\mathrm{i}k\frac{\partial u_m}{\partial z} = 0. \tag{4.3}$$

We can look for solutions which are extensions to the lowest-order Gaussian beams we have previously derived. Hence we write the solutions in the form,

$$u_m(x,z) = A[q(z)]h_m\left[\frac{x}{p(z)}\right]\exp\left[\mathrm{i}\frac{x^2}{2q(z)}\right], \tag{4.4}$$

where $q(z)$ is as previously defined for the lowest-order Gaussian beam, and $h_n[x/p(z)]$, $p(z)$ and $A[q(z)]$ are to be determined. Substitution into (4.3), and a suitable choice of $p(z)$, shows that the h_n are Hermite polynomials. Hence, the Gauss-Hermite solutions to the paraxial wave equation can be written as,

$$\begin{aligned}\mathcal{U}(x,y,z) = U_0\exp\left(\mathrm{i}kz-\omega t\right)\frac{\exp\left(\mathrm{i}\frac{k\rho^2}{2q}\right)}{q}\\ \times H_m\left(\sqrt{2}\frac{x}{w(z)}\right)H_n\left(\sqrt{2}\frac{y}{w(z)}\right)\\ \times\exp\left[-\mathrm{i}\left(m+n\right)\alpha(z)\right],\end{aligned} \tag{4.5}$$

where $q(z)$, $w(z)$ and $\alpha(z)$ are the same as for a lowest-order Gaussian beam.

Figure 4.1 shows the transverse intensity profiles of the first few Gauss-Hermite beams.

4.2 Cavity modes

4.2.1 General considerations

In a laser *oscillator* (as distinct from a laser amplifier), the gain medium is placed within an optical cavity so that waves propagating in the cavity can build up in amplitude by repeatedly passing through the gain medium. The optical cavity imposes boundary conditions to the Helmholtz equation, and hence we expect the allowed solutions to comprise a set of eigenfunctions. An important difference from the standard case of a closed cavity (e.g. a hollow cube with perfectly conducting walls) is that optical cavities are usually *open*, i.e. they are formed from a set of mirrors (usually two) of finite diameter. In this case the radiation can leak out of the sides of the cavity, which means that the cavities are lossy. In a laser oscillator this loss can be made up for by the gain provided by the laser medium

In seeking to find the mode of the cavity we are looking for a solution to the following problem: what are the initial field distributions $U^{(0)}(x,y,z)$ such that after a single round-trip of the optical cavity the field arriving back at the starting point, $U^{(1)}(x,y,z)$, is unchanged (apart from the expected loss in amplitude already mentioned)? Formally this problem can be approached by solving an integral of the form of (2.24):

$$U^{(1)}(X,Y,z_{\mathrm{ref}}) = \exp\left(\mathrm{i}2kL_{\mathrm{c}}\right)\int_{\mathrm{ref}}K(X,Y,x,y)u^{(0)}(x,y,z_{\mathrm{ref}})\mathrm{d}x\mathrm{d}y, \tag{4.6}$$

where L_{c} is the cavity length and $K(X,Y,x,y)$ is the **propagation kernel**. For example, for the simple case of propagation through free space,

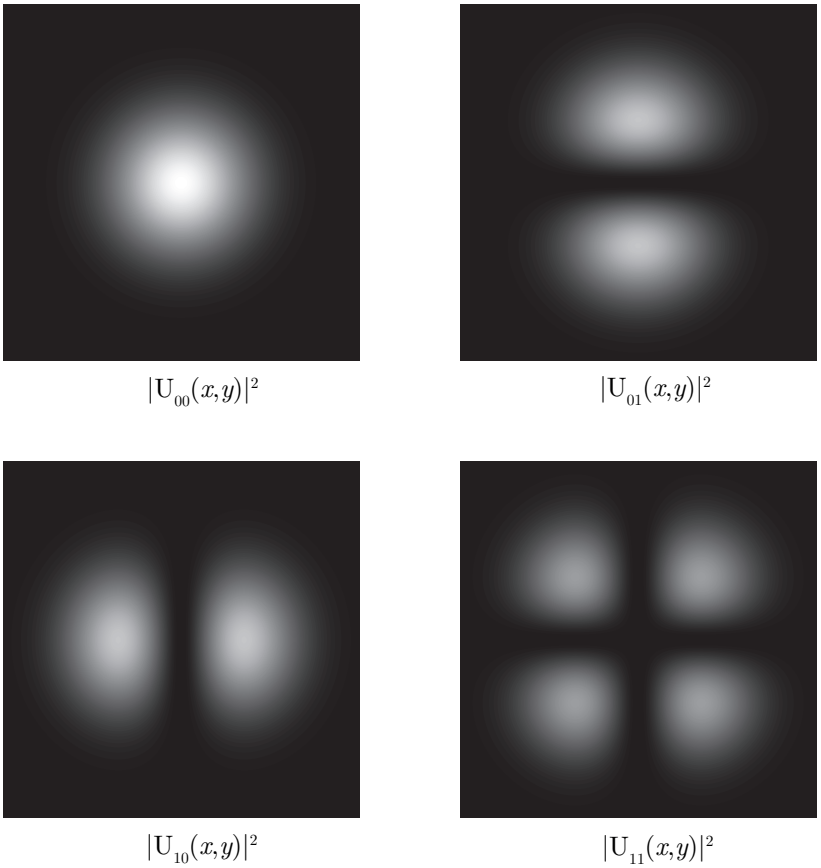


Figure 4.1: Transverse intensity profiles $|U_{mn}(x,y)|^2$ of the first few Gauss-Hermite modes.

$$\exp\left(\mathrm{i}2kL_\mathrm{c}\right) K_{\mathrm{free-space}}(X,Y,x,y) = \frac{1}{\mathrm{i}\lambda L_\mathrm{c}} \exp\left(\mathrm{i}kr_{01}\right). \tag{4.7}$$

It is usual, as we have done above, to the separate out the on-axis phase shift ($2kL_\mathrm{c}$) arising from propagation along the z-axis. this will allow us to concentrate on the **transverse modes**, i.e. the transverse field profiles which reproduce themselves after one round-trip. The transverse modes are then given by those fields $u_{nm}(x,y)$ which satisfy,

$$\int_{\mathrm{ref}} K(X,Y,x,y) u_{nm}(x,y) \mathrm{d}x\mathrm{d}y = \gamma_{nm} u_{nm}(X,Y), \tag{4.8}$$

where the factor γ_{nm} describes the round-trip losses.

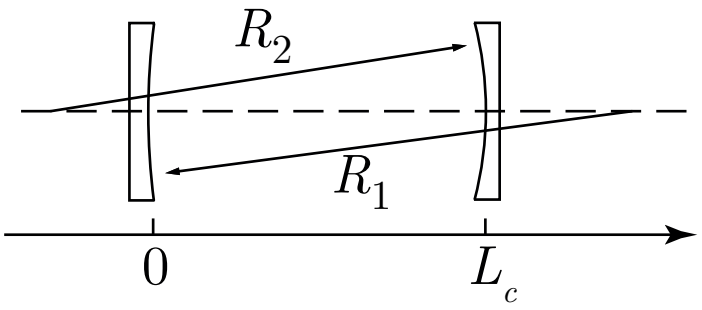


Figure 4.2: A typical low-loss laser cavity.

Comment: For a closed cavity the wave equation can be written in terms of a hermitian operator and hence it can be shown that there exists a complete set of normal modes. However, for open cavities the propagation kernel is often not hermitian, and as such there is no guarantee that there exists a complete set of orthogonal modes.
Very often the modes of an optical cavity are **biorthogonal** to an adjoint set of modes, which we call $E^{\dagger}_{nm}(x,y)$. Then,

$$\int\int u_{nm}(x,y)u^{\dagger}_{pq}(x,y) = \delta_{nm}\delta_{pq}. \tag{4.9}$$

The adjoint functions $u^{\dagger}_{nm}(x,y)$ usually represent the transverse modes travelling in the opposite direction.

4.2.2 Condition for a low-loss cavity

We now turn to the question of the modes of a typical laser cavity, such as that shown in Fig. 4.2, which comprises a pair of curved mirrors separated by a distance $L_{\rm c}$. The cavity modes can be written in the form of the Gauss-Hermite beams since these beams have a radius of curvature given by eqn. (1.63):

$$R(z) = \frac{(z-z_0)^2 + z_{\rm R}^2}{z}.$$

We can therefore match the Hermite-Gauss beams to the cavity by demanding that, at the cavity mirrors, the radii of curvature of the right- and left-going waves in the cavity match the radii of curvature of mirrors. This yields,

$$-R_1 = z_1 + \frac{z_\mathrm{R}^2}{z_1} \tag{4.10}$$

$$R_1 = z_2 + \frac{z_\mathrm{R}^2}{z_2} \tag{4.11}$$

$$L_\mathrm{c} = z_2 - z_1. \tag{4.12}$$

The minus sign in the first equation arises from the fact that (as is standard practice in this context) the radius of curvature of a mirror is taken to be positive if it is convex. The above equations can be solved to find the Rayleigh range of the beam:

$$z_\mathrm{R}^2 = \frac{g_1 g_2 (1 - g_1 g_2)}{g_1(1 - g_2) + g_2(1 - g_1)} L_\mathrm{c}^2, \tag{4.13}$$

where,

$$g_1 = 1 - \frac{L_\mathrm{c}}{R_1} \tag{4.14}$$

$$g_2 = 1 - \frac{L_\mathrm{c}}{R_2}. \tag{4.15}$$

For the modes to be low-loss their spot size must be real, so that the field amplitude decreases (approximately exponentially) with radial distance from the axis. This means that the Rayleigh range Z_R must be positive, and hence we have the **condition for a low-loss cavity**:

$$0 < g_1 g_2 < 1. \tag{4.16}$$

Figure 4.3 illustrates this condition. This condition is sometimes said to be the condition for the optical cavity to be 'stable'. This is an unhelpful term since it suggests that a cavity obeying this condition is resistant to perturbations, whereas one which does not would be particularly sensitive. This is not the case. Equation (4.16) gives the condition for the cavity to exhibit low losses from diffraction of the beam as it propagates between the two mirrors; it says nothing about the sensitivity of the cavity to perturbations or mis-alignment of the mirrors.

4.2.3 Alternative derivation of the condition for a low-loss cavity

The low-loss cavity condition can also be derived using the ray transfer matrices discussed previously. To do this we use matrices to propagate an arbitrary Gaussian beam of unknown complex radius q. We then demand that q is unchanged by a single trip around the cavity, and that q is complex.

The easiest way to perform this calculation is to unfold the cavity of Fig. 4.4(a) to give the optical system shown in Fig. 4.4(b) in which the mirrors of curvature R_1 and R_2 are replaced by lenses of focal lengths $f_1 = R_1/2$ and $f_2 = R_2/2$. The transfer matrix for a round-trip is,

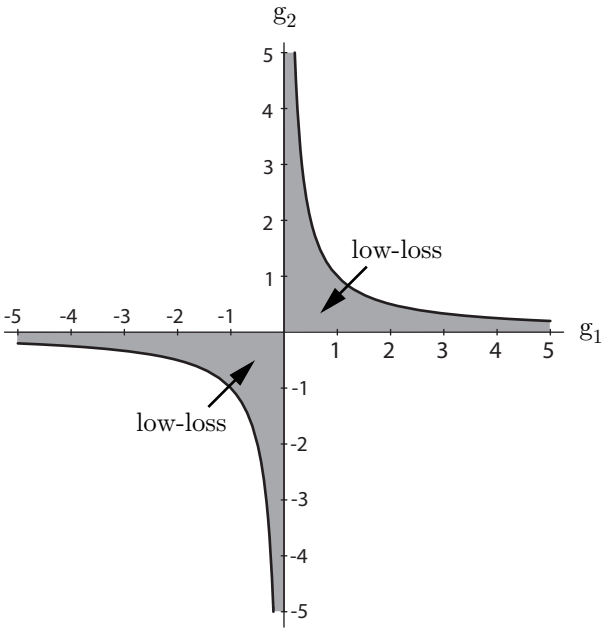


Figure 4.3: The condition for a laser cavity to be low-loss.

$$\boldsymbol{M} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\frac{1}{f_1} & 1 \end{pmatrix} \begin{pmatrix} 1 & L_\mathrm{c} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\frac{1}{f_2} & 1 \end{pmatrix} \begin{pmatrix} 1 & L_\mathrm{c} \\ 0 & 1 \end{pmatrix} \tag{4.17}$$

$$= \begin{pmatrix} 1 & L_\mathrm{c} \\ -\frac{1}{f_1} & 1 - \frac{L_\mathrm{c}}{f_1} \end{pmatrix} \begin{pmatrix} 1 & L_\mathrm{c} \\ -\frac{1}{f_2} & 1 - \frac{L_\mathrm{c}}{f_2} \end{pmatrix} \tag{4.18}$$

$$= \begin{pmatrix} 1 - \frac{L_\mathrm{c}}{f_2} & B \\ -\frac{1}{f_1} - \frac{1}{f_2} + \frac{L_\mathrm{c}}{f_1 f_2} & -\frac{L_\mathrm{c}}{f_1} + \left[1 - \frac{L_\mathrm{c}}{f_1}\right]\left[1 - \frac{L_\mathrm{c}}{f_2}\right] \end{pmatrix}, \tag{4.19}$$

in which we've not evaluated B for the reasons given below.

Using the ABCD rule and demanding that the complex radius is unchanged after one round-trip gives,

$$q = \frac{Aq + B}{Cq + D} \tag{4.20}$$

$$\Rightarrow q = \frac{(A - D) \pm \sqrt{(A + D)^2 - 4}}{2C}, \tag{4.21}$$

where we have used the fact that the transfer matrices are unitary. For a low-loss mode the beam must be a Gaussian beam (not a spherical wave), and hence q must be complex. Hence, for a low-loss mode we require,

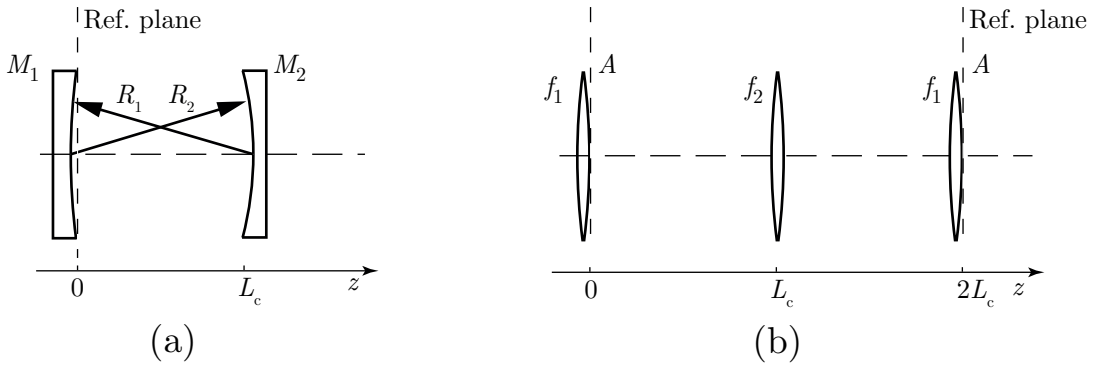


Figure 4.4: An optical cavity (a) and its unfolded equivalent (b).

$$(A + D)^2 < 4. \tag{4.22}$$

Using the matrix evaluated above, and tidying, we find this condition yields eqn (4.16) as expected.

4.3 Longitudinal and transverse modes

The analyses above provide us with conditions the cavity has to satisfy in order for low-loss modes to exist. However eqn. (4.16) does not say anything about the transverse profile of these modes, or their allowed frequencies.

4.3.1 Transverse modes

The shape of the modes can be found by solving eqn. (4.21) for q. To see this, consider the case of a near-hemispherical cavity, i.e. one in which $r_1 \to \infty$ and a cavity length $L_c = R_2 - \Delta L_c$. From (4.19) we have,

$$A = D = 1 - \frac{L_c}{f_2} = 1 - \frac{2L_c}{R_2} = \frac{\Delta L_c - L_c}{\Delta L_c + L_c} \tag{4.23}$$

$$C = -\frac{1}{f_2} = -\frac{2}{R_2} = -\frac{2}{L_c + \Delta L_c}. \tag{4.24}$$

Substituting into (4.21) we find,

$$\begin{aligned} q &= \mp\frac{L_\mathrm{c}+\Delta L_\mathrm{c}}{2}\sqrt{\left(\frac{\Delta L_\mathrm{c}-L_\mathrm{c}}{\Delta L_\mathrm{c}+L_\mathrm{c}}\right)^2-1} \\ &= \mp\frac{1}{2}\sqrt{(\Delta L_\mathrm{c}-L_\mathrm{c})^2-(\Delta L_\mathrm{c}+L_\mathrm{c})^2} \\ &= \mp\frac{1}{2}\sqrt{2\Delta L_\mathrm{c}\,(-2L_\mathrm{c})} \\ &= \mp\mathrm{i}\sqrt{L_\mathrm{c}\Delta L_\mathrm{c}}. \end{aligned} \tag{4.25}$$

Now, $q = z - z_0 - \mathrm{i}z_\mathrm{R}$. Hence we see that the location of the beam waist occurs at $z_0 = 0$, i.e. the waist is located at the reference plane, which we chose to be the location of the plane mirror. We therefore reach the reasonable conclusion that the cavity mode has a waist located at the surface of the plane mirror.

Comparing the real parts we find (taking the positive root),

$$\begin{aligned} z_\mathrm{R} &= \frac{\pi w_0^2}{\lambda} = \sqrt{L_\mathrm{c}\Delta L_\mathrm{c}} \\ \Rightarrow w_0^2 &= \frac{L_\mathrm{c}\lambda}{\pi}\sqrt{\frac{\Delta L_\mathrm{c}}{L_\mathrm{c}}}. \end{aligned} \tag{4.26}$$

Now that we know q in one plane we can evaluate it at any other plane. For example, on mirror 2 it is $q_2 = L_\mathrm{c} - \mathrm{i}z_\mathrm{R}$. The spot size on mirror 2 is given by eqn. (1.64):

$$\begin{aligned} w_2^2 &= w_0^2\left[1+\left(\frac{L_\mathrm{c}}{z_\mathrm{R}}\right)^2\right] \\ &= \frac{L_\mathrm{c}\lambda}{\pi}\sqrt{\frac{\Delta L_\mathrm{c}}{L_\mathrm{c}}}\left(1+\frac{L_\mathrm{c}^2}{L_\mathrm{c}\Delta L_\mathrm{c}}\right) \\ &\approx \frac{L_\mathrm{c}\lambda}{\pi}\sqrt{\frac{\Delta L_\mathrm{c}}{L_\mathrm{c}}}\frac{L_\mathrm{c}}{\Delta L_\mathrm{c}} \\ &\approx \frac{L_\mathrm{c}\lambda}{\pi}\sqrt{\frac{L_\mathrm{c}}{\Delta L_\mathrm{c}}}. \end{aligned} \tag{4.27}$$

Note that the spot sizes at the two mirrors can be controlled by changing the mirror spacing, i.e. changing ΔL_c.

It is worth pointing out that we could, of course, have chosen a different reference plane for our calculation. In this case the matrix $\boldsymbol{M}$ would have been different, as would have the solution for q. This makes sense since q changes with position within the cavity. What would not change, however, would be the deduced position of the beam waist, or the waist size — or indeed the properties of the mode at any point within the cavity.

Now that we have found q, we have found the complete family of Hermite-Gauss modes which are supported by the cavity. We will usually want to restrict laser oscillation to just the lowest-order

transverse mode. This can be done by inserting an aperture in the cavity which will introduce higher round-trip losses for higher order modes. By adjusting the diameter of the aperture suitably it is often possible to restrict oscillation to the lowest-order transverse mode only.

4.3.2 Longitudinal modes

Now that we have found the modes supported by the cavity, what are the allowed frequencies? For the circulating fields to build up, on returning from a cavity round trip the field must have the same transverse profile (which determines the transverse modes) *and phase*. This second requirement determines the allowed frequencies, as we shall see.

From eqn. (4.5), and recognizing the the fact of $1/q$ contains a further factor of α (see eqn. (1.68)), we have the requirement,

$$\phi_{\mathrm{round-trip}} = 2\left[kL_{\mathrm{c}} + (m+n+1)\left(\alpha_2 - \alpha_1\right)\right] = 2\pi p,$$

where $\alpha_{1,2}$ is the phase at the two mirrors, given by eqn. (1.67), and p is an integer. Assuming an empty cavity, i.e. that the refractive index throughout the cavity is unity, we can then find the allowed frequencies as follows:

$$\begin{aligned} kL_{\mathrm{c}} &= p\pi - (m+n+1)\left(\alpha_2 - \alpha_1\right) \\ \Rightarrow \nu_p &= \frac{c}{2L_{\mathrm{c}}}\left[p - (m+n+1)\,\frac{(\alpha_2 - \alpha_1)}{\pi}\right], \end{aligned} \tag{4.28}$$

where ν_p is the frequency (not angular frequency) of the cavity mode (p, m, n).

The integer p indexes the **longitudinal mode**. It is important to realize that in a laser m and n are usually small, since we will restrict laser oscillation to the lowest-order (or lowest few) mode(s). However, p is usually large since $\nu_p = c/\lambda_p$ and hence $p \approx 2L_{\mathrm{c}}/\lambda_p$; e.g. for $L_{\mathrm{c}} = 1\,\mathrm{m}$, $\lambda_p = 500\,\mathrm{nm}$ we find $p \approx 4 \times 10^6$.

The difference in frequency for adjacent longitudinal modes (keeping the transverse mode fixed) is known as the **longitudinal mode spacing**:

$$\Delta\nu_{p,p-1} = \frac{c}{2L_{\mathrm{c}}}. \tag{4.29}$$

This is of the order of hundreds of MHz for metre-scale cavities. In contrast, the difference in frequency for adjacent transverse modes is much smaller. For a given longitudinal mode p, changes in the transverse mode numbers will lead to a fractional shift in the frequency of order $1/p$; this shift is typically in the range of hundreds of Hz.

Lecture 5

Waveguides

Recommended reading: Brooker, Ch 14.

5.1 Geometric optics considerations

Consider a ray of light incident at an angle α to the front face of the optical fibre shown in Fig. 5.1. The ray will experience total internal reflection at the boundary between the core, of refractive index n_1, and the cladding, of refractive index n_2, if $n_1 > n_2$ and if,

$$\begin{aligned}
\frac{n_1 \sin\theta_1}{n_2} &> 1 \\
\Rightarrow \cos\beta &> \frac{n_2}{n_1} \\
\Rightarrow 1 - \sin^2\beta &> \left(\frac{n_2}{n_1}\right)^2 \\
\Rightarrow 1 - \frac{n_0^2 \sin^2\alpha}{n_1^2} &> \left(\frac{n_2}{n_1}\right)^2 \\
\Rightarrow n_0 \sin\alpha &< \mathrm{NA} \equiv \sqrt{n_1^2 - n_2^2},
\end{aligned} \tag{5.1}$$

where we have used the relation $n_0 \sin\alpha = n_1 \sin\beta$, and NA is the numerical aperture of the fibre.

In most cases the difference between the refractive indices of the core and cladding is small, in which case it is helpful to define,

$$\Delta = \frac{n_1 - n_2}{n_2}. \tag{5.2}$$

We can then write,

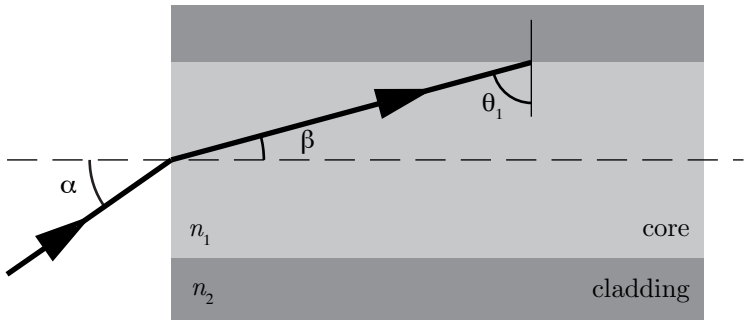


Figure 5.1: Schematic diagram of a step-index fibre showing the geometry used to calculate its numerical aperture.

$$\begin{aligned} \mathrm{NA} &= \sqrt{(n_1 - n_2)\,(n_1 + n_2)} \\ &\approx n_2\sqrt{2\Delta}. \end{aligned} \tag{5.3}$$

Typical values are $\Delta = 0.01$ and $n_2 = 1.4$, from which we find NA $= 0.2$ and $\alpha < 11°$.

These considerations show us that light propagating sufficiently close to the waveguide axis will propagate with low (in principle, zero) loss. However, they give no information on key properties such as the beam profile or the propagation speed. For those we need to find the **waveguide modes**.

5.2 Scalar wave theory of fibre modes

5.2.1 General observations

To find the modes of fibre we should solve Maxwell's equations for the core and cladding regions of the fibre, and match these solutions at the boundary between them. This is not straightforward for the case of a cylindrically symmetric waveguide since the boundary conditions depend on the orientation of the fields with respect to the boundary, and this will change with the azimuthal coordinate.

We will therefore ignore such effects and instead solve the scalar wave equation for the fields. Allowing for the refractive index $n(x, y)$ of the material, which for the moment we assume is a function of the transverse coordinates, this can be written as,

$$\nabla^2 \mathcal{U} = \frac{n(x,y)^2}{c^2}\frac{\partial^2 \mathcal{U}}{\partial t^2}. \tag{5.4}$$

The Helmholtz equation for waves of angular frequency ω is then,

$$\nabla^2 U + \frac{n(x,y)^2\omega^2}{c^2}U = 0. \tag{5.5}$$

We will seek travelling wave solutions of the form $\mathcal{U}(x,y,z,t) = u(x,y)\exp\left[\mathrm{i}\left(\beta z - \omega t\right)\right]$, where we note that the **propagation constant** β may not equal the free-space value ($n_0\omega/c$, where n_0 is the refractive index on axis). Substitution in to the Helmholtz equation then gives,

$$\nabla_\perp^2 u - \beta^2 u + n(x,y)^2\frac{\omega^2}{c^2}u = 0. \tag{5.6}$$

At this point it is worth noting that the transverse dependence of the fields $u(x,y)$ satisfies an equation which has the same form as the time-independent Schrödinger equation, i.e. we can write the above as,

$$\begin{aligned} &\nabla_\perp^2 u - \beta^2 u + \frac{n_0^2\omega^2}{c^2}u + \left[n(x,y)^2 - n_0^2\right]\frac{\omega^2}{c^2}u = 0 \\ \Rightarrow &\left\{-\nabla_\perp^2 + \left[n_0^2 - n(x,y)^2\right]\frac{\omega^2}{c^2}\right\}u = \left[\frac{n_0^2\omega^2}{c^2} - \beta^2\right]u. \end{aligned} \tag{5.7}$$

We see that the term in square brackets on the left-hand side plays the role of an effective potential, and on the right-hand side the term in square brackets plays the role of an energy. We can expect, therefore, the transverse profiles of the allowed modes to resemble the wavefunctions of the equivalent potential. Further, we will find that for the modes to propagate (i.e. for β to be real), the propagation constant will be smaller than the free-space value, and hence that the right-hand side is positive.

As an example, it can be shown that for a *parabolic* **gradient refractive index fibre** (known as a 'GRIN fibre'), the transverse profiles of the modes have the same form as those of a 2D harmonic oscillator.

5.2.2 Step-index fibre modes

For the case of a step-index fibre the refractive index has one of two values, and hence eqn. (5.6) takes the form,

$$\nabla_\perp^2 u - \beta^2 u + n^2\frac{\omega^2}{c^2}u = 0, \tag{5.8}$$

where $n = n_1$ or $n = n_2$. Re-writing this in cylindrical coordinates,

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial u}{\partial\rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 u}{\partial\phi^2} - \beta^2 u + n^2\frac{\omega^2}{c^2}u = 0. \tag{5.9}$$

Let us now assume a solution of the form $u(\rho,\phi) = R(\rho)\Phi(\phi)$. We can then separate the variables in the usual way to give,

$$\frac{\mathrm{d}^2\Phi}{\mathrm{d}\phi^2} = -m^2\Phi \tag{5.10}$$

$$\frac{1}{\rho}\frac{\mathrm{d}}{\mathrm{d}\rho}\left(\rho\frac{\mathrm{d}R}{\mathrm{d}\rho}\right) - \frac{m^2}{\rho^2}R + \left[n^2\frac{\omega^2}{c^2} - \beta^2\right]R = 0, \tag{5.11}$$

where m is a separation constant. We therefore have $\Phi = \exp(\mathrm{i}m\phi)$. Since the solution must be unchanged by the operation $\phi \to \phi + 2\pi$, we conclude that m must be an integer.

We are left with the radial equation, which we can write in the form,

$$\left[\rho^2 \frac{\mathrm{d}^2}{\mathrm{d}\rho^2} + \rho \frac{\mathrm{d}}{\mathrm{d}\rho} + \rho^2 k_\perp^2 - m^2\right] R = 0, \tag{5.12}$$

where the **transverse wavevector** is given by,

$$k_\perp^2 = n^2 \frac{\omega^2}{c^2} - \beta^2. \tag{5.13}$$

Equation (5.12) is Bessel's equation and hence the transverse dependence of the modes will be described by Bessel functions.

We now consider the core and cladding separately. For the core, if $k_\perp^2 = (n_1\omega/c)^2 - \beta^2 > 0$ (i.e. $k_\perp$ is real) then the solutions for the radial profile are $R(\rho) \propto J_m(k_\perp \rho)$, where J_m is Bessel function of the first kind.

For the cladding we expect the modes to have a different form, and in particular we expect the fields to decrease rapidly with radial distance. This type of solution occurs if $k_\perp^2 = (n_2\omega/c)^2 - \beta^2 < 0$. Defining $\gamma^2 = \beta^2 - (n_2\omega/c)^2 > 0$, then the solutions will be of the form $R(\rho) \propto K_m(\gamma\rho)$, where K_m is a modified Bessel function of the second kind.

Notice that for both $k_\perp$ and γ to be positive, the propagation constant must satisfy,

$$n_2 \frac{\omega}{c} < \beta < n_1 \frac{\omega}{c}$$
$$\Rightarrow k_2 < \beta < k_1 \tag{5.14}$$
$$\text{OR} \quad \frac{c}{n_1} < \frac{\omega}{\beta} < \frac{c}{n_2}, \tag{5.15}$$

where $k_{1,2} = n_{1,2}\omega/c$ are the wavevectors of plane waves propagating in the core and cladding. Hence the mode must have a propagation constant β which lies between those of plane waves in the two media, and hence it will propagate with a phase velocity $v_\mathrm{p} = \omega/\beta$ which lies between the wave velocity in the core and cladding.

At this stage the solutions can be written as,

$$\mathcal{U} = \begin{cases} U_0 J_m(k_\perp \rho) \exp(\mathrm{i}m\phi) \exp[\mathrm{i}(\beta z - \omega t)] & \rho < a \\ U_0' K_m(\gamma\rho) \exp(\mathrm{i}m\phi) \exp[\mathrm{i}(\beta z - \omega t)] & \rho > a. \end{cases} \tag{5.16}$$

Boundary conditions

We must now match the solutions in the two regions at the boundary, using the boundary conditions that the fields and their radial derivatives match at the boundary. Doing this gives,

$$U_0 J_m(k_\perp a) = U_0' K_m(\gamma a) \tag{5.17}$$

$$U_0 \frac{k_\perp}{2}\left[2J_{m-1}(k_\perp a) - \frac{2m}{k_\perp a} J_m(k_\perp a)\right] = -U_0' \frac{\gamma}{2}\left[2K_{m-1}(\gamma a) + \frac{2m}{\gamma a} K_m(\gamma a)\right], \tag{5.18}$$

where we have used the recursion relations for the Bessel functions. Eliminating U_0/U_0' then gives the **characteristic equation** for the fibre:

$$k_\perp \frac{J_{m-1}(k_\perp a)}{J_m(k_\perp a)} + \gamma \frac{K_{m-1}(\gamma a)}{K_m(\gamma a)} = 0. \tag{5.19}$$

Since $k_\perp$ and γ are both functions of β, the solutions to eqn. (5.19) gives the allowed propagation constants. For each azimuthal mode m there will in general be several radial modes, each with different β (or, equivalently, $k_\perp$) which satisfy eqn. (5.19).

Recursion relations: The following recursion relations for the Bessel functions are useful:

$$\frac{2\alpha}{x} J_\alpha(x) = J_{\alpha-1}(x) + J_{\alpha+1}(x) \tag{5.20}$$

$$\frac{\mathrm{d}J_\alpha}{\mathrm{d}x} = \frac{1}{2}\left[J_{\alpha-1}(x) - J_{\alpha+1}(x)\right] \tag{5.21}$$

$$\frac{2\alpha}{x} K_\alpha(x) = -K_{\alpha-1}(x) + K_{\alpha+1}(x) \tag{5.22}$$

$$\frac{\mathrm{d}K_\alpha}{\mathrm{d}x} = -\frac{1}{2}\left[K_{\alpha-1}(x) + K_{\alpha+1}(x)\right]. \tag{5.23}$$

Cut-off condition

Before considering in more detail the solutions to eqn. (5.19), we first suppose that we have found such a solution at, say, $k_\perp = k_\perp^{m,l}$. Now, for this mode to have low loss we need $\gamma = \sqrt{\beta^2 - (n_2\omega/c)^2} > 0$, i.e.,

$$\begin{aligned}
&\beta^2 > (n_2\omega/c)^2 \\
\Rightarrow &\left(n_1\frac{\omega}{c}\right)^2 - (k_\perp^{m,l})^2 > (n_2\omega/c)^2 \\
&\Rightarrow k_\perp^{m,l} < (\omega/c)\sqrt{n_1^2 - n_2^2} \\
&\Rightarrow k_\perp^{m,l}\, a < v = (\omega/c)\, a \sqrt{n_1^2 - n_2^2} = (\omega/c)\, a\, \mathrm{NA} \approx (\omega/c)\, a\, n_2\sqrt{2\Delta}
\end{aligned} \tag{5.24}$$

The quantity $v = (\omega/c)\, a \sqrt{n_1^2 - n_2^2}$ is known as the **v-parameter** of the fibre. The v-parameter determines the maximum number of low-loss modes the fibre can support, and it gives the maximum possible value of $k_\perp$.

An important corollary of eqn (5.24) is that, for each mode there is a **cut-off frequency** given by,

$$\omega > \omega_{\rm c} = \frac{c}{\sqrt{n_1^2 - n_2^2}} k_\perp^{m,l}. \tag{5.25}$$

The mode will not propagate (with low loss) if $\omega < \omega_{\rm c}$.

Comment: Our treatment of the boundary conditions reveals limitations of the scalar theory. Really, of course, we need to find the $\boldsymbol{E}$ and $\boldsymbol{H}$ fields of the waves and match their components across the boundary. If we were to do this we would find, to a good approximation, that this is equivalent to matching the amplitudes and radial derivatives of the scalar fields across the boundary.
For example, suppose the field is linearly polarized along the x-axis. The ϕ components of $\boldsymbol{E}$ and the radial components of $\boldsymbol{D} = \epsilon \boldsymbol{E}$ should be continuous:

$$E_1 \sin\phi = E_2 \sin\phi \tag{5.26}$$
$$\epsilon_1 E_1 \cos\phi = \epsilon_1 E_2 \cos\phi, \tag{5.27}$$

where E_1 and E_2 are the amplitudes of the fields either side of the boundary. Now, since $\epsilon_1 \approx \epsilon_2$ these two equations are equivalent to the requiring that the amplitudes of the fields are continuous across the boundary.
The magnetic field can be calculated most easily from the curl equations. In particular, for a harmonic field $\mathrm{i}\omega\boldsymbol{B} = \nabla \times \boldsymbol{E}$. The z-components of $\boldsymbol{H} = \boldsymbol{B}/\mu_0$ must be continuous across the boundary, and hence:

$$\frac{1}{\rho}\left[\frac{\partial(\rho E_\phi)}{\partial\rho} - \frac{\partial E_\rho}{\partial\phi}\right]$$

must also be continuous. Evaluating this (assuming $E_z = 0$), and using the fact that the amplitudes of the fields are continuous across the boundary, we find that $\partial E_\phi/\partial\rho$ must be continuous across the boundary, which gives the boundary condition:

$$U_0 \left.\frac{\mathrm{d}J_m(k_\perp\rho)}{\mathrm{d}\rho}\right|_{\rho=a} = U_0' \left.\frac{\mathrm{d}K_m(\gamma\rho)}{\mathrm{d}\rho}\right|_{\rho=a}, \tag{5.28}$$

i.e. that the radial derivative of the fields should be continuous.

To gain further insight into how the v-parameter limits the number of low-loss modes, we note that we can write,

$$k_\perp^2 + \gamma^2 = \left(\frac{v}{a}\right)^2. \tag{5.29}$$

Hence as $k_\perp$ increases from zero the value of γ decreases. For a mode to be low-loss γ should be positive, and hence the low-loss modes are given by solutions to eqn. (5.19) for which $k_\perp$ lies between zero and v/a.

$m = 0$ modes

It is interesting to ask if there are any modes which can propagate at *any* frequency, i.e. for which $\omega_c \to 0$. In this limit $v \to 0$, $k_\perp \to 0$, and $\gamma \to 0$. Now $K_m(x)$ remains finite as $x \to 0$, but in most cases $J_m(x)$ does *not*, which means that the boundary conditions cannot be satisfied (since the denominator in the first term of eqn. (5.19) blows up). The exception is $J_0(x)$, which *is* finite as $x \to 0$.

Hence there is a solution with $\omega_c \to 0$ for the case $m = 0$. For this case, for any frequency, there will always be at least one solution to the characteristic equation. In general there will be other solutions too, the number of these depending on the v-parameter. However, these higher-order modes will all have finite cut-off frequencies.

To give an example, suppose $n_1 = 1.5$, $\Delta = 0.002$, $a = 10\,\mu$m, $\lambda = 500\,$nm. We find $v = 9.73$ and hence the low-loss modes will correspond to solutions to eqn. (5.19) with $k_\perp a < 9.73$. For $m = 0$ eqn. (5.19) becomes,

$$k_\perp \frac{J_1(k_\perp a)}{J_0(k_\perp a)} + \gamma \frac{K_{-1}(\gamma a)}{K_0(\gamma a)} = 0, \tag{5.30}$$

where we have used the fact that for integer m, $J_{-m}(x) = J_m(x)$. The zeroes of the denominator of the first term play an important role since they mark an asymptote in the first term, and hence as one crosses this asymptote the value of the first term sweeps from $-\infty$ to $+\infty$. Now, since second term is, in general, a slowly-varying function of $k_\perp$, the solutions to the characteristic equation must occur *close* to the zeroes of $J_0(k_\perp a)$.

The zeroes of $J_0(x)$ occur at $x = 2.40, 5.52, 8.65, 11.8, \ldots$, and so for $v = 9.73$ we expect 3 solutions with $k_{\perp,0} a \approx 2.40, 5.52, 8.65$. In fact, numerical solution of eqn. (5.19) gives, for $m = 0$ solutions at $k_{\perp,0} a = 2.18, 4.98, 7.73$.

$m = 1$ modes

We will be interested in finding a condition for single-mode operation. To do this we will want to find the $m \neq 0$ mode with the lowest cut-off frequency, since then frequencies below this will only be able to propagate in the $m = 0$ mode.

For any mode, the cut-off frequency occurs at $\gamma = 0$, in which case the characteristic equation becomes,

$$k_\perp \frac{J_{m-1}(k_\perp a)}{J_m(k_\perp a)} = 0. \tag{5.31}$$

For $m \neq 0$, the solution is,

$$\begin{aligned} & J_{m-1}(k_\perp a) = 0 \\ \Rightarrow\ & k_\perp a = j_{m-1,\ell}, \end{aligned} \tag{5.32}$$

where $j_{m-1,\ell}$ is the ℓth zero of $J_{m-1}(x)$. The smallest value of $j_{m-1,\ell}$ is the first zero of $J_0(x)$, which occurs at $j_{0,1} = 2.405$. Hence the $m = 1$ mode has the smallest cut-off frequency, in which case,

$$\omega_\mathrm{c} = 2.405 \frac{c}{a\sqrt{n_1^2 - n_2^2}}. \tag{5.33}$$

Frequencies below this can only propagate in the $m = 0$ mode. We can re-express this result in terms of the vacuum wavelength λ_0 and the core radius:

$$\begin{aligned} & \frac{2\pi c}{\lambda_0} < 2.405 \frac{c}{a\sqrt{n_1^2 - n_2^2}} \\ \Rightarrow\ & a < \frac{2.405}{2\pi} \frac{\lambda_0}{\sqrt{n_1^2 - n_2^2}} \approx 0.27 \frac{\lambda}{n_2\sqrt{\Delta}}. \end{aligned} \tag{5.34}$$

We see that single-mode operation requires that the fibre radius is of order the wavelength, and that this radius can be increased if Δ is decreased.

5.3 Dispersion in waveguides

Dispersion in waveguides limits the rate at which information can be transmitted through the waveguide, and in particular it determines a **length-bandwidth product**. For example, an optical fibre might be rated as 500 Mbit s^{-1} km: this means that it could transmit information at 500 Mbit s^{-1} over a 1 km link, but only at 250 Mbit s^{-1} over a over 2 km.

The dispersion of a waveguide is determined by the frequency dependence of the propagation constant. From our analysis of optical fibres we found,

$$\beta^2 = \left(\frac{n_1\omega}{c}\right)^2 - k_\perp^2, \tag{5.35}$$

and similar relations will be found for other types of waveguide. It should be recognized that the only term in eqn. (5.35) which *does not* depend on frequency, is the speed of light c.

The frequency dependence of β can be divided into three types of dispersion:

- **Material dispersion:** the frequency dependence of the material of the fibre leads to dispersion in a familiar way. For silica — the material used in most optical fibres — the variation of the group velocity with frequency is zero at a wavelength of 1.3 μm, and hence many optical fibre systems operate at this wavelength.

- **Intermodal dispersion:** the propagation constant of a mode depends on $k_\perp$, and hence different modes have different phase and group velocities. Exciting more than one mode in a waveguide will therefore lead to intermodal dispersion.
- **Intramodal dispersion:** since in eqn. (5.35) $k_\perp$ depends on frequency, even signals propagating in a single mode will experience dispersion in addition to any material dispersion. This is known as intramodal dispersion.

Lecture 6

Interference and coherence

Recommended reading: Brooker, Ch 10; Hecht, Ch 12; Born & Wolf, Ch. 10.

6.1 Introduction

In this lecture we seek to understand, and quantify, how optical waves at different points in space and/or moments in time are *correlated*. We will try to find answers to the following questions:

- To what extent are the oscillations of the field at one point in space (or at one moment) related to the oscillations at another point (or moment)?
- Can this relation be quantified, and, if so, how may it be measured?
- Are there any practical consequences of these correlations?

6.2 General interference

We first consider a generalized interference experiment, as illustrated in Fig. 6.1. The source, which we take to have finite extent, and to emit a narrow range of frequencies, produces a field $E(\boldsymbol{r}, t)$. We will try to measure the correlations in the fields at two points, $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, by placing a screen with two pinholes located at those points. In other words we perform a Young's interference experiment.

The field measured downstream at a point P is then (in terms of the analytic signals),

$$\mathcal{U}_P(t) = K_1 \mathcal{U}_1(t - t_1) + K_2 \mathcal{U}_2(t - t_2), \tag{6.1}$$

where $\mathcal{U}_i(t) = \mathcal{U}(\boldsymbol{r}_i, t)$, and $t_i = r_i/c$ is the time for light to propagate to P from $\boldsymbol{s}_i$. The terms K_1 and K_2 are *propagators*, i.e. they describe how the incident wave is changed in passing through each pinhole and propagating to the observation point. For a pinhole they must be purely imaginary, with

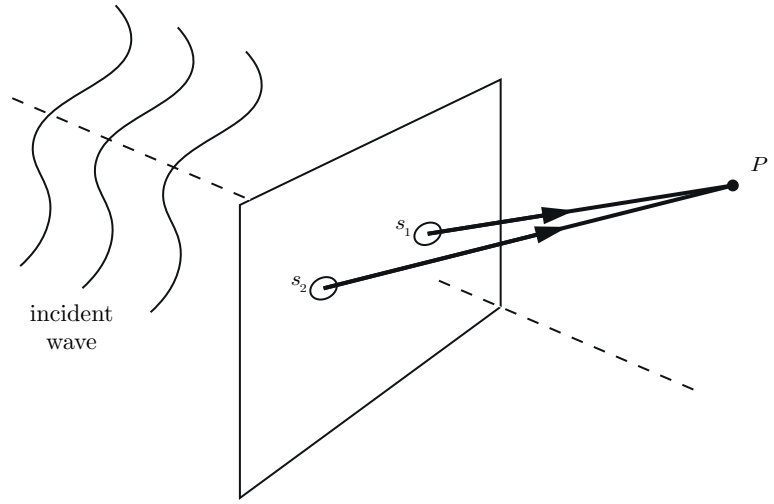


Figure 6.1: Geometry of a generalized interference experiment.

a modulus which depends inversely on the distance r_i. The intensity measured at P is proportional to the time-average of the modulus-squared of the field at P, and hence, ignoring constants,

$$\begin{aligned} I_p &= \langle \mathcal{U}_P(t)\mathcal{U}_P^*(t)\rangle \\ &= \langle [K_1\mathcal{U}_1(t-t_1) + K_2\mathcal{U}_2(t-t_2)]\,[K_1^*\mathcal{U}_1^*(t-t_1) + K_2^*\mathcal{U}_2^*(t-t_2)]\rangle \\ &= |K_1|^2 \left\langle |\mathcal{U}_1(t-t_1)|^2 \right\rangle + |K_2|^2 \left\langle |\mathcal{U}_2(t-t_2)|^2 \right\rangle \\ &\quad + K_1K_2^* \langle \mathcal{U}_1(t-t_1)\mathcal{U}_2^*(t-t_2)\rangle + K_1^*K_2 \langle \mathcal{U}_1^*(t-t_1)\mathcal{U}_2(t-t_2)\rangle\,. \end{aligned} \tag{6.2}$$

If the field is *stationary* we can shift the time: $t \to t + t_2$, so that we have,

$$I_P = |K_1|^2 \left\langle |\mathcal{U}_1(t)|^2 \right\rangle + |K_2|^2 \left\langle |\mathcal{U}_2(t)|^2 \right\rangle + 2\,|K_1|\,|K_2|\,\Re\left\{\langle \mathcal{U}_1(t+\tau)\mathcal{U}_2^*(t)\rangle\right\}, \tag{6.3}$$

where we have used the fact that $< |\mathcal{U}_1(t+\tau)|^2 >=< |\mathcal{U}_1(t)|^2 >$ and that $K_{1,2}$ are purely imaginary. The first two terms are simply the intensities measured at P when light is allowed through one or other of the pinholes; the third term describes the interference between the two waves leaving the pinholes.

We will define the **mutual coherence function** as:

$$\Gamma_{12}(\tau) = \langle \mathcal{U}_1(t+\tau)\mathcal{U}_2^*(t)\rangle\,, \tag{6.4}$$

which is the *cross-correlation* of the fields at the points $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, with a relative delay between the fields of τ. If the peaks and troughs in $\mathcal{U}_1$ coincide with those in $\mathcal{U}_2$, then $\Gamma_{1,2}(\tau)$ will be large, and the two fields are said to be *strongly correlated*. However, the peaks and troughs in $\mathcal{U}_1$ are randomly

associated with those in $\mathcal{U}_2$ then the fields will be *weakly correlated*. As we would expect, we will find that the fields need to be strongly correlated in order to see good interference fringes on the observation screen.

It is useful to normalize the mutual coherence function, to give what is known as the **complex degree of coherence**:

$$\gamma_{12}(\tau) = \frac{\langle \mathcal{U}_1(t+\tau)\mathcal{U}_2^*(t)\rangle}{\sqrt{\left\langle |\mathcal{U}_1|^2 \right\rangle \left\langle |\mathcal{U}_2|^2 \right\rangle}}. \tag{6.5}$$

It should be clear that $0 \leq |\gamma_{12}(\tau)| \leq 1$. With this definition, and writing $I_i = |K_i|^2 < |\mathcal{U}_i|^2 >$, we have,

$$I_P = I_1 + I_2 + 2\sqrt{I_1 I_2}\Re\left[\gamma_{12}(\tau)\right]. \tag{6.6}$$

This is known as the *generalized interference law for partially coherent light.*

6.2.1 Fringe visibility

We quasi-monochromatic light we can write the complex degree of coherence as $\gamma_{12}(\tau) = |\gamma_{12}(\tau)| \exp\left[\mathrm{i}\Phi(\tau)\right]$, where,

$$\Phi(\tau) = \alpha_{12}(\tau) + \phi, \tag{6.7}$$

where $\phi = (2\pi/\lambda)(r_1 - r_2)$ is the geometric phase difference. As we move position P on the observation screen the fringes will change in intensity between local maxima and minima given by,

$$I_{\max} = I_1 + I_2 + 2\sqrt{I_1 I_2}|\gamma_{12}(\tau)| \tag{6.8}$$

$$I_{\min} = I_1 + I_2 - 2\sqrt{I_1 I_2}|\gamma_{12}(\tau)|. \tag{6.9}$$

Hence the fringe visibility is given by,

$$\mathcal{V}_P = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2}|\gamma_{12}(\tau)|. \tag{6.10}$$

We see immediately, as expected, that in order to achieve high-quality fringes it is necessary for the fields to have: (i) a high value of $|\gamma_{12}(\tau)|$, i.e. for the fields to be coherent; and (ii) for the two intensities to be approximately equal.

If the fields have the same intensity (i.e. $I_1 = I_2$), then the fringe visibility is equal to $|\gamma_{12}(\tau)|$. Hence the fringe visibility is related to the coherence of the sources, but $|\gamma_{12}(\tau)|$ is a better measure since it describes the coherence of the fields in a rigorous way, and is not distorted by differences in the source intensities.

6.3 Temporal coherence

The complex degree of coherence $\gamma_{12}(\tau)$ contains the both geometric effects and the effects of the finite bandwidth of the sources. Let us now consider the case when geometric effects can be ignored, such as a Michelson interferometer illuminated by a perfectly plane wave. In this case — assuming an ideal beam-splitter — the detector will be illuminated by two copies of the incident field, one delay relative to the other by $\tau = \ell/c$, where ℓ is the path difference between the two arms of the Michelson. The recorded intensity can then be written,

$$\begin{aligned} I(\tau) &= \langle [\mathcal{U}_P(t) + \mathcal{U}_P(t+\tau)]\,[\mathcal{U}_P^*(t) + \mathcal{U}_P^*(t+\tau)] \rangle \\ &= 2\left\langle |\mathcal{U}_P(t)|^2 \right\rangle + 2\Re\left\{ \langle \mathcal{U}_P^*(t)\mathcal{U}_P(t+\tau)\rangle \right\} \\ &= 2\left\langle |\mathcal{U}_P(t)|^2 \right\rangle \left[1 + \Re\left\{ \frac{\langle \mathcal{U}_P^*(t)\mathcal{U}_P(t+\tau)\rangle}{\left\langle |\mathcal{U}_P(t)|^2 \right\rangle} \right] \right\} \\ &= 2\left\langle |\mathcal{U}_P(t)|^2 \right\rangle \left[1 + \Re\left\{ \gamma_{11}(\tau) \right\} \right], \end{aligned} \tag{6.11}$$

where $\gamma_{11}(\tau)$ is **complex degree of temporal coherence** and is defined in an analogous way to eqn. (6.5).

The phase of $\gamma_{11}(\tau)$ will be dominated by the difference in phase of the waves at t and $t+\tau$, i.e. by $\phi \approx \omega\tau$. Hence, following the reasoning above, the fringe visibility will be,

$$\mathcal{V}(\tau) = |\gamma_{11}(\tau)|. \tag{6.12}$$

Once again, this should make sense: if the fields are coherent then they will generate high-quality fringes, but the opposite will be true if the correlation between the fields will be poor.

Relation between temporal coherence and bandwidth

If the source has a very narrow bandwidth then the field it emits will be close to a sinusoidal wave. In this case, if we know the field at some time t then we can be confident in predicting the field at some time $t+\tau$, even if τ is large. We can do this since we know that field at the later time is just that at t shifted in phase by $\omega\tau$. However, if the field has a broad bandwidth then we can't be sure what to use for ω. There will therefore be some uncertainty in the fields at later times, i.e. $|\gamma_{11}(\tau)|$ will decrease more rapidly with τ.

We can be a little more quantitative. The uncertainty in the shift in phase caused by a shift in time of τ is $\Delta\omega\tau$, where $\Delta\omega$ is a measure of the bandwidth of the source. Clearly if this uncertainty is comparable to π then we've essentially lost all knowledge of the phase of the wave, and in other words the fields are not correlated. The time delay over which the fields are reasonably well correlated should therefore be approximately $\tau_c \approx \pi/\Delta\omega$. If we scan one arm of a Michelson interferometer, we expect the fringes to become poor for delays comparable to τ_c.

6.3.1 Fourier transform spectroscopy

From second year optics we know that, for a monochromatic source, the intensity recorded at the centre of the output of a Michelson interferometer can be written as,

$$I(\tau) = \frac{I(0)}{2}\left[1 + \cos\omega\tau\right], \tag{6.13}$$

where τ is the delay introduced between the two arms. For a polychromatic source we should sum intensities (since the interference terms average out on the timescale of the detector), giving,

$$\begin{aligned} I(\tau) &= \frac{1}{2}\int P(\omega)\left[1 + \cos\omega\tau\right]\mathrm{d}\omega \\ &= \frac{1}{2}\int P(\omega)\mathrm{d}\omega + \frac{1}{2}\int \left[1 + \cos\omega\tau\right]\mathrm{d}\omega \\ &= \frac{1}{2}\int P(\omega)\mathrm{d}\omega + \frac{1}{2}\int P(\omega)\cos\omega\tau\mathrm{d}\omega. \end{aligned} \tag{6.14}$$

The first term is just a constant, and the second term describes the interference fringes. It should be clear that the second term will average to zero over a reasonably large number of fringes, and hence the first term is equal to the mean signal, $\bar{I}$. We therefore have,

$$\begin{aligned} &I(\tau) = \bar{I} + \frac{1}{2}\int P(\omega)\cos\omega\tau\mathrm{d}\omega \\ &\Rightarrow \frac{1}{2}\int P(\omega)\cos\omega\tau\mathrm{d}\omega = I(\tau) - \bar{I} \\ &\Rightarrow P(\omega) = 2\int \left[I(\tau) - \bar{I}\right]\cos\omega\tau\mathrm{d}\tau. \end{aligned} \tag{6.15}$$

In other words the spectrum of the source can be determined by recording the fringe pattern of a Michelson interferometer as a function of delay τ and taking the inverse cosine transform of the recorded signal.

We can reach the same result from the perspective of the correlations between the fields. In terms of the *real* fields, the intensity output by a Michelson interferometer can be written as,

$$\begin{aligned} I(\tau) &= \left\langle \left[V(t) + V(t+\tau)\right]^2 \right\rangle \\ &= \left\langle V(t)^2 + V(t+\tau)^2 + 2V(t)V(t+\tau) \right\rangle \\ &= 2\left\langle V(t)^2 \right\rangle + 2\left\langle V(t)V(t+\tau) \right\rangle, \end{aligned} \tag{6.16}$$

where τ is the delay introduced. The first term is just the sum of the intensities; the second is the interference term, which will oscillate with τ. The first term will therefore be equal to the mean signal averaged over many fringes, which we denote $\bar{I}$. We can therefore write,

$$\langle V(t)V(t+\tau)\rangle = W(\tau) = \frac{1}{2}\left[I(\tau) - \bar{I}\right]. \tag{6.17}$$

By the Wiener-Khintchine theorem (see below), we can write:

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} W(\tau)\exp\left(+\mathrm{i}\omega\tau\right)\mathrm{d}\tau = \sqrt{2\pi}P(\omega), \tag{6.18}$$

where $P(\omega)$ is the power spectrum of the source.

Wiener-Khintchine theorem The considerations above suggest that there is a connection between the spectrum of the source and the temporal coherence. This connection is explained rigorously by the Wiener-Khintchine theorem. Let us find the Fourier transform of the correlation of the real fields:

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\langle V(t)V(t+\tau)\rangle\exp\left(+\mathrm{i}\omega\tau\right)\mathrm{d}\tau \\
&\quad= \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} V(t')V(t'+\tau)\mathrm{d}t'\right\}\exp\left(+\mathrm{i}\omega\tau\right)\mathrm{d}\tau \\
&\quad= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} V(t'+\tau)\exp\left(+\mathrm{i}\omega\tau\right)\mathrm{d}\tau\right\}V(t')\mathrm{d}t' \\
&\quad= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} V(t'')\exp\left(+\mathrm{i}\omega t''\right)\mathrm{d}t''\right\}V(t')\exp\left(-\mathrm{i}\omega t'\right)\mathrm{d}t' \\
&\quad= \sqrt{2\pi}\tilde{V}(\omega)\left\{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} V(t')\exp\left(+\mathrm{i}\omega t'\right)\mathrm{d}t'\right\}^{*} \\
&\quad= \sqrt{2\pi}\left|\tilde{V}(\omega)\right|^2 \\
&\quad= \sqrt{2\pi}P(\omega),
\end{aligned} \tag{6.19}$$

where $P(\omega)$ is the power spectrum of the source. This is the **Wiener-Khintchine theorem**, i.e. the Fourier transform of $< V(t)V(t+\tau) >$ is proportional to the power spectrum of the source.

6.3.2 Optical coherence tomography

Optical coherence tomography (OCT) uses a source with a *short coherence length* to create 3D images of transparent objects. It is in effect "optical ultrasound". It is used in a wide variety of applications, including opthalmology, cardiology, and art conservation.

The sources employed often operate in the near infra-red since these wavelengths can penetrate the object. Femtosecond lasers, and even super-luminescent diodes, are typically used.

The idea is straightforward, as illustrated in Fig. 6.2. Light from the source is directed to a Michelson interferometer, in one arm of which is placed the sample. The other arm contains a standard

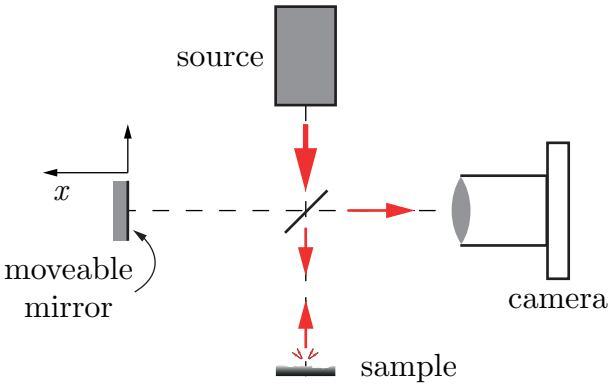


Figure 6.2: Schematic diagram of an apparatus for Optical Coherence Tomography.

mirror, which is scanned longitudinally. Light penetrating the sample will be partially reflected from any boundaries between regions of different refractive index. As such the beam returning along this arm of the Michelson will comprise many, low-amplitude copies of the incident wave which are shifted in time by t_i, where $v_\mathrm{g}\mathrm{t}_i$ is the depth of the layer and v_g is the group velocity in the sample. As the other arm of the Michelson is scanned, high-visibility fringes will occur at mirror positions $x_i = x_0 + v_\mathrm{g}\mathrm{t}_i$, where x_0 is an offset. In a straightforward way, therefore, mapping the positions of high-visibility fringes gives the longitudinal locations of layers in the sample. In the longitudinal direction the resolution will be approximately equal to the longitudinal coherence length $l_\mathrm{c} = c\tau_\mathrm{c} \approx \pi c/\Delta\omega$.

6.4 Transverse coherence

6.4.1 The van Cittert-Zernike theorem

We now seek to determine the complex degree of coherence γ_{12} at a large distance from a *fully incoherent* source. To do this we divide the source into elements $\mathrm{d}\sigma_1, \mathrm{d}\sigma_2, \mathrm{d}\sigma_3, \ldots$ and assume that each element radiates a quasi-monochromatic wave with no statistical correlation with the waves emitted by other parts of the source.

Let us suppose that we measure γ_{12} at two points P_1 and P_2. The fields at each point can be written as,

$$\mathcal{U}_1(t) = \sum_m \mathcal{U}_{m1}(t) \tag{6.20}$$

$$\mathcal{U}_2(t) = \sum_m \mathcal{U}_{m2}(t), \tag{6.21}$$

where $\mathcal{U}_{mi}(t)$ is the field at P_i radiated by $\mathrm{d}\sigma_i$.

We are interested in finding,

$$\begin{aligned}\Gamma_{12} &= \langle \mathcal{U}_1(t)\mathcal{U}_2^*(t)\rangle = \left\langle \sum_m \mathcal{U}_{m1}(t) \sum_n \mathcal{U}_{n2}^*(t) \right\rangle \\ &= \sum_m \langle \mathcal{U}_{m1}(t)\mathcal{U}_{m2}^*(t)\rangle + 2\sum_{m\neq n}\sum_n \langle \mathcal{U}_{m1}(t)\mathcal{U}_{n2}^*(t)\rangle \\ &= \sum_m \langle \mathcal{U}_{m1}(t)\mathcal{U}_{m2}^*(t)\rangle\,, \end{aligned} \tag{6.22}$$

where the last line follows from the fact that the fields radiated from different elements are statistically independent.

Now the field reaching P_i from element $\mathrm{d}\sigma_m$ can be written in the form,

$$\mathcal{U}_{mi} = \frac{1}{r_{mi}} A_m\left(t - \frac{r_{mi}}{c}\right) \exp\left[-\mathrm{i}\bar{\omega}\left(t - \frac{r_{mi}}{c}\right)\right], \tag{6.23}$$

where $A_m(t)$ is the amplitude of the wave emitted by $\mathrm{d}\sigma_m$ at time t, r_{mi} is the distance from the element to P_i, and $\bar{\omega}$ is the mean angular frequency of the source. Substituting into our expression for the mutual coherence function gives,

$$\Gamma_{12} = \sum_m \left\langle A_m\left(t - \frac{r_{m1}}{c}\right) A_m^*\left(t - \frac{r_{m2}}{c}\right)\right\rangle \frac{1}{r_{m1}r_{m2}} \exp\left[-\mathrm{i}\frac{\bar{\omega}}{c}\left(r_{m2} - r_{m1}\right)\right]. \tag{6.24}$$

If the path difference $|r_{m2} - r_{m1}|$ is small compared to the coherence length of the light, then we can neglect the retardation effects in $A_m(t)$, and hence,

$$\Gamma_{12} = \sum_m \left\langle |A_m(t)|^2 \right\rangle \frac{1}{r_{m1}r_{m2}} \exp\left[-\mathrm{i}\frac{\bar{\omega}}{c}\left(r_{m2} - r_{m1}\right)\right]. \tag{6.25}$$

Clearly $\left\langle |A_m(t)|^2 \right\rangle$ is proportional to the intensity emitted by each element. We now assume that the number of such elements is so large that the sum can be replaced by an integral, yielding,

$$\Gamma_{12} = \int_\sigma I(x,y) \frac{1}{r_1 r_2} \exp\left[\mathrm{i}\bar{k}\left(r_1 - r_2\right)\right] \mathrm{d}x\mathrm{d}y, \tag{6.26}$$

where $I(x,y)$ is the intensity emitted from the point (x,y) on the source, $\bar{k} = \bar{\omega}/c$ and r_1 and r_2 are the distances between a typical source point and P_1 and P_2 respectively. The complex degree of coherence is then,

$$\gamma_{12} = \frac{1}{\sqrt{I(P_1)I(P_2)}} \int_\sigma I(x,y) \frac{1}{r_1 r_2} \exp\left[\mathrm{i}\bar{k}\left(r_1 - r_2\right)\right] \mathrm{d}x\mathrm{d}y, \tag{6.27}$$

where $I(P_i)$ is the intensity at P_i, i.e.

$$I(P_i) = \int_\sigma \frac{I(x,y)}{r_i^2} \mathrm{d}x\mathrm{d}y. \tag{6.28}$$

Equation (6.27) has a familiar form: it is the sum of spherical waves reaching P_1, of amplitude,

$$\frac{I(x,y)}{r_2} \exp(-\mathrm{i}\bar{k}r_2)$$

In other words *the complex degree of coherence of an incoherent source of transverse intensity profile* $I(x,y)$ *is proportional to the* amplitude *of the diffraction pattern which would be observed at* P_1 *if the source were replaced by a screen of amplitude transmission function* $I(x,y)$ *and illuminated by a spherical wave converging on* P_2. This is the **van Cittert-Zernike theorem**.

It is clear that if it is measured at sufficiently large distances from the source, then γ_{12} will be proportional to the Fourier transform of the intensity distribution of the source. In this case, from (2.28),

$$r_1 \approx Z - \frac{X_1 x}{Z} - \frac{Y_1 y}{Z} + \frac{X_1^2 + Y_1^2}{2Z} \tag{6.29}$$

$$r_2 \approx Z - \frac{X_2 x}{Z} - \frac{Y_2 y}{Z} + \frac{X_2^2 + Y_2^2}{2Z}, \tag{6.30}$$

where X_i, Y_i are the coordinates of P_i. Then,

$$\gamma_{12} = \exp(\mathrm{i}\psi) \frac{\int_\sigma I(x,y) \exp\left[-\mathrm{i}\bar{k}\left(\frac{\Delta X}{Z}x + \frac{\Delta Y}{Z}y\right)\right] \mathrm{d}x\mathrm{d}y}{\int_\sigma I(x,y)\mathrm{d}x\mathrm{d}y}, \tag{6.31}$$

where $\Delta X = X_1 - X_2$, $\Delta Y = Y_1 - Y_2$, and,

$$\psi = \bar{k}\frac{X_1^2 - X_2^2 + Y_1^2 - Y_2^2}{2Z}. \tag{6.32}$$

In other words, γ_{12} is proportional to the Fourier transform of $I(x,y)$ calculated with the transform pair $(\bar{k}\Delta X/Z, \bar{k}\Delta Y/Z)$.

As an example, applying these results to a disk source of radius a, we find

$$\gamma_{12}^{\mathrm{disk}}(\Delta R) = \frac{2J_1(\bar{k}a\frac{\Delta R}{Z})}{\bar{k}a\frac{\Delta R}{Z}} \exp(\mathrm{i}\psi), \tag{6.33}$$

where $\Delta R = \sqrt{\Delta X^2 + \Delta Y^2}$. This falls from unity to zero when the argument of J_1 equals 3.83, i.e. when $\Delta R = 3.83Z/\bar{k}a \approx 0.61\bar{\lambda}/\alpha$, where $\alpha = a/Z$ is the angle subtended by the source. Note that the quantity determining the transverse coherence is the *angular* size of the source.

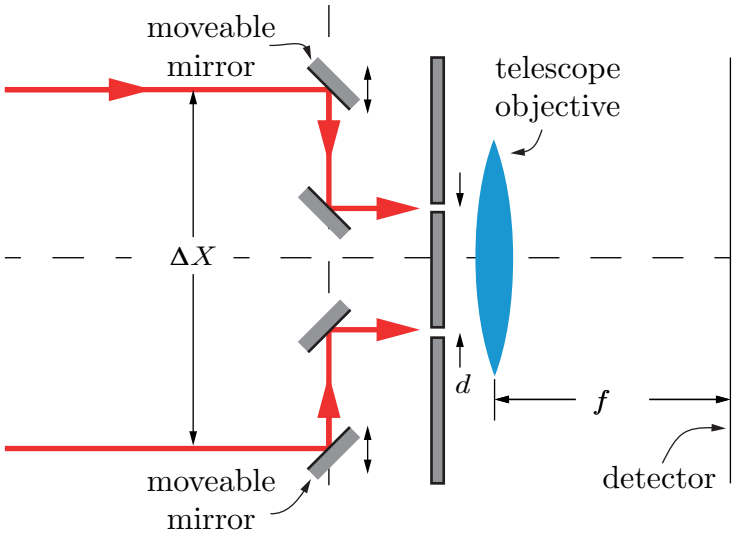


Figure 6.3: Schematic diagram of a Michelson stellar interferometer.

When $v = 1$, $J(v) = 0.88$, and this is often taken as a convenient measure of the transverse distance over which there is "good" coherence. This occurs when $\Delta R = d = 1/\alpha\bar{k} \approx 0.16\lambda/\alpha$. The linear dimension d is sometimes (confusingly) referred to as the "coherence area". Note also that the condition $|\gamma_{12}| = 0.88$ is often used to find the "coherence area" for sources other than disk sources, despite the fact that there is nothing special about the value 0.88 and it was selected in order to make the calculation simple in a particular case.

6.4.2 Michelson stellar interferometer

These ideas form the basis of the Michelson stellar interferometer shown in Fig. 6.3. Light from an astronomical object is collected by two mirrors mounted on rails so that their separation ΔX can be varied. The light reflected from these mirrors is then directed to a pair of slits, of separation d, and then imaged by the objective of the telescope. This arrangement ensures that the light is sampled from two points separated by a large, adjustable distance (ΔX), and that this light is then used to generate interference fringes with a constant spacing determined by d. Hence as ΔX is varied the fringe spacing does not change, but the visibility does.

If the source is completely incoherent then from the van Cittert-Zernike theorem the measured complex degree of coherence (determined from the fringe pattern) is proportional to the Fourier transform of the intensity distribution of the source $I(x, y)$. For example, for a single star of radius a the complex degree of coherence measured along an axis passing through the centre of the image of the star will be,

$$\left|\gamma_{12}^{\text{disk}}(\Delta X)\right| = \left|\frac{2J_1(\bar{k}a\frac{\Delta X}{Z})}{\bar{k}a\frac{\Delta X}{Z}}\right| = \left|\frac{2J_1(\bar{k}\alpha\Delta X}{\bar{k}\alpha\Delta X}\right|, \tag{6.34}$$

where $\alpha = a/Z$ is the angular radius of the star.

The main limitation of the Michelson stellar interferometer is the inconveniently long mirror separation required to determine the angular size of all but the largest stars.

Lecture 7

Electro-optic modulation

Recommended reading: Born & Wolf, Ch. 15, Hooker & Webb Ch 15, Davis Chs 18 & 19, Brooker, Ch 15.

7.1 Introduction

Much of the undergraduate syllabus concerns studies of linear systems. For example, the polarization $\boldsymbol{P}$ of a medium can be described in terms of the electric field $\boldsymbol{E}$ by,

$$\boldsymbol{P} = \epsilon_0 \chi \boldsymbol{E}, \tag{7.1}$$

where χ is the susceptibility of the medium. In fact eqn (7.1) is even more of a special case since χ is a scalar quantity and hence the response of the medium is assumed to be independent of the direction of the electric field.

These assumptions do not hold in many cases, including several of practical importance. For example, nonlinear crystals can be used to generate coherent beams of radiation at frequencies other than those available from laser media; the response of a nonlinear crystal to an applied electric field is neither linear or isotropic.

7.2 Propagation of light in non-isotropic media

We first consider the propagation of light in a linear, but non-isotropic medium. In particular we consider the case of *electrical* anisotropy, and hence we assume that the response of the medium to magnetic fields *is* isotropic (and described by the permeability of free space μ_0).

In general the displacement vector $\boldsymbol{D}$ is related to the electric field by,

$$D_x = \epsilon_0 \left(\epsilon_{xx} E_x + \epsilon_{xy} E_y + \epsilon_{xz} E_z\right) \tag{7.2}$$
$$D_y = \epsilon_0 \left(\epsilon_{yx} E_x + \epsilon_{yy} E_y + \epsilon_{yz} E_z\right) \tag{7.3}$$
$$D_z = \epsilon_0 \left(\epsilon_{zx} E_x + \epsilon_{zy} E_y + \epsilon_{zz} E_z\right) \tag{7.4}$$
$$\Rightarrow \boldsymbol{D} = \epsilon_0 \overline{\overline{\epsilon_\mathrm{r}}} \boldsymbol{E}, \tag{7.5}$$

in which $\overline{\overline{\epsilon_\mathrm{r}}}$ is the dielectric tensor.

If the coordinate system is aligned with the *principal dielectric axes* of the material then then the above simplifies to,

$$D_x = \epsilon_0 \epsilon_x E_x \tag{7.6}$$
$$D_y = \epsilon_0 \epsilon_y E_y \tag{7.7}$$
$$D_z = \epsilon_0 \epsilon_z E_z \tag{7.8}$$

where ϵ_x, ϵ_y and ϵ_z are the principal dielectric constants.

7.2.1 Relative orientation of fields

We now consider the propagation of a monochromatic plane wave in such a medium, assuming a coordinate system aligned with the principal dielectric axes. We assume a plane wave, so that all fields vary in space and time as,

$$\exp\left[\mathrm{i}\left(\boldsymbol{k}\cdot\boldsymbol{r} - \omega t\right)\right].$$

For plane waves of this type the vector operators $\nabla\cdot \equiv \mathrm{i}\boldsymbol{k}\cdot$ and $\nabla\times \equiv \mathrm{i}\boldsymbol{k}\times$. Hence, for a charge- and current-free region Maxwell's equations can be written for the wave as:

$$\boldsymbol{k}\cdot\boldsymbol{D} = 0 \tag{7.9}$$
$$\boldsymbol{k}\cdot\boldsymbol{B} = \boldsymbol{k}\cdot\boldsymbol{H} = 0 \tag{7.10}$$
$$\boldsymbol{k}\times\boldsymbol{E} = \mu_0\omega\boldsymbol{H} \tag{7.11}$$
$$\boldsymbol{k}\times\boldsymbol{H} = -\omega\boldsymbol{D}. \tag{7.12}$$

With a little thought the above equations tell us that:

1. The vectors $\boldsymbol{D}$, $\boldsymbol{H}$, and $\boldsymbol{k}$ are mutually orthogonal
2. The vector $\boldsymbol{D}\times\boldsymbol{H}$ points along the direction of $\boldsymbol{k}$, i.e. along the normals to the wavefronts
3. The Poynting vector $\boldsymbol{S} = \boldsymbol{E}\times\boldsymbol{H}$, which gives the direction of energy flow, does *not* point along $\boldsymbol{k}$. In a ray picture $\boldsymbol{S}$ gives the direction of ray propagation.

7.2.2 The phase velocity

We now find the phase velocity of the wave propagation. Eliminating $\boldsymbol{H}$ from eqns (7.11) and (7.12) we find,

$$\begin{aligned}
\boldsymbol{k}\times\left(\frac{1}{\mu_0\omega}\boldsymbol{k}\times\boldsymbol{E}\right) &= -\omega\boldsymbol{D}\\
\Rightarrow \boldsymbol{D} &= -\frac{1}{\mu_0\omega^2}\boldsymbol{k}\times(\boldsymbol{k}\times\boldsymbol{E})\\
&= -\frac{1}{\mu_0\omega^2}\left[(\boldsymbol{k}\cdot\boldsymbol{E})\,\boldsymbol{k}-k^2\boldsymbol{E}\right]\\
&= -\frac{k^2}{\mu_0\omega^2}\left[\left(\hat{\boldsymbol{k}}\cdot\boldsymbol{E}\right)\hat{\boldsymbol{k}}-\boldsymbol{E}\right].
\end{aligned}$$

Now $\omega/k = c/n = (\epsilon_0\mu_0)^{-1/2}/n$ and hence we can write the above as

$$\boldsymbol{D} = \epsilon_0 n^2\left[\boldsymbol{E}-\left(\hat{\boldsymbol{k}}\cdot\boldsymbol{E}\right)\hat{\boldsymbol{k}}\right].$$

Now using the relations $D_x = \epsilon_0\epsilon_x E_x$ etc. we can write

$$\begin{aligned}
D_x = \epsilon_0\epsilon_x E_x &= \epsilon_0 n^2\left[E_x-\left(\hat{\boldsymbol{k}}\cdot\boldsymbol{E}\right)\hat{k}_x\right]\\
\Rightarrow \left(1-\frac{\epsilon_x}{n^2}\right)E_x &= \left(\hat{\boldsymbol{k}}\cdot\boldsymbol{E}\right)\hat{k}_x\\
\Rightarrow E_x &= \frac{\hat{k}_x}{1-\frac{\epsilon_x}{n^2}}\hat{\boldsymbol{k}}\cdot\boldsymbol{E}. \qquad (7.13)
\end{aligned}$$

The analogous expressions for E_y and E_z can be found in the same way.

We now evaluate $\hat{\boldsymbol{k}}\cdot\boldsymbol{E}$ to give:

$$\begin{aligned}
\hat{\boldsymbol{k}}\cdot\boldsymbol{E} = \hat{k}_x E_x+\hat{k}_y E_y+\hat{k}_z E_z &= \left(\frac{\hat{k}_x^2}{1-\frac{\epsilon_x}{n^2}}+\frac{\hat{k}_y^2}{1-\frac{\epsilon_y}{n^2}}+\frac{\hat{k}_z^2}{1-\frac{\epsilon_z}{n^2}}\right)\hat{\boldsymbol{k}}\cdot\boldsymbol{E}\\
\Rightarrow \frac{\hat{k}_x^2}{1-\frac{\epsilon_x}{n^2}}+\frac{\hat{k}_y^2}{1-\frac{\epsilon_y}{n^2}}+\frac{\hat{k}_z^2}{1-\frac{\epsilon_z}{n^2}} &= 1\\
\Rightarrow \frac{\hat{k}_x^2}{n^2-n_x^2}+\frac{\hat{k}_y^2}{n^2-n_y^2}+\frac{\hat{k}_z^2}{n^2-n_z^2} &= \frac{1}{n^2}, \qquad (7.14)
\end{aligned}$$

where we have written $n_x^2 = \epsilon_x$ etc. This last equation is quadratic in n^2, and hence there are *two* solutions for the refractive index n for a given direction of propagation $\boldsymbol{k}$. (For each such solution there is also the trivial solution obtained from $n\to -n$, but this merely corresponds to a reversal in the direction of $\boldsymbol{k}$). The two solutions for n correspond to each of two possible orientations of the $\boldsymbol{D}$ field; once n is known then (7.13) and (7.5) give the components of $\boldsymbol{E}$ and $\boldsymbol{D}$ respectively.

7.2.3 Propagation in uniaxial crystals

It is useful to apply this last result to the case of a uniaxial crystal, for which $\epsilon_x = \epsilon_y \neq \epsilon_z$. It will be useful to define $n_\mathrm{o} = \sqrt{\epsilon_x}$ and $n_\mathrm{e} = \sqrt{\epsilon_z}$.

One solution to Maxwell's equations in this case arises for waves with $\boldsymbol{D}$ polarized perpendicular to the z-axis (the optic axis). This wave, the *ordinary wave*, will experience a refractive index n_o.

To find the refractive index experienced by the orthogonal polarization, the *extraordinary wave*, let us assume that the coordinate system is rotated such that $\hat{k}_y = 0$, i.e. the wave normals (the k-vector) lies in the xz-plane. Our last result then becomes,

$$\frac{\hat{k}_x^2}{n^2 - n_x^2} + \frac{\hat{k}_z^2}{n^2 - n_z^2} = \frac{1}{n^2}. \tag{7.15}$$

We can write $\hat{k}_x = \sin\theta$ and $\hat{k}_z = \cos\theta$, where θ is the angle between $\boldsymbol{k}$ and the optic axis (the z-axis). Then,

$$\frac{\sin^2\theta}{n(\theta)^2 - n_x^2} + \frac{\cos^2\theta}{n(\theta)^2 - n_z^2} = \frac{1}{n(\theta)^2} \tag{7.16}$$

$$\Rightarrow \left[n(\theta)^2 - n_z^2\right]\sin^2\theta + \left[n(\theta)^2 - n_x^2\right]\cos^2\theta = \frac{\left[n(\theta)^2 - n_x^2\right]\left[n(\theta)^2 - n_z^2\right]}{n(\theta)^2} \tag{7.17}$$

$$\Rightarrow n(\theta)^2 - n_z^2\sin^2\theta - n_x^2\cos^2\theta = n(\theta)^2 - n_x^2 - n_z^2 + \frac{n_x^2 n_z^2}{n(\theta)^2} \tag{7.18}$$

$$\Rightarrow n_x^2\left(1 - \cos^2\theta\right) + n_z^2\left(1 - \sin^2\theta\right) = \frac{n_x^2 n_z^2}{n(\theta)^2} \tag{7.19}$$

$$\Rightarrow \frac{\cos^2\theta}{n_\mathrm{o}^2} + \frac{\sin^2\theta}{n_\mathrm{e}^2} = \frac{1}{n(\theta)^2} \tag{7.20}$$

where we have written $n_\mathrm{o} = n_x = n_y$ and $n_\mathrm{e} = n_z$.

This result gives the refractive index of the extraordinary wave propagating at an angle θ to the optic axis. The extraordinary wave is polarized in the plane containing the direction of propagation and the optic axis. We see that the refractive index experienced by the extraordinary ray lies between n_o and n_e. For example, if $\theta = \pi/2$ then $n(\theta) = n_\mathrm{e}(= n_z)$; if $\theta = 0$ then $n(\theta) = n_\mathrm{o}$ (and the extraordinary wave experiences the same refractive as the ordinary wave). This relation is shown in Fig. 7.1.

7.2.4 The index ellipsoid (a.k.a. the "indicatrix" or the "ellipsoid of wave normals")

An alternative approach to (7.14) for finding the velocities (i.e. refractive indices) and orientations of the fields makes use of a geometric construction. To find this we first note that the energy density of the wave is,

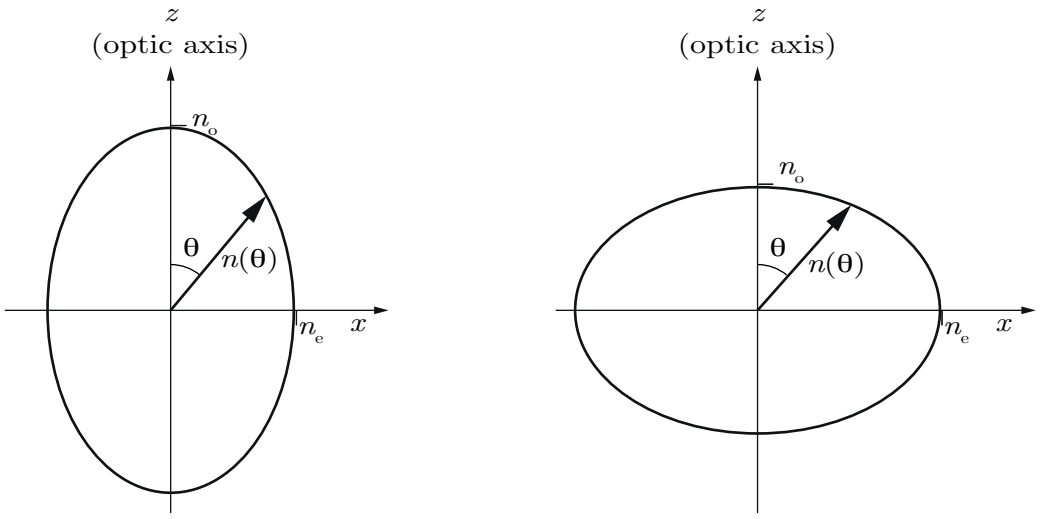


Figure 7.1: The refractive index of the extraordinary ray, as a function of the angle of propagation θ relative to the optic axis, for a uniaxial crystal: (a) positive birefringence case ($n_o > n_e$); (b) negative birefringence case ($n_o < n_e$).

$$U = \frac{1}{2}\boldsymbol{E}\cdot\boldsymbol{D} + \frac{1}{2}\boldsymbol{B}\cdot\boldsymbol{H} = \boldsymbol{E}\cdot\boldsymbol{D}. \tag{7.21}$$

Now since $D_x = \epsilon_0\epsilon_x E_x$ etc, we can write,

$$U = E_x D_x + E_y D_y + E_z D_z \tag{7.22}$$

$$= \frac{D_x^2}{\epsilon_0\epsilon_x} + \frac{D_y^2}{\epsilon_0\epsilon_y} + \frac{D_z^2}{\epsilon_0\epsilon_z}. \tag{7.23}$$

Hence the surface mapped out by the D-vectors of waves of constant energy density is given by,

$$\frac{x^2}{\epsilon_x} + \frac{y^2}{\epsilon_y} + \frac{z^2}{\epsilon_z} = 1 \tag{7.24}$$

$$\text{or,}\quad \frac{x^2}{n_x^2} + \frac{y^2}{n_y^2} + \frac{z^2}{n_z^2} = 1. \tag{7.25}$$

This surface, plotted in Fig. 7.2, is known as the index ellipsoid — also known as "the optical indicatrix", or the "ellipsoid of wave normals" or the "reciprocal ellipsoid". It can be shown that for a wave of given $\boldsymbol{k}$ the two possible refractive indices are proportional to the major and minor axes of

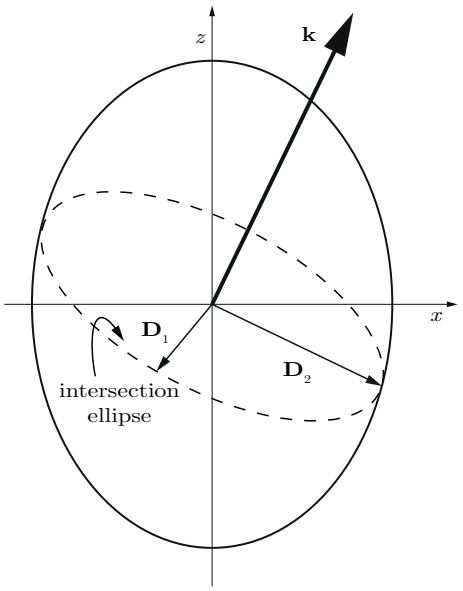


Figure 7.2: The index ellipsoid. For light propagating in the direction $\boldsymbol{k}$ the two orthogonal polarizations, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$, are given by the major and minor ellipses of the intersection ellipse. The refractive indices experienced by these two polarizations are proportional to the lengths of the axes of the intersection ellipse.

the ellipse formed by the intersection of this ellipsoid and plane normal to $\boldsymbol{k}$, and passing through the origin. Further, the two possible orientations of $\boldsymbol{D}$ are parallel to the axes of this ellipse.[1]

Note that the index ellipsoid for a uniaxial crystal has the form,

$$\frac{x^2+y^2}{n_{\rm o}^2}+\frac{z^2}{n_{\rm e}^2}=1. \tag{7.26}$$

This index ellipsoid can then be used to show that the refractive index of the extraordinary ray propagating at angle θ to the optic (z) axis is given by eqn (7.20).

7.3 Nonlinear optics

Most real systems exhibit at least some degree of nonlinear behaviour. To account for this, and non-isotropy, eqn (7.1) must be re-written as,

$$P_i=\epsilon_0\left[\sum_j \chi_{ij}E_j+\sum_{jk}\chi_{ijk}E_jE_k+\ldots\right] \tag{7.27}$$

$$=\epsilon_0\chi_{ij}E_j+\epsilon_0\chi_{ijk}E_jE_k+\ldots, \tag{7.28}$$

[1] For a proof of these statements see Born & Wolf, Chapter 15.

where in the second line we have used the Einstein summation convention, i.e. that summations should be performed over repeated indices. In the above, P_i represents one component of the polarization (e.g. the x-component) at a given frequency. In the next lecture we will return to the question of the possible relations between the frequency of P_i and those of the fields $E_i, E_j, \ldots$.

7.3.1 The electro-optic effect

A constant electric field applied to a crystal will in general lead to (small) movements of the ion positions and distortion of the electron charge clouds. This in turn will change the refractive indices of the crystal, including the orientation and degree of anisotropy.

Crystals for which the response is linear in the applied fields exhibit the *linear* electro-optic effect. In this case, eqn (7.28) can be written as,

$$P_i = \epsilon_0 \chi_{ij} E_j + \epsilon_0 \chi_{ijk} E_j E_k^0 \tag{7.29}$$

$$= \epsilon_0 \left(\chi_{ij} + \chi_{ijk} E_k^0 \right) E_j, \tag{7.30}$$

where E_k^0 is the constant applied field. Note that for $\chi_{ijk} \neq 0$ the crystal must lack inversion symmetry.[2] Crystals *with* inversion symmetry can exhibit a *quadratic* electro-optic effect, in which the response depends on the square of the applied electric field.

In the presence of an applied field the principal axes of the crystal may no longer be aligned with the chosen coordinate system, and hence the index ellipsoid takes the general form,

$$\left(\frac{1}{n^2}\right)_1 x^2 + \left(\frac{1}{n^2}\right)_2 y^2 + \left(\frac{1}{n^2}\right)_3 z^2 + 2\left(\frac{1}{n^2}\right)_4 yz + 2\left(\frac{1}{n^2}\right)_5 xz + 2\left(\frac{1}{n^2}\right)_6 xy = 1, \tag{7.31}$$

where the coefficients $(1/n^2)_i$ determine the shape and orientation of the ellipsoid. For the case of the linear electro-optic effect these coefficients change linearly with the x, y, z components of the applied field, which can be described by the *electro-optic tensor* $\overline{\overline{r}}$:

$$\Delta\left(\frac{1}{n^2}\right)_i = r_{ij} E_j \tag{7.32}$$

Or,

$$\begin{pmatrix} \Delta(1/n^2)_1 \\ \Delta(1/n^2)_2 \\ \Delta(1/n^2)_3 \\ \Delta(1/n^2)_4 \\ \Delta(1/n^2)_5 \\ \Delta(1/n^2)_6 \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \\ r_{41} & r_{42} & r_{43} \\ r_{51} & r_{52} & r_{53} \\ r_{61} & r_{62} & r_{63} \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix}. \tag{7.33}$$

In the above we have used a contracted notation in which $x = 1$, $y = 2$, $z = 3$, $xx = 1$, $yy = 2$, $zz = 3$, $yz = 4$, $xz = 5$, and $xy = 6$.

[2] This requirement follows from the fact that in a system with inversion symmetry the polarization $\boldsymbol{P}$ must change sign if an applied field $\boldsymbol{E}$ is reversed in direction. As a consequence χ_{ijk} — and in fact all coefficients of even powers of the field $\boldsymbol{E}$ — must be zero.

Note that typical magnitudes of the r_{ij} are $10^{-12}\,\mathrm{mV}^{-1}$; i.e. an applied voltage of $1\,\mathrm{kVmm}^{-1}$ will change the coefficients of the ellipsoid by 10^{-6}

In the presence of the applied field the index ellipsoid will take the form,

$$\begin{aligned}&\left[\left(\frac{1}{n^2}\right)_1+\Delta\left(\frac{1}{n^2}\right)_1\right]x^2+\left[\left(\frac{1}{n^2}\right)_2+\Delta\left(\frac{1}{n^2}\right)_2\right]y^2+\left[\left(\frac{1}{n^2}\right)_3+\Delta\left(\frac{1}{n^2}\right)_3\right]z^2\\&+2\left[\left(\frac{1}{n^2}\right)_4+\Delta\left(\frac{1}{n^2}\right)_4\right]yz+2\left[\left(\frac{1}{n^2}\right)_5+\Delta\left(\frac{1}{n^2}\right)_5\right]xz+2\left[\left(\frac{1}{n^2}\right)_6+\Delta\left(\frac{1}{n^2}\right)_6\right]xy\\&=1.\end{aligned}\tag{7.34}$$

In practice many of the elements in the electro-optic tensor are zero, which considerably eases the task of finding the new refractive indices of the crystal.

7.3.2 An example: ADP

In the uniaxial crystal ammonium dihydrogen phosphate (ADP) the only non-zero elements of the electro-optic tensor are r_{41}, r_{52}, and r_{63}. Hence applying a field in an arbitrary direction will in general cause the crystal to become biaxial since the coefficients of the terms in yz, xz, and xy will all be non-zero.

Let us consider the special case in which the applied field is along the z-direction. Then, eqn (7.33) becomes,

$$\begin{aligned}&\begin{pmatrix}\Delta(1/n^2)_1\\\Delta(1/n^2)_2\\\Delta(1/n^2)_3\\\Delta(1/n^2)_4\\\Delta(1/n^2)_5\\\Delta(1/n^2)_6\end{pmatrix}=\begin{pmatrix}0&0&0\\0&0&0\\0&0&0\\r_{41}&0&0\\0&r_{52}&0\\0&0&r_{63}\end{pmatrix}\begin{pmatrix}0\\0\\E_z\end{pmatrix}\\&\Rightarrow\begin{pmatrix}\Delta(1/n^2)_1\\\Delta(1/n^2)_2\\\Delta(1/n^2)_3\\\Delta(1/n^2)_4\\\Delta(1/n^2)_5\\\Delta(1/n^2)_6\end{pmatrix}=\begin{pmatrix}0\\0\\0\\0\\0\\r_{63}E_z\end{pmatrix}.\end{aligned}\tag{7.35}$$

The index ellipsoid then becomes,

$$\frac{x^2}{n_\mathrm{o}^2}+\frac{y^2}{n_\mathrm{o}^2}+\frac{z^2}{n_\mathrm{e}^2}+2xyr_{63}E_z=1.\tag{7.36}$$

Defining new x'- and y'-axes rotated about the z-axis by $+\pi/4$, so that $x=(x'-y')/\sqrt{2}$, $y=(x'+y')/\sqrt{2}$, and $z=z'$ gives,

$$\frac{x'^2 + y'^2 - 2x'y'}{2n_o^2} + \frac{x'^2 + y'^2 + 2x'y'}{2n_o^2} + \frac{z'^2}{n_e^2} + 2\frac{(x'-y')(x'+y')}{2} r_{63}E_z = 1$$

$$\Rightarrow \frac{x'^2 + y'^2}{n_o} + \frac{z'^2}{n_e^2} + (x'^2 - y'^2) r_{63}E_z = 1$$

$$\Rightarrow \left(\frac{1}{n_o^2} + r_{63}E_z\right) x'^2 + \left(\frac{1}{n_o^2} - r_{63}E_z\right) y'^2 + \frac{z'^2}{n_e^2} = 1. \tag{7.37}$$

Light polarized along the x'- and y'-axes therefore experiences refractive indices,

$$n_{x'} = \left(\frac{1}{n_o^2} + r_{63}E_z\right)^{-1/2} \approx n_o - \frac{1}{2}n_o^3 r_{63}E_z + \ldots \tag{7.38}$$

$$n_{y'} = \left(\frac{1}{n_o^2} - r_{63}E_z\right)^{-1/2} \approx n_o + \frac{1}{2}n_o^3 r_{63}E_z + \ldots. \tag{7.39}$$

Application to intensity modulation

Suppose that we want to construct an electro-optic wave plate. We orient the ADP crystal in a longitudinal configuration, i.e. with the applied field E_z along the direction of propagation of the incident light. The induced birefringence and retardance are then,

$$\Delta n = \left|n_{x'} - n_{y'}\right| = n_o^3 r_{63}E_z \tag{7.40}$$

$$\Delta\phi = \frac{2\pi}{\lambda}\Delta n\ell = \frac{2\pi}{\lambda}n_o^3 r_{63}E_z\ell, \tag{7.41}$$

where ℓ is the length of the crystal. Now in this orientation the field is related to the applied voltage by $E_z = V/\ell$, hence the voltage required for a retardance $\Delta\phi$ is,

$$\Delta\phi = \frac{2\pi}{\lambda}n_o^3 r_{63}E_z\ell \tag{7.42}$$

$$\Rightarrow V_{\Delta\phi} = \frac{\lambda}{n_o^3 r_{63}}\frac{\Delta\phi}{2\pi}. \tag{7.43}$$

For example, to make a half-wave plate the *half-wave voltage* is given by,

$$V_\pi = \frac{\lambda}{2n_o^3 r_{63}}. \tag{7.44}$$

The form of the modulation is as follows. Suppose that the modulator is placed between crossed polarizers oriented at 45-degrees to the x'- and y'-axes, as shown in Fig. 7.3. The field entering the modulator is then given by,

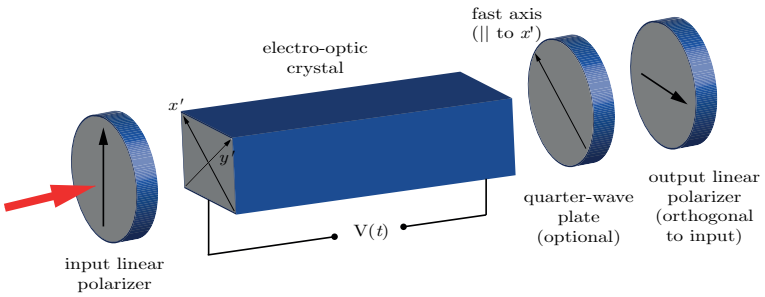


Figure 7.3: A possible geometry for electro-optic intensity modulation.

$$E_{\text{in}} = \frac{E_0}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \exp\left(-\mathrm{i}\omega t\right), \tag{7.45}$$

where we have used the x', y' coordinate system. If the modulator imposes a retardance $\Delta\phi$ then the field after the modulator is,

$$E_{\text{mod}} = \frac{E_0}{\sqrt{2}} \begin{pmatrix} 1 \\ \exp\left(\mathrm{i}\Delta\phi\right) \end{pmatrix} \exp\left(-\mathrm{i}\omega t\right), \tag{7.46}$$

and hence the field transmitted by the second polarizer is

$$E_{\text{trans}} = -\frac{E'_x}{\sqrt{2}} + \frac{E'_y}{\sqrt{2}} \tag{7.47}$$

$$= \frac{1}{2}\left[\exp\left(\mathrm{i}\Delta\phi\right) - 1\right] \tag{7.48}$$

$$\Rightarrow \frac{I_{\text{trans}}}{I_{\text{in}}} = \sin^2\left(\frac{\Delta\phi}{2}\right) = \sin^2\left(\frac{\alpha V(t)}{2}\right) = \frac{1}{2}\left\{1 - \cos\left[\alpha V(t)\right]\right\}, \tag{7.49}$$

where we have written $\Delta\phi = \alpha V(t)$.

Notice that in the above the modulation is not linear in the applied voltage $V(t)$. If the phase modulation is small the modulation can be made linear in the voltage by inserting a quarter-wave plate between the crossed polarizers, oriented with its axes parallel to those of the electro-optic modulator. The effect of this is that,

$$\Delta\phi \to \Delta\phi + \frac{\pi}{2}, \tag{7.50}$$

and hence,

$$\frac{I_{\text{trans}}}{I_{\text{in}}} = \frac{1}{2}\left\{1 - \cos\left[\alpha V(t) + \pi/2\right]\right\} = \frac{1}{2}\left\{1 - \sin\left[\alpha V(t)\right]\right\} \approx \frac{1}{2}\left[1 - \alpha V(t)\right] \tag{7.51}$$

Application to phase modulation

The ADP modulator considered above can be used to modulate the phase of an incident beam of light by ensuring that the incident light is polarized along the x'- or y'-axes. Taking the latter case, the phase shift introduced by the modulator will then be,

$$\begin{aligned}\phi &= \frac{2\pi}{\lambda} n_{y'} \ell \\ &= \frac{2\pi}{\lambda} n_{\rm o} \ell + \frac{\pi}{\lambda} n_{\rm o}^3 r_{63} V(t),\end{aligned} \tag{7.52}$$

where $V(t)$ is the applied voltage. If $V(t) = V_0 \sin \Omega t$ then an incident wave of the form $E_{y'}(0,t) = E_0 \exp(-\mathrm{i}\omega t)$ will leave the modulator with a field,

$$E_{y'}(\ell, t) = E_0 \exp(-\mathrm{i}\omega t) \exp\left[\mathrm{i}\frac{2\pi}{\lambda}\left(n_{\rm o} + \frac{1}{2} n_{\rm o}^3 r_{g3} \frac{V(t)}{\ell}\right)\ell\right] \tag{7.53}$$

$$= E_0 \exp(-\mathrm{i}\omega t) \exp\left(\mathrm{i}\frac{2\pi}{\lambda} n_{\rm o} \ell\right) \exp\left(\mathrm{i}\frac{1}{2} n_{\rm o}^3 r_{g3} V_0 \sin \Omega t\right) \tag{7.54}$$

$$= E_0 \exp(-\mathrm{i}\omega t) \exp\left(\mathrm{i}\frac{2\pi}{\lambda} n_{\rm o} \ell\right) \exp(\mathrm{i}\beta \sin \Omega t) \tag{7.55}$$

$$= E_0 \exp(-\mathrm{i}\omega t) \exp\left(\mathrm{i}\frac{2\pi}{\lambda} n_{\rm o} \ell\right) \sum_{m=-\infty}^{\infty} J_m(\beta) \exp(\mathrm{i}m\Omega t), \tag{7.56}$$

where $\beta = (1/2) n_{\rm o}^3 r_{g3} V_0$. Hence the modulator induces sidebands at frequencies $\omega + m\Omega$. Frequency modulators of this type are used for FM modelocking.

Transverse modulation

It should be obvious that longitudinal modulation, of the type discussed in this section, is not the most practical arrangement since the field E_z is generated by electrodes at each end of the crystal, which the light to be modulated must pass through. Other electro-optic crystals, such as lithium niobate ($LiNbO_3$) can be used to make transverse modulators.

Lecture 8

Frequency generation

Recommended reading: Davis Ch 21; Hooker & Webb Ch 15.

8.1 Preliminary considerations

As we saw in Section 7.3 most media respond nonlinearly to the electric field within them, so that the polarization can be written in the form,

$$P_i = \epsilon_0 \chi_{ij} E_j + \epsilon_0 \chi_{ijk} E_j E_k + \dots . \tag{8.1}$$

We have not so far considered the situation in which the fields, and hence polarization, may have different frequencies. We do so now.

We will deal with real fields, which can be written in the form,

$$\begin{aligned} \boldsymbol{E}^{\omega_i}(t) &= \boldsymbol{E}(\omega_i) \cos\left(\boldsymbol{k}_i \cdot \boldsymbol{r} - \omega_i t\right) && (8.2)\\ &= \frac{\boldsymbol{E}(\omega_i)}{2} \left\{ \exp\left[\mathrm{i}\left(\boldsymbol{k}_i \cdot \boldsymbol{r} - \omega_i t\right)\right] + \exp\left[-\mathrm{i}\left(\boldsymbol{k}_i \cdot \boldsymbol{r} - \omega_i t\right)\right] \right\} && (8.3)\\ &= \frac{\boldsymbol{E}(\omega_i)}{2} \exp\left[\mathrm{i}\left(\boldsymbol{k}_i \cdot \boldsymbol{r} - \omega_i t\right)\right] + \text{c.c.}, && (8.4) \end{aligned}$$

where we have introduced the convention that c.c. denotes the complex conjugate of the previous terms.

Since the nonlinear polarization depends on products of the electric field, it will contain frequencies at the sums and differences of the driving fields. For example, second-order processes will lead to the polarization oscillating at $\omega_3 = \omega_1 + \omega_2$, where ω_1 and ω_2 are the frequencies of the fields, according to:

$$P_i^{\omega_3} = \epsilon_0 \chi^{(2)}_{ijk}(-\omega_3, \omega_1, \omega_1) E_j^{\omega_1}(t) E_k^{\omega_2}(t). \tag{8.5}$$

Note that we have introduced the common convention that the argument of the second-order susceptibility tensor contains the three frequencies, with a sign convention which ensures the sum of arguments is zero. This sign convention expresses energy conservation.

Now the nonlinear polarization can be written as,

$$\boldsymbol{P}^{\omega_3}(t) = \frac{\boldsymbol{P}(\omega_3)}{2} \exp\left[\mathrm{i}\left(\boldsymbol{k}_3 \cdot \boldsymbol{r} - \omega_3 t\right)\right] + \text{c.c.}. \tag{8.6}$$

We can therefore write,

$$\begin{aligned} P_i^{\omega_3}(t) =& \frac{P_i(\omega_3)}{2} \exp\left[\mathrm{i}\left(\boldsymbol{k}_3 \cdot \boldsymbol{r} - \omega_3 t\right)\right] + \text{c.c.} \\ &= \epsilon_0 \chi^{(2)}_{ijk}(-\omega_3, \omega_1, \omega_2) \left\{ \frac{E_j(\omega_1)}{2} \exp\left[\mathrm{i}\left(\boldsymbol{k}_1 \cdot \boldsymbol{r} - \omega_1 t\right)\right] + \text{c.c.} \right\} \left\{ \frac{E_k(\omega_2)}{2} \exp\left[\mathrm{i}\left(\boldsymbol{k}_2 \cdot \boldsymbol{r} - \omega_2 t\right)\right] + \text{c.c.} \right\} \end{aligned} \tag{8.7}$$

$$\Rightarrow P_i(\omega_3) = \frac{\epsilon_0}{2} \chi^{(2)}_{ijk}(-\omega_3, \omega_1, \omega_2) E_j(\omega_1) E_k(\omega_2) \exp\left[\mathrm{i}\left(\boldsymbol{k}_1 + \boldsymbol{k}_2 - \boldsymbol{k}_3\right) \cdot \boldsymbol{r}\right] \tag{8.8}$$

$$= \frac{\epsilon_0}{2} \chi^{(2)}_{ijk}(-\omega_3, \omega_1, \omega_2) E_j(\omega_1) E_k(\omega_2) \exp\left(-\mathrm{i}\boldsymbol{\Delta k} \cdot \boldsymbol{r}\right) \tag{8.9}$$

This gives the relationship between the Fourier components of the nonlinear polarization and those of the fields. Note also the factor $\exp\left[\mathrm{i}\left(\boldsymbol{k}_1 + \boldsymbol{k}_2 - \boldsymbol{k}_3\right) \cdot \boldsymbol{r}\right]$; we will see later that — unless $\boldsymbol{\Delta k} = \boldsymbol{k}_3 - \boldsymbol{k}_1 - \boldsymbol{k}_2 = 0$ — this factor will lead to oscillations in the intensity of the generated frequencies.

The above result is sometimes written in terms of a *nonlinear mixing tensor* $\overline{\overline{d}}^{(2)}$:

$$P_i(\omega_3) = \frac{\epsilon_0}{2} \chi^{(2)}_{ijk}(-\omega_3, \omega_1, \omega_1) E_j^{\omega_1} E_j^{\omega_2} \exp\left(\mathrm{i}\boldsymbol{\Delta k} \cdot \boldsymbol{r}\right) = \epsilon_0 d^{(2)}_{ijk}(-\omega_3, \omega_1, \omega_1) E_j(\omega_1) E_k(\omega_2) \exp\left(\mathrm{i}\boldsymbol{\Delta k} \cdot \boldsymbol{r}\right), \tag{8.10}$$

where $\chi^{(2)}_{ijk}(-\omega_3, \omega_1, \omega_1) = 2d^{(2)}_{ijk}$.

8.2 The nonlinear wave equation

To understand how waves propagate in a nonlinear medium, and hence to see how waves at new frequencies can be generated, we return to Maxwell's equations. In general we can write $\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}$. Now, however, the polarization has both a linear and nonlinear component, so that,

$$\boldsymbol{P} = \epsilon_0 \overline{\overline{\chi}}^{(1)} \boldsymbol{E} + \boldsymbol{P}_{\mathrm{NL}}, \tag{8.11}$$

in which the first term is the usual linear polarization, and the second term is the nonlinear polarization. For a current-free region we can write the Maxwell equation,

$$\nabla \times \boldsymbol{H} = \frac{\partial \boldsymbol{D}}{\partial t} = \epsilon_0 \overline{\overline{\epsilon_\mathrm{r}}} \frac{\partial \boldsymbol{E}}{\partial t} + \frac{\partial \boldsymbol{P}_\mathrm{NL}}{\partial t}, \tag{8.12}$$

where $\overline{\overline{\epsilon_\mathrm{r}}} = 1 + \overline{\overline{\chi^{(1)}}}$ is the usual linear permittivity.

Using the identity $\nabla \times (\nabla \times \boldsymbol{E}) = -\nabla(\nabla \cdot \boldsymbol{E}) - \nabla^2 \boldsymbol{E}$, and assuming the region contains no free charges (so that $\nabla \cdot \boldsymbol{D} = \nabla \cdot \boldsymbol{E} = 0$), then we can write,

$$-\nabla^2 \boldsymbol{E} = \nabla \times \left(-\frac{\partial \boldsymbol{B}}{\partial t} \right) = -\mu \frac{\partial}{\partial t} \nabla \times \boldsymbol{H} \tag{8.13}$$

$$\Rightarrow \nabla^2 \boldsymbol{E} = \mu \epsilon_0 \overline{\overline{\epsilon_\mathrm{r}}} \frac{\partial^2 \boldsymbol{E}}{\partial t^2} + \mu \frac{\partial^2 \boldsymbol{P}_\mathrm{NL}}{\partial t^2}. \tag{8.14}$$

We now restrict ourselves to the case of plane waves propagating along the z-axis. Considering the i component (where i represents x, y, or z) the above then becomes,

$$\frac{\partial^2 E_i}{\partial z^2} = \mu \epsilon_0 \epsilon_\mathrm{r} \frac{\partial^2 E_i}{\partial t^2} + \mu \frac{\partial^2 (P_\mathrm{NL})_i}{\partial t^2}. \tag{8.15}$$

We write the fields in the form,

$$E_j(t) = \frac{E_{j1}(z)}{2} \exp\left[\mathrm{i}\left(k_1 z - \omega_1 t\right)\right] + \text{c.c.} \tag{8.16}$$

$$E_k(t) = \frac{E_{k2}(z)}{2} \exp\left[\mathrm{i}\left(k_2 z - \omega_2 t\right)\right] + \text{c.c.}. \tag{8.17}$$

The nonlinear polarization is then,

$$(P_\mathrm{NL})_i = \epsilon_0 \chi_{kij} \left\{ \frac{E_{j1}(z)}{2} \exp\left[\mathrm{i}\left(k_1 z - \omega_1 t\right)\right] + \text{c.c.} \right\} \left\{ \frac{E_{k2}(z)}{2} \exp\left[\mathrm{i}\left(k_2 z - \omega_2 t\right)\right] + \text{c.c.} \right\}. \tag{8.18}$$

The component oscillating at $\omega_3 = \omega_1 + \omega_2$ is then,

$$\left(P_\mathrm{NL}^{\omega_3}\right)_i = \epsilon_0 \frac{\chi_{kij}}{4} E_{j1}(z) E_{k2}(z) \exp\left[\mathrm{i}\left(k_1 + k_2\right) z - \left(\omega_1 + \omega_2\right) t\right] + \text{c.c.}. \tag{8.19}$$

Substituting this nonlinear polarization into (8.15), and making the slowly-varying envelope approximation, we can derive the following equation for the growth of the amplitude of the generated beam (see appendix):

$$\frac{\partial E_{i3}}{\partial z} = \mathrm{i} \frac{\mu \epsilon_0 \chi_{kij} \omega_3^2}{4 k_3} E_{j1}(z) E_{k2}(z) \exp\left(-\mathrm{i} \Delta k z\right) \tag{8.20}$$

$$= \mathrm{i} \sqrt{\frac{\mu \epsilon_0}{\epsilon_3}} \frac{\omega_3 \chi_{kij}}{4} E_{j1}(z) E_{k2}(z) \exp\left(-\mathrm{i} \Delta k z\right). \tag{8.21}$$

where $\epsilon_3 = \epsilon_\mathrm{r}$. This equation describes the growth with z of the frequency generated at ω_3.

If we ignore pump depletion, i.e. we assume that the intensity of the driving fields at ω_1 and ω_2 do not change with z, then we can integrate this equation straightforwardly:

$$\begin{aligned}
\int_0^{E_{3k}} \mathrm{d}E_{i3} &= \mathrm{i}\sqrt{\frac{\mu\epsilon_0}{\epsilon_3}}\frac{\omega_3\chi_{kij}}{4}E_{j1}(0)E_{k2}(0)\int_0^z \exp\left(-\mathrm{i}\Delta k z\right)\mathrm{d}z \\
\Rightarrow E_{i3}(z) &= \mathrm{i}\sqrt{\frac{\mu\epsilon_0}{\epsilon_3}}\frac{\omega_3\chi_{kij}}{4}E_{j1}(0)E_{k2}(0)\left[\frac{\exp\left(-\mathrm{i}\Delta k z\right)-1}{-\mathrm{i}\Delta k}\right] \\
&= \mathrm{i}\sqrt{\frac{\mu\epsilon_0}{\epsilon_3}}\frac{\omega_3\chi_{kij}}{4}E_{j1}(0)E_{k2}(0)\exp\left(-\mathrm{i}\frac{1}{2}\Delta k z\right)\frac{\sin\left(\frac{1}{2}\Delta k z\right)}{\frac{1}{2}\Delta k} \\
&= \mathrm{i}\sqrt{\frac{\mu\epsilon_0}{\epsilon_3}}\frac{\omega_3\chi_{kij}}{4}E_{j1}(0)E_{k2}(0)\exp\left(-\mathrm{i}\frac{1}{2}\Delta k z\right)z\,\mathrm{sinc}\left(\frac{1}{2}\Delta k z\right). \qquad (8.22)
\end{aligned}$$

Remembering that the intensity of an EM wave is given by $I = (1/2Z)|E|^2$, where $Z = \sqrt{\mu/\epsilon}$ is the impedance of the medium, we see that the intensity of the generated wave is,

$$\begin{aligned}
I_{i3} &= \frac{\mu\epsilon_0}{\epsilon_3}\frac{\omega_3^2\left|\chi_{kij}\right|^2}{4}\frac{2Z_1 2Z_2}{2Z_3}\frac{\left|E_{j1}(0)\right|^2}{2Z_1}\frac{\left|E_{k2}(0)\right|^2}{2Z_2}z^2\,\mathrm{sinc}^2\left(\frac{1}{2}\Delta k z\right) \\
&= \frac{1}{2}\frac{Z_1 Z_2}{Z_3}\frac{\mu\epsilon_0}{\epsilon_3}\omega_3^2\left|\chi_{kij}\right|^2 I_{j1}I_{k2}\,z^2\,\mathrm{sinc}^2\left(\frac{1}{2}\Delta k z\right) \\
&= \frac{1}{2}\left(Z_1 Z_2 Z_3\right)\omega_3^2\epsilon_0^2\left|\chi_{kij}\right|^2 I_{j1}I_{k2}\,z^2\,\mathrm{sinc}^2\left(\frac{1}{2}\Delta k z\right) \\
\Rightarrow I_{i3} &= \frac{1}{2}\frac{1}{\epsilon_0 c^3}\frac{1}{n_1 n_2 n_3}\omega_3^2\left|\chi_{kij}\right|^2 I_{j1}I_{k2}\,z^2\,\mathrm{sinc}^2\left(\frac{1}{2}\Delta k z\right), \qquad (8.23)
\end{aligned}$$

where n_i is the refractive index experienced by wave i, and in deriving the last line we have assumed that the medium is non-magnetic so that $Z_i = \sqrt{\mu_0/\epsilon_0}(1/n_i) = (\epsilon_0 c n_i)^{-1}$.

As we would expect for a second-order process, the intensity at a given point z of the generated radiation is proportional to the product of the intensities of the two driving fields. Further, for non-zero Δk, the generated intensity will oscillate with propagation distance z, as shown in Fig. 8.1. For $\Delta k \neq 0$ the distance at which the generated intensity first reaches a maximum is $L_\mathrm{c} = \pi/\Delta k$, which is known as the *coherence length*.

8.2.1 Phase-matching

The oscillatory behaviour observed for $\Delta k \neq 0$ is undesirable since the generated intensity will also typically be small. Figure 8.2 shows the intensity of the generated beam (at a fixed point z) varies with the phase mis-match Δk. It can be seen that the maximum possible intensity will be generated if things can be arranged so that $\Delta k = 0$. This is known as *phase-matching*.

Achieving phase-matching, i.e. $\Delta k = 0$ amounts to ensuring that the phase velocity of the generated radiation matches the speed at which the nonlinear polarization propagates in the medium. If

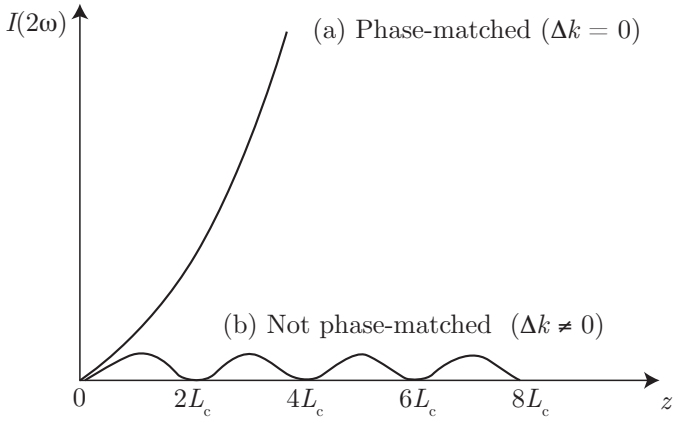


Figure 8.1: Variation of the intensity of the generated radiation with propagation distance z: (a) when phase-matching is achieved ($\Delta k = 0$); (b) when it is not ($\Delta k \neq 0$).

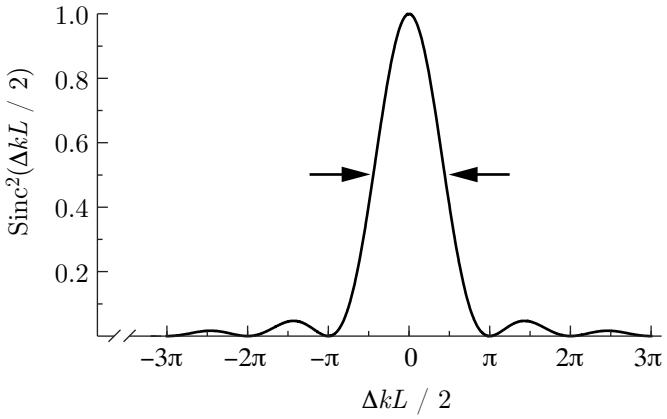


Figure 8.2: The function $\mathrm{sinc}^2\left(\frac{1}{2}\Delta k z\right)$. The arrows indicate the full-width at half-maximum.

this is achieved then radiation generated within the medium is generated in phase with that generated earlier in the medium.

Since nonlinear media are in general dispersive the phase velocities of the driving fields are generally not equal to those of the generated radiation. The usual way of achieving phase-matching in this case is to take advantage of the fact that nonlinear media are usually also birefringent. In this case two phase-matching geometries can be employed:

- **Type I:** in which the two driving waves are of the same type of polarization (i.e. both o-waves or both e-waves) and the generated wave has the other type of polarization. Using an obvious notation, the following are examples of Type I phase-matching: $\mathrm{o} + \mathrm{o} \rightarrow \mathrm{e}$, $\mathrm{e} + \mathrm{e} \rightarrow \mathrm{o}$

- **Type II:** in which the two driving waves have different types of polarization. The following are examples of Type II phase-matching: $\mathrm{o} + \mathrm{e} \rightarrow \mathrm{o}$, $\mathrm{o} + \mathrm{e} \rightarrow \mathrm{e}$.

8.3 Second-harmonic generation (SHG)

If the fields $E_j^{\omega_1}$ and $E_k^{\omega_2}$ have the same frequency, ω, then $\omega_3 = 2\omega$ and the polarization oscillates at the second harmonic. This generates the second-harmonic frequency of the driving radiation. This can be used, for example, to convert the output of a Nd:YAG laser operating at a wavelength of 1064 nm to a beam of wavelength 532 nm.

8.3.1 Phase-matching SHG

In general, phase-matching requires $\boldsymbol{\Delta k} = \boldsymbol{k}_3 - \boldsymbol{k}_1 - \boldsymbol{k}_2 = 0$. If the driving and generate beams are collinear then this becomes $\Delta k = k_3 - k_1 - k_2 = 0$. Now since in general $k = (\omega/c)n$, and $\omega_3 = 2\omega_1$, the condition for SHG phase-matching can be written as,

$$2n_3(2\omega_1) = n_1(\omega_1) + n_2(\omega_1). \tag{8.24}$$

Type I phase-matching

In Type I phase-matching the two driving waves are both o- or e-waves and so phase-matching requires $n_3(2\omega_1) = n_1(\omega_1)$. However, away from regions of absorption most materials exhibit positive dispersion (i.e. the refractive index increases with frequency) and so if the two frequencies have the same polarization we expect $n_3(2\omega_1) > n_1(\omega_1)$.

For crystals with negative birefringence (i.e. $n_\mathrm{e} - n_\mathrm{o} < 0$), the refractive index experienced by the second-harmonic can be decreased by choosing a geometry in which the second harmonic is an e-wave, and hence Type I phase-matching in such materials requires,

$$n_\mathrm{e}^{2\omega}(\theta_\mathrm{m}) = n_\mathrm{o}^{\omega} \tag{8.25}$$

Using eqn. (7.20), this condition can be written,

$$\begin{aligned}
\frac{1}{[n_\mathrm{e}^{2\omega}(\theta_\mathrm{m})]^2} = \frac{\cos^2\theta_\mathrm{m}}{(n_\mathrm{o}^{2\omega})^2} + \frac{\sin^2\theta_\mathrm{m}}{(n_\mathrm{e}^{2\omega})^2} &= \frac{1}{(n_\mathrm{o}^{\omega})^2} \\
\Rightarrow \frac{1-\sin^2\theta_\mathrm{m}}{(n_\mathrm{o}^{2\omega})^2} + \frac{\sin^2\theta_\mathrm{m}}{(n_\mathrm{e}^{2\omega})^2} &= \frac{1}{(n_\mathrm{o}^{\omega})^2} \\
\Rightarrow \sin^2\theta_\mathrm{m}\left[\frac{1}{(n_\mathrm{e}^{2\omega})^2} - \frac{1}{(n_\mathrm{o}^{2\omega})^2}\right] &= \frac{1}{(n_\mathrm{o}^{\omega})^2} - \frac{1}{(n_\mathrm{o}^{2\omega})^2} \\
\Rightarrow \sin^2\theta_\mathrm{m} &= \frac{(n_\mathrm{o}^{\omega})^{-2} - \left(n_\mathrm{o}^{2\omega}\right)^{-2}}{(n_\mathrm{e}^{2\omega})^{-2} - (n_\mathrm{o}^{2\omega})^{-2}}.
\end{aligned} \tag{8.26}$$

Positive birefringent crystal In this case phase-matching is achieved if $n_\mathrm{o}^{2\omega} = n_\mathrm{e}^{\omega}(\theta_\mathrm{m})$. The phase-matching angle can then be found by using the method above.

Type II phase-matching

In the Type II geometry one of the driving beams is an o-wave, and the other is an e-wave. The second-harmonic beam will be an o- or e-wave depending on whether the crystal exhibits positive or negative birefringence.

For a crystal with positive birefringence eqn. (8.24) becomes

$$\begin{aligned}
2n_{\rm o}^{2\omega} &= n_{\rm o}^{\omega} + n_{\rm e}^{\omega}(\theta_{\rm m}) \\
\Rightarrow \left(2n_{\rm o}^{2\omega} - n_{\rm o}^{\omega}\right)^{-2} &= \frac{1}{\left[n_{\rm e}^{\omega}(\theta_{\rm m})\right]^2} = \frac{\cos^2\theta_{\rm m}}{\left(n_{\rm o}^{\omega}\right)^2} + \frac{\sin^2\theta_{\rm m}}{\left(n_{\rm e}^{\omega}\right)^2} \\
\Rightarrow \left(2n_{\rm o}^{2\omega} - n_{\rm o}^{\omega}\right)^{-2} &= \frac{1}{\left(n_{\rm o}^{\omega}\right)^2} + \sin^2\theta_{\rm m}\left[\frac{1}{\left(n_{\rm e}^{\omega}\right)^2} - \frac{1}{\left(n_{\rm o}^{\omega}\right)^2}\right] \\
\Rightarrow \sin^2\theta_{\rm m}\left[\frac{1}{\left(n_{\rm e}^{\omega}\right)^2} - \frac{1}{\left(n_{\rm o}^{\omega}\right)^2}\right] &= \frac{1}{\left(2n_{\rm o}^{2\omega} - n_{\rm o}^{\omega}\right)^2} - \frac{1}{\left(n_{\rm o}^{\omega}\right)^2} = \frac{\left(n_{\rm o}^{\omega}\right)^2 - \left(2n_{\rm o}^{2\omega} - n_{\rm o}^{\omega}\right)^2}{\left(n_{\rm o}^{\omega}\right)^2\left(2n_{\rm o}^{2\omega} - n_{\rm o}^{\omega}\right)^2} = 4\frac{n_{\rm o}^{2\omega}n_{\rm o}^{\omega} - \left(n_{\rm o}^{2\omega}\right)^2}{\left(n_{\rm o}^{\omega}\right)^2\left(2n_{\rm o}^{2\omega} - n_{\rm o}^{\omega}\right)^2} \\
\Rightarrow \sin\theta_{\rm m} &= 2\frac{\left[n_{\rm o}^{2\omega}n_{\rm o}^{\omega} - \left(n_{\rm o}^{2\omega}\right)^2\right]^{1/2}}{n_{\rm o}^{\omega}\left(2n_{\rm o}^{2\omega} - n_{\rm o}^{\omega}\right)}\left[\frac{1}{\left(n_{\rm e}^{\omega}\right)^2} - \frac{1}{\left(n_{\rm o}^{\omega}\right)^2}\right]^{-1/2} \\
\Rightarrow \sin\theta_{\rm m} &= \frac{2n_{\rm e}^{\omega}}{2n_{\rm o}^{2\omega} - n_{\rm o}^{\omega}}\left[\frac{n_{\rm o}^{2\omega}\left(n_{\rm o}^{2\omega} - n_{\rm o}^{\omega}\right)}{\left(n_{\rm e}^{\omega}\right)^2 - \left(n_{\rm o}^{\omega}\right)^2}\right]^{1/2}. \qquad (8.27)
\end{aligned}$$

The phase-matching angle for Type II phase-matching in a negatively birefringent crystal can be obtained by following the same method.

8.3.2 The effective nonlinear coefficient

The amplitude of the component of polarization oscillating at 2ω can be written as,

$$P_i(2\omega) = \epsilon_0 d_{ijk}^{(2)} E_j(\omega) E_k(\omega) \exp\left(\mathrm{i}\Delta k \cdot \boldsymbol{r}\right). \qquad (8.28)$$

Remembering that $d_{xyz}E_y(\omega)E_z(\omega) = d_{xzy}E_z(\omega)E_y(\omega)$, this can be written in matrix form:

$$\begin{pmatrix} P_x(2\omega) \\ P_y(2\omega) \\ P_z(2\omega) \end{pmatrix} = \begin{pmatrix} d_{xxx} & d_{xyy} & d_{xzz} & d_{xyz} & d_{xxz} & d_{xxy} \\ d_{yxx} & d_{yyy} & d_{yzz} & d_{yyz} & d_{yxz} & d_{yxy} \\ d_{zxx} & d_{zyy} & d_{zzz} & d_{zyz} & d_{zxz} & d_{zxy} \end{pmatrix} \begin{pmatrix} E_x^2(\omega) \\ E_y^2(\omega) \\ E_z^2(\omega) \\ 2E_y(\omega)E_z(\omega) \\ 2E_x(\omega)E_z(\omega) \\ 2E_x(\omega)E_y(\omega) \end{pmatrix}. \qquad (8.29)$$

As we found in Section 7.3.1, it is often useful to use a contracted notation in which $x = 1$, $y = 2$, $z = 3$, $xx = 1$, $yy = 2$, $zz = 3$, $yz = 4$, $xz = 5$, and $xy = 6$. Then the above can be written as,

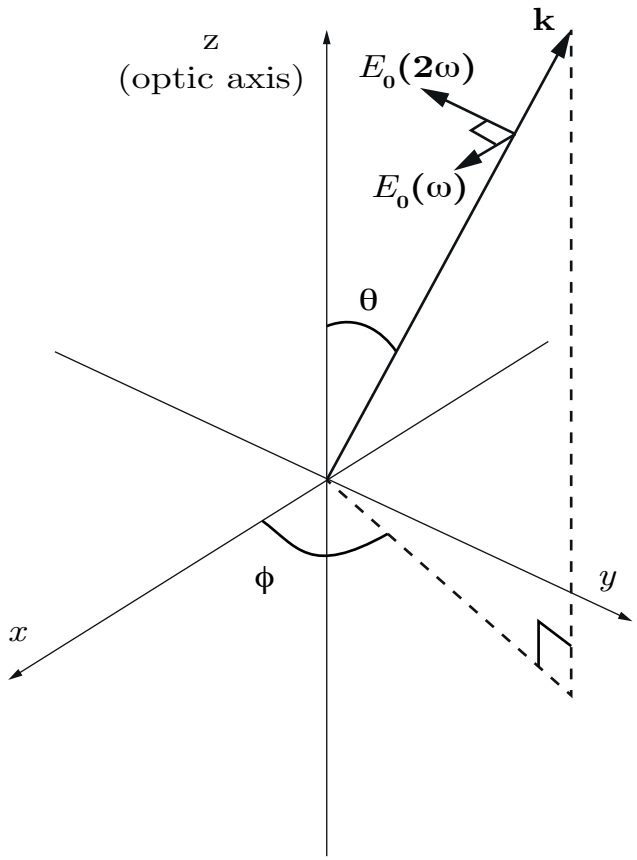


Figure 8.3: Geometry employed in finding the effective nonlinear coefficient for the case of Type I (o + o → e) phase-matching (eqn (8.39)).

$$\begin{pmatrix} P_1(2\omega) \\ P_2(2\omega) \\ P_3(2\omega) \end{pmatrix} = \begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} \end{pmatrix} \begin{pmatrix} E_1^2(\omega) \\ E_2^2(\omega) \\ E_3^2(\omega) \\ 2E_2(\omega)E_3(\omega) \\ 2E_1(\omega)E_3(\omega) \\ 2E_1(\omega)E_2(\omega) \end{pmatrix}. \tag{8.30}$$

For a given geometry the nonlinear polarization depends on a sum over products of the components of the input fields and the nonlinear tensor $\overline{\overline{\chi}}^{(2)}$ or $\overline{\overline{d}}$ (see, for example, eqn. (8.5)). It is often easiest to work with an *effective nonlinear coefficient* $d_{\rm eff}$ which contains all these sums. For example, the second-order nonlinear polarization can be written in the form,

$$P(\omega_3) = d_{\rm eff} E(\omega_1) E(\omega_2). \tag{8.31}$$

As an example, consider frequency doubling in ADP. For this crystal the only non-zero components of $\overline{\overline{d}}$ are d_{14}, d_{25}, and d_{36} so that eqn. (8.30) becomes,

$$\begin{pmatrix} P_1(2\omega) \\ P_2(2\omega) \\ P_3(2\omega) \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{25} & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{36} \end{pmatrix} \begin{pmatrix} E_1^2(\omega) \\ E_2^2(\omega) \\ E_3^2(\omega) \\ 2E_2(\omega)E_3(\omega) \\ 2E_1(\omega)E_3(\omega) \\ 2E_1(\omega)E_2(\omega) \end{pmatrix} = \begin{pmatrix} 2d_{14}E_2(\omega)E_3(\omega) \\ 2d_{25}E_1(\omega)E_3(\omega) \\ 2d_{36}E_1(\omega)E_2(\omega) \end{pmatrix}. \tag{8.32}$$

For Type I phase-matching, and with the second harmonic propagating as an extraordinary wave (i.e. oo $\rightarrow$ e), then $P_1(2\omega) = P_2(2\omega) = 0$ and we are left with,

$$P_3(2\omega) = 2d_{36}E_1(\omega)E_2(\omega) \tag{8.33}$$

$$\text{or,} \quad P_z(2\omega) = 2d_{36}E_x(\omega)E_y(\omega). \tag{8.34}$$

Using the geometry of Fig. 8.3,

$$E_x = E_0(\omega)\sin\phi \tag{8.35}$$

$$E_y = -E_0(\omega)\cos\phi, \tag{8.36}$$

so that,

$$P_z(2\omega) = -2d_{36}E_0^2(\omega)\sin\phi\cos\phi \tag{8.37}$$

$$= -d_{36}E_0^2(\omega)\sin 2\phi. \tag{8.38}$$

Now, the component of the nonlinear polarization parallel to the generated second harmonic wave is $P_z \sin\theta$, and hence,

$$d_{\text{eff}} = -d_{36}E_0^2(\omega)\sin 2\phi \sin\theta. \tag{8.39}$$

The second harmonic intensity is therefore maximzed when:

1. The generated wave propagates at the phase-matching angle, i.e. $\theta = \theta_{\text{m}}$
2. $\phi = \pi/4$, which makes sense since then the *product* of the x- and y-components of the driving field is maximized.

8.3.3 Further comments

A question of key practical importance is how accurately must the crystal be oriented to achieve phase-matching? In other words, what is the tolerance to small errors in the crystal orientation.

It can be shown that the angular tolerance $\Delta\theta \propto 1/\sin 2\theta_{\rm m}$. In practice this tolerance is low unless $\theta_{\rm m} \approx \pi/2$, whereupon it becomes large. Achieving a phase-matching angle close to 90-degrees is highly desirable in practical systems, and is known as *non-critcial phase-matching*. It can be obtained, for example, by adjusting the temperature of the crystal — which will change $n_{\rm e}$ and $n_{\rm o}$ by different amounts, and by differing amounts for each frequency.

A further consideration is beam walk-off. This arises because, for the e-wave, the direction of energy flow (given by the pointing vector) is not parallel to the wave vector. A consequence of this is that the e- and o-waves do not propagate parallel to each other, so that the distance over which they are overlapped is limited. In practice this can limit the useful length of the crystal even if good phase-matching is achieved.

8.4 Three-wave mixing

SHG is a special case of the more general process in which waves of frequencies ω_1 and ω_2 are mixed to form a wave of frequency ω_3. Both sum-frequency generation and difference-frequency generation is possible, as can be seen by inspecting eqn (8.9).

8.4.1 Sum-frequency generation

This corresponds to,

$$\omega_3 = \omega_1 + \omega_2 \tag{8.40}$$

$$\text{or,} \quad \frac{1}{\lambda_3} = \frac{1}{\lambda_1} + \frac{1}{\lambda_2}, \tag{8.41}$$

where the wavelengths λ_i are the wavelengths in vacuo.

The condition for collinear phase-matching is simply,

$$\frac{n_3}{\lambda_3} = \frac{n_1}{\lambda_1} + \frac{n_2}{\lambda_2}. \tag{8.42}$$

8.4.2 Difference-frequency generation

This corresponds to,

$$\omega_3 = \omega_1 - \omega_2 \tag{8.43}$$

$$\text{or,} \quad \frac{1}{\lambda_3} = \frac{1}{\lambda_1} - \frac{1}{\lambda_2}, \tag{8.44}$$

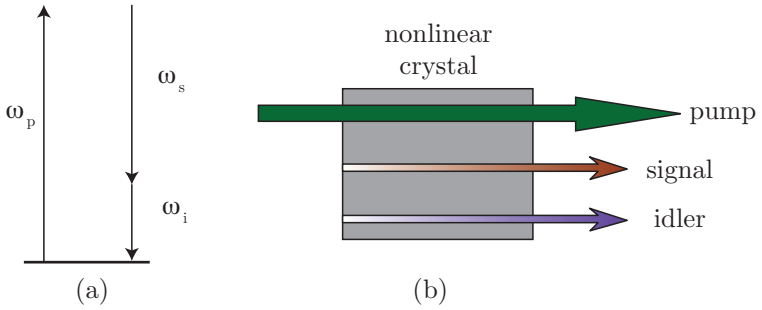


Figure 8.4: Parametric down-conversion. (a) shows the relation between the pump (ω_p), signal (ω_i) and idler (ω_i) frequencies. (b) illustrates how the intensities of the signal and idler beams increase with propagation through the nonlinear crystal.

where we have assumed $\omega_1 > \omega_2$. The condition for collinear phase-matching is,

$$\frac{n_3}{\lambda_3} = \frac{n_1}{\lambda_1} - \frac{n_2}{\lambda_2}. \tag{8.45}$$

Parametric down conversion

Difference frequency generation refers to the process in which two driving beams of angular frequencies ω_1 and ω_2 are mixed in a nonlinear crystal to generate a beam at $\omega_3 = \omega_1 - \omega_2$.

Parametric down conversion is a related process in which a strong pump beam, of frequency ω_p, is converted into a beam of lower frequency, ω_s, together with a second beam of frequency ω_i such that,

$$\omega_\mathrm{p} = \omega_\mathrm{s} - \omega_\mathrm{i} \tag{8.46}$$

$$\text{or,} \quad \frac{1}{\lambda_\mathrm{p}} = \frac{1}{\lambda_\mathrm{s}} - \frac{1}{\lambda_\mathrm{i}}. \tag{8.47}$$

This process is illustrated in Fig. 8.4. The beam at ω_s is known as the *signal* and that at ω_i is the *idler*.[1] Both signal and idler beams are generated from noise and grow through the three-wave mixing process.

In principle there are an infinite number of combinations of ω_s and ω_p which satisfy the energy conservation condition above. However, for efficient generation of either the signal or idler beams it is necessary also to satisfy the (collinear) phase-matching condition:

$$\frac{n_\mathrm{p}}{\lambda_\mathrm{p}} = \frac{n_\mathrm{s}}{\lambda_\mathrm{s}} - \frac{n_\mathrm{i}}{\lambda_\mathrm{i}}. \tag{8.48}$$

Phase-matching therefore determines the wavelength of the signal beam (and hence that of the idler).

[1] By convention the signal has the higher frequency.

Parametric down conversion is important since it can generate coherent beams at wavelengths at which there are no or few lasers, such as the far infrared or terahertz regions.

A second case of importance is that of *degenerate down conversion*, which arises when $\omega_\mathrm{s} = \omega_\mathrm{i}$. This is important in quantum information processing since it provides a means of generating pairs of indistinguishable photons.

8.5 Appendix: Missing steps in derivation of the nonlinear wave equation

Here we provide the missing steps in the derivation of the nonlinear wave equation. The left-hand side of eqn (8.15) can be written as,

$$\frac{\partial^2 E_i}{\partial z^2} = \frac{\partial^2}{\partial z^2}\left[\frac{E_{i3}(z)}{2}\exp\left[\mathrm{i}\left(k_3 z - \omega_3 t\right)\right] + \text{c.c.}\right] \tag{8.49}$$

$$= \frac{1}{2}\left[-k_3^2 E_{i3}(z) + 2\mathrm{i}k_3\frac{\partial E_{i3}}{\partial z} + \frac{\partial^2 E_{i3}}{\partial z^2}\right]\exp\left[\mathrm{i}\left(k_3 z - \omega_3 t\right)\right] + \text{c.c.}. \tag{8.50}$$

We can then make the slowly-varying envelope approximation (analogous to the approximations (1.39) and (1.40)) and neglect the term $\partial^2 E_{3k}/\partial z^2$. Combining this with eqn (8.18), eqn (8.15) becomes,

$$\frac{1}{2}\left[-k_3^2 E_{i3}(z) + 2\mathrm{i}k_3\frac{\partial E_{i3}}{\partial z}\right]\exp\left(\mathrm{i}k_3 z\right) + \text{c.c.} \tag{8.51}$$

$$= \mu\epsilon_0\epsilon_\mathrm{r}(\mathrm{i}\omega_3)^2\frac{E_{i3}(z)}{2}\exp\left(\mathrm{i}k_3 z\right) + \mu\epsilon_0\frac{\chi_{kij}}{4}E_{j1}(z)E_{k2}(z)(-\mathrm{i}\omega_3)^2\exp\left[\mathrm{i}\left(k_1+k_2\right)z\right] + \text{c.c.}$$

$$\Rightarrow \mathrm{i}k_3\frac{\partial E_{i3}}{\partial z} = \left[k_3^2 - \mu\epsilon_0\epsilon_\mathrm{r}\omega_3^2\right]\frac{E_{i3}(z)}{2} + \mu\epsilon_0\frac{\chi_{kij}}{4}E_{j1}(z)E_{k2}(z)(-\mathrm{i}\omega_3)^2\exp\left[\mathrm{i}\left(k_1+k_2-k_3\right)z\right]. \tag{8.52}$$

Now the dispersion relation for the wave at ω_3 is $\omega_3/k_3 = (\mu\epsilon_0\epsilon_\mathrm{r})^{-1/2}$ and so we are left with,

$$\frac{\partial E_{i3}}{\partial z} = \mathrm{i}\frac{\mu\epsilon_0\chi_{kij}\omega_3^2}{4k_3}E_{j1}(z)E_{k2}(z)\exp\left(-\mathrm{i}\Delta k z\right) \tag{8.53}$$

$$= \mathrm{i}\sqrt{\frac{\mu\epsilon_0}{\epsilon_3}}\frac{\omega_3\chi_{kij}}{4}E_{j1}(z)E_{k2}(z)\exp\left(-\mathrm{i}\Delta k z\right). \tag{8.54}$$

where $\epsilon_3 = \epsilon_\mathrm{r}$. This equation describes the growth with z of the frequency generated at ω_3.