THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

Physics 333 — Lasers and EM Waves

2020

Dr Miro Erkintalo
Physics Department
Room 505
m.erkintalo@auckland.ac.nz
ext. 85598

Lectures are Tuesday, Wednesday, and Friday 10-11 am, in Rm 439-G10 (Tue, Fri) or 303-B09 (Wed)

Week	Date		**Lecturer**	**Topic**	**Assessment Dates**
1	Mon	27 July			
	Tue	28	ME	Introduction	
	Wed	29	ME	Basic laser idea	
	Thu	30			
	Fri	31	ME	Tutorial 1	
2	Mon	3 Aug.			
	Tue	4	ME	Light-matter interactions	**Assignment 1 out**
	Wed	5	ME	Rate equations	
	Thu	6			
	Fri	7	ME	Line broadening	
3	Mon	10			
	Tue	11	ME	Maxwell's equations	
	Wed	12	ME	Plane waves I	
	Thu	13			
	Fri	14	ME	Tutorial 2	
4	Mon	17			
	Tue	18	ME	Plane waves II	
	Wed	19	ME	Electron oscillator I	**Assignment 1 due**
	Thu	20			**Assignment 2 out**
	Fri	21	ME	Electron oscillator II	
5	Mon	24			
	Tue	25	ME	Resonators	
	Wed	26	ME	Fabry-Perot lasers I	
	Thu	27			
	Fri	28	ME	Tutorial 3	
6	Mon	31			
	Tue	1 Sept.	ME	Fabry-Perot lasers II	**Assignment 2 due**
	Wed	2	ME	Modelocking	
	Thu	3			
	Fri	4	ME	**TEST 1**	
Mid Semester Break: Mon 7 September – Friday 18 September					
7	Mon	21			
	Tue	22	RL	ABCD propagation	**Assignment 3 out**
	Wed	23	RL	Gaussian beams I	
	Thu	24			
	Fri	25	RL	Tutorial 4	
8	Mon	28			
	Tue	29	RL	Gaussian beams II	
	Wed	30	RL	Polarization optics I	
	Thu	1 Oct.			
	Fri	2	RL	Polarization optics II	
9	Mon	4			
	Tue	6	RL	Basics of Nonlinear Optics	
	Wed	7	RL	Survey of NL optical effects	**Assignment 3 due**
	Thu	8			**Assignment 4 out**
	Fri	9	RL	Tutorial 5	
10	Mon	12			
	Tue	13	RL	NL susceptibilities	
	Wed	14	RL	Second-harmonic generation	
	Thu	15			
	Fri	16	RL	Phase matching	
11	Mon	19			
	Tue	20	RL	Applications of NL optics	**Assignment 4 due**
	Wed	21	RL	Coherence I	
	Thu	22			
	Fri	23	RL	**TEST 2**	
12	Mon	26			
	Tue	27	RL	Coherence II	
	Wed	28	RL	Summary & Hot topics	
	Thu	29			
	Fri	30	RL	Tutorial 6	
Study Break/Exam Period: Monday 2 November – Saturday 21 November					

ME = Dr Miro Erkintalo (m.erkintalo@auckland.ac.nz), Rm 303.505, 09-923.5598
RL = A.-Prof. Rainer Leonhardt (r.leonhardt@auckland.ac.nz), Rm.303.507, 09-923.8835

Contents

1 Introduction

These lectures notes discuss selected topics on (modern) optics and lasers physics. The topics span both fundamental and applied directions, aiming to (i) elucidate the characteristics and behaviours of light in terms of classical electromagnetism, and (ii) to describe the physics underpinning various optical devices that are key to our modern society.

Lasers arguably represent (one of) the most important application(s) of modern optics. They are used in every nook and cranny of modern society, and they play a key role in our everyday lives. Furthermore, many of the other important "photonic" technologies (such as e.g. optical fibres and waveguides), are almost predominantly used with lasers. Compounded by the fact that the physics behind lasers already encapsulates numerous key electromagnetic concepts (e.g. interaction of light and matter, Gaussian beams,...) and optical devices (e.g. resonators, optical amplifiers,...), we will use lasers as an overarching theme to guide our discussion.

Following the philosophy outlined above, we begin by discussing the **basic operation principles behind optical amplifiers and lasers**, starting from the **interactions between light and matter**. These interactions will be examined from two different perspectives: Einstein's photon picture and classical electromagnetism described by **Maxwell's equations** and its **plane wave** solutions. We will then discuss the physics of **optical resonators**, which not only underpin laser operation, but also have many other applications. Subsequently, we will discuss the properties of laser light outside of the resonator. In particular, we will describe the behaviour of so-called **Gaussian beams**, which will allow us to analyse how realistic (laser) light behaves, diffracts, and propagates through optical systems. When (Gaussian) laser beams are tightly focussed, very large electric field amplitudes can be reached. In such situations, we find that light no longer interacts linearly with matter. We will discuss the basic physics and applications of such **nonlinear optical** phenomena. Finally, we will close by discussing briefly methods and devices that allow for the state of **polarisation** of electromagnetic waves to be modified, and we will discuss how fluctuations of electromagnetic waves can be described and analysed using **coherence theory**.

At the moment, these lecture notes cover almost all of the course topics, with the only exceptions being (in 2020) methods of polarisation manipulation, which will be covered in another set of material. (Also, notes pertaining to Gaussian beams, nonlinear optics, and coherence will be handed out at a later date.) There are a number of excellent references from which the lectures, as well as the summarized lecture material contained in this leaflet are derived:

Suggested reading

- *Fundamentals of Photonics* by B. E. A. Saleh and M. C. Teich. Available online via UoA library.
- *Laser Physics* by S. Hooker and C. Webb. Online material available via clever googling.
- *Laser Physics* by P. W. Milonni and J. H. Eberly. Available online via UoA library.
- *Optics* by E. Hecht
- *Physics 326 Lecture Notes* by S. Coen.
- *Physics 726 Lecture Notes* by S. Coen.

2 Special characteristics of laser light

Lasers are devices that emit coherent light through a process of optical amplification via stimulated emission of radiation. They play a key role in our everyday lives, with applications ranging from long-distance telecommunications and bar-code scanners to machining and manufacturing. Their usefulness originates from the many attractive and unique characteristics possessed by laser light compared to other forms of light (generated by e.g. light bulbs or stars). These characteristics include:

1. **Monochromaticity:** Lasers can produce very pure colours, i.e., electromagnetic waves with a very precise frequency. Light emitted by the sun or a light bulb is white, containing frequencies spanning hundreds of THz. In contrast, laser light can have a bandwidth as small as 1 Hz [see Fig. 1].

2. **Directionality:** While most light sources emit light in all directions, lasers can create narrow unidirectional beams that diffract (i.e., broaden) extremely slowly. In radians, the divergence (half-)angle of a diffraction-limited laser beam is [see Section 8]:

$$\theta \approx \frac{\lambda}{\pi w_0}, \tag{2.1}$$

   where λ is wavelength and w_0 is the initial width of the beam. For a typical red laser with $\lambda = 633\ \mathrm{nm}$ and spot size $w_0 = 5\ \mathrm{mm}$, one obtains $\theta \approx 4 \times 10^{-5}\ \mathrm{rad}$. One can easily confirm that a propagation length of more than 120 metres is needed for the spot size to double.

3. **Coherence:** The phases of the electromagnetic waves making up laser beams have fixed relationships in space (spatial coherence) and time (temporal coherence). In contrast, the phases of EM waves corresponding to other forms of light exhibit random fluctuations [see Fig. 2].

   Temporal coherence can be quantified by the coherence time τ_c, which describes the time over which the phase/amplitude does not wander significantly. One finds that the coherence time is inversely proportional to the bandwidth: $\tau_c = \Delta f^{-1}$. Because of their narrow bandwidths, laser light can maintain constant phase for very long times. For example, a laser diode with $\Delta f = 1\ \mathrm{MHz}$ would have a coherence time $\tau_c = 1\ \mu\mathrm{s}$, corresponding to more than 300 million oscillations of the electric field. In contrast, sunlight

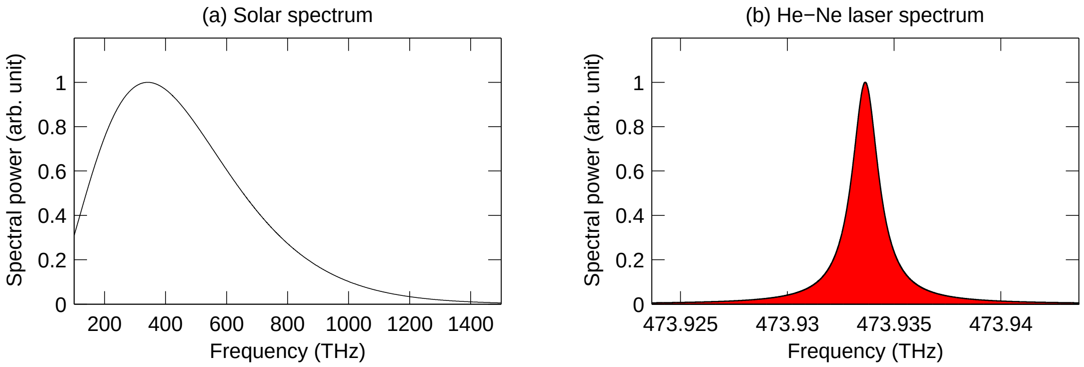

Figure 1: Comparison between spectral intensities of light emanating from (a) the Sun and (b)a He-Ne laser. Note the different x-axis scales. (a) Corresponds to the spectrum of a black body at $T = 5800\ \mathrm{K}$.

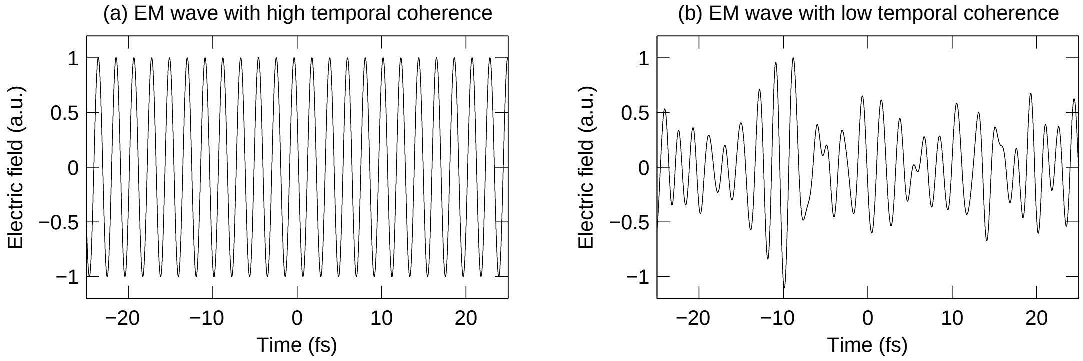

Figure 2: Electric fields for an EM wave with (a) high and (b) low coherence. The spectral bandwidth in (a) is 1 GHz, whilst in (b) the bandwidth is 100 THz.

has a bandwidth $\Delta f \sim 500$ THz and a corresponding coherence time $\tau_c = 2$ fs, which amounts to about one electric field oscillation (for visible light).

4. **Brightness/intensity:** Because of their small beam size and ability to be focussed to very small spots (thanks to spatial coherence), lasers can concentrate light energy very tightly, resulting in extreme brightness / high intensities[1].

Example.
The total solar intensity *on earth* is

$$\frac{P_{\text{sun}}}{A} \approx 1380 \frac{\text{W}}{\text{m}^2}, \tag{2.2}$$

while the intensity of a 1 mW laser focussed to a 80 μm (radius) spot is

$$\frac{P_{\text{laser}}}{A} \approx 50000 \frac{\text{W}}{\text{m}^2}. \tag{2.3}$$

5. **Short durations:** Lasers can create very short bursts of light, i.e. pulses, with durations well below 100 fs (10^{-15} s) [see Fig. 3]. In fact, such ultrashort laser pulses correspond to the shortest events created by humankind. Because of their short duration, ultrashort pulses can also possess extremely large peak powers.

[1] Intensity is the rate of energy transfer per unit area, i.e., power per area.

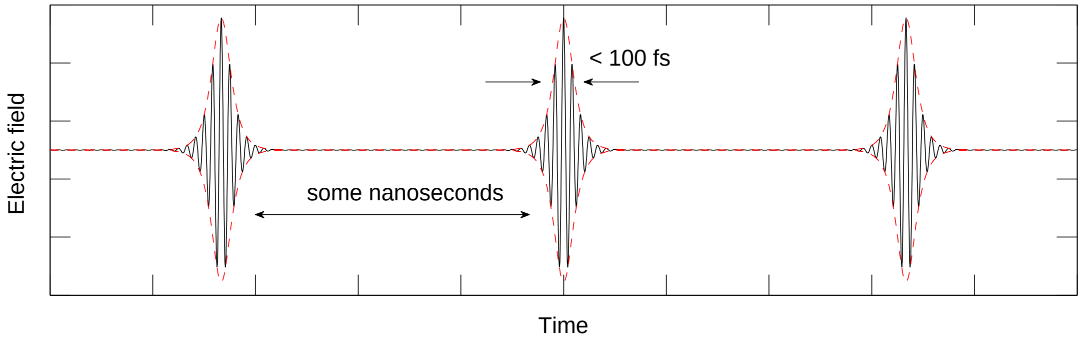

Figure 3: Schematic illustration of the electric field of an EM wave corresponding to a train of ultrashort pulses. Black curve illustrates the electric field, while the red dashed curve depicts the *envelope*.

Example.
Commercial Ti:Sapphire lasers can readily produce $\Delta\tau = 100$ fs pulses at a repetition rate $f_{\text{rep}} = 80$ MHz (i.e., 80 million pulses per second) with average power $P_{\text{avg}} = 1$ W. This corresponds to a peak power (per pulse) of $P_{\text{peak}} = 125$ kW. Assuming a 5 mm beam spot, the corresponding peak intensity is about

$$\frac{P_{\text{peak}}}{A} \approx 6400\,\frac{\text{MW}}{\text{m}^2}. \tag{2.4}$$

This is 100 times larger than the total solar intensity on the surface of the Sun.

All of these distinctive characteristics ensue from the fact that laser light originates from a fundamentally different physical process than any other form of light. Specifically, whereas light generated in stars (e.g., the Sun) or heated filaments (e.g., light bulbs) arise from *spontaneous* emission of photons by excited atoms/molecules, laser light originates from a process known as *stimulated* emission. Hence the name: **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation. As we shall see in Section 3, spontaneous emission results in the generation of photons with random traits (phase, frequency, polarisation, direction), while stimulated emission results in the generation of photons with identical characteristics. It should be evident that this fact (namely the creation of photons with identical phase, frequency, direction) is sufficient to explain the unique characteristics of laser light listed above.

2.1 Ultrashort history of lasers

The story of the laser arguably begins with the quantum hypothesis put forward by Max Planck around 1900. In particular, he proposed that atoms can only absorb and emit radiation in discrete quanta. As we shall see below, this behaviour is paramount for laser operation.

The concept of stimulated emission, which underpins laser operation, was hypothesised in 1917 by Albert Einstein. Classical electromagnetism could only account for the processes of absorption and spontaneous emission (as explained in Section 5), which had left stimulated emission undiscovered. Einstein considered an atom in thermal equilibrium with a black body radiation field, and showed that Plank's law (of black body radiation) could be re-derived only if the atom could undergo the hypothesised process of stimulated emission (in addition to absorption and spontaneous emission).

Stimulated emission was first realised in the laboratory not in the optical but in the microwave domain. The idea for a device utilising stimulated emission to realise coherent sources of microwave radiation (MASER: microwave amplification by stimulated emission of radiation) was first conceived (independently) by Joseph Weber, Charles Townes, Nikolay Basov, and Aleksandr Prokhorov around 1950. The first laboratory demonstration of a maser was reported in 1954 by Townes, together with J. P. Gordon and H. J. Zeiger.

Conceptually, lasers can be considered as masers operating in the optical domain. Their practical construction is, however, technologically different. In 1957, Gordon Gould conceived the idea of building a laser using an open-sided Fabry-Perot resonator (see section 6). Similar ideas were simultaneously pursued by Townes together with A. Schawlow.

The first actual laser was built and demonstrated by Theodore Maiman in 1960. Using a ruby crystal pumped by photographic flash lamps, he realised pulsed laser operation with a wavelength of 694.3 nm (red). From there onwards, numerous different lasers and laser technologies were quickly developed.

In 1964, Townes Basov, and Prokhorov won the Nobel Prize in Physics "for fundamental work in the field of quantum electronics, which has led to the construction of oscillators and amplifiers based on the maser-laser principle."

Problems

2.1 The wavelength of a Neodymium-YAG laser is 1064 nm, and the diameter of its beam is 5 mm. The beam has full coherence and is diffraction limited. What is the diameter of the beam at a 100 km distance from the laser? What should be done to the laser beam in order to significantly reduce its diameter at distance of 100 km?

2.2 The beam from a ruby laser ($\lambda = 694$ nm) is sent to the moon after passing through a telescope of 1 m diameter. Calculate the approximate value of beam diameter on the moon assuming that the beam has perfect spatial coherence (the distance between earth and moon is approximately 384000 km).

2.3 The specifications of a commercial Ti:Sapphire laser (Spectra Physics Tsunami) reveal that the device can readily produce pulses with a duration of $\Delta\tau = 80$ fs at a repetition rate $f_{\text{rep}} = 82$ MHz (i.e., 82 million pulses per second) with average power $P_{\text{avg}} = 1$ W. Give an estimate for the peak power of the individual pulses. **Hint**: you can solve this problem without any actual knowledge of lasers, by just using your brains and a little bit of dimensional analysis.

2.4 The 2014 Nobel prize was awarded for the invention of efficient blue light-emitting diodes. These diodes are based on Gallium Nitride, which has a bandgap energy of about 3.4 eV. What is the frequency and wavelength of light emitted when an electron transitions across the bandgap in Gallium Nitride? What is the colour of the light? **Hint:** $1\text{eV} = 1.602 \cdot 10^{-19}$ J.

2.5 The total intensity (power per unit surface area) radiated by a black body across all wavelengths is given by Stefan-Boltzmann law:

$$\frac{P}{A} = \sigma T^4,$$

where T is the temperature of the black body (in Kelvins) and the Stefan-Boltzmann constant $\sigma = 5.67 \times 10^{-8}\mathrm{Wm^{-2}K^{-4}}$. The emission spectrum of the Sun is approximately that of a black body at temperature $T = 5800$ K.

(a) Show that the solar radiation intensity reaching the surface of the Earth is approximately $1380\ \mathrm{Wm^{-2}}$. **Hint:** No knowledge of laser physics is needed, you just need to use your brains!

(b) Considering a standard laser pointer with a power of 1 mW, at what beam radius does the laser's intensity match that of the solar radiation on Earth?

(c) The spectral bandwidth of solar radiation is about 500 THz, while the bandwidth of a typical laser pointer is 1 GHz. How does the intensity **per unit frequency** of a laser pointer compare with that of solar radiation?

(d) Although the sun emits approximately a continuous range of frequencies (as a black body does), measurements of the solar spectrum on Earth reveals dark features today referred to as Franunhofer lines. Explain the physical origins of these lines.

3 Interaction of light and atoms

The energies of (electrons in) atoms, ions, or molecules are quantized to discrete values: they can only occupy discrete energy levels [see Fig. 4]. The first experimental evidence of this fact was the observation, made already in the 19th century, that a set of spectral lines (now referred to as Fraunhofer lines) were missing from the solar emission spectrum. Full theoretical explanation for the discreteness of atomic energy levels came with the development of quantum mechanics and the Schrödinger wave equation. The existence of discrete energy levels is key to laser operation.

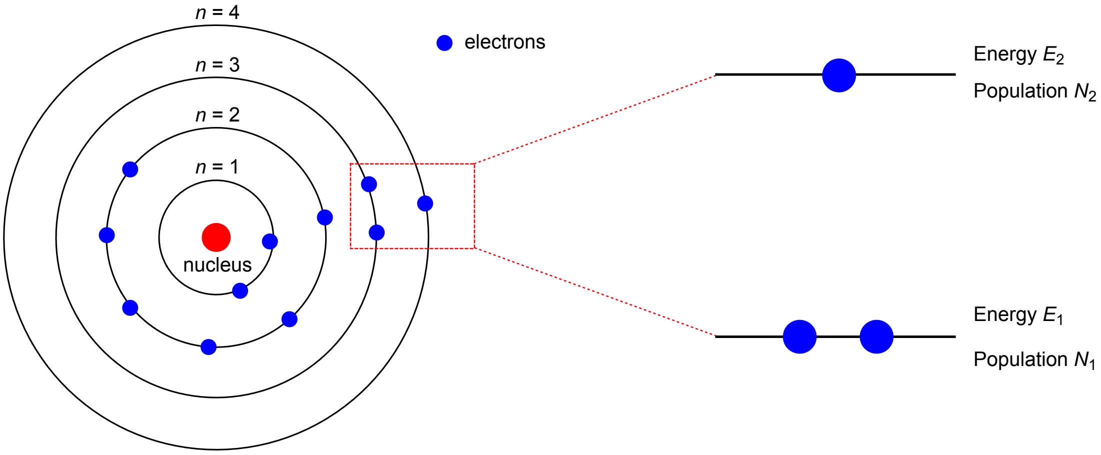

Figure 4: Schematic illustration of discrete energy levels of atoms and molecules.

There are three processes through which light can mediate transitions between atomic energy levels [see Fig. 5]: spontaneous emission, absorption, and stimulated emission. They can be intuitively described in the photon picture, where light at angular frequency ω is understood to comprise of zero-mass particles (photons) with energy $\hbar\omega$, where $\hbar = h/(2\pi)$ is the reduced Planck's constant[2].

Below we will describe each of these processes, considering an ensemble of atoms and a single transition involving two energy levels. The total population density of atoms is denoted N_{tot}, while the population densities in the upper and lower (energy) levels are denoted N_2 and N_1, respectively (the units are $[N_i] = \text{atoms}/\text{m}^3$). For brevity, we will be referring to these quantities as populations, yet emphasise that "density" should be understood.

3.1 Spontaneous emission

In spontaneous emission, an atom in an excited state[3] spontaneously decays to a lower energy state, while simultaneously emitting a photon with energy $\hbar\omega_{21} = E_2 - E_1$, where E_2 and $E_1 < E_2$ are the energies of the atomic states. The emitted photon has random phase, is emitted in a random direction, and at a random time.

Spontaneous emission is behind almost all forms of light, including sun light and artificial lights. (The only real exception is laser light.) It is so prevalent, that it has several names (depending on the mechanism that has excited the atom in the first place). These include luminescence, fluorescence, and phosphorescence. The

[2] The wave picture of light will be discussed in section 4.

[3] An atom is said to be in an excited state if at least one of its electrons sits at an energy level above the ground state.

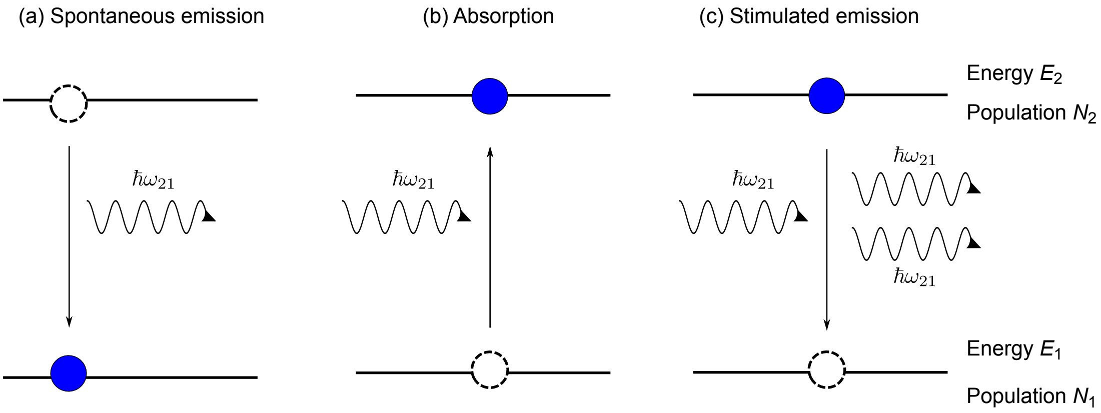

Figure 5: Schematic illustration of discrete energy levels of atoms and molecules.

inherent randomness of spontaneous emission explains the lack of coherence and directionality of these forms of light.

Like radioactive decay, spontaneous emission occurs randomly: the probability that an excited atom will undergo spontaneous emission during any time interval $t ... t + \mathrm{d}t$ is constant, and does not depend on how long the atom has been excited (t). Denoting this probability as $p = A_{21}\,\mathrm{d}t$, we may write

$$p = A_{21}\,\mathrm{d}t = \frac{N_2(t) - N_2(t + dt)}{N_2} = \frac{-dN_2}{N_2}, \tag{3.1}$$

where dN_2 is the change in population N_2. Rearranging yields:

$$\boxed{\left(\frac{dN_2}{dt}\right)_{\mathrm{sp}} = -A_{21}N_2.} \tag{3.2}$$

Thus, the population of the upper state decays exponentially,

$$N_2(t) = N_2(0)e^{-A_{21}t}. \tag{3.3}$$

The coefficient $A_{21} > 0$ above is known as the **Einstein A-coefficient**, and it corresponds to the transition probability per unit time. Its value depends on the characteristics of the atomic transition (i.e., of the energy levels involved), but is independent of the radiation field (number of photons present); spontaneous emission occurs whether or not photons are present.

The inverse of A_{21} describes the radiative lifetime of the transition: $\tau_{21}^{\mathrm{rad}} = A_{21}^{-1}$. The probability that an atom decays from the excited state (2) to the lower state (1) via spontaneous emission in time $t = \tau_{21}^{\mathrm{rad}}$ is $1 - e^{-1}$.

3.2 Absorption

In the presence of a photon with energy $\hbar\omega_{21} = E_2 - E_1$, an atom in the lower level can be excited to an upper level by the absorption (annihilation) of the photon.

Like spontaneous emission, probability per unit time is constant. However, unlike for spontaneous emission, the probability of absorption depends on the radiation field (number of photons). A large photon density implies large absorption probability, while no photons implies zero absorption probability. It is worth emphasising that each absorption event reduces the number of photons by one.

It is found that the probability of absorption depends linearly on the radiation field. Accordingly, the rate of change of the population on the lower energy state (due to absorption) is:

$$\boxed{\left(\frac{dN_1}{dt}\right)_{\mathrm{a}} = -W_{12}N_1 = -B_{12}\rho(\omega_{21})N_1.} \tag{3.4}$$

The variables have the following meaning:

- $W_{12} = B_{12}\rho(\omega_{21})$ describes the probability of absorption per unit time. Units: $[W_{21}] = \mathrm{s}^{-1}$.
- B_{12} is known as the **Einstein B-coefficient**. Like the Einstein A-coefficient, it depends on the characteristics of the transition, but is independent of the radiation field. Note that the radiation-dependence of absorption rather enters explicitly through the $\rho(\omega_{21})$ term. Units: $[B_{21}] = \mathrm{m}^3\mathrm{J}^{-1}\mathrm{s}^{-2}$.
- $\rho(\omega)$ is the energy density per unit frequency of the radiation field at frequency ω . Units: $[\rho(\omega_{21})] = \mathrm{Jm}^{-3}\mathrm{Hz}^{-1}$. Note that $\rho(\omega)$ is directly proportional to the density of photons per unit frequency $\phi(\omega)$. In particular:

$$\rho(\omega) = \phi(\omega)\hbar\omega. \tag{3.5}$$

Accordingly, the total density of photons (in units of m^{-3}) in a frequency interval $\omega ... \omega + \delta\omega$ is:

$$\phi = \frac{\rho(\omega)}{\hbar\omega}\delta\omega \tag{3.6}$$

3.3 Stimulated emission

In the presence of a photon with energy $\hbar\omega_{21} = E_2 - E_1$, an atom in the upper level can be *stimulated* to decay to the lower level state by the emission of a second photon of energy $\hbar\omega_{21}$. The stimulated photon is emitted in the same “radiation mode” as the incident photon, and hence has the exact same frequency, direction, phase, and polarisation of the incident photon.

Just like absorption, the probability of stimulated emission depends on the number of photons. But in contrast to absorption, each stimulated emission event increases the number of photons by one. Because of stimulated emission, the population on the upper state changes with a rate of

$$\boxed{\left(\frac{dN_2}{dt}\right)_{\mathrm{st}} = -W_{21}N_2 = -B_{21}\rho(\omega_{21})N_2.} \tag{3.7}$$

The coefficients have meanings (and units) similar to those listed above: $W_{21} = B_{21}\rho(\omega_{21})$ is the probability of stimulated emission per unit time, while B_{21} is a second Einstein B-coefficient describing the transition. Again, B_{21} depends on the properties of the transition, but it does not depend on the radiation field.

3.4 Summary of interaction processes

In summary, the three processes through which light can mediate transitions between atomic energy levels are:

1. **Spontaneous emission:** An atom in the upper state spontaneously decays to the lower state, while simultaneously emitting a photon with energy $\hbar\omega_{21} = E_2 - E_1$.

2. **Absorption:** An atom in the lower state is excited to the upper state by the absorption (annihilation) of a photon.

3. **Stimulated emission:** An incident photon with energy $\hbar\omega_{21}$ stimulates an atom in the upper state to decay to the lower state by the emission of a second photon of energy $\hbar\omega_{21}$.

The occurrence of any one of the processes changes the populations on the different energy levels. The rates per unit volume at which the populations change are:

Spontaneous emission: $\left(\frac{dN_2}{dt}\right)_{\mathrm{sp}} = -A_{21}N_2$

Absorption: $\left(\frac{dN_1}{dt}\right)_{\mathrm{a}} = -B_{12}\rho(\omega_{21})N_1$

Stimulated emission: $\left(\frac{dN_2}{dt}\right)_{\mathrm{st}} = -B_{21}\rho(\omega_{21})N_2$

Note that the rate equations above only relate to individual processes. To analyse the *total* rate of change of a given population, the rate equations must be appropriately added up, such that the ensuing total rate equation accounts for all of the different processes that change the population. Furthermore, when dealing with just two energy levels, it should be clear that $dN_2/dt = -dN_1/dt$.

3.5 Non-radiative decay

The three processes described above are the only processes through which **light** can mediate transitions between two energy levels. However, it is also possible that atoms decay to lower energy levels without emitting a photon. During such **non-radiative decay**, energy is lost to a *phonon*, which corresponds to a quantum of molecular/lattice vibration, and typically manifests itself as the generation of heat.

The rate at which atoms in the upper state decay via non-radiative decay is governed by en equation similar to Eq. (3.2)

$$\left(\frac{dN_2}{dt}\right)_{\mathrm{nr}} = -C_{21}N_2, \tag{3.8}$$

where C_{21} describes the probability of non-radiative decay per unit time. As for spontaneous emission, $\tau_{21}^{(\mathrm{nr})} = C_{21}^{-1}$ describes the non-radiative lifetime of the transition. It should be clear that the total lifetime of the transition $2 \to 1$ is a combination of radiative and non-radiative lifetimes, and can be obtained as

$$\frac{1}{\tau_{21}} = \frac{1}{\tau_{21}^{(\mathrm{rad})}} + \frac{1}{\tau_{21}^{(\mathrm{nr})}}. \tag{3.9}$$

Likewise the total rate of decay $2 \to 1$ through spontaneous emission and non-radiative decay can be written as

$$\left(\frac{dN_2}{dt}\right)_{\text{decay}} = -\Gamma_{21} N_2, \tag{3.10}$$

where we have defined the total decay constant

$$\Gamma_{21} = A_{21} + C_{21}. \tag{3.11}$$

3.6 Multiple energy levels and rate equations

Atoms always have more than two energy levels, and transitions can occur between many different pairs [see e.g. Fig. 6]. However, we are typically interested only in a single transition – the laser transition – which is responsible for the generation of the desired laser light. As we shall see below, it is often sufficient to just consider the populations on the upper and lower states of the laser transition. This requires nonetheless that the different decay channels are appropriately accounted for.

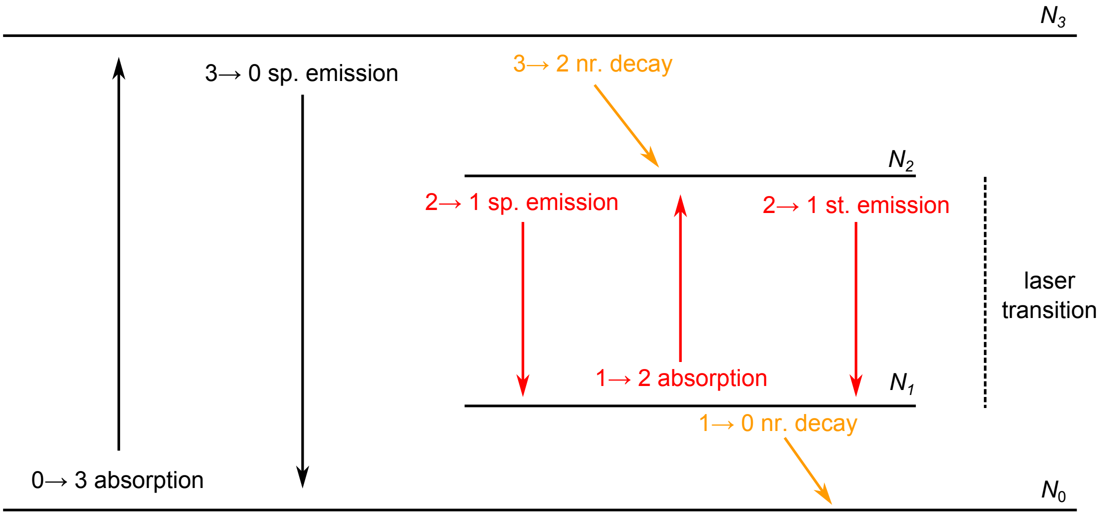

Figure 6: Example transitions between four different energy levels.

To analyse the populations involved in the laser transition, we will often start by writing down so-called **rate equations** that describe the total rate of change of the transition populations. These equations are obtained simply by combining all the different processes that contribute to changes in the population under study.

Example.
Considering the transitions depicted in Fig. 6, the rate equations for populations N_2 and N_1 are:

$$\frac{dN_2}{dt} = B_{12}\rho(\omega_{21})N_1 + C_{32}N_3 - B_{21}\rho(\omega_{21})N_2 - A_{21}N_2, \tag{3.12}$$

$$\frac{dN_1}{dt} = B_{21}\rho(\omega_{21})N_2 + A_{21}N_2 - B_{12}\rho(\omega_{21})N_1 - C_{10}N_1. \tag{3.13}$$

Note that these rate equations are not closed ($N_1 + N_2 \neq N_{\text{tot}}$), and additional equations would be needed for N_3 and N_0 to obtain closed-form solutions. Indeed, because we are no longer dealing with just two energy levels, $dN_2/dt \neq -dN_1/dt$. However, as we shall see, we can often make simplifying approximations that allow us to neglect levels other than those involved in the laser transition.

3.7 Relations between Einstein coefficients

The probabilities of spontaneous emission, absorption, and stimulated emission are governed by the three Einstein coefficients A_{21}, B_{12}, and B_{21}. These coefficients are not independent of one another. Their interrelationship can be found by using the fact that the coefficients do not depend on the radiation field, and by considering one special case in which the relative populations of the upper and lower states are known. The following derivation echoes the original 1917 re-derivation of Planck's radiation law by Einstein, and considers an ensemble of atoms immersed in a bath of blackbody radiation of temperature T and energy density $\rho(\omega)$. Before proceeding, we shall first briefly recall what is meant with blackbody radiation.

3.7.1 Black body radiation

A black body is an idealised physical body that *absorbs* all incident electromagnetic radiation, regardless of frequency or angle of incidence. In thermal equilibrium, a black body must also *emit* radiation (aptly named, black body radiation); in fact, it must emit exactly the same amount of radiation that it absorbs (else it could not maintain thermal equilibrium). While no ideal black bodies exist (the concept is an idealisation), there are many objects that behave approximately as black bodies, and in particular, emit radiation with characteristics similar to a black body. These include our own Sun, all the stars in the sky, and light bulbs. Even the cosmic microwave background resembles black body radiation.

The characteristics of black body radiation depend only on the temperature of the body[4]. The energy spectral density $\rho(\omega)$ of black body radiation is given by Planck's law[5]:

$$\boxed{\rho(\omega) = \frac{\hbar\omega^3}{\pi^2 c^3}\frac{1}{e^{\hbar\omega/(kT)} - 1},} \tag{3.14}$$

where k is the Boltzmann constant and c is the speed of light in vacuum. Equation (3.14) gives the energy spectral density (i.e., energy per unit frequency per unit volume) of radiation emitted by a black body at temperature T at a given (angular) frequency ω. The energy density contained within a narrow frequency band $d\omega$ is $\rho(\omega)d\omega$, and the total energy density emitted by the body is (in units of Joules per cubic metre):

$$\rho_{\text{tot}} = \int_0^\infty \rho(\omega)d\omega. \tag{3.15}$$

[4] For example, cosmic microwave background is like black body radiation at a temperature of 2.7 K.

[5] This is the law that started the whole quantum movement.

3.7.2 2-level atoms in a bath of black body radiation

Consider an ensemble of 2-level atoms in thermal equilibrium, immersed in a bath of black body radiation. We can combine the *rate equations* summarised above to obtain the total rates of change of the two populations:

$$\frac{dN_2}{dt} = -\left(\frac{dN_1}{dt}\right)_{\mathrm{a}} + \left(\frac{dN_2}{dt}\right)_{\mathrm{sp}} + \left(\frac{dN_2}{dt}\right)_{\mathrm{st}}, \tag{3.16}$$

$$= B_{12}\rho(\omega_{21})N_1 - B_{21}\rho(\omega_{21})N_2 - A_{21}N_2 = -\frac{dN_1}{dt}. \tag{3.17}$$

In thermal equilibrium, the populations on the upper and lower energy levels remain constant (on average): $dN_1/dt = -dN_2/dt = 0$. Solving for the radiation field $\rho(\omega_{21})$ yields

$$\rho(\omega_{21}) = \frac{A_{21}/B_{21}}{\dfrac{B_{12}}{B_{21}}\dfrac{N_1}{N_2} - 1}. \tag{3.18}$$

It is known from statistical mechanics that, in thermal equilibrium, the populations $N_{1,2}$ obey Boltzmann statistics. Specifically, the ratio N_2/N_1 between populations in two different levels with energy separation $E_2 - E_1 = \hbar\omega_{21}$ at temperature T is given by:

$$\boxed{\frac{N_2}{N_1} = \frac{g_2}{g_1}e^{-\dfrac{\hbar\omega_{21}}{kT}},} \tag{3.19}$$

where $g_{1,2}$ are the *degeneracies* of the energy levels[6]. Substituting Eq. (3.19) in Eq. (3.18) yields

$$\rho(\omega_{21}) = \frac{A_{21}/B_{21}}{\dfrac{B_{12}}{B_{21}}\dfrac{g_1}{g_2}e^{\hbar\omega_{21}/(kT)} - 1}. \tag{3.20}$$

Because the radiation field corresponds to black body radiation, the energy spectral density must be equal to that given by Planck's law, i.e., Eq. (3.14). Equating the two expressions gives:

$$\frac{\hbar\omega_{21}^3}{\pi^2c^3}\frac{1}{e^{\hbar\omega_{21}/(kT)} - 1} = \frac{A_{21}/B_{21}}{\frac{B_{12}}{B_{21}}\frac{g_1}{g_2}e^{\hbar\omega_{21}/(kT)} - 1}. \tag{3.21}$$

This equation must hold true for all temperatures T. Accordingly, the factors multiplying the exponential terms must be equal:

$$\frac{B_{12}}{B_{21}}\frac{g_1}{g_2} = 1. \tag{3.22}$$

[6] The degeneracy of an energy level is the number of distinguishable levels with the same energy. For example, in a hydrogen atom states with different angular momentum have the exact same energy, governed by the principal quantum number alone. Accordingly, the states are degenerate, as they can be distinguished by their orbital angular momentum.

From now on, we will assume for the sake of simplicity that the degeneracies are the same ($g_1 = g_2$)[7]. The relationship between the two Einstein B-coefficients then becomes particularly simple:

$$\boxed{B_{21} = B_{12}.} \tag{3.23}$$

Recalling that the probabilities of absorption and stimulated emission per unit time are $p_\mathrm{a} = B_{12}\rho(\omega_{21})$ and $p_\mathrm{st} = B_{21}\rho(\omega_{21})$, respectively, Eq. (3.23) reveals that **the probabilities of absorption and stimulated emission are exactly identical.**

Using $B_{21} = B_{12}$ and $g_1 = g_2$ in Eq. (3.21), we can also derive a relationship between the Einstein A- and B-coefficients:

$$\boxed{\frac{A_{21}}{B_{21}} = \frac{\hbar\omega_{21}^3}{\pi^2 c^3}.} \tag{3.24}$$

3.7.3 Generality of relations

The above relations were derived assuming black body radiation and a 2-level atom. However, they are in fact more general, and apply always (bearing in mind the approximation on degeneracies). Firstly, the coefficients are independent of the radiation field by construction, and therefore limiting the analysis to black body radiation does not bear any ill consequences. Secondly, whilst atoms certainly have many energy levels in general, and transitions can occur between any one of them, in thermal equilibrium the transitions between any pair must be in dynamic equilibrium[8]. This can be understood by considering the equilibrium of the radiation field. Specifically, the energy density $\rho(\omega)$ must be constant (on average) at all frequencies, and therefore transitions contributing to changes in photon number at a given frequency must be in dynamic equilibrium. Only transitions between a *pair* of energy levels separated by $\hbar\omega$ contribute to changes at ω, and therefore the populations of the pair must be in dynamic equilibrium. This justifies the analysis based on 2-level atoms. Finally, neglection of non-radiative decay channels is similarly justified by the fact that these transitions do not change the radiation field.

3.8 Necessary condition of optical amplification: population inversion

As their names imply, lasers operate via *light amplification by stimulated emission of radiation*. The medium in which the amplification takes place is known as the *active* or gain medium. Lasing can occur only if light in the active medium experiences net amplification, i.e., a net increase in the number of photons (in the lasing mode[9]). If the number of photons experiences net decrease there will be no lasing as all the photons will eventually be lost. **Key to laser operation is that amplification via stimulated emission exceeds loss.**

We have seen that each stimulated emission event adds one new photon to the lasing mode. On the other hand, each absorption event removes one photon from the lasing mode. Using the rate equations summarised in Section 3, the rate of change of the photon spectral density $\phi(\omega)$ can be written as

$$\frac{d\phi(\omega)}{dt} = [B_{21}\rho(\omega)N_2 - B_{12}\rho(\omega)N_1]\delta(\omega - \omega_{21}). \tag{3.25}$$

[7] This is not of course true in general, but makes conceptual understanding of the principal physics simpler.

[8] This principle is referred to as the principle of detailed balance by Hooker and Webb.

[9] We refer to lasing mode as the bunch of photons with the same phase, polarisation, and direction.

Here $\delta(\omega - \omega_{21})$ is the Dirac delta function[10], which simply enforces the fact that only photons with frequency ω_{21} are generated. Note that the total density of photons (with units of m^{-3}) is simply $\phi_{\text{tot}} = \int_0^\infty \phi(\omega) d\omega$. Because the dirac Delta-function is defined such that $\int_0^\infty \delta(\omega - \omega_{21}) d\omega = 1$, the total photon number evolves as

$$\frac{d\phi_{\text{tot}}}{dt} = B_{21}\rho(\omega_{21})N_2 - B_{12}\rho(\omega_{21})N_1. \tag{3.26}$$

Recalling that $B_{21} = B_{12}$, and thus defining the *probability of a stimulated transition* $W_{\text{st}} = B_{21}\rho(\omega_{21})$, we can simplify the equation above:

$$\frac{d\phi_{\text{tot}}}{dt} = W_{\text{st}}[N_2 - N_1]. \tag{3.27}$$

Given that $W_{\text{st}} > 0$, net amplification occurs precisely when

$$\boxed{N_2 > N_1.} \tag{3.28}$$

A necessary condition for light amplification by stimulated emission of radiation is that the population on the upper energy is larger than the population on the lower energy level. This situation is known as **population inversion**.

3.8.1 The very unnatural nature of population inversion

Population inversion is a necessary condition for optical amplification. It corresponds to a very unusual situation, and one which does not spontaneously occur in nature. To highlight this, let us again consider the ratio of populations at two energy levels E_2 and $E_1 < E_2$. As in Eq. (3.19), Boltzmann statistics tells us

$$\frac{N_2}{N_1} = e^{-\frac{\hbar\omega_{21}}{kT}}. \tag{3.29}$$

We see immediately that, for all temperatures $T > 0$, $N_2 < N_1$, and so population inversion cannot occur in thermal equilibrium. The example below highlights the severity of the problem.

Example. Consider a transition corresponding to red laser light, with a wavelength of $\lambda = 2\pi c/\omega_{21} = 633\ \mathrm{nm}$, where c is the speed of light. At room temperature ($T = 273$ K), the ratios of the populations on the upper and lower energy levels of the laser transition in thermal equilibrium is:

$$\frac{N_2}{N_1} = e^{-\frac{\hbar\omega_{21}}{kT}} \approx 6.3 \times 10^{-37}. \tag{3.30}$$

That is 37 orders of magnitude away from population inversion ($N_2 > N_1$).

[10] Note that the unit of the Dirac delta function is the inverse of the unit of its argument. Thus, $[\delta(\omega - \omega_{21})] = \mathrm{Hz}^{-1}$.

3.8.2 Pumping

To reach population inversion, we must move away from thermal equilibrium. Specifically, we must externally inject energy into the system, so as to force atoms to occupy the upper level. The process of doing so is known as **pumping**.

Pumping is typically achieved either electronically or optically. Electronic pumping is mostly used for gas lasers and semiconductor lasers, and is achieved by passing an electric current through the medium. For example, in semiconductor lasers this results in direct addition of electrons in the conducting band, from which they can transition to the valence band via stimulated emission. In gas lasers, the current causes electrons to collide, with gas atoms, which leads to their excitation.

In optical pumping, atoms are excited directly through absorption: the frequency of **a pump light source** is tuned to match with a transition, and absorption lifts atoms from the lower state to the upper state [see Fig. 7].

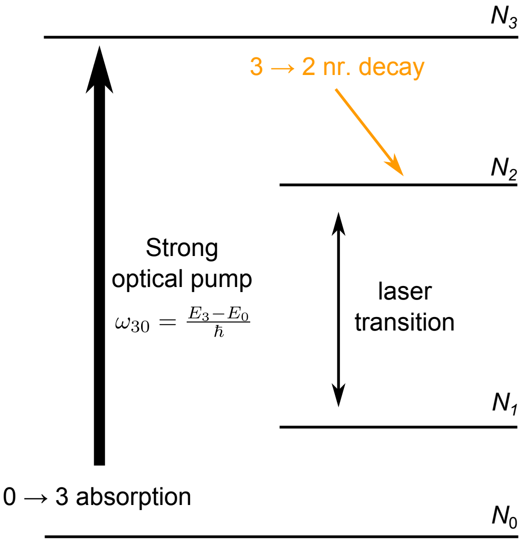

Figure 7: Schematic illustration of optical pumping. The frequency of a strong pump light source (e.g. another laser) is tuned to match with the transition ω_{30}. Atoms from level 0 are then lifted to the state 3 via absorption. In this particular example, the atoms then decay to the upper laser level non-radiatively.

3.9 Propagation of light through an active medium

We have seen that population inversion is required for light amplification. But how do we quantify this amplification? In Eq. (3.27), we have written the rate of change of the photon density. However, this is not a particularly useful quantity for a laser beam that is travelling at the speed of light: the absolute number of photons describes the whole history of the beam, depending on e.g. how long the laser has been on.

Power – particularly the rate at which light energy is transferred through a surface – is a more relevant quantity for laser beams. But power of course depends on the cross-sectional size of the beam (spot size): a beam with a large spot transfers more energy, and hence has larger power, though its apparent "brightness" is the same as that of a beam with a smaller spot (hence, smaller power). An even better quantity is the *optical*

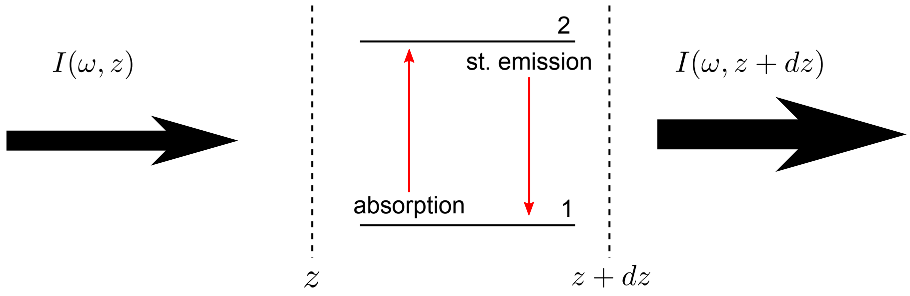

Figure 8: Schematic illustration of a laser beam propagation through a thin segment of active medium.

intensity[11], I, which is the rate at which energy is transferred *per unit area*. Intensity can be loosely[12] defined as

$$I(\omega) = \frac{P(\omega)}{A} = \frac{E(\omega)}{tA}, \tag{3.31}$$

where $E(\omega) = \rho(\omega)V$ is the spectral energy transferred through a surface with area A in time t, and $P(\omega) = E(\omega)/t$ is the corresponding spectral power.

Let us now consider how a beam with cross-sectional area A evolves as it propagates through a thin segment of active medium with length dz [see Fig. 8]. The rate of change of the photon spectral density, $\phi(\omega)$, with units of $\mathrm{photons}/(\mathrm{m}^{-3}\mathrm{Hz})$, is given by Eq. (3.25). Because the beam is travelling at the speed of light, c, the time it takes for it to travel through the active medium is $dt = dz/c$. Thus, we can rewrite Eq. (3.25) as:

$$\frac{d\phi(\omega)}{dz} = \frac{W_{\text{st}}}{c}\Delta N \delta(\omega - \omega_{21}), \tag{3.32}$$

where we introduced the population difference $\Delta N = N_2 - N_1$. We next note that $\phi(\omega) = \rho(\omega)/(\hbar\omega) = I(\omega)/(\hbar\omega c)$, where we used the relationship[13] $\rho(\omega) = I(\omega)/c$. This yields

$$\frac{dI(\omega)}{dz} = \hbar\omega W_{\text{st}}\Delta N \delta(\omega - \omega_{21}), \tag{3.33}$$

Finally, we recognise that the probability of stimulated transitions, W_{st}, depends on the intensity itself. Specifically, $W_{\text{st}} = B_{21}\rho(\omega) = B_{21}I(\omega)/c$. Substituting this to the equation above yields our final result:

$$\boxed{\frac{dI(\omega)}{dz} = \sigma(\omega - \omega_{21})\Delta N I(\omega),} \tag{3.34}$$

where we introduced the interaction cross-section $\sigma(\omega - \omega_{21})$, with units of m^2 as:

$$\boxed{\sigma(\omega - \omega_{21}) = B_{21}\frac{\hbar\omega_{21}}{c}\delta(\omega - \omega_{21}) = \frac{\pi^2 c^2}{\omega_{21}^2 \tau_{\text{sp}}}\delta(\omega - \omega_{21}).} \tag{3.35}$$

[11] Also known as irradiance.

[12] Not being pedantic about possible time/frequency dependencies.

[13] This relationship can be easily derived by considering a beam propagation for a length l in time t with a speed $c = l/t$. Specifically, $\rho(\omega) = E(\omega)/V = I(\omega)At/(Al) = I(\omega)/c$.

The latter form is obtained by using Eq. (3.24) and the fact that A_{21} is the inverse of the spontaneous emission lifetime, i.e., $A_{21} = \tau_{\text{sp}}^{-1}$.

Assuming the population inversion ΔN to remain constant, Eq. (3.34) can be solved trivially[14]:

$$\boxed{I(z,\omega) = I(0,\omega)e^{\sigma\Delta N z} = I(0,\omega)e^{gz},} \tag{3.36}$$

where we omitted the frequency dependence of $\sigma(\omega - \omega_{21})$ for brevity, and defined the *gain coefficient* $g = \sigma\Delta N$. We see that the intensity evolves exponentially as it propagates through the medium. Depending on the population inversion, three different behaviours can be identified:

1. $\Delta N > 0$**:** Population inversion leads to exponential amplification as stimulated emission prevails over absorption.
2. $\Delta N < 0$**:** Absorption prevails over stimulated emission, giving rise to exponential loss of intensity.
3. $\Delta N = 0$**:** Rates of absorption and stimulated emission are exactly equal, and so the intensity remains constant. The material is totally transparent.

3.10 Population inversion in a 2-level system

And so we have seen that a population inversion is needed for optical amplification. Is it possible to achieve population inversion in a 2-level system using optical pumping? To answer this question, we write and solve the **rate equations** for the level populations $N_{1,2}$:

$$\frac{dN_2}{dt} = W_{\text{st}}(N_1 - N_2) - A_{21}N_2, \tag{3.37}$$

$$\frac{dN_1}{dt} = -W_{\text{st}}(N_1 - N_2) + A_{21}N_2. \tag{3.38}$$

Note that, in a 2-level system, the lower level corresponds to the ground level, and hence cannot undergo spontaneous emission. In steady-state, the populations are constant, such that $dN_1/dt = -dN_2/dt = 0$. It is easy to solve for N_2:

$$N_2 = \frac{W_{\text{st}}N_1}{W_{\text{st}} + A_{21}}. \tag{3.39}$$

We can thus write

$$\frac{N_1}{N_2} = 1 + \frac{A_{21}}{W_{\text{st}}}. \tag{3.40}$$

Clearly, $N_1 > N_2$ always, implying that **population inversion cannot be achieved in a 2-level system (via optical pumping)**. The best you get is half of the total population in the upper level (and the other half on the ground level). This occurs as $W_{\text{st}} \to \infty$, i.e., as the optical pumping strength approaches infinity.

[14] As we shall soon see, this is not the case in general.

3.11 Population inversion in multi-level systems

As shown above, population inversion cannot be achieved (with optical pumping) in systems with just two energy levels. Intermediate states are needed. Furthermore, at least one of these intermediate states must be *metastable*: an excited state with a very long lifetime (e.g. microseconds). This metastable state will act as the upper level of the laser transition. Its long lifetime ensures that a large number of atoms can be stored on the level, thus favouring population inversion. Depending on the lower level of the lasing transition, lasers can generally be divided into three categories: 3-level systems, 4-level systems, and quasi-3-level systems (see Fig. 9).

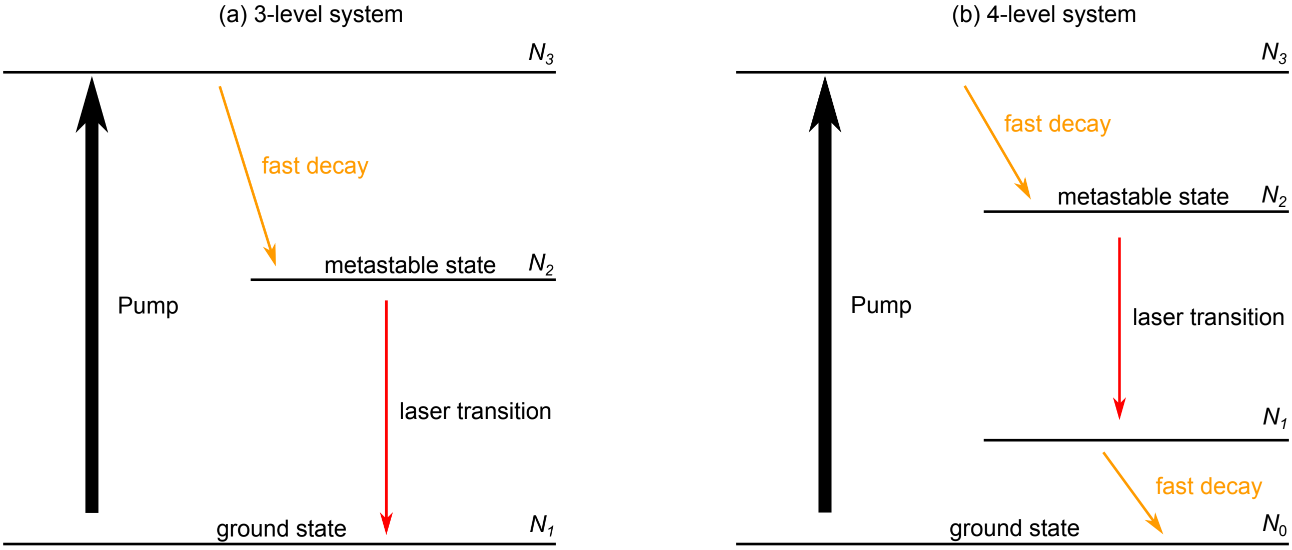

Figure 9: Schematic illustrations of energy levels for (a) 3-level system and (b) 4-level system. The key difference is that, in a 3-level system the laser transition ends at the ground state, while in a 4-level state the transition ends at an intermediate level above the ground state.

3.11.1 3-level systems

In 3-level systems [c.f. Fig. 9(a)], the ground state acts as the lower level of the laser transition. The upper level of the laser transition (2) is a metastable state whose energy lies in between the ground state (1) and a second excited state (3). Population inversion can be achieved by pumping atoms from the ground state to the higher lying level, followed by a rapid (radiative or non-radiative) decay into the upper state of the laser transition. Thanks to the fast decay, the third state remains empty all the time. This assists population inversion, since (i) spontaneous emission 3 → 1 does not populate the ground state and (ii) pump absorption 1 → 3 prevails over corresponding stimulated emission.

Taking $N_3 \approx 0$, it is clear that $N_{\text{tot}} = N_1 + N_2$. Thus, inversion requires $N_2 > N_{\text{tot}}/2$; more than half of the atoms are needed at the excited upper laser level. Compounded by the fact that the lower laser level is prone to be highly-populated due to it being the ground state[15], very strong pumping is needed to reach population

[15] Remember Boltzmann...

inversion in 3-level systems. This makes 3-level systems inefficient.

Because of their inefficiency, 3-level lasers are not in wide use. There is one important historical exception: the first laser ever demonstrated (by Maiman in 1960) used a 3-level laser medium, namely ruby.

3.11.2 4-level systems

4-level lasers overcome the inefficiency of 3-level lasers. In these systems, the lower level of the laser transition is an excited state well-above the ground state ($E_1 \gg kT$). This ensures that, in thermal equilibrium, $N_1 \approx 0$, thus making it easier to reach inversion. Ideally, the lower laser level also decays rapidly to the ground state, so that $N_1 \approx 0$ even during laser operation. In this way, inversion is achieved by getting just a few atoms to the upper lasing state.

Atoms are pumped from the ground state (0) to an excited state (3) that lies above the upper laser level (2). As in 3-level systems, the pumped atoms quickly decay to the metastable upper laser level. This ensures that (i) the upper laser level has large population and (ii) that atoms do not fall back to the ground state via spontaneous emission and/or stimulated emission.

A good example of a 4-level active medium is the popular Nd:YAG (neodymium-doped yttrium aluminium garnet), which typically emits light at 1064 nm. Such lasers have numerous applications in medicine, manufacturing, and basic science.

In some "nominally" 4-level systems, the lower laser level is so close to the ground state that it possesses an appreciable population even in thermal equilibrium. Such lasers are sometimes called quasi-3-level. Examples include all ytterbium-doped active materials (e.g. Yb:glass used to realise fiber lasers at 1030 nm) as well as erbium-doped media for emission around 1500 nm, including the extremely important Er:glass used for fiber amplifiers and lasers around 1550 nm.

3.12 Quantitative analysis of 4-level rate equations

Population inversion can be reached in both systems described above. Here we will quantitatively analyse the rate equations for a 4-level system to show that this is indeed the case. We will also find that the net gain that can be achieved exhibits curious behaviour for high intensities.

3.12.1 Rate equations

For simplicity, we make the following assumptions:

- The population of the ground level $N_0 \gg N_1, N_2, N_3$. Therefore N_0 will not be depleted, and can be assumed constant.
- Atoms are pumped from the ground level to the third level at a constant rate of R_3.
- The third level decays immediately, such that $N_3 \approx 0$. The atoms decay to the upper laser level (2) at a rate of R_2, but some of them may end up on the lower laser level (1) at a rate of $R_1 = R_3 - R_2$.
- We assume that the rate of spontaneous (radiative and non-radiative) decay from the upper lasing level to the ground state is very slow[16]. Accordingly, the corresponding Einstein coefficient A_{20} is very small,

[16] As expected for a metastable state.

and the transition $2 \to 0$ can be neglected[17].

With these assumptions, the rate equations for N_1 and N_2 read[18]:

$$\frac{dN_2}{dt} = R_2 + W_{\text{st}}(N_1 - N_2) - A_{21}N_2, , \tag{3.41}$$

$$\frac{dN_1}{dt} = R_1 - W_{\text{st}}(N_1 - N_2) + A_{21}N_2 - A_{10}N_1. \tag{3.42}$$

We are interested in steady-state behaviour, and set $dN_1/dt = dN_2/dt = 0$. Adding the two equations together under these conditions yields:

$$R_1 + R_2 = A_{10}N_1 \tag{3.43}$$

Solving for N_1 yields

$$N_1 = \frac{R_1 + R_2}{A_{10}}. \tag{3.44}$$

On the other hand, from Eq. (3.41), we can solve for N_2:

$$N_2 = \frac{R_2 + W_{\text{st}}N_1}{A_{21} + W_{\text{st}}} \tag{3.45}$$

Using the expression for N_1, we can now calculate the population inversion $\Delta N = N_2 - N_1$. After some algebra, one finds:

$$\Delta N = \frac{R}{A_{21} + W_{\text{st}}}, \tag{3.46}$$

where

$$R = R_2\left[1 - \frac{A_{21}}{A_{10}}\left(1 + \frac{R_1}{R_2}\right)\right] \tag{3.47}$$

corresponds to a *reduced effective pump rate*, accounting for the fact that some of the atoms pumped to level 3 decay to level 1, as well as the fact that the decay out of level 1 is not instantaneous. In the limit $R_1 \to 0$ or $A_{10} \to \infty$, $R \to R_2$.

Rearranging Eq. (3.46) yields important insights:

$$\Delta N = \frac{R/A_{21}}{1 + W_{\text{st}}/A_{21}}. \tag{3.48}$$

Recalling that $W_{\text{st}} = B_{21}\rho(\omega_{21}) = B_{21}I(\omega_{21})/c$, we can write the population inversion as

$$\boxed{\Delta N = \frac{\Delta N_0}{1 + I(\omega_{21})/I_{\text{sat}}},} \tag{3.49}$$

[17] Note that the spontaneous transition $2 \to 1$ should not be neglected, since we do not want A_{21} to be too small. This is because the rate of stimulated emission $W_{\text{st}} \propto B_{21} \propto A_{21}$.

[18] Note that we neglect here non-radiative decays by using the Einstein coefficients A_{ij} instead of the full transition decay rates Γ_{ij} defined in section 3.5. This does not change any of the conclusions.

where we defined the *small-signal* inversion $\Delta N_0 = R/A_{21}$ and the *saturation intensity*

$$\boxed{I_{\text{sat}} = \frac{cA_{21}}{B_{21}} = \frac{\hbar\omega_{21}^3}{\pi^2 c^2}.} \tag{3.50}$$

Several important conclusions can be drawn from Eq. (3.49):

1. The "small-signal" inversion ΔN_0 can be positive, implying that population inversion can indeed be reached in a 4-level configuration.
2. For small laser intensities $I(\omega_{21}) \approx 0$, the population inversion corresponds to the small-signal inversion $\Delta N = \Delta N_0$.
3. As the intensity $I(\omega_{21})$ grows, the population inversion decreases. In the limit $I(\omega_{21}) \gg I_{\text{sat}}$, population inversion $\Delta N \approx 0$; the medium becomes transparent and there is no more amplification[a]. This behaviour is known as **gain saturation**. Ultimately, gain saturation is the limiting factor for laser intensities.

[a] Recall that $I(z,\omega) \propto \exp(\sigma \Delta N z)$.

3.13 Line broadening

So far, we have assumed that the electronic energy levels are infinitely sharp. This is not true in practice: all transitions have a finite spectral width. As a consequence, atoms do not only interact with photons whose frequency matches *exactly* with the transition frequency ω_{21}. Rather, they interact with photons whose frequency falls within a certain (finite) range of frequencies [see Fig. 10]. Specifically:

- An atom can absorb a photon even if the photon's frequency is slightly detuned from the (mean) transition frequency ω_{21}.
- An atom decaying to a lower energy level can emit a photon whose frequency is slightly detuned from the (mean) transition frequency ω_{21}.

The "strength" of the interaction[19] depends on the detuning between the transition frequency and the photon frequency. It is strongest when the the photon frequency matches exactly with the transition frequency, and decreases as the detuning increases. For very large detunings, there is no interaction.

Line broadening can be quantitatively taken into account by multiplying the Einstein-coefficients with a suitable "lineshape function" $g(\omega - \omega_{21})$, which has a maximum at $\omega = \omega_{21}$ and decays to zero as $\omega - \omega_0 \to \pm\infty$. Using spontaneous emission as an example, the physical meaning is as follows:

[19] For example, probability of absorption.

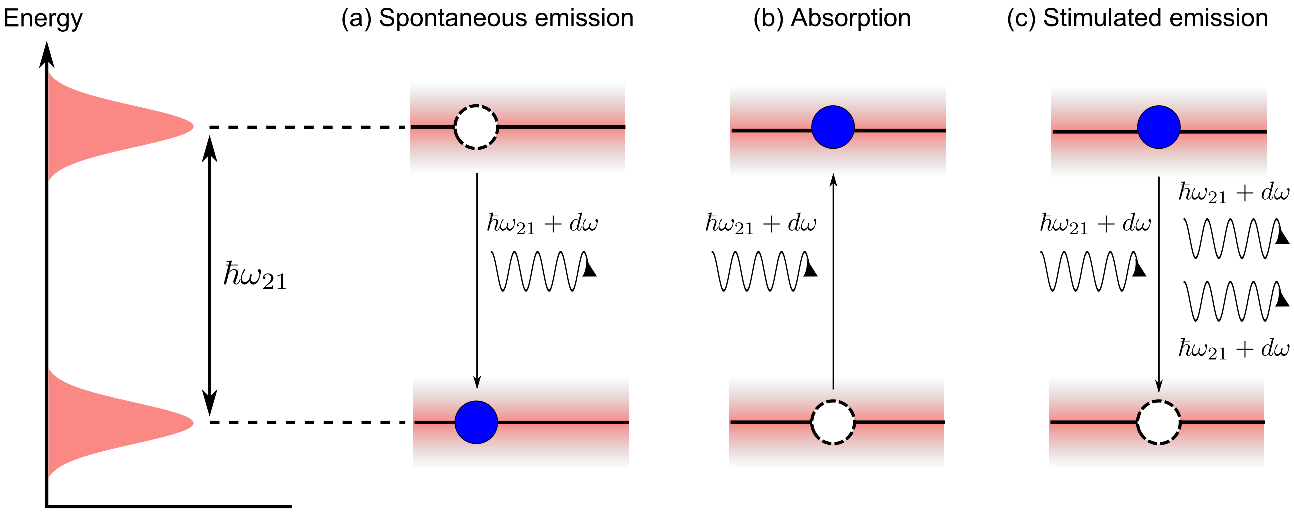
Energy
(a) Spontaneous emission
(b) Absorption
(c) Stimulated emission
$\hbar\omega_{21}$
$\hbar\omega_{21} + d\omega$
$\hbar\omega_{21} + d\omega$
$\hbar\omega_{21} + d\omega$
$\hbar\omega_{21} + d\omega$
$\hbar\omega_{21} + d\omega$

Figure 10: Schematic illustrations of interactions between photons and atoms with broadened energy levels.

The probability that an excited atom will undergo spontaneous decay $2 \to 1$ (with energy separation $E_2 - E_1 = \hbar\omega_{21}$) during a time interval dt, while emitting a photon in the frequency interval $\omega + d\omega$ is

$$p(\omega) = [A_{21} g(\omega - \omega_{21}) d\omega] dt. \tag{3.51}$$

The lineshape function must be normalised such that

$$\int_0^\infty g(\omega - \omega_{21}) d\omega = 1. \tag{3.52}$$

This is because the total probability must remain constant:

$$p = \int_0^\infty p(\omega) d\omega = A_{21} dt. \tag{3.53}$$

To understand this, note that the atom does not care what frequency the photon has. It will decay with probability $p = A_{21}dt$ during time interval dt, no matter what. Thus the infinitesimal probabilities associated with the emission of photons with different frequencies must all add up so that the overall probability of decay is $p = A_{21}dt$.

Mechanisms that cause deviations from infinitely sharp transitions are known as “line broadening mechanisms”.[20] They can be divided into two categories with distinct lineprofiles: homogeneous and inhomogeneous.

[20] For the plain reason that they broaden the absorption and emission lines observed for the atom.

3.13.1 Homogeneous broadening: Lorentzian lineshapes

Broadening is said to be homogeneous if each atom is affected in the same way (on average), and if each atom is associated with the same resonance frequency ω_{21}. In this case, the lineshape function corresponds to a Lorentzian[21]:

$$\boxed{g_{\rm H}(\omega-\omega_{21}) = \frac{1}{\pi}\frac{\gamma/2}{(\omega-\omega_{21})^2+(\gamma/2)^2},} \tag{3.54}$$

where γ corresponds to the full-width at half maximum of the profile. An example is shown in Fig. 11.

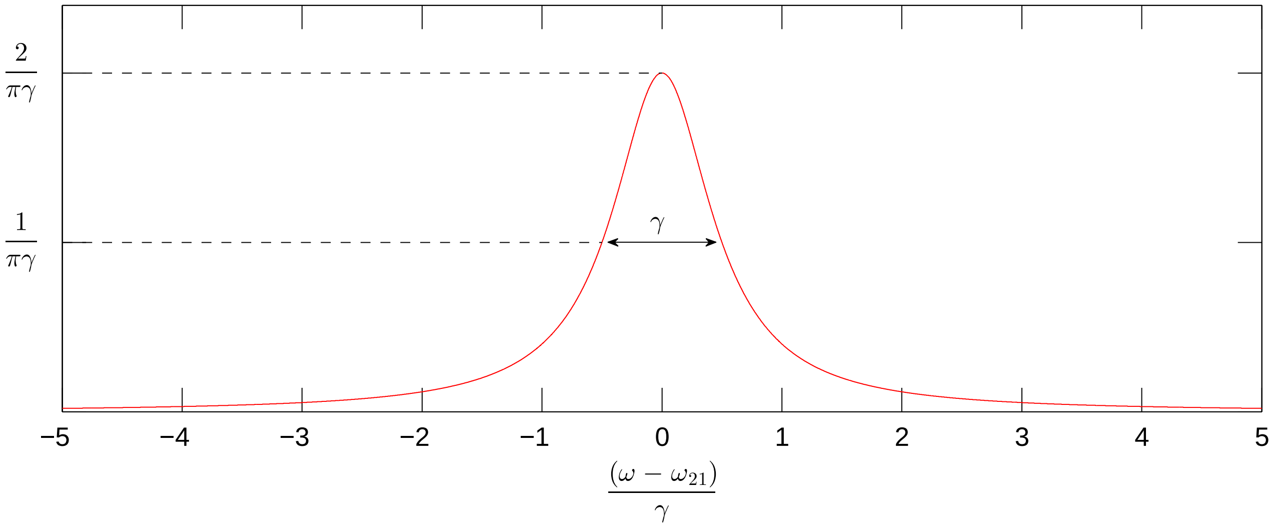

Figure 11: Lorentzian lineshape function $g_{\rm H}(\omega-\omega_{21})$ as given by Eq. (3.54).

The most important example of homogeneous broadening is *lifetime broadening*. Specifically, *all* energy levels have a finite width simply as a result of their finite (i.e. nonzero) lifetimes. Heisenberg's uncertainty principle states

$$\Delta E \Delta\tau \approx \hbar \Rightarrow \Delta E \approx \frac{\hbar}{\Delta\tau}, \tag{3.55}$$

where $\Delta\tau$ is the lifetime of the excited state. We see that the larger the lifetime, the smaller the uncertainty in the state's energy (hence, width of the level). Considering now a transition between two states, we see that the transition frequency will have an uncertainty

$$\Delta\omega = \frac{\Delta E_2+\Delta E_1}{\hbar} = \frac{1}{\Delta\tau_1}+\frac{1}{\Delta\tau_2}. \tag{3.56}$$

And so photons with a range of frequencies $\omega_{21} \pm \Delta\omega$ can be absorbed or emitted. A complete calculation shows that the transition lineshape is given by Eq. (3.54) with $\gamma = \Delta\omega$. There are several mechanisms that can contribute to lifetime broadening:

[21] The physics will be explained in Section 5.

Processes responsible for lifetime broadening

1. **Spontaneous emission** naturally limits the lifetime of a state. Considering all possible spontaneous emission pathways, the *radiative* lifetime of a state j can be obtained from

$$\frac{1}{\tau_{\mathrm{j,rad}}} = \sum_{i \neq j} A_{ji}. \tag{3.57}$$

For any transition $2 \rightarrow 1$, spontaneous emission sets the absolute minimum linewidth:

$$\gamma_{\mathrm{N}} = \frac{1}{\tau_{2,\mathrm{rad}}} + \frac{1}{\tau_{1,\mathrm{rad}}}. \tag{3.58}$$

Broadening due to spontaneous emission is typically known as *natural broadening*, since it is always present and cannot be avoided.

2. **Non-radiative decay.** As we have seen, atoms can also decay without the emission of a photon. Just like spontaneous emission, this limits the lifetime of the state.

3. **Collision broadening.** When matter is in gaseous or liquid form, atoms (or ions or molecules) tend to collide with each other. This can lead to the decay of excited states, thus affecting their lifetimes.

3.13.2 Inhomogeneous broadening: Gaussian lineshapes

Sometimes different atoms in the ensemble are associated with slightly different resonance frequencies ω_{21}. Because the total absorption/emission spectrum is the average over millions of atoms in the ensemble, presence of a range of resonance frequencies gives rise to an apparent linewidth. When broadening arises in this way, it is said to be *inhomogeneous*.

Doppler broadening in gas lasers is a classic example of inhomogeneous broadening. Consider the spontaneous emission from atoms in a gaseous sample that is being detected by a spectrometer [see Fig. 12]. If the atoms are all stationary, they will all emit radiation at the same resonance frequency ω_{21}, leading to a narrow naturally broadened lineshape. However, for nonzero temperatures the atoms will exhibit thermal motion: each atom is moving in a random direction with a speed that is linked to the temperature. If a given atom has total velocity $\vec{\mathbf{v}}$, such that its component in the direction of the spectrometer is v_z, then the frequency of radiation emitted by the atom will be shifted from ω_{21} due to the Doppler effect. Specifically, the spectrometer will detect the Doppler-shifted frequency:

$$\omega = \left(1 + \frac{v_z}{c}\right)\omega_{21}. \tag{3.59}$$

Since the ensemble consists of millions of atoms with random motions (hence, different v_z), a continuous distribution of frequencies will be emitted (and detected). This is known as **Doppler broadening**.

We can obtain an expression for the lineshape function $g_{\mathrm{D}}(\omega - \omega_{21})$ associated with Doppler broadening by noting that, in thermal equilibrium, the velocities of atoms in gases obey the Maxwell-Bolzmann distribution. Specifically, the probability that an atom picked at random has a velocity component between v_z and $v_z + dv_z$

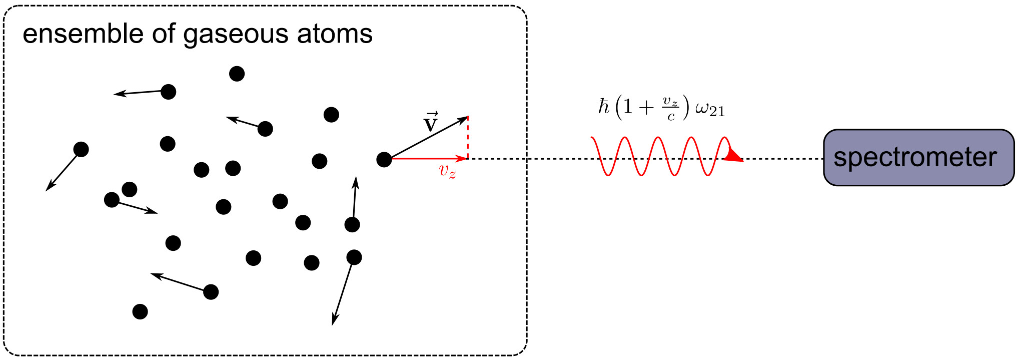
ensemble of gaseous atoms
$\vec{v}$
v_z
$\hbar\left(1+\frac{v_z}{c}\right)\omega_{21}$
spectrometer

Figure 12: Schematic illustration of Doppler broadening. Atoms in a gaseous sample exhibit random thermal motion. Because of the Doppler effect, the frequency of a photon emitted by an atom depends on the emission direction relative to the velocity of the atom.

is

$$f(v_z)dv_z = \sqrt{\frac{M}{2\pi kT}}e^{-\frac{Mv_z^2}{2kT}}, \tag{3.60}$$

where M is the mass of the atom/molecule. Because the resonance frequency of the atom is determined by v_z [see Eq. (3.59)], this probability also equals the probability that an atom picked at random has resonance frequency between ω and $\omega + d\omega$. On the other hand, the latter probability corresponds by definition to the lineshape function multiplied by $d\omega$, and so we get:

$$g_{\mathrm{D}}(\omega-\omega_0)d\omega = f(v_z)dv_z. \tag{3.61}$$

From Eq. (3.59), we obtain $v_z = (\omega - \omega_{21})c/\omega_{21}$ such that $dv_z/d\omega = c/\omega_0$, allowing us to rewrite the equation above as:

$$g_{\mathrm{D}}(\omega-\omega_0)d\omega = \sqrt{\frac{Mc^2}{2\pi kT\omega_{21}^2}}e^{-\frac{Mc^2}{2kT\omega_{21}^2}(\omega-\omega_{21})^2}d\omega. \tag{3.62}$$

From here it easy to find a form for the lineshape function. It is typically written in the following form:

$$g_{\mathrm{D}}(\omega-\omega_0) = \frac{2}{\Delta\omega_D}\sqrt{\frac{\ln 2}{\pi}}e^{-4\ln 2\left(\frac{\omega-\omega_{21}}{\Delta\omega_{\mathrm{D}}}\right)^2}, \tag{3.63}$$

$$\Delta\omega_D = 2\sqrt{2\ln 2}\frac{\omega_{21}}{c}\sqrt{\frac{kT}{M}}. \tag{3.64}$$

As can be seen, Doppler broadening gives rise to a Gaussian lineshape, with a full-width at half-maximum of $\Delta\omega_D$. Example profile is shown in Fig. 13.

Another example of inhomogeneous broadening occurs for ions doped in crystals or glasses. Here the ions experience a local electric field produced by the surrounding atoms of the material. This results in the shifting

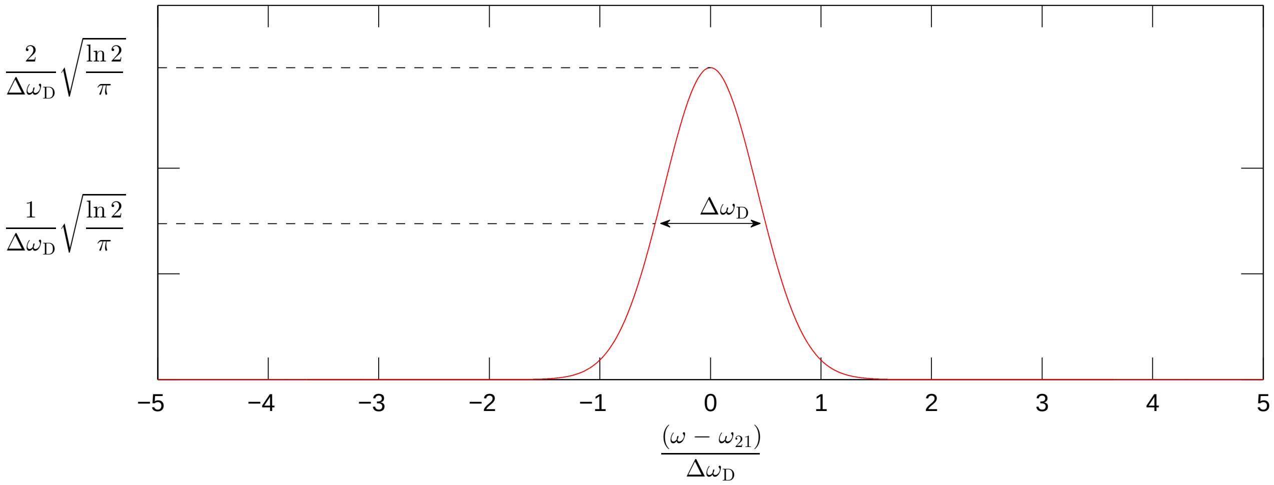

Figure 13: Gaussian lineshape function $g_D(\omega - \omega_{21})$ as given by Eq. (3.63).

and splitting of the ions' spectral lines through the so-called **Stark effect**.[22] Because of inhomogeneities in the host medium (crystal or glass), the local electric field is different from ion to ion, giving rise to a distribution of resonance frequencies. If the local field variations are fully random, one finds that the lineshape function is again a Gaussian, following the general form given by Eq. (3.63). Of course, the full-width at half-maximum is not given by Eq. (3.64), but is rather governed by the extent of variation in the resonance frequencies due to the local field variations.

3.13.3 More complex lineshapes

Homogeneous broadening arises because the atoms' energy levels are broadened due to their finite lifetimes, while inhomogeneous (e.g. Doppler) broadening arises because different atoms are associated with different resonance frequencies. But of course, these two broadening mechanisms can in general be present simultaneously; in fact, this is always the case for inhomogeneously broadened systems since natural broadening can never be avoided. The total lineshape then arises as a superposition of multiple Lorenzians whose centres are distributed according to the inhomogeneous lineshape function [see Fig. 14]. Mathematically, the composite lineshape is given by a convolution between the natural and inhomogeneous profiles:

$$g(\omega - \omega_{21}) = \int_{-\infty}^{\infty} g_D(\omega')g_H(\omega - \omega_{21} - \omega')d\omega' = g_D(\omega) * g_H(\omega), \tag{3.65}$$

where we avoided the introduction of a new variable and used $g_D(\omega)$ to indicate a generic inhomogeneous lineshape function. In the limit where the inhomogeneous broadening mechanism dominates (i.e., $\Delta\omega_D \gg \gamma$), we can replace $g_H(\omega - \omega_{21}) \approx \delta(\omega - \omega_{21})$ and obtain $g(\omega - \omega_{21}) \approx g_D(\omega - \omega_{21})$, as expected. When the widths of the homogeneous and inhomogeneous lineshape functions are of the same order of magnitude, similar approximations cannot be performed, and the overall lineshape must be obtained from the convolution above.

[22] Stark effect described the phenomenon that atoms/ions/molecules subject to an external electric field will experience shifting and splitting of their spectral lines.

In the special case where $g_{\mathrm{D}}(\omega - \omega_{21})$ is a Gaussian (as is the case for Doppler broadening), the total lineshape emerging from the convolution is known as a **Voigt profile**[23].

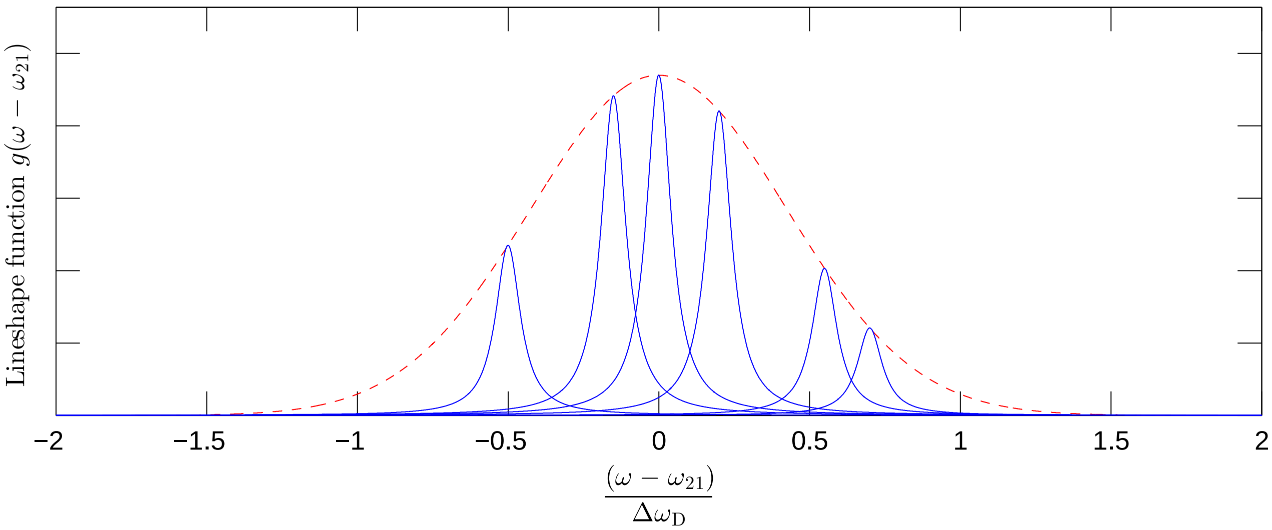

Figure 14: Schematic illustration of how inhomogeneous broadening can be understood as a superposition of multiple homogeneous lineshapes with different centre frequencies. Here, the dashed curve is the probability of finding an atom with a particular centre frequency, while the solid blue curves correspond to the underlying homogeneous lineshapes.

The broadening mechanisms described above apply for a single transition. However, in many lasers – particularly those based on transition metal ions[24] – there is a very large number of possible energy levels that are very close to each other. A situation like this can arise when each electronic energy level is associated with a large number of *vibrational* states associated with "mechanical" vibrations of the ionic lattice. Optical transitions can then occur between many different pairs of energy levels, accompanied by the emission of both photons and phonons. Because the different vibrational states are so close to each other (in energy), and because each of them exhibits some lifetime broadening, the result is a broad continuum of energy states. Lasers that exhibit broad linewidths due to vibrational energy states are commonly referred to as **vibronic lasers**. Their linewidths can be very large, exceeding hundreds of nanometres, which makes them extremely useful for the realisation of tunable lasers and ultrashort-pulsed lasers. Because of the complexity of the underlying line broadening, there are no simple analytic expressions for the lineshape functions of vibronic lasers.

[23] In other words, the convolution between a Gaussian and a Lorentzian yields a Voigt ptofile.

[24] Some important examples include Titanium-doped sapphire, alexandrite lasers, chromium forsterite lasers.

3.13.4 Line broadening in amplification and rate equations

As we had some foresight, incorporating lineshape functions in our analysis of light propagation through an active medium [Section 3.9] is very simple. Specifically, we only need to replace the Dirac delta function in Eq. (3.35) with the desired lineshape function. Could not be easier.

To see why this is the case, one notes that, for transitions that are not infinitely narrow, photons can be generated even at frequencies detuned from the transition frequency ω_{21}. Thus, the Dirac delta function in (3.25) should be replaced with something that characterises the relative likelihood that a transition results in the generation of a photon with a given frequency: this is just the lineshape function. Thus, we may rewrite (3.25) as:

$$\frac{d\phi(\omega)}{dt} = \left[B_{21}\rho(\omega)N_2 - B_{12}\rho(\omega)N_1\right] g_\mathrm{H}(\omega-\omega_{21}). \tag{3.66}$$

Strictly speaking, this expression is only correct for homogeneously broadened transitions, which is why we have assumed the lineshape function to be Lorentzian. For inhomogeneously broadened transitions, photons at different frequencies may interact with different classes of atoms associated with different centre frequencies. Thus, a complete description would require several equations similar to Eq. (3.66), with the total level populations replaced by "sub-populations" describing atoms with different centre frequencies.[25]

It is straightforward to repeat the analysis in Section 3.9 using Eq. (3.66). One finds that, for atomic transitions with nonzero linewidth, the evolution of the intensity spectral density obeys the following equations:

$$\frac{dI(\omega)}{dz} = \sigma(\omega-\omega_{21})\Delta N I(\omega), \tag{3.67}$$

$$\Delta N = N_2 - N_1, \tag{3.68}$$

$$\sigma(\omega-\omega_{21}) = B_{21}\frac{\hbar\omega}{c}g_\mathrm{H}(\omega-\omega_{21}) = \frac{\pi^2c^2}{\omega_{21}^2\tau_\mathrm{sp}}g_\mathrm{H}(\omega-\omega_{21}). \tag{3.69}$$

Thus, the only difference is that the Dirac delta function in the transition cross-section is replaced with the lineshape function. We must note that a similar expression can be derived even for an inhomogeneously broadened transition (except with $g_\mathrm{H} \to g_\mathrm{D}$), provided that the underlying homogeneous linewidth is much narrower than the inhomogeneous width (see e.g. Fig. 14).

A finite linewidth also affects rate equation analyses. Consider photons in the frequency interval $\omega ... \omega + d\omega$. These photons induce stimulated emission and absorption events that change the population densities on the upper and lower laser levels. Focussing on stimulated emission (and assuming a homogeneously broadened transition), the rate at which photons in the frequency interval $\omega ... \omega + d\omega$ change the upper level population is

$$\left(\frac{dN_{2,\omega}}{dt}\right)_\mathrm{st} = -B_{21}g_\mathrm{H}(\omega-\omega_{21})\rho(\omega)N_2 d\omega, \tag{3.70}$$

where the subscript ω in $N_{2,\omega}$ highlights the fact that we are only interested in atoms emitting photons in the frequency interval $\omega ... \omega + d\omega$. To now find the rate of change of the total population, we must integrate over all frequencies, yielding

$$\left(\frac{dN_2}{dt}\right)_\mathrm{st} = -B_{21}\left[\int_0^\infty g_\mathrm{H}(\omega-\omega_{21})\rho(\omega)d\omega\right]N_2. \tag{3.71}$$

[25] The sub-populations would then be linked to the total populations through the inhomgoeneous lineshape function.

A similar equation can be argued to hold for absorption. Thus, the lineshape function can be included in rate equation analyses by simply rewriting the stimulated transition probability as[26]:

$$W_{\mathrm{st}} = B_{21}\left[\int_0^\infty g_{\mathrm{H}}(\omega-\omega_{21})\rho(\omega)d\omega\right]. \tag{3.72}$$

Again, it should emphasized that this approach is only valid for homogeneously broadened transitions, and things get somewhat more complicated for inhomogeneous lineshapes.

If the transition is infinitely narrow, such that $g_{\mathrm{H}}(\omega-\omega_{21}) = \delta(\omega-\omega_{21})$, we recover our earlier transition probability, i.e., $W_{\mathrm{st}} = B_{21}\rho(\omega_{21})$. On the other hand, often we are interested in situations where laser light has a much narrower linewidth than the transition itself. In this case, we can write $\rho(\omega) = \rho_{\mathrm{tot}}\delta(\omega-\omega_L)$, where ω_{L} is the centre frequency of the laser, obtaining

$$W_{\mathrm{st}} = B_{21}g_{\mathrm{H}}(\omega_{\mathrm{L}}-\omega_{21})\rho_{\mathrm{tot}} = \sigma(\omega_{\mathrm{L}}-\omega_{21})\frac{I_{\mathrm{tot}}}{\hbar\omega_{\mathrm{L}}}. \tag{3.73}$$

It is then straightforward to – for example – repeat the analysis in Section 3.12 using this stimulated transition probability. Significantly, one finds that the steady-state population inversion can again be written in the form

$$\boxed{\Delta N = \frac{\Delta N_0}{1+I_{\mathrm{tot}}/I_{\mathrm{sat}}},} \tag{3.74}$$

but now with the saturation intensity incorporating the lineshape function:

$$\boxed{I_{\mathrm{sat}} = \frac{cA_{21}}{g_{\mathrm{H}}(\omega_{\mathrm{L}}-\omega_{21})B_{21}} = \frac{\hbar\omega_{\mathrm{L}}A_{21}}{\sigma(\omega_{\mathrm{L}}-\omega_{21})}.} \tag{3.75}$$

3.14 Gain coefficient and summary of equations

Often we may not have direct (experimental) access to the transition cross-section σ and/or the inversion ΔN. To simplify the notation in this case, it is customary to define a *gain coefficient* $g = \sigma\Delta N$. Note that (i) g should not be confused with the lineshape function and (ii) it exhibits a frequency-dependence due to the frequency-dependence of the cross-section (omitted to simplify notation). Furthermore, the gain coefficient saturates because the inversion saturates:

$$g = \frac{g_0}{1+I/I_{\mathrm{sat}}}. \tag{3.76}$$

Here, $g_0 = \sigma\Delta N_0$ is the small-signal gain coefficient.

Summary of equations for optical amplification of total intensity of narrowband laser light

$$\frac{dI_{\mathrm{tot}}}{dz} = g(\omega_{\mathrm{L}})I_{\mathrm{tot}} = \sigma(\omega_{\mathrm{L}}-\omega_{21})\Delta N I_{\mathrm{tot}}, \tag{3.77}$$

$$g(\omega_{\mathrm{L}}) = \frac{g_0(\omega_{\mathrm{L}})}{1+I_{\mathrm{tot}}/I_{\mathrm{sat}}(\omega_{\mathrm{L}})}. \tag{3.78}$$

[26] Note that spontaneous emission does not require any tweaks, since the total probability per unit time remains A_{21}.

Problems

3.1 Consider a cylindrical laser beam with diameter $d = 5$ mm propagating through an active crystal of length $L = 2$ cm. The energy spectral density of the light field in the active medium has a Lorentzian shape, and is given by

$$\rho(\omega) = \frac{\rho_0}{\pi}\frac{\gamma/2}{(\omega-\omega_{21})^2+(\gamma/2)^2},$$

where $\rho_0 = 2.5\cdot 10^{-5}$ Jm^{-3} describes the peak energy spectral density, $\gamma = 2\pi \cdot 1$ GHz is the full-width at half maximum of the Lorentzian, and the centre frequency of the beam is $\omega_{21} = 2\pi \cdot 474$ THz. What is the number of photons inside the active crystal? **Hint:** it may not be a bad idea to Google "Lorentzian function" and to approximate all the photons to have the same frequency ω_{21}; if you're brave enough, you can numerically check the validity of this approximation.

3.2 An atomic transition occurs at 633 nm. What is the ratio between the populations of the upper and lower states of the transition at room temperature (298 K)? Determine the temperature T so that 15% of the atoms are in the upper states. Other states and degeneracies can be neglected.

3.3 The first maser (microwave "laser") was demonstrated in 1954, and it operated at a frequency $f =$ 24 GHz. The first laser, on the other hand, was demonstrated in 1960 at the wavelength $\lambda = 694$ nm. Estimate the ratio N_2/N_1 between the populations of the upper and lower states of the active transition in both cases, assuming room temperature (298 K) and thermal equilibrium. How can the results explain the historical development of masers and lasers?

3.4 An excited state of an atom has a radiative lifetime $\tau^{(\mathrm{rad})}$ and a non-radiative lifetime $\tau^{(\mathrm{nr})}$. These lifetimes account for decay to all energy levels below the excited state. Show that the total lifetime τ_2 of the state (i.e., the time after which the population density has reduced by a factor of e^{-1}) is given by:

$$\frac{1}{\tau_2} = \frac{1}{\tau^{(\mathrm{rad})}} + \frac{1}{\tau^{(\mathrm{nr})}}.$$

3.5 The wavelength of the D-transition of sodium atoms is 589 nm, and the spontaneous emission lifetime of the upper level of the transition is 10 ns. At what rate does a sodium atom absorb photons from the field of a black body at 1200 K? You can neglect degeneracies of the energy levels.

3.6 When deriving the differential equation for optical amplification in a gain medium, we used the relationship $\rho(\omega) = I(\omega)/c$. By considering a laser beam with cross-sectional area A propagating at the speed of light for time Δt, prove this relationship.

3.7 Consider a typical 3-level system where the laser transition is between states 2 and 1, as shown in the figure below. Atoms are pumped into state 3 at a rate $W_\mathrm{p}N_1$, where N_1 is the population density of the ground state. The spontaneous lifetime of level 3 is short, so atoms decay immediately to the upper laser level 2. Accordingly, you can assume $N_3 \approx 0$.

(a) Write down the rate equations for the population densities in states 2 and 1.

(b) Show that the rate equations are closed, i.e., that

$$\frac{d}{dt}(N_2 + N_1) = 0.$$

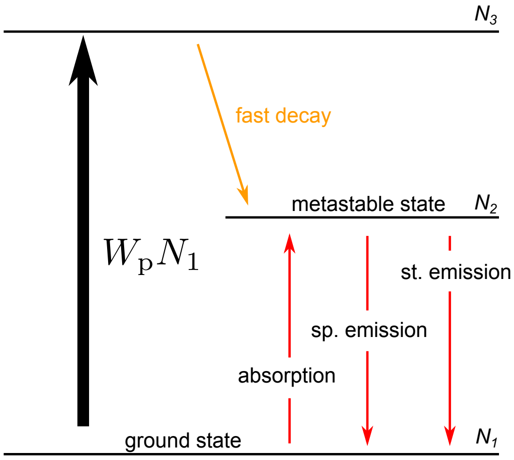

Figure 15: Schematic illustration of a 3-level system.

(c) Assuming steady-state operation, derive an expression for N_2/N_1, i.e., the ratio of atoms in state 2 and state 1. Show that there exists a pump threshold above which population inversion can be reached, and derive an expression for this condition.

(d) Assuming steady-state, derive an explicit expression for the population inversion $\Delta N = N_2 - N_1$. In particular, show that the inversion can be written in the exact same form as for a 4-level system, and derive appropriate expressions for the reduced pump rate and saturation intensity.

3.8 A spectrally narrowband laser beam with small initial intensity $I(0) = 0.2\ \mathrm{Wm}^{-2}$ propagates through an active medium and interacts with a homogeneously broadened transition. The interaction cross-section at the laser line is $\sigma(\omega_\mathrm{L} - \omega_{21}) = 3.7 \times 10^{-19}\ \mathrm{cm}^2$, the small-signal inversion is $\Delta N_0 = 1.35 \times 10^{17}\ \mathrm{cm}^{-3}$, and the saturation intensity $I_\mathrm{sat} = 10\ \mathrm{Wm}^{-2}$. What is the intensity of the amplified beam at the medium output, when

(a) The gain medium has a length of $L = 2\ \mathrm{cm}$? **Hint:** Be careful with the units...

(b) The gain medium has a length of $L = 1\ \mathrm{m}$? **Hint:** You'll need a computer to solve this one... Why?

3.9 The transition energy between the ground state of helium and the first excited state is 21.3 eV

(a) Calculate the Doppler linewidth for this transition at room temperature (300 K), and compare it with the corresponding natural linewidth. (The radiative lifetime of the excited state is 0.57 ns.)

(b) Calculate the interaction cross section at the line center.

3.10 A homogeneously broadened atomic transition has a centre frequency $\omega_{21} = 2\pi \times 282$ THz and a linewidth $\gamma = 2\pi \times 25$ GHz. The lifetime of the upper state against spontaneous emission is $\tau_\mathrm{sp} = 1.3\ \mathrm{ms}$, and the populations in steady-state operation are $N_2 = 10^{15}\ \mathrm{m}^{-3}$ and $N_1 = 10^{14}\ \mathrm{m}^{-3}$. The transition interacts with a laser beam with a Gaussian energy spectral density:

$$\rho(\omega) = \rho_0 e^{-(\omega-\omega_{21})^2/(2\Delta\omega^2)},$$

where $\Delta\omega = 2\pi \times 5$ GHz characterises the width of the Gaussian and $\rho_0 = 10^{-19}\ \mathrm{Jm^{-3}Hz^{-1}}$ is the peak energy spectral density. Calculate the rate of change of the total photon density in the active medium due to stimulated emission and absorption. **Hint:** computer...

3.11 In the lectures we examined the 4-level rate equations assuming that the states only decay via spontaneous emission. Here, we will consider the more general case of arbitrary decay channels. Furthermore, we will assume that the laser beam has narrow bandiwdth, centre frequency ω_L, and that the atomic transition is homogeneously broadened.

(a) Show that the rate equations for a 4-level system can be written as

$$\begin{aligned}\frac{dN_2}{dt} &= R_2 + \frac{I_\mathrm{tot}}{\hbar\omega_\mathrm{L}}(N_1 - N_2)\sigma(\omega_\mathrm{L} - \omega_{21}) - \frac{N_2}{\tau_2}\\ \frac{dN_1}{dt} &= R_1 - \frac{I_\mathrm{tot}}{\hbar\omega_\mathrm{L}}(N_1 - N_2)\sigma(\omega_\mathrm{L} - \omega_{21}) + A_{21}N_2 - \frac{N_1}{\tau_1},\end{aligned}$$

where τ_1 and τ_2 are the total *fluorescnce* lifetimes of the energy levels.

(b) Derive an expression for population inversion and saturation intensity.

(c) Comment on how the obtained expressions differ from those derived in the lecture notes.

3.12 Consider a 4-level system as in Fig. 9(b) of the lecture notes, where the rates of spontaneous emission between levels $0, 1, 2$ are governed by the Einstein coefficients A_{21}, A_{20}, A_{10}. You can neglect other decay channels.

(a) What should be the relative magnitudes of the coefficients to ensure efficient laser operation? In particular, which of the coefficients should be largest, and which of them should be smallest?

(b) Is it possible to achieve population inversion in a 4-level system where the spontaneous emission lifetimes of the upper and lower states are $\tau_{21} = 1\ \mu$s and $\tau_{10} = 5\ \mu$s?

3.13 In the absence of population inversion, i.e., when $\Delta N = N_2 - N_1 < 0$, a light beam propagating through (an active) medium is absorbed according to the Beer-Lambert law:

$$\frac{dI}{dz} = -\alpha I, \tag{3.79}$$

where $\alpha > 0$ is known as the liner absorption coefficient with units of m^{-1}. You can assume the light beam to be spectrally narrowband.

(a) Derive an expression for α in terms of the interaction cross-section $\sigma(\omega - \omega_{21})$ and population inversion ΔN.

(b) Especially in fibre optics, it is conventional to express loss with units of $\mathrm{dB/km}$. Find an expression for loss α_dB with units of $\mathrm{dB/km}$ as a function of α (with units of m^{-1}). Remember that dBs are defined as

$$10\log_{10}\frac{I_\mathrm{out}}{I_\mathrm{in}}.$$

4 Maxwell's equations

So far, we have considered the interactions of light and matter in the photon picture, where light is described as a bunch of zero-mass particles. But of course, we know that light can also be described as an **electromagnetic wave**, consisting of electric and magnetic fields that are oscillating in time and space perpendicular to one another. In the framework of classical physics, the most general description of any phenomena involving electric or magnetic fields is given by Maxwell's equations[27]: a set of four equations that form the foundations of classical electromagnetism.

Maxwell's equations can be written in many different forms. Here we will use the so-called macroscopic Maxwell equations in differential form:

Maxwell's equations

Gauss' law (electricity):
$$\nabla \cdot \vec{\mathbf{D}} = \rho_{\mathrm{f}} \tag{4.1}$$

Gauss' law (magnetism):
$$\nabla \cdot \vec{\mathbf{B}} = 0 \tag{4.2}$$

Maxwell-Faraday law of induction:
$$\nabla \times \vec{\mathbf{E}} = -\frac{\partial \vec{\mathbf{B}}}{\partial t} \tag{4.3}$$

Ampere's law (with Maxwell's correction):
$$\nabla \times \vec{\mathbf{H}} = \vec{\mathbf{J}}_{\mathrm{f}} + \frac{\partial \vec{\mathbf{D}}}{\partial t} \tag{4.4}$$

The physical quantities have the following meaning:

- $\vec{\mathbf{E}}(\vec{\mathbf{r}}, t)$: Electric field (in units of $\mathrm{V/m}$).
- $\vec{\mathbf{B}}(\vec{\mathbf{r}}, t)$: Magnetic field (in units of T).
- $\vec{\mathbf{D}}(\vec{\mathbf{r}}, t)$: Electric displacement field (in units of $\mathrm{C/m^2}$).
- $\vec{\mathbf{H}}(\vec{\mathbf{r}}, t)$: Magnetic displacement field (in units of $\mathrm{A/m}$) [28].
- $\vec{\mathbf{J}}_{\mathrm{f}}(\vec{\mathbf{r}}, t)$: Density of free current (in units of $\mathrm{Am^{-2}}$).
- ρ_{f}: Density of free charge (in units of $\mathrm{Cm^{-3}}$).

Each of the field variables in Eqs. (4.1)– (4.4) is a vector with three cartesian components. The symbol ∇ describes a vector differential operator, defined in three cartesian coordinates (with unit vectors $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$) as

$$\nabla = \frac{\partial}{\partial x}\hat{\mathbf{x}} + \frac{\partial}{\partial y}\hat{\mathbf{y}} + \frac{\partial}{\partial z}\hat{\mathbf{z}}. \tag{4.5}$$

[27] An even more general description is offered by quantum electrodynamics (QED), which is beyond the scope of these notes. Maxwell's equations correspond to an approximation of some aspects of QED. This approximation is extremely good; deviations typically arise only for extremely small light intensity levels, when behaviours of individual photons becomes important.

[28] Historically, $\vec{\mathbf{H}}$ is often called the magnetic field, while $\vec{\mathbf{B}}$ is referred to as the magnetic flux density. This is a remnant of times when magnetism was described to arise from magnetic poles. Today, it is understood that currents generate magnetism, such that $\vec{\mathbf{B}}$ is the more fundamental quantity, and hence referred to as the magnetic field here.

In Eqs. (4.1) and (4.2) we take the dot product between ∇ and one of the field vectors. The result $\nabla \cdot \vec{\mathbf{A}}$ is known as the *divergence* of $\vec{\mathbf{A}}$[29], and roughly speaking describes the vector field's tendency to converge toward ($\nabla \cdot \vec{\mathbf{A}} < 0$) or diverge from ($\nabla \cdot \vec{\mathbf{A}} > 0$) a point. In Eqs. (4.3) and (4.4) we take the cross product between ∇ and one of the field vectors. The result $\nabla \times \vec{\mathbf{A}}$ is known as the *curl* of $\vec{\mathbf{A}}$, and roughly speaking measures the degree to which the vector field is rotating about a given point.

The physical meaning of the different Maxwell's equations may not be so apparent from their differential form. The integral forms are perhaps clearer in that sense, but unfortunately not so useful for the analysis of EM waves[30]. The meanings are roughly as follows:

- **Gauss's law for electricity:** Net electric flux through a closed surface is proportional to the electric charge enclosed within the surface.

- **Gauss's law for magnetism:** As above, but since there are no magnetic monopoles, the net magnetic flux through a closed surface is zero.

- **Maxwell-Faraday law of induction:** A changing magnetic field gives rise to an electric field.

- **Ampere's law (with Maxwell's addition):** A changing electric field (or a current) gives rise to a magnetic field.

4.1 Polarization and magnetization

The Maxwell's equations (4.1)–(4.4) correspond to the "macroscopic" form of the equations: unlike their "microscopic" counterparts, they do not explicitly include electric charges or magnetic dipoles *bound* to individual atoms (hence the only charges and currents appearing are "free"). Rather, charges and currents associated with individual atoms are averaged into electric and magnetic dipoles described by macroscopic polarization and magnetization fields $\vec{\mathbf{P}}$ and $\vec{\mathbf{M}}$, respectively. This averaging considerably simplifies the analysis of macroscopic situations[31], as we can ignore details on a fine (atomic) scale.

The macroscopic polarization and magnetization fields come into play through the definitions of the displacement fields $\vec{\mathbf{D}}$ and $\vec{\mathbf{H}}$:

$$\vec{\mathbf{D}}(\vec{\mathbf{r}},t) = \varepsilon_0 \vec{\mathbf{E}}(\vec{\mathbf{r}},t) + \vec{\mathbf{P}}(\vec{\mathbf{r}},t) \tag{4.6}$$

$$\vec{\mathbf{H}}(\vec{\mathbf{r}},t) = \frac{1}{\mu_0} \vec{\mathbf{B}}(\vec{\mathbf{r}},t) - \vec{\mathbf{M}}(\vec{\mathbf{r}},t) \tag{4.7}$$

The variables have the following meaning:

- $\vec{\mathbf{P}}(\vec{\mathbf{r}},t)$: density of permanent or induced electric dipole moments.

- $\vec{\mathbf{M}}(\vec{\mathbf{r}},t)$: density of permanent or induced magnetic dipole moments.

- ε_0: vacuum permittivity.

- μ_0: vacuum permeability.

[29] Here $\vec{\mathbf{A}}$ is just a generic field vector.

[30] Note nevertheless that the integral and differential forms are fully equivalent.

[31] Note that "macroscopic" here can still be extremely small; we are only averaging over individual atoms, whose spacing is several thousand times smaller than typical wavelength of light.

4.1.1 Constitutive relations in linear optics

To apply the macroscopic Maxwell's equations, we need to specify the dependence of the polarization $\vec{\mathbf{P}}$ (hence the bound charges) and the magnetisation $\vec{\mathbf{M}}$ (hence the bound currents) on the applied electric and magnetic fields, as well as the distributions of free charges ρ_f and currents $\vec{\mathbf{J}}_\mathrm{f}$. The equations specifying these relations are known as *constitutive relations*. In general, the relations can be complicated, but luckily we are interested only in materials relevant to optics, and light intensities that are not outrageous[32]. Under a wide range of typical conditions encountered in the study of optics, the constitutive equations read

$$\boxed{\begin{aligned} \rho_\mathrm{f}(\vec{\mathbf{r}},t) &= 0 && (4.8)\\ \vec{\mathbf{M}}(\vec{\mathbf{r}},t) &= 0 && (4.9)\\ \vec{\mathbf{P}}(\vec{\mathbf{r}},t) &= \varepsilon_0\chi\vec{\mathbf{E}}(\vec{\mathbf{r}},t) && (4.10)\\ \vec{\mathbf{J}}_\mathrm{f}(\vec{\mathbf{r}},t) &= \sigma\vec{\mathbf{E}}(\vec{\mathbf{r}},t) && (4.11) \end{aligned}}$$

These relations describe materials that have no free charges ($\rho_\mathrm{f} = 0$) and that are non-magnetic ($\vec{\mathbf{M}} = 0$). The materials are also homogeneous and isotropic[33], and respond linearly to the electric field ($\vec{\mathbf{P}} = \varepsilon_0\chi\vec{\mathbf{E}}$), where χ is known as the linear electric susceptibility. This last condition implies that the electric dipoles induced by the electric field align (linearly) with the electric field, with χ describing the degree of polarization in response to an applied electric field. Finally, the free current is assumed to obey Ohm's law in the form $\vec{\mathbf{J}}_\mathrm{f} = \sigma\vec{\mathbf{E}}$, where σ is the conductivity of the material. We will be mostly interested in dielectrics (i.e., insulators), for which $\sigma = 0$, but include the possibility of $\sigma \neq 0$ for completeness.

4.2 EM wave equation

Equations (4.3) and (4.4) show that a changing magnetic field can induce an electric field and vice versa, already hinting at the possibility of a self-sustaining EM wave. To show that Maxwell's equations indeed have solutions in the form of waves, we shall first manipulate the equations using the constitutive relations introduced above so as to derive a wave equation. We assume here the medium to be dielectric ($\sigma = 0$) for simplicity, and leave it as an exercise for the reader to analyse the situation $\sigma \neq 0$.

We first take the *curl* of both sides of Eq. (4.3):

$$\nabla \times \nabla \times \vec{\mathbf{E}} = -\nabla \times \frac{\partial \vec{\mathbf{B}}}{\partial t} \tag{4.12}$$

We then change the order of *curl* and time derivative[34], and with the help of Eq. (4.9) write $\vec{\mathbf{B}} = \mu_0\vec{\mathbf{H}}$. This yields:

$$\nabla \times \nabla \times \vec{\mathbf{E}} = -\mu_0\frac{\partial}{\partial t}(\nabla \times \vec{\mathbf{H}}). \tag{4.13}$$

[32] With large intensities, the material polarization $\vec{\mathbf{P}}$ becomes a *nonlinear* function of the applied electric field; the resulting nonlinear optical effects will be considered later.

[33] Such that χ and σ do not depend on position.

[34] Both operators are linear and commute, allowing us to do this.

Using now Eq. (4.4) with $\vec{\mathbf{J}}_\mathrm{f} = 0$[35], we obtain

$$\nabla \times \nabla \times \vec{\mathbf{E}} = -\mu_0 \frac{\partial^2 \vec{\mathbf{D}}}{\partial t^2}. \tag{4.14}$$

Using the definition of the displacement field [Eq. (4.6)] and the constitutive Eq. (4.10) we get

$$\nabla \times \nabla \times \vec{\mathbf{E}} = -\mu_0 \varepsilon \frac{\partial^2 \vec{\mathbf{E}}}{\partial t^2}, \tag{4.15}$$

where

$$\boxed{\varepsilon = \varepsilon_0 \varepsilon_\mathrm{r} = \varepsilon_0 (1 + \chi)} \tag{4.16}$$

is the permittivity of the material and $\varepsilon_\mathrm{r} = 1 + \chi$ is the relative permittivity. Finally, we make use of the following vector calculus identity[36]:

$$\nabla \times \nabla \times \vec{\mathbf{E}} = \nabla(\nabla \cdot \vec{\mathbf{E}}) - \nabla^2 \vec{\mathbf{E}}, \tag{4.17}$$

which together with Eqs. (4.1), (4.6), and (4.10) allows us to rewrite Eq. (4.15) as

$$\boxed{\nabla^2 \vec{\mathbf{E}} - \mu_0 \varepsilon \frac{\partial^2 \vec{\mathbf{E}}}{\partial t^2} = 0.} \tag{4.18}$$

The operator ∇^2 is known as the *Laplacian*, and is defined in Cartesian coordinates as

$$\nabla^2 = \frac{\partial}{\partial x^2} + \frac{\partial}{\partial y^2} + \frac{\partial}{\partial z^2}. \tag{4.19}$$

Note that an identical equation can be derived for the magnetic field $\vec{\mathbf{B}}$.

4.2.1 Link to wave equations in general

Equation (4.18) has the generic form of a wave equation. Analogous equations arise for the description of all kinds of different waves, such as sound waves, waters waves, waves on a string and so on. To gain some general insights before investigating EM waves in particular, we write the wave equation in one space dimension, describing e.g. the mechanical disturbance on a guitar string [denoted $u(x,t)$]:

$$\frac{\partial^2 u}{\partial x^2} - \frac{1}{v^2} \frac{\partial^2 u}{\partial t^2} = 0, \tag{4.20}$$

where v is a fixed constant. The most general solution of the one-dimensional wave equation has the following form[37]:

$$u(x,t) = F(x - vt) + G(x + vt), \tag{4.21}$$

[35] Recall that we are considering a dielectric, such that $\sigma = 0$.

[36] Valid for arbitrary vector fields, not just $\vec{\mathbf{E}}$.

[37] This solution is only valid when v is constant. As we shall see below, in optics this is not typically the case as v depends on frequency. In this case, Eq. (4.21) is not valid in general, but only applies for monochromatic waves.

where F and G are arbitrary disturbance profiles. These profiles travel along the x-direction, with a speed given by v; assuming $v > 0$, F moves towards positive x, while G moves toward negative x. To see why this is the case, consider an arbitrary point x_0 at time $t = 0$, where the disturbance has the value $F(x_0)$. After some time t, this point along the disturbance moves to $x = x_0 + d$ [see Fig. 16] such that

$$F(x_0) = F(x_0 + d - vt). \tag{4.22}$$

From this, we obtain

$$v = \frac{d}{t}. \tag{4.23}$$

The right-hand side is obviously the definition of velocity $(\text{displacement}/\text{time})$.

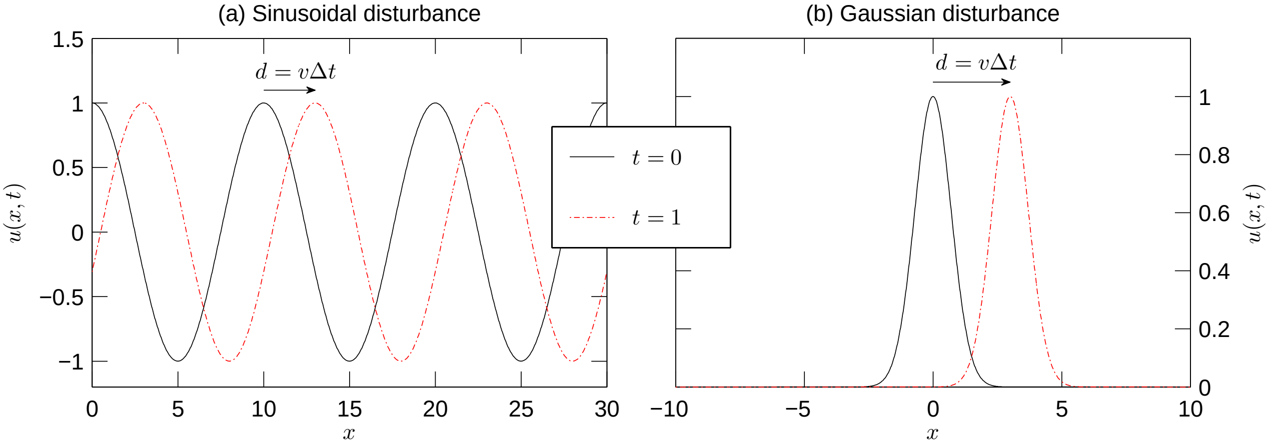

Figure 16: Two example solutions to the wave equation. In (a) the disturbance is sinusoidal, $u(x, t) = \cos[2\pi f(x - vt)]$, with $f = 0.1\ \mathrm{m}^{-1}$ while in (b) the disturbance is Gaussian, $u(x, t) = e^{-(x-vt)^2}$. In both cases, the wave velocity $v = 3\ \mathrm{m/s}$ and the disturbances are plotted at two different times $t = 0$ s and $t = 1$ s as indicated.

4.3 Speed of light and refractive index

Comparing Eqs. (4.18) and (4.20), we see that the solutions of the EM wave equation will travel at a speed of

$$v = \frac{1}{\sqrt{\mu_0 \varepsilon}} = \frac{1}{\sqrt{\mu_0 \varepsilon_0 \varepsilon_\mathrm{r}}} \tag{4.24}$$

In vacuum $\chi = 0$ and $\varepsilon_\mathrm{r} = 1$, and so we get the speed of EM waves in vacuum:

$$\boxed{c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} \approx 2.998 \times 10^8\ \mathrm{m/s}.} \tag{4.25}$$

This is, of course, the speed of light[38]

[38] We arrived at this observation from an educated background, so the surprise may not be so great. But consider what it must have been like in 1862, when people did not know that light is an EM wave, and Maxwell showed that general equations of EM fields can be recast into a wave equation, with a speed very close to the known speed of light. Physics really doesn't get more beautiful than that!

In media (i.e., outside of a vacuum), the linear susceptibility $\chi \neq 0$, and accordingly $\varepsilon_\mathrm{r} \neq 1$. From Eq. (4.24), the velocity of EM waves becomes:

$$v = \frac{c}{\sqrt{\varepsilon_\mathrm{r}}} = \frac{c}{\sqrt{1+\chi}}. \tag{4.26}$$

We can recognise the refractive index of the medium as

$$\boxed{n = \sqrt{\varepsilon_\mathrm{r}} = \sqrt{1+\chi}.} \tag{4.27}$$

4.3.1 Frequency-dependence of refractive index

The analysis above suggests that refractive index is a constant. This is not true in real life, however, as refractive indices are known to depend on frequency [see e.g. Fig. 17]. The discrepancy arises from the fact that the constitutive equation (4.10) approximates polarization to react instantaneously to the electric field. In reality, polarization arises with some delay, and can be more formally represented as a convolution between a time-dependent susceptibility $\chi(t)$ and the electric field:

$$\boxed{\vec{\mathbf{P}}(\vec{\mathbf{r}}, t) = \varepsilon_0 \int_{-\infty}^{t} \chi(t-t')\vec{\mathbf{E}}(\vec{\mathbf{r}}, t')dt',} \tag{4.28}$$

where $\chi(t)$ is a time-dependent susceptibility, describing the material's temporal response to an electric field.

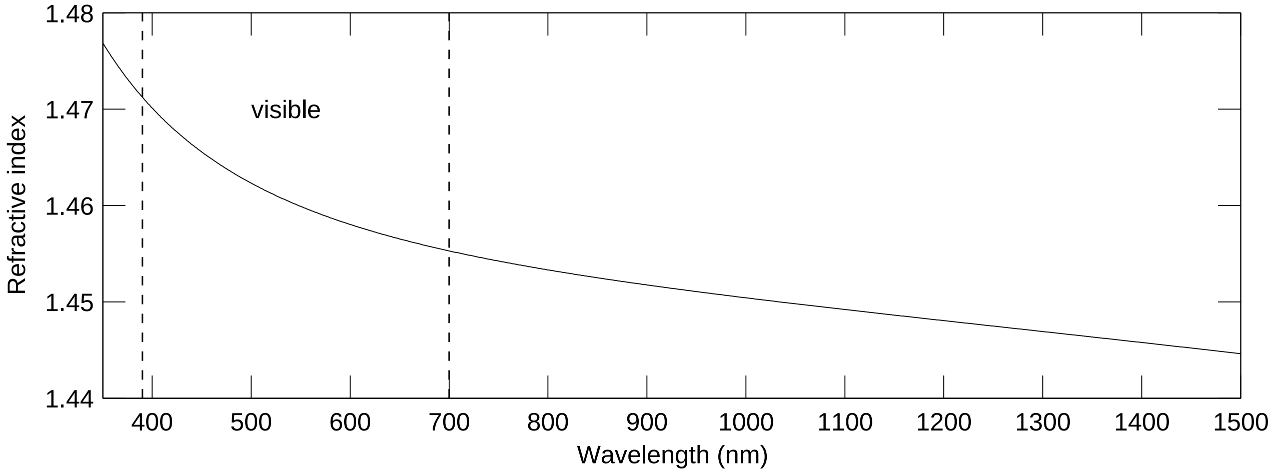

Figure 17: Refractive index of fused silica glass as a function of wavelength $\lambda = f/c = \omega/(2\pi)$.

The relationship can be written in a prettier form in the frequency domain. Specifically, considering a monochromatic EM wave with frequency ω and complex amplitude $\tilde{E}(\omega)$, such that $E(t) = \tilde{E}(\omega)e^{i\omega t}$, the convolution can be easily computed to yield $P(t) = \tilde{P}(\omega)e^{i\omega t}$, where the polarisation amplitude

$$\tilde{P}(\omega) = \varepsilon_0\tilde{\chi}(\omega)\tilde{E}(\omega), \tag{4.29}$$

and $\tilde{\chi}(\omega)$ is the Fourier transform of $\chi(t)$. Thus, in general, the simple constitutive Eq. (4.10) applies at all different frequencies ω, connecting the complex amplitudes of the polarisation and the electric field oscillating at that frequency. However, in general the susceptibility varies with frequency; in the time domain, the frequency-dependent susceptibility manifests itself as a convolution. Note that mathematically these relationships are simply manifestations of the **convolution theorem:**

> The Fourier transform of a convolution of two functions is equal to the product of the function's Fourier transforms.

Using the above formalism, the refractive index is also found to have a frequency-dependence:

$$n(\omega) = \sqrt{1 + \tilde{\chi}(\omega)} \tag{4.30}$$

In what follows, we will for simplicity assume polarization to respond instantaneously, but reserve the right to make use of frequency-dependent refractive indices. While this may appear phenomenological, the author assures that identical results are obtained when using the full polarization convolution above.

4.4 Plane EM waves

We will now proceed to derive an important class of EM wave solutions: plane waves. To this end, we look for monochromatic plane wave solutions of Eq. (4.18) by using the following ansatz:

$$\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = \vec{\mathbf{E}}_{0r} \cos\left(\omega t - \vec{\mathbf{k}} \cdot \vec{\mathbf{r}} + \phi_E\right), \tag{4.31}$$

$$\vec{\mathbf{B}}(\vec{\mathbf{r}}, t) = \vec{\mathbf{B}}_{0r} \cos\left(\omega t - \vec{\mathbf{k}} \cdot \vec{\mathbf{r}} + \phi_B\right). \tag{4.32}$$

Here $\vec{\mathbf{E}}_{0r}$ and $\vec{\mathbf{B}}_{0r}$ are the amplitudes of the electric and magnetic fields, respectively, ω is the (angular) frequency of the wave, $\phi_{E,B}$ are arbitrary phases, and $\vec{\mathbf{k}} = k_x\hat{\mathbf{x}} + k_y\hat{\mathbf{y}} + k_z\hat{\mathbf{z}}$ is known as the *wave vector*. As we shall soon see, the wave vector defines the direction of propagation of the plane wave. Note that the magnitude of the wave vector, $k = \sqrt{k_x^2 + k_y^2 + k_z^2}$, is known as the *wave number*. In what follows, we will only examine the electric field for simplicity.

Dealing with trigonometric functions is a little bit cumbersome, and a better way is to express the electric field in (complex) phasor form:

$$\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = \mathrm{Re}\left[\vec{\mathbf{E}}_0 e^{i(\omega t - \vec{\mathbf{k}} \cdot \vec{\mathbf{r}})}\right], \tag{4.33}$$

where we defined the complex amplitude $\vec{\mathbf{E}}_0 = \vec{\mathbf{E}}_{0r} e^{i\phi_E}$. Because the Maxwell's equations are linear and only involve simple derivatives[39], we can neglect the real part in our calculations and simply analyse the complex field:

$$\boxed{\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = \vec{\mathbf{E}}_0 e^{i(\omega t - \vec{\mathbf{k}} \cdot \vec{\mathbf{r}})}} \tag{4.34}$$

When dealing with EM waves, we almost always use this complex notation, as it greatly simplifies the analyses. It should nevertheless be borne in mind that the electric field must be a real quantity, and so in any calculations, the *physical values* will be given by the *real part* of the solution.

[39] Derivatives of the fields are equal to the derivatives of the real part of the complex field.

For the plane wave defined by Eq. (4.34) to be a solution of the Maxwell's equations, the different variables ($\vec{\mathbf{E}}_0$, $\vec{\mathbf{B}}_0$, $\vec{\mathbf{k}}$, ω) must satisfy a set of conditions. The conditions can be found by injecting the ansatz (4.34) into the Maxwell's equations. This calls for a big bunch of differential operations. Luckily, for complex plane waves of the form $\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = \vec{\mathbf{E}}_0 e^{i(\omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}})}$, the differential operators are very easy to deal with[40]:

$$\boxed{\begin{aligned} \frac{\partial \vec{\mathbf{E}}}{\partial t} &= i\omega \vec{\mathbf{E}}, \\ \nabla \cdot \vec{\mathbf{E}} &= -i\vec{\mathbf{k}} \cdot \vec{\mathbf{E}}, \\ \nabla \times \vec{\mathbf{E}} &= -i\vec{\mathbf{k}} \times \vec{\mathbf{E}}, \\ \nabla^2 \vec{\mathbf{E}} &= -k^2 \vec{\mathbf{E}}. \end{aligned}} \qquad \begin{aligned} &(4.35) \\ &(4.36) \\ &(4.37) \\ &(4.38) \end{aligned}$$

Or in more concise form:

$$\boxed{\begin{aligned} \frac{\partial}{\partial t} &\to i\omega, \\ \nabla &\to -i\vec{\mathbf{k}}. \end{aligned}} \qquad \begin{aligned} &(4.39) \\ &(4.40) \end{aligned}$$

4.4.1 Dispersion relation

The first condition to derive is the relationship between the wave vector $\vec{\mathbf{k}}$ (and in particular the wave number) and the frequency ω, i.e., the *dispersion relation*[41]. To this end, we inject our ansatz (4.34) into Eq. (4.18). Using the differential operations summarised above, we obtain

$$-k^2 \vec{\mathbf{E}} + \mu_0 \varepsilon \omega^2 \vec{\mathbf{E}} = 0. \tag{4.41}$$

For nontrivial solutions ($\vec{\mathbf{E}} \neq 0$), we must have

$$\boxed{k^2 = \mu_0 \varepsilon \omega^2.} \tag{4.42}$$

This is the dispersion relation of EM waves in dielectric media. Using the definition of the refractive index, we have

$$\boxed{k(\omega) = \pm n \frac{\omega}{c}.} \tag{4.43}$$

Note that ± simply implies that the wave can be travelling either along or against $\vec{\mathbf{k}}$[42]. For our plane wave to be a solution to the Maxwell's equations, the wave number and frequency must be related by Eq. (4.43).

4.4.2 Other requirements

By injecting our ansatz into the full Maxwell's equations, we find additional conditions (note that Eqs. (4.35) – (4.38) significantly simplify the equations):

[40] It is left as an exercise to verify these relations.

[41] Dispersion relations are of principal importance to all forms of waves, as they specify the relationship between frequency and wave number.

[42] Note that $\vec{\mathbf{k}} = k\hat{\mathbf{k}}$.

Conditions between field amplitudes

$\nabla \cdot \vec{\mathbf{E}} = 0$	$\Rightarrow$	$-i\vec{\mathbf{k}} \cdot \vec{\mathbf{E}}_0 = 0$	$\vec{\mathbf{E}}_0$ is perpendicular to $\vec{\mathbf{k}}$.
$\nabla \cdot \vec{\mathbf{B}} = 0$	$\Rightarrow$	$-i\vec{\mathbf{k}} \cdot \vec{\mathbf{B}}_0 = 0$	$\vec{\mathbf{B}}_0$ is perpendicular to $\vec{\mathbf{k}}$.
$\nabla \times \vec{\mathbf{E}} = -\dfrac{\partial \vec{\mathbf{B}}}{\partial t}$	$\Rightarrow$	$-i\vec{\mathbf{k}} \times \vec{\mathbf{E}}_0 = -i\omega\vec{\mathbf{B}}_0$	$\vec{\mathbf{E}}_0$ is perpendicular to $\vec{\mathbf{B}}_0$.
$\nabla \times \vec{\mathbf{B}} = \dfrac{\partial \vec{\mathbf{D}}}{\partial t}$	$\Rightarrow$	$-i\vec{\mathbf{k}} \times \vec{\mathbf{B}}_0 = i\omega\mu_0\varepsilon\vec{\mathbf{E}}_0$	$\vec{\mathbf{B}}_0$ is perpendicular to $\vec{\mathbf{E}}_0$.

The first two conditions show that $\vec{\mathbf{E}}$ and $\vec{\mathbf{B}}$ must oscillate in directions perpendicular to $\vec{\mathbf{k}}$; the last two conditions show that $\vec{\mathbf{E}}$ and $\vec{\mathbf{B}}$ must be mutually perpendicular. Given that $\vec{\mathbf{k}}$ defines the direction of propagation of the EM wave, the conditions show that the following well-known fact derives directly from Maxwell's equations:

Electromagnetic waves consist of electric and magnetic fields that oscillate perpendicular to one another and perpendicular to the direction of wave propagation.

Note that the direction of electric field oscillations by convention defines the **polarization** of the EM wave. The polarization of a plane EM wave can in general have any direction (i.e., the vector $\vec{\mathbf{E}}_0$ can point to any direction), provided that direction is transverse to the direction of the wave vector $\vec{\mathbf{k}}$.

The final two conditions listed above also enforce that the complex amplitudes $\vec{\mathbf{E}}_0$ and $\vec{\mathbf{B}}_0$ must be related to each other. The example below illustrates the conditions in action.

Example.
Consider an EM wave propagating in the $+z$-direction, so that $\vec{\mathbf{k}} = k\hat{\mathbf{z}} = [0, 0, k]^T$. Since the fields must be perpendicular to $\vec{\mathbf{k}}$, this implies that $\vec{\mathbf{E}}_0 = [E_x, E_y, 0]^T$ and $\vec{\mathbf{B}}_0 = [B_x, B_y, 0]^T$. From the third (or fourth) condition above, $\vec{\mathbf{k}} \times \vec{\mathbf{E}}_0 = k[-E_y, E_x, 0]^T = \omega[B_x, B_y, 0]^T = \omega\vec{\mathbf{B}}_0$. Thus we conclude that the cartesian components of the field amplitudes must satisfy the following ratios:

$$\frac{E_y}{B_x} = -\frac{E_x}{B_y} = -\frac{\omega}{k}. \tag{4.44}$$

Provided that these ratios are fulfilled together with the dispersion relation, then the following EM waves satisfy the Maxwell's equations:

$$\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = \vec{\mathbf{E}}_0 e^{i(\omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}})}, \tag{4.45}$$

$$\vec{\mathbf{B}}(\vec{\mathbf{r}}, t) = \vec{\mathbf{B}}_0 e^{i(\omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}})}. \tag{4.46}$$

4.4.3 Properties of plane waves

1. At any fixed time t, the field $\vec{\mathbf{E}}(\vec{\mathbf{r}},t) = \vec{\mathbf{E}}_0 e^{i(\omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}})}$ has the same value at each $\vec{\mathbf{r}}$ for which the phase is constant, i.e., $\omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}} = \text{cnst}$. Consider now two such points, $\vec{\mathbf{r}}_{1,2}$. They both satisfy $\vec{\mathbf{k}}\cdot\vec{\mathbf{r}}_{1,2} = \omega t - \text{cnst}$, and so
$$\vec{\mathbf{k}}\cdot\vec{\mathbf{r}}_1 - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}}_2 = \vec{\mathbf{k}}\cdot(\vec{\mathbf{r}}_1 - \vec{\mathbf{r}}_2) = 0. \tag{4.47}$$
This defines a plane that is perpendicular to $\vec{\mathbf{k}}$ [see Fig. 18]. In other words, **the points of equal field value of $\vec{\mathbf{E}}$ form planes in space.**[43] Along each plane in space that is perpendicular to $\vec{\mathbf{k}}$, the EM field takes the same value.

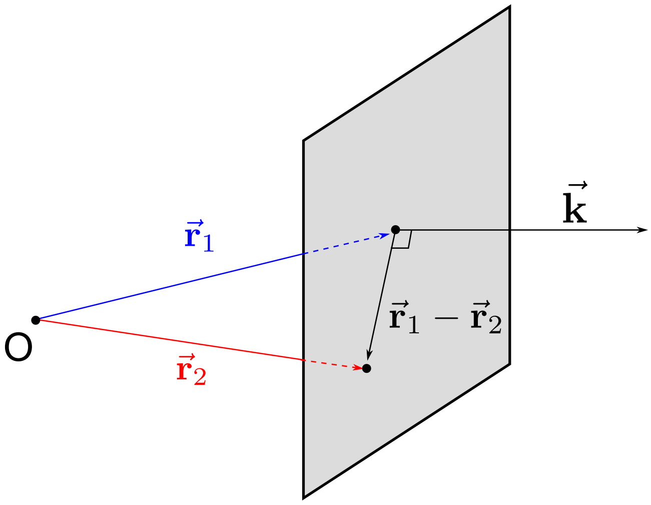

Figure 18: Schematic illustration of a plane of constant phase.

2. Planes of equal phase repeat periodically in space[44]. The period – i.e., the distance over which the wave's shape repeats itself – is simply the wavelength λ. In other words, if the phase at a point $\vec{\mathbf{r}}$ is $\phi = \omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}}$, then at a point $\vec{\mathbf{r}} + \lambda\,\hat{\mathbf{k}}$ the phase must be $\phi - 2\pi$. Mathematically:
$$\omega t - \vec{\mathbf{k}}\cdot(\vec{\mathbf{r}} + \lambda\,\hat{\mathbf{k}}) = \omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}} - 2\pi, \tag{4.48}$$
$$\lambda k(\hat{\mathbf{k}}\cdot\hat{\mathbf{k}}) = 2\pi. \tag{4.49}$$
And so we get
$$\boxed{\lambda = \frac{2\pi}{k} = \frac{2\pi c}{n\omega} = \frac{\lambda_0}{n},} \tag{4.50}$$
where $\lambda_0 = 2\pi\omega/c$ is the wavelength in vacuum.

[43] Hence the name plane waves.

[44] Recall that $e^{\omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}}}$ is periodic with a period of 2π.

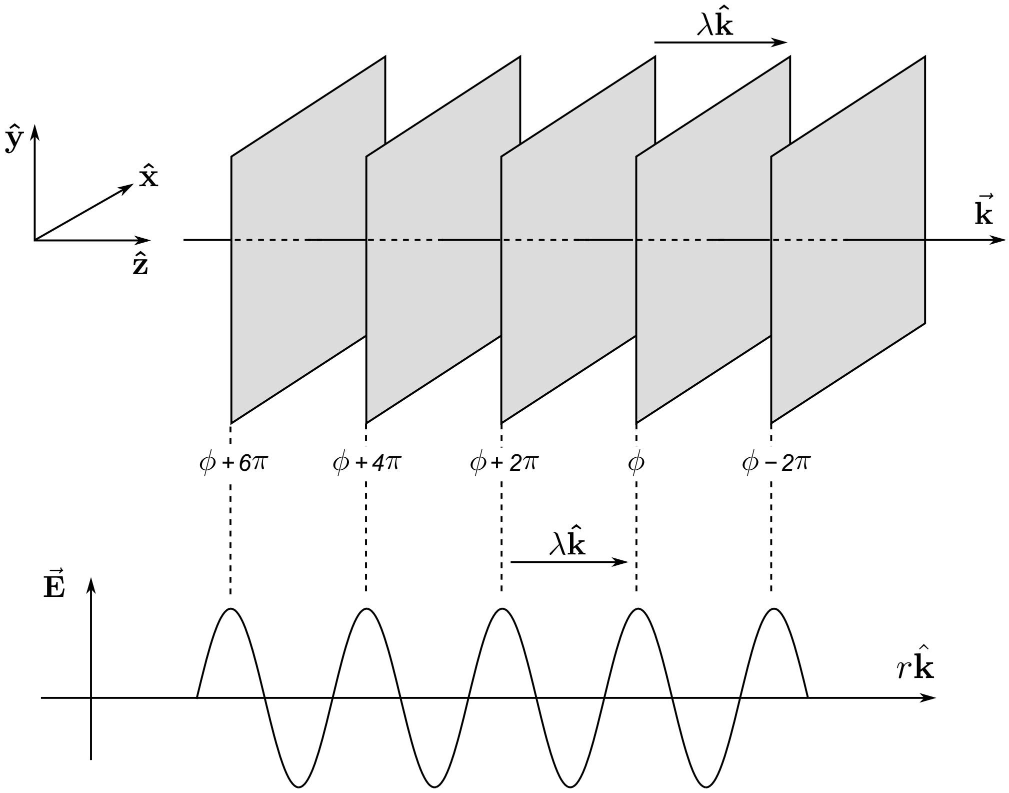

Figure 19: Planes of constant phase repeat in space with a period of one wavelength. For simplicity, here the wave vector is aligned with the z-axis, i.e., $\hat{\mathbf{k}} = \hat{\mathbf{z}}$

3. The wave fronts (planes of constant phase) are moving in the direction of the wave vector. In time dt, the fronts travel across the displacement $\vec{\mathbf{dr}} = dr\,\hat{\mathbf{k}}$ such that the phase remains constant:

$$\omega t - \vec{\mathbf{k}} \cdot \vec{\mathbf{r}} = \omega(t + dt) - \vec{\mathbf{k}} \cdot (\vec{\mathbf{r}} + \vec{\mathbf{dr}}), \tag{4.51}$$

$$0 = \omega\, dt - k\, dr(\hat{\mathbf{k}} \cdot \hat{\mathbf{k}}). \tag{4.52}$$

And so the wave's *phase velocity* v is

$$\boxed{v = \frac{dr}{dt} = \frac{\omega}{k} = \frac{c}{n}} \tag{4.53}$$

4.4.4 Intensity of EM waves

EM waves transport energy as they travel. The rate of energy transfer per unit area is described by the following vector:

$$\vec{\mathbf{S}} = \frac{1}{\mu_0}\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) \times \vec{\mathbf{B}}(\vec{\mathbf{r}}, t), \tag{4.54}$$

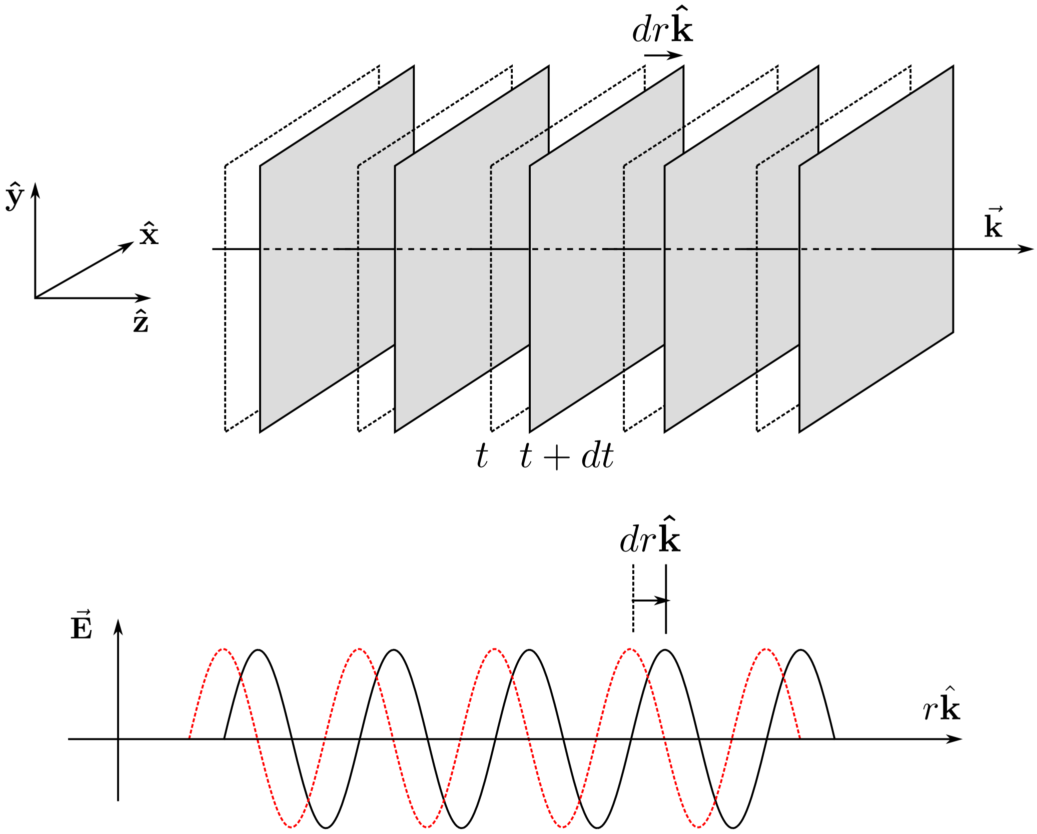

Figure 20: Planes of constant phase move in space as time progresses. During a short time interval dt, the planes are displaced along the wave vector by the amount dr.

which is known as the *Poynting vector*, and comes with units of Wm^{-2}. Given that $\vec{\mathbf{S}}$ arises as the cross product of $\vec{\mathbf{E}}$ and $\vec{\mathbf{B}}$, it is perpendicular to both; for a plane wave, the Poynting vector must therefore be parallel to the wave vector $\vec{\mathbf{k}}$. Furthermore, since $\vec{\mathbf{E}}$ and $\vec{\mathbf{B}}$ are perpendicular, the magnitude of the Poynting vector is

$$S(\vec{\mathbf{r}},t) = \frac{1}{\mu_0}E(\vec{\mathbf{r}},t)B(\vec{\mathbf{r}},t), \tag{4.55}$$

where E and B represent the corresponding scalar magnitudes of the electric and magnetic field vectors[45]. The conditions between field amplitudes for plane waves, outlined in section 4.4.2, show that

$$B(\vec{\mathbf{r}},t) = \frac{k}{\omega}E(\vec{\mathbf{r}},t) = \frac{n}{c}E(\vec{\mathbf{r}},t), \tag{4.56}$$

and the magnitude of the Poynting vector is thus

$$S(\vec{\mathbf{r}},t) = \frac{n}{c\mu_0}E^2(\vec{\mathbf{r}},t), \tag{4.57}$$

[45] In other words, if $\vec{\mathbf{A}}(\vec{\mathbf{r}},t) = [A_x(\vec{\mathbf{r}},t), A_y(\vec{\mathbf{r}},t), A_z(\vec{\mathbf{r}},t)]^T$ then $A = \sqrt{A_x(\vec{\mathbf{r}},t)^2 + A_y(\vec{\mathbf{r}},t)^2 + A_z(\vec{\mathbf{r}},t)^2}$.

Because the electric field oscillates very fast (with typical optical frequencies in the 300 THz range), we are not typically interested in the instantaneous magnitude of the Poynting vector. Rather, we are interested in the *time averaged* value, obtained by averaging the Poynting vector over several oscillation cycles. This time-average corresponds precisely to optical *intensity*:

$$I(\vec{\mathbf{r}}, t) = \langle S \rangle_T = \frac{1}{T} \int_{-T/2}^{T/2} S(\vec{\mathbf{r}}, t)\, dt \tag{4.58}$$

Because the Poynting vector involves a product of the field, we cannot use complex notation[46]. Instead, we must use the form

$$\vec{\mathbf{E}} = \vec{\mathbf{E}}_{0r} \cos(\omega t - \vec{\mathbf{k}} \cdot \vec{\mathbf{r}} + \phi). \tag{4.59}$$

The time average can be easily calculated[47], yielding

$$\boxed{I(t) = \frac{n}{c\mu_0} \frac{E_{0r}^2}{2}.} \tag{4.60}$$

And so the intensity of an EM wave is simply proportional to the square of the wave's amplitude. **In complex (phasor) notation, the field amplitude squared is replaced by the absolute value squared of the complex amplitude.**

4.4.5 Complex wave number

Our analysis above has assumed the wave number to be real. This is not always the case, as the material permittivity ε (hence, refractive index) can in general be complex [see Eq. (4.43)]. We can then write

$$k = k_{\mathrm{R}} - ik_{\mathrm{I}}. \tag{4.61}$$

The real part of the wave number, k_{R}, defines the wavelength and phase velocity of the wave. To see the effect of the imaginary part, we consider a plane wave propagating in the positive z direction:

$$\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = \vec{\mathbf{E}}_0 e^{i(\omega t - kz)}, \tag{4.62}$$

$$\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = \vec{\mathbf{E}}_0 e^{-k_{\mathrm{I}} z} e^{i(\omega t - k_{\mathrm{R}} z)}. \tag{4.63}$$

And so we see that the wave amplitude decreases ($k_{\mathrm{I}} > 0$) or increases ($k_{\mathrm{I}} < 0$) as the wave propagates. The wave's intensity would correspondingly decrease or increase as $I \propto \exp(-2k_{\mathrm{I}} z)$. In other words, the imaginary part of the wave number accounts for attenuation or amplification of the wave.

4.5 Do plane EM waves really exist

No they do not. Plane EM waves are idealizations that can never be created in practice. This can be readily understood by noting that the "planes" of constant field value $\vec{\mathbf{E}}$ of the wave extend everywhere: they have infinite extent, and hence carry infinite energy. Such a situation is clearly *unphysical*. Real light beams have some

[46] This is because $z_1 z_2 \neq Re[z_1 z_2]$.

[47] $\langle \cos^2(\omega t - \vec{\mathbf{k}} \cdot \vec{\mathbf{r}}) \rangle_T = 1/2$.

finite transverse structure (e.g. Gaussian profile) and they diffract: the beam width broadens with propagation distance. Moreover, plane EM waves are monochromatic. As we have already argued, there is no such thing as a monochromatic wave. All waves have some temporal structure, which gives rise to a finite spectral width; to the very least, the wave must start somewhere and end somewhere.

So why do we bother with plane waves? The reason is that plane waves are good approximations of real light beams (with finite width and nonzero bandwidth) when the beam width is much larger than the wavelength, and the wave is quasi-monochromatic. Because plane waves are fairly simple solutions of the Maxwell's equations, they allow us to gain important (and simple) insights into the behaviour of real light beams.

4.5.1 Introduction to Fourier transforms

Another important feature of plane waves is the fact that **any real beam can be represented as a superposition of plane waves**. This is because any arbitrary function $E(t)$[48] can be represented as a sum of sine and cosine functions:

$$E(t) = \int_{-\infty}^{\infty} \tilde{E}(\omega) e^{i\omega t}\, d\omega, \tag{4.64}$$

where the complex amplitudes $\tilde{E}(\omega)$ are obtained from the **Fourier transform** of the original function:

$$\tilde{E}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} E(t) e^{-i\omega t}\, dt. \tag{4.65}$$

Essentially, the Fourier transform decomposes the original time-domain signal into different frequency components, with $\tilde{E}(\omega)$ describing the relative amplitudes of the different monochromatic waves that make up the original signal. That original signal can then be written as a superposition of monochromatic waves with amplitudes $\tilde{E}(\omega)$. This is precisely what is done in Eq. (4.64), which is typically referred to as the **inverse Fourier transform**.

Equations (4.64) and (4.65) should make immediate sense in terms of constructing temporal profiles of real light beams with a finite bandwidth from purely monochromatic plane waves. Indeed, (4.64) shows that an electric field with arbitrary temporal profile can be written as the superposition of monochromatic (plane) waves with different frequencies. Arbitrary temporal profiles, on the other, can have arbitrary spectral widths and profiles; these profiles are fully described by the complex amplitudes $\tilde{E}(\omega)$, given by the Fourier transform Eq. (4.65). **Significantly, the spectrum (i.e., frequency content) of any light source is simply governed by the Fourier transform Eq.** (4.65). If you know the time-domain wave profile, you can figure out the spectrum by calculating the Fourier transform (and vice versa). In practical terms, the spectrum that can be easily measured (using e.g. spectrometers or spectrum analysers) corresponds to the absolute value squared of the Fourier transform, i.e.,

$$S(\omega) = |\tilde{E}(\omega)|^2. \tag{4.66}$$

4.5.2 Spatial Fourier transforms

Fourier transforms are most commonly used with functions of time t as in Eqs. (4.64) and (4.65). However, there is nothing that prevents us from doing a Fourier transform with functions of spatial coordinates. In this case, we can write a signal, e.g. $E(x)$, that exhibits some *spatial* dependence as a superposition of plane waves

[48] Here we use the symbol resembling the electric field for familiarity, but the arguments hold any function.

associated with different spatial frequencies k_x. Furthermore, when our function depends on several spatial coordinates, e.g. $E(x, y, z)$, we can generalize the superposition to multiple dimensions. Long story short, a **monochromatic** light wave $E(x, y, z, t) = A(x, y, z)\exp(i\omega t)$ with an arbitrary spatial distribution $A(x, y, z)$ (no longer a simple plane wave) can be written as

$$E = \iiint\limits_{-\infty}^{\infty} \tilde{A}(k_x, k_y, k_z) e^{i\omega t} e^{-ik_x x} e^{-ik_y y} e^{-ik_z z} dk_x\, dk_y\, dk_z, \tag{4.67}$$

$$= \iiint\limits_{-\infty}^{\infty} E_0 e^{i\omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}}} dk_x\, dk_y\, dk_z, \tag{4.68}$$

where the second equation highlights that the whole construction is nothing but superpositions of plane waves with different wave vectors $\vec{\mathbf{k}}$ and amplitudes $E_0 = \tilde{A}(k_x, k_y, k_z)$. This picture should give an intuitive explanation as to why real beams diffract: they are made out of plane waves that propagate in different directions determined by their different wave vectors $\vec{\mathbf{k}}$! Furthermore, a (spatially) narrower beam has a broader spatial spectrum, i.e., its composition will include plane waves traveling at large angles. A narrower beam will therefore broaden faster than a broad beam [see Eq. (2.1)]!

4.6 So where do EM waves come from?

In the preceding discussion, we have covered the basic characteristics and properties of (plane) EM waves. But where do these EM waves come from? As with all electromagnetic phenomena, the answer is (of course) that they arise from electric charges (somehow). In fact, it turns out that all EM waves can be understood to originate from *accelerating* electric charges (or equivalently, *changing* electric currents).

EM waves correspond to a form of *radiation*: energy flows irreversibly away from the source and out "to infinity". For energy to "reach infinity", the total power passing through a spherical surface centred at a radiation source must stay constant as the radius r of the sphere (distance from the source) increases. If the power would increase with r, energy would not be conserved; if the power would decrease with r, energy would not be able to reach infinity, as the total radiated power $P_{\text{rad}} = \lim_{r\to\infty} P(r)$ would approach zero.

The total EM power passing through a surface is given by the surface integral over the Poynting vector:

$$P(r) = \oint \vec{\mathbf{S}} \cdot d\vec{\mathbf{a}} = \frac{1}{\mu_0} \oint \vec{\mathbf{E}} \times \vec{\mathbf{B}} \cdot d\vec{\mathbf{a}}. \tag{4.69}$$

Because the area of a sphere increases as $4\pi r^2$, we see that for radiation to occur ($P_{\text{rad}} \neq 0$), the Poynting vector must decrease (at large r) no faster than $1/r^2$. We know, however, from Coulomb's and Biot-Savart's laws that *electrostatic* and *magnetostatic* fields fall off like $1/r^2$, implying $S \sim 1/r^4$ for static configurations. Thus, we see that neither static charges nor currents radiate EM waves.

Detailed analysis of Maxwell's equations shows that, when charges are accelerating, terms that decay as $1/r$ appear in the formulae describing the electric and magnetic fields. It is these terms that are responsible for the generation of EM waves. In fact, for very large r, all other terms can be neglected, since they fall off as $1/r^2$ or faster. Although a detailed analysis is beyond the scope of the present course (the keen reader is recommended to open "Introduction to Electrodynamics" by Griffiths), we present and discuss below the equations describing EM radiation from a time-varying electric dipole, which arguably represents the most important form of "elementary" EM radiation.

4.6.1 Dipole radiation

Let us consider a charge distribution characterised by a time-varying electric dipole moment $\vec{\mathbf{p}}(\tau)$ located at the origin.[49] At a position $\vec{\mathbf{r}}$ that is far away from the charge (such that all terms decaying faster than $1/r$ can be neglected), the electric and magnetic fields can be (approximately) written as [see Griffiths]:

$$\boxed{\begin{aligned}\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) &= \frac{\mu_0}{4\pi r}\left[(\hat{\mathbf{r}} \cdot \vec{\mathbf{p}}_{\tau\tau})\hat{\mathbf{r}} - \vec{\mathbf{p}}_{\tau\tau}\right], \qquad (4.70)\\ \vec{\mathbf{B}}(\vec{\mathbf{r}}, t) &= -\frac{\mu_0}{4\pi r c}\left[\hat{\mathbf{r}} \times \vec{\mathbf{p}}_{\tau\tau}\right], \qquad (4.71)\end{aligned}}$$

where $r = \|\vec{\mathbf{r}}\| = \sqrt{x^2 + y^2 + z^2}$ is the distance from the source, $\hat{\mathbf{r}}$ is a unit vector along $\vec{\mathbf{r}}$, and $\vec{\mathbf{p}}_{\tau\tau}$ is the second derivative of the dipole moment, evaluated at a *retarded* time $\tau = t - r/c$:

$$\vec{\mathbf{p}}_{\tau\tau} = \left.\frac{d^2\vec{\mathbf{p}}(\tau)}{d\tau^2}\right|_{\tau=t-r/c}. \qquad (4.72)$$

The expressions can be significantly simplified by considering a dipole moment $\vec{\mathbf{p}}(\tau) = [0, 0, p(\tau)]^{\mathrm{T}}$ pointing along the z-axis and expressing the fields in spherical coordinates. We find:

$$\boxed{\begin{aligned}\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) &= \frac{\mu_0 p_{\tau\tau}}{4\pi}\left(\frac{\sin\theta}{r}\right)\hat{\boldsymbol{\theta}}, \qquad (4.73)\\ \vec{\mathbf{B}}(\vec{\mathbf{r}}, t) &= \frac{\mu_0 p_{\tau\tau}}{4\pi c}\left(\frac{\sin\theta}{r}\right)\hat{\boldsymbol{\phi}}. \qquad (4.74)\end{aligned}}$$

In these coordinates, the Poynting vector reads

$$\boxed{\vec{\mathbf{S}} \approx \frac{1}{\mu_0}(\vec{\mathbf{E}} \times \vec{\mathbf{B}}) = \frac{\mu_0 p_{\tau\tau}^2}{16\pi^2 c}\left(\frac{\sin^2\theta}{r^2}\right)\hat{\mathbf{r}}.} \qquad (4.75)$$

It is clear that, in the far-field radiation zone, the electric and magnetic fields are non-zero only when the second-derivative of the electric dipole moment is non-zero. On the other hand, this situation requires accelerating charges. For example, for a single moving point charge (with displacement $\vec{\mathbf{d}}(\tau)$ relative to the origin), we have $\vec{\mathbf{p}} = q\vec{\mathbf{d}}(\tau)$, and thus $\vec{\mathbf{p}}_{\tau\tau} = q\vec{\mathbf{a}}(t - r/c)$, where $\vec{\mathbf{a}}$ is the acceleration of the charge. Moreover, from the expressions in spherical coordinates, it is clear that the electric and magnetic fields are (i) mutually perpendicular and (ii) transverse to the direction of propagation (Poynting vector). This is quite satisfying.

It is interesting to emphasise that $\vec{\mathbf{p}}_{\tau\tau}$ is evaluated at the "retarded" time $\tau = t - r/c$. This is simply because electromagnetic news travel at the speed of light. Specifically, when evaluating the fields at time t and displacement $\vec{\mathbf{r}}$ that is a distance r away from the source, the state of the source "right now" does not matter, as

[49] For a physical dipole consisting of two electric charges $\pm q$ separated by a displacement $\vec{\mathbf{d}}$, we have $\vec{\mathbf{p}} = q\vec{\mathbf{d}}$. For a single point charge, we have $\vec{\mathbf{p}} = q\vec{\mathbf{d}}$, where $\vec{\mathbf{d}}$ is the displacement from the origin.

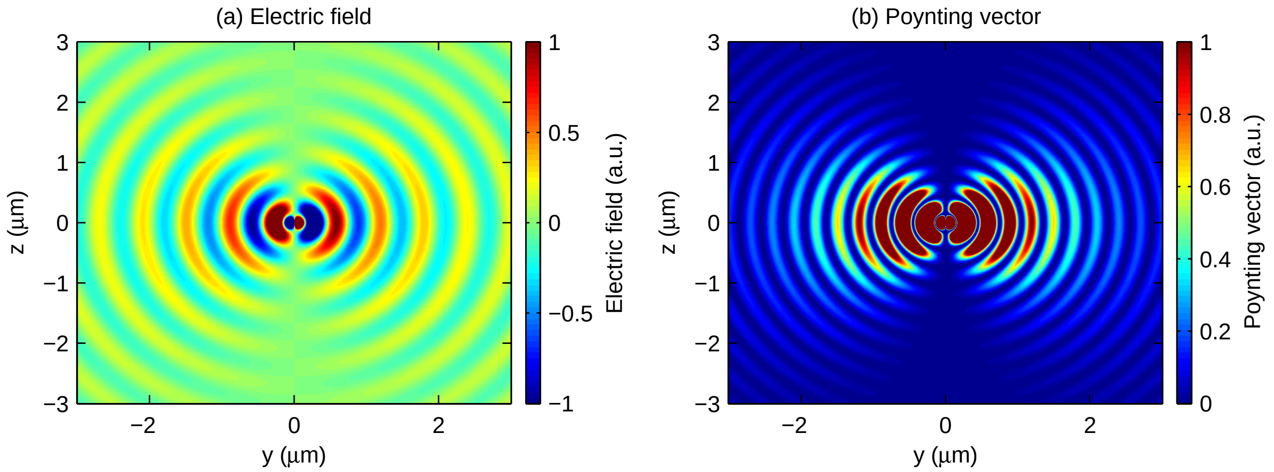

Figure 21: Snapshot of the (a) electric field and (b) Poynting vector of an EM wave with wavelength $\lambda = 600 \text{ nm}$ emitted by a harmonically oscillating dipole. The dipole is oscillating along the z-axis, and the fields are calculated in the yz-plane.

it will take a time r/c for the electromagnetic news to travel from the source to the position of interest. Rather, it is the source's state at an earlier time $t - r/c$ that contributes to the fields at time t.

Arguably the simplest source of EM waves is that of an oscillating electric dipole. In this arrangement, two opposite charges are separated by a distance, and either their values or their separation exhibits harmonic oscillation. The dipole moment $\vec{\mathbf{p}}(\tau) = \vec{\mathbf{p}}_0 \cos(\omega t)$. Figure 21 shows a typical radiation pattern emitted by such an oscillating dipole, calculated from Eqs. (4.70) and (4.71). It is noteworthy that the frequency of the emitted radiation is equal to the oscillation frequency of the dipole.

In practice, it is straightforward to make charges accelerate: one can simply connect conducting rods to a source of alternating voltage. This will give rise to an alternating current in the conductors, which amounts to accelerating charges and hence emission of EM radiation (with frequency identical to that of the alternating voltage source). In fact, this is how antennas work[50]. This simple scheme underpins the generation of EM radio- and microwaves, whose frequencies range from kilohertz to gigahertz. It cannot, however, be used to create EM waves in the infrared or visible spectral regions, where the EM wave frequencies are measured in (hundreds of) terahertz. This is simply because it is not possible to create alternating voltages (or currents) that would oscillate at such high frequencies: electronic circuits are far too slow. Rather, EM waves in the infrared and visible spectral regions arise from transitions between different atomic or molecular energy levels, and they can be coherently generated only by leveraging those transitions (as is done with lasers).

[50] The simplest and most widely used class of antenna is the so-called "dipole antenna" whose radiation pattern is very similar to that of an ideal electric dipole.

Problems

4.1 During the lecture, we derived the wave equation for the electric field $\vec{\mathbf{E}}$.

(a) Show that an identical wave equation can be derived for the magnetic field $\vec{\mathbf{B}}$.

(b) Explicitly write the wave equation for each of the different cartesian components of the magnetic field vector.

(c) When studying linearly polarized electromagnetic waves, it is sufficient to only consider a single cartesian component of the electric and magnetic fields. Assuming an EM wave polarized along the x-direction and propagating along the z-direction, write down a *scalar* wave equation for the magnetic field.

4.2 Show that a complex refractive index gives rise to attenuation or amplification of an EM wave.

4.3 Consider an EM plane wave polarized along the z-direction and propagating in the $+x$-direction in a material whose refractive index $n = 1.4$. The wave's angular frequency is $\omega = 2\pi \cdot 200$ THz and the amplitude of the electric field is

$$\vec{\mathbf{E}}_0 = 10\ \mathrm{Vm}^{-1}\hat{\mathbf{z}}.$$

Write the full plane EM wave solution, defining the numerical values of all parameters.

4.4 The electric field of a monochromatic (not necessarily plane) EM wave can be written as $\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = \vec{\mathbf{A}}(\vec{\mathbf{r}})e^{i\omega t}$.

(a) Show that the vector amplitude $\vec{\mathbf{A}}$ obeys the Helmholtz equation

$$(\nabla^2 + k^2)\vec{\mathbf{A}} = 0.$$

(b) Explicitly write down the Helmholtz equation for each of the cartesian components of the EM field.

(c) Assuming the EM wave is polarized along the x-direction, write the Helmholtz equation in scalar form.

4.5 Consider a plane wave in free space defined by

$$\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = \vec{\mathbf{E}}_0 e^{i(\omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}})}.$$

(a) Write down explicit expressions for the x, y, and z components of the electric field.

(b) Show that the electric field satisfies $\nabla \cdot \vec{\mathbf{E}} = 0$ if $\vec{\mathbf{E}}_0 \cdot \vec{\mathbf{k}} = 0$.

(c) Find an associated magnetic field such that Faraday's law is satisfied.

(d) Find a dispersion relationship such that Ampere's law is satisfied.

4.6 Consider the vector function given by

$$\vec{\mathbf{F}}(\vec{\mathbf{r}}, t) = \vec{\mathbf{F}}_0 e^{i(\omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}})}.$$

(a) What is the relationship between ω and $\vec{\mathbf{k}}$ that would allow this to be a plane wave solution to Maxwell's equations for an electric field in a dielectric medium with refractive index n?

(b) If the vector field above is to represent a plane EM wave, what should the value of the dot product $\vec{\mathbf{F}} \cdot \vec{\mathbf{k}}$ be?

(c) Suppose that $\vec{\mathbf{F}}$ given above and a vector field $\vec{\mathbf{G}}$ represent the electric and magnetic fields for a plane wave in a vacuum. What is the value of the dot product $\vec{\mathbf{F}} \cdot \vec{\mathbf{G}}$?

(d) Are plane waves physical or not? Provide a reason why?

4.7 Consider a plane EM wave with frequency ω that is propagating along the $+x$ direction and that is polarized along the z-direction.

(a) Show that the electric field can be written (in complex notation) as

$$\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = E_0 e^{i(\omega t - kx)} \hat{\mathbf{z}}.$$

(b) Find a corresponding magnetic field such that the resulting pair satisfies all of Maxwell's equations.

4.8 Consider a plane electromagnetic wave $\vec{\mathbf{E}}_0 e^{i(\omega t - \vec{\mathbf{k}} \cdot \vec{\mathbf{r}})}$. Prove the following differential operations that were used to derive the properties of plane EM waves:

$$\begin{aligned} \frac{\partial \vec{\mathbf{E}}}{\partial t} &= i\omega \vec{\mathbf{E}} \\ \nabla \cdot \vec{\mathbf{E}} &= -i\vec{\mathbf{k}} \cdot \vec{\mathbf{E}} \\ \nabla \times \vec{\mathbf{E}} &= -i\vec{\mathbf{k}} \times \vec{\mathbf{E}} \\ \nabla^2 \vec{\mathbf{E}} &= -k^2 \vec{\mathbf{E}}. \end{aligned}$$

4.9 Consider a plane EM wave propagating in the $+z$ direction and that is polarized along the x-axis:

$$\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = E_0 e^{i(\omega t - kz)} \hat{\mathbf{x}}.$$

(a) For the electric field above, find a suitable magnetic field so that the resulting pair satisfies all of Maxwell's equations.

(b) Write down an equation for the electric and magnetic fields for a wave travelling in the $-z$ direction, but with a different polarization and arbitrary phase to the one above.

4.10 In the lecture notes, we have derived the wave equation for a dielectric material, i.e., assuming that the electrical conductivity $\sigma = 0$.

(a) Derive the electromagnetic wave equation in the general case when $\sigma \neq 0$.

(b) Derive the dispersion relation for this more general case.

(c) Show that electromagnetic waves experience strong losses in conductors, i.e., in materials for which $\sigma \neq 0$. In particular, derive an expression for the imaginary part of the wavenumber.

(d) The distance over which the field amplitude impinging on a conductor decreases by a factor of e is known as the skin-depth. This represents a typical "penetration" depth for the EM wave, i.e., the depth to which the EM wave can survive without being totally dissipated. Derive an expression for the skin-depth of a good conductor, i.e., one for which $\sigma \gg \epsilon\omega$.

(e) Calculate the skin-depth (in nanometers) of a good conductor like copper ($\sigma \approx 10^7\ \Omega^{-1}\mathrm{m}^{-1}$) in the visible range ($\omega \approx 10^{15}\ \mathrm{s}^{-1}$).

4.11 Because electrons do not move instantaneously in response to an applied electric field, the simple constitutive relation

$$\vec{\mathbf{P}}(\vec{\mathbf{r}}, t) = \varepsilon_0 \chi \vec{\mathbf{E}}(\vec{\mathbf{r}}, t)$$

is not correct in general. Indeed, as argued in the lecture notes, this simple relationship cannot model dispersion, i.e., the fact that refractive index depends on frequency. In this exercise, you will consider the more general constitutive relation

$$\vec{\mathbf{P}}(\vec{\mathbf{r}}, t) = \varepsilon_0 \chi(\tau) * \vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = \varepsilon_0 \int_{-\infty}^{t} \chi(t - t')\vec{\mathbf{E}}(\vec{\mathbf{r}}, t')dt', \tag{4.76}$$

where $*$ denotes convolution.

(a) Show that for a monochromatic wave $\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = \vec{\mathbf{E}}_0(\vec{\mathbf{r}})e^{i\omega t}$, the polarization can be written as

$$\vec{\mathbf{P}}(\vec{\mathbf{r}}, t) = \varepsilon_0 \tilde{\chi}(\omega)\vec{\mathbf{E}}(\vec{\mathbf{r}}, t),$$

where $\tilde{\chi}(\omega)$ is the Fourier transform of $\chi(t)$. **Hint:** the response must be causal, such that $\chi(t - t') = 0$ for $t' > t$.

(b) Derive the EM wave equation for a monochromatic wave using the general constitutive relationship in Eq. (4.76).

(c) Show that the refractive index now depends on frequency.

4.12 Show that, at a fixed spatial position $\vec{\mathbf{r}}_0$, the electric field of a monochromatic EM wave polarised along the $\hat{\mathbf{x}}$ direction can be written as

$$\vec{\mathbf{E}}(\vec{\mathbf{r}}_0, t) = E_0 e^{i\omega t}\hat{\mathbf{x}}.$$

4.13 (a) Use Eqs. (4.73) and (4.74) to derive the Poynting vector of the EM wave generated by a time-varying electric dipole [Eq. (4.75)].

(b) Use the Poynting vector to derive an expression for the total power radiated by the dipole.

(c) Derive the total power radiated by an oscillating electric dipole $p(t) = p_0 \cos(\omega t)\hat{\mathbf{z}}$.

(d) For a point charge, the electric dipole moment $\vec{\mathbf{p}} = q\vec{\mathbf{d}}(t)$, where $\vec{\mathbf{d}}(t)$ is the (time-dependent) displacement of the charge. Derive an expression for the total power radiated by an accelerating point charge. The result is the famous **Larmor formula**.

5 Classical electron oscillator

In this Section, we will examine light-matter interactions from the classical perspective, where light is interpreted as an EM wave. The key parameter that describes the interaction is the refractive index: it fixes the wavelength of an EM wave in a medium, the wave's phase velocity, and even gives rise to attenuation or amplification if complex. But physically, where does refractive index come from? Well, we have already seen that $n = \sqrt{1+\chi}$, where χ is the susceptibility that determines the relationship between the material polarisation $\vec{\mathbf{P}}$ and the electric field, viz. $\vec{\mathbf{P}} = \varepsilon_0\chi\vec{\mathbf{E}}$. So what we really need to explain is: how does the electric polarisation $\vec{\mathbf{P}}$ arise? If we can physically model how a given electric field $\vec{\mathbf{E}}$ induces a polarisation $\vec{\mathbf{P}}$, we should be able to extract χ and subsequently the refractive index n. This is the aim of this Section.

To simplify the analysis, we will assume the electric field to be linearly polarised along the x-axis: $\vec{\mathbf{E}} = E\hat{\mathbf{x}} \to \vec{\mathbf{P}} = P\hat{\mathbf{x}}$. It is then sufficient to only consider the scalar amplitudes E and P.

5.1 Origins of polarisation

Polarisation P is the average electric dipole moment per unit volume in the medium:

$$P = \frac{\Delta p}{\Delta V}. \tag{5.1}$$

At a microscopic level, an electric field E applied to a dielectric material displaces the electron charge clouds – bound to individual atomic nuclei – from their equilibrium positions [see Fig. 22]. On the other hand, the displacement of an electron out of its equilibrium position induces an electric dipole moment

$$p_e(t) = -ex(t), \tag{5.2}$$

where e is the elementary charge and $x(t)$ is the displacement of the electron[51]. Assuming we have N displaced electrons per unit volume, then the total polarisation

$$P(t) = Np_e(t) = -Nex(t). \tag{5.3}$$

Thus, **to model the relationship between polarisation and the applied electric field, we must model how an electron is displaced by a given applied (EM) field.**

5.2 Electron on a spring

We model the motion of an electron very classically, using Newton's second law:

$$m\frac{dx^2(t)}{dt^2} = F_{\text{net}} = \sum F. \tag{5.4}$$

Below we list the different forces acting on the electron.

[51] Note that for EM waves the electric field is continuously oscillating in time, and hence the displacement will similarly depend on time.

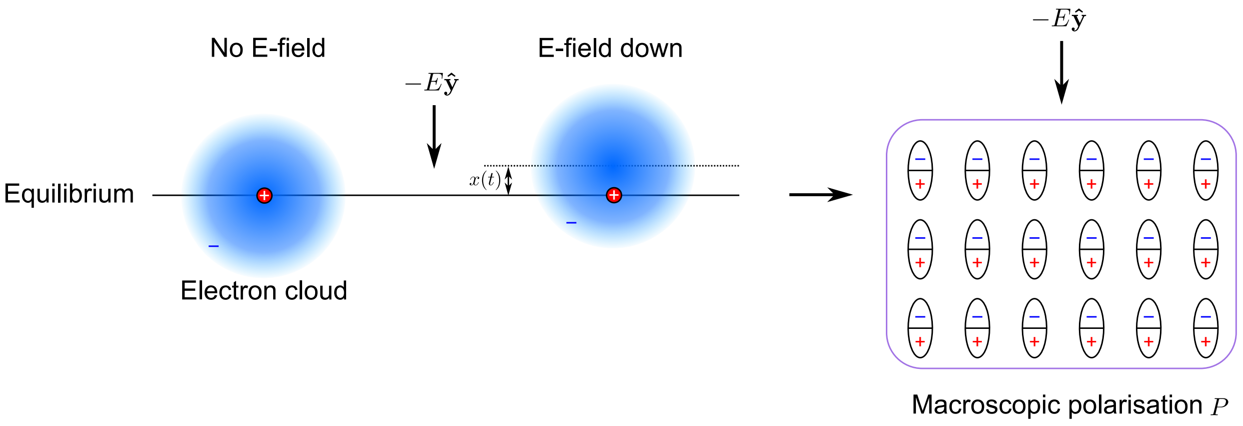

Figure 22: Schematic illustration of the microscopic origins of material polarization. An electric field displaces the elctron cloud of an atom, giving rise to an atomic dipole moment. A bunch of such dipole moments then give rise to a macroscopic polarization.

1. First, the electron is very comfortable at its equilibrium position; it does not want to move away. We therefore assume that there is a restoring force that pulls the electron back to its equilibrium position. We take this restoring force to be directly proportional to the displacement:

$$F_{\text{restoring}} = -m\omega_0^2 x \tag{5.5}$$

   Here the frequency ω_0 should be understood as the transition frequency between the atom's two energy levels.

2. Second, as for all charges in an electric field, there is a simple driving force due to the applied electric field:

$$F_{\text{driving}} = -eE(t). \tag{5.6}$$

3. Third, because the applied EM wave oscillates in time, so does the electron. But all oscillating charges emit electromagnetic radiation, and therefore there has to be some loss of energy. This is accounted for by a damping term analogous to friction/drag:

$$F_{\text{damping}} = -\gamma m \frac{dx}{dt}, \tag{5.7}$$

   where γ is the damping coefficient.

With these forces, the equation of motion for the electron becomes

$$\boxed{m\frac{dx^2(t)}{dt^2} = -m\omega_0^2 x - \gamma m \frac{dx}{dt} - eE(t).} \tag{5.8}$$

This equation is analogous to the equation of motion for a *driven damped simple harmonic oscillator*. In other words, we are modelling the electron as if it is attached to a spring and driven by an oscillating electric field. One might think that such a spring-electron model is an extremely naive approach to the analysis of light-matter interactions. However, it turns out that the model is extremely successful, faithfully reproducing the results of considerably more complex quantum mechanical calculations. To quote Feynman:

> *You may think that this is a funny model of an atom if you have heard about electrons whirling around orbits. But that is just an oversimplified picture. The correct picture of an atom, which is given by [quantum mechanics] says that, so far as problems involving light are concerned, the electrons behave as though they were held by springs.*

5.3 No driving field

Let us first consider the case where there is no applied electric field, i.e., $E = 0$. The general solution for our driven damped harmonic oscillator is:

$$x(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}, \tag{5.9}$$

where $c_{1,2}$ are constants that depend on the initial conditions while $\lambda_{1,2}$ solve the characteristic equation

$$m\lambda^2 + \gamma m\lambda + m\omega_0^2 = 0. \tag{5.10}$$

Assuming the damping to be light, such that $\gamma \ll \omega_0$, one obtains after some algebra:

$$\lambda_{1,2} = -\frac{\gamma}{2} \pm i\omega_0. \tag{5.11}$$

The general solution is thus

$$x(t) = e^{-\gamma t/2}\left(c_1 e^{i\omega_0 t} + c_2 e^{-i\omega_0 t}\right). \tag{5.12}$$

Since the displacement has to be real, we choose the constants as $c_1 = \frac{1}{2}A\exp{(i\delta)} = c_2^*$, yielding

$$\boxed{x(t) = Ae^{-\gamma t/2}\cos{(\omega_0 t + \delta)}.} \tag{5.13}$$

The equation above shows that, in the absence of an applied electric field, an initially displaced electron ($A \neq 0$) will undergo exponentially damped oscillations. This gives rise to an electric dipole moment $p(t) = -ex(t)$ that is similarly exhibiting damped oscillations. On the other hand, as described in Section 4.6.1, oscillating dipole moments emit EM waves. Far away from the dipole, the amplitude of the EM wave will be proportional to $d^2p(t)/dt^2$ [see Section 4.6.1]. Thus, for $t \geq 0$ the EM wave will have the form

$$\boxed{E(t) \propto Ae^{-\gamma t/2}\cos{(\omega_0 t + \delta)},} \tag{5.14}$$

where we again used $\gamma \ll \omega_0$.

We have just learnt that (i) an initially displaced atom will undergo exponentially decaying oscillations with frequency ω_0, while simultaneously (ii) emitting EM waves with frequency ω_0 that similarly undergo exponential decay. **This is the classical description of spontaneous emission**: an excited atom decays back to its equilibrium displacement while emitting EM waves.

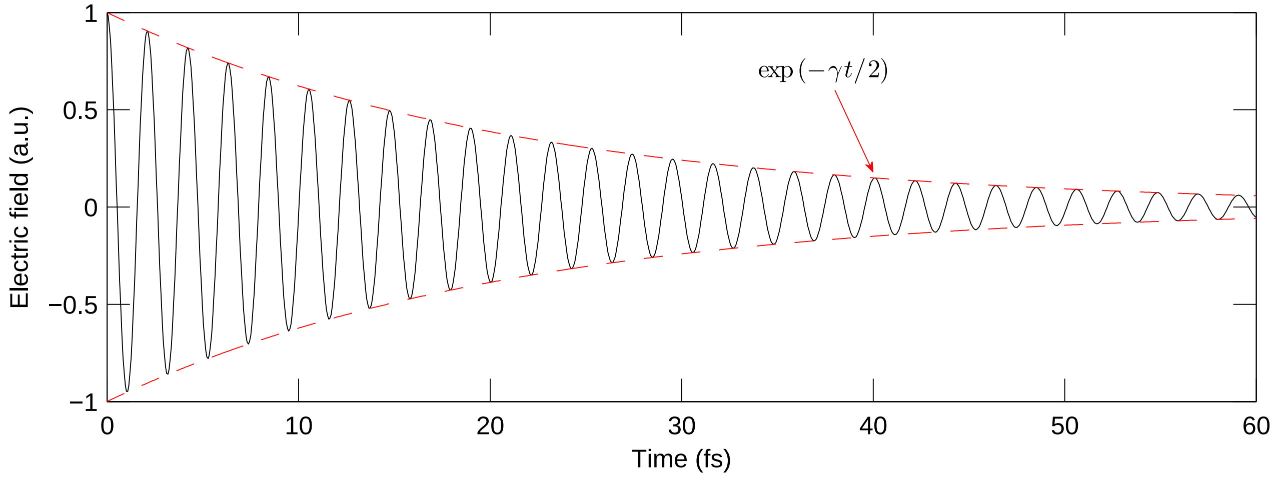

Figure 23: Electric field corresponding to the far-field EM wave radiated by an initially excited atom.

5.3.1 Spectrum of spontaneous emission

Figure 23 shows the EM wave described by Eq. (5.14). Because of the exponential decay, this wave does not correspond to a monochromatic wave[52]. Rather, it corresponds to a superposition of a range of monochromatic waves with different frequencies. It is interesting to consider the *spectrum* of the wave, i.e., the relative magnitudes of the different frequency components that superpose to create the wave. The spectrum can be obtained by taking the Fourier transform of $E(t)$:

$$\begin{aligned}\tilde{E}(\omega) &= \frac{1}{2\pi}\int_0^\infty E(t)e^{-i\omega t}\,dt \qquad (5.15)\\ &= \frac{1}{2\pi}\int_0^\infty E_0 e^{-\gamma/2t}\cos{(\omega_0 t+\delta)}e^{-i\omega t}\,dt. \qquad (5.16)\end{aligned}$$

The integrals can be calculated without too much tears. Doing this, and using $\gamma \ll \omega_0$, one finds

$$\tilde{E}(\omega) = \frac{E_0 e^{i\delta}}{4\pi i}\frac{1}{(\omega-\omega_0)-i\gamma/2}. \qquad (5.17)$$

The corresponding spectral intensity is

$$\begin{aligned}S(\omega) &\propto |\tilde{E}(\omega)|^2 \qquad (5.18)\\ &= S_0 g_{\mathrm{H}}(\omega-\omega_0), \qquad (5.19)\end{aligned}$$

where S_0 is a constant amplitude and

$$\boxed{g_{\mathrm{H}}(\omega-\omega_0) = \frac{1}{\pi}\frac{\gamma/2}{(\omega-\omega_0)^2+(\gamma/2)^2}.} \qquad (5.20)$$

The shape of g_{H} is Lorentzian, centred at ω_0 with a full-width at half maximum given by γ.

[52] A monochromatic wave would just oscillate away forever; only a pure (co)sinusoidal oscillation is monochromatic.

The spring-electron model predicts that EM waves emitted via spontaneous emission follow a Lorentzian frequency spectrum – exactly as confirmed by experiments. **Spring-electron 1 - Everyone else 0**.

5.4 Yes driving field

Let us now consider the response of a spring-atom to a monochromatic EM wave, $E(t) = E_0 e^{i\omega t}$. Note that we omit the $\vec{\mathbf{k}} \cdot \vec{\mathbf{r}}$ phase term since the atom is not moving: we can set its fixed position as $\vec{\mathbf{r}} = 0$.[53] Equation (5.8) becomes:

$$m\frac{dx^2(t)}{dt^2} + \gamma m\frac{dx(t)}{dt} + m\omega_0^2 x(t) = -eE_0 e^{i\omega t}. \tag{5.21}$$

We anticipate a harmonic solution, i.e., that the electron oscillates at the same frequency as the driving:

$$x(t) = X(\omega)e^{i\omega t}, \tag{5.22}$$

where $X(\omega)$ is a complex amplitude that can, in general, depend on frequency. Injecting this ansatz into Eq. (5.21), evaluating the derivatives, and removing the common exponential terms, we derive

$$-m\omega^2 X(\omega) + i\gamma\omega m X(\omega) + m\omega_0^2 X(\omega) = -eE_0. \tag{5.23}$$

The solution is easy to find:

$$\boxed{X(\omega) = \frac{e}{m}\frac{E_0}{(\omega^2 - \omega_0^2) - i\gamma\omega}.} \tag{5.24}$$

And so we see that the amplitude of the electron oscillations depends on the frequency ω of the driving field. The amplitude has the largest magnitude when $\omega = \omega_0$. In this case, the frequency of the driving laser matches the atom's *resonance frequency*. However, nonzero oscillation amplitudes can be achieved even if the driving frequency is detuned from the resonance frequency. **This is the classical description of the fact that atomic transitions have finite linewidths.**

5.5 Susceptibility

With the electron oscillation amplitude evaluated, we can calculate the polarisation using Eq. (5.3):

$$P(t) = -Nex(t) \tag{5.25}$$

$$= -NeX(\omega)e^{i\omega t} \tag{5.26}$$

$$= -\frac{Ne^2}{m}\frac{1}{(\omega^2 - \omega_0^2) - i\gamma\omega}E_0 e^{i\omega t}. \tag{5.27}$$

Thus, we see that, just like everything else, also the polarisation will oscillate at the driving frequency ω. Furthermore, we clearly see that the polarization amplitude is linearly proportional to the incident electric field

[53] Also note that we will be using complex notation, and so only the real part of the solution describes the physical displacement.

$E(t) = E_0 e^{i\omega t}$. From our earlier relationship $P = \varepsilon_0 \chi(\omega) E$, we can identify the linear susceptibility that we set out to find:

$$\boxed{\chi(\omega) = -\frac{Ne^2}{m\varepsilon_0} \frac{1}{(\omega^2 - \omega_0^2) - i\gamma\omega}.} \tag{5.28}$$

It should be clear that the susceptibility depends on the frequency of the incident plane wave, thus faithfully predicting the effects of dispersion. Furthermore, the susceptibility is a complex quantity: $\chi(\omega) = \chi_{\mathrm{R}}(\omega) - i\chi_{\mathrm{I}}(\omega)$. Below we first discuss the physical significance of the real and imaginary parts, and then consider the values of the susceptibility in two different limits: far from a transition frequency and close to a transition frequency.

5.5.1 Physical interpretation

We have seen that a plane EM wave in a medium has the form[54]

$$E(z,t) = E_0 e^{i(\omega t - kz)}, \tag{5.29}$$

$$E(z,t) = E_0 e^{-k_{\mathrm{I}} z} e^{i(\omega t - k_{\mathrm{R}} z)}, \tag{5.30}$$

where the wave number satisfies the dispersion relation

$$k(\omega) = \frac{\omega}{c}\sqrt{1 + \chi(\omega)}. \tag{5.31}$$

Typically we are interested in frequencies far from an atomic resonance; the only exception comes when we deal with lasers, but in this case the number of active atoms/ions is typically small, $N \sim 10^{12}\ \mathrm{m}^{-3}$. In either case, we have $\chi_{\mathrm{R,I}} \ll 1$, allowing us to approximate $\sqrt{1+\chi} \approx 1 + \chi/2$, yielding

$$k(\omega) = k_{\mathrm{R}}(\omega) - ik_{\mathrm{I}}(\omega) \approx \frac{\omega}{c}\left(1 + \frac{\chi_{\mathrm{R}}(\omega)}{2}\right) - i\frac{\omega}{c}\frac{\chi_{\mathrm{I}}(\omega)}{2}. \tag{5.32}$$

To summarise:

$$k_{\mathrm{R}}(\omega) = \frac{\omega}{c}\left(1 + \frac{\chi_{\mathrm{R}}(\omega)}{2}\right), \tag{5.33}$$

$$k_{\mathrm{I}}(\omega) = \frac{\omega}{2c}\chi_{\mathrm{I}}(\omega). \tag{5.34}$$

We can now assign physical significance to the real and imaginary parts of the susceptibility:

- By defining k_{R}, the real part of $\chi(\omega)$ sets the wavelength and phase velocity of an EM wave propagating in the medium. In other words, $\chi_{\mathrm{R}}(\omega)$ defines the usual (real) refractive index of the medium: $n(\omega) \approx 1 + \frac{\chi_{\mathrm{R}}(\omega)}{2}$.
- By defining k_{I}, the imaginary part of $\chi(\omega)$ is responsible for attenuation/amplification of the EM wave.

[54] For simplicity, we assume linear polarisation and propagation along z.

5.5.2 Off-resonance susceptibility

Far from the resonance frequency ($|\omega - \omega_0| \gg \gamma$), we can approximate the susceptibility as[55]

$$\boxed{\chi(\omega) = -\frac{Ne^2}{m\varepsilon_0}\frac{1}{\omega^2 - \omega_0^2}.} \tag{5.35}$$

The susceptibility is completely real:

If the frequency of an EM wave is far from a material resonance frequency, it will experience neither amplification or attenuation. The material is transparent.

The refractive index will, however, depend on frequency, with a profile given by

$$n^2(\omega) = 1 + \chi(\omega) = 1 - \frac{Ne^2}{m\varepsilon_0}\frac{1}{\omega^2 - \omega_0^2}. \tag{5.36}$$

An example of such a curve is shown in Fig. 24(a). Real materials have strong resonances at ultraviolet frequencies, and so the refractive indices around optical (e.g. visible and infrared) frequencies typically follow curves very similar to that shown in Fig. 24(a), i.e., they monotonically decrease (increase) with wavelength (frequency), diverging at frequencies close to resonance. This is illustrated in Fig. 24(b), where we plot $n(\lambda)$ for several different materials.

Of course, real materials have resonances at many different frequencies, each corresponding to a transition between a different pair of energy levels. In this case, the full susceptibility can be approximated as a superposition of susceptibilities arising from different transitions. Significantly, for frequencies far from the resonance frequencies, the refractive indices of real materials can be well approximated by a **Sellmeier equation**[56]:

$$\boxed{n^2(\omega) = 1 - \sum_j \frac{a_j}{\omega^2 - \omega_j^2},} \tag{5.37}$$

where ω_j is an atomic resonance and a_j is a constant that describes the strength of the transition.[57] By comparing with Eq. (5.36), it should be evident that each term of the Sellmeier equation represents an atomic transition. The Sellmeier equation is of key importance in optical system design. At this point we highlight one key conclusion that derives from our analysis so far:

Resonances between atomic energy levels underlie the fact that refractive indices depend on frequency. In other words, chromatic dispersion originates from the fact that electrons can be excited from one discrete energy level to another.

[55] $\omega^2 - \omega_0^2 - i\gamma\omega = \omega[(\omega - \omega_0)(\omega + \omega_0)/\omega - i\gamma]$. Since $(\omega + \omega_0)/\omega > 1$ and $|\omega - \omega_0| \gg \gamma$, the γ term is negligible.

[56] More often the Sellmeier equation is given in terms of wavelength λ.

[57] a_j is similar to $Ne^2/(m\varepsilon_0)$, but also has a quantum mechanical correction term. Normally both a_j and ω_j are obtained empirically.

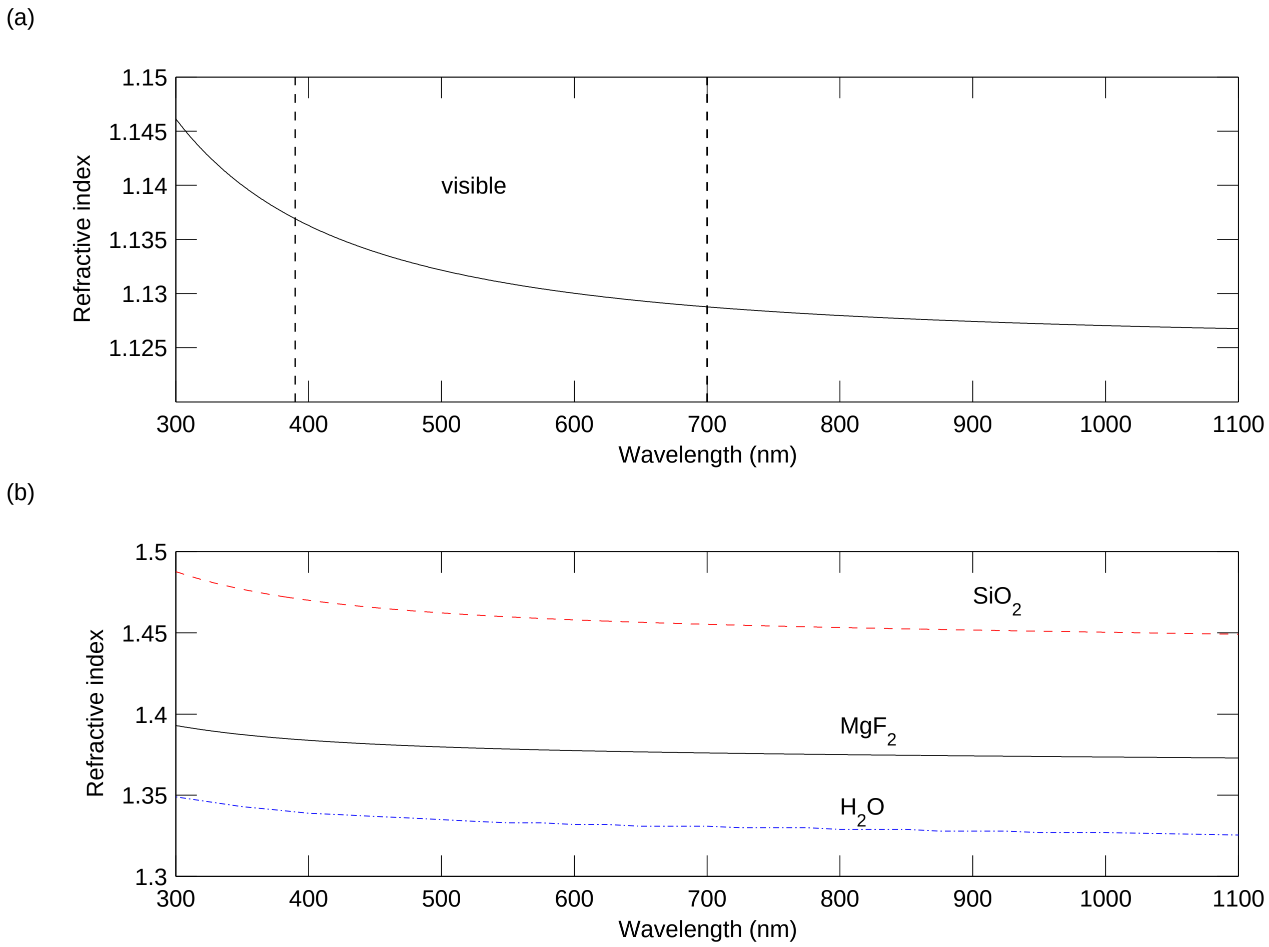

Figure 24: (a) Refractive index arising from a single Lorentzian resonance. Calculated for a resonance wavelength $\lambda_0 = 116\ \mathrm{nm}$ and macroscopic density $\rho = 2200\ \mathrm{kgm}^{-3}$. These values correspond to fused silica (SiO_2). (b) Refractive indices of various real materials.

Equation 5.36 predicts a curious phenomenon: for frequencies $\omega > \omega_0$, the refractive index will be less than unity. As resonances typically occur in the ultraviolet, this would suggest that the speed of an EM wave in a medium can exceed the speed of light in vacuum ($v = c/n$) for frequencies beyond the UV. It turns out that this is indeed possible, and X-rays for example can routinely travel faster than the speed of light. No violation of special relativity or causality is taking place, however, as $v = c/n$ is only the wave's *phase velocity*; information, it turns out, always travels slower than this, happily respecting relativity. Of course, it should be emphasised that real atoms have many resonances and the actual refractive index arises through the interplay of all the different resonances (as in the Sellmeier equation). Accordingly, one cannot conclude that the phase velocity always exceeds the speed of light for frequencies below a resonance (another resonance may lift the refractive index above unity).

5.5.3 On-resonance susceptibility

Close to a resonance frequency $\omega \approx \omega_0$, allowing us to approximate $(\omega^2 - \omega_0^2) = (\omega - \omega_0)(\omega + \omega_0) \approx 2\omega_0(\omega - \omega_0)$. With this approximation, the susceptibility becomes

$$\boxed{\chi(\omega) = -\frac{Ne^2}{2m\varepsilon_0\omega_0}\frac{1}{(\omega - \omega_0) - i\gamma/2}} \tag{5.38}$$

This time, the imaginary part is nonzero. Using the convention $\chi(\omega) = \chi_\mathrm{R}(\omega) - i\chi_\mathrm{I}(\omega)$, the real and imaginary parts read

$$\chi_\mathrm{R}(\omega) = -\frac{Ne^2}{2\varepsilon_0 m\omega_0}\frac{(\omega - \omega_0)}{(\omega - \omega_0)^2 + (\gamma/2)^2}, \tag{5.39}$$

$$\chi_\mathrm{I}(\omega) = \frac{Ne^2}{2\varepsilon_0 m\omega_0}\frac{\gamma/2}{(\omega - \omega_0)^2 + (\gamma/2)^2}, \tag{5.40}$$

with typical profiles depicted in Fig. 25. The first thing we note is that the imaginary part has the exact same Lorentzian profile as the one we derived for spontaneous emission. To gain more insight, we write the evolution of the intensity of our EM wave as

$$I(z, \omega) = |E(t, z)|^2 = |E_0|^2 e^{-2k_\mathrm{I}z} = |E_0|^2 e^{-\alpha(\omega)z}, \tag{5.41}$$

where we used Eq. (5.34) to define

$$\alpha(\omega) = \frac{\omega_0}{c}\chi_\mathrm{I}(\omega). \tag{5.42}$$

Because all the coefficients multiplying the Lorentzian profile in Eq. (5.40) are positive, we have $\chi_\mathrm{I}(\omega) > 0$ for all ω. This implies that $\alpha(\omega) > 0$, allowing us to conclude:

Atoms modelled by the spring-electron model absorb EM waves whose frequencies are close to the atom's transition frequency. This constitutes the classical description of absorption.

The Lorentzian absorption profile signifies that, if a range of EM waves with different frequencies (but identical amplitudes) are injected into the medium, then the spectrum coming out of the medium will have a dip following the Lorentzian profile.

5.6 What about stimulated emission?

We have now seen that our simple spring-electron model correctly accounts for the (i) frequency-dependence of the refractive index, (ii) spontaneous emission, (iii) the fact that atomic transitions have finite linewidths, and (iv) absorption. But what about *stimulated emission*?

If the imaginary part of the susceptibility would be negative, the intensity of our EM wave would increase exponentially with propagation ($\alpha(\omega) < 0$). As we have seen earlier in section 3.9, this would be a clear signature of light amplification via stimulated emission. But unfortunately, we have just argued that, because

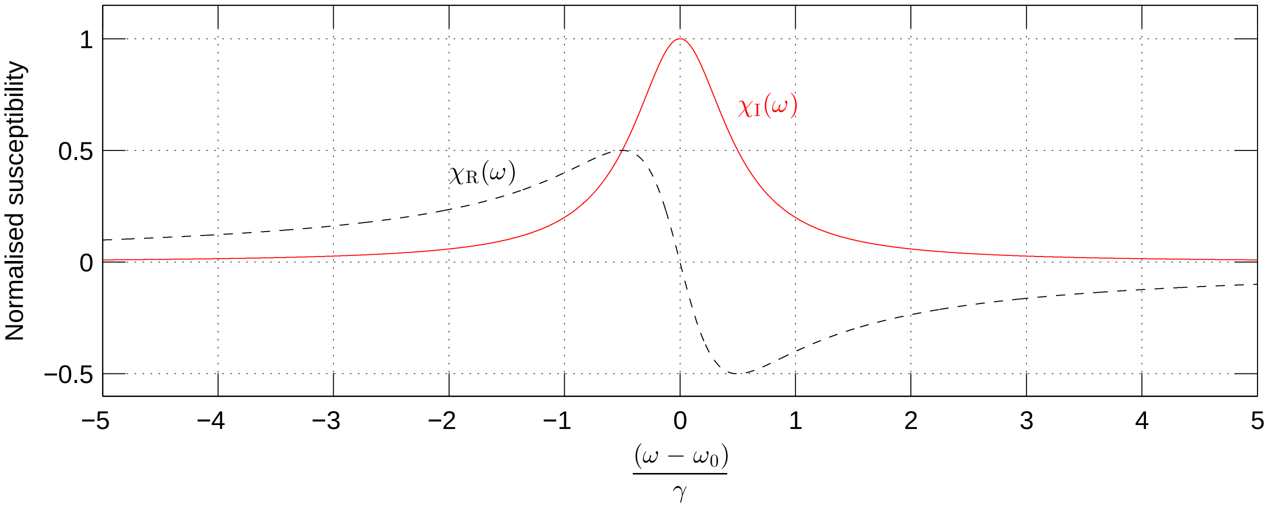

Figure 25: Real (dashed black curve) and imaginary (solid red curve) parts of the normalised susceptibility $\chi(\omega)$ close to an atomic resonance. The normalisation is such that the maximum of the imaginary part is unity.

all the coefficients multiplying the Lorentzian in Eq. (5.40) are positive, $\chi_I(\omega) < 0$. No wonder that the stimulated emission remained hidden until Einstein hypothesised it in 1917 using the photon description of light![58]

But wait a minute, what if N should not be interpreted as the density of atoms (which is always strictly positive), but rather as the difference between populations in two energy levels: $N = N_1 - N_2$? Then we would find that, under conditions of population inversion ($N_2 > N_1$), $N < 0$: hence the imaginary part of the susceptibility becomes negative and our EM wave will be exponentially amplified! We get stimulated emission!

There is more. Using $N = N_1 - N_2 = -\Delta N$, where ΔN is the population inversion as defined in Section 3.9, we can re-write the expression for exponential amplification from above:

$$I(z,\omega) = |E_0|^2 e^{\sigma \Delta N z}, \tag{5.43}$$

where

$$\boxed{\sigma(\omega-\omega_0) = \frac{\omega_0}{c}\frac{\chi_I(\omega)}{N} = \frac{e^2}{2\varepsilon_0 cm}\frac{\gamma/2}{(\omega-\omega_0)^2 + (\gamma/2)^2}.} \tag{5.44}$$

[58] Note that the electron oscillator model is much older.

By interpreting $N = N_1 - N_2$ as the difference in populations at the two energy levels of the atomic transition, the spring-electron model not only predicts exponential amplification via stimulated emission, but it even provides a classical prediction for the interaction cross-section $\sigma(\omega)$ defined in terms of Einstein coefficients in Section 3.9! And the best part is that the prediction is very close to reality: a fully quantum mechanical calculation yields

$$\sigma_{\mathrm{QM}}(\omega) = \frac{\sigma(\omega)}{f}, \tag{5.45}$$

where the constant factor f is the so-called "transition strength". This shows that the spring-model provides the correct qualitative picture, of light-matter interactions. Furthermore, for strong transitions $f \approx 1$, highlighting that even quantitative agreement is not beyond reach.

Problems

5.1 Consider a co-sinusoidal wave with angular frequency ω:

$$E(t) = E_0 \cos(\omega t).$$

(a) Analytically compute the Fourier transform of the wave.

(b) Write a computer code that uses FFT (Fast Fourier transform) to calculate the wave's Fourier transform when $\omega = 2\pi \times 200$ THz and $E_0 = 1$ V/m. **Hint:** the time grid used to sample the wave should be an integer multiple of the wave period, i.e., $t_{\mathrm{span}} = N \times 2\pi/\omega$. Can you explain why?

(c) Plot the numerically computed power spectrum as a function of frequency. Throughout all exercises and assignments, you can normalize the spectrum to unity.

(d) Consider now a corresponding complex field

$$E(t) = E_0 e^{i\omega t}.$$

Repeat parts (a) and (b) and comment on the difference.

5.2 Consider a co-sinusoidal *carrier* wave with frequency $\omega_0 = 2\pi \times 200$ THz that is *modulated* with another co-sinusoidal wave with frequency $\Delta\omega = 2\pi \times 10$ THz:

$$E(t) = E_0 \cos(\omega_0 t) \times \cos(\Delta\omega t).$$

For the amplitude E_0, you can pick whatever you want...

(a) Write a computer that code that plots the field as a function of time.

(b) Write a computer code that uses FFT to calculate the fields's Fourier transform. **Hint:** the time grid used to sample the wave should be an integer multiple of of both wave periods...

(c) Plot the numerically computed power spectrum as a function of frequency. You only need to consider positive frequencies.

(d) Interpret the power spectrum in light of what you know about beat signals / notes / frequencies...

5.3 Consider the classical electron oscillator model in the absence of a driving electric field:

$$m\frac{dx^2(t)}{dt^2} = -m\omega_0^2 x - \gamma m\frac{dx}{dt}.$$

(a) Show that, in the limit of small damping ($\omega_0 \gg \gamma$), the general solution of the equation can be written as:

$$x(t) = Ae^{-\gamma t/2}\cos(\omega_0 t + \delta). \tag{5.46}$$

(b) The oscillating electron gives rise to an oscillating dipole moment $p(t) = -ex(t)$. It is well-known that dipoles emit EM radiation, and that far away from the dipole, the radiation field obeys

$$E(t) \propto \frac{d^2p(t)}{dt^2}.$$

Show that, in the small damping limit, Eq. (5.46) results in the following far-field EM radiation profile:

$$E(t) = E_0 e^{-\gamma t/2}\cos(\omega_0 t + \delta).$$

(c) Analytically prove that the spectrum of $E(t)$ given above is a Lorenzian.

(d) Write a computer code that explicitly calculates the Fourier transform of $E(t)$, and show that the spectrum is a Lorentzian. For the numerical part, you can use the following parameters: $\gamma = 2\pi \times 15$ THz, $\omega = 2\pi \times 300$ THz, and $\delta = 0$.

5.4 (a) Write down the equation of motion for the classical electron oscillator model.

(b) Define all the terms in the equation, and describe their physical interpretation.

(c) Show mathematically that, for plane wave driving at frequency ω, this equation yields an expression for the electric susceptibility χ given by

$$\chi(\omega) = -\frac{Ne^2}{m\varepsilon_0}\frac{1}{(\omega^2 - \omega_0^2) - i\gamma\omega}.$$

(d) Derive expressions for the real and imaginary components of the electric susceptibility, following the definition $\chi = \chi_{\mathrm{R}} - i\chi_{\mathrm{I}}$, when $\omega \approx \omega_0$, i.e., near resonance.

(e) Sketch carefully labelled graphs of χ_{R} and χ_{I} as a function of $(\omega - \omega_0)/\gamma$ and describe, in one or two sentences only, the physical effect each of these has on a propagating plane wave.

(f) With respect to the "real" world, what are the limitations of the electron oscillator model and why? How can we extend its validity?

5.5 Consider a Gaussian pulse:

$$E(t) = E_0 e^{-t^2/T_0^2}, \tag{5.47}$$

where T_0 characterises the width of the pulse.

(a) Analytically compute the Fourier transform of the pulse.

(b) How does the frequency bandwidth of the power spectrum (absolute value squared of the Fourier transform) depend on the pulse duration?

(c) Interpret the relationship between the frequency bandwidth and pulse duration in light of Heisenberg's uncertainty principle.

5.6 In the lecture notes, we derived an expression for the electron displacement $x(t)$ in the absence of a driving field [Eq. (5.13)]. In particular, we showed that, if the atom is initially excited ($A \neq 0$), it will undergo exponentially damped oscillations. Furthermore, we showed that the electron's Fourier spectrum has a Lorentzian shape. Here we will perform an alternative derivation of all this.

(a) Show that the Fourier transform, as defined by Eq. (4.65), satisfies the following property:

$$\mathcal{F}\left[\frac{dx(t)}{dt}\right] = (i\omega)\tilde{X}(\omega),$$

where $\tilde{X}(\omega) = \mathcal{F}[x(t)]$ is the Fourier transform of $x(t)$.

(b) For the atom to be initially excited, there must some form of initial excitation. We will assume this excitation occurs at $t = 0$, and we will model it using a Dirac delta function. In this case, the equation of motion reads:

$$m\frac{dx^2(t)}{dt^2} = -m\omega_0^2 x - \gamma m\frac{dx}{dt} - eE_0\delta(t).$$

Take the Fourier transform of both sides of the equation, and derive an expression for the Fourier transform $\tilde{X}(\omega)$.

(c) Show that, with the usual assumption $\gamma \ll \omega_0$, the Fourier spectrum of $x(t)$ (i.e., $|X(\omega)|^2$) has a Lorentzian shape.

(d) If you are brave enough, take the inverse Fourier transform of $X(\omega)$ to obtain the time-domain field.

5.7 Consider the equation of motion of the classical electron oscillator in the presence of a plane EM wave:

$$m\frac{dx^2(t)}{dt^2} = -m\omega_0^2 x - \gamma m\frac{dx}{dt} - eE(t).$$

(a) Take the Fourier transform of both sides of the equation, and derive an expression for the Fourier transform $\tilde{X}(\omega)$.

(b) Take the inverse Fourier transform of $\tilde{X}(\omega)$ to show that Eq. (5.22) [together with Eq. (5.24)] of the lecture notes provides a general solution to the equation.

5.8 Write a computer code that uses the Sellmeier equation to calculate the refractive index of a given material at a given wavelength. In particular, the code should take the relevant Sellmeier coefficients and wavelength as inputs, and then return the correct refractive index. Use the code to plot the refractive indices of the following materials as a function of wavelength ($\lambda = 350\ \mathrm{nm} \ldots 1600\ \mathrm{nm}$): fused silica (normal glass), magnesium fluoride ($\mathrm{MgF_2}$), and water ($\mathrm{H_2O}$). **Hint:** The website www.refractiveindex.info is extremely useful for finding Sellmeier coefficients!

6 Resonators

We have now seen from two different perspectives that the intensity of light propagating in an absorbing or amplifying medium evolves as

$$\frac{dI(\omega)}{dz} = gI(\omega). \tag{6.1}$$

where we defined the **gain coefficient**[59]

$$\boxed{g = \sigma_{21}(\omega - \omega_{21})\Delta N.} \tag{6.2}$$

In general, the gain coefficient depends on both frequency (through the frequency dependence of the cross-section) and on intensity (through the intensity dependence of the inversion ΔN), but we omit both for simplicity. In the small-signal limit, where $I \gg I_{\mathrm{sat}}$, the population inversion is constant [see Section 3.12] and the attenuation or amplification is simply exponential:

$$I(z,\omega) = I(0,\omega)e^{gz}. \tag{6.3}$$

Laser operation typically starts from spontaneous emission, which creates a single "seed" photon that is subsequently amplified via stimulated emission. A single photon carries insignificant energy, implying $I(0,\omega) \ll 1$. Compounded by the fact that the gain coefficient is typically close to unity, this would suggest that the active medium must be enormously long ($z \gg 1$ m) to reach significant levels of intensity. As this is not practical, an alternative solution is needed. That solution is to employ an **optical resonator**[60], which is a device in which light can be made to circulate back and forth. In this way, light can pass through the active medium numerous times, being slightly amplified at each transit until some steady-state situation with significant intracavity intensity is reached.

In addition to lasers, resonators have many other applications. For example, they can be used in absorption spectroscopy to enhance the sensitivity of detection.[61] They can also be used as very high-resolution spectrometers that allow the wavelength of light to be determined with great precision. Other applications include sensing and nonlinear optics.

In this Section, we describe the basic physics of optical resonators. In particular, we will look at a resonator that does not contain an amplifying or an attenuating medium, and that is simply comprised of two partially reflective parallel mirrors as shown in Fig. 27: **a Fabry-Perot resonator**. To describe the characteristic resonator behaviour, we consider the situation where light is externally injected into the resonator. Because partially reflective mirrors are key to the operation of optical resonators, we begin by briefly considering their **complex reflection and transmission coefficients**.

6.1 Phase shifts in lossless reflection and transmission

We consider a partially reflecting mirror, illuminated from left and right by two electromagnetic waves with complex scalar amplitudes E_{a} and E_{b} (see Fig. 26). Both input waves will be partially transmitted and reflected,

[59] Be wary not to confuse the gain coefficient with the lineshape function.

[60] Also known as optical cavity.

[61] Even small absorption features can become visible when light passes several times through the sample.

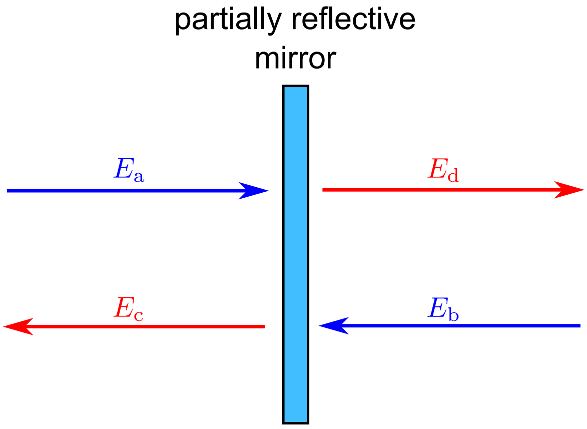

Figure 26: Partially reflecting mirror which transforms two input EM waves E_a and E_b (blue) into two output waves E_c and E_d (red).

giving rise to output waves E_c and E_d. The output fields are related to the input fields as

$$E_\mathrm{c} = r_\mathrm{ac} E_\mathrm{a} + t_\mathrm{bc} E_\mathrm{b}, \tag{6.4}$$

$$E_\mathrm{d} = r_\mathrm{bd} E_\mathrm{b} + t_\mathrm{ad} E_\mathrm{a}, \tag{6.5}$$

where r_nm and t_nm are the **complex** reflection and transmission coefficients describing the reflection/ transmission $E_\mathrm{n} \to E_\mathrm{m}$. Assuming that the mirror is lossless, energy must be conserved, which also implies that the total intensity must be conserved. We can thus write

$$|E_\mathrm{out}|^2 = |E_\mathrm{in}|^2, \tag{6.6}$$

$$|E_\mathrm{c}|^2 + |E_\mathrm{d}|^2 = |E_\mathrm{a}|^2 + |E_\mathrm{b}|^2. \tag{6.7}$$

Expressing the reflection and transmission coefficients in polar form [e.g. $r_{nm} = |r_{nm}| \exp{(i\phi_{nm})}$], we find that the following two conditions must hold:

$$|r_\mathrm{ac}|^2 + |t_\mathrm{ad}|^2 = |r_\mathrm{bd}|^2 + |t_\mathrm{bc}|^2 = 1, \tag{6.8}$$

$$\phi_\mathrm{bd} + \phi_\mathrm{bc} - \phi_\mathrm{ac} - \phi_\mathrm{ad} = \pi. \tag{6.9}$$

Because only relative phases are important, we can choose the phases associated with transmission to be zero. The condition above then reduces to $\phi_\mathrm{bd} = \phi_\mathrm{ac} + \pi$, and we obtain our final result:

$$\boxed{\begin{aligned} t_\mathrm{ad} &= t_\mathrm{bc}, & (6.10)\\ r_\mathrm{ac} &= -r_\mathrm{bd}. & (6.11) \end{aligned}}$$

And so we see that the transmission coefficients are identical irrespective of the direction, while a π (relative) phase shift occurs depending on the direction of reflection.

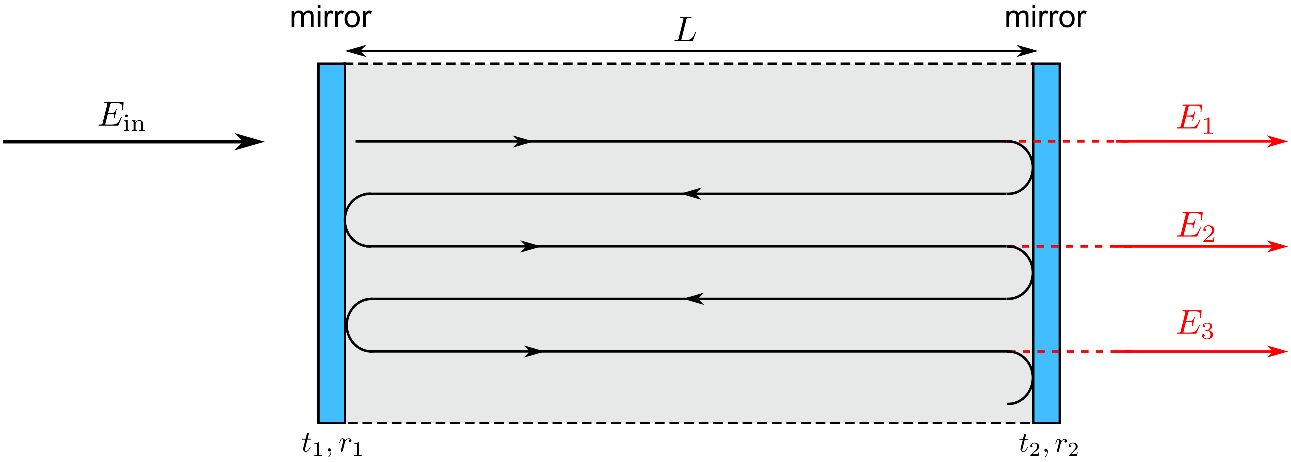

Figure 27: Schematic illustration of a driven Fabry-Perot resonator, consisting of two partially reflective mirrors separated by a length L. The mirrors' complex transmission and reflection coefficients are $t_{1,2}$ and $r_{1,2}$, respectively. For simplicity, the field components reflected by the resonator, i.e., those exiting on the side of the driving field, are not shown.

It is worth highlighting that the condition above is an "abstract" requirement for energy conservation in any partially reflecting mirror. In real mirrors the phases cannot be chosen arbitrarily, but they depend on the construction of the mirror; however, if the mirror is lossless, the phase shifts must satisfy the general condition above.

6.2 Fabry-Perot intensity transmission

We now describe the physics of a simple Fabry-Perot resonator, consisting of two partially reflecting mirrors with amplitude reflection and transmission coefficients $r_{1,2}$ and $t_{1,2}$, respectively. To ensure that the mirrors conserve energy (as per discussion above), we insist that reflections from the inner surfaces are positive ($r_{1,2} > 0$) while reflections from the outer surfaces are negative ($-r_{1,2} < 0$).[62]

A monochromatic EM wave $E = E_{\text{in}}e^{i(\omega t - kz)}$ is continuously injected into the resonator, giving rise to an intracavity field bouncing back and forth between the two mirrors. Some part of the injected field only completes one transit before it escapes, while another part completes multiple bounces back and forth. The field at the cavity output corresponds to the superposition of all the fields that are partially transmitted after different numbers of transits around the cavity. In steady-state operation, we can find an expression for that output by summing up all the fields that are partially transmitted after different round trips.

Letting E_n correspond to the complex amplitude[63] of the EM wave that is transmitted after n cavity transits, we can write the total complex amplitude of the transmitted field as

$$E_{\text{T}} = E_1 + E_2 + E_3 + \cdots \tag{6.12}$$

During a full cavity transit (i.e., from mirror 2 to mirror 1 and back), the amplitude of the intracavity field is

[62] As we shall see, this decision does not play any role when evaluating the transmitted field, but becomes important when considering the reflected field.

[63] The full field is $E = E_n e^{i(\omega t - kz)}$.

reduced by reflections at the two mirrors. Furthermore, the phase of the field changes as $e^{-i\delta}$, where $\delta = 2kL$ and L is the length of the resonator[64]. Expressions for the complex amplitudes E_n are easy to find iteratively:

Complex amplitudes transmitted after n round trips in a Fabry-Perot resonator

$E_1 = t_1 t_2 e^{-i\delta/2} E_{\text{in}}$ The injected field is transmitted by both mirrors and exits immediately.

$E_2 = r_2 r_1 e^{-i\delta} E_1$ The field reflects from mirror 2 and mirror 1 and is then transmitted.

$E_3 = r_2 r_1 e^{-i\delta} E_2$ The field reflects twice from both mirrors and is then transmitted.

$\vdots$

It is easy to see that the total transmitted field can be written as a geometric series:

$$E_{\text{T}} = t_1 t_2 e^{-i\delta/2} E_{\text{in}} \left(1 + r_1 r_2 e^{-i\delta} + (r_1 r_2)^2 e^{-i2\delta} + \cdots\right). \tag{6.13}$$

As $|r_1 r_2| < 1$, the sum converges and we obtain

$$E_{\text{T}} = t_1 t_2 e^{-i\delta/2} E_{\text{in}} \frac{1}{1 - r_1 r_2 e^{-i\delta}}. \tag{6.14}$$

Typically we are more interested in intensities than field values, and so we evaluate $I_{\text{T}} = |E_{\text{T}}|^2$:

$$\frac{I_{\text{T}}}{I_{\text{in}}} = \frac{T_1 T_2}{1 + R_1 R_2 - 2\sqrt{R_1 R_2}\cos(\delta)}. \tag{6.15}$$

Here $T_{1,2} = |t_{1,2}|^2$ and $R_{1,2} = |r_{1,2}|^2$ are the intensity transmission and reflection coefficients of the two mirrors, respectively.[65] After some algebra, we obtain the "standard" form for the intensity transmission of a Fabry-Perot resonator:

$$\frac{I_{\text{T}}}{I_{\text{in}}} = \frac{T_1 T_2}{(1 - \sqrt{R_1 R_2})^2} \frac{1}{1 + F\sin^2(\delta/2)}, \tag{6.16}$$

$$F = \frac{4\sqrt{R_1 R_2}}{(1 - \sqrt{R_1 R_2})^2}. \tag{6.17}$$

[64] This phase shift simply arises because of the usual $\exp(-ikz)$ factor associated with the propagation of EM waves. Note that we do not need to keep track of the $\exp(i\omega t)$ terms since this phase is common to all the waves; the propagation phase is not, since different waves travel different distances (round trips).

[65] Energy conservation requires $R_{1,2} + T_{1,2} = 1$.

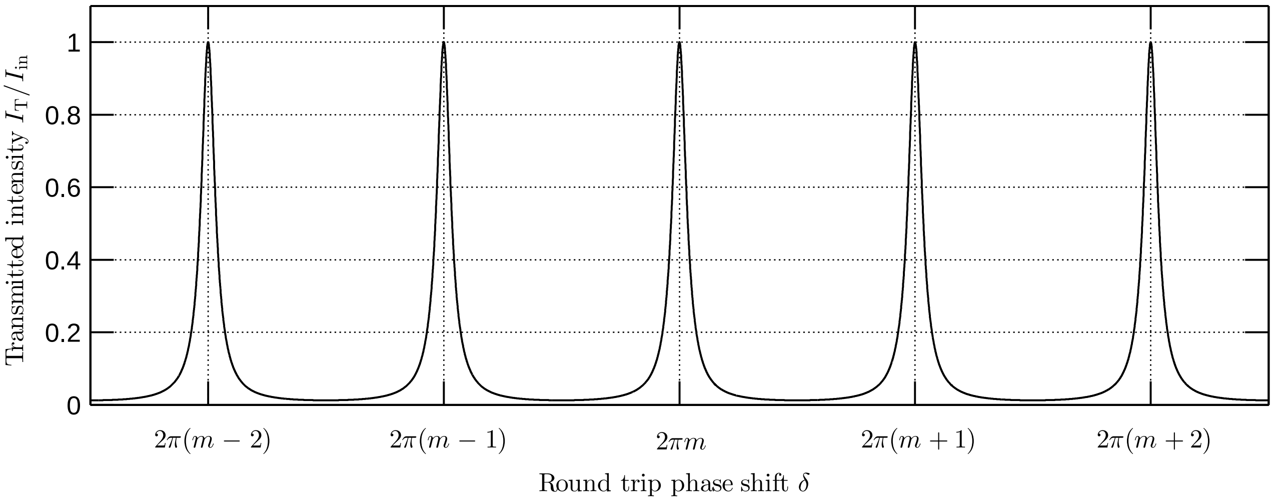

Figure 28: Fabry-Perot transmission for two identical mirrors with $R_1 = R_2 = R = 0.8$.

When the mirrors are identical, such that $R_1 = R_2 = R$, the expressions simplify considerably:

$$\frac{I_\mathrm{T}}{I_\mathrm{in}} = \frac{1}{1 + F\sin^2(\delta/2)}, \tag{6.18}$$

$$F = \frac{4R}{(1-R)^2}. \tag{6.19}$$

6.3 Longitudinal modes

For fixed mirror parameters, the intensity transmission of a Fabry-Perot depends only on the round trip phase shift δ. Figure 28 shows a typical transmission curve, calculated for parameters listed in the caption. As expected based on the sinusoidal dependence on δ, the transmission is periodic. The transmission is maximised when $\sin^2(\delta/2) = 0$, i.e., for $\delta = 2m\pi$ where m is an integer. This condition means that **the roundtrip phase shift is an integer multiple of 2π, implying that all the different waves corresponding to different round trips will be in phase at each position along the cavity.** As a consequence, the waves reinforce each other via constructive interference, giving rise to a large intensity.

Because $\delta = 2kL = 4\pi L/\lambda$, the transmission is maximised for wavelengths

$$\lambda_m = \frac{2L}{m}. \tag{6.20}$$

These wavelengths are known as **resonance wavelengths**, and the corresponding frequencies

$$f_m = \frac{c}{n\lambda_m} = \frac{c}{2nL}m \tag{6.21}$$

are known as **resonance frequencies** (n is the refractive index). Just like for waves on a string, the resonances correspond to arrangements in which an integer number of half-wavelengths fits in the resonator. Furthermore, as the EM wave is bouncing between the two mirrors, it is travelling in both directions in the cavity, thus forming a standing wave pattern (just like waves on a string). These standing wave patterns are known as the **longitudinal modes** of the cavity.

6.4 Reflected field

The analysis above reveals that only those wavelengths that are on-resonance[66] will be transmitted by a Fabry-Perot. If the wavelength of the injected field falls outside of a resonance, it will not be transmitted. In other words, a Fabry-Perot acts as a spectral filter with a comb-like response. But of course, energy must be conserved, so where does the light go if it is not transmitted? Let us look at the reflection side of things.

The analysis proceeds exactly as before. The only thing to note now is that the first partially reflected field arises immediately from the reflection of the injected field from the outer surface of mirror 1. In keeping with our previous convention, we assign a π phase shift to that reflection to ensure energy conservation. Therefore the first reflected field is $E_0 = -r_1 E_{\text{in}}$.[67] As we are lazy, we assume for simplicity that the mirrors are identical, such that $r_1 = r_2$ and $r_1 = r_2$.

Complex amplitudes reflected after n round trips in a Fabry-Perot resonator

$E_0 = -rE_{\text{in}}$	The injected field is reflected immediately by mirror 1.
$E_1 = trte^{-i\delta}E_{\text{in}}$	The field transmits through mirror 1, bounces from mirror 2, and exits.
$E_2 = rre^{-i\delta}E_1$	The field reflects twice from both mirrors and then exits.
$\vdots$	

The total reflected field is the sum of all of the partially reflected waves,

$$E_{\text{R}} = E_0 + E_1 + E_2 + E_3 \cdots , \tag{6.22}$$

$$= -rE_{\text{in}} + trtE_{\text{in}}e^{-i\delta}\Big[1 + rre^{-i\delta} + (rr)^2 e^{-i2\delta}\cdots\Big] . \tag{6.23}$$

The term in brackets can again be identified as a converging geometric series. After some algebra, we obtain the following expression for the total reflected intensity $I_{\text{R}} = |E_{\text{R}}|^2$:

$$\boxed{\frac{I_{\text{R}}}{I_{\text{in}}} = \frac{F\sin^2(\delta/2)}{1 + F\sin^2(\delta/2)},} \tag{6.24}$$

[66] Or close to a resonance.

[67] Notice that, in contrast to the analysis of the transmitted wave, there will now be waves transmitted and reflected in both directions at mirror 1. Energy conservation requires that there is a π phase difference between reflection/transmission from the inner resonator side and reflection/transmission from the outer side. We choose the reflection from the outer surface to carry that full phase shift, but emphasize that this is an arbitrary choice.

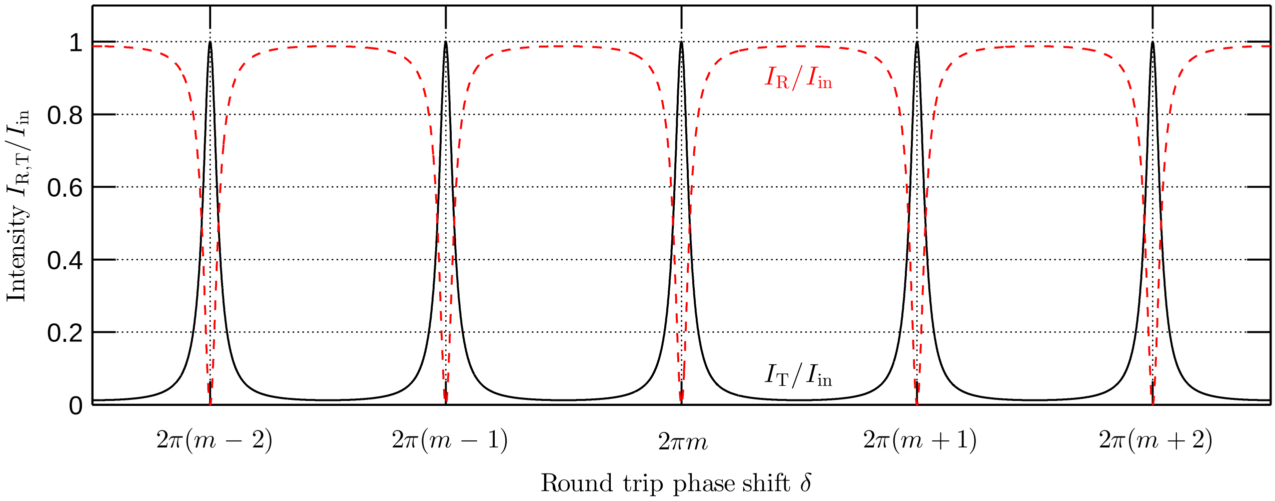

Figure 29: Fabry-Perot transmission and reflection curves for two identical mirrors with $R_1 = R_2 = R = 0.8$.

where F is given by Eq. (6.19). In Fig. 29 we plot the transmitted and reflected intensities given by Eqs. (6.18) and (6.24), respectively, for parameters listed in the caption. As can be seen, wavelengths that are not transmitted are reflected (and vice versa). Mathematically, it is a simple task to verify that $I_{\mathrm{T}} + I_{\mathrm{R}} = I_{\mathrm{in}}$.

6.5 Basic properties of (Fabry-Perot) resonators

In this Subsection we discuss some key properties of (Fabry-Perot) resonators. For simplicity, we consider the case where the two mirrors are identical, such that the reflected and transmitted intensities are governed by Eqs. (6.18) and (6.24).

Energy storage

On resonance, the transmitted intensity is equal to the injected intensity, $I_{\mathrm{T}} = I_{\mathrm{in}}$. On the other hand, the intracavity intensity is given by

$$I_{\mathrm{cav}} = \frac{I_{\mathrm{T}}}{T}. \tag{6.25}$$

Because $T < 1$, this implies that, on resonance, the intracavity intensity exceeds the injected intensity. Thus, a (Fabry-Perot) resonator stores energy, acting as a light capacitor.

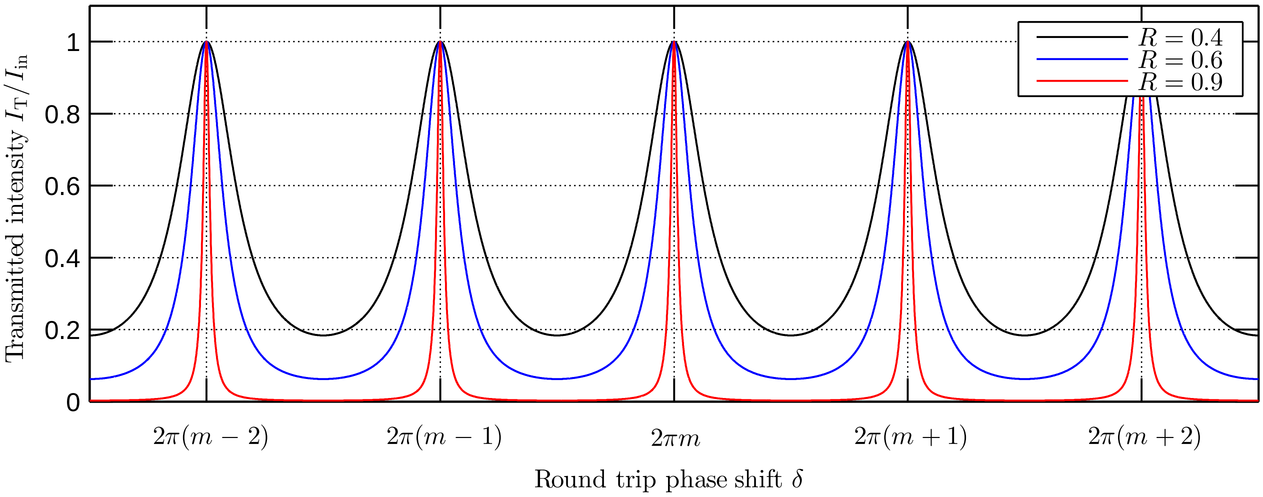

Figure 30: Transmission for a symmetric Fabry-Perot resonator ($R_1 = R_2 = R$) with three different mirror reflectivity.

Free-spectral range

The resonance frequencies are equally-spaced, and the separation between two adjacent resonances is known as the **free spectral range** (FSR). From Eq. (6.21), we find

$$\mathrm{FSR} = f_{m+1} - f_m \tag{6.26}$$

$$= \frac{c}{2nL}(m+1) - \frac{c}{2nL}m \tag{6.27}$$

$$= \frac{c}{2nL}. \tag{6.28}$$

And so the frequency spacing between adjacent longitudinal modes (i.e. the FSR) is fully determined by the length of the resonator and the refractive index of the material inside it. Note that, strictly speaking, the above equation is only true for a resonator consisting of two mirrors between which light is bouncing back and forth. For arbitrary resonators, the FSR is given by

$$\mathrm{FSR} = \frac{c}{nL_{\mathrm{c}}}, \tag{6.29}$$

where L_{c} is the *round trip length* of the resonator. It should be clear that $\mathrm{FSR} = t_{\mathrm{R}}^{-1}$, where t_{R}^{-1} is the time that it takes for light to complete one round trip. This is a general result that is applicable to all kinds of resonators, even those that are made out of many different materials with different refractive indices: the FSR is given by the inverse of the total round trip time.

Resonance widths

The half-width at half-maximum of a single transmission resonance peak can be solved from

$$\frac{I_\mathrm{T}}{I_\mathrm{in}} = \frac{1}{1 + F\sin^2(\delta_{0.5}/2)} = \frac{1}{2}. \tag{6.30}$$

Rearranging, we obtain

$$\sin^2(\delta_{0.5}/2) = \frac{1}{F}. \tag{6.31}$$

This expression shows that the larger F, the narrower the resonances. On the other hand, $F \to \infty$ as the mirror reflectivity $R \to 1$. The resonances are thus narrow for cavities made of mirrors that have high reflectivity. Such cavities are said to have low loss, since only a small amount of light is lost each round trip.[a] Figure 30 illustrates the trend, showing the transmission curves for a variety of R.

When $F \gg 1$, the resonance width $\delta_{0.5} \ll 1$. We can then approximate $\sin^2(\delta_{0.5}/2) \approx \delta_{0.5}^2/4$, obtaining the full-width at half-maximum

$$\delta_\mathrm{FWHM} = 2\delta_{0.5} \approx \frac{4}{\sqrt{F}}. \tag{6.32}$$

[a] Recall that the transmission $T = 1 - R$.

Finesse and quality factor

The **finesse** of a resonator is defined as the ratio between the FSR and the resonance full-width at half maximum (expressed in frequency):

$$\mathcal{F} = \frac{\mathrm{FSR}}{\Delta f} = \frac{2\pi}{2\delta_{0.5}}, \tag{6.33}$$

where the second expression comes from the fact that one FSR corresponds to 2π in phase. Using $2\delta_{0.5} \approx 4/\sqrt{F}$, we can write

$$\mathcal{F} \approx \frac{\pi\sqrt{F}}{2} = \frac{\pi\sqrt{R}}{1-R}. \tag{6.34}$$

For $R \approx 1$, we can further approximate

$$\mathcal{F} \approx \frac{\pi}{1-R} = \frac{\pi}{T}. \tag{6.35}$$

The expression above shows that the finesse is governed by the mirror reflectivity R. In particular, we see that a large reflectivity gives rise to large finesse; in others words, a low-loss cavity has high finesse, and accordingly sharp resonances. This is a general property of all resonators. Indeed, one finds in general that finesse is fully governed by the cavity losses. For arbitrary resonators, it can be approximately written as

$$\mathcal{F} \approx \frac{2\pi}{1-\rho}, \tag{6.36}$$

where $1-\rho$ is the total intensity lost per round trip. For our symmetric cavity, the total loss $1-\rho = (1-R^2) = (1-R)(1+R) \approx 2(1-R) = 2T$, yielding our earlier expression. It is worth highlighting that all of the approximations above are valid in the *good-cavity limit*, for which the finesse is large: $\mathcal{F} \gg 1$

The finesse of an optical resonator is a very important characteristic. In fact, any given resonator is fully characterised by its FSR and finesse; FSR describes the spacing of the longitudinal modes, while the finesse characterises the cavity losses and the sharpness of the resonances. Furthermore, both FSR and the finesse can be easily accessed experimentally, by simply measuring the transmission as the frequency of the injected laser is scanned.

Another quantity – closely related to the finesse – that is often used to describe the losses of a resonator is the resonator quality-factor (Q-factor). There are two definitions for Q-factor, which can be shown to be equivalent in the limit $Q \gg 1$ (this is always the case for optical resonators). First, the Q-factor can be defined as the ratio between energy stored in the resonator and the energy lost per oscillation cycle; second, Q-factor can be defined as the ratio between the resonance frequency and the width of that resonance. Mathematically:

$$Q = \frac{\omega_p E_{\mathrm{st}}}{P_{\mathrm{dis}}} = \frac{\omega_p}{\Delta\omega}, \tag{6.37}$$

where E_{st} is the energy stored, P_{dis} is the rate at which energy is lost, and ω_{p} is the p^{th} resonance (angular) frequency with width $\Delta\omega$. It is easy to see that the relationship between the Q-factor and the finesse is simply

$$Q = \frac{\omega_p \mathcal{F}}{2\pi \mathrm{FSR}} \tag{6.38}$$

Cavity photon lifetime

Yet a third quantity describing the cavity losses is the cavity photon lifetime. It can be understood as the average time that a photon persist in the resonator after the injected field is turned off. If we have ϕ_0 photons in the cavity before the injected field is turned off, then after one roundtrip we are left with $\phi_1 = R^2\phi_0$, after two roundtrips $\phi_2 = R^2\phi_1$ and so on. We can thus write a rate equation for the photon number:

$$\frac{\Delta\phi}{\Delta t} = \frac{R^2\phi - \phi}{t_{\mathrm{r}}}, \tag{6.39}$$

where $t_{\mathrm{R}} = 2nL/c$ is the round trip time (i.e., the time it takes for light to complete one round trip). In the limit $\Delta t \to 0$, the finite differences become differentials:

$$\frac{d\phi}{dt} = -\frac{\phi}{\tau_{\mathrm{ph}}}, \tag{6.40}$$

where the **cavity photon lifetime**

$$\tau_{\mathrm{ph}} = \frac{t_{\mathrm{R}}}{1-R^2} = \frac{1}{\mathrm{FSR}(1-R^2)}. \tag{6.41}$$

Clearly the photon number decays exponentially, with a time constant corresponding to the photon lifetime:

$$\phi(t) = \phi(0)e^{-t/\tau_{\mathrm{ph}}}. \tag{6.42}$$

Thus, indeed, the photon lifetime describe the lifetime of photons in the cavity. For low-loss cavities ($R \approx 1$), the photon lifetime is related to the cavity finesse as

$$\tau_{\mathrm{ph}} = \frac{t_{\mathrm{R}}}{2\pi}\mathcal{F}. \tag{6.43}$$

The larger the finesse the larger the photon lifetime, as expected. Last but not least, recalling that $t_{\mathrm{R}} = \mathrm{FSR}^{-1}$ and $\mathcal{F} = \mathrm{FSR}/\Delta f$, we find that

$$\tau_{\mathrm{ph}} = \frac{1}{2\pi\Delta f}, \tag{6.44}$$

where Δf is the width of the resonance.

Problems

6.1 Consider electromagnetic waves E_{a} and E_{b} incident on a partially reflecting mirror from different sides of the mirror. Through reflection and transmission, the incident waves transform into output waves E_{c} and E_{d} [see Fig. 26]. Show that, for a lossless mirror, the complex transmission and reflection coefficients must satisfy the following two conditions:

$$|r_{\mathrm{ac}}| + |t_{\mathrm{ad}}| = |r_{\mathrm{bd}}| + |t_{\mathrm{bc}}| = 1$$

$$\phi_{\mathrm{bd}} + \phi_{\mathrm{bc}} - \phi_{\mathrm{ac}} - \phi_{\mathrm{ad}} = \pi,$$

where the subscript ij denotes reflection or transmission from input waves E_i onto output wave E_j.

6.2 Consider a Fabry-Perot etalon made from two identical mirrors, with nothing but air in between the mirrors. Answer the following questions:

(a) The Fabry-Perot's free-spectral range is measured to be 10 GHz. What is its length, i.e., the distance between the mirrors?

(b) The resonator's finesse is $\mathcal{F} = 100$. What is the spectral full-width at half-maximum of its transmission peaks?

(c) Calculate the reflectivity of the mirrors.

(d) The measured power transmitted by the resonator is 1 mW. What is the optical power inside the Fabry-Perot?

6.3 A Fabry-Perot etalon can be used as a very high-resolution optical spectrum analyzer, i.e., a device that can resolve closely-spaced optical frequencies. Devise a way to use a Fabry-Perot for this purpose. Considering a Fabry-Perot composed of two identical mirrors with $R = 0.99$ separated by $d = 10$ cm, answer the questions below. You can assume the light examined to be visible, such that $\lambda_0 \approx 500$ nm and $f_0 = c/\lambda_0 \approx 600$ THz

(a) What is the frequency resolution of the Fabry-Perot -based spectrum analyzer?

(b) What is the corresponding wavelength resolution?

(c) What is the resolving power of the spectrum analyzer? (You may need to google the definition.)

6.4 Consider a Fabry-Perot resonator composed of two identical mirrors. Write a computer code that shows that the profile of a single cavity resonance can be approximated with a Lorenzian curve:

$$\left(\frac{I_\mathrm{T}}{I_\mathrm{in}}\right)_{\delta\ll 2\pi} \approx \frac{(\gamma/2)^2}{\delta^2 + (\gamma/2)^2}.$$

What should the variable γ be in terms of the cavity parameters? Compare the expression above with the full “Airy” function [Eq. (6.18)] by plotting the two curves together.

6.5 Researchers at the University of Auckland are very interested in ultra-high-quality microresonators. These are tiny *ring* resonators which exhibit characteristics similar to a Fabry-Perot resonator. The expression for the optical power inside such a ring resonator is given by:

$$\frac{P_\mathrm{cavity}}{P_\mathrm{in}} = \frac{T}{(1-\sqrt{\rho})^2\left[1 + F\sin^2(\delta/2)\right]},$$

where T is the power transmission coefficient of the coupler used to inject light into the resonator (akin to a partially reflective mirror), P_in is the input power, $F = 4\sqrt{\rho}/(1-\sqrt{\rho})^2$, and the coefficient ρ corresponds to the fraction of power left after one round trip: accordingly, $1-\rho$ represents the total power lost per round trip.

(a) In a recent study,[68] the researchers investigated a silica microsphere resonator whose major diameter $d = 250\ \mu$m. Calculate the resonator's free-spectral range. **Hint:** Consider the round trip length and remember that the refractive index of silica $n \approx 1.45$.

(b) To measure the finesse of the microsphere resonator, the researchers scanned the frequency of a tunable laser across a cavity resonance. From the measured signal, they inferred the resonance linewidth to be about $\Delta f \approx 5.2$ MHz. Calculate the finesse of the resonator.

(c) Calculate the value of the parameter ρ, and comment on the amount of power lost per round trip.

(d) Calculate the photon lifetime of the resonator. How many round trips does a photon spend inside the cavity on average without decaying?

(e) In their experiments, the power of the driving field was set to $P_\mathrm{in} = 80$ mW. What is the intracavity power when the driving laser is exactly on-resonance? You can assume that the coupling is *critical*, such that $T = (1-\rho)/2$.

(f) When writing the constitutive relation $P = \varepsilon_0\chi E$, we assumed light intensities not to be humongous. Do you think this is a reasonable assumption in this experiment?

[68] K. E. Webb et al., Opt. Lett. **41**, 4613 (2016).

7 Fabry-Perot Lasers

The preceding Section surveyed the basic physics and key characteristics of (Fabry-Perot) resonators. In the analysis, the cavity was assumed to be filled with a material of refractive index n that was neither absorbing nor amplifying. In this Section, we consider a full laser system, i.e., the situation in which an active medium is placed inside the resonator. Without loss of generality, we assume the reflection coefficients $r_{1,2}$ to be real and positive.[69]

Figure 31 schematically illustrates the general situation. An active medium, with population inversion ΔN and length l_c, is separated from mirror 1 (mirror 2) by a distance l_1 (l_2). As the light passes through the active medium, it will be exponentially amplified with a gain coefficient $g = \sigma_{21}(\omega - \omega_{21})\Delta N$. For the sake of generality, we also include the possibility of absorption or scattering losses in the active medium; these will exponentially attenuate the beam with a loss coefficient α. Note that the frequency-dependence of the gain and loss coefficients is not explicitly written for clarity. Furthermore, we assume that, during a single transit through the active medium, the inversion ΔN stays constant; however, we allow the inversion to change over consecutive round trips, implying the same for the gain coefficient g.

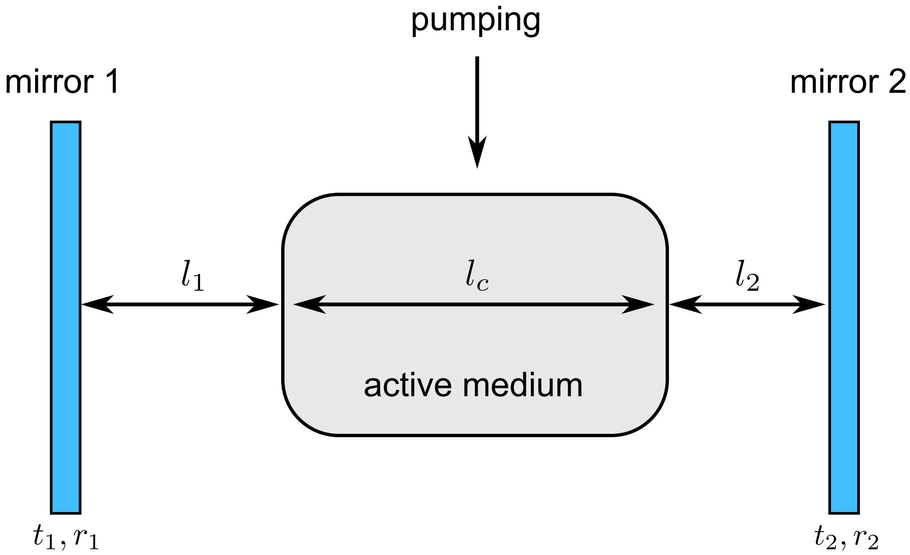

Figure 31: Schematic illustration of a simple Fabry-Perot laser. An active medium is placed in between two mirrors that form a Fabry-Perot resonator. External pumping provides energy needed to achieve population inversion.

7.1 Threshold gain and inversion

Laser operation begins with spontaneous emission creating a small field with broad bandwidth[70]. We denote the initial complex amplitude at frequency ω as $E_0(\omega)$. During one round trip, the field is exponentially amplified (or attenuated) in the active medium, reflected by both mirrors, and it acquires a phase shift due to propagation

[69] In other words, the π phase shift occurs somewhere else.

[70] Photons across the whole lineshape are created.

in air and in the active medium. Mathematically, after one round trip the field component is

$$E_1(\omega) = r_1 r_2 E_0(\omega) e^{(g-\alpha)l_c} e^{-i\delta}. \tag{7.1}$$

Here $\delta = \delta_1 + \delta_2 + \delta_c$ is the total phase shift per round trip, with $\delta_{1,2} = 2k_0 l_{1,2} = 4\pi l_{1,2}/\lambda_0$ representing the phase shift due to propagation in air, and $\delta_c = 2kl_c = 4\pi n l_c/\lambda_0$ the corresponding phase shift due to propagation in the active medium.[71]

For the spontaneous emission seed to be amplified, the gain must be sufficiently high to overcome losses arising from scattering in the active medium as well as mirror reflections. In particular, we require $|E_1(\omega)| > |E_0(\omega)|$. This condition is achieved when

$$r_1 r_2 e^{(g-\alpha)l_c} > 1, \tag{7.2}$$

Solving for the gain coefficient yields

$$\boxed{g > g_t = \alpha - \frac{1}{l_c} \ln{(r_1 r_2)},} \tag{7.3}$$

where we defined the threshold gain g_t coefficient: **the minimum gain required for laser oscillation**.

The threshold gain allows us to obtain a corresponding threshold population inversion, i.e., the minimum inversion needed for laser operation:

$$\boxed{\Delta N_t = \frac{g_t}{\sigma_{21}(0)} = \frac{1}{\sigma_{21}(0)} \left[\alpha - \frac{1}{l_c} \ln{(r_1 r_2)}\right] = \frac{\alpha_{\text{tot}}}{2 l_c \sigma_{21}(0)},} \tag{7.4}$$

where we introduced the **total exponential loss** α_{tot}[72] and set $\omega = \omega_{21}$ as we are interested in the *minimum* inversion required.[73]

[71] Note that the factors of two arise because the beam travels twice through the whole resonator length in one round trip; this factor does not arise for the gain term, since we are writing the amplification of the electric field and not the intensity.

[72] The total loss per roundtrip is $\exp{(-\alpha_{\text{tot}})}$. Check for yourself!

[73] max$[\sigma_{21}(\omega - \omega_{21})] = \sigma_{21}(0)$.

Good exercise

To get an order of magnitude estimate for the threshold inversion, we use Eq. (3.35) to write out $\sigma_{21}(0) = \pi^2 c^2 n g_{\text{line}}(0)/(\omega_{21}^2 \tau_{\text{sp}})$, where $g_{\text{line}}(0)$ is the (maximum) value of the transition's lineshape function. As a first order estimate, we can approximate $g_{\text{H}}(0) \approx 1/\Delta\omega$, where $\Delta\omega$ is the width of the lineshape function, obtaining

$$\Delta N_{\text{t}} = \frac{\alpha_{\text{tot}} \omega_{21}^2 \tau_{\text{sp}} \Delta\omega}{2 l_c \pi^2 c^2 n} \tag{7.5}$$

This equation shows that, to minimise the inversion needed, the following characteristics help:

1. Short upper state lifetime τ_{sp}. On the other hand, this makes it more difficult to achieve the inversion.
2. Small transition bandwidth.
3. Long wavelength (small frequency): lasers at blue are hard.
4. Small losses α_{tot}.

7.2 Steady-state operation

An initial spontaneous emission seed will be amplified towards full laser operation when $g > g_t$. This amplification occurs over several round trips, but must ultimately stop; otherwise the laser intensity would unphysically diverge to infinity. After N round trips, the laser oscillation reaches a **steady-state**, whereby the field inside the resonator no longer changes. To see what the steady-state situation looks like, we note that the field in the beginning of the $(n+1)^{\text{th}}$ round trip is related to the field in the beginning of the $(n)^{\text{th}}$ round trip as:

$$E_{n+1}(\omega) = r_1 r_2 E_n(\omega) e^{(g-\alpha)l_c}\, e^{-i\delta}. \tag{7.6}$$

In steady-state, the field does not change from round trip to round trip, hence $E_{n+1}(\omega) = E_n(\omega)$. Injecting this to the equation above yields

$$E_n(\omega)\left[1 - r_1 r_2 e^{(g-\alpha)l_c}\, e^{-i\delta}\right] = 0. \tag{7.7}$$

The trivial solution $E_n(\omega) = 0$ is not particularly interesting. The non-trivial solution is:

$$\boxed{r_1 r_2 e^{(g-\alpha)l_c}\, e^{-i\delta} = 1.} \tag{7.8}$$

The condition above must hold in steady-state operation. It can be split into two separate conditions – one for phase and one for amplitude.

7.2.1 Phase condition

In steady-state, the per roundtrip phase shift δ must be such that $e^{-i\delta} = 1$. This implies that

$$\delta = 2\pi m. \tag{7.9}$$

But this condition is exactly the same as the resonance condition for the Fabry-Perot cavity:

Steady-state phase condition

Steady-state lasing occurs at wavelengths that are exactly on-resonance with the laser cavity. Thus, the longitudinal modes of the cavity correspond to the longitudinal modes of the laser, and the laser wavelength(s) are given by

$$\lambda_m = \frac{2L}{m}. \tag{7.10}$$

7.2.2 Amplitude condition

Equation (7.8) shows that, in steady-state, the gain and loss must be balanced such that

$$r_1 r_2 e^{(g-\alpha)l_c} = 1. \tag{7.11}$$

This condition should make intuitive sense: for the amplitude (intensity) to be constant (as expected in steady-state), the gain and loss must be fully balanced[74]. However, what is less intuitive is that this condition is actually exactly the same as the condition for threshold gain.

Steady-state gain condition

In steady-state operation, gain is just enough to overcome losses; the gain coefficient must be equal to the threshold gain coefficient

$$g = g_t = \alpha - \frac{1}{l_c} \ln{(r_1 r_2)}. \tag{7.12}$$

Similarly, in steady-state, the population inversion must be equal to the threshold inversion ΔN_t given by Eq. (7.4).

7.2.3 Laser dynamics and gain saturation

For laser operation to begin, we need $g > g_t$, yet in steady-state we must have $g = g_t$. How is this possible? The answer is simple: recalling our analysis of population inversion in a 4-level system (Section 3.12), the inversion saturates for high intensities:

$$\Delta N = \frac{\Delta N_0}{1 + I/I_{\text{sat}}}. \tag{7.13}$$

Seeing as $g = \sigma_{21}\Delta N$, we can similarly conclude that the **gain coefficient saturates with intensity**:

$$g = \frac{g_0}{1 + I/I_{\text{sat}}}, \tag{7.14}$$

where g_0 is the **small-signal gain coefficient**. For laser oscillation to start, we must have $g_0 > g_t$. As the intensity grows, the gain will become saturated, until it eventually reaches the threshold value g_{t} and remains clamped at that value. (The same applies for the population inversion.) This **gain clamping** is easy to understand by considering what happens if the gain would not be clamped (assuming $g_0 > g_{\text{t}}$):

[74] Otherwise the intensity would either decrease or increase.

1. If $g > g_\mathrm{t}$, intensity will grow. This, in turn, reduces the gain. The cycle goes on until g has reduced to g_t.
2. If $g < g_\mathrm{t}$, intensity will be reduced. This, in turn, increases the gain. The cycle goes on until g has increased to g_t.

Dynamics of laser build-up and gain saturation

We can qualitatively examine the dynamical build-up of laser intensity and gain clamping by considering how the intracavity intensity I_n evolves from roundtrip-to-roundtrip when the gain saturates as above. As a simplifying assumption, we assume that the gain reacts instantaneously to changes in intensity. Denoting as I_n and g_n the intensity and gain coefficients in the beginning of roundtrip n, the dynamics can be obtained by iterating the following map:

$$g_n = \frac{g_0}{1 + I_n/I_\mathrm{sat}}, \tag{7.15}$$

$$I_{n+1} = R_1 R_2 I_n e^{2g_n l_c}, \tag{7.16}$$

where $R_{1,2} = r_{1,2}^2$ and we assumed $\alpha = 0$ for simplicity. Furthermore, we assume a homogeneously broadened laser transition, such that the intensities can be understood as total intensities. Figure 32 shows the dynamics of the intensity and gain for parameters listed in the caption. In the beginning, $I \ll I_\mathrm{sat}$, and so the gain coefficient $g_n \approx g_0$. As the intensity grows, the gain starts to saturate. After numerous round trips, a steady state is reached with the intensity remaining constant. In that steady-state, the saturated gain coefficient coincides precisely with the threshold coefficient, as highlighted.

7.2.4 Pumping and slope efficiency

In steady-state, the intracavity intensity can be simply solved from Eq. (7.14) with $g = g_\mathrm{t}$:

$$I_\mathrm{c} = \begin{cases} I_\mathrm{sat}\left(\dfrac{g_0}{g_t} - 1\right) = I_\mathrm{sat}\left(\dfrac{\Delta N_0}{\Delta N_t} - 1\right) & g_0 > g_t \\ 0 & g_0 \le g_t, \end{cases} \tag{7.17}$$

where we highlight the fact that lasing occurs only if $g_0 > g_t$. The threshold gain and inversion only depend on the losses in the active medium and in the resonator. Their small-signal counterparts depend, however, on how hard the atoms are pumped to the excited laser level: $g_0 \propto \Delta N_0 \propto R_\mathrm{p}$, where R_p is the pump rate [see e.g. Section 3.12[75]]. **As a consequence, the laser intensity increases linearly with the pump rate.**

It should be emphasized that Eq. (7.17) represents the intensity inside the laser cavity. The intensity at the laser output is given by

$$I_\mathrm{out} = \frac{1}{2} T_2 I_\mathrm{c}, \tag{7.18}$$

where T_2 is the transmission coefficient of the mirror used for output coupling. The factor $1/2$ accounts for the fact that the intracavity intensity is a combination of intensities moving in opposite directions at any given

[75] Do not confuse R_p with the mirror reflectivity!

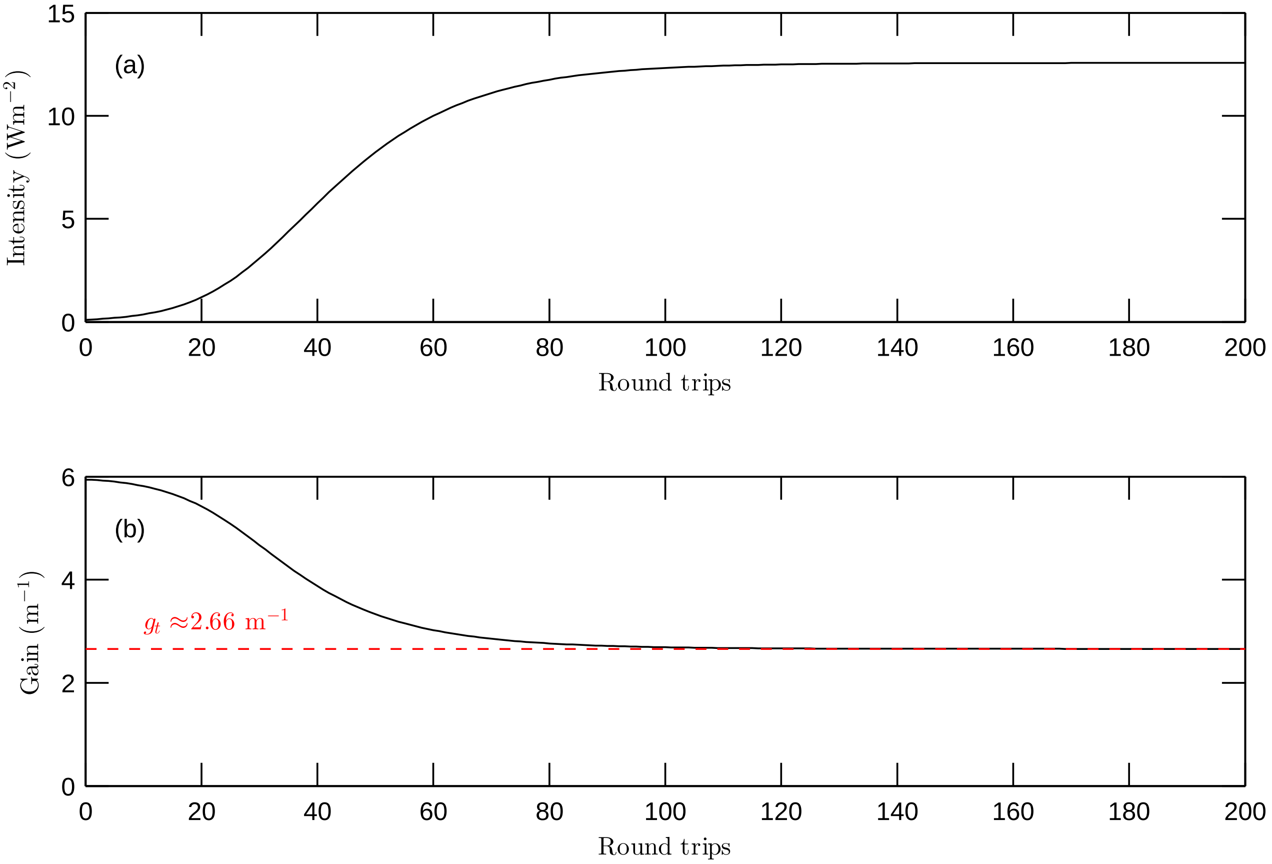

Figure 32: Laser start-up dynamics. In the beginning the intensity is low, and so the small-signal gain leads to exponential amplification. As the intensity grows, the gain saturates, ultimately leading to steady-state operation. In steady-state, the gain is exactly clamped to the threshold value g_t. The data was obtained by iterating Eqs. (7.16) and (7.15) with the following parameters: $l_c = 2$ cm, $R_1 = 0.999$, $R_2 = 0.9$, $g_0 = 6\ \mathrm{m}^{-1}$, $I_{\mathrm{sat}} = 10\ \mathrm{Wm}^{-2}$.

time: in practice one only extracts the half moving towards the output coupler. In experimental situations, one typically measures the output power rather than intensity. Assuming a uniform beam profile with area A, we have $P_{\mathrm{out}} = I_{\mathrm{out}}A$. Figure 33 shows typical dependencies of the laser output power and population inversion on the pump rate (assuming steady-state operation). As can be seen, the intensity is zero until the pump rate is sufficiently high to push the population inversion to its threshold value. The inversion itself increases linearly until that point (recall that for $I = 0$, $\Delta N = \Delta N_0 \propto R_{\mathrm{p}}$). After the threshold inversion has been reached, the laser output power begins to grow linearly with the pump rate, while the inversion remains clamped at its threshold value. The slope of the resulting curve is known as the laser's **slope efficiency** (or differential efficiency). As the name implies, it describes how efficiently pump energy is converted into useful laser energy. The zero-crossing of the curve is the laser threshold pump power (or current).

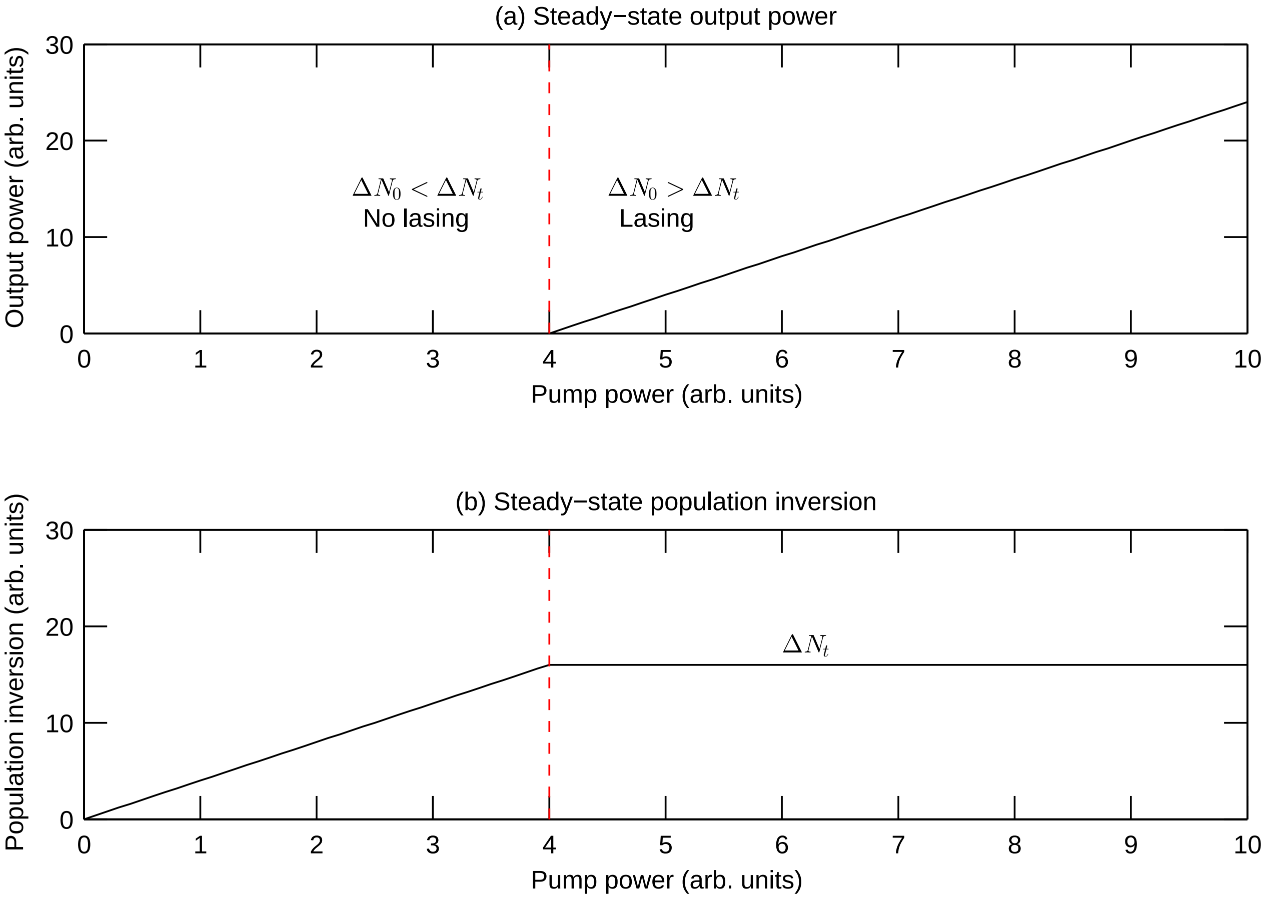

Figure 33: Dependence of laser output power (a) and population inversion (b) on the pump power in steady-state operation. As long as the small-signal inversion $\Delta N_0 < \Delta N_t$, the device is not lasing and hence $I \approx 0$ and $\Delta N = \Delta N_0$. Once the small-signal inversion reaches the threshold inversion, it remains clamped at that value, while the laser output intensity grows linearly.

7.3 Multimode operation

According to the steady-state (phase) condition, lasing can in principle occur at any of the resonator's longitudinal modes. However, in practice only those frequencies that experience gain larger than the threshold will be amplified. The numbers of modes for which this (amplitude) condition is satisfied depends on the resonator free-spectral range and the frequency-dependence of the gain coefficient; the latter being governed by the lineshape function $g_{\text{line}}(\omega - \omega_{21})$ of the laser transition. Because the lineshape function is maximised at the (mean) transition frequency ω_{21}, **lasing will always occur preferentially at the longitudinal mode closest to peak of the lineshape function at ω_{21}.**

Figure 34 illustrates a situation where the unsaturated gain curve is above threshold for several longitudinal modes. All of these modes will be amplified, but whether they remain oscillating in steady-state depends on the line broadening mechanism. The laser is to be in **multimode operation** when several longitudinal modes oscillate simultaneously in steady-state. In contrast, when only a single longitudinal mode is lasing, operation is said to be **single-mode**.

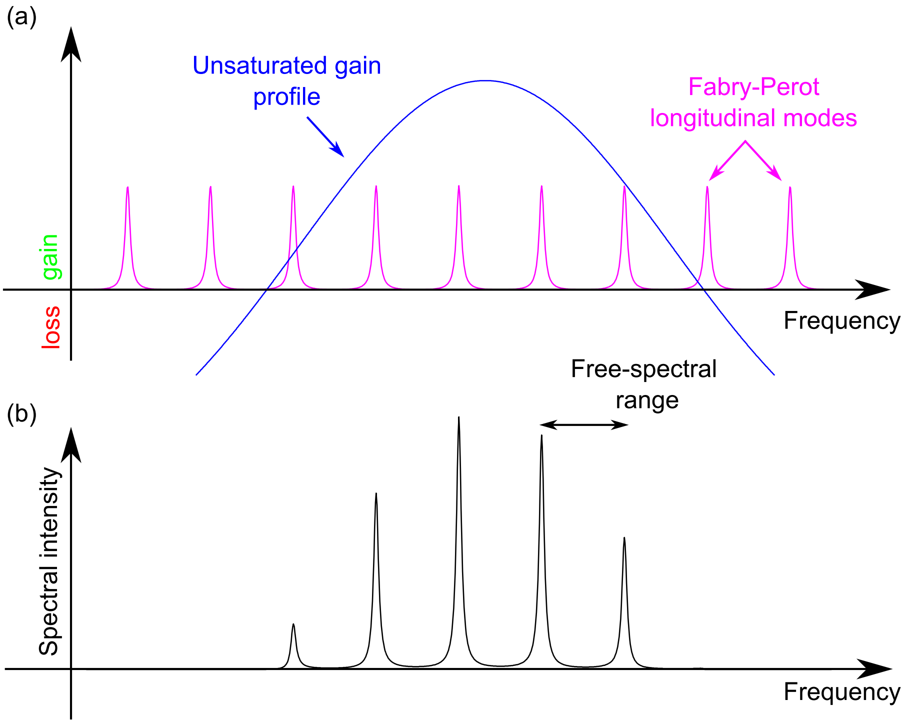

Figure 34: Schematic illustration of the origins of multimode operation. (a) The active medium has a broad spectral line profile, such that several longitudinal cavity modes experience small-signal gain in excess of the threshold gain. (b) Before the gain saturates, all the modes experiencing net gain may be amplified, leading to an output spectrum with multiple emission peaks. What happens in steady-state depends on the lineshape broadening mechanism.

7.3.1 Avoiding multimode operation

Single-mode operation is typically more desirable than multimode operation. This is because the latter does not possess the attractive monochromacity that underpins many laser applications. Furthermore, the different longitudinal modes generally oscillate with random phase relationships, resulting in a chaotic intensity profile in the time domain [see Fig. 38], which is again not very desirable.

There are many ways to force single-mode operation:

1. One can reduce the pump rate, thus reducing the small signal gain g_0, until only the longitudinal mode closest to ω_{21} is above threshold. The price to pay is that the intensity will be small.

2. One can reduce the cavity length, thus pushing the longitudinal modes further apart. [Recall that FSR $= c/(2nL)$.] Also this method can lead to a reduced intensity, as there will be less space for active atoms.

3. One can introduce an additional filter inside the laser cavity (for example, another shorter Fabry-Perot resonator). Through careful design, this filter can be made to increase the losses for all except one longitudinal mode, thus pushing them below threshold.

7.3.2 Homogeneously broadened lasers are (almost) immune to multimode operation

If the laser transition is homogeneously broadened, then all the atoms are identical and they all interact with all the different frequencies present. Accordingly, all of the different longitudinal modes "experience" **the same population inversion ΔN, and all of them contribute to saturating that one population inversion.** However, the gain coefficients associated with the different longitudinal modes will **not** be the same; they will be scaled by the Lorentzian lineshape function $g_{\mathrm{H}}(\omega - \omega_{21})$. The longitudinal mode closest to ω_{21} will experience the largest gain, $g_{\max}$, while all the other modes will experience a smaller gain $g < g_{\max}$ (both saturated and unsaturated).

During the initial build-up of laser radiation, all modes whose small-signal gain coefficient[76] exceeds the gain threshold will be amplified. However, in steady-state the inversion will be saturated such that the gain for the mode closest to ω_{21} is equal to the threshold value, i.e., $g_{\max} = g_{\mathrm{t}}$. If this would not be the case (and $g_{\max} > g_{\mathrm{t}}$), then the intensity of the longitudinal mode closest to ω_{21} would be amplified; this would further saturate the inversion until $g_{\max} = g_{\mathrm{t}}$.

Now if the gain experienced by the longitudinal mode closest to ω_{21} is clamped at threshold ($g_{\max} = g_{\mathrm{t}}$), then it should be clear that all of the other longitudinal modes will have gain below the threshold value (recall that $g \leq g_{\max}$). As a consequence **only the longitudinal mode closest to the (mean) transition frequency will oscillate in steady-state**. Figure 35 illustrates this behaviour.

Spatial hole burning

In practice, it is still possible for homogeneously broadened laser to emit multiple longitudinal modes. This is because the dominant oscillating mode (one closest to ω_{21}) forms a standing wave within the cavity. Population inversion will be saturated only close to the antinodes, where intensity is high; in contrast, the mode will not saturate the inversion close to its antinodes. Thus, it may be possible for another mode, whose standing wave antinodes match the nodes of the first mode, to oscillate. This phenomenon is known as **spatial hole burning**.

7.3.3 Multimode operation for inhomogeneously broadened lasers

Multimode operation more generally occurs for inhomogeneously broadened lasers. This is because different longitudinal modes do not all interact with the same ensemble of atoms. Rather, a longitudinal mode at ω_m only interacts with the sub-set of atoms whose central transition frequency ω_{21} satisfies $|\omega_m - \omega_{21}| < \Delta\omega_{\mathrm{H}}$, where $\Delta\omega_{\mathrm{H}}$ is the natural linewidth of the transition. One can thus qualitatively understand that each longitudinal mode is associated with its own population inversion ΔN_m, which is only being saturated by that longitudinal mode. Accordingly, the dynamics of the m^{th} mode is not greatly influenced by any other longitudinal mode; each longitudinal mode behaves as if it was a totally independent laser.

During laser build-up, all modes whose gain exceeds the threshold value will be amplified. As the intensities of the different modes grow, each of them starts to saturate their own population inversions. Ultimately, a steady-state is reached, where the inversion of each longitudinal mode is independently clamped to the threshold value. In this case, the laser emission spectrum consist of several peaks, each corresponding to a longitudinal mode of the laser [see Fig. 36].

[76] $g_0(\omega) \propto \Delta N_0 g_{\mathrm{H}}(\omega - \omega_{21})$.

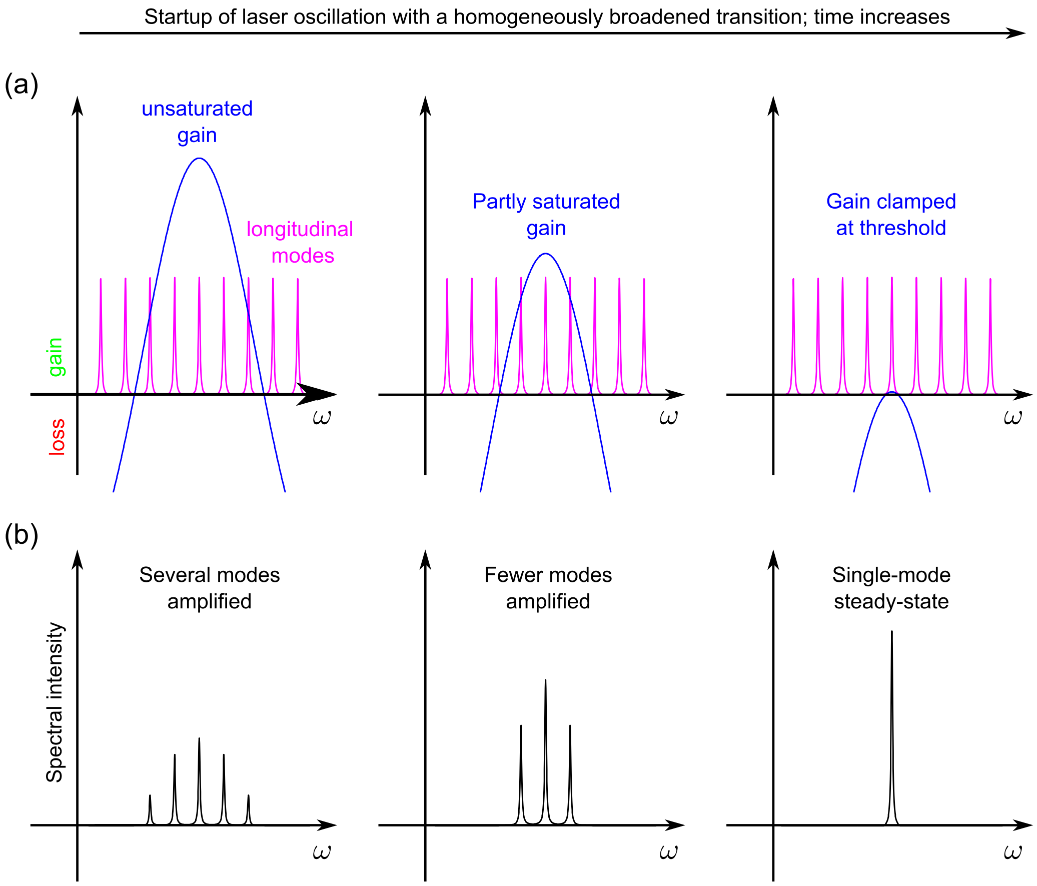

Figure 35: Schematic illustration of spectral startup dynamics in a laser with a homogeneously broadened transition. Initially, the unsaturated gain is able to amplify several longitudinal modes, but as the total intensity grows, the gain becomes saturated and the number of oscillating modes decreases. In steady-state, the gain is clamped to the threshold value at the mode experiencing strongest gain (closes to line peak). (a) Gain and longitudinal modes. (b) Emission spectrum.

7.4 Mode-locking

Mode-locking is a technique that allows for the generation of ultrashort laser pulses with huge peak intensities. In brief, this is achieved by taking a multimode laser and forcing the different longitudinal modes to oscillate with fixed phase relationships. Here we describe briefly the basic physics behind mode-locking.

As we have seen, the spacing between longitudinal modes is

$$\Delta\omega = 2\pi\,\mathrm{FSR} = 2\pi\frac{c}{2nL} = \frac{2\pi}{t_\mathrm{R}}, \tag{7.19}$$

where t_R is the cavity round trip time. The frequency of the p^{th} mode is then simply $\omega_\mathrm{p} = \omega_0 + p\Delta\omega$. Let us now assume a laser that is oscillating at N longitudinal modes, $\omega_{q0}...\omega_{q0} + (N-1)\Delta\omega$, such that the total emission bandwidth $\Delta\omega_\mathrm{G} = (N-1)\Delta\omega$. Furthermore, we assume for simplicity that the intensities of each longitudinal mode are the same, $I_0(\omega) = |E_0|^2$. The laser then has a power spectrum similar to that shown in

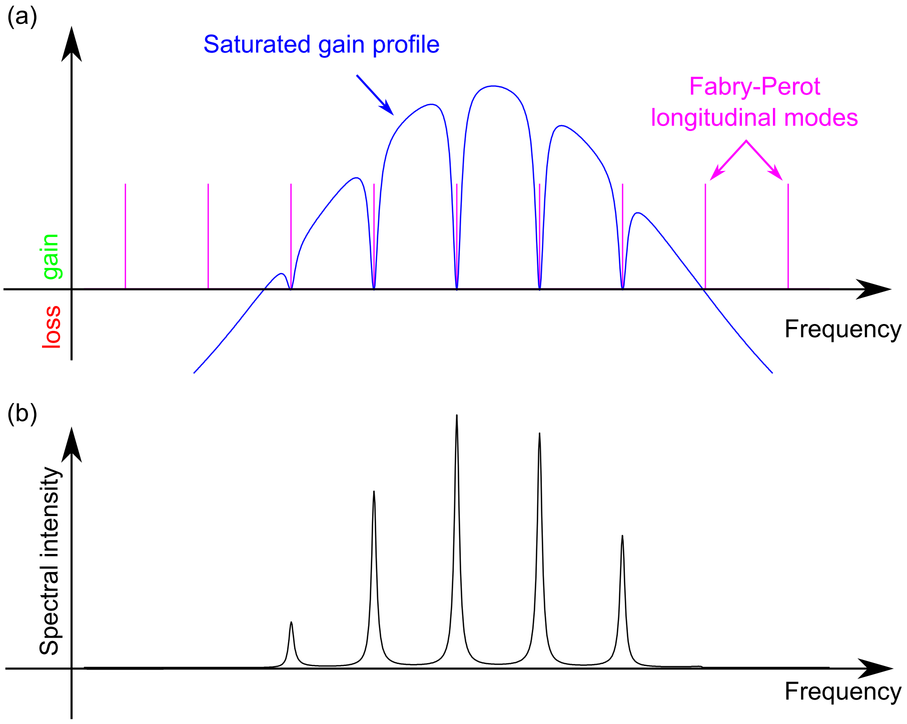

Figure 36: (a) Saturated gain profile during steady-state operation of an inhomogeneously-broadened laser. Different longitudinal modes interact with different atoms associated with different resonance frequencies, leading to "spectral holes" in the inversion and gain. (b) Corresponding steady-state emission spectrum.

Fig. 37.

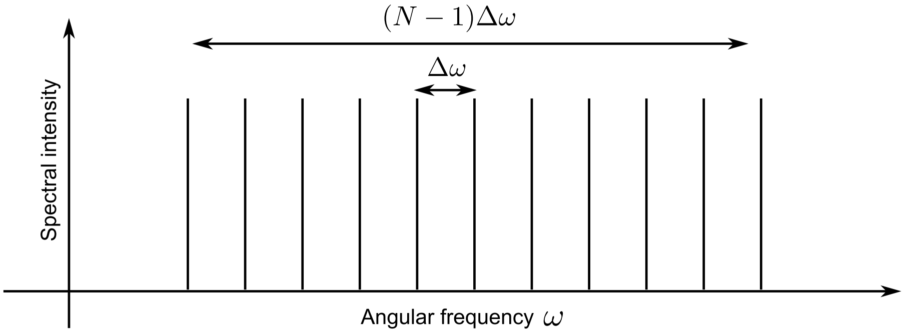

Figure 37: Schematic illustration of spectral intensities of N oscillating modes with identical spectral identities.

We can write the total electric field at the output of the laser as a superposition of the electric fields correspond to all of the longitudinal modes:

$$\boxed{E_{\text{out}}(t) = \sum_{m=q_0}^{q_0+N-1} E_0 e^{i(\omega_0+m\Delta\omega)t+i\phi_m}.} \tag{7.20}$$

For typical multimode lasers, the phases ϕ_m are totally random.[77] In this case, the intensity profile $I_{\text{out}}(t) = |E_{\text{out}}(t)|^2$ is a chaotic pattern that repeats periodically. The period of the pattern corresponds to the cavity round trip time $t_{\text{R}} = 2nL/c$, while the shortest fluctuations have a duration of the order of $2\pi/\Delta\omega_{\text{G}}$, as shown in Fig. 38.

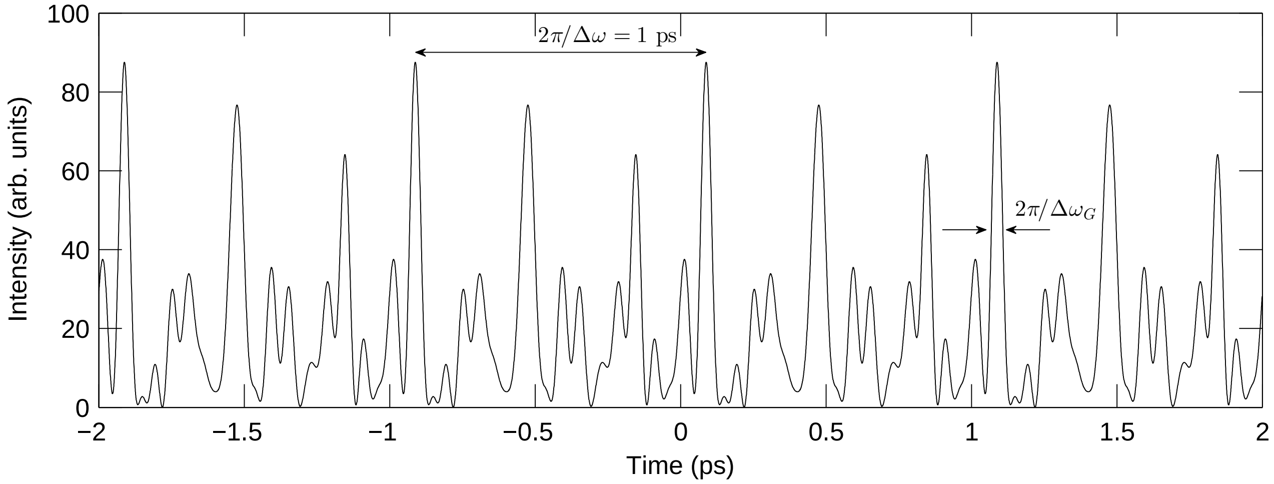

Figure 38: Temporal intensity profile corresponding to the oscillation of 21 longitudinal modes with random phase relationships. The mode spacing is $\Delta\omega = 2\pi \cdot 1$ THz. The shortest temporal features have a duration $\Delta\tau \approx 2\pi/\Delta\omega_G$, where $\Delta\omega_G = (N-1)\Delta\omega$ is the total bandwidth encompassed by the oscillating modes.

If the phases are not random, but instead correlated as $\phi_m = m\phi$, we get something very different. Without loss of generality, we can take $\phi = 0$ such that the output field becomes

$$E_{\text{out}}(t) = E_0 e^{i\omega_0 t} \sum_{m=q_0}^{q_0+N-1} e^{im\Delta\omega t}. \tag{7.21}$$

The sum is a converging geometric progression, and can be readily evaluated. After some algebra, we obtain

$$\boxed{E_{\text{out}}(t) = E_0 e^{i\bar{\omega}t} \left[\frac{\sin\left(N\Delta\omega t/2\right)}{\sin\left(\Delta\omega t/2\right)}\right],} \tag{7.22}$$

[77] The phases can even randomly fluctuate with time.

where $\bar{\omega}$ corresponds to the average frequency across the oscillating longitudinal modes:

$$\bar{\omega} = \frac{\omega_{q0} + \omega_{q0} + (N-1)\Delta\omega}{2} \tag{7.23}$$

The field described by Eq. (7.22) corresponds to a wave with a *carrier* frequency $\bar{\omega}$ multiplied by an envelope. The corresponding intensity profile is simply

$$I_{\text{out}}(t) = |E_0|^2 \left[\frac{\sin^2(N\Delta\omega t/2)}{\sin^2(\Delta\omega t/2)}\right]. \tag{7.24}$$

A typical intensity profile is shown in Fig. 39, obtained using the same parameters as in Fig. 38 (except that we assume the longitudinal modes to oscillate in-phase). As can be seen, the intensity profile now corresponds to a perfectly periodic train of pulses, separated by the cavity round trip time t_{R}. One should understand that there is a single pulse circulating in the cavity, bouncing back and forth between the two mirrors, while a part of the pulse is output after each round trip.

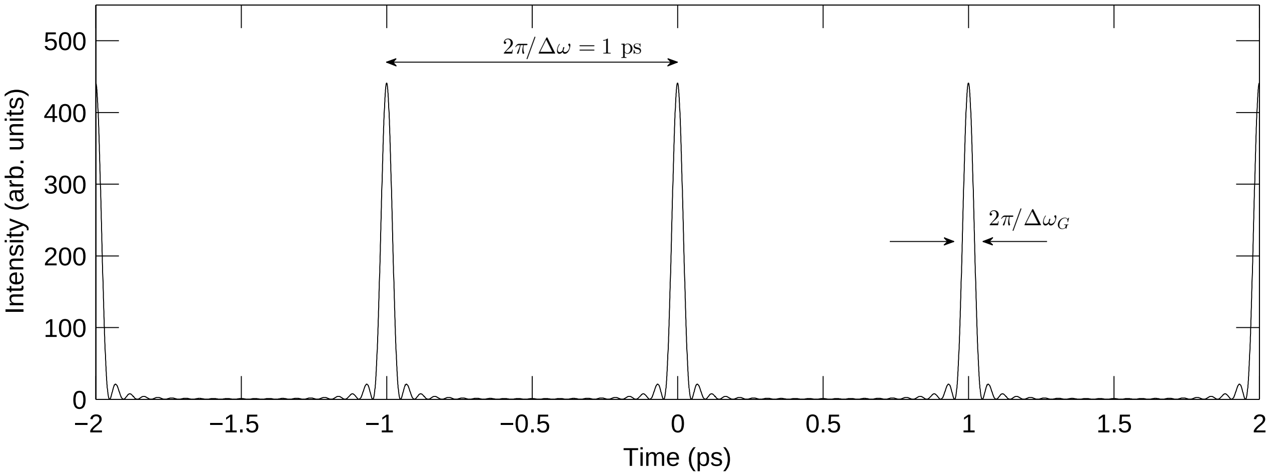

Figure 39: Temporal intensity profile corresponding to the oscillation of 21 longitudinal modes with fixed phase relationships. Other than the phases, the parameters are the same as in Fig. 38.

We can evaluate the peak intensity of the pulses by noting that, for $t \approx 0$, we can approximate

$$I_{\max} \approx |E_0|^2 \left[\frac{N^2\Delta\omega^2 t^2/4}{\Delta\omega^2 t^2/4}\right] = |E_0|^2 N^2. \tag{7.25}$$

Thus, the peak intensity is N times larger than the average intensity of N modes with random phases ($N|E_0|^2$). We can also estimate the duration of the pulses by looking for the time $t = \tau_p$ when the numerator in Eq. (7.22) becomes zero, i.e.:

$$\frac{N}{2}\Delta\omega\tau_{\text{p}} = \pi. \tag{7.26}$$

Solving for the duration yields

$$\boxed{\tau_{\text{p}} = \frac{2\pi}{N\Delta\omega} \approx \frac{2\pi}{\Delta\omega_{\text{G}}}.} \tag{7.27}$$

Thus, we see that the larger the frequency bandwidth covered by the oscillating modes, the shorter the pulses. Or alternatively, short pulses require large bandwidth, in agreement with Fourier limits, Heisenberg's uncertainty and whatnot.

Real lasers do not have flat spectra like that in Fig. 37, and so discrepancies arise in terms of the precise formulae quoted above. However, the general conclusions remain the same. For example, mode-locked lasers whose emission spectrum is Gaussian typically create pulses whose temporal shape is also Gaussian, with a full-width at half-maximum duration

$$\tau_{\mathrm{p}} = \frac{2.77}{\Delta\Omega_{\mathrm{G}}}. \tag{7.28}$$

Ti:Sapphire example

Ti:Sapphire lasers have a very broad gain bandwidth of about $\Delta\omega_{\mathrm{G}} \approx 2\pi \times 15$ THz. They are typically implemented with laser cavities that are about 3 m long, and thus have an FSR ≈ 80 MHz. The number of oscillating longitudinal modes is then

$$N = \frac{\Delta\omega_{\mathrm{G}}}{2\pi\,\mathrm{FSR}} \approx 187500, \tag{7.29}$$

resulting in pulses as short as

$$\tau_{\mathrm{p}} = \frac{2.77}{\Delta\omega_{\mathrm{G}}} \approx 30 \text{ fs}. \tag{7.30}$$

7.4.1 Techniques for mode-locking

Mode-locking requires that all of the longitudinal modes oscillate with fixed phase relationships, and there are several techniques to achieve this. The easiest way to understand their operation is perhaps not to consider the locking of the different longitudinal modes, but rather the evolution of the pulse that must be circulating back and forth in the laser cavity. In particular, **mode-locking can be generically achieved by placing elements in the laser cavity that favour the formation of ultrashort pulses and hinder the formation of "quasi"-continuous wave radiation.** Based on the precise operation of these elements, the mode-locking is said to be either **active** or **passive**.

Active mode-locking

In active mode-locking, an element is placed in the cavity that modulates at least one of the cavity parameters at a frequency determined by the user. For example, an amplitude modulator (e.g. a shutter) can be used to induce time-dependent loss. If the shutter is synchronised with the round trip time, it promotes the formation of a pulse that hits the shutter only when the shutter is open.

Passive mode-locking

In passive mode-locking, an element whose transmission depends on the intensity of light is inserted in the cavity. For example, a saturable absorber is an element whose absorption saturates with intensity:

$$\alpha = \frac{\alpha_0}{1 + I/I_{\rm sat}}. \tag{7.31}$$

High intensity pulses experience low losses and their formation is thus favoured. In contrast, any small radiation components will experience large losses, and their formation thus hindered.

Problems

7.1 Consider a Fabry-Perot laser whose active medium has a length $l_{\rm c} = 0.1$ m, a small-signal gain coefficient $g_0 = 1$ m^{-1}, and saturation intensity $I_{\rm sat} = 10$ kWm^{-2}. One of the mirrors of the Fabry-Perot cavity is perfectly reflecting ($R = 1$), while the other acts as the output coupler, reflecting 95 % of light and transmitting 5 %. You can assume that there is no scattering losses in the active medium, such that $\alpha = 0$.

(a) Calculate the threshold gain of the laser cavity.

(b) Based on your calculation in (a), explain whether steady-state laser oscillation is expected.

(c) Calculate the intracavity intensity in steady-state.

(d) Assuming that the laser beam (diameter 10 mm) has a uniform intensity distribution, calculate the laser's output power in steady-state.

7.2 The cross-section σ_{21} of a laser transition is $\sigma_{21} = 2.8 \times 10^{-19}$ cm^2 and the length of the active medium is $L = 2.5$ cm. Pumping results in a small-signal population inversion $\Delta N_0 = 7 \times 10^{17}$ cm^{-3}, and the saturation intensity is $I_{\rm sat} = 100$ Wcm^{-2}. The active medium is placed inside a Fabry-Perot resonator with mirror reflectivities $R_1 = 0.99$ and $R_2 = 0.9$.

(a) What is the maximum value allowed for the internal losses α for the laser to oscillate?

(b) Assuming $\alpha = 0$, what is the gain during steady-state operation?

(c) Further assume that the laser beam (diameter 10 mm) has a uniform intensity distribution, and that the output is extracted with mirror R_2. Calculate the output power in steady-state.

7.3 An active medium with an internal loss coefficient $\alpha = 0.1$ cm^{-1}, refractive index $n = 1.5$, and length $l_{\rm c} = 4$ cm is placed inside a Fabry-Perot resonator with mirror reflectivities $R_1 = 0.99$ and $R_2 = 0.9$. The laser transition has a central frequency $\omega_{21} = 2\pi \times 280$ THz, linewidth $\Delta\omega = 10$ GHz and a spontaneous emission lifetime $\tau_{\rm sp} \approx 1.3$ ms. Estimate the threshold inversion for laser oscillation.

7.4 Consider a multimode laser that is oscillating at N longitudinal modes with equal amplitudes. Show that, under mode-locked operation, where all the longitudinal modes oscillate in-phase, the temporal intensity profile emitted by the laser is given by Eq. (7.24).

8 Ray transfer matrices and resonator stability

In the preceding sections, we have examined resonators and the physics that underpins the operation of simple Fabry-Perot lasers. In the following sections, we will now start to examine how laser light propagates and behaves outside (and inside) of the resonator. First, however, in this section we present a brief summary of matrix methods for ray tracing, as this formalism turns out to be very useful for describing how actu al laser beams propagate. Furthermore, the use of ray transfer matrices allow us to answer a simple question which we have ignored so far: what kind of mirrors should we use to build our Fabry-Perot resonator. Specifically, it should be clear that not just all mirrors will do. For example, consider a ray of light bouncing back and forth between two convex mirrors facing each other. It should be clear that the ray cannot remain inside the resonator for very long, but instead, will quickly escape to a position where the mirrors no longer extend. Indeed, it turns out that only some mirror pairs allow us to construct *stable* Fabry-Perot resonators where light can persist forever.

8.1 Ray optics

As seen in Section 4, light is an electromagnetic wave consisting of two vector fields (electric and magnetic fields). In some situations, it is necessary to consider all the vectorial components of those fields. However, in many situations light polarization remains constant, and in this case it is sufficient to only consider a single scalar wavefunction. This approximate way of treating light is known as (scalar) *wave optics.*

Wave optics details how the wave character of light influences the behaviour and observation of light, encompassing effects such as interference and diffraction. However, these wave phenomena only become important when light interacts with objects whose dimensions are comparable with the wavelength of light. For objects much larger than the wavelength, the behaviour of light can be described by *rays* that obey a set of geometrical rules. This model of light is called *ray optics*, and it can be understood as the limit of wave optics when the wavelength is very small. It is worth emphasising that the rays in ray optics can be understood as the wave vectors of plane EM waves.

Ray optics is the simplest description of light. It is very much an approximation, but a useful one, for it allows us to straightforwardly describe how different optical components (e.g. lenses, mirrors) work and how light behaves at interfaces. It is worth emphasising that ray optics is founded on a set of (more or less) empirically derived geometrical laws (e.g. law of reflection, Snell's law); the first principles origins of those laws can only be derived through rigorous analysis of Maxwell's equations.

8.2 Ray transfer matrices

Ray transfer matrix analysis is a type of ray tracing technique that is used in the design of some optical systems. It involves the construction of a ray transfer matrix which describes the optical system; tracing of a light path through the system can then be performed by multiplying this matrix with a vector that represents the light ray.

In ray transfer matrix analysis, a light ray at a certain position is characterized by two parameters [see Fig. 40]:

$$\boxed{\begin{aligned} r &\equiv \text{Distance from optical axis}\\ r' &\equiv \text{Angle between ray and optical axis}\end{aligned}}$$

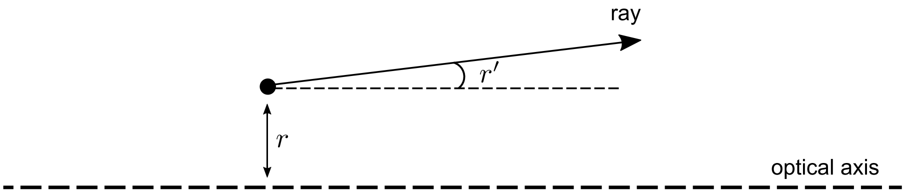

Figure 40: In ray transfer analysis, a ray at a given position along the optical axis is characterised by the ray's displacement from the optical axis, r, and the angle between the ray and the optical axis, r'.

The ray is then described by a 2×1 vector:

$$\vec{\mathbf{r}} = \begin{bmatrix} r \\ r' \end{bmatrix} \tag{8.1}$$

We assume that the light is propagating "paraxially", i.e., that the angle between the ray and the optical axis is small. In this case we have $r' \ll 1$, and therefore

$$\tan(r') \approx \sin(r') \approx r'. \tag{8.2}$$

The sign convention is such that an angle corresponding to an anti-clockwise rotation (from the optical axis) is positive, whilst a clockwise rotation is negative.

The ray vector $\vec{\mathbf{r}}_1$ at some location z_1 can be found by multiplying the ray vector $\vec{\mathbf{r}}_0$ at some initial position z_0 with an appropriate ray transfer matrix. Different optical components are described by different matrices, and they are referred to as ABCD matrices:

$$\mathrm{M} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{8.3}$$

ABCD matrix for straight propagation

Consider the straight propagation of a ray over a distance d, as schematically illustrated in Fig. 41(a). It should be clear that the ray transforms as

$$\begin{aligned} r_1 &= r_0 + \tan(r_0')d \approx r_0 + r_0'd \\ r_1' &= r_0' \end{aligned}$$

Accordingly, the ray transfer can be written as a simple matrix product $\vec{\mathbf{r}}_1 = \mathbf{M}\vec{\mathbf{r}}_0$ where

$$\mathrm{M} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \tag{8.4}$$

ABCD matrix for a thin lens

To find the ABCD matrix for a thin lens with focal length f, we consider two rays: one that is parallel to the optical axis, and another one that starts a focal distance away from the lens [see Fig. 41(b)]. First, it should be clear that passing through the lens does not affect the ray's displacement, hence $r_1 = r_0$. For our ABCD matrix, this implies A $= 1$ and B $= 0$.

Consider now the ray that is parallel to the optical axis. The lens will change the angle of the ray such that the ray crosses the optical axis at the focal point. Thus, we have

$$r_1' \approx \tan(r_1') = -\frac{r_0}{f}, \tag{8.5}$$

where the minus sign comes from the sign conventions: the angle is negative whilst f for a converging lens is positive. From this, we see that, for our ABCD matrix, $C = -1/f$.

Considering finally the ray that originates from the focal point, we know that right before the ray hits the mirror, its displacement will be $r_0 = \tan(r_0')f \approx r_0'f$. On the other hand, after the passing the lens, the ray's angle will be zero. Hence

$$r_1' = Cr_0 + Dr_0' = -\frac{1}{f}r_0 + \frac{D}{f}r_0 = 0. \tag{8.6}$$

From this, we see that $D = 1$. Thus, the ABCD matrix for a thin lens is

$$\mathrm{M} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix} \tag{8.7}$$

ABCD matrix for a mirror

When dealing with mirrors in the framework of ray transfer matrices, we always ignore the change in direction of propagation. Rather, we "fold" the ray by taking the mirror image of the output ray [see Fig. 41]. Considering a planar mirror [Fig. 41(c)], it should be clear that the mirror changes neither the displacement nor the angle. Thus, the ABCD matrix of a planar mirror corresponds to the identity matrix:

$$\mathrm{M} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{8.8}$$

Considering then the more general case of a spherical mirror [Fig. 41(d)], we note that such mirrors focus rays of light with a focal length $f = 2/R$, where R is the curvature of the mirror. (The sign convention is such that $R > 0$ ($R < 0$) for a concave (convex) mirror.) Because we ignore the change in direction by folding the ray, it should be evident that the ABCD matrix for a spherical mirror is the same as for a thin lens:

$$M = \begin{bmatrix} 1 & 0 \\ -\dfrac{2}{R} & 1 \end{bmatrix} \tag{8.9}$$

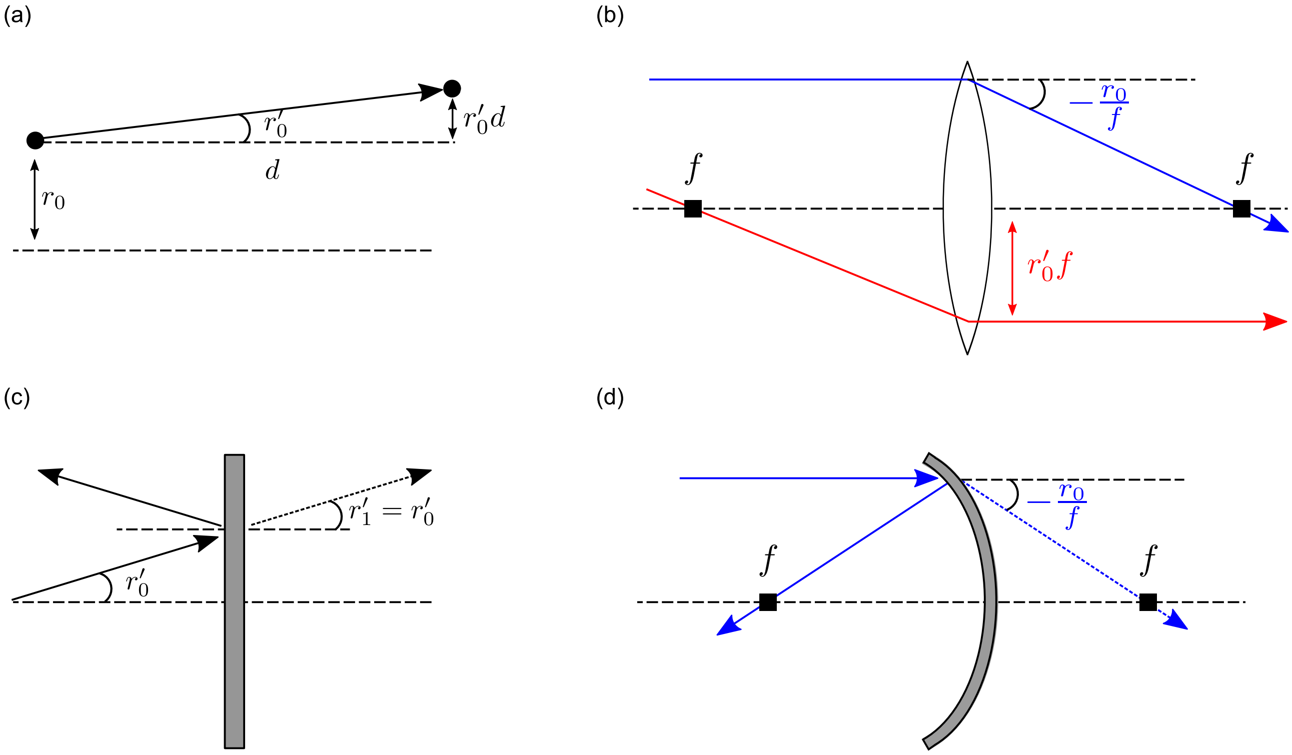

Figure 41: Schematic illustrations for deriving ray transfer matrices for (a) free-space propagation through a distance d, (b) thin lens, (c) planar mirror, and (d) spherical mirror with focal length $f = 2/R$, where R is the mirror's radius of curvature. Note that for the mirrors, we fold the ray by taking the mirror image of the true output ray, thus avoiding the awkwardness of changing direction.

To summarise, arbitrary light rays can be represented as vectors $\vec{\mathbf{r}} = [r, r']^{\mathrm{T}}$, while the effect of optical components can be accounted for by simple matrix multiplications. Furthermore, propagation of rays through complex systems consisting of multiple optical elements can also be simply achieved by just multiplying together the the ray matrices for all the different elements:

$$\boxed{\vec{\mathbf{r}}_n = \mathrm{M}_n \mathrm{M}_{n-1} \cdots \mathrm{M}_1 \vec{\mathbf{r}}_0.} \tag{8.10}$$

Of course, because matrix multiplications do not commute, one must pay attention to the order in which the different matrices are multiplied. In the above equation, for example, M_1 would correspond to the first element encountered by the ray, whilst M_n would be the last element in the sequence.

8.3 Resonator stability

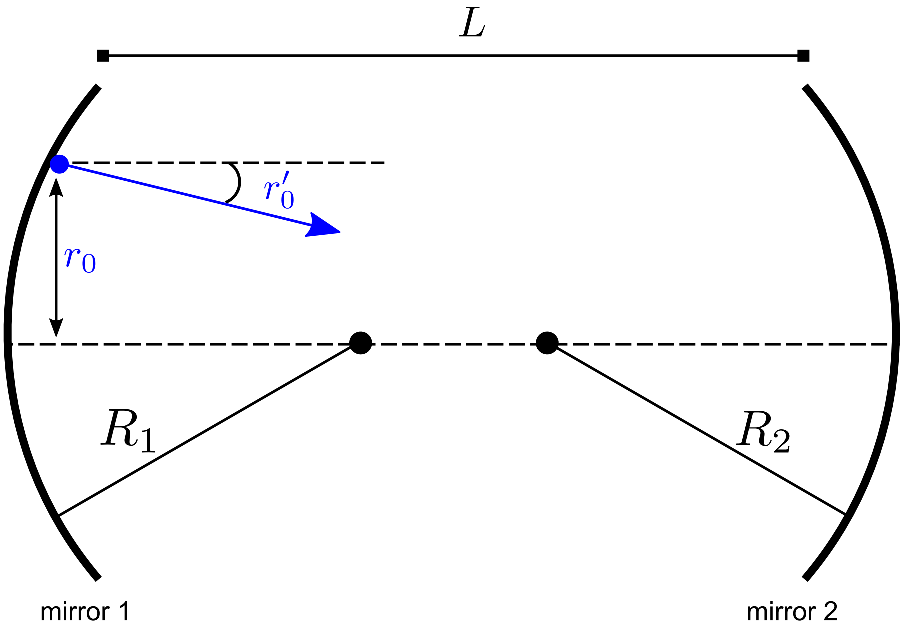

Figure 42: Schematic illustration of ray matrix analysis applied to examine resonator stability.

A particularly useful application of ray matrix analysis is that it allows for straightforward assessment of laser resonator stability, i.e., whether a given mirror configuration allows for rays to persistently remain confined to the cavity. Let us consider an arbitrary (laser) resonator [c.f. Fig. 42] consisting of two mirrors with curvatures R_1 and R_2 (not to be confused with mirror reflectivities) spaced by length L. Considering a ray $\vec{\mathbf{r}}_0$ that is just reflected from the surface of mirror 1, we can obtain an expression for the ray $\vec{\mathbf{r}}_1$ after one full round trip:

$$\vec{\mathbf{r}}_1 = \mathrm{M}_4\mathrm{M}_3\mathrm{M}_2\mathrm{M}_1\vec{\mathbf{r}}_0. \tag{8.11}$$

Here, M_2 and M_4 correspond to reflections from mirrors 2 and 1, respectively, whilst $\mathrm{M}_3 = \mathrm{M}_1$ each describe propagation in free-space through a distance L. Writing out the ABCD matrices for the different components, we have

$$\vec{\mathbf{r}}_1 = \begin{bmatrix} 1 & 0 \\ -\frac{2}{R_1} & 1 \end{bmatrix} \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{2}{R_2} & 1 \end{bmatrix} \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \vec{\mathbf{r}}_0. \tag{8.12}$$

Performing the matrix multiplications yields

$$\vec{\mathbf{r}}_1 = \begin{bmatrix} 1 - \frac{2L}{R_2} & 2L - \frac{2L^2}{R_2} \\ -\frac{2}{R_1} + \frac{4L}{R_1 R_2} - \frac{2}{R_2} & -\frac{2L}{R_1} + (1 - \frac{2L}{R_1})(1 - \frac{2L}{R_2}) \end{bmatrix} \vec{\mathbf{r}}_0 \tag{8.13}$$

This equation can be written in a simpler form as

$$\vec{\mathbf{r}}_1 = \mathrm{M}_{\text{cavity}}\vec{\mathbf{r}}_0, \tag{8.14}$$

where $\mathrm{M}_{\text{cavity}}$ is the ABCD matrix describing one full cavity round trip. Its elements are

$$\boxed{\begin{aligned} A &= 1 - \frac{2L}{R_2} && (8.15)\\ B &= 2L - \frac{2L^2}{R_2} && (8.16)\\ C &= -\frac{2}{R_1} + \frac{4L}{R_1R_2} - \frac{2}{R_2} && (8.17)\\ D &= -\frac{2L}{R_1} + (1 - \frac{2L}{R_1})(1 - \frac{2L}{R_2}) && (8.18)\end{aligned}}$$

A given resonator is “stable” only if systematically $|r_{n+1}| < |r_n|$. If this is not the case, the ray’s displacement r_n increases from round trip to round trip, implying that the ray will eventually escape from the cavity. The question of whether a given resonator is stable can be cast into an eigenvalue problem. Specifically, an arbitrary ray $\vec{\mathbf{r}}_0$ can be expressed as a linear superposition of the eigenvectors $\vec{\mathbf{r}}_\pm$ of the cavity ABCD matrix:

$$\vec{\mathbf{r}}_0 = a\vec{\mathbf{r}}_+ + b\vec{\mathbf{r}}_-, \tag{8.19}$$

After n round trips, the ray vector will have transformed into

$$\vec{\mathbf{r}}_n = (\mathrm{M}_{\text{cavity}})^n[a\vec{\mathbf{r}}_+ + b\vec{\mathbf{r}}_-]. \tag{8.20}$$

Using now the fact that $\vec{\mathbf{r}}_\pm$ are eigenvectors of $\mathrm{M}_{\text{cavity}}$, such that $\mathrm{M}_{\text{cavity}}\vec{\mathbf{r}}_\pm = \lambda_\pm\vec{\mathbf{r}}_\pm$, where $\lambda_\pm$ are the corresponding eigenvalues, Eq. (8.20) can be simplified into

$$\vec{\mathbf{r}}_n = a\lambda_+^n\vec{\mathbf{r}}_+ + b\lambda_-^n\vec{\mathbf{r}}_-. \tag{8.21}$$

It should be clear that, if $|\lambda_\pm| > 1$, the ray’s displacement will diverge as $n \to \infty$. Thus, the resonator stability requires that the eigenvalues satisfy:

$$\boxed{|\lambda_\pm| \leq 1.} \tag{8.22}$$

The eigenvalues of the cavity matrix can be found from the usual equation $\det(\mathrm{M}_{\text{cavity}} - \lambda I) = 0$. Using the (easily verifiable[78]) fact that the determinant $\det(\mathrm{M}_{\text{cavity}}) = 1$, one can derive:

$$\begin{aligned}\lambda_\pm &= \frac{(A+D) \pm \sqrt{(A+D)^2 - 4}}{2} && (8.23)\\ &= m \pm \sqrt{m^2 - 1}, && (8.24)\end{aligned}$$

where $m = (A+D)/2$. It is easy to see that $|\lambda_\pm| > 1$ iff $|m| > 1$. Accordingly, the condition for resonator stability is $|m| \leq 1$, or equivalently

$$\boxed{\left|\frac{A+D}{2}\right| \leq 1.} \tag{8.25}$$

[78] $\det(\mathrm{M}_{\text{cavity}}) = \det(\mathrm{M}_1)\det(\mathrm{M}_2)\det(\mathrm{M}_3)\det(\mathrm{M}_4)$

Substituting the matrix elements from Eqs. (8.15) and (8.18) yields, after some algebra, the following condition:

$$\boxed{\begin{aligned} &0 \leq \left(1-\frac{L}{R_1}\right)\left(1-\frac{L}{R_2}\right) \leq 1 && (8.26)\\ &0 \leq g_1 g_2 \leq 1, && (8.27)\end{aligned}}$$

where we defined the coefficients $g_{1,2} = 1 - L/R_{1,2}$.

Equations Eq. (8.26) and (8.27) can be used to deduce whether a given resonator (with given mirror curvatures $R_{1,2}$ and separation L) is stable or not. For example, the stability condition immediately reveals that a cavity made out of two convex mirrors (with $R_{1,2} < 0$) is always unstable ($g_1 g_2 > 1$). Of course, this result is expected based on simple geometrical considerations. Figure 43 shows a map of stability regimes in the plane formed by the parameters g_1 and g_2, together with configurations illustrative of the different regimes.

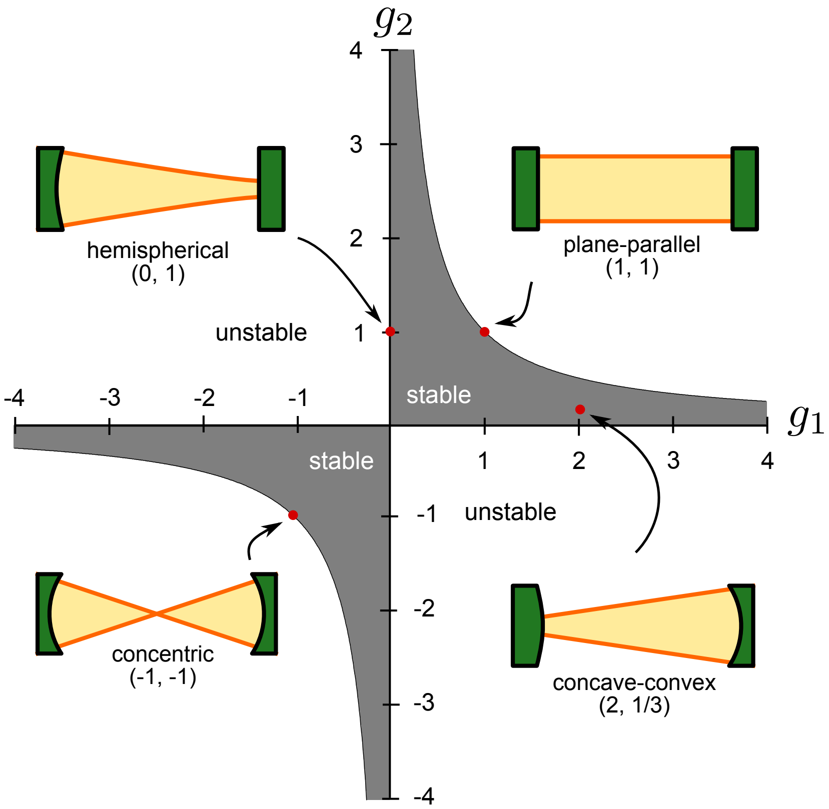

Figure 43: Map of resonator stability as a function of $g_1 = 1 - L/R_1$ and $g_2 = 1 - L/R_2$. Gray areas correspond to stable resonator configurations, while white areas are unstable. Four examples of stable resonators are shown. These example configurations are from a figure by FDominec (Own work) [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons.

It is worth emphasizing that the stability condition derived above is an approximation (in keeping with the fact that ray optics is an approximation). However, the approximation is very good. In fact, a more rigorous analysis based on wave optics yields the exact same stability condition, expect that the limits become strict inequalities: $0 < g_1 g_2 < 1$. This implies that, whilst the stability condition derived using ray optics predicts that a plane-parallel mirror, comprised of two planar mirrors with $R_{1,2} = \infty$, is stable, this is not the case in reality. Rather, at least one of the mirrors has to be curved.

9 Paraxial wave equation and Gaussian beams

So far we have not discussed the transverse profile of real laser beams in any great detail. In Section 4, we derived the plane wave solutions for Maxwell's equations, and found that they can explain many of the salient characteristics attributed to light (and other forms of EM waves). However, it was also noted that plane waves cannot exist in reality. One of the reasons was that their transverse profile corresponds to an infinite plane, implying plane waves to carry infinite energy. This is clearly unphysical, and indeed, real beams have a finite transverse profile[79]. In this section, we discuss the transverse profiles of real laser beams, and describe how they propagate through different optical components.

9.1 Paraxial wave equation

Our starting point is the EM wave equation:

$$\nabla^2 \vec{\mathbf{E}} - \mu_0 \varepsilon \frac{\partial^2 \vec{\mathbf{E}}}{\partial t^2} = 0. \tag{9.1}$$

We assume a linearly polarized field $\vec{\mathbf{E}} = E\hat{\mathbf{x}}$ and that the field does not interact with birefringent materials (such that the polarization stays constant). In this case, we can simply consider the scalar amplitude $E(x, y, z, t)$. Furthermore, we assume the EM wave to be monochromatic, such that

$$E(x, y, z, t) = A(x, y, z)e^{i\omega t}. \tag{9.2}$$

Injecting this ansatz into our wave equation yields

$$\nabla^2 A + k^2 A = 0, \tag{9.3}$$

where we used the dispersion relation $k = \omega^2 \mu_0 \varepsilon$. This equation is known as the scalar **Helmholtz** equation. It is a time-independent wave equation, obtained by separating the spatial and temporal dependencies.

Next, we assume that the the field is propagating "principally" in the z-direction. Recalling how arbitrary fields can be represented as superpositions of plane waves with different wave vectors, this assumption implies that the dominant plane wave contribution to the field comes from a plane wave with wave vector $\vec{\mathbf{k}} = k\hat{\mathbf{z}}$. We can then write the field's spatial dependence as:

$$A(x, y, z) = \psi(x, y, z)e^{-ikz}, \tag{9.4}$$

where $\psi(x, y, z)$ is an *envelope* that evolves slowly (relative to kz) and $k = n\omega/c = 2\pi/\lambda$ is the usual wave number. Injecting this decomposition into the Helmholtz equation yields

$$\nabla_\perp^2 \psi + \frac{\partial^2 \psi}{\partial z^2} - 2ik\frac{\partial \psi}{\partial z} = 0, \tag{9.5}$$

where $\nabla_\perp^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is a transverse Laplacian. Finally, we invoke the **paraxial approximation**: the field is propagating almost paraxially along z. Under this approximation, ψ evolves slowly with z. (Indeed, the

[79] Note that transverse profile here refers to the spatial distribution of the beam in a plane perpendicular to the direction of propagation (e.g. $x - y$ plane). This is the "spot" you see when a lases beam is shined on a wall.

main z evolution has already been factored out with the $\exp(-ikz)$ term.) The meaning of "slow" evolution is that, over a distance of one wavelength, the field does not change much. In particular, we have:

$$|\delta\psi| = \left|\frac{\partial\psi}{\partial z}\right|\lambda \ll |\psi|, \tag{9.6}$$

which implies that

$$\left|\frac{\partial^2\psi}{\partial z^2}\right| \ll \frac{1}{\lambda}\left|\frac{\partial\psi}{\partial z}\right| \sim 2k\left|\frac{\partial\psi}{\partial z}\right|. \tag{9.7}$$

Thus, $|\partial^2\psi/\partial^2 z| \ll 2k|\partial\psi/\partial z|$, allowing us to drop the second term in Eq. (9.5). This yields the so-called **paraxial wave equation**:

$$\boxed{\nabla_\perp^2\psi - 2ik\frac{\partial\psi}{\partial z} = 0.} \tag{9.8}$$

Paraxial beam propagation

We can obtain more insights by writing the paraxial wave equation in "beam propagation" form:

$$\frac{\partial\psi}{\partial z} = -\frac{i}{2k}\nabla_\perp^2\psi \tag{9.9}$$

$$= -\frac{i}{2k}\left(\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2}\right) \tag{9.10}$$

From this form, it should be evident that the paraxial wave equation describes how a beam with a given transverse profile $\psi(x, y, z)$ evolves as a function of z. Specifically, if we know what the transverse field profile is at some longitudinal coordinate $z_0 = 0$, we can use the paraxial wave equation to deduce what the profile will be like at another coordinate z_1. As we shall see, the paraxial wave equation indeed describes effects such as diffraction, i.e., the natural tendency of light beams to broaden as they propagate.

It is worth highlighting that the paraxial wave equation is completely analogous to some fairly famous equations encountered in other physical contexts. Indeed, one should recognise that the paraxial wave equation has the exact same form as the time-dependent Schrödinger equation in empty space (the longitudinal coordinate z representing time t of the Schrdöinger equation). In addition, an analogous equation can be found to describe propagation of light pulses in optical fibres and waveguides. These analogies should make clear that the ability to understand and solve the paraxial wave equation is extremely valuable, not only because it allows us to infer how real laser beams behave, but also because identical equations can be found in many other contexts.

9.2 General solution to the paraxial wave equation

As mentioned above, if we know the transverse profile $\psi(x, y, 0)$ at some $z_0 = 0$, we can use the paraxial wave equation to obtain the transverse profile $\psi(x, y, z)$ at any other location z. This is accomplished by using

Fourier transforms. Specifically, let us consider the two-dimensional Fourier transform of $\psi(x, y, 0)$. Referring to Subsection 4.5.1, the transverse profile $\psi(x, y, 0)$ can be written as an inverse Fourier transform:

$$\psi(x, y, 0) = \iint\limits_{-\infty}^{\infty} \widetilde{\psi}(k_x, k_y, 0) e^{-i(k_x x + k_y y)} dk_x \, dk_y, \tag{9.11}$$

where the spectral coefficients are obtained from the forward Fourier transform $\widetilde{\psi} = \mathcal{F}[\psi]$:

$$\widetilde{\psi}(k_x, k_y, 0) = \frac{1}{(2\pi)^2} \iint\limits_{-\infty}^{\infty} \psi(x, y, 0) e^{i(k_x x + k_y y)} dx \, dy. \tag{9.12}$$

It is easy to verify that Fourier transforms satisfy a derivative property, which in our notation reads:

$$\boxed{\begin{aligned} \mathcal{F}\left[\frac{\partial \psi}{\partial x}\right] &= -ik_x \mathcal{F}[\psi] \qquad (9.13) \\ \mathcal{F}\left[\frac{\partial \psi}{\partial y}\right] &= -ik_y \mathcal{F}[\psi]. \qquad (9.14) \end{aligned}}$$

In other words, the Fourier transform of a derivative of a function is equal to the Fourier transform of the original function multiplied by $-i\omega$, where ω is the spectral variable corresponding to the variable in the derivative. This property allows us to solve the wave equation. Specifically, taking the transverse Fourier transform of both sides of Eq. 9.10 yields

$$\frac{\partial \widetilde{\psi}(k_x, k_y, z)}{\partial z} = \frac{i}{2k} \left[k_x^2 + k_y^2\right] \widetilde{\psi}(k_x, k_y, z). \tag{9.15}$$

Note that, since our transverse 2-dimensional Fourier transform does not involve the z-coordinate, it commutes with the z-derivative on the left-hand-side of the equation. We can now integrate Eq. (9.15) trivially:

$$\boxed{\widetilde{\psi}(k_x, k_y, z) = \widetilde{\psi}(k_x, k_y, 0) \exp\left(\frac{i}{2k} \left[k_x^2 + k_y^2\right] z\right),} \tag{9.16}$$

where $\widetilde{\psi}(k_x, k_y, 0) = \mathcal{F}[\psi(x, y, 0)]$ is the 2D Fourier transform of the field profile at $z_0 = 0$. To now finally obtain the transverse field profile at arbitrary z, we just take the inverse Fourier transform of $\widetilde{\psi}(k_x, k_y, z)$:

$$\boxed{\psi(x, y, z) = \mathcal{F}^{-1}\left[\widetilde{\psi}(k_x, k_y, 0) \exp\left(\frac{i}{2k} \left[k_x^2 + k_y^2\right] z\right)\right].} \tag{9.17}$$

Solving the paraxial wave equation using Fourier transforms

The procedure below allows us to find the transverse beam profile at arbitrary z when the profile is known at some initial $z_0 = 0$.

1. Take the 2D Fourier transform of the initial field profile:
$$\widetilde{\psi}(k_x, k_y, 0) = \mathcal{F}[\psi(x, y, 0)].$$
2. Use Eq. (9.16) to "propagate" the spatial spectrum to an arbitrary location:
$$\widetilde{\psi}(k_x, k_y, z) = \widetilde{\psi}(k_x, k_y, 0) \exp\left(\frac{i}{2k}\left[k_x^2 + k_y^2\right] z\right)$$
3. Take the inverse Fourier transform to obtain the final transverse field profile:
$$\psi(x, y, z) = \mathcal{F}^{-1}[\widetilde{\psi}(k_x, k_y, z)].$$

This procedure is general, and it works for arbitrary initial field profiles. The same procedure can also be used to examine the time-evolution of wave packets in the Schrodinger equation, or pulse propagation in optical fibers.

9.3 Gaussian beams

In principle, the transverse beam profile $\psi(x, y, z)$ can have whatever shape, provided that the profile decays to zero as $x, y \to \infty$ so as to ensure that the beam does not carry infinite energy. There is, however, one profile that is of particular importance, as lasers typically emit beams with that profile. This profile is the Gaussian beam:

$$\boxed{\psi(x, y, 0) = \psi_0 \exp\left(-\frac{x^2 + y^2}{w_0^2}\right).} \tag{9.18}$$

Gaussian beams are defined by a Gaussian function (surprise surprise), and Fig. 44 shows an example profile. The radius of the beam is described by the variable w_0.

To see how a Gaussian beam propagates as a function of z, we must use the procedure outlined above. As the Fourier transform of a Gaussian function can be calculated analytically, this procedure can be worked out on pen and paper. First, the Fourier transform of $\psi(x, y, 0)$ is:

$$\widetilde{\psi}(k_x, k_y, 0) = \psi_0 \frac{\pi w_0^2}{(2\pi)^2} \exp\left[-\frac{w_0^2}{4}\left(k_x^2 + k_y^2\right)\right]. \tag{9.19}$$

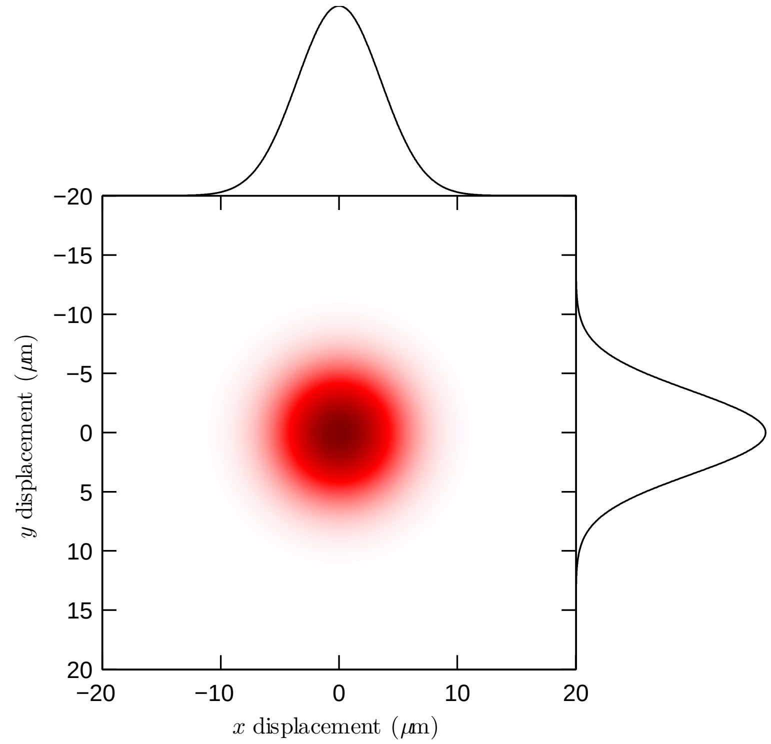

Figure 44: Gaussian beam profile with $w_0 = 5\ \mu\text{m}$. Line profiles show projections along the x and y directions.

Next we evolve the spatial spectrum to arbitrary z using Eq. (9.16):

$$\widetilde{\psi}(k_x, k_y, z) = \psi_0 \frac{\pi w_0^2}{(2\pi)^2} \exp\left[-\frac{w_0^2}{4}\left(k_x^2 + k_y^2\right)\right] \exp\left[\frac{i}{2k}\left(k_x^2 + k_y^2\right) z\right] \tag{9.20}$$

$$= \psi_0 \frac{\pi w_0^2}{(2\pi)^2} \exp\left[\left(-\frac{w_0^2}{4} + \frac{i}{2k} z\right)\left(k_x^2 + k_y^2\right)\right] \tag{9.21}$$

We can see that the form of Eq. (9.21) is exactly the same as that of Eq. (9.19), but the beam width has transformed as $\omega_0^2 \to \omega_0^2 - 2iz/k$ inside the exponential. Accordingly, it should be clear that the transverse profile $\psi(x, y, z) = \mathcal{F}^{-1}[\widetilde{\psi}(k_x, k_y, z)]$ can be written as[80]

$$\boxed{\psi(x, y, z) = \psi_0 \frac{w_0^2}{w_0^2 - 2iz/k} \exp\left(-\frac{x^2 + y^2}{w_0^2 - 2iz/k}\right).} \tag{9.22}$$

Equation (9.22) is a general expression that describes the transverse field profile at arbitrary z for a beam that has a simple Gaussian shape at $z = 0$ given by Eq. (9.18). Unfortunately, the expression is not too easy

[80] Probably the easiest way to confirm the result is to just compare the forms of the initial and transformed profiles and their Fourier transform.

to interpret. To obtain a more intuitive form, we must separate the real and imaginary parts inside and outside of the exponential. After some serious although straightforward algebra, we obtain the standard form for a Gaussian beam at arbitrary z

$$\boxed{\psi(x,y,z) = \psi_0 \frac{w_0}{w(z)} \exp\left[-\frac{x^2+y^2}{w^2(z)}\right] \exp\left[-ik\frac{x^2+y^2}{2R(z)}\right] \exp[i\phi(z)].} \tag{9.23}$$

Here we have defined the following important parameters

Gaussian beam parameters

Beam radius:
$$w(z) = w_0\sqrt{1+\frac{z^2}{z_\mathrm{R}^2}} \tag{9.24}$$

Radius of wavefront curvature:
$$R(z) = z\left(1+\frac{z_\mathrm{R}^2}{z^2}\right) \tag{9.25}$$

Guoy phase shift:
$$\phi(z) = \tan^{-1}\left(\frac{z}{z_\mathrm{R}}\right) \tag{9.26}$$

Rayleigh range:
$$z_\mathrm{R} = \frac{w_0^2 k}{2} = \frac{w_0^2\pi}{\lambda} \tag{9.27}$$

9.4 Properties of Gaussian beams

The expressions above allow us to gain physical insights into the evolution of Gaussian beams.

1. The beam radius $w(z)$ describes the transverse localisation of the beam. The maximum intensity occurs at $x = y = 0$, while at $x^2 + y^2 = r^2 = w^2(z)$ the intensity ($I \propto |\psi|^2$) drops to e^{-2}. Equation (9.24) shows that the beam radius is smallest at $z = 0$, where $w(z) = w_0$. This location, where the beam attains its minimum value, is known as the **beam waist**.

2. A Gaussian beam remains Gaussian as it travels through free space. However, its spot size increases [see Fig. 45(a)] due to diffraction as we move away from the beam waist. The factor $w_0/w(z)$ in the front of the Gaussian formula accounts for energy conservation: as the beam gets wider, its maximum amplitude decreases.

3. When $z = z_\mathrm{R}$, we have $w(z) = \sqrt{2}w_0$. Thus, the Rayleigh range z_R corresponds to the distance from the beam waist over which the beam radius increase by a factor of $\sqrt{2}$. As the beam area is approximately $\pi w(z)^2$, the Rayleigh range also corresponds to the distance from the beam waist over which the beam area doubles. Overall, Rayleigh range describes how fast the beam loses collimation (or diverges): a small z_R implies that the beam diverges quickly and vice versa. Recalling that $z_\mathrm{R} = w_0^2\pi n/\lambda$, we see that a narrow beam (small w_0) diverges more rapidly than a broad beam (large w_0). Finally, it is worth noticing that the beam waist can be understood as a focal point. The distance between the points $\pm z_\mathrm{R}$ is known as the *depth of focus* or the *confocal parameter*.

4. Far away from the beam waist, $z \gg z_{\mathrm{R}}$, allowing us to approximate

$$w(z) \approx w_0 \frac{z}{z_{\mathrm{R}}} = \frac{\lambda}{w_0 \pi} z. \tag{9.28}$$

We see that the beam radius increases linearly with z [see Fig. 45(a)]. We can compute the **angle of beam divergence** as:

$$\boxed{\theta \approx \tan(\theta) = \frac{w(z)}{z} = \frac{\lambda}{w_0 \pi}.} \tag{9.29}$$

This is precisely the divergence angle that was quoted as Eq. (2.1). As can be seen, a small spot diverges strongly.

5. The wavefronts of Gaussian beams are not in general flat as they are for plane waves. Indeed, Fig. 45(b) shows the full electric field, Re$[\psi(x, 0, z) \exp(-ikz)]$ in the $x - z$ plane, and we can readily see how the points of equal field value are curved. The radius of curvature of the wavefronts is given by $R(z)$ as defined in Eq. (9.25). Indeed, the phase at a given point can be written as [see Eq. (9.23)]:

$$\phi = \frac{kr^2}{2R} + kz, \tag{9.30}$$

where $r = \sqrt{x^2 + y^2}$. We can now solve for the points z at which the field assumes the same phase:

$$z = -\frac{r^2}{2R} + \phi. \tag{9.31}$$

This is clearly a parabola with a radius of curvature R.

6. At the beam waist, the radius of curvature $R(0) = \infty$. Accordingly, **the wavefronts are flat at the beam waist.**

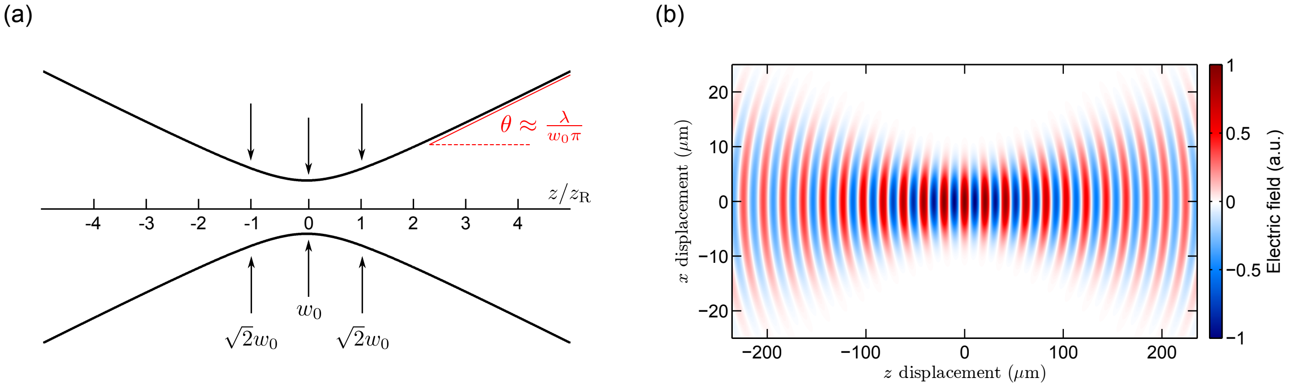

Figure 45: (a) Contours of $\exp(-1)$ field values, showing the evolution of the beam radius $w_0(z)$. (b) Electric field Re$[\psi(x, 0, z) \exp(-ikz)]$, showing the curved wavefronts. To facilitate visualisation, the $\exp(-ikz)$ term has been scaled as $\exp(-ikz/20)$. This increases the wavelength by a factor of 20, making it easier to visualise the wavefronts.

9.5 Complex q parameter

Equation 9.23 describes how a Gaussian beam propagates in free-space. But of course, laser beams are often shaped by various optical components, and so we are interested in knowing how those components affect the beam parameters. For instance, knowing where the beam waist lies after a Gaussian beam passes through a lens is crucial for figuring out where the focal point shall be.

When analysing the transformations of Gaussian beams, it is useful to introduce a complex q parameter that describes the "state" of a Gaussian beam at a certain position z. This parameter is defined as:

$$\boxed{q(z) = (z - z_{\mathrm{w}}) + iz_{\mathrm{R}},} \tag{9.32}$$

where z_{w} is the position of the beam waist. (Above we assumed $z_{\mathrm{w}} = 0$, but it should be clear that z in Eqs. (9.24) – (9.26) can be changed to $z - z_{\mathrm{w}}$ when the waist is not located at $z = 0$.) It is easy to show that the reciprocal of the q-parameter is given by

$$\boxed{\frac{1}{q(z)} = \frac{1}{R(z)} - i\frac{\lambda}{\pi w^2(z)}.} \tag{9.33}$$

As can be seen, the q parameter fully describes the state of the beam. Specifically:

1. If we know the value of $q(z)$ at any point z, the real part immediately tells us how far the point is from the beam waist, while the imaginary part gives us the Rayleigh range. On the other hand, from the Rayleigh range we can compute the beam radius at the waist, w_0. Together, these quantities will then allow us to calculate the beam radius $w(z)$ and radius of curvature $R(z)$ using Eqs. (9.24) and (9.25), respectively.

2. The real part of $1/q(z)$ immediately gives us the reciprocal of the radius of curvature $R(z)$, while its imaginary part allows us to calculate the beam radius $w(z)$.

9.6 ABCD law for Gaussian beams

When a Gaussian beam propagates through any optical element (e.g. free space, mirror, lens), its beam parameters change. These changes can be succinctly described by considering the complex q parameters before and after the element. To get warmed up, let us first consider the propagation through free-space from a point z_i to a point z_f. The initial and final q-parameters are $q_i = (z_i - z_{\mathrm{w}}) + iz_{\mathrm{R}}$ and $q_f = (z_f - z_{\mathrm{w}}) + iz_{\mathrm{R}}$, respectively. Denoting $d = z_f - z_i$, it is evident that

$$q_f = q_i + d. \tag{9.34}$$

Interestingly, this expression can be written as

$$q_f = \frac{Aq_i + B}{Cq_i + D}, \tag{9.35}$$

where the coefficients A, B, C, D are the elements of the ABCD matrix that describes the propagation of light rays through free-space. Maybe this was a coincidence.

Let us now consider a Gaussian beam passing through a thin lens, and look at the complex q parameter just before (q_i) and after (q_f) the lens. Because the lens is thin, there is no change in the beam width: $w_i(z_0) = w_f(z_0)$, where z_0 indicates the lens' position. But what does a lens actually do to an EM wave? The answer is that it imparts a quadratic phase shift on the transverse wavefront. Indeed, consider an EM wave propagating through a lens made of a material with refractive index n, with the two sides of the lens being curved with radii of curvatures R_1 and R_2, and with the lens' thickness as a function of displacement from the optical axis described by the function $d(r)$ [see Fig. 46]. Because light passing through the off-axis portion of the lens propagates through less glass, it acquires a smaller phase shift. Indeed, the phase shift as a function of the radial distance is simply

$$\Delta\phi = -\left[k_0 l_1 + k_0 l_2 + k d(r)\right] = -k_0 n d(r) - k_0[d(0) - d(r)]. \tag{9.36}$$

where $k_0 = \omega/c$ and $k = n\omega/c = nk_0$ are the wavenumbers in air and in the lens, respectively, and l_1 and l_2 are the distances that an off-axis portion of the EM wave propagate in air compared to an on-axis EM wave[81]. Note that these phase shifts simply ensue from the usual $\exp(-ikz)$ phase evolution of an EM wave.

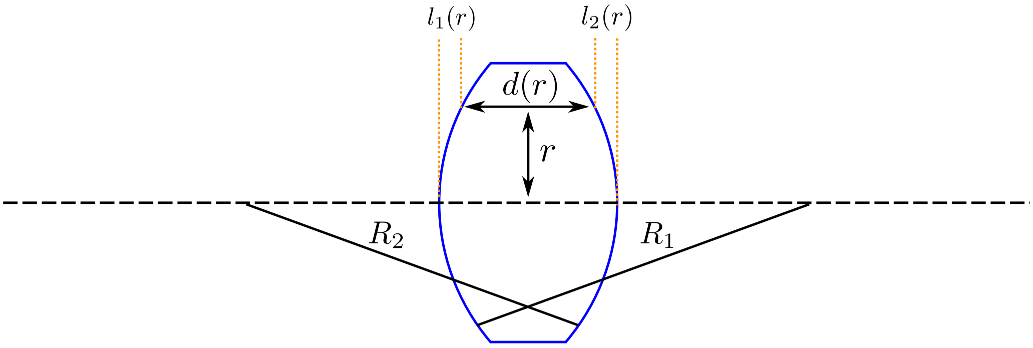

Figure 46: Schematic illustration of a lens with surface radii R_1 and R_2. $d(r)$ is the thickness of the lens as a function of the vertical displacement from the optical axis, r.

Simple geometrical arguments yield the following relationships between the lens curvatures $R_{1,2}$ and distances $l_{1,2}$:

$$R_1^2 = r^2 + (R_1 - l_1)^2 \approx r^2 + R_1^2 - 2R_1 l_1 \tag{9.37}$$

$$R_2^2 = r^2 + (R_2 - l_2)^2 \approx r^2 + R_2^2 - 2R_2 l_2, \tag{9.38}$$

where the approximations use $l_{1,2} \ll r$. From this we obtain the following expressions:

$$l_1 \approx \frac{r^2}{2R_1} \tag{9.39}$$

$$l_2 \approx \frac{r^2}{2R_2}. \tag{9.40}$$

[81] The second form of Eq.(9.36) comes from the relationship $l_1 + l_2 + d(r) = d(0)$.

Accordingly, we can write the the thickness as $d(r)$ as

$$d(r) = d(0) - l_1 - l_2 = d(0) - \frac{r^2}{2}\left[\frac{1}{R_1} + \frac{1}{R_2}\right]. \tag{9.41}$$

For a thin lens, the lensmakers formula applies:

$$\frac{1}{f} = (n-1)\left[\frac{1}{R_1} + \frac{1}{R_2}\right]. \tag{9.42}$$

Substituting Eq. (9.41) and (9.42) in Eq. (9.36), we obtain

$$\boxed{\Delta\phi = -k_0 n d(0) + \frac{k_0 r^2}{2f}.} \tag{9.43}$$

The first term on the right hand side of Eq. (9.43) is an absolute phase shift that does not influence the Gaussian beam parameters. The second term, however, gives rise to different phase shifts across the transverse wave front. If the Gaussian beam has the form $\psi_i(x, y, z_0)$ given by Eq. (9.23) (with $k = k_0$) right before the lens, then right after the lens the beam will have the form $\psi_f(x, y, z_0) = \psi_i(x, y, z_0)\exp(i\Delta\phi)$. Neglecting the absolute phase shift, this yields:

$$\psi_f(x, y, z_0) = \psi_0 \frac{w_0}{w_0(z_0)} \exp\left[-\frac{r^2}{w^2(z)}\right] \exp\left[-ik_0\frac{r^2}{2R(z_0)} + ik_0\frac{r^2}{2f}\right] \exp[i\phi(z_0)] \tag{9.44}$$

It should be clear that the the lens modifies the radius of wavefront curvature. Specifically, we have

$$\boxed{\frac{1}{R_f(z_0)} = \frac{1}{R_i(z_0)} - \frac{1}{f},} \tag{9.45}$$

where R_i and R_f are the wavefront curvatures right before and after the lens, respectively. Accordingly, the inverse of the complex q parameter transforms as

$$\frac{1}{q_f(z_0)} = \frac{1}{R_f(z_0)} - i\frac{\lambda}{\pi w^2(z_0)} = \frac{1}{q_i(z_0)} - \frac{1}{f}. \tag{9.46}$$

Taking the reciprocal, we can write

$$\boxed{q_f = \frac{q_i}{-q_i/f + 1} = \frac{Aq_i + B}{Cq_i + D},} \tag{9.47}$$

where A, B, C, D are the elements of the ABCD matrix derived for a thin lens!

The above analyses for a thin lens and for free-space propagation both show that the transformation of the complex beam parameter q can be expressed using the ABCD matrices of the respective optical components. This law turns out to be very general:

ABCD law for Gaussian beams

The passage of Gaussian beams through various optical components can be described using a simple ABCD law for Gaussian beams. Specifically, if the complex q-parameter of the Gaussian beam right before the component is q_i, the corresponding parameter right after the component is

$$q_f = \frac{Aq_i + B}{Cq_i + D}, \tag{9.48}$$

where A, B, C, and D are the elements of the ray transfer matrix describing the component.

9.7 Gaussian beams and resonators

Lasers typically generate Gaussian beams. This is because Gaussian beams correspond to *eigenmodes* of optical cavities, i.e., transverse field arrangements that can persist in the cavity without changing from round trip to round trip. This can actually be understood quite simply by recalling that the wavefronts of a Gaussian beam are close to spherical. Indeed, if a Gaussian beam hits a spherical mirror whose radius of curvature matches the beam's radius of wavefront curvature, the mirror will simply return the beam exactly from where it came from. Accordingly, if the beam's radius of curvatures at both resonator mirrors matches the curvature of the mirrors, the beam can happily bounce back and forth between the mirrors forever.

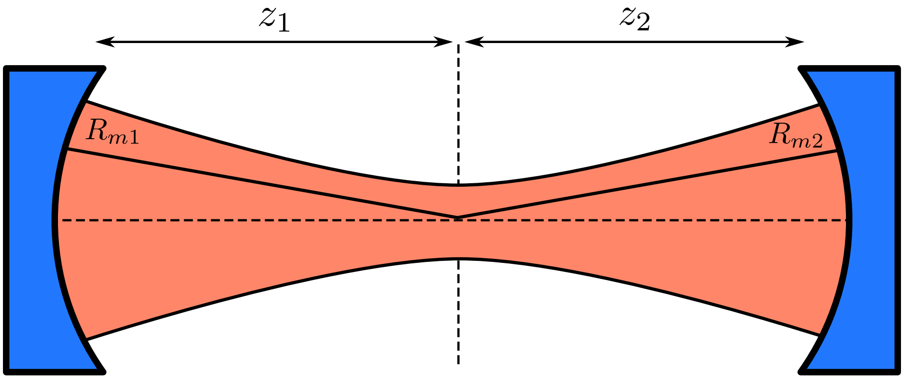

Figure 47: Schematic illustration of a Gaussian beam in a laser resonator. For a cavity eigenmode, the beam's radii of curvatures must match the mirrors' radii of curvatures.

The requirement that the beam's radius of wavefront curvature must match the radii of the resonator mirrors allows us to deduce what the beam will be like inside and outside the cavity. Referring to Fig. 47, we have

$$-R_{m1} = R(z_1) \tag{9.49}$$

$$R_{m2} = R(z_2), \tag{9.50}$$

where R_{m1} and R_{m2} denote mirror curvatures, z_1 and z_2 are the displacements of the two mirrors from the waist (the mirror separation is $L = |z_1| + |z_2|$), and the minus sign comes from sign conventions[82]. Injecting

[82] For example, concave mirrors always have positive curvature, while the sign of our Gaussian beam curvature depends on relative position to the waist.

the expressions for the wavefront curvatures, we obtain

$$\boxed{\begin{aligned} -R_{m1} &= z_1 + \frac{z_{\mathrm{R}}^2}{z_1} && (9.51)\\ R_{m2} &= z_2 + \frac{z_{\mathrm{R}}^2}{z_2} && (9.52)\\ L &= z_2 - z_1. && (9.53)\end{aligned}}$$

If we know the resonator parameters (mirror curvatures and their separation), we can solve for the waist position (z_1 and z_2) and the Rayleigh range z_{R} (three equations, three unknowns). One finds (exercise):

$$\boxed{\begin{aligned} z_{\mathrm{R}}^2 &= \frac{g_1 g_2(1-g_1 g_2)}{(g_1+g_2-2g_1 g_2)^2}L^2 && (9.54)\\ z_1 &= \frac{-g_2(1-g_1)}{g_1(1-g_2)+g_2(1-g_1)}L && (9.55)\\ z_2 &= \frac{g_1(1-g_2)}{g_1(1-g_2)+g_2(1-g_1)}L, && (9.56)\end{aligned}}$$

where $g_i = 1 - L/R_{\mathrm{i}}$. From these equations, we can obtain expressions for the spot size at the waist w_0 as well as spot sizes at both mirrors.

An alternative way of finding the characteristics of a cavity eigenmode is to note that, after one full round trip, the beam must come back to its original state. If the eigenmode's q-parameter in the beginning of the round trip is q_i, then after one full round trip the parameter must transform as

$$\boxed{q_f = \frac{Aq_i + B}{Cq_i + D} = q_i,} \tag{9.57}$$

where A, B, C and D are the elements of the ray transfer matrix M_{cavity} that describes one full round trip.

From this, we can see that the eigenmode's q-parameter must satisfy

$$\boxed{Cq_i^2 + (D-A)q_i - B = 0.} \tag{9.58}$$

Solving for q_i yields

$$\begin{aligned} q_i &= \frac{(A-D) \pm \sqrt{(D-A)^2 + 4BC}}{2C} && (9.59)\\ &= \frac{(A-D) \pm \sqrt{(D+A)^2 - 4}}{2C}, && (9.60)\end{aligned}$$

where we used $\det[\mathrm{M}_{\text{cavity}}] = AD - BC = 1$. Noting that the *q-parameter must have a nonzero imaginary part*, it follows that

$$\boxed{\left|\frac{A+D}{2}\right| < 1.} \tag{9.61}$$

This is the same resonator stability condition as we found before [Eq. (8.13)], but now with a strict inequality.

9.7.1 Higher order transverse modes

Gaussian beams are not the only beams with suitable wavefront curvatures to be sustained in a given laser resonator. In fact, there can be infinitely many such eigenmodes, and they are collectively referred to as the *transverse modes* of the cavity. The particular transverse modes sustained by a cavity depends on the cavity's symmetry. For example, if the symmetry is restricted by a rectangular element, the modes can most conveniently be described using Hermite-Gaussian polynomials[83]:

$$\boxed{\begin{aligned}\psi_{lm}(x,y,z) = \psi_0 \frac{w_0}{w_0(z)} \exp\left[-\frac{x^2+y^2}{w^2(z)}\right] \exp\left[-ik\frac{x^2+y^2}{2R(z)}\right] \exp[i\phi(z)(1+l+m)] \\ \times H_l\left[\sqrt{2}\frac{x}{w(z)}\right] H_m\left[\sqrt{2}\frac{y}{w(z)}\right],\end{aligned}} \tag{9.62}$$

where H_n is the n^{th} Hermite polynomial and l and m are nonnegative integers.

Each combination of integers l and m gives rise to a distinct transverse field pattern. These modes are commonly labelled as TEM$_{lm}$. In Fig. 48, we show intensity profiles corresponding to an assortment of TEM$_{lm}$ modes. It should be evident that the TEM$_{00}$ mode corresponds to our standard Gaussian beam. It is worth noting that the Hermite-Gaussian polynomials form a complete set: an arbitrary transverse profile can be expressed as a linear superposition of TEM$_{lm}$ modes.

The Gaussian TEM$_{00}$ transverse mode is also known as the *fundamental mode*. In most applications, it is desirable to have the laser output as close as possible to a clean TEM$_{00}$. This is because the Gaussian profile is clean and uniform, and does not contain any gaps. Furthermore, out of all the modes, the Gaussian beam has the smallest spot size at the waist, and it exhibits the smallest angle of divergence.

To force the laser to emit a clean TEM$_{00}$ mode, one can introduce a a narrow aperture inside the cavity. This induces extraneous losses to the higher-order modes, which are spatially wider than the TEM$_{00}$ mode. Yet unfortunately, in practice lasers never produce 100 % true Gaussian beams, but there is always some residual contributions from higher-order modes. The degree of variation of a beam from a Gaussian beam can be quantified using the so-called M^2 parameter, defined as

$$\boxed{M^2 = \theta\frac{\pi w_0}{\lambda},} \tag{9.63}$$

where θ is the far-field divergence half-angle and w_0 is the beam radius at the waist. For a pure Gaussian beam, $M^2 = 1$, whilst for all higher order modes $M^2 > 1$. Thus, measuring the M^2 of a real laser beam tells us

[83] If the cavity is cylindrically symmetric, the modes can be more conveniently described with Laguerre-Gaussian polynomials.

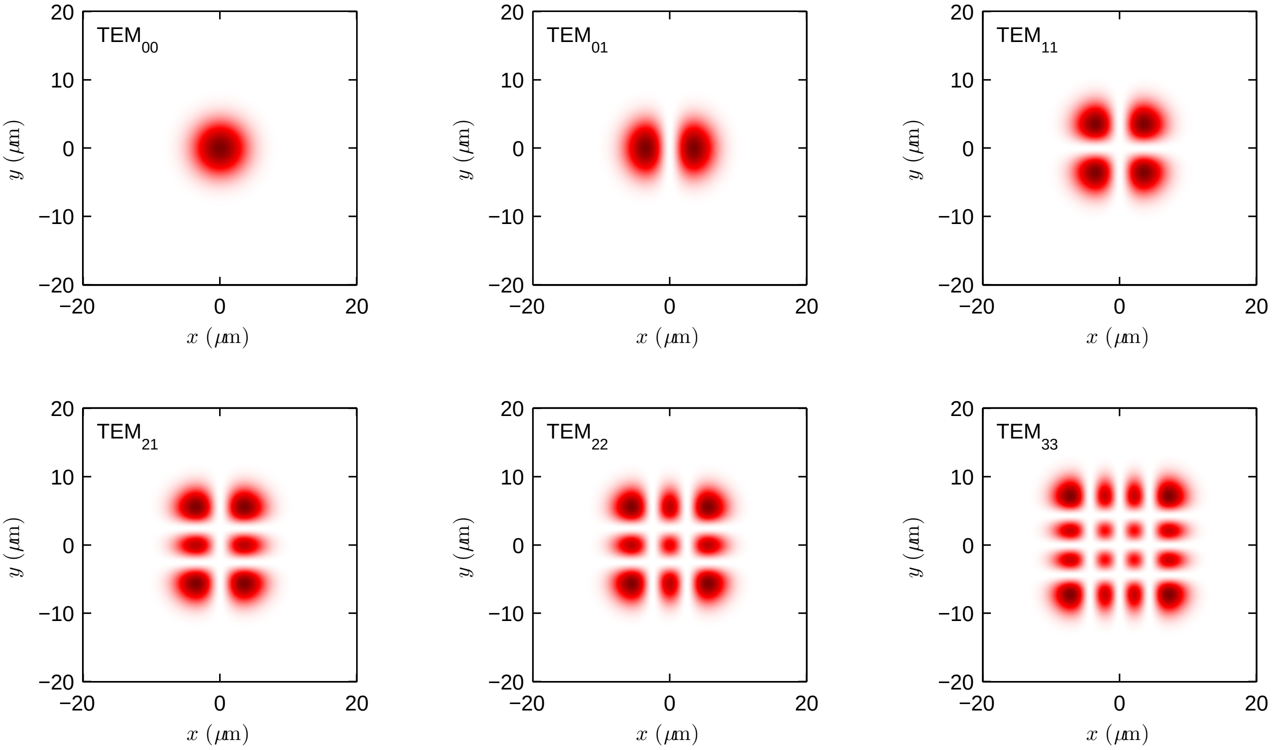

Figure 48: Examples of Hermite-Gaussian transverse modes, $|\psi(x, y)|^2$ at the beam waist. In all cases, $w_0 = 5\ \mu$m.

how close to a Gaussian the beam is: if $M^2 \approx 1$, the beam is almost a Gaussian, while higher values indicate significant deviations.

10 Nonlinear optics

Long time ago, all the way in Chapters 4 and 5, we considered the propagation of electromagnetic waves in media. In Chapter 4, we introduced the constitutive equation

$$\vec{\mathbf{P}}(\vec{\mathbf{r}},t) = \varepsilon_0 \chi \vec{\mathbf{E}}(\vec{\mathbf{r}},t) \tag{10.1}$$

to describe the macroscopic polarization $\vec{\mathbf{P}}(\vec{\mathbf{r}},t)$ induced in a dielectric medium by an applied electric field $\vec{\mathbf{E}}(\vec{\mathbf{r}},t)$. With the help of this constitutive equation, we found that dielectric media can support EM waves with velocity c/n, where the refractive index depends on the susceptibility χ viz. $n = \sqrt{1+\chi}$. Then, in Chapter 5, we used the classical electron oscillator model to describe the interactions between light and matter, and managed to *derive* the constitutive equation quoted above starting from fundamental "first principles".

The constitutive equation (10.1) states that the induced polarization is linearly proportional to the electric field. This simple linear relationship is sufficient to explain all optical phenomena that we encounter in our daily lives: lenses, mirrors, refraction, reflection, diffraction, prisms, telescopes, and so on. Yet, as it turns out, Eq. (10.1) is not universally valid: it breaks down when the amplitude of the electric field is sufficiently large. A more general constitutive equation expresses the induced polarization as a *power series* with respect to the electric field:

$$\vec{\mathbf{P}}(\vec{\mathbf{r}},t) = \varepsilon_0 \left[\chi^{(1)}\vec{\mathbf{E}}(\vec{\mathbf{r}},t) + \chi^{(2)}\vec{\mathbf{E}}^2(\vec{\mathbf{r}},t) + \chi^{(3)}\vec{\mathbf{E}}^3(\vec{\mathbf{r}},t) + ...\right], \tag{10.2}$$

where $\chi^{(n)}$ is known as the n^{th} order electric susceptibility of the material[84].

The magnitudes of the electric susceptibilities decrease rapidly with order, which implies that terms beyond the first order are significant only when the electric field amplitude is sufficiently large. When these higher-order terms are *not* negligible, the dielectric polarization $\vec{\mathbf{P}}(\vec{\mathbf{r}},t)$ is a *nonlinear* function of the electric field. This nonlinear relationship results in altogether new – and sometimes unexpected – physical phenomena with a rich variety of applications. Nonlinear optics is the branch of optics that describes the behaviour of light in such *nonlinear* media.

Common light sources (such as stars, light bulbs etc.) do not emit radiation with electric fields sufficiently strong to excite nonlinear optical effects, explaining why nonlinear optical effects are not encountered in our daily lives. Only laser light can be sufficiently strong. As a consequence, the birth of nonlinear optics coincides almost exactly with the birth of lasers: the first laser was demonstrated by Maiman in 1960, and the first nonlinear optical effect – so-called second-harmonic generation [discussed below] – was observed by Franken and co-workers in 1961. In what follows, we present a brief introduction to nonlinear optics, explaining some of the salient physics and describing some of the key nonlinear phenomena and their applications. We begin by briefly commenting on the mathematical structure that underpins the nonlinear constitutive Eq. (10.2), and then underline the simplifications that we will be employing throughout our analyses.

10.1 The complex mathematics of nonlinear optics

The nonlinear constitutive Eq. (10.2) looks innocent enough, but in fact hides a rather complex mathematical structure. The observant reader may already be cautious of the fact that the electric field is a vector and one cannot (unambiguously) raise vectors to powers. Equation (10.2) should indeed be taken as a gross notational

[84] Do not confuse the index n as an exponent.

simplification. In actuality, the electric susceptibility $\chi^{(n)}$ is a tensor of rank $(n+1)$, and the i^{th} Cartesian component of the dielectric polarization $\vec{\mathbf{P}} = [P_x, P_y, P_z]^T$ is given by

$$P_i = \varepsilon_0 \left[\sum_j \chi_{ij}^{(1)} E_j + \sum_{j,k} \chi_{ijk}^{(2)} E_j E_k + \sum_{j,k,l} \chi_{ijkl}^{(3)} E_j E_k E_l + ... \right], \tag{10.3}$$

Physically, this means that different components of the electric field vector (or products thereof) can contribute to the different Cartesian components of the dielectric polarization.

The susceptibility tensor $\chi^{(n)}$ can in general have 3^{n+1} non-zero components, but most materials exhibit symmetries that reduce that number. Moreover, all of the susceptibilities depend on frequency – and not just one frequency but several. All in all, the full mathematical formalism that underpins nonlinear optics [e.g. Eq. (10.3)] is rather complicated. In this course, we will ignore much of these complications and focus on a simplified scalar description. Specifically, we assume that the susceptibilities do not depend on frequency, and that only their diagonal components are non-zero. Assuming further that the electric field is linearly polarized along $\hat{\mathbf{x}}$, such that $\vec{\mathbf{E}}(\vec{\mathbf{r}}, t) = E(\vec{\mathbf{r}}, t)\hat{\mathbf{x}}$ we can write the dielectric polarization as $\vec{\mathbf{P}}(\vec{\mathbf{r}}, t) = P(\vec{\mathbf{r}}, t)\hat{\mathbf{x}}$, where

$$\boxed{P(\vec{\mathbf{r}}, t) = \varepsilon_0 \left[\chi^{(1)} E(\vec{\mathbf{r}}, t) + \chi^{(2)} E^2(\vec{\mathbf{r}}, t) + \chi^{(3)} E^3(\vec{\mathbf{r}}, t) + ... \right].} \tag{10.4}$$

Here the effective susceptibilities $\chi^{(n)}$ correspond to complex-valued scalars that describes the diagonal component of the susceptibility tensor along the x direction [e.g. $\chi_{xxx}^{(2)}$].

10.2 Classical anharmonic electron oscillator

In Chapter 5, we used the classical electron oscillator model to show that the dielectric polarization is linearly proportional to the electric field, and we derived a classical prediction for the linear electric susceptibility. So where does the nonlinearity originate from at a microscopic level? The answer is simple. When writing the equation of motion for the electron in Chapter 5, we assumed a restoring force identical to Hooke's law:

$$F_{\text{restoring}} = -kx, \tag{10.5}$$

where x is the electron displacement, and $k = m\omega_0$ with m the mass of the electron and ω_0 the characteristic frequency. However, Hooke's law is only a first-order linear approximation to the real response of springs and other elastic bodies. If the displacement is sufficiently large, the linear approximation becomes inaccurate, and it becomes necessary to consider higher-order terms of the underlying Taylor series [exercise 10.1]:

$$\boxed{F_{\text{restoring}} = -kx - max^2 - mbx^3 - ...,} \tag{10.6}$$

where a and b are constant coefficients.

Considering only the first-order correction term, the classical electron oscillator model would take the form

$$\boxed{m\frac{dx^2(t)}{dt^2} = -m\omega_0^2 x - max^2 - \gamma m\frac{dx}{dt} - eE(t).} \tag{10.7}$$

This equation cannot be solved in closed form with the usual plane wave driving field $E(t) = E_0 \exp(i\omega t)$; perturbation expansion is needed [exercise 10.2]. We can nevertheless glean some obvious insights. First, we can expect that an electric field oscillating at ω will to first order cause the electrons to oscillate at ω. But on the other hand, the x^2 correction term will in this case oscillate at 2ω, highlighting how the electron displacement cannot *just* oscillate at ω: there also has to be some oscillations at 2ω.

A comprehensive analysis of the classical-electron oscillator model with a restoring force given by Eq. (10.6) reveals that the electron displacement can be written as a power series

$$x(t) = c_1 E(t) + c_2 E^2(t) + c_2 E^3(t) + ..., \tag{10.8}$$

where c_n are constants. The power series representation of the macroscopic polarization, given in Eq. (10.4), then follows from the definition of the dielectric polarization, $P(t) = -Nex(t)$. The fact that the electron displacement (and the polarization) depend on different powers of the electric field imply that the corresponding waveforms do not simply mimic the electric field, but rather exhibit more complex behaviours [see Fig. 49(a)].

It is worth noting that the restoring force is related to the underlying potential energy function $U(x)$ as $F_{\text{restoring}} = -dU/dx$. Hooke's law is clearly associated with a harmonic potential $U = kx^2/2$, while a more general restoring force of the form given by Eq. (10.6) is a signature of an *anharmonic* potential. This situation in fact corresponds to the physical situation of electrons in real materials, as the actual potential well experienced by the electrons is not perfectly parabolic [see Fig. 49]. It is only in the vicinity of the potential minimum where the parabolic approximation is suitable [exercise 10.1]: for larger displacements, the anharmonicity must be taken into account.

10.3 Electromagnetic wave equation with a source term

In Chapter 4, we derived the EM wave equation and wrote it in a simple form that clearly discloses the fact that the speed of EM waves in a medium with refractive index n is given by c/n [see Eq. (4.18)]. It is insightful to write the wave equation in a slightly different (and more general) form:

$$\boxed{\nabla^2\vec{\mathbf{E}} - \mu_0\varepsilon_0\frac{\partial^2\vec{\mathbf{E}}}{\partial t^2} = \mu_0\frac{\partial^2\vec{\mathbf{P}}}{\partial t^2}.} \tag{10.9}$$

Equation (4.18) is recovered with the substitution of the linear constitutive Eq. (10.1).

The left-hand side of Eq. (10.9) corresponds to the EM wave equation in vacuum, while the term on the right-hand side can be understood as a *source* term that gives rise to EM waves. As an example, consider a situation where the electric field and the polarization are both zero for times $t < 0$. If the polarization then starts to undergo oscillations at time $t \geq 0$, the right-hand side of the equation will become non-zero. But the equality in Eq. (10.9) stipulates that also the left-hand side must become nonzero, which implies that an oscillating electric field has been generated. Of course, this should not be particularly surprising: in Section 4.6.1, we have argued that oscillating dipoles emit EM radiation and that the fields they generate are proportional to the second time-derivative of the dipole moment. The right-hand side of Eq. (10.9) is nothing but the second time-derivative of the macroscopic dipole moment.

Equation (10.9) leads to a peculiar interpretation for the propagation of EM radiation in media.

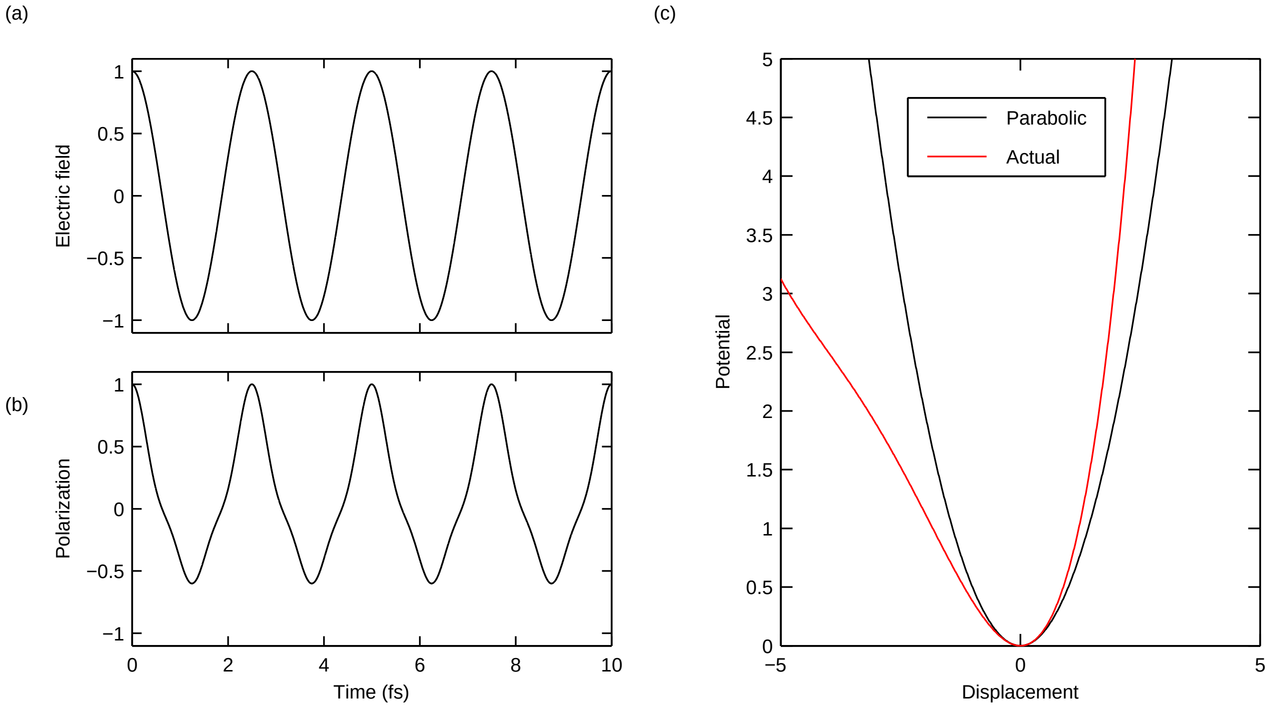

Figure 49: Waveforms corresponding to an (a) electric field with a frequency of 400 THz and (b) the corresponding nonlinear dielectric polarization. (c) Comparison of a parabolic potential characteristic to a harmonic oscillator (black) and an anharmonic potential characteristics to a real atom (red).

Ewald-Oseen extinction theorem

We know that EM radiation incident from vacuum to a medium with refractive index n slows down and propagates with the speed c/n. But why and how does this happen from a microscopic perspective? Well, the electric field of the EM wave causes the bound electrons of the medium to oscillate in time, giving rise to oscillating electric dipoles and a corresponding macroscopic polarization $\vec{\mathbf{P}}$. All of the oscillating dipoles will emit EM radiation of their own, as described in Section 4.6.1 and evident from Eq. (10.9). The total electric field inside the material can be obtained as the superposition of the incident field and the fields generated by all of the oscillating dipoles:

$$\vec{\mathbf{E}}_{\mathrm{m}}(\vec{\mathbf{r}},t)=\vec{\mathbf{E}}_{\mathrm{vac}}(\vec{\mathbf{r}},t)+\vec{\mathbf{E}}_{\mathrm{dip}}(\vec{\mathbf{r}},t), \tag{10.10}$$

where $\vec{\mathbf{E}}_{\mathrm{vac}}$ is the incident field and $\vec{\mathbf{E}}_{\mathrm{dip}}$ is the field radiated by the dipoles. Re-arranging, we obtain

$$\vec{\mathbf{E}}_{\mathrm{dip}}(\vec{\mathbf{r}},t)=\vec{\mathbf{E}}_{\mathrm{m}}-\vec{\mathbf{E}}_{\mathrm{vac}}(\vec{\mathbf{r}},t). \tag{10.11}$$

Now, we know that the wave inside the material ($\vec{\mathbf{E}}_{\mathrm{m}}$) corresponds to a single EM wave that propagates with the speed c/n. The above expression shows that the oscillating dipoles radiate a field that exactly cancels out the incident field and creates the transmitted field travelling within the medium at speed c/n. This rather remarkable fact is known as the Ewald-Oseen extinction theorem.

10.4 Nonlinear polarization

Let us divide the dielectric polarization in a linear and a nonlinear part:

$$\boxed{\vec{\mathbf{P}}(\vec{\mathbf{r}},t)=\vec{\mathbf{P}}_{\mathrm{L}}(\vec{\mathbf{r}},t)+\vec{\mathbf{P}}_{\mathrm{NL}}(\vec{\mathbf{r}},t),} \tag{10.12}$$

where $\vec{\mathbf{P}}_{\mathrm{L}}(\vec{\mathbf{r}},t)=\varepsilon_0\chi^{(1)}\vec{\mathbf{E}}(\vec{\mathbf{r}},t)$ and the *nonlinear* polarization $\vec{\mathbf{P}}_{\mathrm{NL}}$ contains all the nonlinear terms. Substituting this expression in Eq. (10.9) allows us to derive the following nonlinear EM wave equation:

$$\boxed{\nabla^2\vec{\mathbf{E}}-\frac{n_0^2}{c^2}\frac{\partial^2\vec{\mathbf{E}}}{\partial t^2}=\mu_0\frac{\partial^2\vec{\mathbf{P}}_{\mathrm{NL}}}{\partial t^2},} \tag{10.13}$$

where $n_0=\sqrt{1+\chi^{(1)}}$. Here the left-hand side accounts for linear propagation effects while the right-hand side corresponds to a nonlinear source term. The nonlinear EM wave equation allows us to rigorously study the propagation of light through a nonlinear optical medium. In what follows, we provide a brief description of various nonlinear optical effects by considering different forms of the nonlinear polarization. For the sake of simplicity, we will consider the scalar approximation, where $\vec{\mathbf{E}}(\vec{\mathbf{r}},t)=E(\vec{\mathbf{r}},t)\hat{\mathbf{x}}$, $\vec{\mathbf{P}}(\vec{\mathbf{r}},t)=P(\vec{\mathbf{r}},t)\hat{\mathbf{x}}$, and the constitutive Eq. (10.4) holds.

10.5 Second-order nonlinear effects I: second-harmonic generation and rectification

Consider a second-order nonlinear polarization of the form

$$\boxed{P_{\mathrm{NL}}^{(2)}(\vec{\mathbf{r}},t)=\varepsilon_0\chi^{(2)}E^2(\vec{\mathbf{r}},t).} \tag{10.14}$$

We first note that second-order effects can arise only in materials that do not possess a center of inversion, i.e., materials that are non-centrosymmetric [exercise10.3]. The most important materials of this type are crystal materials with certain symmetry, such as lithium niobate, lithium tantalate, and β-barium borate.

Let us assume that the incident wave is a plane EM wave propagating in the positive z direction such that

$$E(z,t)=E_0e^{i(\omega t-kz)}+\text{c.c.}, \tag{10.15}$$

where c.c stands for complex conjugate. Note that, since we will be calculating squares of the fields, it is imperative that we express them as real quantities: it is for this reason that we include the complex conjugate in our expression for the electric field.

Using Eq. (10.15) in Eq. (10.14), we obtain [exercise 10.4]

$$\boxed{P_{\mathrm{NL}}^{(2)}(\vec{\mathbf{r}},t)=\varepsilon_0\chi^{(2)}\left[E_0^2e^{2i(\omega t-kz)}+2|E_0|^2+\text{c.c}\right].} \tag{10.16}$$

We see that the nonlinear polarization contains two terms; in what follows, we describe their physical meaning.

10.5.1 Second-harmonic generation

The second time-derivative of the nonlinear polarization, which acts as a source in the nonlinear wave Eq. (10.13), is given by

$$\frac{\partial^2 P_{\mathrm{NL}}^{(2)}}{\partial t^2} = -4\varepsilon_0 \chi^{(2)} \omega^2 E_0^2 e^{2i(\omega t - kz)} + \text{c.c.} \tag{10.17}$$

We see that this source term oscillates at frequency 2ω, and as a consequence, it drives the emission of EM waves with that frequency. This observation is quite surprising. An incident EM wave causes electrons to oscillate in a way that their motion contains frequency components both at ω and 2ω; the latter results in the generation of new EM radiation at 2ω. In essence, an EM wave with frequency ω enters the medium, and two EM waves with frequencies ω and 2ω exit the medium [see Fig. 50(a)]. This phenomenon is known as *second-harmonic generation* (SHG), and it was the first nonlinear optical phenomenon observed. SHG is widely used to generate laser light at frequencies where it is difficult to directly build a laser (e.g. due to unavailability of suitable active media). Green laser pointers are good examples: they start from a neodymium-doped active medium that generates light at 1064 nm (infrared), and SHG is then used to convert that to 532 nm (green).

Second-harmonic generation can be understood in the photon picture as the annihilation of two photons at frequency ω and creation of a new photon at frequency 2ω [see Fig. 50(b)]. In this picture, the incident photons excite *virtual* energy levels of the atom, while the second-harmonic photons are emitted when the atom relaxes down to its ground state. Energy is clearly conserved in the process: $2\omega = \omega + \omega$.

10.5.2 Optical rectification

The nonlinear polarization given by Eq. (10.16) also contains a DC (non-oscillating) term $2\varepsilon_0\chi^{(2)}|E_0|^2$. The second derivative of this term vanishes, and so it does not generate any EM waves. Rather, this term describes the fact that, through the second-order nonlinearity, an EM wave oscillating at frequency ω induces a *static* polarization in the material. Such a static polarization results in a static (internal) electric field: a voltage appears across the medium [see Fig. 51(a)]. This phenomenon is known as *optical rectification*.

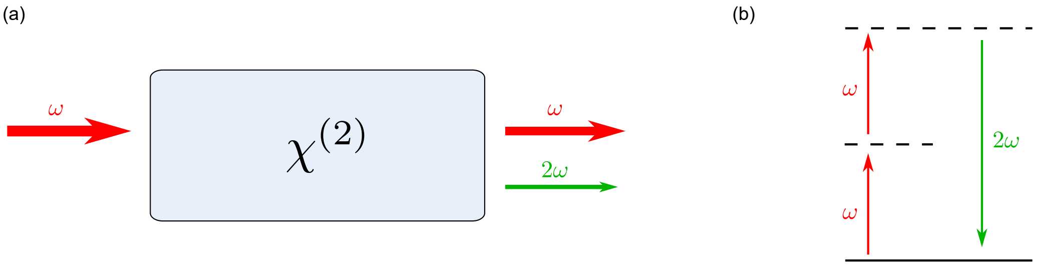

Figure 50: (a) Schematic illustration of SHG. An incident beam at ω is frequency-doubled by the $\chi^{(2)}$ nonlinearity to produce a new signal at 2ω. (b) SHG can be understood in the photon picture as the annihilation of two photons at ω and the creation of a single photon at 2ω. The dashed lines correspond to *virtual* energy levels.

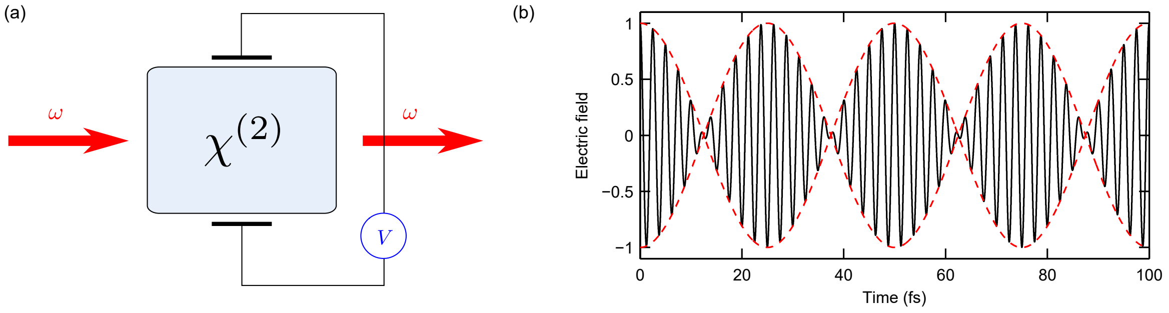

Figure 51: (a) Because of the $\chi^{(2)}$ nonlinearity, an intense incident beam can induce a static voltage across the material. Similarly, if an external voltage is applied across the material, the refractive index experienced by the beam can be changed. Modulating the external voltage in time allows the phase of the incident beam to be modulated. When combined with polarizers or interferometers, also amplitude modulation can be achieved (b). In (b), the red dashed curve represents the modulation envelope, while black solid curve is the actual electric field.

10.6 Second-order nonlinear effects II: Pockels effect and modulators

Above we saw that a strong beam propagating through a second-order nonlinear material will give to a static voltage across the material. But what if we externally apply such a voltage to begin with? In this case, the electric field can be written [ignoring second-harmonic generation] as

$$E(z,t) = E_{\mathrm{DC}} + E_0 e^{i(\omega t - kz)} + \text{c.c.}, \tag{10.18}$$

where the first term is real and corresponds to the static electric field, and the second (and third) term(s) corresponds to the EM wave. Substituting Eq. (10.25) into Eq. (10.14), we find that, in addition to the SHG and rectification terms considered above, the nonlinear polarization contains a term that oscillates at frequency ω. The total polarization at the incident frequency ω, including the linear term, then reads [exercise 10.4]

$$P(z,t) = \varepsilon_0 \chi^{(1)} E_0 e^{i(\omega t - kz)} + 2\varepsilon_0 \chi^{(2)} E_{\mathrm{DC}} E_0 e^{i(\omega t - kz)} + \text{c.c} \tag{10.19}$$

$$= \varepsilon_0 \left[\chi^{(1)} + 2\chi^{(2)} E_{\mathrm{DC}}\right] E_{\mathrm{EM}}(z,t) + \text{c.c}, \tag{10.20}$$

where we defined the EM wave $E_{\mathrm{EM}}(z,t) = E_0 \exp[i(\omega t - kz)]$. Substituting this form of the total polarization in Eq. (10.9), we can derive the following form for the EM wave equation [exercise 10.5]:

$$\nabla^2 \vec{\mathbf{E}} - \frac{n^2}{c^2}\frac{\partial^2 \vec{\mathbf{E}}}{\partial t^2} = 0, \tag{10.21}$$

where the refractive index n now obeys

$$n^2 = 1 + \chi^{(1)} + 2\chi^{(2)} E_{\mathrm{DC}}. \tag{10.22}$$

Noting that the usual, *linear* refractive index $n_0 = \sqrt{1+\chi^{(1)}}$, and that the term involving the susceptibility $\chi^{(2)}$ is generally much smaller than n_0^2, we can express the refractive index in the form

$$\boxed{n \approx n_0 + \frac{\chi^{(2)}}{n_0}E_{\text{DC}}.} \tag{10.23}$$

And so we see that an external voltage (electric field) applied across a medium can be used to modify the refractive index experienced by light. This effect is known as the *Pockel's effect* after the German physicist Friedrich Pockels who studied the process around 1895. It is interesting to note that the Pockel's effect was discovered several decades before lasers were invented; this is because the effect is induced by a static electric field rather than an EM wave. It is often considered a precursor of nonlinear optics, but the analysis above nevertheless shows that the Pockel's effect corresponds to a second-order nonlinear optical effect.

Pockel's effect has important practical applications. Specifically, by modulating the DC voltage at some radiofrequency ω_m, such that $E_{\text{DC}} \equiv E_{\text{DC}}(t) = A\sin(\omega_m t)$ the optical beam propagating through the material can be *phase-modulated*:

$$E_{\text{out}} = E_{\text{in}}e^{-ikL} = E_{\text{in}}e^{-\frac{\omega L}{c}[n_0 + AK\sin(\omega_m t)]}, \tag{10.24}$$

where $K = \chi^{(2)}/n_0$ and we used $k = n\omega/c$ with n given by Eq. (10.23). By using additional polarizers – or by placing a Pockels phase modulator inside an interferometer – intensity modulation can be achieved as well [see Fig. 51(b) and exercise 10.6]. Modulators based on the Pockel's effect are widely used to modulate data on an optical carrier wave: they are vital components in modern telecommunication systems.

10.7 Second-order nonlinear effects III: sum- and difference-frequency generation

Consider now a situation where two beams with different frequencies ω_1 and ω_2 are simultaneously launched into a second-order nonlinear medium [see Fig. 52(a)]. The incident field is given by

$$E(z,t) = E_{01}e^{i(\omega_1 t - k_1 z)} + E_{02}e^{i(\omega_2 t - k_2 z)} + \text{c.c.}. \tag{10.25}$$

Calculating the nonlinear polarization using Eq. (10.14), we find DC rectification terms and second-harmonic terms induced by both waves. But in addition, we find two new terms that oscillate at frequencies that correspond to the sum ($\omega_1 + \omega_2$) and difference ($\omega_1 - \omega_2$) of the two incident waves. Since an oscillating polarization results in the generation of EM waves whose frequency coincides with that of the polarization, these terms will give rise to new EM waves at the sum- and difference-frequencies. The processes that result in the generation of these two new EM waves are aptly known as sum- and difference-frequency generation.

Sum-frequency generation (SFG) can be interpreted in the photon picture as the annihilation of photons with frequency ω_1 and ω_2 and the subsequent creation of a single new photon with frequency $\omega_3 = \omega_1 + \omega_2$ [see Fig. 52(b)]. Difference-frequency generation (DFG) can similarly be interpreted as the annihilation of one energetic photon at ω_1 and the subsequent generation of lower-energy photons at ω_2 and ω_3 such that $\omega_2 + \omega_4 = \omega_1$, where $\omega_4 = \omega_1 - \omega_2$ [see Fig. 52(c)]. Both processes clearly satisfy energy conservation.

As seen above, the nonlinear polarization can in general contain numerous terms that are oscillating at different frequencies. However, not all of these terms result in the generation of EM radiation with significant intensity. Indeed, it turns out that a stringent condition – known as the phase-matching condition – needs to be satisfied for the wave to grow to significant intensity as it propagates. Such a phase-matching condition can be

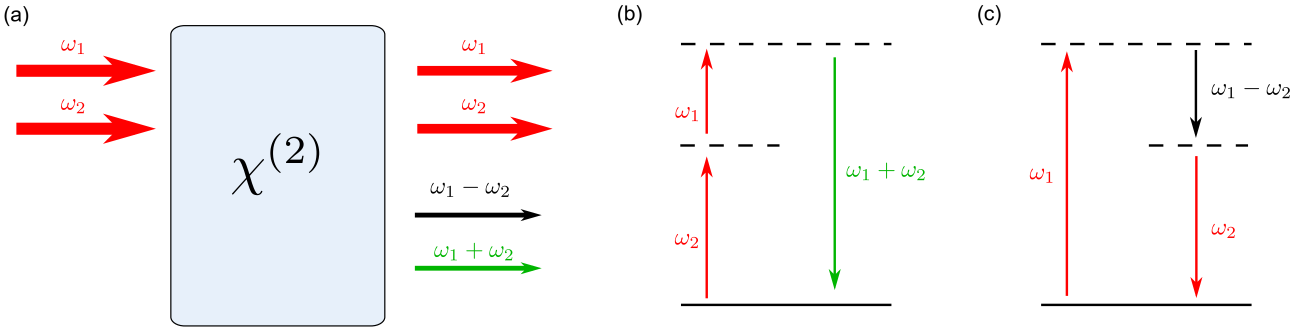

Figure 52: (a) Schematic illustration of sum- and frequency-difference generation. Possible second-harmonic waves are not shown for clarity. (b, c) Photon energy diagrams for SFG and DFG, respectively. In SFG, photons at ω_1 and ω_2 merge to produce one photon at $\omega_1 + \omega_2$. In DFG, a photon at ω_1 splits into a photon at ω_2 and $\omega_1 - \omega_2$.

satisfied for only one process at a time (if even). Accordingly, we may expect only one of the generated new frequencies to be significant. Phase-matching will be considered in detail at a latter stage.

10.7.1 Parametric amplification

The photon interaction picture for DFG [see Fig. 52(c)] reveals that, for every interaction, a photon is generated at the new frequency $\omega_4 = \omega_1 - \omega_2$ and at one of the original frequencies ω_2. This means that the incident beam associated with frequency ω_2 is in fact amplified as it propagates through the medium (at the expense of the beam at frequency ω_1). It is important to note that this amplification does not require an active gain medium; rather, it occurs entirely via the nonlinear optical interaction between the incident (and the generated) beams. Such nonlinear optical amplification is commonly referred to as *parametric amplification*.

Parametric amplification can occur even without the “signal” beam at ω_2 present at the input [see Fig. 53(a)]. Here, a single intense pump $\omega_1 = \omega_2 + \omega_4$ launched into a nonlinear medium spontaneously breaks into new photons at ω_2 and ω_4. It may feel strange that this can occur, as the nonlinear polarization with a single input does not contain the DFG term. The conundrum is resolved by noting that the vacuum is not exactly empty: it exhibits quantum fluctuations that manifest themselves as the appearance (and disappearance) of photons. Photons at ω_2 can emerge from such fluctuations and be parametrically amplified. The resulting spontaneous process is known as *parametric fluorescence*.

Parametric amplification immediately implies an application. Specifically, it can be used to conveniently convert laser light from one frequency to another. The process can be made very efficient by placing the nonlinear medium inside a resonator. By driving the system at some pump frequency ω_1, so-called *signal* and *idler* beams that satisfy $\omega_2 + \omega_3 = \omega_1$ can be generated [see Fig. 53(b)]. The precise signal and idler frequencies that are generated depend on the phase-matching conditions, which can be modified by controlling the angle or temperature of the nonlinear medium. In this way, it is possible to generate laser light whose frequency can be tuned over broad spectral range. An oscillator device that achieves such operation is known as an *optical parametric oscillator* (OPO).

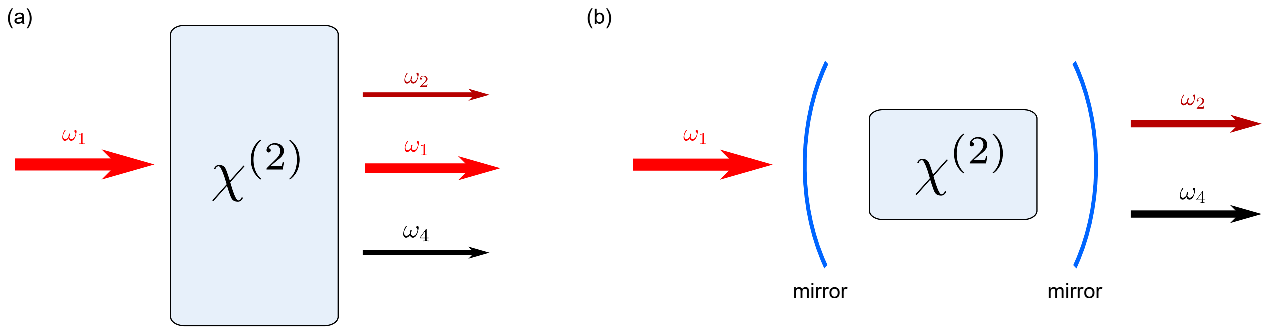

Figure 53: (a) Schematic illustration of parametric fluorescence: an intense beam at ω_1 propagating in a $\chi^{(2)}$ nonlinear medium can spontaneously generate photons at ω_2 and ω_4 such that $\omega_2+\omega_4=\omega_1$. (b) Schematic illustration of an optical parametric oscillator (OPO). By driving a resonator that contains a $\chi^{(2)}$ nonlinear medium with an intense pump at ω_1, new beams at ω_2 and ω_4 can be efficiently generated. For historical reasons, the generated beam with the higher frequency is known as the “signal” beam whereas the beam with the lower frequency is known as the “idler”.

10.8 Third-order nonlinear effects

Consider now the third-order nonlinear polarization of the form

$$\boxed{P_{\rm NL}^{(3)}(\vec{\mathbf{r}},t)=\varepsilon_0\chi^{(3)}E^3(\vec{\mathbf{r}},t).} \tag{10.26}$$

In contrast to second-order nonlinearities, which only manifest themselves in materials without inversion symmetry, the third-order nonlinearity is ubiquitous: it is present in all materials. Accordingly, for materials (such as glass) that are centrosymmetric, it corresponds to the lowest-order nonlinearity.

In general, third-order nonlinear interactions can involve up to four different waves with different frequencies. Their general analysis would require us to consider an incident wave made out of three different frequency components. This analysis is quite tedious, as the number of terms in the nonlinear polarization is quite large. Here we only consider the simplest situation, where a single monochromatic wave is incident on the material. Substituting Eq. (10.15) into Eq. (10.26), we obtain [exercise 10.4]:

$$\boxed{P_{\rm NL}^{(3)}(z,t)=\varepsilon_0\chi^{(3)}\left[E_0^3e^{3i(\omega t-kz)}+3|E_0|^2E_0e^{i(\omega t-kz)}+\text{c.c}\right].} \tag{10.27}$$

We again see that the nonlinear polarization contains two terms.

10.8.1 Third-harmonic generation

The third-order nonlinear polarization exhibits a term that oscillates at frequency 3ω. This term will give rise to EM radiation at that frequency. The process is known as *third-harmonic generation*, and it allows for an incident wave with frequency ω to drive the generation of a secondary beam at the third-harmonic frequency 3ω [see Fig. 54]. Similarly to second-harmonic generation, three photons at the fundamental frequency are

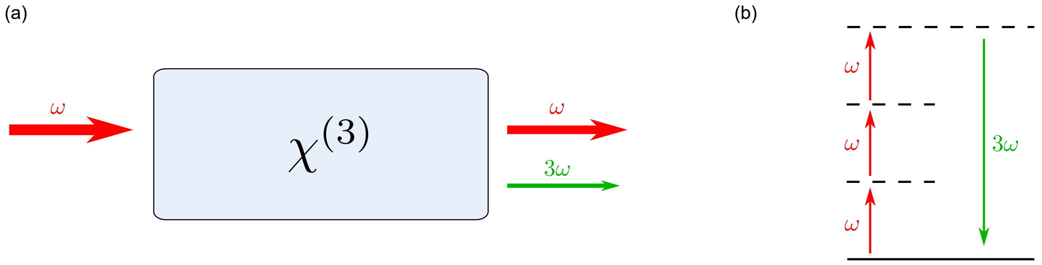

Figure 54: (a) Schematic illustration of THG. An incident beam at ω is frequency-tripled by the $\chi^{(3)}$ nonlinearity to produce a new signal at 3ω. (b) THG can be understood in the photon picture as the annihilation of three photons at ω and the creation of a single photon at 3ω.

annihilated during the process, and a single photon at the third-harmonic frequency is generated, ensuring energy conservation.

Second- and third-harmonic generation should highlight that each term in the total nonlinear polarization [see Eq. (10.4)] is responsible for the generation of harmonics with higher and higher order. Of course, we must emphasise again that efficient generation of any of the harmonics requires stringent phase-matching conditions to be satisfied [more on that later]. Moreover, even under conditions of (almost) perfect phase-matching, the efficiency with which harmonics can be generated tends to decrease with the order because of the decreasing magnitude of the electric susceptibilities; in practice, harmonics beyond the third-order are rarely generated with common incident intensities. When the incident intensities become sufficiently large, Eq. (10.4) actually becomes invalid. In this regime, the electric fields are so strong that they can ionize the atoms, i.e., pull the electrons out of their potential well altogether. Interestingly, as the electrons eventually recombine with the positive nuclei, harmonics of very high-order can be efficiently generated – a phenomenon known as *high-harmonic generation*. High-harmonic generation is beyond the scope of this course; here we stick to the *perturbative* regime, where Eq. (10.4) is valid.

10.8.2 Nonlinear refractive index

The second term in the third-order nonlinear polarization [Eq. (10.16)] oscillates at the fundamental frequency ω and is in fact proportional to the EM wave. This gives rise to an interesting phenomenon. Let us assume that all other terms in the nonlinear polarization are negligible (which in fact is often the case unless special care is taken to ensure phase-matching). The total polarization, including the linear term, then reads

$$\boxed{\begin{aligned} P(z,t) &= \varepsilon_0\chi^{(1)}E_0 e^{i(\omega t-kz)} + 3\varepsilon_0\chi^{(3)}|E_0|^2E_0e^{i(\omega t-kz)} + \text{c.c} && (10.28)\\ &= \varepsilon_0\left[\chi^{(1)} + 3\chi^{(3)}|E_0|^2\right]E(z,t) + \text{c.c}, && (10.29)\end{aligned}}$$

where the latter form uses $E(z,t) = E_0\exp[i(\omega t-kz)]$. Following the analysis presented in Section 10.6, when describing the Pockel's effect, we find that the refractive index now obeys

$$n^2 = 1+\chi^{(1)}+3\chi^{(3)}|E_0|^2. \tag{10.30}$$

Recalling that the usual, *linear* refractive index $n_0 = \sqrt{1+\chi^{(1)}}$, and that the intensity of an EM wave is related to the absolute value squared of the complex amplitude as $I = 2n|E_0|^2/(c\mu_0)$, we can express the refractive index in the form

$$\boxed{n \approx n_0 + n_2 I.} \tag{10.31}$$

Equation 10.31 shows that, because of the third-order nonlinearity, the refractive index depends (linearly) on the light intensity. This phenomenon is known as the *Kerr effect* or *self-phase modulation*, since the propagating wave is nonlinearly modifying its own phase. In deriving Eq. (10.31), the nonlinear perturbation to the refractive index was assumed small, which indeed is typically the case. The coefficient n_2 is known as the *nonlinear refractive* index and it typically has very small value. For example, in silica glass, $n_2 \approx 3\times10^{-20}\mathrm{m}^2/\mathrm{W}$, highlighting how large intensities are required for the effect to be significant. Such large intensities are nevertheless often encountered, and effects of self-phase modulation observed. They are particularly important in optical fibres, which can readily be several hundreds of kilometres long, such that the effect can accumulate [exercise 10.7].

10.8.3 Self-focusing and solitons

Self-phase modulation underpins many important effects. Consider a Gaussian beam propagating in a third-order nonlinear material. Assuming that the waist of the beam occurs before the material, we would expect the beam to spread (diffract) as it propagates [see Fig. 55]. However, because of the Kerr effect, the refractive index close to the intensity maximum of the beam will be larger than at the wings (assuming $n_2 > 0$ which is often but not always the case). In other words, the medium through which the beam propagates exhibits a refractive index profile that varies along the transverse spatial dimension, reaching a maximum at the centre of the beam. Such an index variation effectively acts as a (focusing) lens: the phase shift close to the beam centre is larger than at the wings [see also discussion in Section 9.6]. As a consequence, if the intensity of the beam is sufficiently large, the beam can undergo *self-focusing* in the medium [see Fig. 55]. We call this phenomenon self-focusing, since the beam itself is responsible for the refractive index change.

If a beam undergoes self-focusing, the intensity will further increase (as energy is being confined to a smaller area), which will give rise to even stronger self-focusing. However, as the beam gets narrower, the impact of diffraction will also increase. A balance can be reached where the self-focusing exactly balances

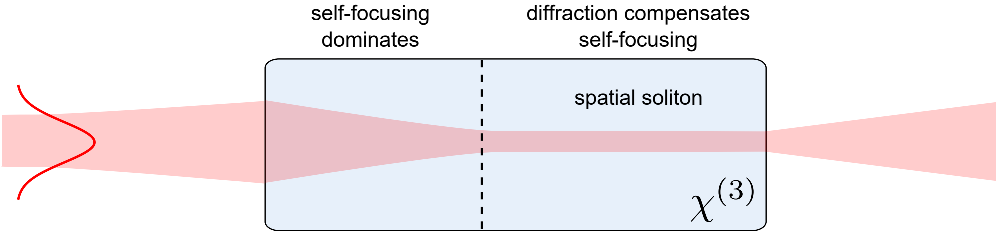

Figure 55: Schematic illustration of the formation of a spatial soliton. A beam propagating in a $\chi^{(3)}$ medium undergoes self-induced focusing, which can eventually be balanced by diffraction. This balance can give rise to a spatial *soliton*: beam that maintains constant shape. After the medium, self-focusing vanishes and ordinary diffraction takes over.

diffraction. In this case, the beam will propagate unchanged, i.e., without spreading or narrowing. Such a beam is known as a spatial *soliton*: a wave that maintains constant shape as it propagates. Solitons are universal in nature, manifesting themselves in a variety of nonlinear systems ranging from cold atoms and hydrodynamics to optics. In addition to the spatial domain, in optics they can also manifest themselves in the temporal domain, where the Kerr nonlinearity balances temporal broadening due to dispersion, allowing a short pulse to propagate unchanged through a nonlinear material (such as an optical fibre).

Problems

10.1 Consider an arbitrary potential function $U(x)$ with a minimum at x_0. Use Taylor series expansion to show that, for displacements close to the minimum ($x \approx x_0$), Hooke's law arises.

10.2 Consider the anharmonic electron oscillator model described by Eq. (10.7) with a driving field given by Eq. (10.15) Use perturbation theory to derive an expression for the second-order nonlinear susceptibility $\chi^{(2)}(2\omega)$. To this end, replace the driving term $E(t) \to \lambda E(t)$, where λ is a control parameter that will be set to unity at the end of the calculation and expand the solution as a power series with respect to λ,

$$x(t) = \lambda x^{(1)} + \lambda^2 x^{(2)} + \lambda^3 x^{(3)} + ... \tag{10.32}$$

For the above equation to be a solution for any value of λ, we require that the terms in Eq. (10.7) proportional to λ, λ^2, λ^3 and so on satisfy the equation separately. You can thus solve for the terms $x^{(n)}$ iteratively by first solving the equation for $x^{(1)}$ and using your solution to solve for $x^{(2)}$ and so on.

10.3 Use simple symmetry arguments to show that, for all centrosymmetric materials, $\chi^{(2)} = 0$. To this end, use the fact that an inversion symmetry stipulates that inverting the direction of the electric field $E(t) \to -E(t)$ must result in a similar inversion of the direction of the dielectric polarization, $P(t) \to -P(t)$.

10.4 Explicitly show that the nonlinear polarizations described by Eq. (10.16), Eq. (10.20), and (10.27) are correct for the incident waves quoted.

10.5 Derive Eq. (10.21) for the incident wave quoted, and show that the refractive index appearing in the equation is given by Eq. (10.23).

10.6 Consider a device, where light is equally split into two paths with equal lengths, but one of the paths contains a Pockels cell that gives rise to a phase-modulation as described by Eq. (10.24). Show that when the two paths are combined, the superposition field exhibits an amplitude modulation. Devices of this kind are known as Mach-Zehnder modulators.

10.7 Consider a 100 mW laser beam at 1550 nm propagating in a standard silica optical fibre with a core diameter of 10.4 $\mu\mathrm{m}$. Assuming the beam exhibits uniform spatial profile for the sake of simplicity, estimate the propagation length L that results in a nonlinear phase shift of 2π. The nonlinear refractive index of silica is about 3×10^{-20} m^2/W.

11 Coherence

Coherence is an important property of all waves, and it is intimately related to waves' ability to produce stable interference patterns. It is, however, difficult to define in any concise and unambiguous manner. Indeed, the concept of coherence broadly encompasses "all properties of the correlation between physical quantities of a single wave, or between several waves or wave packets". Here, we will only describe some basic principles of coherence.

A single light field is said to be **coherent** when there is a fixed phase relationship between the electric field values at different locations in space or at different times. If the phase relationships are not fixed, but change randomly in time, we say that the light field is **incoherent**. If the phase relationship changes somewhat with time (but is not totally random), we say that the field is **partially coherent**. Depending on whether we are interested in the phase-relationships at constant time but different positions in space (or vice versa), we typically distinguish between **spatial** and **temporal** coherence. Simple physical interpretations of these quantities are as follows:

Interpreting coherence

- **Temporal coherence:** describes how well we can predict the phase (or value) of a wave at some future time if we know the wave's phase (or value) at some initial time. For a temporally coherent field, we can easily predict the value of the field at any time: there is little or no randomness.
- **Spatial coherence:** describes how well we can predict the phase (or value) of a wave at some spatial location $\vec{\mathbf{r}}_2$ if we know the wave's phase (or value) at some other location $\vec{\mathbf{r}}_1$. For a spatially coherent field, we can easily predict the value at any spatial location $\vec{\mathbf{r}}_2$ based on the knowledge the field value at $\vec{\mathbf{r}}_1$ at all times.

In this Section, we will be mostly interested in temporal coherence.

In more general terms, coherence always involves two waves, and basically describes how well those two waves are "correlated", or what they have in common. When talking about the coherence of single light fields, we are really talking about the coherence between the light field and a delayed replica of itself (e.g. temporal coherence) or the coherence between two waves corresponding to different spatial locations (spatial coherence). So always two waves. In a nutshell, **two beams of light are coherent when the phase difference between their waves is constant; they are incoherent if there is a random or changing phase relationship.** As we shall see, stable interference patterns are formed only between fields that are coherent with one another.

11.1 Example of temporal coherence

Let us consider a monochromatic plane wave with frequency ω at some point $\vec{\mathbf{r}}_0$ in space. For brevity, we drop the $\vec{\mathbf{r}}$-dependence and assume linear polarization, which then allows us to consider a scalar field:

$$E(t) = E_0 e^{i\omega t}. \tag{11.1}$$

This wave [see Fig. 56(a)] is perfectly coherent. Indeed, if we know that the phase at some time t_0 is $\phi_0 = \omega t_0$ we can predict the phase at any other time: $\phi(t) = \omega t = \omega(t - t_0) + \phi_0$. Alternatively, consider the phase

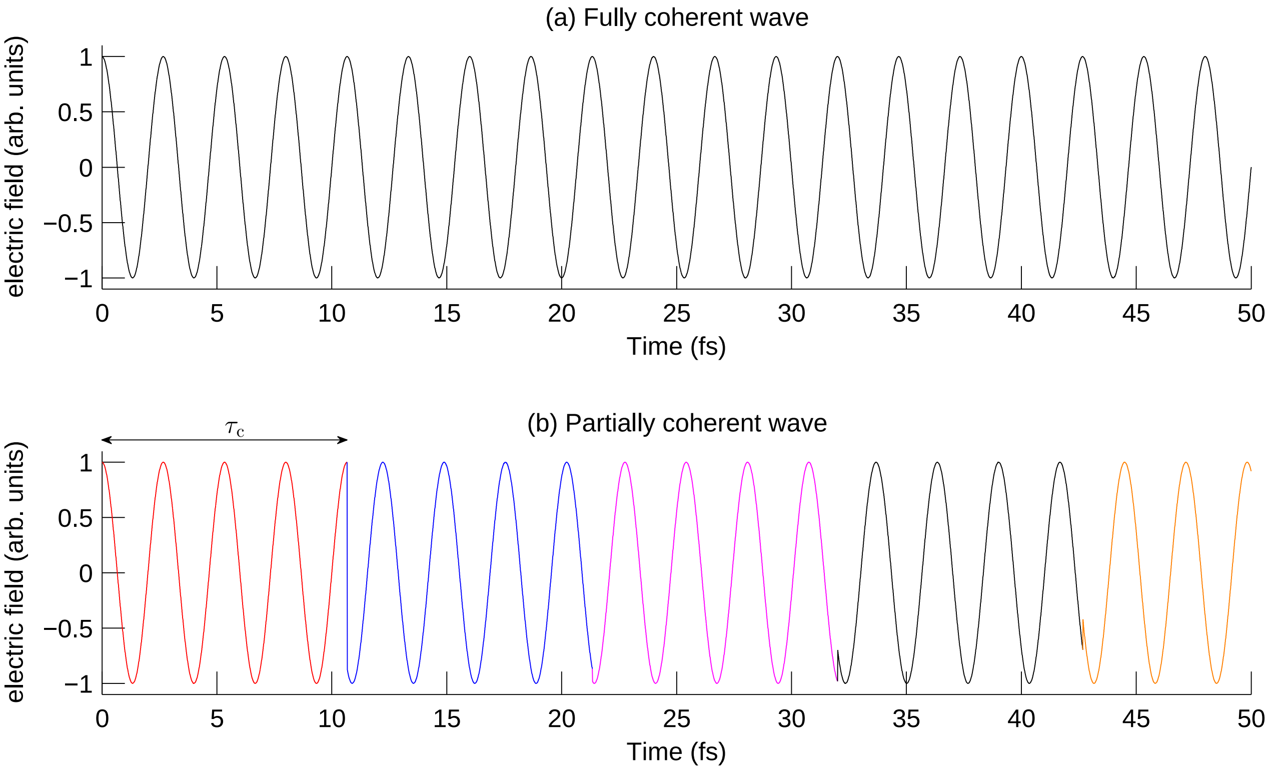

Figure 56: Electric fields corresponding to EM waves with (a) perfect and (b) partial temporal coherence. The electric field in (b) is constructed such that, every τ_c in time, the wave undergoes a random phase shift. Different colours highlight oscillations after different phase jumps.

difference between the wave and a replica of itself delayed by a time $\Delta\tau$. At any given time, we have

$$\Delta\phi(t) = \omega t - \omega(t - \Delta\tau) = \omega\Delta\tau. \tag{11.2}$$

As can be seen, the phase difference is constant with time: the field is perfectly coherent with itself, and is thus said to be perfectly coherent.

Let us now consider a wave of the form

$$E(t) = E_0 e^{i\omega t + i\phi(t)}, \tag{11.3}$$

where $\phi(t)$ is a random process[85] that changes the phase of the wave once every τ_c. In other words, after a time τ_c has elapsed, the phase of the wave randomly changes to a new value [see Fig. 56(b)]. This wave no longer displays perfect coherence. Indeed, if we know the phase at some time t_0, there is no way for us to predict what it will be after a time τ_c, since the phase will have changed randomly. We can, however, predict what the phase will be after a time interval $t - t_0 < \tau_c$, so there is still some coherence in the wave. Accordingly, we say that the wave is **partially coherent**. This observation allows us to roughly define what we mean with "partial" temporal coherence:

[85] Function probably doesn't make sense in this context!

Classification of temporal coherence

- **Perfect coherence:** the wave is perfectly correlated with itself at all times. In other words, knowing the phase at some time t_0 allows us to predict the phase at any future time.
- **Partial coherence:** the wave is correlated with itself for times $t < \tau_c$. In other words, knowing the phase at some time t_0 allows us to predict the phase for times $t < t_0 + \tau_c$.
- **Full incoherence:** the wave is not correlated with itself at any times t. In other words, we can make no predictions of future times.

An alternative way to interpret the above is to note that the phase difference between the original wave and a delayed replica of itself exhibits constant phase difference precisely for delays smaller than τ_c. For delays larger than τ_c, the phase difference wanders in time.

11.2 Coherence time and coherence length

The time τ_c referred to above corresponds to a **coherence time**. In general, real light waves are influenced by complex random processes (the random process introduced above was very simple), which stochastically change the phase and amplitude of the wave. For arbitrary light waves, the coherence time τ_c is defined as the delay over which the wave's phase or amplitude of wanders by a significant amount (and hence the correlation decreases by significant amount). Certainly, τ_c used in our little example above fits this definition, and hence can be understood as a coherence time.

The distance that light travels in vacuum during one coherence time is known as the coherence length. It is defined as

$$\boxed{l_c = c\tau_c.} \tag{11.4}$$

We can rewrite our conditions for coherence/partial coherence/incoherence above in terms of the coherence time and the coherence length:

Classification of temporal coherence using coherence time

- **Perfect coherence:** coherence time $\tau_c = \infty$.
- **Partial coherence:** coherence time τ_c finite but non-zero.
- **Full incoherence:** coherence time $\tau_c = 0$.

It should be clear that all real light waves are partially coherent.

11.3 Interference

Coherence plays a key role in interference. Let us consider two partially coherent waves with identical amplitudes and frequencies:

$$E_1(t) = E_0 e^{i\omega t + i\phi_1(t)}, \tag{11.5}$$

$$E_2(t) = E_0 e^{i\omega t + i\phi_2(t)}, \tag{11.6}$$

where $\phi_{1,2}(t)$ can be constant or they can depend on time in random (or not) fashion. They can also evolve independently or in a perfectly correlated manner. Imagine that these two waves are superimposed together, and we detect the resulting intensity using our eyes or a photodetector. The superposition wave $E = E_1 + E_2$ can be written as:

$$\boxed{E(t) = E_0 e^{i\omega t + i\phi_1(t)} \left[1 + e^{i\Delta\phi(t)}\right],} \tag{11.7}$$

where $\Delta\phi(t) = \phi_2(t) - \phi_1(t)$ is the difference between the phases of the two waves.

Our eyes or photodetectors do not detect electric fields directly, but rather the intensity of the electromagnetic wave. A long time ago, we have argued that the intensity of a complex EM wave can be obtained by taking the absolute value squared of the complex electric field. Thus, the intensity corresponding to our superposition field is given by

$$\boxed{I(t) = 2I_0 \left[1 + \cos(\Delta\phi(t))\right],} \tag{11.8}$$

where $I_0 = |E_0|^2$. Below we will consider two different limit situations.

Assume that the waves are mutually coherent. This implies that the phase difference $\Delta\phi(t) = \Delta\phi$ is a constant: indeed, by definition, two beams of light are coherent when the phase difference between their waves is constant. In this case, the superposition intensity $I(t)$ will also be constant and not fluctuate with time. If the constant phase difference is equal to $\Delta\phi = 2m\pi$, with m an integer, we see that the waves will **interfere constructively**, yielding a resultant intensity of $I(t) = 4I_0$. On the other hand, if the phase difference is $\Delta\phi = (m+1)\pi$, the waves **interfere destructively**, yielding a constant resultant intensity of $I(t) = 0$. The key observation is this: the interference between two mutually coherent waves is stable and does not fluctuate with time. If we can systematically change the phase-difference $\Delta\phi$, using for example a delay line for one of the waves, we can trace an interference pattern consisting of periodically repeating minima or maxima. Another key point to note is that the ability for two waves to stably interfere does not depend on how the individual fields fluctuate: so long as they fluctuate in an identical way (e.g. such that the phase difference is constant), stable interference ensues. This highlights the fact that coherence truly describes how well a given wave is correlated with another.

Let us now assume that fields are **not** at all coherent with one another. The phase difference $\Delta\phi(t)$ is no longer constant. Furthermore, because $\phi_{1,2}$ are random processes, the phase difference $\Delta\phi(t)$ will also change randomly. This of course implies similar random fluctuations for the intensity $I(t)$ [see Eq. (11.8)]. In practice, we can only resolve intensity fluctuations that occur over sufficiently long timescales. This is because all detectors come with some finite response time. For our eyes, the response is some milliseconds, whilst for state-of-the art photodetectors we are talking about some tens of picoseconds. Mathematically, we can say the

Eq. (11.8) gives us the “instantaneous” intensity, whilst the intensity we can actually detect is given by the time average:

$$\boxed{I_\mathrm{d}(t) = \frac{1}{2T}\int_{-T}^{T} I(t),} \tag{11.9}$$

where T is a characteristic response time of a detector. Note that, in theoretical analyses, we often assume T to be very large, and define the detected intensity as the limit when $T \to \infty$.

If the two waves are not mutually coherent, such that the phase difference $\Delta\phi$ is totally random, the $\cos(\Delta\phi(t))$ term averages to zero. In this case, the detected intensity from the superposition of our two waves reads simply:

$$\boxed{I_\mathrm{d}(t) = 2I_0.} \tag{11.10}$$

In other words, there is no interference observed at all for waves that are mutually incoherent!

11.4 Characterising the temporal coherence of a wave of light

The above discussion suggests an experimental method to quantitatively examine the temporal coherence of a light field. Specifically, let us consider a situation where a light wave is passed through an interferometer, such as the Michelson interferometer shown in Fig. 57. Here, a beam splitter divides the input wave into two arms. Each of those waves is then reflected back to the beam splitter, and the superposition wave is detected using e.g. a photodetector. We denote the distances between the beam splitter and the mirrors as l_1 and l_2, and further assume that one of the mirrors is fixed while the other one can be moved towards or away from the beam splitter. In this way, the distance l_2 can be continuously adjusted.

We write the EM wave just before it is split at the beam splitter as:

$$E_i(t) = E_0 e^{i\omega t + i\phi(t)}, \tag{11.11}$$

where the additional phase factor $\phi(t)$ can be constant or depend on time. As the beams travel back and forth their respective arms, they accumulate phase shifts $\exp(2ikl_n)$. Ignoring the common phase factor arising from propagation to the photodetector, the superposition field at the detector will be

$$E(t) = \frac{E_0}{2} e^{i\omega t + i\phi(t-\tau_1) - 2ikl_1} + \frac{E_0}{2} e^{i\omega t + i\phi(t-\tau_2) - 2ikl_2} \tag{11.12}$$

$$= \frac{E_0}{2} e^{i\omega t + i\phi(t-\tau_1) - 2ikl_1}\left[1 + e^{i\phi(t-\tau_2) - i\phi(t-\tau_1) - 2ik(l_2 - l_1)}\right], \tag{11.13}$$

where $\tau_n = 2l_n/c$ is the time taken for each beam to travel to its respective mirror and back. Note that the factor $E_0/2$ ensures that the beam splitter divides *intensity* equally (also note that both beams pass through the splitter twice). Further note that phase shifts from reflection and transmission through the beam splitter cancel out since both arms will be reflected and transmitted once.

Calculating the instantaneous intensity is straightforward:

$$\boxed{I(t) = \frac{I_0}{4}\left[2 + 2\cos(\Delta\phi(t) + k\Delta l)\right],} \tag{11.14}$$

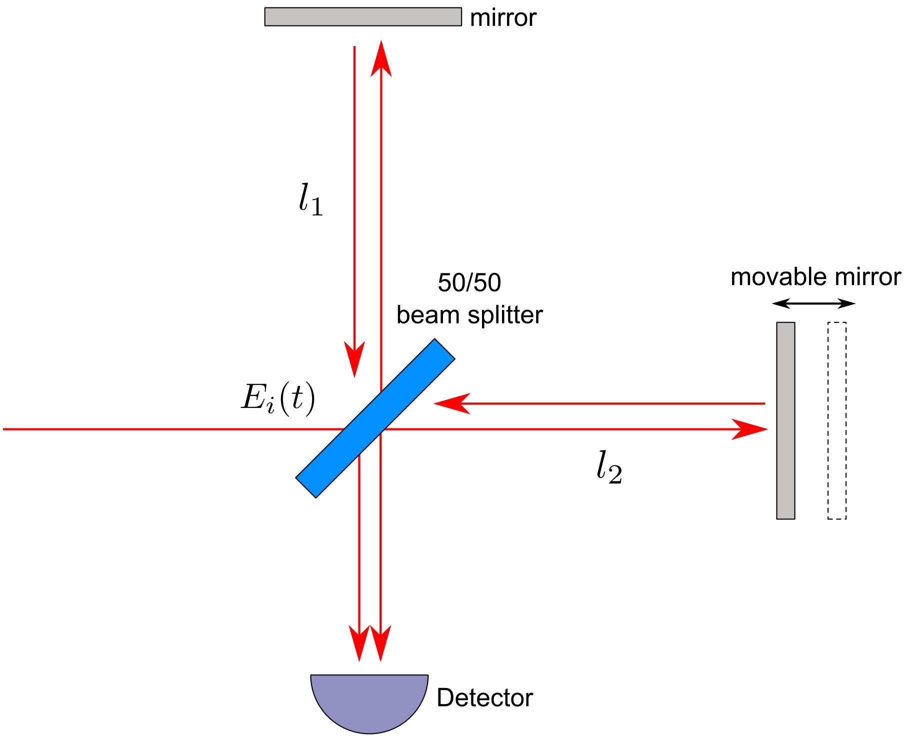

Figure 57: Schematic illustration of a Michelson interferometer. A beam splitter divides an input EM wave into two arms with lengths l_1 and l_2, and subsequently combines the two reflected waves. Time-averaged intensity of the superposition field is then detected with a photodetector. By translating one of the mirrors, it is possible to probe the interference between the input wave and its time-delayed replica. From the resulting interferogram, we can deduce the wave's temporal coherence.

where $I_0 = |E_0|^2$ and $\Delta l = 2(l_2 - l_1)$ is the length difference between the two arms and $\Delta\phi(t) = \phi(t - \tau_1) - \phi(t - \tau_2)$. Note that, if the interferometer arms have exactly the same length, we have $\tau_1 = \tau_2$ and hence $\Delta\phi(t) = 0$. It is traditional to write this equation in terms of the relative temporal delay between the two interferometer arms rather than the path difference as:

$$\boxed{I(t) = \frac{I_0}{4}\left[2 + 2\cos(\Delta\phi(t) + \omega\Delta\tau)\right],} \tag{11.15}$$

where we used $k = \omega/c$ and defined the relative temporal delay $\Delta\tau$ as

$$\boxed{\Delta\tau = \tau_2 - \tau_1 = 2c(l_2 - l_1).} \tag{11.16}$$

As in the preceding section, the actual detected intensity will be the time-average of Eq. (11.15). Depending on the nature of the phase fluctuations $\phi(t)$ (and hence $\Delta\phi(t)$), we can obtain very different behaviours when the delay is scanned by manually moving one of the mirrors:

Temporal interference of partially coherent fields

- If the field exhibits perfect temporal coherence, $\phi(t)$ is constant. Accordingly, $\Delta\phi(t) = 0$ for all t. Scanning one of the interferometer arms (i.e., continuously changing e.g. l_2) results in the detection of a perfect interference pattern, as shown in Fig. 58(a).

- If the field is exhibits total temporal incoherence, $\phi(t)$ is fluctuating at infinitely short timescales ($\tau_\mathrm{c} = 0$). In this case $\Delta\phi(t) = 0$ only for $\Delta\tau = 0$, i.e., when the interferometer arms have precisely the same lengths. For any other delay, the $\cos(\Delta\phi(t) + \omega\Delta\tau)$ term will average to zero. Accordingly, our detected intensity will be $I(0) = I_0$ over an infinitesimally narrow range of delays around $\Delta\tau = 0$, and $I(t) = I_0/2$ for all other $\Delta\tau$ [see Fig. 58(b)].

- For real, partially coherent light fields, $\phi(t)$ is fluctuating but not totally randomly. Rather, $\phi(t)$ stays almost constant for delays much smaller than the coherence time τ_c, and only exhibits erratic changes over delays larger than the coherence time. Accordingly, the phase-difference $\Delta\phi(t) \approx 0$ for delays $\Delta\tau \ll \tau_\mathrm{c}$, giving rise to beautiful interference fringes [see Fig. 58(c)]. In contrast, for very large delays $\Delta\tau \gg \tau_\mathrm{c}$, the phase-difference $\Delta\phi(t)$ oscillates randomly, such that the cosine term in Eq. (11.15) averages to zero and we lose the interference fringes. In between these two limits (very small and very large delay), we find a uniform transition whereby the interference fringes gradually disappear, as shown in Fig. 58(c). A key observation is the following: **We can measure the coherence time of a light source by passing it through a (Michelson) interferometer, and by measuring at what delay do the interference fringes disappear!**

11.5 Interferograms and fringe visibility

All real light fields are partially coherent, and therefore exhibit an interference pattern similar to the one shown in Fig. 58(c). The time-scale over which the interference fringes vanish corresponds to the source's coherence time. Accordingly, the coherence time of a source can be measured by using a Michelson interferometer with a variable arm-length, and gauging at what delay do the interference fringes vanish.

The time-averaged interference pattern observed when the relative delay between the interferometer arms is changed [c.f. Fig. 58] is known as an **interferogram**. It can be shown (left as an exercise), that the time-average of Eq. (11.15) can be written as:

$$\boxed{I_\mathrm{d}(\Delta\tau) = \frac{I_0}{2}\left[1 + |g(\Delta\tau)|\cos(\omega\Delta\tau + \phi)\right],} \tag{11.17}$$

where $\phi = \arg[g(\Delta\tau)]$ and the quantity $g(\Delta\tau)$ is known as the **complex degree of first-order coherence**. This quantity governs *the ability of a wave to interfere with a time delayed replica of itself.* It has a maximum value of unity at zero delay, and vanishes for $\Delta\tau \gg \tau_\mathrm{c}$, i.e., when the optical path difference between the interferometer arms is much greater than the coherence length $l_\mathrm{c} = c\tau_\mathrm{c}$.

The temporal periodicity of the interferogram is equal to the period of the EM wave: $\Delta\tau_\mathrm{per} = 2\pi/\omega$. If the coherence time is much larger than this (as is typically the case), then the degree coherence $g(\tau)$ varies slowly in comparison to that period. In this case, the magnitude $|g(\tau)|$ can be measured by monitoring the **visibility** of

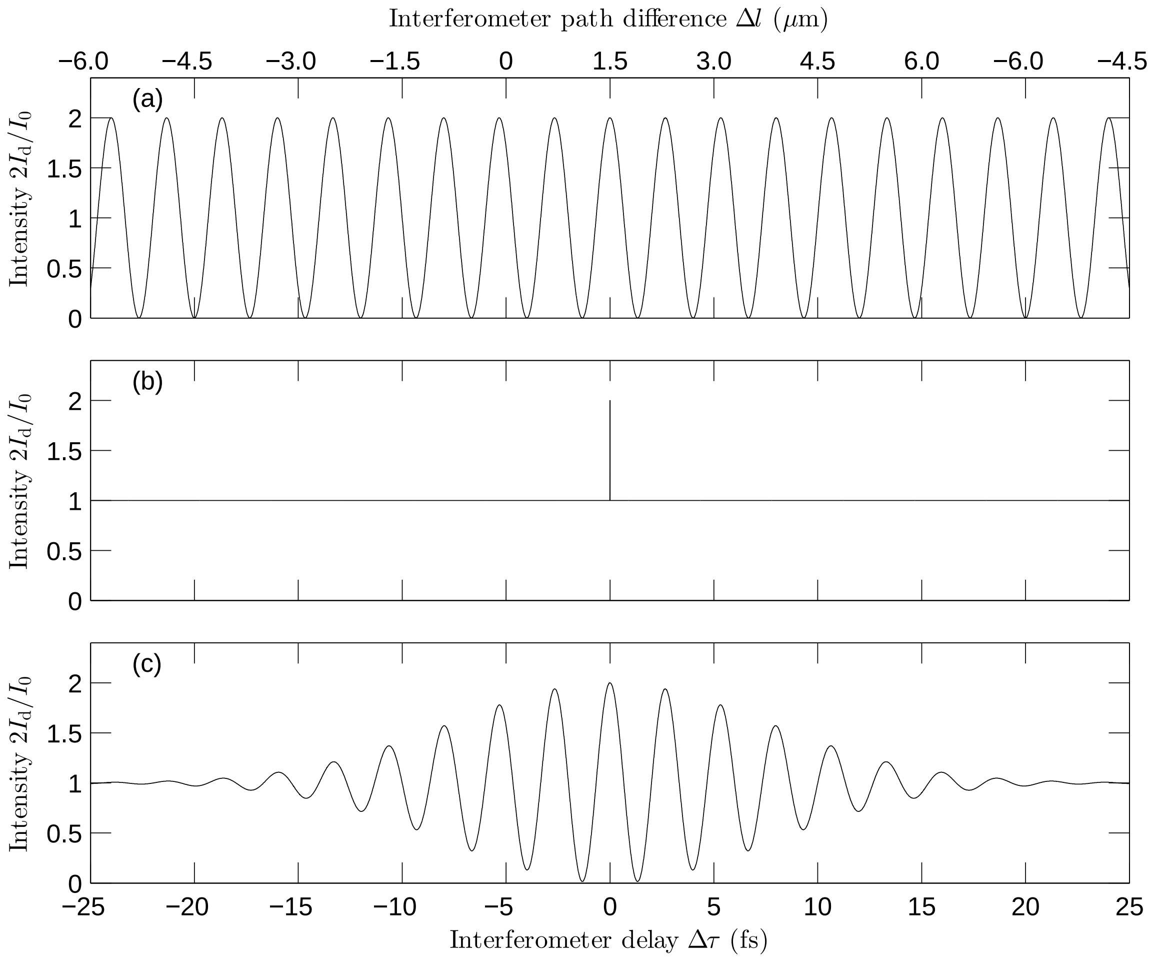

Figure 58: Detected (time-averaged) intensity as a function of interferometer delay $\Delta\tau$ (bottom x-axis) and corresponding interferometer path length difference Δl (top x-axis). (a) Field with perfect temporal coherence ($\tau_\mathrm{c} = \infty$), (b) totally incoherent field ($\tau_\mathrm{c} = 0$) and (c) partially coherent field with $\tau_\mathrm{c} \approx 15$ fs. In general, interference fringes are observed for delays $\Delta\tau < \tau_\mathrm{c}$.

the interference pattern as a function of time delay. Here, the visibility of the interference fringes is defined as:

$$\boxed{\mathcal{V} = \frac{I_\mathrm{max} - I_\mathrm{min}}{I_\mathrm{max} + I_\mathrm{min}},} \tag{11.18}$$

where I_max and I_in are the maximum and minimum values that I_d takes as $\Delta\tau$ is slightly varied in the vicinity of some fixed delay $\Delta\tau_0$. Using $|g(\Delta\tau)| \approx |g(\Delta\tau_0)|$ in Eq. (11.17), we easily obtain [exercise]:

$$\boxed{\mathcal{V}(\Delta\tau_0) = |g(\Delta\tau_0)|.} \tag{11.19}$$

It should be clear that $|g(\Delta\tau)|$ can be understood as the *envelope* of the rapidly oscillating function $2I_{\mathrm{d}}(\Delta\tau)/I_0$ [see Fig. 59].

11.6 Spectrum and coherence

The power spectrum and coherence of a light source are closely intertwined. In fact, it turns out that *the power spectrum of a light source is proportional to the Fourier transform of the complex degree of coherence*[86]*:*

$$\boxed{S(\omega) \propto \mathcal{F}[g(\tau)].} \tag{11.20}$$

Furthermore, the proportionality coefficient is nothing but the time-averaged intensity of the light wave. This relationship immediately raises interesting implications. Recalling from Fourier analysis that the more localized a signal the more spread out its Fourier transform, we can draw the following conclusion:

Relationship between coherence time and spectral width

A field with a short coherence time [i.e., a highly localized $g(\tau)$] must be associated with a broad spectrum [i.e., a spread out Fourier transform of $g(\tau)$]. Accordingly, the spectral width of a partially coherent light source is inversely proportional to its coherence time:

$$\Delta f \approx \frac{1}{\tau_{\mathrm{c}}}. \tag{11.21}$$

The approximately equal sign ensues from the fact that we have not really defined what kind of widths we are talking about. If we are referring to full-width at half maxima, then the relationship should include a multiplicative factor (of the order of unity) that depends on the shape of the spectrum. Nevertheless, from the inverse relationship, we can readily draw the following conclusions:

- The bandwidth of a perfectly coherent sources (with $\tau_{\mathrm{c}} = \infty$) is $\Delta f = 0$. In other words, a perfectly coherent source is monochromatic.
- The bandwidth of a totally incoherent sources (with $\tau_{\mathrm{c}} = 0$) is $\Delta f = \infty$. In other words, a totally incoherent source would need to have infinite bandwidth.
- The bandwidth of a real partially coherent light field is finite (but non-zero). The source's bandwidth governs its coherence time.

The relationship between the power spectrum and the interferogram of a EM wave is illustrated in Fig. 59.

[86] The reason it so turns out is that, formally, the complex degree of coherence corresponds to the autocorrelation of the field divided by the total time-averaged intensity. Because the Fourier transform of a field autocorrelation is the power spectrum, the "turning out" follows.

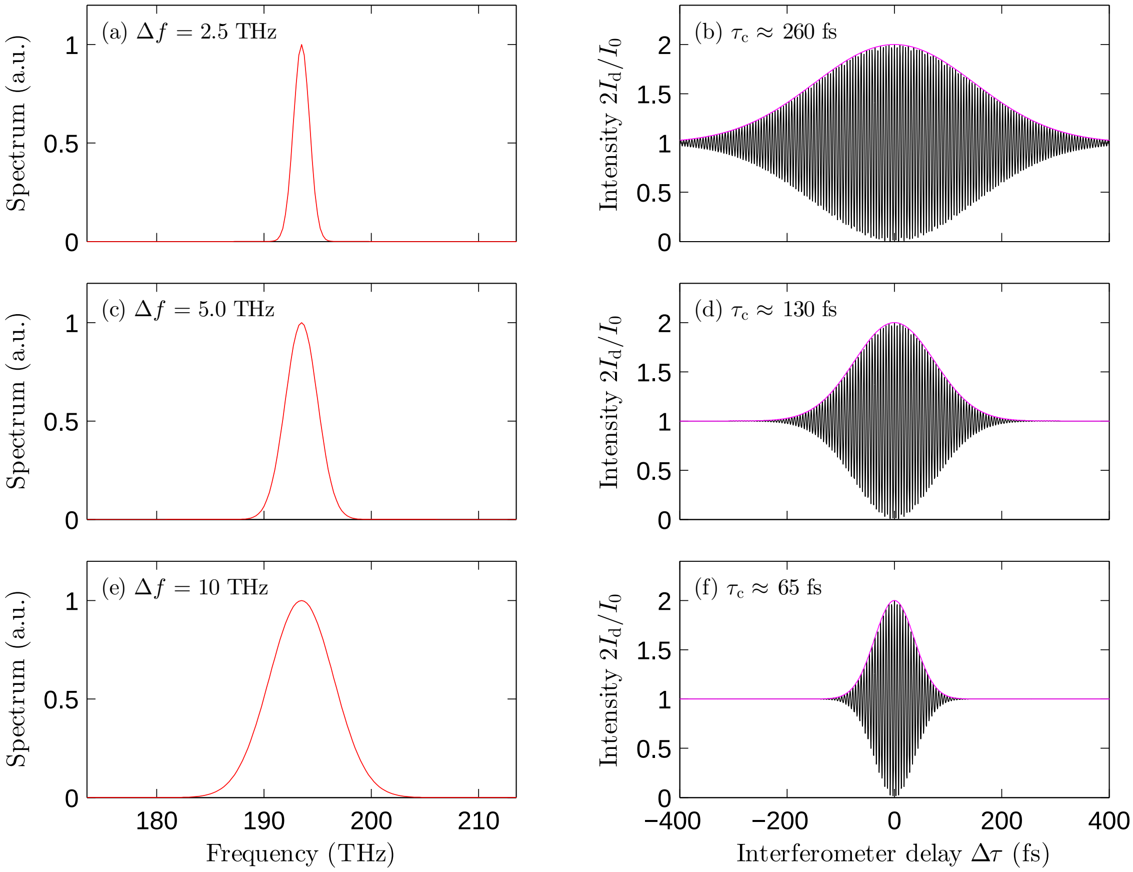

Figure 59: (a,c,e) Power spectra and (b,d,f) corresponding interferograms for three different spectral widths (and hence, coherence time): (a,b) $\Delta f = 2.5$ THz and $\tau_c \approx 260$ fs, (c,d) $\Delta f = 5$ THz and $\tau_c \approx 130$ fs, (e,f) $\Delta f = 10$ THz and $\tau_c \approx 65$ fs. As can be seen, a broad spectrum is associated with a localized interferogram (i.e., short coherence time) and vice versa. The magenta curves in (b,d,f) correspond to $1 + |g(\Delta\tau)|$, highlighting how $|g(\Delta\tau)|$ describes the envelope of the interferogram.

11.7 Applications

The ideas presented above can be harnessed for practical applications. For example, a Michelson interferometer can be used to measure the spectrum of a light wave. Here, one simply measures the interferogram, and subsequently computes its Fourier transform to obtain the spectrum. This technique is known as Fourier transform spectroscopy, and it is particularly popular in the infrared spectral region.

Optical coherence tomography is another interesting application of coherence. Here, one uses a source with a low-coherence (i.e., short coherence time) in conjunction with a Michelson interferometer where one (reference) arm can be axially scanned, whilst the other arm contains a sample under test. Because the source has such low-coherence, interference is observed only when the interferometer path-lengths are equal. When the sample is moved laterally, the interferometer path length changes due to the sample's non-uniform surface (or wherever the reflection occurs from); recovering the interference requires the reference arm to be axially scanned. Thus, the surface features can be mapped by laterally moving the sample in discrete steps, and by adjusting the reference arm to recover the interference at each step. The motion of the reference arm immediately gives the surface features.

11.8 Correlation functions

Our discussion above was fairly qualitative. Quantitative (or formal) analysis of coherence requires the use of correlation functions. For example, considering the temporal coherence of a single EM wave, correlations between the field and its time-delayed replica is described by the *autocorrelation* function, defined as:

$$\boxed{G(\Delta\tau) = \langle E^*(t)E(t+\Delta\tau)\rangle.} \tag{11.22}$$

Here the angle brackets $\langle\cdot\rangle$ denote an *ensemble average* taken over many realizations of the random field $E(t)$. This means that the wave is produced repeatedly under the same conditions, with each trial yielding a different wavefunction. For waves that are statistically *stationary*[87], the ensemble average can be shown to be equal to a time average:

$$\boxed{G(\Delta\tau) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} E^*(t)E(t+\Delta\tau).} \tag{11.23}$$

The autocorrelation function describes how correlated (i.e., "similar") is the field with its own time-delayed replica. If the two fields (original and replica) are fluctuating without any correlations, the autocorrelation is zero. In the language of coherence theory, the autocorrelation function is known as the *temporal coherence function*. It is easy to see that $G(0) = I_\mathrm{d}$. This means that the autocorrelation function carries information on both the coherence (correlations) and averaged intensity of the field. To obtain a measure that describes coherence alone, we can normalise the autocorrelation function to the average intensity. This turns out to produce the complex-degree of temporal coherence used above:

$$\boxed{g(\Delta\tau) = \frac{G(\Delta\tau)}{G(0)} = \frac{\langle E^*(t)E(t+\Delta\tau)\rangle}{\langle E^*(t)E(t)\rangle}} \tag{11.24}$$

From this definition, it is easy to see that $0 \leq |g(\Delta\tau)| \leq 1$. Furthermore, we can now clearly appreciate that $|g(\Delta\tau)|$ is indeed a measure of the degree of correlations between $E(t)$ and $E(t+\Delta\tau)$. Following these ideas, it should be evident that formal examination of correlations between any two fields at any two points can be achieved by considering appropriate *cross-correlation* functions.

[87] This means that the statistical properties of the wave do not depend on time. With the exception of pulsed light, most light fields obey this condition fairly faithfully.