1

Introduction

1.1 Masers and lasers

The words 'maser' and 'laser' were originally acronyms: MASER standing for microwave amplification by stimulated emission of radiation, and LASER for the very similar phrase with 'light' substituted for 'microwave'. The important point is that masers and lasers are both derived from the stimulated emission process, and the only difference between them is a rather arbitrary distinction, based on the frequency of radiation they emit. Masers, as laboratory instruments, in fact pre-dated lasers by several years, and both had been completed as practical instruments before the discovery of astrophysical maser sources.

Although this book is about masers, most people are probably more familiar with lasers, so keeping in mind that the two things are very similar, we will begin by considering lasers. Most people probably own several lasers: lasers are used to interpret the information stored on CD and DVD discs; they are also used in many computer printers. Even if they have only a vague idea about how they work, and view lasers as some sort of 'black box', tube, or chip that emits light, most people will probably be aware that this light is in some way 'special' – that is, it has properties that make it different from the light emitted by, say, a filament electric light bulb. What are these important characteristics? Given time to ponder on this question, most people would proably come up with a list something like this to summarize the important properties of laser light:

- The laser light is extremely bright, or intense.
- The beam is tightly focussed.
- The light has a distinct colour: it is not composed of many colours, spanning the whole visible spectrum, like sunlight or the light from a filament lamp.
- There is an additional property, harder to define, which is based on the idea that light waves from a laser line up, peak to peak and trough to trough, while such arrangements are random in ordinary light. This property allows lasers to make almost magical pictures, called holograms, which somehow retain three-dimensional (3-D) information in a two-dimensional (2-D) photograph.

To understand these properties, it is necessary to peer inside the black-box, the laser itself, and look at what goes on inside. As with the properties of the light it emits, a lot of people would probably be able to list the main components of a laser:

- An active ingredient (a tube of gas, a crystal, or semiconductor wafer) emits and amplifies the laser radiation.

- Usually, a pair of mirrors is present, which bounce the laser light back and forth many times through the active material.
- One of the mirrors must be ‘half-silvered’, so that some of the light can escape to form the external beam.
- An electrical power supply provides the energy to drive, or ‘pump’, the laser, perhaps by firing a flashlight onto the active ingredient, or by some other means.
- What is the mysterious active ingredient? Again, a lot of people would probably have some idea that these are often atoms, ions or molecules, though not in semiconductor lasers.
- Assuming that the active ingredients are atoms or molecules, many people would know that they need to have something called a ‘population inversion’, which is rather unusual, and that it is the job of the power supply, and its ‘pumping’, to sustain this.

How is this arcane collection of components supposed to produce light with the peculiar properties listed in the paragraph above? Considering only gas lasers, which are the type most akin to the astrophysical masers we shall encounter later, we will begin by looking in more detail at their special ingredient: the atoms or molecules that emit the radiation.

1.2 Atoms and molecules

What is it about the atoms and molecules in a gas laser that makes them able to produce laser light? The key is in their internal structure of allowed states and energy levels. The rules of quantum mechanics only allow certain shapes for the arrangement of the electrons in atoms and molecules. Associated with each shape is an allowed energy. Other internal motions, such as the the vibration and rotation of molecules, are governed by similar rules, which also lead to patterns of allowed energy levels. We consider particular molecules that form astrophysical masers in detail in Chapter 5. Real atoms and molecules have a very large number of energy levels, and we will usually call this number N. In fact, N tends to infinity, but a laser can be understood on the basis of a simple model, with just three levels, or $N = 3$. A diagram of this model is drawn in Fig. 1.1, with the levels labelled 1, 2 and 3, in order of increasing energy. The laser light comes from radiation emitted in a transition between the highest pair of energy levels, that is between levels 2 and 3 in Fig. 1.1. This idea of a transition between levels means that an additional explanation is required. When a molecule undergoes a transition, it changes from one quantum-mechanically allowed state to another. In doing this, it changes energy, from the energy associated with the original state, to the energy associated with the new one. If the new level is higher than the original, the molecule must aquire the necessary energy from some source; if it is lower, energy is released by the molecule. An example is the laser transition in Fig. 1.1. If we follow the solid arrow, the molecule changes from state 3 to state 2; in doing this it changes energy by the difference, $E_3 - E_2$, between the original and final energy levels that correspond to these states. As this example is a downward transition, the energy $\Delta E_{32} = E_3 - E_2$ is released. I shall adopt a convention, when referring to energy differences in transitions, that the order of subscripts implies the direction of transition. Thus ΔE_{32} is the energy change in shifting from level 3 to level 2. As level 3 is the higher, in our example, ΔE_{32} is a positive quantity. I shall also only count downward transitions, noting that an upward counterpart accompanies every one of these. With these conventions, the level scheme in Fig. 1.1 has three possible

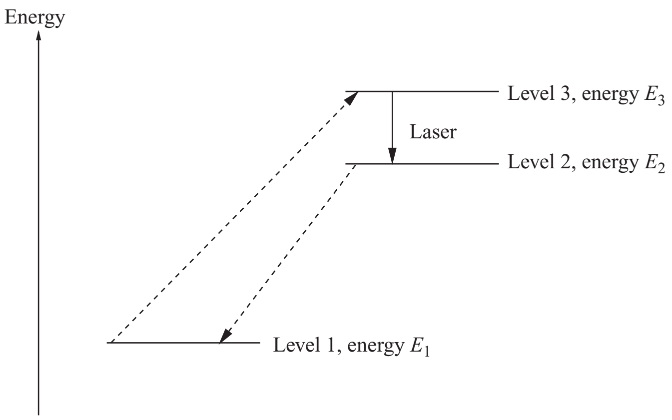

Fig. 1.1. The simplest possible laser system with three levels of energies, E_1, E_2 and E_3. If population can be transferred from level 1 to level 3, and drained from level 2 to level 1, rapidly enough to maintain a population inversion between levels 3 and 2, then a laser can form between between these two levels.

transitions, with energy changes equal to ΔE_{32}, ΔE_{31} and ΔE_{21}. As the levels themselves have fairly precise energies, so do the energy differences between the transitions.

What causes a transition to occur? There are several possibilities, and we consider a few here. A molecule can collide with another atom or molecule in the gas. If the parameters of the collision are suitable, it can cause a transition. A similar process can occur with radiation: if the molecule interacts with light waves of suitable wavelength the molecule can undergo a transition of some energy difference, and take an equal amount of energy from, or add energy to, the radiation field. We deal with these interactions in more detail in Section 1.4. It is worth noting, however, that quantum-mechanical 'selection rules' can prohibit radiative changes via some transitions: only a subset are allowed. Another process that causes transitions, which does not involve interaction, is spontaneous decay: only one energy level, the *ground state*, is actually stable, with a lifetime equal to that of the atom or molecule itself. All other levels have characteristic lifetimes for decay via each radiatively allowed transition to lower levels. Again, we treat spontaneous decay in more detail in Section 1.4. An important consequence of such decays is that all energy levels except the ground state have a finite width. This follows from Heisenberg's uncertainty principle for energy and time. For our example transition, we can write this as

$$\delta E_{32}\delta t_{32} \sim \hbar/2, \qquad (1.1)$$

where δt_{32} is the characteristic lifetime, and δE_{32} is the uncertainty in the transition energy. We assume that this transition, being the laser transition, is radiatively allowed. All transitions have some uncertainty in the transition energy, even if the lower level is the ground state, because of the finite lifetime of the upper state. This uncertainty in energy due to the Uncertainty Principle is known as the natural, or lifetime, width of the transition, but in real systems it is often masked by other broadening processes, which we shall encounter in Chapter 3.

With no further explanation, we can now explain one of the properties of laser light in the first list. A laser emits light of a distinct colour because it comes from a radiatively allowed transition between energy levels of precise energy – as far as is allowed by Heisenberg's Uncertainty Principle. This is one of the characteristics of laser light, but it is not specific to lasers: various lamps, for example sodium street lamps, also have colours based on transitions between atomic energy levels, but this does not make them lasers. A second property of laser light, the narrow, focussed beam, can also be explained at this point. The main factor in controlling beaming is the cavity formed from the two mirrors. These ensure that the path light must take before it can escape is vastly longer between the mirrors than in any other direction, and in this long path length, it can interact with the molecules many times. However, like the colour of the light, the mirrored cavity is not specific to lasers; we could build one around any form of light source.

In the second list, we suggested that a population inversion was necessary for the generation of laser light. What exactly is a population inversion, and in what sense is it unusual or special? First we introduce the concept of population with regard to energy levels. A typical gas laser has a very large number of active atoms or molecules. If the gas obeys the standard gas laws, the number is $N = pV/(kT)$, where k is Boltzmann's constant. For a tube of length 10 cm, radius 0.5 cm containing gas at room temperature and at 1/1000 of atmospheric pressure, we have a volume $V = \pi r^2 l$, equal to $7.85 \times 10^{-6}\,\mathrm{m}^{-3}$, and a total of some 1.9×10^{17} active molecules. This vast number of molecules must be distributed between the available energy levels in some way. In the absence of any other process, collisions between the gas molecules will eventually share them amongst the energy levels according to the equilibrium distribution derived by Boltzmann. The subject of equilibria is discussed in more detail in Section 1.5, but taking the laser transition in Fig. 1.1 as an example, we expect[1] that the ratio of the number of molecules in the upper level to the number in the lower level in equilibrium to be

$$\frac{N_3}{N_2} = \frac{g_3}{g_2} e^{-\frac{\Delta E_{32}}{kT}}. \tag{1.2}$$

The population of the upper level is simply the number of molecules in it: N_3. Similarly, the population of the lower level is N_2. The numbers g_2, g_3 are called the statistical weights, or degeneracies, of the levels. These allow for the possibility that more than one quantum-mechanical state can have the same energy. The other undefined numbers that appear in Eq. (1.2) are k, Boltzmann's constant, and T, the temperature. Section 1.5 gives more detail concerning the meaning of this temperature, but for the moment we consider it to be the same temperature that would be measured by placing a thermometer inside the laser tube. We can now define the population difference, ΔN_{32}, as

$$\Delta N_{32} = \frac{N_3}{g_3} - \frac{N_2}{g_2} = \frac{N_2}{g_2}\left(e^{-\frac{\Delta E_{32}}{kT}} - 1\right). \tag{1.3}$$

A population inversion is present if ΔN_{32} is positive, that is, there is more population in the upper level than in the lower, when divided by their appropriate statistical weights. It is easy to see from Eq. (1.3) that a population inversion is impossible in the Boltzmann equilibrium we have assumed: the exponential is always smaller than 1, leaving the bracket negative. Since the population of level 2 cannot be negative, and the minimum value of g_2

[1] For the reader who does not expect this, and would like more explanation, a justification of Boltzmann's formula is given in Appendix A.

is 1 (when the level is not degenerate), the multiplier cannot reverse the sign. Even if we drop the requirement of equilibrium, allowing the population to be distributed in some other way between the levels, it can be proved (see Section 1.4) that it is not possible to obtain an inversion in a two-level system. It is, then, the presence of the third level in Fig. 1.1, and the additional transitions with energies ΔE_{31} and ΔE_{21}, that make a population inversion possible; an inversion is unusual because, to obtain one, we need to drive the system away from equilibrium. The pumping system, using energy from the electrical power supply, needs to use the upward transition from level 1 to level 3 and the downward transition from level 2 to level 1 in a manner which keeps more population in level 3 than level 2. This is not easy, because the counterparts to these two transitions tend to empty level 3 into level 1. We therefore expect the laser system, as a whole, to be rather inefficient: for each joule of energy emitted in laser light, many joules of electrical energy must be expended. It is also not possible to use just any system of atoms or molecules to create a laser. Only a relatively small number of materials have suitable arrangements of energy levels and transitions to make maintaining an inversion practical. Now that the concept of a population inversion has been explained, we need to consider some properties of electromagnetic radiation.

1.3 Electromagnetic radiation

The propagation of electromagnetic (EM) radiation is governed by Maxwell's equations, which in SI units are

$$\nabla \cdot \boldsymbol{B} = 0, \tag{1.4}$$

$$\nabla \cdot \boldsymbol{E} = \rho_q / \epsilon_0, \tag{1.5}$$

$$\nabla \wedge \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{1.6}$$

$$\nabla \wedge \boldsymbol{H} = \frac{\partial \boldsymbol{D}}{\partial t} + \boldsymbol{j}, \tag{1.7}$$

where $\boldsymbol{E}$ and $\boldsymbol{B}$ are the electric and magnetic fields, ρ_q is the charge density, ϵ_0 is the permittivity of the vacuum and $\boldsymbol{j}$ is the current per unit area. The magnetic field is related to the magnetic field intensity, $\boldsymbol{H}$, by the equation

$$\boldsymbol{B} = \mu_0(\boldsymbol{H} + \boldsymbol{M}), \tag{1.8}$$

where μ_0 is the vacuum permeability and $\boldsymbol{M}$ is the macroscopic magnetization of the material to which the magnetic field is applied. The electric displacement, $\boldsymbol{D}$, is related to the electric field via the relation

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P}, \tag{1.9}$$

where $\boldsymbol{P}$ is known as the macroscopic polarization of the medium on which $\boldsymbol{E}$ acts.

1.3.1 Response of media

We begin by making the assumption that any medium we consider has a linear response to the electric and magnetic fields. This means we can write $\boldsymbol{P} = \epsilon_0 \chi_{\mathrm{E}} \boldsymbol{E}$, where

$\chi_{\rm E}$ is a constant, known as the electric susceptibility. This allows us to write the electric displacement as

$$\boldsymbol{D} = \epsilon_0(1 + \chi_{\rm E})\boldsymbol{E} = \epsilon_0\epsilon_{\rm r}\boldsymbol{E}, \tag{1.10}$$

where $\epsilon_{\rm r} = 1 + \chi_{\rm E}$ is known as the relative permittivity of the medium.

Similarly, we define the magnetic susceptibility, $\chi_{\rm M}$, such that $\boldsymbol{M} = \chi_{\rm M}\boldsymbol{H}$, so that we can re-write Eq. (1.8) as

$$\boldsymbol{B} = \mu_0(1 + \chi_{\rm M})\boldsymbol{H} = \mu_0\mu_{\rm r}\boldsymbol{H}, \tag{1.11}$$

where $\mu_{\rm r}$ is the relative permeability of the medium. If, in addition, we assume that the medium has a constant conductivity, σ, we can represent $\boldsymbol{j}$ through a microscopic form of Ohm's law,

$$\boldsymbol{j} = \sigma\boldsymbol{E}. \tag{1.12}$$

Now, with the aid of Eq. (1.10), Eq. (1.11) and Eq. (1.12), the fourth Maxwell equation, Eq. (1.7), can be written only in terms of the electric and magnetic fields as

$$\nabla \wedge \boldsymbol{B} = \mu\epsilon\frac{\partial \boldsymbol{E}}{\partial t} + \mu\sigma\boldsymbol{E}, \tag{1.13}$$

where $\mu = \mu_0\mu_{\rm r}$ and $\epsilon = \epsilon_0\epsilon_{\rm r}$.

It is worth noting that in all known astrophysical environments where masers form, the electric field in EM waves is never strong enough to invalidate the linear approximations made in Eq. (1.10) and Eq. (1.11). It is almost always possible, also, to ignore the magnetic susceptibility. However, in the presence of a significant local magnetic field, the electric susceptibility will not be isotropic, and this needs to be considered in any theory of the transport of polarized radiation.

1.3.2 *The wave equation*

In this section, we develop a wave equation from Maxwell's equations, which describes the propagation of an EM wave through a medium. Taking the partial time derivative of Eq. (1.13), and swapping the order of the space and time derivatives on the left-hand side, we obtain

$$\frac{\partial}{\partial t}(\nabla \wedge \boldsymbol{B}) = \nabla \wedge \frac{\partial \boldsymbol{B}}{\partial t} = \mu\epsilon\frac{\partial^2 \boldsymbol{E}}{\partial t^2} + \mu\sigma\frac{\partial \boldsymbol{E}}{\partial t}. \tag{1.14}$$

The magnetic field can be eliminated with the help of Eq. (1.6), leaving $\nabla \wedge (\nabla \wedge \boldsymbol{E})$ on the left-hand side. A vector identity, Eq. (B.14) of Appendix B, transforms this expression, so that we are left with the wave equation,

$$\nabla^2\boldsymbol{E} = \frac{1}{c_m^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} + \mu\sigma\frac{\partial \boldsymbol{E}}{\partial t} + \frac{1}{\epsilon_0}\nabla\rho_q, \tag{1.15}$$

where the speed of light in the medium is

$$c_m = 1/\sqrt{\mu\epsilon}. \tag{1.16}$$

The last term on the right-hand side of Eq. (1.15) disappears if the medium has a constant charge density, which we will assume. If we can also ignore the conductivity of the medium, a conventional assumption for the regions where astrophysical masers are generated, we can reduce Eq. (1.15) to the standard EM wave equation,

$$\nabla^2 \boldsymbol{E} = \frac{1}{c_m^2} \frac{\partial^2 \boldsymbol{E}}{\partial t^2}. \tag{1.17}$$

If σ is not assumed to be zero, the second term on the right-hand side of Eq. (1.15) results in very rapid attenuation of the wave. It is also usual to take the magnetization of any astrophysical maser medium to be negligible compared with the macroscopic polarization, so that $\mu = \mu_0$. Such a medium is said to be dielectric.

Where possible in this book, a wave description of electromagnetic radiation will be used, though from time to time it will be useful to treat radiation through its particle description, as a stream of photons. In the wave description, electromagnetic (EM) radiation is formed from mutually perpendicular oscillating electric and magnetic fields. EM waves are of the transverse type, so the electric and magnetic fields are both perpendicular to the direction of propagation of the wave.

1.3.3 Travelling solutions

We now solve Eq. (1.17) for a dielectric medium. As a trial solution, insert

$$\boldsymbol{E}(\boldsymbol{r}, t) = \boldsymbol{E}_0 e^{-i(\omega t - \boldsymbol{k} \cdot \boldsymbol{r})}, \tag{1.18}$$

where the electric field varies with position, $\boldsymbol{r}$, and time, t, and has amplitude $\boldsymbol{E}_0$. The wave has frequency ν (angular frequency $\omega = 2\pi\nu$) and wavevector $\boldsymbol{k}$. A first time differential of Eq. (1.18) yields $-i\omega\boldsymbol{E}$, so differentiating again gives

$$\frac{\partial^2 \boldsymbol{E}}{\partial t^2} = -\omega^2 \boldsymbol{E}, \tag{1.19}$$

and similar operations with the space differentials yield

$$\nabla^2 \boldsymbol{E} = -(k_x^2 + k_y^2 + k_z^2)\boldsymbol{E} = -k^2 \boldsymbol{E}. \tag{1.20}$$

When Eq. (1.19) and Eq. (1.20) are substituted into Eq. (1.17), the electric field cancels, so the trial expression is a solution, provided that

$$\omega^2 = k^2 c_m^2 = k^2 c^2 / (1 + \chi_{\mathrm{E}}), \tag{1.21}$$

which is known as the dispersion relation of the medium. The second form on the right-hand side follows from the use of a dielectric medium, and the introduction of the vacuum speed of light,

$$c = 1/\sqrt{\mu_0 \epsilon_0}. \tag{1.22}$$

If we take the square root of Eq. (1.21), we obtain

$$\omega = kc/(1 + \chi_{\mathrm{E}})^{1/2}. \tag{1.23}$$

Define the complex refractive index through the relation

$$kc/\omega = n + iq, \tag{1.24}$$

where the real part, n, is the usual refractive index, and q is the attenuation coefficient. In an isotropic medium, these quantities are constants, independent of the direction of radiation propagation. The squared form of Eq. (1.24) is

$$k^2c^2/\omega^2 = n^2 - q^2 + 2inq. \tag{1.25}$$

If we assume that χ_E is also complex, so that $\chi_E = \chi_r + i\chi_i$, then we can equate the real and imaginary parts of Eq. (1.21) and Eq. (1.25), linking the refractive index to the electric susceptibility with the formulae

$$\begin{aligned} n^2 - q^2 &= 1 + \chi_r, \\ nq &= \chi_i/2. \end{aligned} \tag{1.26}$$

If we write the wavevector as $\boldsymbol{k} = k\hat{\boldsymbol{k}}$, where $\hat{\boldsymbol{k}}$ is a unit vector in the direction of the propagation of the EM wave, then Eq. (1.24) can be used to eliminate the wavevector from our wave solution, Eq. (1.18), which now has the form demanded by Eq. (1.21):

$$\boldsymbol{E}(\boldsymbol{r}, t) = \boldsymbol{E}_0 \exp\left\{i\omega\left(\frac{n}{c}\hat{\boldsymbol{k}}\cdot\boldsymbol{r} - t\right) - \frac{\omega q}{c}\hat{\boldsymbol{k}}\cdot\boldsymbol{r}\right\}. \tag{1.27}$$

The important points to notice in Eq. (1.27) are, firstly, that the real part of the refractive index appears in an oscillatory term, which cannot change the amplitude of the wave. In the diffuse gases which support astrophysical masers, it is almost always acceptable to set $n = 1$. Secondly, we see that the imaginary part appears in a real exponential. It is this part of the refractive index which can attenuate the wave away, or, with a reversed sign, lead to amplification of the wave. Therefore, attenuation, and its reverse, amplification, of radiation are linked to the properties of the medium through which the wave propagates via the imaginary part of its refractive index, or alternatively the imaginary part of its electric susceptibility.

1.3.4 Wave modes in a cavity

This digression shows the limitation of the wave theory of light when applied to the statistics of radiation inside some enclosed cavity. The cavity will be taken to be a cube, for simplicity, with conducting sides of length L. The volume of the cavity is therefore $V = L^3$. The cavity will also be assumed to be evacuated, so EM waves inside it must obey Eq. (1.17) with $c_m = c$, or

$$\nabla^2\boldsymbol{E} = \frac{1}{c^2}\frac{\partial^2\boldsymbol{E}}{\partial t^2}. \tag{1.28}$$

A Cartesian coordinate system can be constructed with its origin at one corner of the cube, and one axis directed along each of the edges which meet at this corner. The geometrical arrangement of the cavity is also set out in Fig. 1.2. Equation (1.28) has travelling solutions, like Eq. (1.27) but with the refractive index, n, equal to 1, and the extinction coefficient, q, equal to zero. However, inside a cavity, we can form standing-wave solutions, with the additional constraints of Eq. (1.5), in a form with $\rho_q = 0$, and the requirement that the tangential component of the electric field must vanish at any boundary, which follows from making the walls conducting. An acceptable solution of Eq. (1.28) with these constraints and boundary conditions is $\boldsymbol{E}(\boldsymbol{r}, t) = \sum_{i=1}^{3} \hat{\boldsymbol{e}}_i E_i(\boldsymbol{r}, t)$, where $i = 1, 2, 3$ can correspond to the

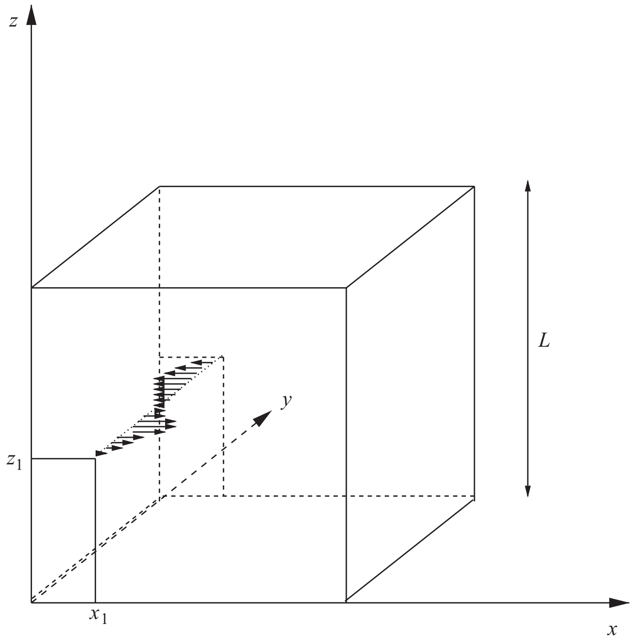

Fig. 1.2. An evacuated cavity containing electromagnetic radiation. The cavity is a cube of side L, with a Cartesian axis system defined as shown. The behaviour of the x-component of the electric field is shown as a function of y along the line $(x, y) = (x_1, z_1)$, where $\cos(k_x x_1) = 1$ and $\sin(k_z z_1) = 1$. The field oscillation shown corresponds to a wavevector with the y-component, $k_y = 2\pi/L$.

ordered sequence of axes, x, y, z, and $\hat{\boldsymbol{e}}_i$ is a unit vector along the appropriate axis. The ith Cartesian electric field component then takes the form

$$E_i(\boldsymbol{r}, t) = E_i(t) \prod_{j=1}^{3} \mathrm{sc}(k_j r_j), \tag{1.29}$$

where the function sc denotes a sine when the axis labels i and j are different, and a cosine when $i = j$. If this solution, with components given by Eq. (1.29), is substituted into Eq. (1.28), the wave equation breaks down into scalar wave equations, for each component, for example,

$$-k^2 E_i = \frac{1}{c^2} \frac{d^2 E_i(t)}{dt^2}, \tag{1.30}$$

where $k^2 = \sum_i k_i^2$. Equation (1.30) is solved by a time-varying part of the electric field which oscillates as

$$E_i(t) = E_{0i} e^{-i\omega t}, \tag{1.31}$$

where $\omega^2 = k^2 c^2$.

If components of the form given by Eq. (1.29) are substituted into $\nabla \cdot \boldsymbol{E} = 0$, the result is that we require $\boldsymbol{k} \cdot \boldsymbol{E} = 0$: in the cavity, the wavevector is perpendicular to the electric field, and lies along the direction of propagation of the wave. We can also see that components satisfy the conducting boundary constraint: taking component $i = 1$, corresponding to the coordinate x, we have

$$E_x(\boldsymbol{r}, t) = E_x(t)\cos(k_x x)\sin(k_y y)\sin(k_z z), \tag{1.32}$$

and this component is transverse for waves travelling in the y- or z-directions. Because of the sine functions in Eq. (1.32), we can see that E_x will indeed be zero when y or z are zero. However, at the other extreme, where y or $z = L$, the sine will only satisfy the boundary condition if an integer number of half-wavelengths fit across the cavity. In symbols, we have, $L = m_x\lambda/2 = \pi m_x/k_x$, where the wavelength, $\lambda = 2\pi/k_x$, so we require

$$k_x = \pi m_x/L, \tag{1.33}$$

and similarly for the other dimensions. Equation (1.33) also dictates that the lattice spacing between adjacent modes along any of the Cartesian axes is π/L. We therefore have three positive, or zero, integers, (m_x, m_y, m_z), which control the set of standing waves that satisfy Eq. (1.29), with appropriate boundary conditions. These numbers are not quite independent, in the sense that only one of them can be zero if we want to have any EM waves at all inside the cavity. We define an EM wave mode of the cavity as a set of standing waves with a valid group of integers, (m_x, m_y, m_z).

There are several important points to note about the cavity modes introduced above:

- Each mode is individually a solution of the wave equation, Eq. (1.28), which is a linear differential equation. Therefore, any sum of modes is also a solution, and any field of EM waves that can be excited in the cavity can be represented as a sum of modes.
- As the cavity becomes larger, relative to the wavelength, the influence of the boundary conditions becomes progressively less important, and the distribution of modes approaches a continuum.
- Although unproven here, the results above still hold for a cavity of any size and shape, in the sense that EM waves will still be confined to a set of allowed modes. The formula for the density of modes, derived below, is truly independent of the cavity shape.
- A mode solution has a degeneracy of order 2. This follows from the requirement that $\boldsymbol{k}\cdot\boldsymbol{E} = 0$, and the definition of the dot product, Eq. (B.1): the cosine of the angle between these two vectors has two zeros in a 2π rotation. Electric fields at these angles, with respect to $\boldsymbol{k}$, represent two independent polarizations of the electric field.
- The number of available modes rises with the upper limiting value for the mode integers, $m_{\max}$. For example, only four modes (with two polarizations) are available for $m_{\max} = 1$, but there are 20 with $m_{\max} = 2$. In general, the number of modes rises as $(m_{\max} + 1)^3 - 3m_{\max} - 1$.

The structure of modes can be represented as a series of points in mode space, as shown in Fig. 1.3. Note the absence of allowed modes where more than one of m_x, m_y and m_z are zero. For a large enough cavity, or small enough wavelength, this distribution of points tends to a continuous density of modes, in which the number of modes rises as $m_{\max}^3$. How many modes are there altogether? As the minimum value of any mode integer is zero, the available space is a cube of volume equal to 1/8 of the volume of the larger cube we would have if we