Optics Lectures

1 Lecture 1: The wave equation and diffraction

1.1 The wave equation

We know that optics is about waves (or at least the bit that isn't about particles!). Wave optics was invented by Huygens, but formalized by Maxwell. These are Maxwell's equations:

$$\boldsymbol{\nabla}.\boldsymbol{E} = \rho, \qquad (1) \qquad \boldsymbol{\nabla}.\boldsymbol{B} = 0, \qquad (3)$$

$$\boldsymbol{\nabla}\times\boldsymbol{E} = -\partial_t\boldsymbol{B} \qquad (2) \qquad \boldsymbol{\nabla}\times\boldsymbol{B} = \mu_0\boldsymbol{J} + \epsilon_0\mu_0\partial_t\boldsymbol{E}. \qquad (4)$$

Here ρ is the charge density and $\boldsymbol{J}$ is the current density. $\boldsymbol{E}$ and $\boldsymbol{B}$ are the electric and magnetic fields. Optical waves are hidden somewhere here. We consider fields in empty space, so we set $\rho = \boldsymbol{J} = 0$. Differentiate (4) with respect to time, and then substitute (2) into it:

$$\boldsymbol{\nabla}\times\partial_t\boldsymbol{B} = \epsilon_0\mu_0\partial_t^2\boldsymbol{E},$$

$$\Rightarrow -\boldsymbol{\nabla}\times\boldsymbol{\nabla}\times\boldsymbol{E} = \epsilon_0\mu_0\partial_t^2\boldsymbol{E}.$$

This is some kind of differential equation for the electric field. We can decode it using the vector calculus identity

$$\boldsymbol{\nabla}\times\boldsymbol{\nabla}\times\boldsymbol{F} = \boldsymbol{\nabla}\left(\boldsymbol{\nabla}.\boldsymbol{F}\right) - \nabla^2\boldsymbol{F},$$

for any vector field $\boldsymbol{F}$. Let us define the constant c via the relation $c^2 = 1/(\epsilon_0\mu_0)$. Finally, (1) tells us that the electric field *in vacuo* is divergence free. Putting all this together yields the wave equation

$$\left[\nabla^2 - \frac{1}{c^2}\partial_t^2\right]\boldsymbol{E} = 0, \qquad (5)$$

where c turns out to be the speed of light. The above is actually a pretty weird result. The speed of light is just a combination of constants quantifying the strength of the electromagnetic interaction. The full appreciation of the weirdness of Maxwell's equations ultimately led Einstein to the theory of relativity, which takes the constancy of c, independent of sources or their motion, as a fundamental postulate. This is a wave equation, which admits sinusoidally varying solutions, like

$$\boldsymbol{E}(\boldsymbol{r}, t) = \boldsymbol{E}_0 e^{i(\boldsymbol{k}.\boldsymbol{r} - \omega t)}, \qquad (6)$$

where $\boldsymbol{E}_0$ is some constant vector, $\boldsymbol{k}$ is called the *wavevector*, and ω is the *angular frequency* (or just the frequency).

1.2 Fourier Transform

The linearity of the wave equation means that a general solution to the wave equation can be constructed from sums of exponentials as in (6). For this reason the Fourier transform is critical to the analysis of optical signals. The Fourier transform pair is defined as follows,

$$S(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \widetilde{S}(\omega) e^{i\omega t} \, d\omega, \tag{7}$$

and

$$\widetilde{S}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} S(t) e^{-i\omega t} \, dt. \tag{8}$$

If the coordinate t is a temporal coordinate, then The Fourier transform $\widetilde{S}(\omega)$ of the signal $S(t)$ is called the *spectrum* of the signal, and the coordinate ω is a frequency (an angular frequency). The symmetry of the above relations means that the spectrum is 'just as good' as the temporal signal as a representation of the signal. It's good to become comfortable with switching between these two complementary viewpoints of the same signal.

The properties of Fourier transforms are well-documented and familiarity with manipulating transforms is extremely useful in nearly all optical problems.

1.3 Green's function

We derived the wave equation for the electric field *in vacuo*,

$$\left[\nabla^2 - \frac{1}{c^2} \partial_t^2 \right] \boldsymbol{E} = 0. \tag{9}$$

This equation can be used to compute the way light diffracts from an aperture, which is of fundamental importance in optics and imaging science, with obvious applications in acoustics, medical diagnostics, astronomy, radar sensing *etc...* To proceed, we first simplify the calculation by neglecting the vector character of the electric field: we consider the evolution of just a single component, whose magnitude we denote by the scalar function E. We also remove the time dependence by substituting the following ansatz, corresponding to a monochromatic field with frequency ω, into (9):

$$E(\boldsymbol{r}, t) = E(\boldsymbol{r}) e^{i\omega t}.$$

This yields the *Helmholtz equation*

$$\left[\nabla^2 + k^2 \right] E = 0, \tag{10}$$

where $k = \omega/c$ is the (magnitude of the) wavevector of the field. We know plane waves can be used to build up a solution, but we are interested in the *Green's function* for this equation. The Green's function $G(\boldsymbol{r})$ associated with a differential equation is defined as follows,

$$\mathcal{L}_{\boldsymbol{r}} G(\boldsymbol{r}) = \delta(\boldsymbol{r}),$$

where $\delta(\boldsymbol{r})$ is the three dimensional Dirac delta function, and where $\mathcal{L}_{\boldsymbol{r}}$ represents the differential operator describing the particular differential equation we are concerned with. Green's functions are very useful for constructing the solutions to boundary value problems of the form $\mathcal{L}_{\boldsymbol{r}}\psi(\boldsymbol{r}) = U(\boldsymbol{r})$, since we can write $\psi(\boldsymbol{r}) = \int G(\boldsymbol{r} - \boldsymbol{r}')U(\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}'$. For the Helmholtz equation, we have

$$\left[\nabla^2 + k^2\right]G(\boldsymbol{r}) = \delta(\boldsymbol{r}). \tag{11}$$

To find G, we Fourier transform both sides from $\boldsymbol{r} \longrightarrow \boldsymbol{q}$, which gives

$$\left(k^2 - q^2\right)\widetilde{G}(\boldsymbol{q}) = \frac{1}{(2\pi)^3}.$$

Solving for $\widetilde{G}$ and Fourier transforming back, we get

$$
\begin{aligned}
G(\boldsymbol{r}) &= \frac{1}{(2\pi)^3}\int \frac{e^{\mathrm{i}\boldsymbol{r}\cdot\boldsymbol{q}}}{k^2 - q^2}\,\mathrm{d}\boldsymbol{q} \\
&= \frac{1}{(2\pi)^3}\int_0^\infty \int_{-1}^1 \int_0^{2\pi} \frac{e^{\mathrm{i}rq\cos\theta}q^2}{k^2 - q^2}\,\mathrm{d}q\mathrm{d}(\cos\theta)\mathrm{d}\phi \\
&= \frac{2\pi}{(2\pi)^3}\int_0^\infty \left(\frac{e^{\mathrm{i}rq}}{\mathrm{i}rq} - \frac{e^{-\mathrm{i}rq}}{\mathrm{i}rq}\right)\frac{q^2}{k^2 - q^2}\,\mathrm{d}q \\
&= \frac{1}{2\mathrm{i}r(2\pi)^2}\left\{\int_{-\infty}^\infty \frac{qe^{\mathrm{i}rq}}{(k-q)(k+q)}\,\mathrm{d}q - \int_{-\infty}^\infty \frac{qe^{-\mathrm{i}rq}}{(k-q)(k+q)}\,\mathrm{d}q\right\} \tag{12}
\end{aligned}
$$

where in the last line we extended the integral over q to the range $-\infty \longrightarrow \infty$, and divided the result by 2. These integrals can be done via contour integration. First, we apply a small 'regularization', meaning that we shift the singularities in the integrands off the real line by making the transformation $k \longrightarrow k + \mathrm{i}\epsilon$, where ϵ is a small real number. For example, the integral

$$I = \int_{-\infty}^\infty \frac{qe^{\mathrm{i}rq}}{(k + \mathrm{i}\epsilon - q)(k + \mathrm{i}\epsilon + q)}\,\mathrm{d}q$$

can be closed in the upper half of the complex plane. Using the residue theorem, we then get

$$I = 2\pi\mathrm{i}\frac{(k + \mathrm{i}\epsilon)e^{\mathrm{i}r(k+\mathrm{i}\epsilon)}}{k + \mathrm{i}\epsilon + k + \mathrm{i}\epsilon},$$

which becomes

$$I = \mathrm{i}\pi e^{\mathrm{i}rk}, \tag{13}$$

when we take the limit $\epsilon \longrightarrow 0$. For a given choice of regularization (choosing $\epsilon > 0$), each integral in (12) above contains one pole, and so each integral contributes a result like (13). The Green's function then works out to be

$$G(\boldsymbol{r}) = \frac{1}{4\pi}\frac{e^{\mathrm{i}k|\boldsymbol{r}|}}{|\boldsymbol{r}|}.$$

This describes spherical waves expanding out from the origin. This is consistent with the solution we expect for a wave equation with an impulse (*i.e.* a delta function) as a source term: G describes the ripples on a pond after a pebble is thrown in. In the next section, we will use its properties to study diffraction from an aperture.

1.4 Kirchhoff diffraction

Consider the quantity
$$\boldsymbol{\chi} = \phi\boldsymbol{\nabla}\psi - \psi\boldsymbol{\nabla}\phi,$$
defined for arbitrary functions ϕ and ψ. The divergence of this quantity is given by
$$\begin{aligned}\boldsymbol{\nabla}.\boldsymbol{\chi} &= \boldsymbol{\nabla}\phi.\boldsymbol{\nabla}\psi + \phi\nabla^2\psi - \boldsymbol{\nabla}\psi.\boldsymbol{\nabla}\phi - \psi\nabla^2\phi \\ &= \phi\nabla^2\psi - \psi\nabla^2\phi.\end{aligned}$$

Integrating over some volume, and applying the divergence theorem, we get
$$\int \boldsymbol{\nabla}.\boldsymbol{\chi}\, \mathrm{d}\boldsymbol{r} = \int \boldsymbol{\chi}.\mathrm{d}\boldsymbol{S},$$
where the latter integral runs over the surface bounding the volume. Substituting $\psi \longrightarrow E(\boldsymbol{r})$ and $\phi \longrightarrow G(\boldsymbol{r} - \boldsymbol{R})$, where $\boldsymbol{R}$ is some point of interest, we get
$$\begin{aligned}\int \left[E(\boldsymbol{r})\boldsymbol{\nabla}G(\boldsymbol{r} - \boldsymbol{R}) - G(\boldsymbol{r} - \boldsymbol{R})\boldsymbol{\nabla}E(\boldsymbol{r})\right].\mathrm{d}\boldsymbol{S} &= \int \left[E(\boldsymbol{r})\nabla^2 G(\boldsymbol{r} - \boldsymbol{R}) - G(\boldsymbol{r} - \boldsymbol{R})\nabla^2 E(\boldsymbol{r})\right]\,\mathrm{d}\boldsymbol{r} \\ &= \int \left\{E(\boldsymbol{r})\left[\delta(\boldsymbol{r} - \boldsymbol{R}) - k^2 G(\boldsymbol{r} - \boldsymbol{R})\right]\right. \\ &\quad \left. + k^2 G(\boldsymbol{r} - \boldsymbol{R})E(\boldsymbol{r})\right\}\,\mathrm{d}\boldsymbol{r} \\ &= E(\boldsymbol{R}),\end{aligned}$$
where in the second step we used the fact that E and G satisfy the wave equations (10) and (11). This relationship gives us a prescription for finding the field E at some point $\boldsymbol{R}$ from the field and its derivatives on some surface enclosing $\boldsymbol{R}$. We now consider the case that this surface is a large hemisphere, with infinite radius, centred on a diffracting aperture surrounded by an opaque screen. We neglect contributions to the surface integral from the curved portion, since this is infinitely far away, and from the opaque screen. The only contribution to the surface integral is then from the transparent portion of the aperture itself. We take the normal to this plane aperture to point along the z axis, and so we obtain
$$E(\boldsymbol{R}) = \int_{\text{aperture}} \left[E(\boldsymbol{r})\partial_z G(\boldsymbol{r} - \boldsymbol{R}) - G(\boldsymbol{r} - \boldsymbol{R})\partial_z E(\boldsymbol{r})\right]\,\mathrm{d}S.$$
Differentiating the Green's function, we get
$$\begin{aligned}\partial_z G(\boldsymbol{r} - \boldsymbol{R}) &= -\left(\mathrm{i}k - \frac{1}{|\boldsymbol{r} - \boldsymbol{R}|}\right)G(\boldsymbol{r} - \boldsymbol{R})\cos\theta \\ &\approx -\mathrm{i}k G(\boldsymbol{r} - \boldsymbol{R})\cos\theta,\end{aligned}$$

where θ is the angle between the vector $\boldsymbol{R} - \boldsymbol{r}$ and the positive z axis, and where in the last step we assumed that $k \gg 1/|\boldsymbol{r} - \boldsymbol{R}|$ (*i.e.* that the distance from the aperture to $\boldsymbol{R}$ is much larger than an optical wavelength). Finally, we assume that the field illuminating the aperture is composed of plane waves travelling along the positive z axis,

$$E(\boldsymbol{r}) = E_{\mathrm{in}}(x, y)e^{\mathrm{i}kz}.$$

Differentiating this, we end up with the *Kirchhoff diffraction integral*

$$E(\boldsymbol{R}) = -\frac{\mathrm{i}k}{4\pi} \int_{\mathrm{aperture}} \frac{e^{\mathrm{i}k|\boldsymbol{r}-\boldsymbol{R}|}}{|\boldsymbol{r} - \boldsymbol{R}|} E_{\mathrm{in}}(x, y) \left(\cos\theta + 1\right) \, \mathrm{d}S.$$

The factor $(\cos\theta + 1)$ is known as the *obliquity factor*. Let us write $\boldsymbol{R} = (X, Y, Z)$, and let us denote the diffracted field in the plane at $z = Z$ by $E(\boldsymbol{R}) = E(X, Y, Z) = E_{\mathrm{diff}}(X, Y)$. Generally we will be concerned with small-angle diffraction for which $Z \gg (X, Y)$, and $\cos\theta = Z/\sqrt{Z^2 + X^2 + Y^2} \approx 1$. Then we can write the diffraction integral in the form

$$E_{\mathrm{diff}}(X, Y) = -\frac{\mathrm{i}k}{2\pi Z} \int_{\mathrm{aperture}} e^{\mathrm{i}k\sqrt{(x-X)^2 + (y-Y)^2 + Z^2}} E_{\mathrm{in}}(x, y) \, \mathrm{d}S.$$

1.5 Fresnel and Fraunhofer diffraction

The square root in the exponent of the integrand in Kirchhoff's integral — thanks to Pythagoras — is inconvenient, and rather obfuscates the features of diffraction. Since Z is generally large compared to the aperture size, and to the transverse size of the region over which we would like to compute the diffracted field, we can expand the square root using a Taylor series,

$$\begin{aligned}
\sqrt{(x-X)^2 + (y-Y)^2 + Z^2} &\approx Z + \frac{(x-X)^2 + (y-Y)^2}{2Z} \\
&= Z + \frac{X^2 + Y^2}{2Z} + \frac{x^2 + y^2}{2Z} - \frac{X}{Z}x - \frac{Y}{Z}y.
\end{aligned}$$

For simplicity, lets proceed with a one dimensional treatment, and drop the y coordinates (its easy to put them back in at any time). Then we can write

$$E_{\mathrm{diff}}(X) = -\frac{\mathrm{i}ke^{\mathrm{i}kZ}e^{\mathrm{i}k\varphi(X)}}{2\pi Z} \int_{\mathrm{aperture}} e^{\mathrm{i}k\varphi(x)}e^{-\mathrm{i}(kX/Z)x} E_{\mathrm{in}}(x) \, \mathrm{d}x,$$

where $\varphi(x) = x^2/2Z$ is a transverse phase factor that is quadratic in x. This is known as the Fresnel diffraction kernel. Clearly it is nicer than the square-root business above, but it is not easy to work with. Fortunately, there are many situations in which the kernel can be simplified even further. The most obvious is just the case where the quadratic phase $\varphi(x)$ in the integrand is so small that it can be neglected. If φ is small, we can expand the exponential in the integrand above as

$$e^{\mathrm{i}k\varphi(x)} \approx 1 + \mathrm{i}k\varphi(x) + \dots.$$

To ignore the contribution of φ to the diffraction kernel, we simply require that $k\varphi(x) \ll 1$ for all values of x appearing in the integral. If the aperture has a size a, this gives the condition $k\varphi(a) \ll 1$. Sometimes the quantity $\mathcal{F} = k\varphi(a) = \pi a^2/\lambda Z$ is known as the *Fresnel number*. Setting $\mathcal{F} \ll 1$ gives

$$\Rightarrow Z \gg \frac{\pi a^2}{\lambda}. \tag{14}$$

So if we look far enough away — in what we term the *far field* — we can neglect the quadratic contributions to the diffraction kernel. With this simplification, the diffracted field is closely related to the Fourier transform of the field in the plane of the aperture. To see this, note that we can convert the integral over the aperture to an integral over the entire plane containing the aperture, as long as we require that the field $E_{\text{in}}(x, y)$ is zero outside the aperture. With this in mind, we can write the diffraction kernel as

$$
\begin{aligned}
E_{\text{diff}}(X) &= -\frac{ik e^{ikZ} e^{ik\varphi(X)}}{2\pi Z} \int e^{-i(kX/Z)x} E_{\text{in}}(x)\, dx \\
&= -\frac{i e^{ikZ} e^{ik\varphi(X)}}{\sqrt{2\pi}} \times \frac{k}{Z} \widetilde{E}_{\text{in}}\left(\frac{k}{Z}X\right).
\end{aligned}
$$

This is known as the *Fraunhofer* diffraction kernel. It is approximately correct in the limit that the distance between the diffracting aperture and plane of observation is large. Perhaps you recall that Fraunhofer diffraction from a rectangular slit produces a sinc^2 intensity distribution (found by Fourier transforming the 'rect' function). The first zero of this distribution appears at an angle of $\theta = \lambda/a$, and this angle serves as a good characterization for the angle subtended by light diffracting from a general aperture of size a in the Fraunhofer limit. Note that the Fresnel number can be expressed (neglecting a factor of π) as

$$\mathcal{F} \sim \frac{a}{\theta Z}.$$

That is, the Fresnel number is roughly the width of the aperture over the size of the diffraction pattern. Fraunhofer diffraction is therefore valid in the limit $\mathcal{F} \ll 1$ that the aperture is much smaller than the size of its diffraction pattern.

2 Lecture 2: Lenses and interference

2.1 Lenses

We make a lens using a lump of transparent material with a refractive index different from that of air. One way to understand the action of a lens is to trace the paths of rays refracted at its two surfaces and show how diverging rays striking one surface of the lens are brought together to form an image downstream from the lens. A more powerful way to think about lenses is as *phase shifters*. For a sufficiently thin lens, the transverse phase shift added by the presence of a lens is

$$\phi(x) = k(n - 1)t(x),$$

where $t(x)$ is the thickness of the lens at transverse coordinate x, and where n is the refractive index of the lens material. For a spherical plano-convex lens with radius of curvature R and maximum thickness T, the thickness a distance x away from centre is given by

$$\begin{aligned} t(x) &= T - (R - R\cos\theta)\,, \qquad \text{where} \qquad \sin\theta = \frac{x}{R} \\ &\approx T - \frac{x^2}{2R}. \end{aligned}$$

The constant T has no important effects, so we neglect it in what follows. Suppose we have a point source emitting spherical waves a distance u from the face of the lens. The spherical wavefronts striking the lens have the transverse phase

$$\begin{aligned} \phi_u(x) &= k\sqrt{u^2 + x^2} \\ &\approx ku + \frac{kx^2}{2u}. \end{aligned}$$

After the lens, then, and neglecting the constant phases, we have the phase

$$\phi_u(x) + \phi(x) = \frac{k}{2}\left(\frac{1}{u} - \frac{n-1}{2R}\right)x^2.$$

But this phase is just equal to the phase fronts associated with a spherical wave converging on a point a distance v downstream from the lens, where

$$\frac{1}{u} + \frac{1}{v} = \frac{n-1}{R}.$$

This is readily identified with the lens-maker's formula by defining the focal length $f = R/(n-1)$.

Question: Can you think of a way to make a lens, composed of two glass plates, such that the focal length can be adjusted by sliding the plates over along the x-axis? Suppose one plate has thickness $t(x,y)$, and other has thickness $T - t(x,y)$. Now one is moved $x \longrightarrow x - \delta$ and other other is moved $x \longrightarrow x + \delta$. What form should the function $t(x,y)$ have such that the phase imparted by propagation through the pair is quadratic?

Now, since a lens imparts a negative quadratic transverse phase, we can construct another situation in which Fraunhofer diffraction applies. When a lens cancels the Fresnel phase $\varphi(x)$ in the diffraction kernel. Placing a lens immediately behind the diffracting aperture (or indeed, one can consider the lens itself — with its finite size — as an aperture), we see that $\varphi(x)$ is cancelled exactly if

$$\begin{aligned} \phi(x) + \varphi(x) &= 0, \\ \Rightarrow f &= Z. \end{aligned}$$

That is, Fraunhofer diffraction describes the diffraction pertaining to the image plane of any optical instrument. It is this property that makes Fraunhofer so important.

Now, we showed the diffracted field is related to the Fourier transform of the incident field in the Fraunhofer limit, but it is not quite *equal* to the Fourier transform, because of the transverse quadratic phase $\varphi(X)$. To remove even this phase factor, the lens can be moved from immediately after the aperture, to halfway between the aperture and the observation plane. If the focal length of the lens is now equal to $Z/2$, then the diffraction pattern is exactly (at least up to quadratic terms in the phase) given by a scaled Fourier transform of the field transmitted through the aperture. This is not that difficult to show, but *two* diffraction integrals are required:

$$E_{\text{diff}} \propto \int \int e^{\mathrm{i}k\psi(X,\xi,x)} \, \mathrm{d}\xi \, E_{\text{in}}(x) \, \mathrm{d}x,$$

where

$$\psi(X,\xi,x) = \frac{X^2 + \xi^2 + x^2}{Z} - 2\frac{X+x}{Z}\xi.$$

where ξ is the transverse coordinate in the plane of the lens, and where we have neglected multiplicative constants and phases that are independent of any transverse coordinates. Performing the integral over ξ (by 'completing the square' on ψ and applying the formula for a Gaussian integral), yields,

$$E_{\text{diff}}(X) \propto \widetilde{E}_{\text{in}}\left(\frac{2kX}{Z}\right).$$

That is, the diffracted field is truly a Fourier transform of the incident field.

2.2 Interference

Generally, in optics, we deal with electric fields that are varying quickly — typically at THz frequencies — so that it is not possible to directly measure the electric field amplitude $\boldsymbol{E}$: no electronic detector is fast enough. Instead, we use detectors that measure the energy in the field, like power meters or photodiodes, CCDs, or indeed the human eye. The quantity we have access to is then the intensity $I = |\boldsymbol{E}|^2$. Given such detectors, how do we surmise that light propagates as waves? The key feature is *interference*, which can be broadly characterized as the modulation of an intensity distribution arising when two light sources are combined. But interference is not always apparent. We do not see interference fringes in daylight or lamp light, but we certainly do if we use lasers. Clearly there are differences between these light sources, but without delving into the physics of them, it is useful to define a property associated with each source — *coherence* — that describes the quality of the interference fringes one can produce using it. This is a very operational definition, but we will see that it is straightforward to turn this into a mathematical quantity with simple properties. We will see that coherence is degraded by non-uniformity of the interfering fields, and also by statistical fluctuations.

In any interference, at least two fields are combined, with some relative phase whose variation produces intensity modulations. For simplicity, let us consider just a single polarization component of the electric field (so we can treat all fields as scalar quantities). Then we are concerned with a field of the form

$$E = E_1 + E_2,$$

where E_1 and E_2 are complex numbers representing the two interfering fields. The intensity is found by squaring E,

$$\begin{aligned} I &= |E|^2 \\ &= |E_1|^2 + |E_2|^2 + E_1 E_2^* + E_2 E_1^* \\ &= I_1 + I_2 + 2\Re\{E_1 E_2^*\}. \end{aligned}$$

The first two terms are what one would expect if light were described by a 'corpuscular' theory: we just add intensities to get the total. But the last term does not have to be positive, depending on the relative phase between E_1 and E_2 it can be either positive or negative. Therefore as the relative phase of E_1 and E_2 changes, the intensity I will be modulated. This is interference. It is *the* characteristic feature of waves.

Now, since our detectors are slow, our measured signal will take the form of a time average of the intensity,

$$\langle I \rangle = \frac{1}{T} \int_{-T/2}^{T/2} I(t)\,\mathrm{d}t,$$

where T is the response time of the detector. Applying this average to our interference pattern gives

$$\langle I \rangle = \langle I_1 \rangle + \langle I_2 \rangle + 2\Re\{\langle E_1 E_2^* \rangle\}.$$

Now, the average of the interference term is always smaller in magnitude than the largest magnitude appearing in its un-averaged form. Therefore the interference we see on our detector may be of reduced quality, depending on the magnitude of $\langle E_1 E_2^* \rangle$. Let us define the *coherence* as the normalized interference term

$$\gamma_{12} = \frac{\langle E_1 E_2^* \rangle}{\sqrt{\langle I_1 \rangle \langle I_2 \rangle}},$$

so that our interference signal can be written as

$$\langle I \rangle = \langle I_1 \rangle + \langle I_2 \rangle + 2\sqrt{\langle I_1 \rangle \langle I_2 \rangle}\,\Re\{\gamma_{12}\}.$$

Now γ_{12} cannot have a magnitude larger than 1 (this follows from the Cauchy-Schwarz inequality: think of the average as the scalar product of the two 'wave-functions' E_1 and E_2; the result can be seen by analogy with the vector scalar product, where $|\boldsymbol{a}.\boldsymbol{b}| = |ab\cos(\theta)| \leq ab$). When $|\gamma_{12}| = 1$, we say that the fields E_1 and E_2 are coherent or perfectly coherent. If $|\gamma_{12}| = 0$, we say that the fields are *incoherent*. Clearly in the first case, interference is maximal, while in the second case it is absent. If $|\gamma_{12}|$ takes an intermediate value, than we say the fields are *partially coherent* with one another.

2.3 Interferometers

In an interferometer, we generally combine two fields with a controllable relative phase that we can vary to observe fringes. The maxima and minima of the interference pattern will have intensities

$$\langle I \rangle_{\max} = \langle I_1 \rangle + \langle I_2 \rangle + 2\sqrt{\langle I_1 \rangle \langle I_2 \rangle} |\gamma_{12}|,$$
$$\text{and} \quad \langle I \rangle_{\min} = \langle I_1 \rangle + \langle I_2 \rangle - 2\sqrt{\langle I_1 \rangle \langle I_2 \rangle} |\gamma_{12}|.$$

The *visibility* of the interference is defined as

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}$$
$$= \frac{2\sqrt{\langle I_1 \rangle \langle I_2 \rangle}}{\langle I_1 \rangle + \langle I_2 \rangle} |\gamma_{12}|.$$

And if we interfere two fields with equal average intensities $\langle I_1 \rangle = \langle I_2 \rangle$, we have

$$V = |\gamma_{12}|.$$

Therefore, the visibility of interference fringes observed on a slow detector gives direct information about the coherence of the interfering fields. Perfect visibility, with $V = 1$, occurs only when $I_{\min} = 0$, so the presence of completely dark fringes is a signature of perfect coherence.

2.4 Temporal coherence

Suppose that we build a Michelson interferometer. This splits a single incident field into two parts; they travel different path lengths and are then recombined. Therefore, in the language of the above discussion, we have $E_2 = E_2(t) = E_1(t - \tau)$. That is, the two interfering fields are just time-shifted copies of one another. Now the coherence γ_{12} can be written simply as $\gamma(\tau)$,

$$\gamma(\tau) = \frac{\langle E(t) E^*(t - \tau) \rangle}{\langle |E|^2 \rangle} = \frac{1}{\langle I \rangle T} \int_{-T/2}^{T/2} E(t) E^*(t - \tau) \, dt.$$

This coherence function is called the *temporal* coherence, and it describes the coherence of a light beam with itself, at a later time. If we set the time shift τ equal to zero (by balancing the arms of the interferometer), then clearly we have $\gamma(0) = 1$. In general, there will be some finite timescale τ_c over which the coherence remains roughly constant, so that $\gamma(\pm\tau_c) \approx 1$. This time is called the *coherence time* of the field.

The temporal coherence is related to the *spectrum* of the light. To see how, note that the temporal coherence has the form of a *convolution* of the field E with the field E^*, as long as T is much larger than the timescales over which E varies. Recall that the Fourier transform of a convolution is simply the product of the Fourier transform of the individual fields. Therefore we have

$$\mathcal{F}_\tau \left\{ \gamma(\tau) \right\}(\omega) \propto \widetilde{E}(\omega) \times \widetilde{E}^*(\omega) = I(\omega).$$

This is known as the *Wiener-Kinchine* theorem. It tells you that the visibility of interference fringes produced by a Michelson interferometer is directly related to the *power spectrum* of the light entering the interferometer, by a Fourier transform. From our knowledge of Fourier transforms, it's clear that the broader the spectrum, the narrower the coherence function, so the coherence time τ_c characterizing the temporal width of the coherence function is inversely related to the spectral bandwidth.

2.5 Transverse coherence

Now suppose we build a Young's slits apparatus. Here, a field illuminates two spatially separated slits, and then the light emerging from each slit is recombined on a screen. The angular position on the screen is related to the difference in optical path lengths connecting points on the screen to the two slits, and interference fringes are therefore observed across the screen. Now the two fields are observed at the same time (or at least, at times separated by much less than the coherence time of the field), but they are 'picked' from two different points in space, in the plane transverse to the propagation direction. Now the coherence function takes the form

$$\gamma(\xi) = \frac{\langle E(x)E^*(x - \xi)\rangle}{\langle |E|^2 \rangle}.$$

Now, suppose that both of these fields have propagated from a source far away. Fraunhofer diffraction tells us that they can be related to the field E_s in the plane of the source by a Fourier transform,

$$E(x) \propto \int E_s(y) e^{ikxy/D}\,\mathrm{d}y.$$

The coherence function is then given by

$$\gamma(\xi) \propto \int\int \langle E_s(y)E_s^*(y')\rangle e^{ikxy/D} e^{-ik(x-\xi)y'/D}\,\mathrm{d}y\mathrm{d}y'.$$

Now, suppose that the light fields coming from the source are completely *incoherent.* Then we have

$$\langle E_s(y)E_s^*(y')\rangle = E_s(y)E_s^*(y)\delta(y - y') = I_s(y)\delta(y - y').$$

That is, they show no correlation unless you're looking at exactly the same point on the source. Inserting the delta function, we get

$$\gamma(\xi) \propto \int I(y) e^{ik\xi y/D}\,\mathrm{d}y.$$

That is, the visibility of interference fringes in a Young's slit interferometer is related to the transverse intensity distribution of an incoherent source by a Fourier transform. This is known as the *van Cittert-Zernike* theorem. It underlies the operation of stellar interferometry, where the transverse profiles of stars can be found by observing the interference from two telescopes.

3 Lecture 3: Ray Transfer matrices

The invention of the laser introduced an entirely new kind of optics, based on the manipulation of coherent, collimated beams of light. Such beams behave very much as thin pencils of light, and often a description in terms of rays is appropriate. On the other hand, their coherence makes interference effects arising from the wave nature of light particularly easy to observe. Fortunately the results obtained for rays remain useful even when a wave description is used.

We will be concerned with *paraxial* optics, in which any ray (or wavefront) propagates at a small angle $\theta \ll 1$ to the optical axis (or z axis), such that $\sin\theta \approx \tan\theta \approx \theta$. Assuming cylindrical symmetry about the z axis, we can specify a ray at any point z uniquely by its displacement x away from the z axis, and its direction θ. Now consider the specification of the same ray at a later position $z + L$. The change in its displacement over the distance L is given by the relation

$$\tan(\theta) = \frac{x' - x}{L},$$

where x' is the new displacement. Therefore, in the paraxial approximation, we have

$$x' = x + L\theta.$$

Since the ray's direction is unchanged, we have

$$\theta' = \theta,$$

where θ' is the new direction. Putting these results together gives the relation

$$\begin{pmatrix} x' \\ \theta' \end{pmatrix} = \begin{pmatrix} 1 & L \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ \theta \end{pmatrix}.$$

The ray at both positions is described by a vector quantity, whose components specify its displacement and its direction, and the connection between the two — which is a linear transformation in the paraxial approximation — is represented by a matrix known as a *ray transfer matrix*. In general, any optical component that maintains paraxial propagation can be represented by some transfer matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where A, B, C and D are real constants. These matrices are sometimes known as "$ABCD$ matrices". What about a lens? If the lens is thin, the displacement is not changed by propagation through the material of the lens, so $x' = x$, but the angles of rays either side of the lens are altered. To derive the relationship between these angles, recall the lens maker's formula

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f},$$

where f is the focal length of the lens. We can then write

$$\tan\theta = \frac{x}{u}\,; \qquad \tan(-\theta') = \frac{x}{v},$$

where the minus sign in the second relation accounts for the fact that the ray emerging from the lens is propagating towards the z axis. Invoking the paraxial approximation gives

$$\theta' = \theta - \frac{x}{f},$$

so that the ray transfer matrix for a lens is

$$\begin{pmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{pmatrix}. \tag{15}$$

It is also possible to derive transfer matrices for reflection and refraction at curved and plane surfaces. Very complicated optical systems can be analyzed easily in this formalism, simply by multiplying the transfer matrices for each component (and for the free propagation between them) together to produced a single, final matrix for the whole system. This is therefore a very powerful technique!

In general, if the refractive indices of the initial and final medium are given by n_i and n_f, the determinant $AD - BC$ of the transfer matrix is equal to n_i/n_f, so that transfer matrices describe unitary transformations if $n_i = n_f$. It is easy to check for the cases shown above that $AD - BC = 1$. In some formulations, the refractive index of the medium is included with the direction θ in the description of the ray, and then the transfer matrices are *always* unitary. For simplicity, we'll stick with the current formulation.

3.1   Spherical waves

Consider a point source, located on the z-axis. The wavefronts produced by such a point source are spherical, and can be characterized by their radius of curvature R. How are they transformed by a general optical system? Consider the ray defined by the normal to the wavefronts. We have that $R\sin\theta = x$, and therefore, in the paraxial approximation, $R = x/\theta$. After the optical system, we can calculate x' and θ' using the ray transfer matrix, and then the new radius of curvature for the wavefronts is given by $R' = x'/\theta'$. Putting all this together we get

$$R' = \frac{AR + B}{CR + D}. \tag{16}$$

This is known as a Möbius transformation. For the case of free space propagation, we have $A = D = 1$, $B = L$ and $C = 0$, so that

$$R' = R + L. \tag{17}$$

This result will be useful in analyzing the propagation of laser beams.

13

3.2 Generalized diffraction integral

The $ABCD$ matrices can be applied to the study of wave propagation in the paraxial limit, as well as to the study of rays. This is why they are worth using! To see how this works, recall the Fresnel-Kirchoff integral,

$$E(x') = \frac{\mathrm{i}k}{2\pi} \int K(x', x) E(x) \, \mathrm{d}x, \tag{18}$$

where the integral kernel K propagates the incident field to the outgoing field, and takes the form

$$K(x', x) = \frac{e^{\mathrm{i}kl(x', x)}}{l(x', x)}, \tag{19}$$

where $l(x', x)$ is the optical path length from x at the input to x' at the output. Is there a way to use the $ABCD$ matrix to evaluate the path length l through a general optical system? If there is, we can use these matrices to implement diffractive propagation through very complicated optical systems in 'one step'. To see how to do this, note that the initial direction θ of a ray can be expressed in terms of its initial and final displacements x and x' using the relation $x' = Ax + B\theta$, which gives

$$\theta = \frac{x' - Ax}{B}. \tag{20}$$

Similarly, for the final direction we have $\theta' = Cx + D\theta = [(BC - AD)x + Dx']/B$. Using the fact that the determinant is $AD - BC = 1$, we have

$$\theta' = \frac{Dx' - x}{B}. \tag{21}$$

Now, the centres of the spherical waves on either side of an optical system are *conjugate points*: waves emerging from one centre are brought together at the other, so they are images of each other. The optical path length of any ray connecting two such points is the same (as required by Fermat's principle). Therefore the optical path along the optical axis is equal to the optical path along an oblique incident ray, through the optical system, and then along the emerging ray. Let the first distance be equal to $D_{\mathrm{direct}} = R + L_0 - R'$, where L_0 is the optical path through the optical system along the optical axis (the minus sign accounts for the fact that R' is a negative number for a converging spherical wave). The second distance $D_{\mathrm{ray}} = \sqrt{R^2 + x^2} + l(x', x) + \sqrt{R'^2 + x'^2}$. Setting $D_{\mathrm{direct}} = D_{\mathrm{ray}}$ and applying the paraxial approximation ($x \ll R$ etc...), we get

$$l = R - R' + L_0 - R - \frac{x^2}{2R} + R' + \frac{x'^2}{2R'}.$$

Note again that we have adjusted the signs to account for the fact that R' is a negative number. Using the $ABCD$ matrix for the system gives

$$R = \frac{x}{\theta} = \frac{Bx}{x' - Ax},$$

$$\text{and} \quad R' = \frac{x'}{\theta'} = \frac{Bx'}{Dx' - x}.$$

Putting this all together yields the expression

$$l(x', x) = L_0 - \frac{1}{2B} \left[Dx'^2 + Ax^2 - 2xx' \right].$$ (22)

Inserting this expression into (19) and (18) results in a simple, universal expression, that enables one to propagate a paraxial field through an arbitrarily complicated optical system in a single calculation, simply using the $ABCD$ matrix for the system. Nice.

3.3   Gaussian Beams

Now, a Gaussian beam is about the simplest kind of beam one could think of. Fortunately, it is also precisely the type of beam that is actually produced by lasers in the laboratory. Therefore understanding the Gaussian beam will get you 90% of the way to being a laser physicist. So, where does the idea come from? We start by returning to the time-independent scalar wave equation — the Helmholtz equation — describing one component of the electric field in free space.

$$\left[\nabla^2 + k^2 \right] E = 0.$$ (23)

Now, in the paraxial approximation, we have a beam that essentially looks like a plane wave, propagating along the optical axis, except that it has some finite extent in the transverse direction:

$$E(x, y, z) = U(x, y, z) e^{-\mathrm{i}kz}.$$ (24)

The function U is *slowly varying*, meaning that it can be considered constant on the scale of an optical wavelength. Therefore we have $|\boldsymbol{\nabla} U| \ll kU$. On substituting (24) into (23), we obtain the *paraxial wave equation*

$$
\begin{aligned}
e^{\mathrm{i}kz} \left[\partial_x^2 + \partial_y^2 + \partial_z^2 - 2\mathrm{i}k\partial_z - k^2 + k^2 \right] U &= 0, \\
\Rightarrow \nabla_\perp^2 U &\approx 2\mathrm{i}k\partial_z U.
\end{aligned}
$$ (25)

Note the similarity of this to the time-dependent Schrödinger equation — the existence of a complete set of eigenfunctions satisfying this equation should not be a surprise. Anyhow, for the moment, consider the trial solution

$$U = e^{-\mathrm{i}\left[P + \frac{k}{2q} r^2 \right]},$$ (26)

where $P(z)$ and $q(z)$ are two — in general *complex* — functions of z, and where $r = \sqrt{x^2 + y^2}$ is the perpendicular distance from the optical axis. Substituting this into our paraxial wave equation (25), we find

$$
\begin{aligned}
\nabla_\perp^2 U &= -\frac{\mathrm{i}k}{q} \left[2 - \frac{\mathrm{i}kr^2}{q} \right] U, \\
\text{and} \quad 2\mathrm{i}k\partial_z U &= \left[2k\partial_z P - \left(\frac{kr}{q} \right)^2 \partial_z q \right] U.
\end{aligned}
$$

Comparing coefficients of r, we get the conditions

$$\partial_z P = \frac{1}{\mathrm{i}q}, \tag{27}$$

$$\text{and} \quad \partial_z q = 1. \tag{28}$$

This second equation is easily solved to yield

$$q(z) = q(0) + z.$$

That is to say, the function q simply grows linearly with distance. Now, since distances are always real, the imaginary part of q does not change. For some position, however, the real part of q can be cancelled by z, and at this position q is then purely imaginary. Let us choose our coordinate system such that the position $z = 0$ coincides with this point, so that we have

$$q(0) = \mathrm{i}\Im\{q\} = \mathrm{i}z_{\mathrm{R}}. \tag{29}$$

We will see that the quantitiy z_{R}, which has the dimensions of length, has a clear interpretation. To bring this out, we make a cosmetic change to the solution (26) by defining a pair of real functions $R(z)$ and $w(z)$ according to the relation

$$\frac{1}{q} = \frac{1}{R} - 2\mathrm{i}\frac{1}{kw^2}. \tag{30}$$

We then have

$$U = e^{-\mathrm{i}P} \times e^{-\mathrm{i}kr^2/2R} \times e^{-(r/w)^2}.$$

We'll address the first term, involving P, shortly. The second term is a quadratic phase factor with precisely the form expected for spherical waves in the paraxial approximation, with radius of curvature R. The last term is not a phase factor, but a real transverse Gaussian, describing damping of the field amplitude with increasing distance from the optical axis. This term defines the solution as a 'beam', with a width characterized by w. Comparing the relations (29) and (30), we find

$$R(z) = z\left[1 + \left(\frac{z_{\mathrm{R}}}{z}\right)^2\right], \tag{31}$$

$$\text{and} \quad w(z) = w_0\sqrt{1 + \left(\frac{z}{z_{\mathrm{R}}}\right)^2}, \tag{32}$$

where $w_0 = w(0) = \sqrt{2z_{\mathrm{R}}/k}$ is the beam size at $z = 0$. Note that this beam size is the smallest the beam ever gets: either side of $z = 0$ the beam grows wider. Note also that at the beam's narrowest point at $z = 0$, the radius of curvature becomes infinite, $R(0) \longrightarrow \infty$, so that the phase fronts are flat at this point. The distance z_{R} is known as the *Rayleigh range*, because it quantifies the distance either side of $z = 0$ over which the Gaussian beam is approximately collimated (*i.e.* the distance over which $w(z)$ is approximately constant). In

fact, setting $z = z_R$ gives $w(z_R) = \sqrt{2}w_0$, so the Rayleigh range is the distance either side of $z = 0$ over which the beam size lies between and 1 and $\sqrt{2}$ times its minimum size. At distances much greater than the Rayleigh range, the beam width rapidly diverges. The radius of curvature R grows linearly with z for $z \gg z_R$, so that from afar, a Gaussian beam looks like a spherical wave centred on $z = 0$.

The reason for the divergence of the beam is diffraction: a bundle of waves cannot remain localized indefinitely, and their spreading is responsible for the 'de-focussing' of the beam. The trade off between spatial localization and collimation is reflected by the behaviour of the Rayleigh range $z_R = kw_0^2/2$, which can be expressed as

$$z_R = \pi \times \frac{w_0^2}{\lambda}.$$

If we consider Fraunhofer diffraction from a slit of width w_0, the diffraction angle would be, roughly $\theta_{\text{diff}} \sim \lambda/w_0$, and the size of the diffraction pattern after a distance z would be, roughly, $z\theta_{\text{diff}} \sim z\lambda/w_0$. If we require that this size is approximately the same as slit itself, we get,

$$w_0 \quad \sim \quad \frac{z\lambda}{w_0},$$

$$\Rightarrow z \approx \frac{w_0^2}{\lambda}.$$

From this perspective, one can interpret the spreading of a Gaussian beam in terms of Fraunhofer diffraction from its own focus.

What about the phase P? Integrating (27) gives

$$
\begin{aligned}
P(z) &= -i \int_0^z \frac{1}{z' + iz_R} \, dz' \\
&= -i \ln \left(\frac{z}{z_R} + i \right) \\
&= -i \ln \left\{ \left[1 + \left(\frac{z}{z_R} \right)^2 \right]^{1/2} e^{i \tan^{-1}(z_R/z)} \right\}.
\end{aligned}
$$

Substituting this result into (26), we get

$$U(r, z) = \frac{w_0}{w(z)} e^{-i\zeta(z)} \times e^{-ikr^2/2R(z)} \times e^{-[r/w(z)]^2}. \tag{33}$$

The pre-factor $w_0/w(z)$ accounts for the reduction in amplitude of the field as it spreads out. The longitudinal phase $\zeta(z) = \tan^{-1}(z_R/z)$ is known as the *Guoy phase*.

3.4 Gaussian beam propagation

What happens to a Gaussian beam when it encounters a lens? The beam size is not changed, since the lens is assumed to be thin. Only the transverse phase is

affected. We have already identified the transverse phase of a Gaussian beam as being that of spherical waves with radius of curvature R. Using the "$ABCD$" law (16) and the lens transfer matrix (15), along with the relation (29), we derive the transformation law for the q parameter

$$\frac{1}{q'} = \frac{1}{q} - \frac{1}{f}.$$

We also have derived $q' = q+L$ for free space propagation, which compares with the transformation $R' = R + L$ derived in (17) above. Therefore the complex radius of curvature q transforms in precisely the same way as the real radius of curvature R does, upon propagation through lenses and free space. This property is summarized by the $ABCD$ law

$$q' = \frac{Aq + B}{Cq + D}.$$

That is, the propagation of a Gaussian beam through an arbitrary optical system can be described using the ray transfer matrix for the system, and applying the above law to find the new q parameter. *Note*: this law can be explicitly verified by inserting (33) into the propagation integral (18), and using the kernel (19) with the path length given in the $ABCD$ formalism by (22).

3.5 Focussing a Gaussian beam

How tightly can a Gaussian beam be focussed by a lens? This can be answered easily by considering the problem in reverse: to what beam size does a Gaussian beam spread over the focal length f of a lens, if its focus has a beam waist w_0. A simple geometric argument yields

$$\begin{aligned}
w(f) &\approx f\theta_{\text{diff}} \\
&= \frac{f\lambda}{w_0}, \\
\text{or} \quad w_0 &\approx \frac{f\lambda}{w(f)}.
\end{aligned}$$

That is, the size of the focal spot w_0 is limited by the size of the beam $w(f)$ on the lens: to get a tighter focus, you need to start with a wider beam, and thus a wider lens. Also, shorter focal lengths and shorter wavelengths allow for tighter focussing.

4 Lecture 4: Optical cavities and resonators

4.1 Cavity stability

A cavity is essentially a trap for electromagnetic fields, formed by enclosing a volume with reflective surfaces. Trapping fields in this way stops them from

spreading and dissipating, so that a high intensity, and a large electromagnetic energy, can be concentrated in a prescribed region. A microwave oven is just a cavity for microwaves: when it is switched on, the field strength grows so large that you can zap your lasagne. In a laser, an optical cavity is employed to raise the intensity of the field within the laser gain medium: this stimulates the extraction of more energy, resulting in a bright, coherent field, which leaks out of the cavity as a beam. Optical cavities are generally not completely enclosed on all sides: instead, it's possible to confine a beam using just a pair of mirrors to fold a beam back on itself. If this beam is well collimated, losses due to diffraction of the beam out of the sides of the cavity can be made very small. It is very instructive to analyze this kind of very simple cavity, which is comprised of just two curved mirrors, aligned to face each other. We can get a long way by considering the ray transfer matrices associated with propagation in such a system. As an initial simplification, we start by 'unfolding' the cavity: instead of two curved mirrors, M_1 and M_2, think instead of an infinite series of equally-spaced lenses L_1 and L_2. Start by launching a ray $\boldsymbol{r} = (x, \theta)^{\mathrm{T}}$ just after the first lens L_1, and then consider: propagation, L_2, propagation, L_1, propagation, L_2, etc.... After n lenses (i.e. after n round-trips of the real cavity), the ray is described by the vector

$$\boldsymbol{r}_n = M^n \boldsymbol{r},$$

where $M = L_1 P L_2 P$ is the matrix formed from the product of the ray transfer matrix P for propagation over the cavity length with the matrices L_1, L_2 for propagation through the lenses. To analyze the behaviour of the propagator M^n, we can diagonalize M, which yields the characteristic equation

$$\lambda^2 - \mathrm{tr}\,\{M\} + |M| = 0,$$

$$\Rightarrow \lambda_\pm = \frac{\mathrm{tr}\,\{M\}}{2} \pm \sqrt{\left(\frac{\mathrm{tr}\,\{M\}}{2}\right)^2 - |M|},$$

where $\mathrm{tr}\,\{M\} = A + D$ and $|M| = AD - BC = 1$ are the trace and the determinant of M, respectively (recall that our ray transfer matrices are unitary). Now, we can distinguish two cases. First, suppose that $-2 \leq \mathrm{tr}\,\{M\} \leq 2$. In this case, we can define an angle θ such that $\mathrm{tr}\,\{M\}/2 = \cos(\theta)$, and then we can write

$$\lambda_\pm = \cos(\theta) \pm \mathrm{i}\sin(\theta) = e^{\pm \mathrm{i}\theta}.$$

On the other hand, if $|\mathrm{tr}\,\{M\}| > 2$, then the cosine function is not an appropriate parameterization. Instead write $\mathrm{tr}\,\{M\}/2 = \cosh(\psi)$. This yields

$$\lambda_\pm = \cosh(\psi) \pm \sinh(\psi) = e^{\pm \psi}.$$

Now, if we express the initial ray $\boldsymbol{r}$ in terms of the eigenvectors $\boldsymbol{r}_\pm$ associated with the two eigenvalues $\lambda_\pm$ as $\boldsymbol{r} = c_- \boldsymbol{r}_- + c_+ \boldsymbol{r}_+$, then the final ray, after n cavity roundtrips, can be simply written as

$$\boldsymbol{r}_n = c_- \lambda_-^n \boldsymbol{r}_- + c_+ \lambda_+^n \boldsymbol{r}_+.$$

19

It is now clear that the two situations we distinguished above predict drastically different behaviour as n grows large. In the first case, we have

$$\boldsymbol{r}_n = c_- e^{-\mathrm{i}\theta n} \boldsymbol{r}_- + c_+ e^{\mathrm{i}\theta n} \boldsymbol{r}_+, \tag{34}$$

whereas in the second case, we have

$$\boldsymbol{r}_n = c_- e^{-\psi n} \boldsymbol{r}_- + c_+ e^{\psi n} \boldsymbol{r}_+. \tag{35}$$

The first relation (34) describes an oscillatory behaviour, in which each component of the ray experiences a phase shift upon completion of a roundtrip. This is *stable* behaviour, because the ray continues indefinitely, evolving in a periodic fashion. On the other hand, the second relation (35) describes exponential damping of one component, and exponential growth of the other. This is clearly *unstable* behaviour. The displacement and direction of the eigenray $\boldsymbol{r}_+$ grows larger and more oblique with each roundtrip; the other eigenray $\boldsymbol{r}_-$ becomes increasingly confined to the optical axis. Such cavities do not form an efficient trap for radiation, because rays are quickly ejected from the sides of the cavity or confined to its axis.

Let's analyze this stability condition for the sake of our two-mirror cavity. Having unfolded it, we can easily express the $ABCD$ matrices for the components as

$$P = \begin{pmatrix} 1 & L \\ 0 & 1 \end{pmatrix}, \quad L_1 = \begin{pmatrix} 1 & 0 \\ -\frac{1}{f_1} & 1 \end{pmatrix}, \quad L_2 = \begin{pmatrix} 1 & 0 \\ -\frac{1}{f_2} & 1 \end{pmatrix}.$$

Therefore M is given by

$$M = \begin{pmatrix} \left[1 - \frac{L}{f_1}\right]\left[1 - \frac{L}{f_2}\right] - \frac{L}{f_2} & L\left[2 - \frac{L}{f_1}\right] \\ -\frac{1}{f_1}\left[1 - \frac{L}{f_2}\right] - \frac{1}{f_2} & 1 - \frac{L}{f_1} \end{pmatrix}.$$

Taking the trace of M, the stability condition is

$$-2 \leq 1 - \frac{L}{f_1} + \left[1 - \frac{L}{f_1}\right]\left[1 - \frac{L}{f_2}\right] - \frac{L}{f_2} \leq 2,$$

which can be re-written as

$$0 \leq \left[2 - \frac{L}{f_1}\right]\left[2 - \frac{L}{f_1}\right] \leq 4.$$

One can then classify the stability of two-mirror resonators of this type with reference to a fairly simple diagram.

4.2   Gaussian and Hermite-Gaussian modes

We have already seen that the Gaussian beam is a solution of the paraxial wave equation, given by

$$U(r,z) = \frac{w_0}{w(z)} e^{-\mathrm{i}\zeta(z)} \times e^{-\mathrm{i}kr^2/2R(z)} \times e^{-[r/w(z)]^2},$$

where
$$\zeta(z) = \tan^{-1}(z/z_R)$$
is the *Guoy phase*, and where $w(z)$ and $R(z)$ are the beam waist and radius of curvature, respectively.

In fact, the Gaussian beam is the simplest of a *family* of solutions. The most commonly encountered family is the set of *Hermite-Gaussian* modes. These can be found by separating the solution $U(x, y, z)$ into two parts,

$$U(x, y, z) = u_n(x, z) \times u_m(y, z).$$

The integers n and m will play the role of mode indices. Substituting this trial solution into the paraxial wave equation

$$\nabla_\perp^2 U = 2ik\partial_z U,$$

one can separate variables and obtain equations for u_n and u_m which are both of precisely the same form, namely

$$\partial_x^2 u_n = 2ik\partial_z u_n. \tag{36}$$

If we substitute the trial solution

$$u_n(x, z) = h_n\left(\frac{x}{p}\right) e^{-i\left[P + \frac{k}{2q}x^2\right]}, \tag{37}$$

into (36), where p, q and P are all functions of z, and as before $\partial_z q = 1$, we obtain the differential equation

$$h_n'' - 2ik\left[\frac{p}{q} - p'\right]xh_n' - \frac{ikp^2}{q}\left[1 - 2iqP'\right]h_n = 0, \tag{38}$$

where primes indicate differentiation with respect to the argument of each function, and where we have used the fact derived previously for the case of the simple Gaussian mode that $q' = \partial_z q = 1$. The key observation is that this rather messy looking equation is similar in structure to the following differential equation,

$$H_n'' - 2\frac{x}{p}H_n' + 2nH_n = 0, \tag{39}$$

whose solutions $H_n(x/p)$ are known as the *Hermite polynomials*. The first few polynomials are given by

$$
\begin{aligned}
H_0(x) &= 1, \\
H_1(x) &= x, \\
H_2(x) &= 4x^2 - 1, \\
H_3(x) &= 8x^3 - 12x.
\end{aligned}
$$

These polynomials appear in the description of the eigenfunctions of the quantum harmonic oscillator, and they form a complete orthonormal set in which

any function can be expanded. Therefore they form a basis of modes into which any paraxial optical field can be decomposed. There are a number of ways to bring the equation (38) into correspondence with (39). The most commonly used definitions involve setting $p(z) = w(z)/\sqrt{2}$, which satisfies the equation

$$\mathrm{i}kp\left(\frac{p}{q} - p'\right) = 1,$$

associated with matching the coefficients of h'_n and H'_n in (38) and (39). Matching the coefficients of h_n and H_n, we also require that

$$\mathrm{i}kp^2\left(2\mathrm{i}P' - \frac{1}{q}\right) = 2n,$$

so that P must depend on n. Rearranging gives

$$P'_n = -\frac{1}{z^2 + z_\mathrm{R}^2}\left[\left(n + \frac{1}{2}\right)z_\mathrm{R} + \frac{\mathrm{i}}{2}z\right],$$

which can be solved to yield

$$
\begin{aligned}
P_n(z) &= -\left(n + \frac{1}{2}\right)\tan^{-1}\left(\frac{z}{z_\mathrm{R}}\right) - \frac{\mathrm{i}}{4}\ln\left(\frac{z^2 + z_\mathrm{R}^2}{z_\mathrm{R}^2}\right) \\
&= -\left(n + \frac{1}{2}\right)\tan^{-1}\left(\frac{z}{z_\mathrm{R}}\right) + \frac{\mathrm{i}}{2}\ln\left[\frac{w_0}{w(z)}\right].
\end{aligned}
$$

Substituting this back into (37), and including the corresponding result for the y coordinate, one can write a Gaussian beam in the form

$$U_{nm}(x, y, z) = \frac{w_0}{w(z)}e^{-\mathrm{i}(n+m+1)\tan^{-1}(z/z_\mathrm{R})}H_n\left(\sqrt{2}\frac{x}{w(z)}\right)H_m\left(\sqrt{2}\frac{y}{w(z)}\right)e^{-\mathrm{i}k(x^2+y^2)/2q(z)}.$$

The mode indices n and m count the number of nodal lines in the x and y directions, respectively. Note that for these higher order modes, the Guoy phase is modified and becomes $\zeta_{nm}(z) = (n + m + 1)\tan^{-1}(z/z_R)$.

4.3   Longitudinal modes

In a real optical resonator, the fields are reflected back on top of each other, and therefore they must add in phase if a significant field is to build up. Having derived the conditions for stable operation using the ray transfer matrices for the 'unfolded' cavity, we now need to impose this additional condition, which manifests itself as a restriction on the optical frequencies supported by the cavity. To see this, we impose the periodic boundary condition that after one roundtrip, the longitudinal phase of the optical field is unchanged. The length of a roundtrip is $2L$, where L is the cavity length, so this condition can be written as

$$\phi(2L) = 2\pi p,$$

22

where $p = 0, 1, 2, ...$ is some integer, and where

$$\phi(z) = kz + \zeta_{nm}(z)$$

is the optical phase, found by adding the phase associated with the plane wave carrier and the Guoy phase for the transverse mode we are considering. Using the free space dispersion relation $\omega = ck$, we find the allowed frequencies to be

$$
\begin{aligned}
\omega_p &= \frac{c}{2L}\left[2\pi p + \zeta_{nm}\left(\frac{2L}{2}\right) - \zeta_{nm}\left(-\frac{2L}{2}\right)\right] \\
&= \frac{c}{L}\left[\pi p + \zeta_{nm}(L)\right].
\end{aligned}
$$

These frequencies define the *longitudinal modes* of the laser oscillator. Note that the frequency spacings between the modes is just set by the cavity length, but the Guoy phase associated with the transverse mode contributes a small offset. For $L \gg z_R$, we have $\zeta(L) \approx (n + m + 1)\pi/2$, which is usually very small, since $m \sim n \sim 1$, while $p \sim L/\lambda \sim 10^6$.

4.4 Solving for the q parameter

The real advantage of using the Hermite-Gaussian modes as a means of describing laser oscillation, is that the $ABCD$ law for the transformation of the q parameter,

$$q' = \frac{Aq + B}{Cq + D}, \tag{40}$$

can be used for any of the Hermite-Gaussian modes, in order to analyze their propagation through an arbitrary optical system. To show that the condition for a stable resonator applies for Gaussian and Hermite-Gaussian modes, this law can be expressed as

$$\begin{pmatrix} q' \\ 1 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} q \\ 1 \end{pmatrix}.$$

Which shows that the condition on the eigenvalues of the $ABCD$ matrix applied in the case of rays, is also required here to prevent divergence of the q parameter over multiple roundtrips.

The condition (40) can also be used to find the shape of a given optical mode, once its indices n, m and p are specified. One simply imposes the requirement that $q' = q$, and solves the resulting quadratic equation, which is of the form

$$
\begin{aligned}
Cq^2 + (D - A)q - B &= 0, \\
\Rightarrow q &= \frac{(A - D)}{2C} - i\frac{\sqrt{4 - (A + D)^2}}{2|C|}.
\end{aligned}
$$

This is, explicitly, a solution for $q(z_{\text{ref}})$ at the reference plane we used to define the $ABCD$ matrix for the cavity (recall that we 'started' the propagation just

after the first mirror). Use of the propagation formula $q(z_{\text{ref}} + z) = q(z_{\text{ref}}) + z$ allows us to find the q parameter at any other position in the cavity. And from q we are able to find the beam waist $w(z)$ and the radius of curvature $R(z)$. Therefore the beam is completely specified.

5   Lecture 5: Waveguides

5.1   Optical Fibres

Light travels quickly, does not interact with itself, and is easily generated and detected. Therefore light is extremely appealing as an information carrier, and data transmission using optical signals now forms the bedrock of modern telecommunications systems. But these innovations rely on the ability to send light down 'wires'. How is this done?

In an optical fibre, light is guided down a thin glass filament whose refractive index is higher than the surrounding material. The simplest picture for how this guiding is accomplished is to consider rays propagating at small angles to the symmetry axis of the fibre: when they encounter the boundary of the central core, they experience total internal reflection, provided that they subtend an angle larger than the critical angle with the normal to the boundary. These rays make their way down the fibre by repeated reflections at the boundary. If the fibre is bent or coiled, the light can be guided around corners, provided that the curvature of the fibre is not so large as to allow rays to escape the core.

Based on this picture, one can calculate the acceptance angle of such a 'step-index' fibre — the largest angle that an incident ray may make with the fibre axis such that it will be guided down the fibre. Suppose the ray is launched into the fibre from the surrounding air at an angle α. Snell's law tells us that, when the ray inside the fibre encounters the core boundary at the critical angle θ_{c}, we have

$$
\begin{aligned}
\sin \alpha &= n_1 \sin \left(\pi/2 - \theta_{\text{c}} \right) \\
&= n_1 \sqrt{1 - \sin^2 \theta_{\text{c}}} \\
&= \sqrt{n_1^2 - n_2^2},
\end{aligned}
$$

where in the second line we used the relation $\sin \theta_{\text{c}} = n_2/n_1$ for the critical angle, and where n_1 and n_2 are the refractive indices of the core and the surrounding cladding respectively. The quantity $\sqrt{n_1^2 - n_2^2}$ is known as the *numerical aperture* (NA) of the fibre. Note how larger refractive index differences between the core and the cladding increase the NA and allow larger acceptance angles: they guide light more strongly. A convenient expression for the NA can be found when the difference between the refractive indices is small, so that

$$
\Delta = \frac{n_1 - n_2}{n_2} \ll 1.
$$

In that case, we can write

$$
\begin{aligned}
\mathrm{NA} &= \sqrt{(n_1 - n_2) \times (n_1 + n_2)} \\
&\approx \sqrt{\Delta n_2 \times 2 n_2} \\
&= n_2 \sqrt{2\Delta}.
\end{aligned}
$$

Typical values for Δ and n_2 are 0.01 and 1.4, which yields a numerical aperture of ~ 0.2, and therefore a maximum acceptance angle of $\sim 11°$.

5.2 Fibre modes

The ray-optics treatment of fibres is clearly instructive, but — as always — is rather incomplete. Of course, the light in an optical fibre is wavelike, and is described by Maxwell's equations. The confinement of the light to the fibre core can be thought of rather like the confinement of an electron wave in a Coulomb potential as described by Schrödinger's equation. The refractive index step makes it energetically favourable for the optical field to propagate within the core, and diffraction out of the sides of the fibre is suppressed, as if the cladding were repulsive. To see how this works, consider the wave equation for the electric field,

$$
\left[\nabla^2 - \frac{n_1^2}{c^2} \partial_t^2 \right] \boldsymbol{E} = 0. \tag{41}
$$

The symmetry of the fibre naturally suggests that we employ cylindrical coordinates, so that the equation becomes

$$
\frac{1}{r} \partial_r \left(r \partial_r \boldsymbol{E} \right) + \frac{1}{r^2} \partial_\phi^2 \boldsymbol{E} + \partial_z^2 \boldsymbol{E} - \frac{n_1^2}{c^2} \partial_t^2 \boldsymbol{E} = 0. \tag{42}
$$

Assuming a monochromatic field with frequency ω, and a linear polarization, which we'll define to be along the x-axis, we proceed by substituting in the following ansatz,

$$
\boldsymbol{E} = \hat{\boldsymbol{x}} R(r) \Phi(\phi) e^{\mathrm{i}(\beta z - \omega t)}. \tag{43}
$$

Here β is the *propagation constant* — essentially the wavevector of the propagating mode — and R and Φ represent the radial and azimuthal dependence of the field, respectively. Using (43) in (42), we find

$$
\Phi \frac{1}{r} \partial_r \left(r \partial_r R \right) + \frac{R}{r^2} \partial_\phi^2 \Phi - R\Phi \left[\beta^2 - \left(\frac{n_1 \omega}{c} \right)^2 \right] = 0. \tag{44}
$$

This simplifies further if we assume a very simple azimuthal dependence,

$$
\Phi(\phi) = e^{\pm i l \phi},
$$

where l is an integer. This yields,

$$
\frac{1}{r} \partial_r \left(r \partial_r R \right) + \left[\left(\frac{n_1 \omega}{c} \right)^2 - \beta^2 - \frac{l^2}{r^2} \right] = 0,
$$

25

or

$$\left[r^2\partial_r^2 + r\partial_r + r^2k_{\mathrm{T}}^2 - l^2\right] R = 0, \tag{45}$$

where we have defined the *transverse wavevector*

$$k_{\mathrm{T}}^2 = \frac{n_1^2\omega^2}{c^2} - \beta^2.$$

This radial equation is recognizable as Bessel's equation, and it is solved by Bessel functions. In the fibre core we expect solutions that are well-behaved, oscillatory functions, and so we use the *ordinary* Bessel functions of the first kind, $R_{\mathrm{core}} \propto J_l(k_{\mathrm{T}}r)$.

Now, in the cladding, the wave satisfies essentially the same equation (45), except that the refractive index becomes n_2 instead of n_1. The character of the solution should change too, since the mode is damped rather than propagating. To see how this arises, we write the radial equation in the cladding in the form

$$\left[r^2\partial_r^2 + r\partial_r - \left(r^2\gamma^2 + l^2\right)\right] R = 0, \tag{46}$$

where $\gamma^2 = \beta^2 - \frac{n_2^2\omega^2}{c^2}$ is a positive quantity, because the refractive index n_2 of the cladding is small. This equation is solved by the *modified* Bessel functions, and of those, the functions with the correct damped behaviour are the modified Bessel functions of the second kind, denoted by $R_{\mathrm{cladding}} \propto K_l(\gamma r)$.

At the boundary between the core and the cladding, $r = a$, these two solutions should 'match', meaning that the field E should be continuous across the boundary. The full solution is therefore found by stitching the two solutions together,

$$\begin{aligned}
\boldsymbol{E}_{\mathrm{core}}(r < a, \phi) &= \hat{\boldsymbol{x}}E_0 \frac{J_l(k_{\mathrm{T}}r)}{J_l(k_{\mathrm{T}}a)} \cos(l\phi)e^{\mathrm{i}(\beta z - \omega t)}, \\
\boldsymbol{E}_{\mathrm{cladding}}(r > a, \phi) &= \hat{\boldsymbol{x}}E_0 \frac{K_l(\gamma r)}{K_l(\gamma a)} \cos(l\phi)e^{\mathrm{i}(\beta z - \omega t)},
\end{aligned} \tag{47}$$

where both the 'plus' and 'minus' solutions for $\Phi(\phi)$ were combined to guarantee that the solution for the field is real. In addition to matching the magnitudes of the fields at the core/cladding interface, the derivatives of the fields should also be matched. Differentiating both the fields with respect to r and setting the derivatives equal to one another at $r = a$ for all ϕ yields the condition

$$k_{\mathrm{T}} \frac{J_{l-1}(k_{\mathrm{T}}a)}{J_l(k_{\mathrm{T}}a)} = -\gamma \frac{K_{l-1}(\gamma a)}{K_l(\gamma a)}, \tag{48}$$

where we have used the recursion relations that hold for Bessel functions to re-express their derivatives in terms of functions with different orders. (48) is known as the *characteristic function* of the fibre, which connects k_{T} and γ, given the fibre radius a. Since k_{T} and γ both contain the propagation constant β and the 'vacuum wavevector' $k = \omega/c = 2\pi/\lambda$, (48) expresses the *dispersion relation* of the fibre that determines the dependence of β on k. Since β parameterizes

26

the velocity of optical signals propagating along the fibre, this relation is critical in determining what signals the fibre carries, and how much these signals are distorted by dispersive spreading.

Here we briefly note that the solution (47) does not strictly satisfy Maxwell's equations — in order that the field is divergence free, the y and z components of $\boldsymbol{E}$ cannot be exactly zero. But they are small, and are generally neglected. The same is true of the paraxial description of Gaussian beams in free space.

It is also worth commenting that we are not exactly right when we enforce continuity of the electric field $\boldsymbol{E}$ across the core/cladding boundary. Recall that in a dielectric medium free of free charges, it is the divergence of the *displacement field* $\boldsymbol{D}$ that vanishes, and therefore it is the radial component of D that is continuous across the boundary. However, there is not much difference between enforcing the continuity of E and that of D. Inserting $D = \epsilon_0 \epsilon_{\mathrm{r}} E = \epsilon_0 n^2 E$, we find that continuity of D causes a relative difference between the electric fields at the boundary of

$$
\begin{aligned}
\frac{E_{\mathrm{core}}(r=a) - E_{\mathrm{cladding}}(r=a)}{E_{\mathrm{cladding}}(r=a)} \;&=\; \frac{n_1^2 - n_2^2}{n_2^2} \\
&=\; \frac{(n_1 - n_2)(n_1 + n_2)}{n_2^2} \\
&=\; 2\Delta \\
&\ll\; 1.
\end{aligned}
$$

5.3   Fibre modes

The solution (47) describes an optical mode that propagates without significant loss, provided that the field does not penetrate far into the cladding. That is to say, the field inside the cladding should be damped, otherwise optical power can propagate freely out of the core, and the light is eventually absorbed at or lost from the outer edge of the fibre. The condition that the cladding field is damped is simply that γ is positive, and so the *cutoff condition* $\gamma = 0$ expresses when a mode becomes lossy. Recall that the transverse wavevector is given by $k_{\mathrm{T}}^2 = n_1^2 \omega^2/c^2 - \beta^2$. In order that a mode is freely propagating, and not lossy, we have $\gamma = \beta^2 - n_2^2 \omega^2/c^2 > 0$, so that $\beta^2 > n_2^2 \omega^2/c^2$, and therefore

$$
\begin{aligned}
k_{\mathrm{T}}^2 \;&<\; \frac{\omega^2}{c^2}\left(n_1^2 - n_2^2\right), \\
\Rightarrow k_{\mathrm{T}}a \;&<\; ka\sqrt{n_1^2 - n_2^2}.
\end{aligned}
$$

That is, for a mode to propagate, the argument $k_{\mathrm{T}}a$ of the Bessel function on the left hand side of the characteristic equation (48) must be smaller than ka times the numerical aperture of the fibre. The close connection between the cutoff condition and the fibre NA is not surprising, since a wide acceptance angle for incident rays translates into the ability to support a large number of optical modes. The quantity $ka\mathrm{NA}$ is sometimes called the *v-parameter* of the

fibre:

$$v = ka\text{NA} = ka\sqrt{n_1^2 - n_2^2} \approx kan_2\sqrt{2\Delta}.$$

Note that the characteristic equation at cutoff becomes (setting $\gamma = 0$ on the right hand side of (48)),

$$\frac{J_{l-1}(k_\text{T}a)}{J_l(k_\text{T}a)} = 0. \tag{49}$$

This condition determines the values of $k_\text{T}a$ at which modes switch from being freely propagating to being lossy.

5.4 $l = 0$ mode

The lowest order LP (linear polarization) fibre mode is the $l = 0$ mode, which has a field pattern given by

$$\boldsymbol{E}_\text{core}(r, \phi, z, t) = \hat{\boldsymbol{x}}E_0 \frac{J_0(k_\text{T}r)}{J_0(k_\text{T}a)} e^{\text{i}(\beta z - \omega t)}. \tag{50}$$

The characteristic equation at cutoff (49) is solved when $J_{-1}(k_\text{T}a) = 0$. Since $J_{-1}(k_\text{T}a) = J_1(k_\text{T}a)$, we require that $k_\text{T}a$ is a zero of J_1. The first such zero is at $k_\text{T}a = 0$. Therefore the lowest order LP_0 mode can propagate without loss provided that $v > 0$. That is, any v-parameter (any numerical aperture) is sufficient to support the LP_0 mode. Therefore a step-index fibre of any radius and any index contrast can support a lowest order mode. This lowest order mode is two-fold degenerate because there are two possible polarizations.

5.5 $l = 1$ mode

The LP_1 mode has an azimuthal dependence, so in cross section it has the appearance of a pair of lobes — this is of course very similar to the TEM_{01} Hermite-Gaussian mode supported by a laser cavity. The mode is 4-fold degenerate because there are two polarizations and two orthogonal orientations for the lobes. The cutoff condition of the $l = 1$ mode is found from (49) to be $J_0(k_\text{T}a) = 0$. The first zero of J_0 occurs at $k_\text{T}a \approx 2.4$, and so the cutoff condition is $v > 2.4$. For typical telecoms wavelengths of $\lambda \sim 1.5$ μm, and typical fibre contrasts, this translates into a minimum core radius of 2 μm $< a < 5$ μm. That is to say, fibres with smaller core radii cannot support the $l = 1$ mode. Higher modes are similarly unsupported in such fibres. For this reason, such narrow fibres are known as *single mode fibres*, since only the lowest order, $l = 0$ mode can propagate. In general, different fibre modes propagate with different speeds (they have different values for β), which causes data pulses to spread out and blur if they are carried over multiple modes. Therefore single mode fibres, in which all optical data is confined to a single mode, are extremely useful.

5.6   Dispersion

Pulses propagating in a fibre will travel at the group velocity determined by the gradient of β with respect to k. If the dispersion relation $\beta(k)$ is non-linear, so that the curvature $\partial_k^2\beta$ is non-zero, then in addition to a group delay, pulses will be distorted and broadened as they propagate in the fibre. The dispersion due to the curvature of β depends on the characteristic equation (48), but also on the material dispersion (the extent to which the refractive indices n_1 and n_2 depend on k). In general, the former becomes significant only for modes close to cutoff, where β becomes significantly different from $n_1 k$. If one defines the quantity b according to the relation

$$b = \frac{(\beta/k)^2 - n_2^2}{n_1^2 - n_2^2}, \tag{51}$$

one can re-write the characteristic equation in terms of b and v,

$$v\sqrt{1-b}\frac{J_{l-1}(v\sqrt{1-b})}{J_l(v\sqrt{1-b})} = -v\sqrt{b}\frac{K_{l-1}(v\sqrt{b})}{K_l(v\sqrt{b})}.$$

A choice of v implies k, and then choosing b fixes β. β can be written in terms of b by re-arranging (51) to get

$$
\begin{aligned}
\beta^2 &= k^2\left[n_2^2 + b(n_1^2 - n_2^2)\right] \\
&\approx n_2^2 k^2\left[1 + 2\Delta b\right],
\end{aligned}
$$

where in the second line we used the approximation that $\Delta \ll 1$. Taking the square root gives

$$\beta \approx n_2 k + n_2 k \Delta b.$$

Expressing this in terms of $v = ka\sqrt{n_1^2 - n_2^2} \approx kan_2\sqrt{2\Delta}$, we have

$$\beta \approx n_2 k + \frac{vb}{a}\sqrt{\frac{\Delta}{2}}.$$

Now we can express the derivative of β with respect to k as follows,

$$
\begin{aligned}
\partial_k \beta &= \partial_k\left(n_2 k\right) + \frac{1}{a}\sqrt{\frac{\Delta}{2}}\partial_k\left(vb\right) \\
&= \partial_k\left(n_2 k\right) + n_2\Delta\partial_v\left(vb\right),
\end{aligned}
$$

where we have neglected the derivative of Δ, which is small when the dispersive behaviour of n_1 and n_2 are similar. The first term represents *material dispersion*, arising from the variation of n_2 with k. The second term represents the *waveguide dispersion*, arising from the multiple reflections of the fibre mode from the core/cladding boundary. This latter term becomes particularly important close to a cut-off frequency, when β approaches $n_2 k$ and becomes very dispersive.