THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

Physics 743 — Waves and Potentials

2021

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

Tuesday, Wednesday, Friday 12 noon – 1 pm in Rm 303.610

Week	Date		Lecturer	Topic	Assessment Dates
1	Mon	1 Mar.			
	Tue	2	ME	Introduction	
	Wed	3	ME	Fourier analysis	
	Thu	4			
	Fri	5	ME	Tutorial/example 1	
2	Mon	8			
	Tue	9	ME	Discrete Fourier transform	**Assignment 1 out**
	Wed	10	ME	From oscillations to waves	
	Thu	11			
	Fri	12	ME	Tutorial/example 2	
3	Mon	15			
	Tue	16	ME	Waves and wave equations	
	Wed	17	ME	Normal modes	
	Thu	18			
	Fri	19	ME	Tutorial/example 3	
4	Mon	22			
	Tue	23	ME	Fourier wave propagation	**Assignment 1 due**
	Wed	24	ME	Diffraction	**Assignment 2 out**
	Thu	25			
	Fri	26	ME	Tutorial/example 4	
5	Mon	29			
	Tue	30	ME	Waves at interfaces I	
	Wed	31	ME	Waves at interfaces II	
	Thu	1 April			
	Fri	2	ME	Holiday	
Mid Semester Break: Monday 5 April – Friday 16 April					
6	Mon	19			**Assignment 2 due**
	Tue	20	ME	Waveguides I	**Assignment 3 out**
	Wed	21	ME	Waveguides II	
	Thu	22			
	Fri	23	ME	Tutorial/example 5	
7	Mon	26			
	Tue	27	KvW	Elasticity	**Assignment 3 due**
	Wed	28	KvW	Layered media	**Assignment 4 out**
	Thu	29			
	Fri	30	KvW	Tutorial 6 on surface waves	
8	Mon	3 May			
	Tue	4	KvW	Rayleigh waves	
	Wed	5	KvW	Dispersion	
	Thu	6			
	Fri	7	KvW	Tutorial 7 on Group vs Phase velocity	
9	Mon	10			
	Tue	11	KvW	Heat equation	**Assignment 4 due**
	Wed	12	KvW	Green function	**Assignment 5 out**
	Thu	13			
	Fri	14	KvW	Tutorial 8 on Heat in 1D	
10	Mon	17			
	Tue	18	KvW	Heat in n-D	
	Wed	19	KvW	A cooling slab	
	Thu	20			
	Fri	21	KvW	Tutorial 9 on climate change	
11	Mon	24			
	Tue	25	KvW	Intro to potentials	**Assignment 5 due**
	Wed	26	KvW	Multipoles	**Assignment 6 out**
	Thu	27			
	Fri	28	KvW	Tutorial/example 10	
12	Mon	31			
	Tue	1 June	KvW	Shells	**Assignment 6 due**
	Wed	2	KvW	Analytic functions	
	Thu	3			
	Fri	4	KvW	Tutorial/example 11 on biomedical imaging	
Study Break/Exam Period: Mon 7 June – Mon 28 June					

KvW = A. Prof. Kasper van Wijk (k.vanwijk@auckland.ac.nz), Rm.303.702, 09-923.5754
ME = A. Prof. Miro Erkintalo (m.erkintalo@auckland.ac.nz), Rm 303.505, 09-923.5598

Contents

1 Introduction

Waves are disturbances that transport energy through matter or space. They are ubiquitous, manifesting themselves in a wide variety of different physical systems: from optics and geophysics to acoustics and hydrodynamics. Remarkably, regardless of the specific physical system under consideration, the behaviour of waves is always very similar; they interfere, diffract, reflect from boundaries and so on. The purpose of these lecture notes is to describe salient topics of interdisciplinary wave physics. While practical examples are predominantly linked to optics and geophysics (due to the specialisation of the host department), we emphasise the universality of the topics discussed.

We begin by briefly recounting the description of **oscillating signals**, and how **Fourier transforms** allow us to describe arbitrary signals as superpositions of simple oscillations with different frequencies. Particular emphasis will be given to computing Fourier transforms numerically using an algorithm known as **Fast Fourier Transformation (FFT).** We will subsequently show how arrays of coupled oscillators result in **wave equations** that describe the **behaviour and characteristics of waves**, and we will analyse the ensuing equations and their solutions. We will then investigate the **propagation of waves** and how waves are **reflected and transmitted at interfaces that separate different media.** Finally, we will describe **waveguides** that allow light to be transversely confined and longitudinally transported over great distances. The silica-glass optical fibres that form the backbone of modern telecommunications are an example of such waveguides (for light waves).

The focus of these lecture notes is to apply the mathematical ideas pertinent to oscillations and waves to gain better understanding on salient physical phenomena. Given the very limited time available, this focus on physics inevitably somewhat limits the mathematical depth. There are a number of excellent references that provide more details on the underlying mathematics, and that have acted as references for the material contained in this booklet:

References

- *A guided tour of mathematical methods for the physical sciences* by R. Snieder and K. Van Wijk.
- *Linear Systems and Noise with Applications* by S. Tan and C. Fox. Available as PDF.
- *Linear Systems* by John A. Scales.
- *Fundamentals of Photonics* by B. E. A. Saleh and M. C. Teich. Available online via UoA library.
- *Optics* by E. Hecht.

2 Oscillations, signals, and Fourier analysis

Waves are intimately intertwined with oscillations, and a wave is often described as oscillations that occur simultaneously in space and time. Moreover, one may argue that oscillations form the building blocks that allow waves to exist and propagate. Because of their intimate connection, many of the concepts and tools we use to describe waves stem from the description of oscillating signals. We therefore begin our discussion from the beginning, i.e., from oscillations and on the description of arbitrary signals as superpositions of simple oscillations.

Oscillations abound in nature, ranging from the periodic motion of a pendulum to the time-evolution of the electric field of an electromagnetic wave at a fixed position. We write a generic signal that is oscillating as a function of t (which may or may not represent time) as

$$y(t) = A_{\mathrm{r}} \cos(\omega t + \phi), \tag{2.1}$$

where A_{r} is the real oscillation amplitude, $\omega = 2\pi f$ is the angular frequency (f is the ordinary frequency), and ϕ is the phase of the oscillation (at $t = 0$). Mathematically, it is often easiest to deal with the equivalent complex form

$$y(t) = \mathrm{Re}\left[Ae^{i\omega t}\right], \tag{2.2}$$

where $A = A_{\mathrm{r}} \exp[i\phi]$ is a complex amplitude. In this course, we are interested in systems that are "linear"[1]. For such systems, one can perform all the mathematical analyses using the phasor form $y(t) = Ae^{i\omega t}$ and then just take the real part in the end. This simplifies things considerably.

The oscillating signal given by Eq. (2.1) has no beginning or an end; it is a mathematical idealisation that does not strictly speaking exist in physical reality. All real signals are more complicated than simple harmonic motion, and by the very least, they will begin somewhere and end somewhere else. Figure 1 qualitatively illustrates the difference between a real signal and the idealisation given by Eq. (2.1). In the context of electromagnetic waves, we could say that the signal described by Eq. (2.1) is perfectly *monochromatic*, containing just a single frequency ω. Real signals are never fully monochromatic, but instead are composed of a range of frequencies.

Although real signals never coincide exactly with simple harmonic oscillations, they can be described as a superpositions of oscillations with different frequencies. This is the idea behind Fourier analysis, which forms the backbone of modern signal processing.[2] Fourier analysis has numerous applications, ranging from image processing to pricing of financial instruments, and in fact plays a key role in our daily lives, enabling e.g. compression of audio and video files for streaming via Spotify or Netflix. Moreover, as we shall see, Fourier analysis plays a key role in the description of wave physics. There are numerous variants of Fourier analysis, and in what follows, we will focus on two of them: (continuous) Fourier transformation and the discrete Fourier transformation.

[1] This means that the superposition principle holds, and that the equations describing the systems are linear with the oscillation amplitude.

[2] Fourier analysis is named after Joseph Fourier, who showed in 1822 that representing a function as a sum of trigonometric functions could significantly simplify the analyses of physical situations. Interestingly, however, Carl Friedrich Gauss came up with similar ideas about 20 years earlier, and even devised a clever algorithm to compute Fourier transforms that was rediscovered in 1965 and has subsequently become the cornerstone of modern digital signal processing. More about that later!

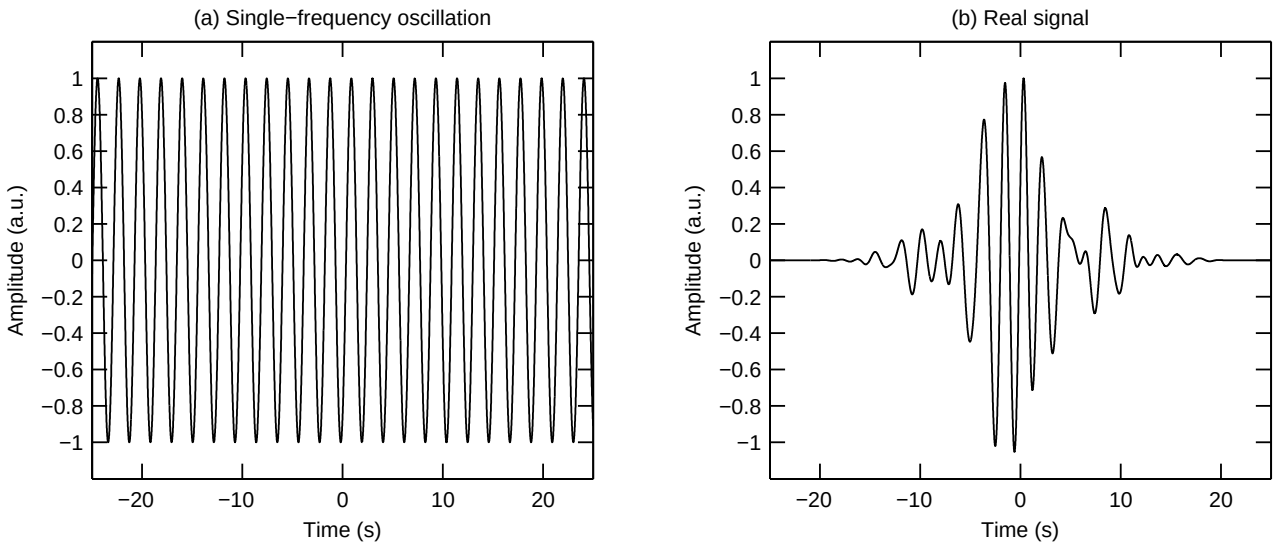

Figure 1: (a) Idealized monochromatic oscillation described by Eq. (2.1). (b) A possible real signal.

2.1 Fourier transform

Let us consider an arbitrary[3] function $f(t)$ that could describe e.g. the pressure at your ear when you are listening someone playing the chord C major on a guitar. We define the Fourier transform of $f(t)$ as:

$$\tilde{F}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t}\, dt. \tag{2.3}$$

Note that, in these lecture notes, we typically use a capital letter and a tilde to indicate the Fourier transform. The original function $f(t)$ can be recovered via the *inverse* Fourier transform:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{F}(\omega) e^{i\omega t}\, d\omega. \tag{2.4}$$

The physical meaning of $\tilde{F}(\omega)$ should be clear from this inverse transformation. Specifically, the function $f(t)$ is made out of a (possibly) continuous range of frequency components, with $\tilde{F}(\omega)$ representing the complex amplitude of the signal oscillating at frequency ω. In other words, $\tilde{F}(\omega)$ can be understood as the "weight" with which a harmonically oscillating signal $\exp(i\omega t)$ contributes to the function $f(t)$.

You may wonder why on earth would we want to express a perfectly fine signal $f(t)$ as a superposition of different frequency components. The reason is that it is often much easier to analyse and manipulate the signal in the frequency domain rather than time domain. For example, some signals may appear very complicated in

[3]Mathematically speaking, "arbitrary" might be stretching it, as functions have to satisfy certain conditions for their Fourier transforms to exist. Functions describing real-life signals always satisfy these conditions, and for some important functions for which the conditions are not met, the limitations can be overcome by using generalized functions, such as the Dirac delta function.

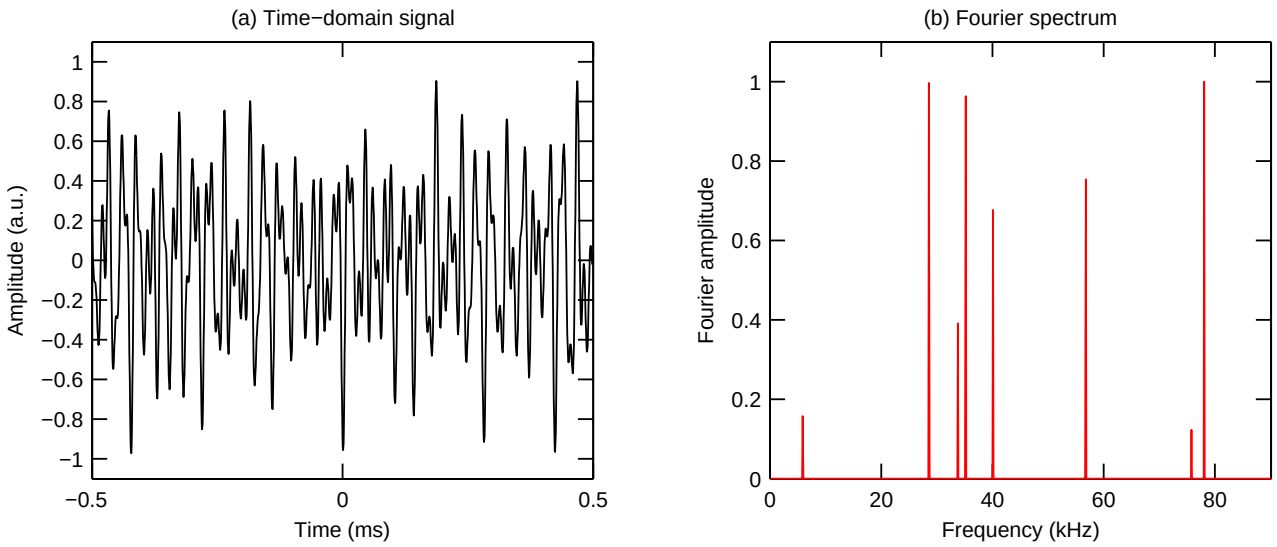

Figure 2: Example of a Fourier transform pair. (a) Time-domain signal and (b) corresponding Fourier spectrum, i.e., $|\tilde{F}(\omega)|$ plotted as a function of the ordinary frequency $f = \omega/(2\pi)$.

the time domain, even though they are very simple in the frequency domain. Figure 2 shows an example. Here, Fig. 2(a) shows a time-domain signal $f(t)$ which appears very chaotic and provides little insights. In contrast, Fig. 2(b) shows the corresponding Fourier transform, and we now see that in fact the signal is nothing but a superposition of eight waves with different frequencies and complex amplitudes. Clearly, Fig. 2(b) provides much more insights into the soul of the signal.

Two notes about Fourier transform

1. Both the function $f(t)$ and its Fourier transform $\tilde{F}(\omega)$ can in general be *complex* functions of a *real* variable. Even if the original function $f(t)$ is real-valued, the Fourier transform will be complex. This is nothing to be scared about: $\tilde{F}(\omega)$ simply corresponds to the complex amplitude introduced in Eq. (2.2), with its amplitude $|\tilde{F}(\omega)|$ and phase $\arg[\tilde{F}(\omega)]$ describing the amplitude and phase of the corresponding sinusoid. Mathematically, the reason that the Fourier transform is complex (even for real signals) is simply because we elect to use phasors of the form $\exp(i\omega t)$ rather than sines and cosines. In this context, one also notes that the (inverse) Fourier integral involves both positive and *negative* frequencies. The latter may seem a bit strange from a physical perspective, but again simply arises because of the complex phasor representation. For example, considering a simple sinusoid $y(t) = A_\mathrm{r}\cos(\omega t + \phi)$, we may write

$$y(t) = \frac{1}{2}\left[Ae^{i\omega t} + A^* e^{-i\omega t}\right], \tag{2.5}$$

where $A = A_\mathrm{r}\exp(i\phi)$ and A^* is the complex conjugate of A. This example shows that the complex representation of the signal $y(t)$ involves both negative and positive frequency components; the same is true for all real signals.

2. There are several different definitions of the Fourier transform, with Eqs. (2.3) and (2.4) constituting just one of them. Another definition would associate the negative sign in the exponential with the inverse rather than the forward transform. Moreover, some definitions use the ordinary frequency f rather than the angular frequency ω. The benefit of this notation is that the equations become symmetric, as the scaling factor $1/(2\pi)$ disappears from Eq. (2.4) through the change of variable. We have opted to use the form defined above because it is more relatable to the physics that follows. It must be emphasised that, regardless of the definition, the results are always the same, though one must be careful with signs and factors of 2π.

2.1.1 Basic properties of Fourier transform

Below we list some important properties of Fourier transform that will be useful for the our subsequent discussion. A more complete list can be found from various sources, including Wikipedia. The notation used below is such that Fourier transform pair is represented with a double-arrow, e.g. $f(t) \leftrightarrow \tilde{F}(\omega)$. The properties can be proven using the definitions of the Fourier transform given above, and is left as exercise.

1. **Linearity.** For any real or complex constants c_1 and c_2,

$$c_1 f_1(t) + c_2 f_2(t) \leftrightarrow c_1 \tilde{F}_1(\omega) + c_2 \tilde{F}_2(\omega). \tag{2.6}$$

2. **Shifting.** For a function $f(t - t_0)$ that is *shifted* by an amount t_0 compared to the function $f(t)$,

$$f(t - t_0) \leftrightarrow \tilde{F}(\omega) e^{-i\omega t_0}. \tag{2.7}$$

   Thus, the Fourier transform of $f(t - t_0)$ is the same as the Fourier transform of the un-shifted function $f(t)$ multiplied with $\exp(-i\omega t_0)$. In other words, the shift does not affect the magnitude of the spectral components, but introduces a linear phase shift whose slope depends on the amount of shift t_0. The *dual* of this property states that a linear phase in the time-domain corresponds to a shift in the frequency domain:

$$f(t) e^{i\omega_0 t} \leftrightarrow \tilde{F}(\omega - \omega_0). \tag{2.8}$$

3. **Scaling.**

$$f(at) \leftrightarrow \frac{1}{|a|} \tilde{F}\left(\frac{\omega}{a}\right). \tag{2.9}$$

4. **Differentiation.** Fourier transforms are extremely useful for numerically calculating derivatives and for solving differential equations. This is due to following handy property:

$$\frac{\mathrm{d}f(t)}{\mathrm{dt}} \leftrightarrow (i\omega)\tilde{F}(\omega). \tag{2.10}$$

   The derivative of a function can thus be obtained by (1) calculating the Fourier transform of the function, (2) multiplying the Fourier transform with $(i\omega)$, and (3) taking the inverse Fourier transform.

5. **Symmetry.** Fourier transforms display several symmetry properties depending on whether the original function $f(t)$ is real, imaginary, even, or odd. Arguably the most important symmetry is the fact that the Fourier transform of a real function $f(t)$ is Hermitian, i.e.,

$$\tilde{F}(-\omega) = \tilde{F}^*(\omega). \tag{2.11}$$

This implies that the real (imaginary) part of the Fourier transform is even (odd). Since real-world measured signals are always real, their Fourier transforms are Hermitian. An important consequence is that negative frequency components are "redundant", as they can be fully reconstructed from the positive frequency components. For this reason, when plotting Fourier transforms of real-world measured signals, it is conventional to only show positive frequency components [see e.g. Fig. 2(b)].

6. **Convolution.** The Fourier transform of a convolution of two functions is the product of the functions' Fourier transforms:

$$(f * g)(t) \leftrightarrow \tilde{F}(\omega)\tilde{G}(\omega). \tag{2.12}$$

This property is extremely useful for calculating convolutions: one can simply (1) calculate the Fourier transforms of the two functions, (2) multiply them together, and (3) take an inverse Fourier transform. The *dual* of this property states that a product in the time-domain corresponds to a convolution in the frequency domain:

$$f(t)g(t) \leftrightarrow (\tilde{F} * \tilde{G})(\omega). \tag{2.13}$$

7. **Parseval's theorem.** For a Fourier transform pair $f(t) \leftrightarrow \tilde{F}(\omega)$,

$$\int_{-\infty}^{\infty} |f(t)|^2 \, \mathrm{d}t = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\tilde{F}(\omega)|^2 \, \mathrm{d}\omega. \tag{2.14}$$

The interpretation of this property is that the total "energy" calculated in the time and in the frequency domains is the same. As an explicit example, if $|f(t)|^2$ represents instantaneous power with units of $\mathrm{J/s}$, the Fourier transform $|\tilde{F}(\omega)|^2$ will represent energy per unit frequency with units of $\mathrm{J/Hz}$. From this, it should be clear that both integrals will have units of Joules, hence describing energy.

2.1.2 Dirac delta function

Strictly speaking, Fourier transforms can be calculated only for a narrow class of functions which decrease sufficiently rapidly to zero as time goes to infinity. Indeed, it is only for such "localised" functions that the Fourier integral given by Eq. (2.3) can be evaluated. This is rather annoying, since it suggests that one cannot evaluate Fourier transforms of many important functions such as standard trigonometric functions that extend to infinity. Luckily, the problem can be resolved by referring to a function known as the Dirac delta function $\delta(x)$. Despite its name, the Dirac delta function is rigorously speaking not really a function at all. Rather, it corresponds to a *generalised* function or a distribution.

Loosely speaking, the Dirac delta function $\delta(t)$ can be understood as an infinitely narrow and infinitely tall peak centred at $x = 0$. You can think of it as the limit of a rectangle with width Δ and amplitude $1/\Delta$ as Δ

approaches zero. Alternatively, you can think of it as a distribution that is non-zero at one point only, but in such a way that its integral is equal to one:

$$\int_a^b \delta(x)\,\mathrm{d}x = \begin{cases} 1 & \text{if } 0 \in (a,b) \\ 0 & \text{if } 0 \notin [a,b] \end{cases} \tag{2.15}$$

More rigorously, the delta function is perhaps best considered in relation to how it affects other "test" functions when integrated against them[4]. In particular, the delta function is defined to exhibit the *substitution* or *sifting* property, which essentially allows for "extraction" of the value of a function at a given point:

$$\int_{-\infty}^{\infty} \delta(x-a) f(x)\,\mathrm{d}x = f(a). \tag{2.16}$$

You can think of this formula as showing how to *decompose* the function $f(t)$ into a linear combination of delta functions. Alternatively, the formula essentially says that the convolution between the Dirac delta function and some other function $f(t)$ is equal to the function $f(t)$ itself.

The above sifting property can be used together with the Fourier transformations defined above to derive an alternative representation for the delta function [exercise]:

$$\boxed{\delta(x-a) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ip(x-a)}\,\mathrm{d}p.} \tag{2.17}$$

This interpretation is of critical importance for Fourier analysis, and its interpretation will be described shortly.

2.1.3 Some important Fourier transforms

We now summarise the Fourier transforms of some particularly important functions with the help of the Dirac delta function. Complete proofs are again left as exercises.

1. **Complex harmonic oscillation.** Consider the complex function $f(t) = \exp(i\omega_0 t)$ describing simple harmonic oscillation at frequency ω_0. Using Eq. (2.17), it is easy to see that the Fourier transform will be given by

   $$\boxed{e^{i\omega_0 t} \leftrightarrow 2\pi\delta(\omega-\omega_0).} \tag{2.18}$$

   This expression shows that the Fourier transform of a monochromatic oscillation is a delta function, which is rather satisfying: monochromatic oscillation only contains one single frequency component, so the Fourier "spectrum" must accordingly have non-zero value only at one frequency.

2. **Sines and cosines.** The Fourier transforms of sine and cosine functions can be readily obtained by using (1) Euler's formula, (2) the linearity of Fourier transforms, and (3) Eq. (2.17):

   $$\boxed{\begin{aligned} \sin(\omega_0 t) &\leftrightarrow -i\pi\left[\delta(\omega-\omega_0) - \delta(\omega+\omega_0)\right], \qquad (2.19)\\ \cos(\omega_0 t) &\leftrightarrow \quad \pi\left[\delta(\omega-\omega_0) + \delta(\omega+\omega_0)\right]. \qquad (2.20) \end{aligned}}$$

[4] Similarly to other generalized functions or distributions.

These expressions show that sines and cosines are made out two frequency components at ω_0 and $-\omega_0$, which is in line with our expectation based on Euler's formula. It is also evident that the only thing distinguishing sine and a cosine is that the phases of the frequency components are different. Recalling that $\pm i = \exp(\pm i\pi/2)$, we see that sine and cosine are 90° out of phase as expected.

3. **Gaussian.** The Fourier transform of a Gaussian function $f(t) = \exp(-\alpha t^2)$ can be shown to be (very good exercise!):

$$\exp(-at^2) \leftrightarrow \sqrt{\frac{\pi}{\alpha}} e^{-\omega^2/(4\alpha)}. \tag{2.21}$$

In other words, the Fourier transform of a Gaussian is a Gaussian! Importantly, the e^{-1} width of the Gaussian[5] in the time domain is given by $\Delta t = 1/\sqrt{\alpha}$, while in the frequency domain it is given by $\Delta\omega = 2\sqrt{\alpha}$. We thus have

$$\Delta t \Delta\omega = 2. \tag{2.22}$$

This implies that the width in the frequency domain is inversely proportional to the width in the time domain (and vice versa). The feature is not unique to Gaussian functions, but is rather a general property of any localised signal. It follows that short signals in the time domain must be associated with broad spectra in the frequency domain. Heisenberg's uncertainty theorem is a particular manifestation of this relationship, as is the fact that ultrashort pulses of laser light require broad spectral widths.

2.2 Fourier transform of a periodic function

Let us consider a periodic function $f_\mathrm{p}(t)$ with a period T_p, and define $f(t)$ to be a single period (or cycle) of this function [see Fig. 3]. Mathematically,

$$f(t) = \begin{cases} f_\mathrm{p}(t) & \text{if } t \in (-T_\mathrm{p}/2, T_\mathrm{p}/2) \\ 0 & \text{otherwise,} \end{cases} \tag{2.23}$$

By definition, the Fourier transform of the periodic function $f_\mathrm{p}(t)$ is

$$\tilde{F}_\mathrm{p}(\omega) = \int_{-\infty}^{\infty} f_\mathrm{p}(t) e^{-i\omega t}\, \mathrm{d}t. \tag{2.24}$$

We can divide the total integral into a sum of integrals evaluated over each period of the signal:

$$\tilde{F}_\mathrm{p}(\omega) = \sum_{n=-\infty}^{\infty} \int_{nT_\mathrm{p}-T_\mathrm{p}/2}^{nT_\mathrm{p}+T_\mathrm{p}/2} f_\mathrm{p}(t) e^{-i\omega t}\, \mathrm{d}t. \tag{2.25}$$

[5] This width is the separation between the peak of the Gaussian at $t = 0$ and the point where the Gaussian has decayed to the value e^{-1}.

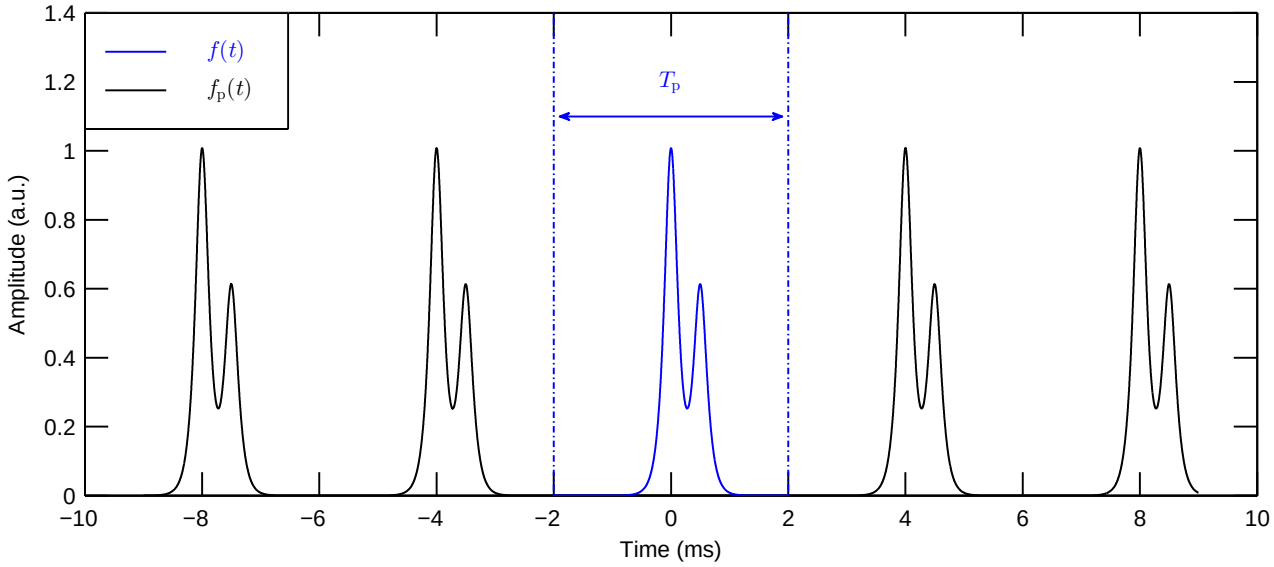

Figure 3: Example of a periodic function with period T_p. The blue curve highlights a single period (or cycle) of the function.

Then, using the periodicity of the signal and doing a change of variable $t = t' + nT_\mathrm{p}$, we can replace the periodic signal $f_\mathrm{p}(t)$ with the function $f(t)$ describing a single period:

$$\tilde{F}_\mathrm{p}(\omega) = \sum_{n=-\infty}^{\infty} \int_{-T_\mathrm{p}/2}^{T_\mathrm{p}/2} f(t') e^{-i\omega(t'+nT_\mathrm{p})} \, \mathrm{d}t' \tag{2.26}$$

$$= \sum_{n=-\infty}^{\infty} e^{-i\omega n T_\mathrm{p}} \int_{-T_\mathrm{p}/2}^{T_\mathrm{p}/2} f(t') e^{-i\omega t'} \, \mathrm{d}t' \tag{2.27}$$

$$= \sum_{n=-\infty}^{\infty} e^{-i\omega n T_\mathrm{p}} \tilde{F}(\omega), \tag{2.28}$$

where $\tilde{F}(\omega)$ is the Fourier transform of a single period of the original function, i.e., $\tilde{F}(\omega) \leftrightarrow f(t)$. To shed more light into this expression, we must discuss the prefactor $\sum_{n=-\infty}^{\infty} \exp(-i\omega n T_\mathrm{p})$. This surprisingly corresponds to an interesting generalized function known as a *Dirac comb*.

Dirac comb

The Dirac comb is a generalized function that essentially corresponds to a sequence of periodically repeating Dirac delta functions. It is defined as

$$\mathrm{III}_X(x) = \sum_{n=-\infty}^{\infty} \delta(x - nX), \tag{2.29}$$

where X denotes the period with which the delta functions repeat. It can be shown (exercise) that the Dirac comb can also be expressed as an infinite sum of simple exponential functions:

$$\mathrm{III}_X(x) = \frac{1}{X} \sum_{n=-\infty}^{\infty} e^{-i2\pi n \frac{x}{X}}. \tag{2.30}$$

The validity of this expression can easily be demonstrated graphically by considering the limit of the sum of a finite number of exponentials as the number of terms in the sum approaches infinity. Formal proof is a bit trickier [exercise].

With the help of the Dirac comb, the Fourier transform given by Eq. (2.28) can be written as

$$\tilde{F}_{\mathrm{p}}(\omega) = \tilde{F}(\omega)\Delta\omega \sum_{n=-\infty}^{\infty} \delta(\omega - n\Delta\omega), \tag{2.31}$$

$$= \Delta\omega \sum_{n=-\infty}^{\infty} \tilde{F}\left(n\Delta\omega\right) \delta(\omega - n\Delta\omega) \tag{2.32}$$

where $\Delta\omega = 2\pi/T_{\mathrm{p}}$, and the second term utilises the fact that the Dirac comb is non-zero only at frequencies $\omega = n\Delta\omega$. These expressions show that **the spectrum of a periodic function $f_{\mathrm{p}}(t)$ with period T_{p} consists of discrete components separated by $\Delta\omega = 2\pi/T_{\mathrm{p}}$, enveloped by the spectrum of a single period of the function (see Fig. 4).**

Practical example: the 2005 Nobel Prize in Physics

Mode-locked lasers are devices that emit ultrashort bursts of laser light. These pulses are typically emitted as a periodic sequence, with the temporal separation between two pulses corresponding to the so-called "repetition time" t_{r}. In light of the discussion above, the spectrum of the entire mode-locked pulse train must consist of discrete components separated by $\Delta\omega = 2\pi/t_{\mathrm{rep}}$, as shown in Fig. 4. This is indeed found to be the case. The implications are tremendous: each of the discrete frequency components, which there are typically hundreds of thousands, essentially corresponds to an ultrastable laser in its own right. The resulting "optical frequency combs" have found numerous applications, and were recognised by the 2005 Nobel Prize in physics.

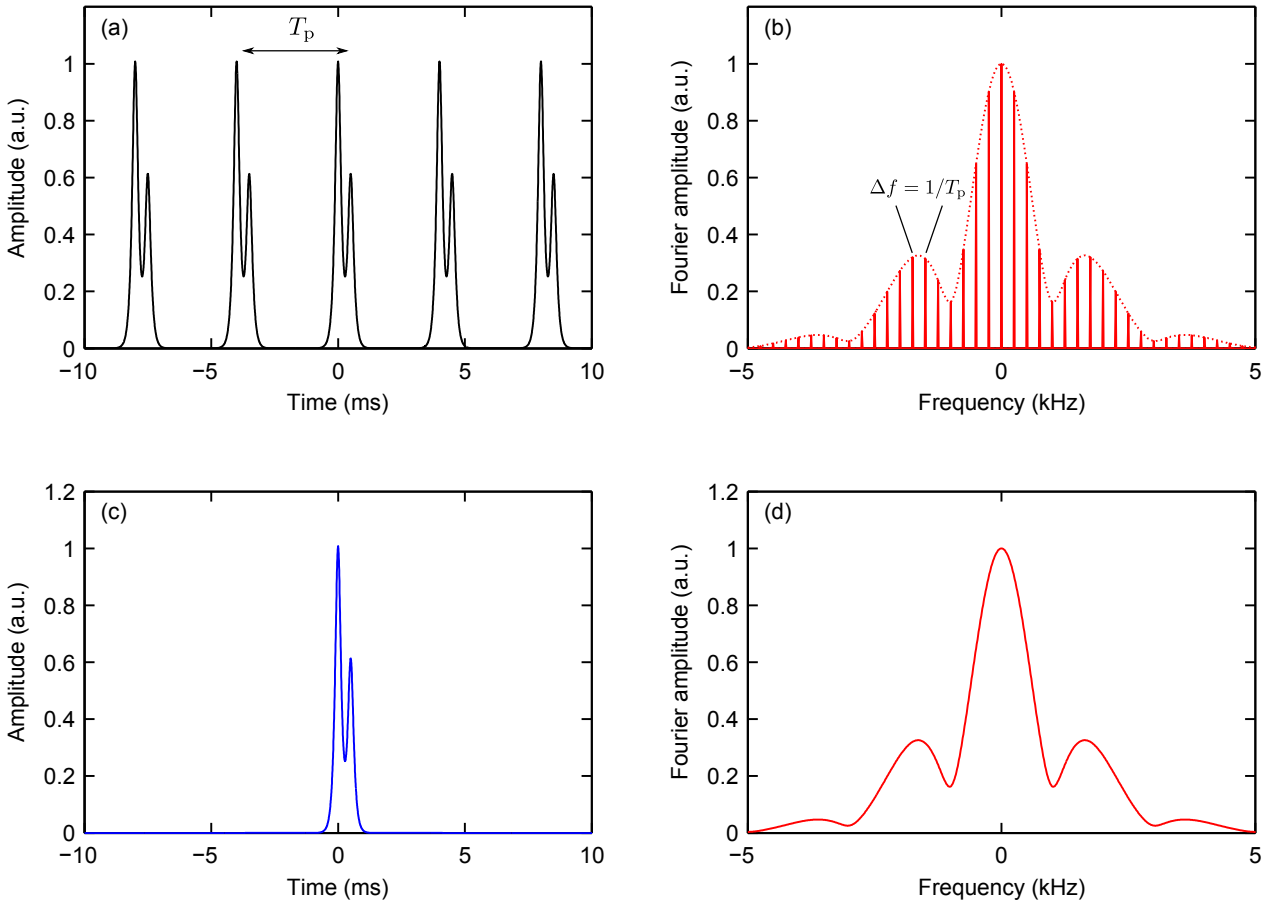

Figure 4: Comparison of the Fourier transforms of a periodic function (a, b) and a single cycle of that function (c, d). Solid curves in (a) and (c) show the time-domain signals while solid curves in (b) and (d) show the corresponding Fourier transforms. The dotted curve in (b) highlights that the discrete components that make up the spectrum of the periodic signal is *enveloped* by the specturm of a single period of that function. Note: Fourier spectra plotted as a function of the ordinary frequency $f = \omega/(2\pi)$. Also note that, since the signals are real, the Fourier spectra are symmetric.

2.3 Fourier series

It is instructive to take the inverse Fourier transform of $\tilde{F}_\mathrm{p}(\omega)$ given by Eq. (2.32) to recover the original function:

$$f_\mathrm{p}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{F}_\mathrm{p}(\omega)e^{i\omega t}\,\mathrm{d}\omega \tag{2.33}$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} \Delta\omega \sum_{n=-\infty}^{\infty} \tilde{F}\left(n\Delta\omega\right)\delta(\omega - n\Delta\omega)e^{i\omega t}\,\mathrm{d}\omega \tag{2.34}$$

$$= \frac{1}{2\pi}\sum_{n=-\infty}^{\infty} \Delta\omega\tilde{F}(n\Delta\omega)e^{in\Delta\omega t}. \tag{2.35}$$

Recalling $\Delta\omega = 2\pi/T_\mathrm{p}$, we can write this expression in a slightly different form:

$$f_\mathrm{p}(t) = \sum_{n=-\infty}^{\infty} c_n e^{i\frac{2\pi n}{T_\mathrm{p}}t}, \tag{2.36}$$

where we also introduced the coefficient $c_n = \Delta\omega\tilde{F}(n\Delta\omega)/(2\pi)$. Writing out the Fourier integral, this coefficient can be written as

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} f(t)e^{-i\frac{2\pi n}{T_\mathrm{p}}t}. \tag{2.37}$$

We can now recognise Eq. (2.36) as the familiar *Fourier series*, with Eq. (2.37) providing the corresponding Fourier coefficients. To summarise:

1. A periodic function can be written as a sum of sines and cosines[a] with different frequencies. Historically, this was the first form of Fourier (series) representation of a (periodic) function, introduced by Joseph Fourier in 1822.

2. The Fourier series representation arises naturally from the theory of continuous Fourier transforms. In particular, the Fourier transform of a periodic function is equal to the Fourier series. It follows that the Fourier transform is more general, being applicable also to functions (and distributions) that are not periodic.

[a] Equation (2.36) can be readily transformed into sines and cosines using Euler's formula.

Problems

2.1 Use Eqs. (2.3) and (2.4) to prove the basic properties of Fourier transform listed in Subsection 2.1.1.

2.2 Derive the sifting property of the Dirac delta function [see Eq. (2.16)] from the definition given by Eq. (2.15).

2.3 Use the sifting property of the Dirac delta function [see Eq. (2.16)], together with the definitions of the Fourier transform, to derive Eq. (2.17).

2.4 Explicitly prove all of the Fourier transforms listed in Subsection (2.1.3). Use the definitions of the Dirac delta function where needed.

2.5 Show that the Dirac comb, defined by Eq. (2.29), can be expressed in the form given by Eq. (2.30). Hint: since the Dirac comb is periodic, it can be expressed as a Fourier series.

2.6 Consider a function $f(t)$ comprising of two Gaussian features with identical durations that are separated from one another by time t_0. Derive an expression for the *power spectrum* of the function, i.e., $|\tilde{F}(\omega)|^2$. Figure 5 shows an example of the power spectrum for some unknown separation t_0. Estimate t_0 for this example.

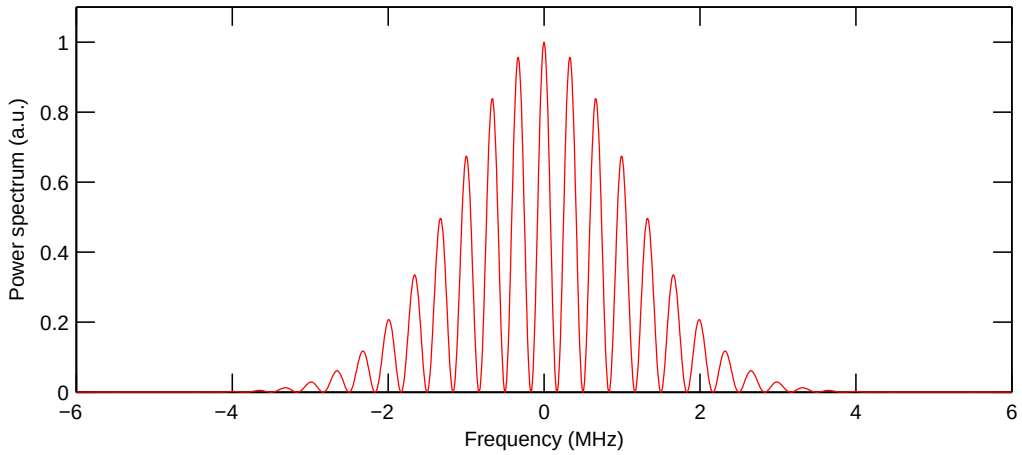

Figure 5: Power spectrum of a time-domain function consisting of two Gaussian features temporally separated from one another by t_0.

2.7 Consider a damped, driven harmonic oscillator. As will be shown in Section 4, the Fourier transform of the oscillator displacement [$x(t) \leftrightarrow \tilde{X}(\omega)$] is given by:

$$\tilde{X}(\omega) = \frac{\tilde{F}(\omega)}{\omega_0^2 - \omega^2 + 2i\gamma\omega}, \tag{2.38}$$

where $\tilde{F}(\omega)$ is the Fourier transform of the time-domain driving function $f(t)$, ω_0 is the resonance frequency, and γ is the damping coefficient. Assume that an oscillator that is initially at rest ($x(t) = 0$ for $t < 0$) is perturbed with an impulse at $t = 0$, such that $f(t) = \delta(t)$. Calculate the time-domain response $x(t)$ of the oscillator by explicitly evaluating the inverse Fourier transform of $\tilde{X}(\omega)$.

3 Sampling in time and discrete Fourier transform

The Fourier transform takes a continuous function $f(t)$ and gives out another continuous function $\tilde{F}(\omega)$ that describes the frequency spectrum of the original function. However, in most real-world applications, we do not have knowledge of the functional form of the signal we are interested in. Rather, we only have access to (digital) data that *samples* a continuous (analog) waveform at discrete points [see Fig. 6]. For example, while sound is a continuous waveform of pressure variations, the only way we can store and process it on a computer is by taking discrete samples of the pressure variations at finite points in time. This raises three questions: what is the Fourier spectrum of a discretely sampled signal, how does it relate to the spectrum of the original analog waveform, and how can we efficiently compute the spectrum on a computer?

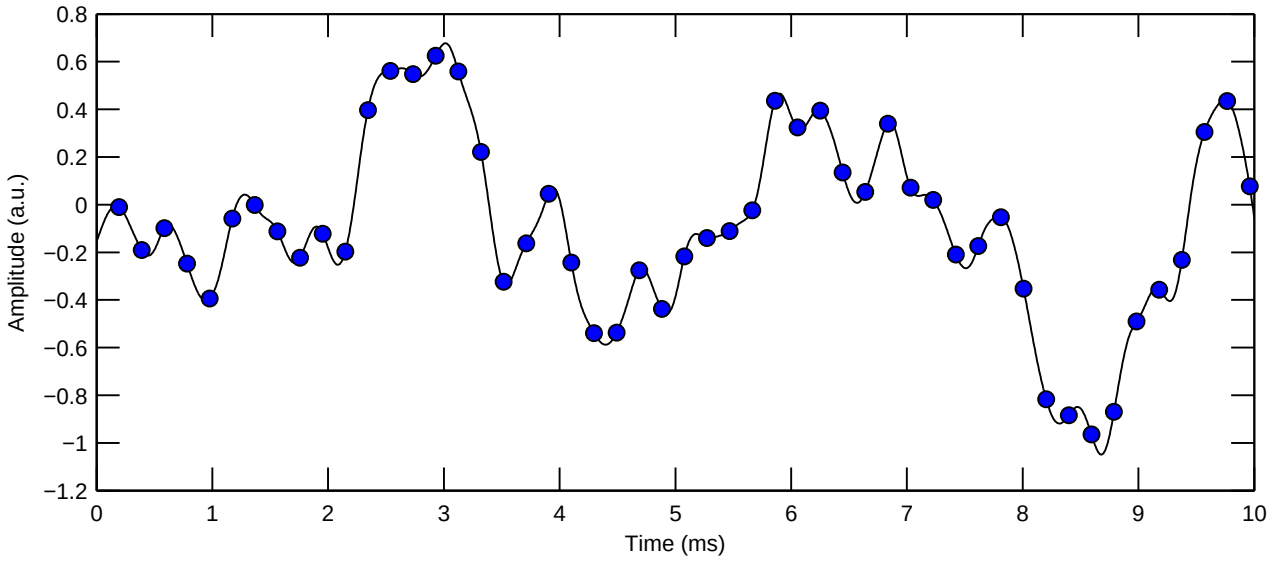

Figure 6: Illustration of sampling in time. The blue circles represent discrete and equally-spaced samples taken from the underlying continuous waveform.

3.1 Sampling in time and discrete-time Fourier transform

Let us first consider a continuous waveform described by a function $f(t)$ that is sampled at equally-spaced points in time. Mathematically, we may represent the process of sampling by multiplying the function $f(t)$ with a Dirac comb to give a new (generalised) function $f_{\mathrm{s}}(t)$:

$$f_{\mathrm{s}}(t) = f(t) \sum_{n=-\infty}^{\infty} \delta(t - nT_{\mathrm{s}}) = \sum_{n=-\infty}^{\infty} f(nT_{\mathrm{s}})\delta(t - nT_{\mathrm{s}}), \tag{3.1}$$

where T_{s} is the *sampling period*, i.e., the time-separation between two sampling points.

The Fourier transform of $f_{\mathrm{s}}(t)$ is known as the *discrete-time Fourier transform*[6], and is given by[7]

$$\tilde{F}_{\mathrm{s}}(\omega) = \int_{-\infty}^{\infty} f_{\mathrm{s}}(t) e^{-i\omega t}\, dt, \tag{3.2}$$

$$= \frac{1}{T_{\mathrm{s}}} \sum_{n=-\infty}^{\infty} \tilde{F}\left(\omega - 2\pi n/T_{\mathrm{s}}\right), \tag{3.3}$$

where $\tilde{F}(\omega)$ is the Fourier transform of the un-sampled function $f(t)$ [see Fig. 7].

And so we see that the Fourier spectrum of a sampled function is *periodic*, with the period given by $2\pi/T_{\mathrm{s}}$ in angular frequency. This should not be particularly surprising, since the situation is simply the *dual* of the Fourier transform of the periodic signal considered in Section 2.2:

1. The Fourier transform of a periodic function is an enveloped Dirac comb.
2. The Fourier transform of an enveloped Dirac comb is a periodic function.

We also see that the Fourier spectrum of a sampled waveform can be constructed by shifting and adding up copies of the Fourier transform of the underlying continuous waveform [see Fig. 7]. This observation leads to a fundamental requirement for the period T_{s} with which the waveform must be sampled for there to be no information loss:

Nyquist-Shannon sampling theorem

To recover the original waveform from sampled data, the latter must contain the entire undistorted spectrum of the former. This requires that the shifted spectra added up in Eq. (3.3) do not overlap. If the spectra do overlap, they will add up in the overlap region, resulting in distortions that are not present in the original waveform, which prevents us from recovering the original spectrum. To avoid this issue, the sampling time T_{s} must be sufficiently short so as to ensure that the shifted spectra do not overlap. Quantitatively, if the largest angular frequency contained in the original waveform is $\omega_{\mathrm{c}} = 2\pi f_{\mathrm{c}}$, then the amount with which the spectra are shifted in Eq. (3.3) must be at least $2\omega_{\mathrm{c}}$[a]. This gives the following requirement for the sampling time:

$$T_{\mathrm{s}} < \frac{1}{2f_{\mathrm{c}}}. \tag{3.4}$$

The highest frequency that can be reconstructed from the sampled signal [$f_{\mathrm{c}} = 1/(2T_{\mathrm{s}})$] is known as the Nyquist frequency. Digitisation of music (or sound) is an excellent example. Common audio tracks stored on CDs or in other digital formats are sampled at 44.1 kHz, corresponding to a sampling period $T_{\mathrm{s}} \approx 23$ μs. This yields a Nyquist frequency of $f_{\mathrm{c}} \approx 22.1$ kHz, allowing the full range of frequencies heard by humans to be represented (human hearing extends to 20 kHz). On the other hand, there is no point to use a larger sampling frequency, since human hearing does not extend beyond 20 kHz.

As an example, imagine we sample a simple sine wave with frequency f_0 and corresponding pe-

[6] Note that discrete-time Fourier transform is not the same as discrete Fourier transform. Very confusing but not my fault.

[7] To show this, one first uses Eq. (2.30) to replace the sum of Dirac delta functions with exponential functions, then changes the order of summation and integration, and finally does a change of variable $\omega' = \omega - 2\pi n/T$.

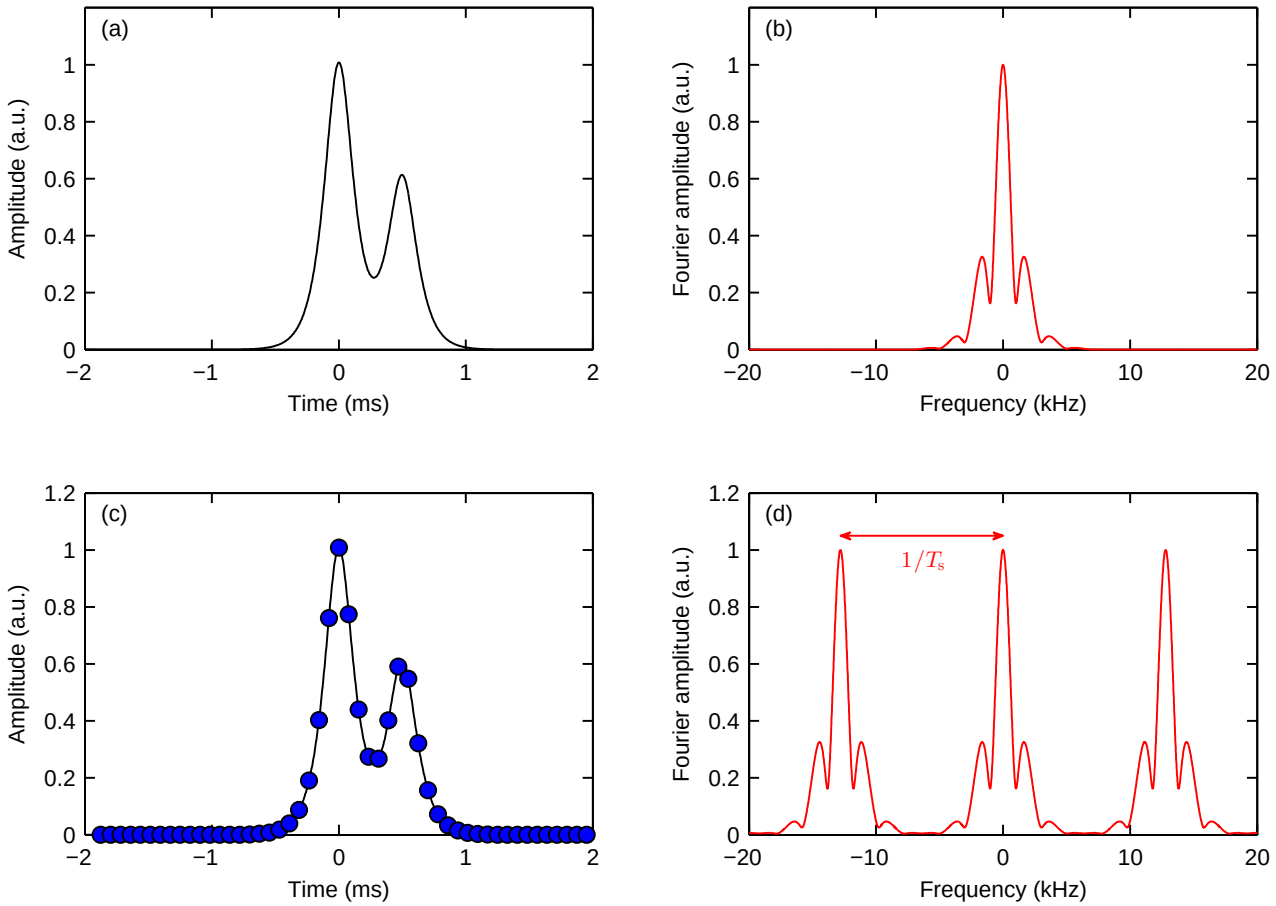

Figure 7: Illustration of discrete-time Fourier transform. (a, b) Continuous-time waveform (a) and corresponding Fourier transform (b). (c) Discretely sampled version of the waveform shown in (a). (d) Discrete-time Fourier transform corresponding to the sampled waveform shown in (c). The Fourier transform of the sampled signal repeats in frequency with a period $1/T_s$. Note: spectra plotted as a function of the ordinary frequency.

riod $t_0 = 1/f_0$. Equation (3.4) shows that reconstruction of this waveform requires that the sampling time $T_s < t_0/2$. Thus, we need at least two sample points per period.

[a] The factor of two comes from the fact that the Fourier spectrum of real waveforms is symmetric versus the zero frequency. In this case, the sampling time must be sufficiently short so that the ω_c component does not overlap with the $-\omega_c$ component of the next term in the sum. Hence the factor of two.

3.2 Discrete Fourier transform

Our discussion above imposed no limits on the number of points in our sampled waveform (the sum in Eq. (3.1) extends to infinity). This is clearly not physical in terms of actual real-world data stored on a computer, which

is not only discrete but also finite. Indeed, real-world data constitutes a sampled version of an analog waveform *over some finite region with duration* T_p. To calculate the Fourier spectrum of such data, we need to make an assumption on what the waveform is doing outside the region with which we have sampled it (recall that the Fourier integral goes to infinity).

A particularly good approach is to *assume* that the waveform is periodic with a period T_p, as shown in Fig. 8. Indeed, with this assumption, we know that the Fourier spectrum will only be composed of discrete components spaced by $2\pi/T_\mathrm{p}$ [see Section 2.2]. Moreover, because the original data is sampled with a sampling time of T_s, we know that the Fourier spectrum will be periodic with a period of $2\pi/T_\mathrm{s}$ (see previous subsection). This implies that all of the spectral information is contained at a discrete set of angular frequencies between zero and $2\pi/T_\mathrm{s}$ that are spaced by $2\pi/T_\mathrm{p}$, i.e., at frequencies $\omega_n = n2\pi/T_\mathrm{p}$ where $n = 0...(N-1)$ and $N = T_\mathrm{p}/T_\mathrm{s}$ is the total number of sampling points [see Fig. 8]. The implication is that we can take a waveform sampled at N evenly-spaced points and obtain the Fourier spectrum at N frequencies without having to calculate a single integral going to infinity. The resulting procedure is known as the discrete Fourier transform (DFT) and is arguably the most important variant of Fourier analysis since it can be efficiently calculated on a computer for real-world data that is discrete and finite.

Mathematically, we can represent a waveform sampled with N points as

$$f_\mathrm{s}(t) = \sum_{m=0}^{N-1} f(t)\delta(t - mT_\mathrm{s}). \tag{3.5}$$

Taking the continuous Fourier transform yields

$$\tilde{F}_\mathrm{s}(\omega) = \sum_{m=0}^{N-1} f(mT_\mathrm{s})e^{-i\omega mT_\mathrm{s}}. \tag{3.6}$$

Since we are assuming that $f_\mathrm{s}(t)$ corresponds to a sampled version of the single cycle $[f(t)]$ of an underlying periodic waveform $[f_\mathrm{p}(t)]$, we know that the total spectrum is nonzero only at frequencies $\omega = n\Delta\omega$ [see Section 2.2]. Thus, we only need to evaluate $\tilde{F}_\mathrm{s}(\omega)$ at $\omega = n\Delta\omega$ to obtain the spectrum of our (assumed) periodic waveform:

$$\tilde{F}_\mathrm{p}(n\Delta\omega) = \sum_{m=0}^{N-1} f_\mathrm{s}(mT_\mathrm{s})e^{-in\Delta\omega mT_\mathrm{s}}. \tag{3.7}$$

Note that we have chosen to ignore the prefactor $\Delta\omega$ that arises in the Fourier transform of a periodic signal; this is not particularly important. Recalling that $\Delta\omega = 2\pi/T_\mathrm{p}$ and that $N = T_\mathrm{p}/T_\mathrm{s}$, we finally obtain

$$\tilde{F}_\mathrm{p}(n\Delta\omega) = \sum_{m=0}^{N-1} f_\mathrm{s}(mT_\mathrm{s})e^{-\frac{2\pi i}{N}nm}. \tag{3.8}$$

This corresponds to the *discrete Fourier transform* (DFT), which allows us to obtain a (finite) discrete Fourier spectrum from a (finite) sampled data $f_\mathrm{s}(t)$.

We can express the DFT relationship more algorithmically by referring exclusively to the indices n and m, without referring to the actual time and frequency at all. Specifically, given a sequence of (complex) numbers with components $x[k]$ for $k = 0, 1, ..., N-1$, the DFT is another sequence $X[n]$ for $n = 0, 1, ..., N-1$ defined

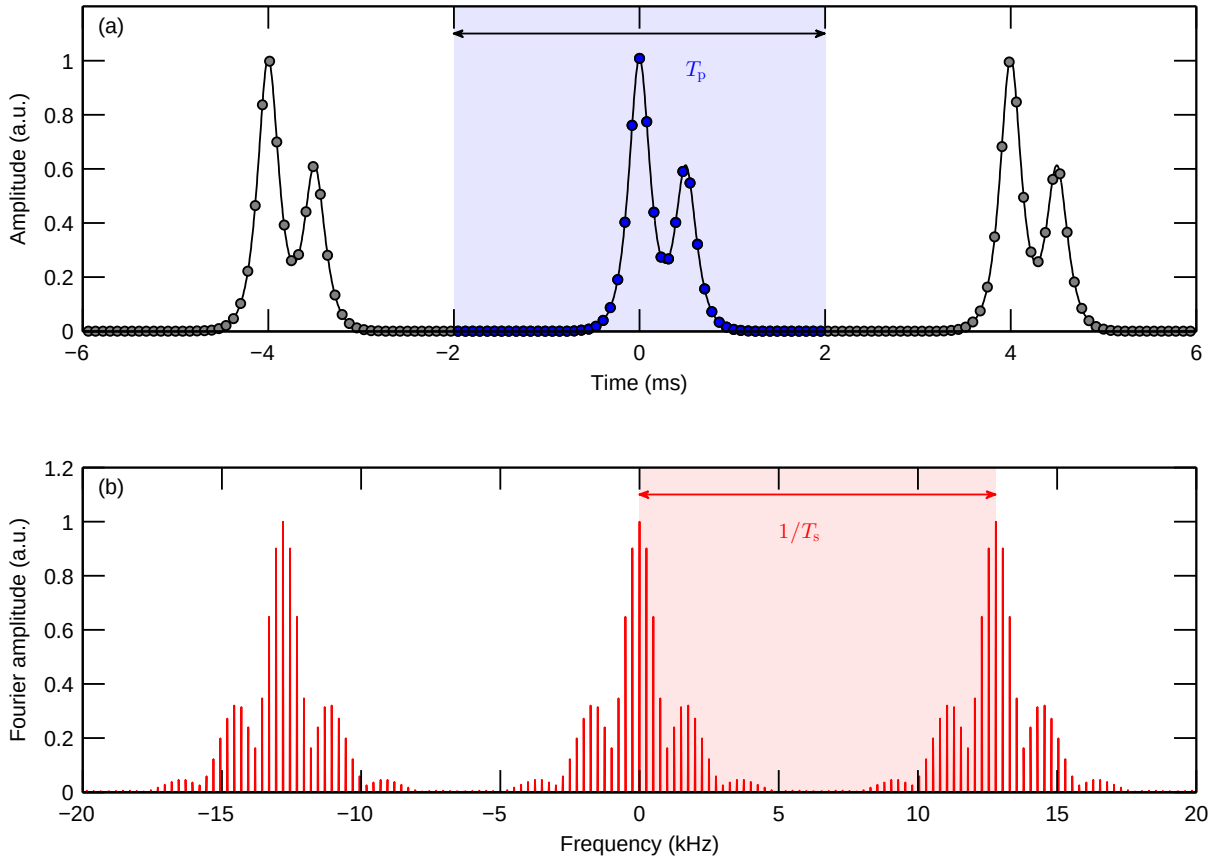

Figure 8: Schematic illustration of discrete Fourier transform. (a) A time-domain signal composed of a finite number of points sampled at even intervals separated by T_s (blue) is *assumed* to be part of a larger waveform that repeats periodically with a period T_p. (b) The Fourier spectrum of the *assumed* periodic waveform is periodic with period $1/T_\mathrm{s}$ and non-zero only at frequencies separated by $1/T_\mathrm{p}$. A single period of the Fourier spectrum contains all the information available (red shaded area). Hence, we only need to evaluate the Fourier components at the discrete frequencies within the shaded region.

by

$$X[n] = \sum_{m=0}^{N-1} x[m] e^{-\frac{2\pi i}{N} nm}. \tag{3.9}$$

This expression is formally identical to the one derived above. The corresponding inverse Fourier transform is

$$x[m] = \frac{1}{N} \sum_{n=0}^{N-1} X[n] e^{\frac{2\pi i}{N} nm}. \tag{3.10}$$

Note that different conventions exist for the normalisation factor $1/N$.

3.3 DFT with a computer

The beauty of the DFT lies in the fact that it can be easily computed as simple sums, without the need for any infinite integrals. Morever, it can be very efficiently calculated on a computer using a clever algorithm known as Fast Fourier Transform (FFT). The modern development of the algorithm is credited to the 1965 work by James Cooley and John Tukey, but in fact, it was already alluded to by Carl Friedrich Gauss in 1805 – more than a decade before Fourier came up with his series! We will not discuss the FFT algorithm itself, but we will summarise how it is used in typical computational environments (e.g. Python or Matlab).

FFT is truly one of the cornerstones of modern signal processing, and it is widely used in image, sound, and video compression and editing, analysis of scientific data, numerical integration and differentiation and so on. Because of its universal usefulness, most platforms of scientific computation include routines for performing FFT. This is typically very simple:

1. **Python.** In Python, FFT is included both in the numpy and and scipy libraries (among others). Given a sequence of numbers with N components $x[k]$:

   ```
   from scipy.fftpack import fft
   X = fft(x)
   ```

2. **Matlab.** FFT is a standard function in Matlab. Given a sequence of numbers with components $x[k]$:

   ```
   X = fft(x);
   ```

Both examples return a sequence of numbers $X[n]$ with N components according to Eq. (3.9). The only difficulty is then to interpret what frequencies do these different numbers represent. Comparison of Eqs. (3.8) and (3.9) make this simple: the n^{th} component corresponds to the frequency

$$\boxed{\omega[n] = n\Delta\omega = n\frac{2\pi}{T_{\text{p}}},} \tag{3.11}$$

where $T_{\text{p}} = NT_{\text{s}}$ is the entire duration of the sampled data signal. Thus, to construct the FFT *frequency grid*, the only thing we need to know is the sampling time T_{s}. We can then extract the number of points N from the data and construct both the time- and frequency grids.

There is one important point to note. Specifically, because the DFT is essentially a Fourier transform of discretely sampled data, the Fourier spectrum is periodic [see Fig. 8]. This implies that the latter half of the frequencies can be interpreted as negative frequency components, and this indeed is the convention. Assuming N is even[8], the frequency grid should be interpreted as

$$\omega[n] = [0, 1, ...n/2 - 1, -n/2, ... - 1]\Delta\omega. \tag{3.12}$$

Note how the frequencies turn negative after the mid-point. If the signal being analysed is real, then the second half of the DFT spectrum (corresponding to negative frequencies) is of course redundant thanks to the Hermitian symmetry. However, if the signal is not real, the negative frequency components contain new information.

The numpy library in Python contains a handy function that returns the FFT frequency grid in terms of the ordinary frequency $f = \omega/(2\pi)$:

[8] FFT typically performs better for even numbers of sampling points

```
import numpy as np
f = np.fft.fftfreq(N, d = Ts)
```

where `N` is the number of sample points and `Ts` is the sampling time. Below we give an explicit example of the use of FFT with Python:

```
import numpy as np
from scipy.fftpack import fft
import matplotlib.pyplot as plt

Tper = 4
f0 = 1/Tper

Tp = 8*Tper
N = 2**10
Ts = 40*Tper/N

t = np.arange(0,N)*Ts
x = np.cos(2*np.pi*f0*t)
f = np.fft.fftfreq(N,d = Ts)
X = fft(x)

plt.plot(f,X)
```

Two common problems

There are two important things to be aware of when using DFT to compute spectra of real-world signals on a computer. Both of these things result directly from the theoretical foundations outlined above:

1. For the spectrum to be representative of the appropriate physical waveform, the sampling period T_s must satisfy the Nyquist-Shannon sampling theorem. If the sampling period is too large, the spectrum will not accurately represent the spectrum of the actual waveform. The resulting artifacts are referred to as "aliasing".

2. The DFT explicitly assumes that the data being transformed is a single period of a fully periodic waveform with period T_p. However, if the first and the last points of the sampled data are not identical, then an unphysical discontinuity must be present in the underlying periodic waveform. This results in artifacts in the evaluated spectrum. Such artifacts can be readily seen when considering the example above with e.g. `Tp = 8.2*Tper` such that the time window does not contain an integer multiple of sinusoidal oscillations. The FFT no longer contains infinitely sharp peaks, but there is rather some energy leakage to neighbouring Fourier frequencies. Bottom line is that, when considering data sequences that fill the entire sampling window, one has to bear in mind the fact that DFT assumes periodic boundaries.

Problems

3.1 In Canvas, you will find the file `tohoku.npy`, which contains a numpy array of seismometer data recorded during the 2011 Tohoku earthquake. The sampling rate of the data is 1 Hz.

(a) Write a Python code that loads the data and plots the seismometer data as a function of time. Describe your observations.

(b) Estimate the speed of the seismic wave corresponding to the earthquake.

(c) Plot the Fourier spectrum of the seismic wave. Describe your observations.

3.2 In Canvas, you will find the audio files `32a.wav` and `32b.wav`. The former (latter) is a recording of the the lecturer playing a free note (chord) on an acoustic guitar. Use Fourier analysis (and the internet) to answer the following questions.

(a) What is the free note recorded on file `32a.wav`?

(b) What are the dominant frequencies that make up the the recording on file `32b.wav`? With dominant frequencies, we mean those that do not correspond to harmonics (or overtones).

(c) What is the chord played on the recording on file `32b.wav`?

(d) Has the lecturer been careful with tuning the instrument?

Hint: the following commands allow you to load audio files with Python:

```
import scipy.io.wavfile
rate, signal = scipy.io.wavfile.read('32a.wav')
```

Here, the variable `rate` will contain the *sampling rate* $1/T_\mathrm{s}$ in units of Hz, while the variable `signal` contains the actual audio file (i.e., the pressure wave generated by a loudspeaker).

3.3 Consider the hyperbolic secant function, which is of particular significance in many areas of physics,

$$f(t) = \mathrm{sech}\left(\frac{t}{T_0}\right) = \frac{2}{e^{t/T_0} + e^{-t/T_0}}, \tag{3.13}$$

where T_0 is a parameter that defines the width (or duration) of the function. Write a Matlab/Python code that uses FFT to numerically demonstrate that the Fourier transform

$$\tilde{F}(\omega) \propto \mathrm{sech}\left(\frac{\pi T_0}{2}\omega\right). \tag{3.14}$$

Note the proportionality: for simplicity, you can ignore prefactors and normalize the FFT result to unity.

3.4 Write a Matlab/Python code that evaluates the derivative of the hyperbolic secant function given by Eq. (3.13). Check your result against the analytical formula

$$\frac{df(t)}{dt} = -\frac{1}{T_0}\mathrm{sech}\left(\frac{t}{T_0}\right)\tanh\left(\frac{t}{T_0}\right). \tag{3.15}$$

Note: evaluating derivatives in this manner is very handy, especially when an analytical result is not available. It is also very straightforward to evaluate higher-order derivatives, which may be cumbersome to do analytically. For example, consider the second or third order derivatives of the hyperbolic secant function. Would you rather evaluate them analytically or numerically?

3.5 In Canvas, you will find the file `35.mat`. This file contains the complex amplitude of the slowly-varying envelope of an ultra-short light pulse that has propagated through a highly nonlinear material, as well as the corresponding time base. Mathematically, if we write the complex electric field as

$$E(t) = A(t)e^{i\omega_0 t}, \tag{3.16}$$

the amplitude contained in the file corresponds to $A(t)$. Assuming $\omega_0/(2\pi) = 400$ THz, write a Python/-Matlab code that plots the power spectrum $|\tilde{E}(\omega)|^2$ of the light pulse in *decibel* scale as a function of *wavelength* from 500 nm to 1200 nm. Note that, since $A(t)$ is complex, you cannot discard negative frequencies. Rather consider what they mean, and consider shifting things around so that the "zero" frequency lies in the middle of the frequency grid. Python and Matlab have functions `fftshift` that do this shifting for you. Note that the data is stored in a Matlab file format: to open this on python, use the following code

```
import scipy.io
mat = scipy.io.loadmat('35.mat',squeeze_me=True)

t = mat['t']
A = mat['A']
```

4 From oscillations to waves

In this Section, we will describe how wave equations arise naturally from the analysis of coupled oscillators. We will start with a brief recap into harmonic oscillations, and apply our newly-developed knowledge of Fourier transforms to solve the pertinent equations.

4.1 Damped, driven harmonic oscillator

In the preceding two Sections, we have mathematically described how arbitrary signals can be represented as superpositions of simple oscillatory functions. But of course, (superpositions of) oscillations are not only mathematical constructs, but they arise very naturally in numerous physical systems. A familiar example — that also happens to be of particular significance — is that of the *harmonic oscillator*, which is typically first encountered in undergraduate physics in the form of a mass attached to a spring. It turns out, however, that the harmonic oscillator model is not limited to masses on springs, but is rather remarkably universal, appearing in numerous situations such as electronic circuits, acoustical systems, light-matter interaction, optical resonators and so on.

The equation of motion of any second-order linear oscillatory system can be cast into the following *universal oscillator equation* form:

$$\boxed{\frac{d^2x(t)}{dt^2} + 2\gamma\frac{dx}{dt} + \omega_0^2 x = f(t),} \tag{4.1}$$

where γ is a damping coefficient, ω_0 is the characteristic frequency of the oscillator, and $f(t)$ is the driving function. We can obtain the familiar equation of motion for a mass m attached to a spring by replacing $f(t) = F(t)/m$, where $F(t)$ is the driving force with units of Newtons[9]. In this case, the characteristic frequency $\omega_0^2 = k/m$, where k is the spring constant.

In the absence of damping ($\gamma = 0$) and driving ($f(t) = 0$), Eq. (4.1) admits solutions in the form of simple harmonic oscillations: $x(t) = A\cos(\omega_0 t)$. If damping (but not driving) is present, the solutions depend on the ratio γ/ω_0; in the case of weak damping ($\gamma \ll \omega_0$), the solutions correspond to exponentially damped oscillations. In the general case, where both damping and driving are present, Eq. (4.1) can be easily solved using Fourier transforms. Specifically, Fourier transforming both sides of the equation and using the *differentiation rule* given by Eq. (2.10), we obtain

$$-\omega^2\tilde{X}(\omega) + 2i\gamma\omega\tilde{X}(\omega) + \omega_0^2\tilde{X}(\omega) = \tilde{F}(\omega), \tag{4.2}$$

where $\tilde{X}(\omega)$ and $\tilde{F}(\omega)$ are the Fourier transforms of $x(t)$ and $f(t)$, respectively. Simple manipulation yields

$$\boxed{\tilde{X}(\omega) = \frac{\tilde{F}(\omega)}{\omega_0^2 - \omega^2 + 2i\gamma\omega} = \tilde{S}(\omega)\tilde{F}(\omega),} \tag{4.3}$$

where we defined the *frequency response function*

$$\tilde{S}(\omega) = \frac{1}{\omega_0^2 - \omega^2 + 2i\gamma\omega}. \tag{4.4}$$

[9]The exact same form, with $F(t) = -qE(t)$, would describe the motion of an electron in a time-varying electric field.

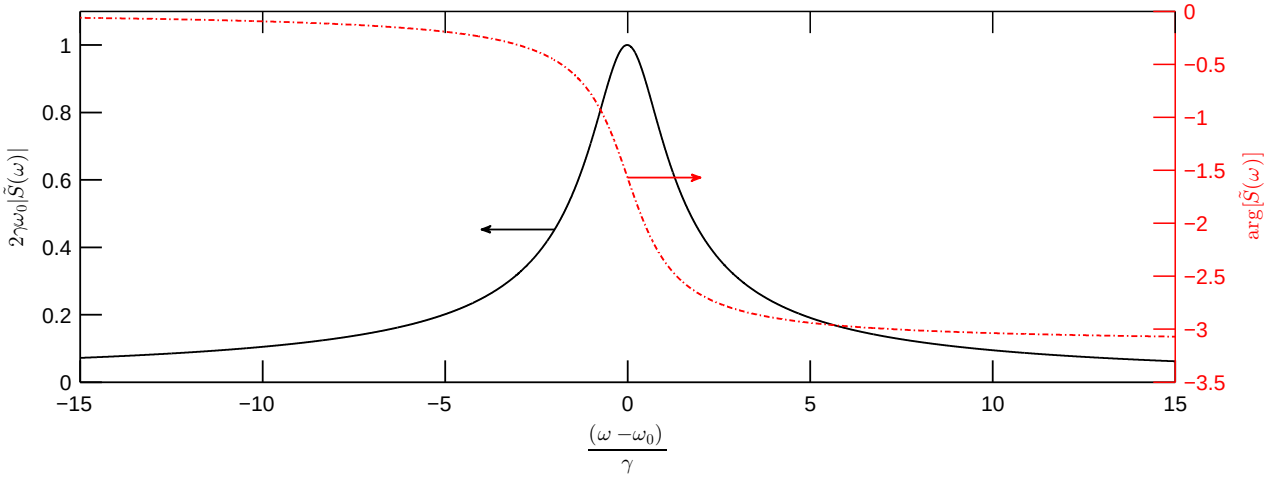

Figure 9: Amplitude (black solid curve) and phase (red dash-dotted curve) of the frequency-response function $\tilde{S}(\omega)$ of a damped driven harmonic oscillator in the vicinity of $\omega \approx \omega_0$. Note how the y- and x-axes are normalised. Also note that the response function is Hermitian, i.e., $\tilde{S}^*(\omega) = \tilde{S}(-\omega)$, which means that there is another resonance peak in the vicinity of $\omega \approx -\omega_0$ with inverted phase.

Equation (4.3) shows that the Fourier spectrum of $x(t)$ can be obtained by multiplying each frequency component of the Fourier spectrum of the driving function $f(t)$ with an appropriate complex number [i.e., $\tilde{S}(\omega)$] [10]. The corresponding time-domain behaviour can then be obtained via inverse Fourier transform, regardless of the form of the driving function. In the particular case where the driving function is harmonic, such that $f(t) = \cos(\omega_f t)$, it can be shown that the oscillator motion is also harmonic, and given by [exercise]

$$x(t) = |\tilde{S}(\omega_f)| \cos(\omega_f t + \phi), \tag{4.5}$$

where $\phi = \arg[\tilde{S}(\omega_f)]$. In other words, the magnitude (phase) of the response function sets the amplitude (phase) of the oscillator motion relative to the driving function.

Figure 9 shows an example of the phase and amplitude profiles of the response function defined in Eq. (4.3) in the vicinity of $\omega \approx \omega_0$. As can be seen, the amplitude of oscillation is maximised when $\omega = \omega_0$; this is the concept of *resonance*. In contrast, for frequencies that satisfy $|\omega - \omega_0| \gg \gamma$, the oscillation amplitude is virtually zero; the driving is out-of- resonance.

In the more general case where the driving function is not harmonic, the amplitude and phase of each spectral component of $\tilde{F}(\omega)$ must be scaled with the amplitude and phase of the response function $S(\omega)$. The resulting changes in the time-domain can be very convolved[11], but in the frequency domain everything is nice and linear.

[10] This frequency-domain relationship is in fact a universal feature of *linear systems*.

[11] Pun intended: the time-domain response can be understood as a convolution between a time-domain response function and the driving function.

4.1.1 Another perspective: Green's functions

Our approach to solving the universal oscillator equation consisted of solving the equation in the Fourier domain [Eq. (4.3)] and then taking the inverse Fourier transform to obtain the time-domain response. It is instructive to approach the problem from a slightly different perspective, although the end-result will of course be equivalent. Let us consider the special situation where the driving function is an impulse $\delta(t-t')$ centred at t', and let us denote the solution to the oscillator equation in this case as $G(t-t')$. (Note how the value of this function at time t depends on the delay between t and the centre position t' of the impulse.) By definition(s), we have

$$\left[\frac{d^2}{dt^2} + 2\gamma\frac{d}{dt} + \omega_0^2\right] G(t-t') = \delta(t-t'). \tag{4.6}$$

Let us now multiply both sides by an arbitrary driving function $f(t')$ and integrate from $-\infty$ to ∞. We obtain

$$\left[\frac{d^2}{dt^2} + 2\gamma\frac{d}{dt} + \omega_0^2\right] \int_{-\infty}^{\infty} f(t')G(t-t')\,\mathrm{d}t' = \int_{-\infty}^{\infty} f(t')\delta(t-t')\,\mathrm{d}t'. \tag{4.7}$$

Note that we can swap the order of the differential and integral operators since t and t' are independent variables. Now, the right-hand side of Eq. (4.7) is equal to $f(t)$ by virtue of the sifting property of the Dirac delta function. What this means is that the solution to the universal oscillator equation in the presence of an arbitrary driving function $f(t)$ can be written as

$$\boxed{x(t) = \int_{-\infty}^{\infty} f(t')G(t-t')\,\mathrm{d}t' = (f * G)(t).} \tag{4.8}$$

This implies that, as soon as $G(t)$ is known, we can readily obtain the solution to the universal oscillator for *any* driving function as a simple convolution! The function $G(t)$ is an example of Green's functions, which are essential for understanding the solutions to inhomogeneous linear differential equations.

Green's functions

Green's function is the *impulse response* of an inhomogeneous linear differential equation. Specifically, given a linear differential operator $L(t)$, a Green's function solves the equation

$$LG(t,t') = \delta(t-t'). \tag{4.9}$$

If the operator exhibits translational invariance, such that $L(t)$ has constant coefficients with respect to t, then the functional dependence of the Green's function can be written as $G(t,t') = G(t-t')$. For the case of the universal oscillator equation considered above, it is easy to see that the operator

$$L(t) = \frac{d^2}{dt^2} + 2\gamma\frac{d}{dt} + \omega_0^2. \tag{4.10}$$

Since the coefficients are constant with respect to t, we used the form $G(t-t')$.

As described above, Green's functions allow us to study inhomogeneous differential equations of the form

$$Lx(t) = f(t), \tag{4.11}$$

which manifest themselves in many branches of physics. For example, as we shall see, the generation and behaviour of waves obeys an equation of this form, with the driving function $f(t)$ representing a *source* term that is actually responsible for the generation of the waves. To find solutions to such equations, one must first find the Green's function for the particular problem by solving the equation in the case of an impulse driving function. Solutions for arbitrary driving functions can then be obtained by using Eq. (4.8). It must be highlighted that each differential equation is associated with a different Green's function. Moreover, Green's functions may not be unique: some systems may admit infinitely many Green's functions corresponding to different initial and/or boundary conditions.

4.1.2 Green's function for damped, driven oscillator

To find the Green's function $G(t)$ for the damped, driven oscillator equation, we must solve Eq. (4.6) with $t' = 0$. This can be accomplished in the Fourier domain as before. Considering $f(t) = \delta(t)$, we have $\tilde{F}(\omega) = 1$. Using Eq. (4.3), the Fourier transform of the Green's function satisfies [see Eq. (4.3)]

$$\tilde{G}(\omega) = \frac{1}{\omega_0^2 - \omega^2 + 2i\gamma\omega} = \tilde{S}(\omega). \tag{4.12}$$

And so we see that the Green's function corresponds to the inverse Fourier transform of the frequency-response function $S(\omega)$ defined in Eq. (4.4). This should not be particularly surprising, as Eq. (4.3) and Eq. (4.8) correspond to Fourier transform pairs by virtue of the convolution property. To re-iterate, the Green's function $G(t)$ can be understood as the impulse response of the system, which corresponds to the inverse Fourier transform of the system's frequency response.

We can explicitly evaluate the inverse Fourier transform using contour integration [exercise]. Assuming $\omega_0 > \gamma$, we obtain

$$G(t-t') = \frac{e^{-\gamma(t-t')}}{\sqrt{\omega_0^2 - \gamma^2}} \sin\left[\sqrt{\omega_0^2 - \gamma^2}(t-t')\right] H(t-t'), \tag{4.13}$$

where we defined the Heaviside step-function

$$H(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{if } t < 0. \end{cases} \tag{4.14}$$

As shown in Fig. 10, the Green's function given by Eq. (4.13) corresponds to damped oscillation that begins at $t = t'$. Physically, an impulse excites the oscillator from rest and causes it to oscillate; however, because of dissipation and the absence of any subsequent driving, the oscillations are exponentially damped.

4.2 Coupled oscillators and normal modes

Let us now consider the situation of two coupled oscillators. Our goal is to introduce the concept of *normal modes*[12], and to pave the way for a simple physical derivation of a *wave equation*. To these ends, we focus here on the simplest possible situation and the most salient physics, and ignore the effects of damping and driving.

[12] We will be looking at normal modes of some complex systems later on, so consider this section as a warm-up.

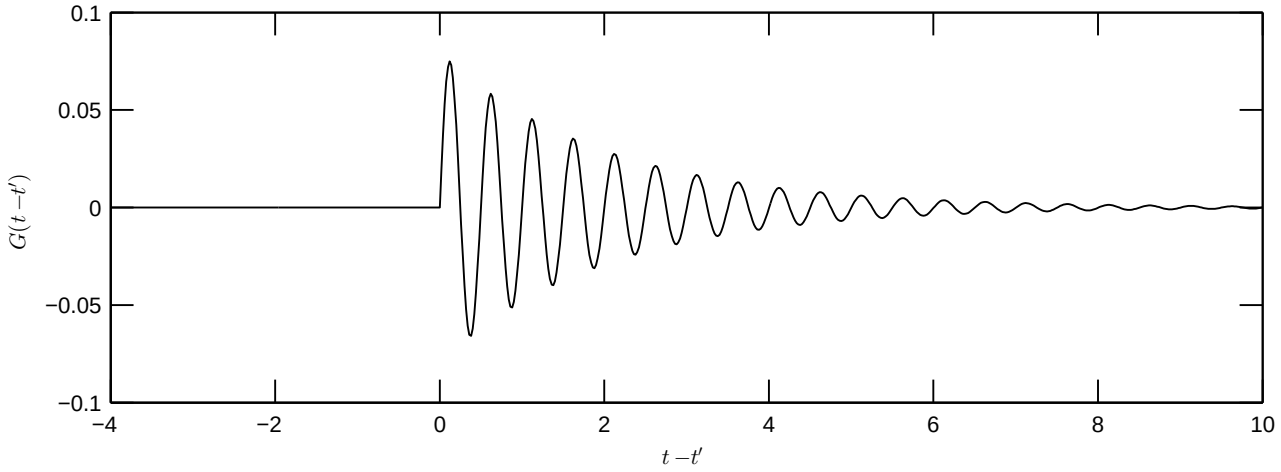

Figure 10: Green's function $G(t-t')$ for a damped, driven, harmonic oscillator with $\gamma = 0.5\ \mathrm{s}^{-1}$ and $\omega_0/(2\pi) = 2$ Hz.

As shown in Fig. 11, we consider two masses that are attached to fixed walls with springs with spring constants are k_1 and k_3, respectively. Moreover, we assume they are connected to one another by a spring k_2. It should be clear that (i) the force applied to a given mass is transmitted by the two springs it is connected to, and (ii) the force each spring transmits is governed by the extent to which the spring is compressed or extended.

Spring 1 (3) can only be compressed or extended if mass 1 (2) is displaced from its equilibrium (we denote the displacements as u_1 and u_2). On the other hand, spring 2 is compressed or stretched depending on whether the quantity $u_1 - u_2$ is positive (compression) or negative (stretched). If that spring is compressed, it will exert a force pushing mass 1 (2) to the left (right) and vice versa. The equations of motion following from Hooke's and Newton's laws are:

$$m_1 \frac{\mathrm{d}^2 u_1}{\mathrm{d}t^2} = -k_1 u_1 - k_2(u_1 - u_2), \tag{4.15}$$

$$m_2 \frac{\mathrm{d}^2 u_2}{\mathrm{d}t^2} = -k_3 u_2 + k_2(u_1 - u_2). \tag{4.16}$$

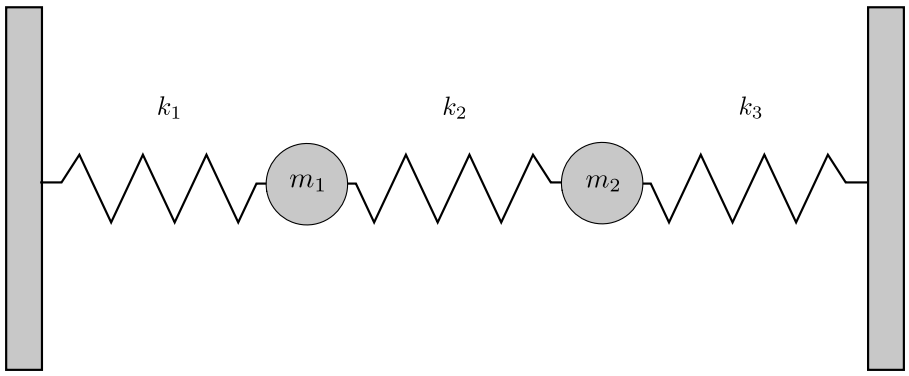

Figure 11: Schematic illustration of two coupled oscillators.

These equations of motion can be written in matrix form as

$$\frac{\mathrm{d}^2\vec{\mathbf{u}}}{\mathrm{d}t^2} = \mathrm{M}\vec{\mathbf{u}}, \tag{4.17}$$

where $\vec{\mathbf{u}} = [u_1, u_2]^{\mathrm{T}}$ and the matrix

$$\mathrm{M} = \begin{bmatrix} -\frac{k_1+k_2}{m_1} & \frac{k_2}{m_1} \\ \frac{k_2}{m_2} & -\frac{k_3+k_2}{m_2} \end{bmatrix}. \tag{4.18}$$

The general solution of the system given by Eq. (4.17) can be expressed as a linear combination of four terms of the form

$$\vec{\mathbf{u}}_{\pm 1,2} = \vec{\mathbf{a}}_{1,2} e^{\pm i\omega_{1,2}t} \tag{4.19}$$

where $\vec{\mathbf{a}}_{1,2}$ are the two linearly independent eigenvectors of the matrix M and $\omega_{1,2}$ are related to the corresponding eigenvalues $\lambda_{1,2}$ viz. $\omega_1 = \sqrt{|\lambda_{1,2}|}$. Assuming for simplicity that the masses and springs are all identical[13], the frequencies and corresponding eigenvectors are

$$\omega_1 = \omega_0 \quad \leftrightarrow \quad \vec{\mathbf{a}}_1 = [1, 1]^{\mathrm{T}} \tag{4.20}$$

$$\omega_2 = \sqrt{3}\omega_0 \quad \leftrightarrow \quad \vec{\mathbf{a}}_2 = [1, -1]^{\mathrm{T}}, \tag{4.21}$$

where $\omega_0^2 = k/m$. We can draw important conclusions from the above results.

Normal modes

A single harmonic oscillator has a single characteristic frequency ω_0. In the case of a simple spring-mass system, this frequency is related to the spring constant and mass viz. $\omega_0^2 = k/m$. However, when we couple two such oscillators together, as in our analysis above, we find that the overall system exhibits **two** characteristic frequencies; in our example above, those frequencies are ω_0 and $\sqrt{3}\omega_0$. Importantly, all the constituents of the system (in our case two springs) simultaneously oscillate with the same frequency (either ω_0 or $\sqrt{3}\omega_0$ in our example). However, the oscillations can occur in different directions, as was found for our two-spring system: the eigenvectors show that, when the system is oscillating at ω_0, both masses are displaced in the same direction, while the opposite is true for $\sqrt{3}\omega_0$.

The characteristic motion patterns (oscillation frequencies, directions) identified above are known as the *normal modes* of the system. They can be understood as the patterns at which the overall system "wants" to oscillate. While our example of coupled springs is not particularly practical, normal modes are an important concept that arise in many different branches of science and engineering. For example, the resonant frequencies of a laser cavity, oscillations of the Earth after an earthquake, or the motion of a drum[a] all represent examples of normal mode behaviour.

[a] You know, the instrument you hit with a stick.

[13] The writer is sitting on a plane while writing this, and solving the eigenvalues and eigenvectors for the more general case seems too much for this environment!

4.3 Continuum limit

Our analysis above can be readily extended to as many oscillators as one desires. Here we are interested in the "continuum limit", i.e., the limit where we have an infinitely many oscillators that are infinitely close to each other. (The physical interpretation is presented in the end; it is obvious.) In this case, each oscillator is coupled to its two neighbouring oscillators. Assuming all the springs and masses to be identical, the equation of motion for the displacement u_i reads

$$m\frac{\mathrm{d}^2 u_i}{\mathrm{d}t^2} = k(u_{i+1} - u_i) + k(u_{i-1} - u_i) \tag{4.22}$$

$$= k(u_{i+1} - 2u_i + u_{i-1}). \tag{4.23}$$

The right-hand side of Eq. (4.23) resembles the finite-difference approximation of a second derivative. Specifically, given a function $u(x)$, we have

$$\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} \approx \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}, \tag{4.24}$$

with the approximation getting better as h gets smaller small. Since we are considering a linear chain of discrete oscillators, h in our case should be understood as the oscillator spacing (or lattice spacing). We can rearrange Eq. (4.23) to include this quantity (in a way that makes sense shortly):

$$\frac{\mathrm{d}^2 u_i}{\mathrm{d}t^2} = \frac{k/h}{m/h^3}\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}. \tag{4.25}$$

In the continuum limit, the spacing of the oscillators goes to zero, the number of oscillators goes to infinity, and the mass of a single oscillator must go to zero (otherwise the mass would un-physically diverge to infinity). In this limit, the discrete set of displacements u_n can be replaced with the continuous function $u(x,t)$. Moreover, we can identify the limits $\lim_{h\to 0} m/h^3$ and $\lim_{h\to 0} k/h$ as the density (ρ) and stiffness per unit length (E) of the resulting continuous medium. Taking the limit thus yields

$$\boxed{\frac{\partial^2 u(x,t)}{\partial t^2} = \frac{E}{\rho}\frac{\partial^2 u(x,t)}{\partial x^2}.} \tag{4.26}$$

And so we have derived the *wave equation* for a thin elastic rod! Let us now put some physics into the derivation.

Waves and coupled oscillators

The derivation above may feel very abstract: when would we ever encounter coupled oscillators? The answer is "all the time". In fact, different media support waves precisely because they are essentially (the continuum limits of) coupled oscillators. Consider a string. All parts of the string are connected (or coupled), and so displacing one region upwards forces the neighbouring regions to similarly shift upwards. And so a wave can propagate through a string. Similarly consider sound. Air molecules move about their equilibrium positions, colliding with each other in a manner that gives rise to (spring-like) restoring forces.

This simple idea of coupling provides significant insights. For example, in fluids such as air, longitudinal (sound) waves can propagate because different positions are coupled through collisions between diffrent molecules. Fluids do not, however, support transverse waves since there is no coupling in the transverse direction (an air molecule displaced upwards will not pull its neighbours with it). Simple as that.

Problems

4.1 Consider a damped, driven harmonic oscillator described by Eq. (4.1). Show that, for a harmonic driving $f(t) = \cos(\omega_0 t)$, the oscillator response can be written as

$$x(t) = |\tilde{S}(\omega_0)| \cos(\omega_0 t + \phi), \tag{4.27}$$

where $\tilde{S}(\omega_0)$ is the frequency-domain response function given by Eq. (4.4) and $\phi = \arg[\tilde{S}(\omega_0)]$.

4.2 The quantity $|\tilde{S}(\omega)|^2$ is often of more interest than $|\tilde{S}(\omega)|$, as the former is directly related to the energy of oscillation. Show that, for $\omega_0 \gg \gamma$, the quantity $|\tilde{S}(\omega)|^2$ can be written as a Lorentzian function. Derive an expression for the full-width at half maximum of this Lorenzian. **Hint:** When $\omega_0 \gg \gamma$, $|\tilde{S}(\omega)|^2$ is highly localised around ω_0. Thus, is may be a good idea to approximate $\omega \approx \omega_0$ somewhere.

4.3 Show that Eq. (4.13) corresponds to the Green's function of the damped, driven, harmonic oscillator equation [Eq. (4.1)].

4.4 Show that a linear combination of four terms of the form given by Eq. (4.19) satisfy the differential Eq. (4.17). Further show that, for $k_1 = k_2 = k_3$ and $m_1 = m_2$, the angular frequencies $\omega_1 = \omega_0$ and $\omega_2 = \sqrt{3}\omega_0$ where $\omega_0^2 = k/m$, and that the corresponding eigenvectors $\vec{\mathbf{a}}_1 = [1, 1]^{\mathrm{T}}$ and $\vec{\mathbf{a}}_2 = [1, -1]^{\mathrm{T}}$.

4.5 Equation (4.26) can be identified as the wave equation for a thin solid rod, where longitudinal deformations propagate through the rod. In this case, the coefficient E should be understood as the Young's modulus defined by

$$E = \frac{\sigma}{\epsilon} = \frac{F/A}{\Delta L/L_0}, \tag{4.28}$$

where $\sigma = F/A$ is the force applied along the rod per unit area (stress) and $\epsilon = \Delta L/L_0$ is the change in length (ΔL) of the rod divided by the original length (L) when a force F compresses or extends the rod (strain). Suppose that a longitudinal wave causes an x-directed displacement $u(x, t)$ of the various elements of the rod from their equilibrium position. By considering the equation of motion of a thin section of the rod, derive Eq. (4.26) using the concepts of stress and strain.

4.6 Consider an elastic string (like a guitar string) of length L and mass m that is held taught with tension T. Show that the vertical displacement $u(x, t)$ of the string satisfies the following wave equation

$$\frac{\partial^2 u(x,t)}{\partial t^2} = \frac{T}{\mu} \frac{\partial^2 u(x,t)}{\partial x^2}. \tag{4.29}$$

5 Waves and wave equations

In the previous Section, we presented a simple derivation of a wave equation describing the oscillator displacements in a system of coupled oscillators. Of course, an immediate conclusion is that the system supports waves: a disturbance excited at one spatial position will propagate across the oscillator chain. In this Section, we discuss the basic solutions of the wave equation – first in one dimension and subsequently in three dimensions.

We can write Eq. (4.26) in a more general form as

$$\boxed{\frac{\partial^2 u(x,t)}{\partial t^2} = v^2 \frac{\partial^2 u(x,t)}{\partial x^2},} \tag{5.1}$$

where v is a free parameter whose meaning will be evident shortly[14]. This is the general form of a wave equation in one spatial dimension. The simplest solution of Eq. (5.1) is

$$u(x,t) = Ae^{i\omega t \pm ikx}, \tag{5.2}$$

where ω and k are constants related via the *dispersion relation*

$$\boxed{\omega = kv.} \tag{5.3}$$

It is straightforward to verify (e.g. via direct substitution) that Eq. (5.2) indeed solves the wave equation provided that ω and k are related via the above dispersion relation. (Note that also the complex conjugate of Eq. (5.2) is a solution.) It should be clear that ω is the angular frequency of the wave, describing the number of periods that pass a fixed point per second. The parameter k is known as the *wave number*, and can easily be shown to be related to the wavelength λ as

$$\boxed{k = \frac{2\pi}{\lambda}.} \tag{5.4}$$

As illustrated in Fig. 12, the solution given by Eq. (5.2) (or its physically relevant real part) describes a sinusoidal wave that propagates along the spatial x dimension as time t evolves. Consider the value[15] $u(x,t)$ of the wave at an arbitrary time t and position x. After an infinitesimal time increment $t \to t + \mathrm{d}t$, the wave assumes this same value but at some slightly different position $x \to x + \mathrm{d}x$. Equating $u(x,t) = u(x + \mathrm{d}x, t + \mathrm{d}t)$ yields

$$\omega t \pm kx = \omega(t + \mathrm{d}t) \pm k(x + \mathrm{d}x) \tag{5.5}$$

$$0 = \omega \mathrm{d}t \pm k\mathrm{d}x. \tag{5.6}$$

From the second expression, we obtain

$$\boxed{\frac{\mathrm{d}x}{\mathrm{d}t} = \mp\frac{\omega}{k} = \mp v.} \tag{5.7}$$

The parameter v therefore corresponds to the speed of the wave, whilst the sign in the exponential of Eq. (5.2) determines the direction of wave propagation.

[14] I'm sure you already now the meaning...

[15] or displacement or whatever the wave describes

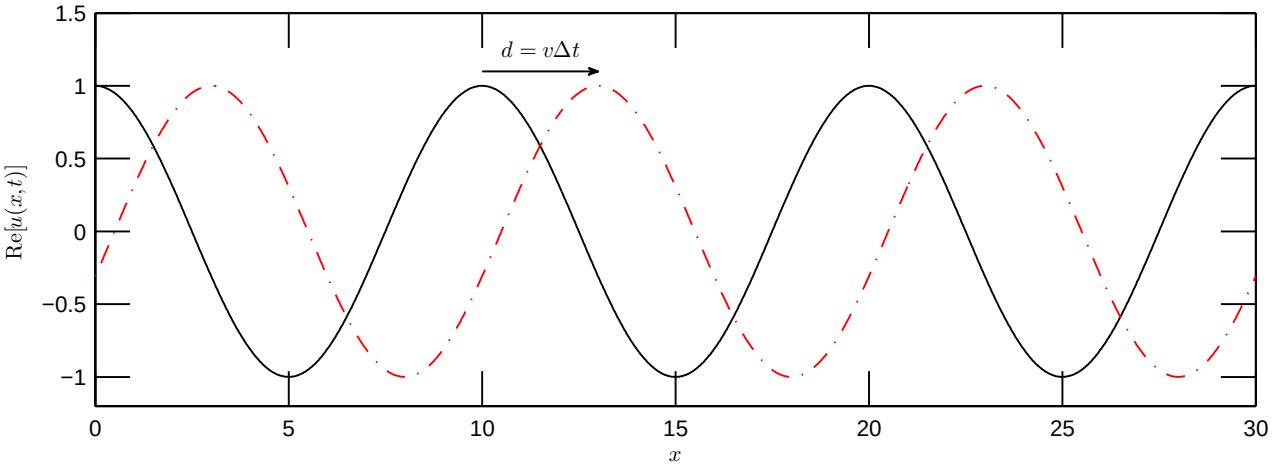

Figure 12: Schematic illustration of a sinusoidal wave propagating with speed v. In time Δt, the wave advances the spatial distance $d = v\Delta t$.

5.1 General solution in one dimension

The wave equation is a linear equation, which implies that superpositions of elementary solutions also solve the equation. In particular, we may write a general solution as an integral of elementary waves of the form given by Eq. (5.2) with different frequencies and amplitudes[16]:

$$u(x,t) = \frac{1}{2\pi}\left[\int_{-\infty}^{\infty} \tilde{A}(\omega) e^{i(\omega t + k(\omega)x)}\, d\omega + \int_{-\infty}^{\infty} \tilde{B}(\omega) e^{i(\omega t - k(\omega)x)}\, d\omega\right] \tag{5.8}$$

Note how the general solution comprises of two linearly independent integrals corresponding to two different signs in the exponential. Substituting $k(\omega) = \omega/v$ from the dispersion relation yields

$$u(x,t) = \frac{1}{2\pi}\left[\int_{-\infty}^{\infty} \tilde{A}(\omega) e^{i\omega(t + x/v)}\, d\omega + \int_{-\infty}^{\infty} \tilde{B}(\omega) e^{i\omega(t - x/v)}\, d\omega\right]. \tag{5.9}$$

Introducing new variables $t' = t + x/v$ and $t'' = t - x/v$, it is easy to see that the two terms in Eq. (5.9) have the general form of an inverse Fourier transform as defined in Eq. (2.4). Accordingly, we may write the general solution to the wave equation in the following time-domain form

$$\boxed{u(x,t) = A(t + x/v) + B(t - x/v).} \tag{5.10}$$

Physically, the two terms in Eq. (5.10) represent disturbances with arbitrary profiles that are moving with uniform speed v along the $-x$ $[A(t + x/v)]$ or $+x$ $[B(t - x/v)]$ directions. **In other words, any function that is uniformly moving with a speed of v solves the wave equation.**

[16] the factor of 2π will be clear shortly

5.2 Wave equation in three dimensions

The wave equation of Eq. (5.1) only includes a single spatial dimension (x), and therefore only describes waves in systems where wave propagation is possible along one single direction (such as a guitar string). This is of course not particularly general, as there are numerous systems[17] where waves can propagate in arbitrary directions [see exercises]. Moreover, Eq. (5.1) corresponds to a *scalar* wave equation, as the wave displacement is aligned along a single direction y; in general, wave displacements can point in arbitrary directions. We can write a more general form of the wave equation by permitting the directions of wave propagation and oscillation to point in any direction in a three-dimensional Euclidean space:

$$\frac{\partial^2 \vec{\mathbf{u}}(\vec{\mathbf{r}},t)}{\partial t^2} = v^2 \nabla^2 \vec{\mathbf{u}}(\vec{\mathbf{r}},t). \tag{5.11}$$

Here, $\vec{\mathbf{u}} = [u_x, u_y, u_z]^{\mathrm{T}}$ is a vector that describes the direction and magnitude of the wave displacement at some time t and position $\vec{\mathbf{r}} = [x, y, z]^{\mathrm{T}}$, while ∇^2 is the Laplace operator defined in Cartesian coordinates as

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{5.12}$$

We must note that, as for our derivation of the one-dimensional wave equation, the three-dimensional wave equation arises from the laws of physics that govern the particular system under study. It will be left as an exercise to derive the wave equation in different contexts.

The solutions of Eq. (5.11) are vectorial generalizations of the solutions of the one-dimensional wave equation. In particular, the simplest solution assumes the form

$$\vec{\mathbf{u}}(\vec{\mathbf{r}},t) = \vec{\mathbf{u}}_0 e^{i(\omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}})}, \tag{5.13}$$

where $\vec{\mathbf{u}}_0 = [u_{0x}, u_{0y}, u_{0z}]^{\mathrm{T}}$ is a constant *vector* amplitude of the wave, while $\vec{\mathbf{k}} = [k_x, k_y, k_z]^{\mathrm{T}}$ is known as the *wave vector*. The wave vector describes both the direction of wave propagation as well as the spatial periodicity (i.e. wavelength) of the wave. In particular, it can be shown that the wave propagates in the direction defined by $\vec{\mathbf{k}}$, and that the magnitude of the wave vector corresponds to the wave number

$$k = ||\vec{\mathbf{k}}|| = \sqrt{k_x^2 + k_y^2 + k_z^2} = \frac{2\pi}{\lambda}. \tag{5.14}$$

It can also be shown via direct substitution that, for Eq. (5.13) to be a solution to the three-dimensional wave equation, the dispersion relation described by Eq. (5.3) must still hold [but with the wave number defined by Eq. (5.14)]. In this regard, we must highlight that, for a given angular frequency ω, the 3D wave equation admits infinitely many solutions of the form Eq. (5.13), each associated with a *different* wave vector $\vec{\mathbf{k}}$; indeed, the only condition the wave vector must satisfy is that its magnitude fulfills the dispersion relation.

[17] Consider for example electromagnetic waves propagating through space or seismic waves propagating through the Earth.

Plane waves and polarization

Waves of the form given by Eq. (5.13) are known as *plane waves*. This is because the wave function $\vec{\mathbf{u}}(\vec{\mathbf{r}}, t)$ takes the same value at all points $\vec{\mathbf{r}}$ that lie on a two-dimensional plane perpendicular to $\vec{\mathbf{k}}$. Moreover, the planes of equal wave value repeat periodically in space. The period — i.e., the distance over which the wave's shape repeats itself along the direction of the wave vector $\vec{\mathbf{k}}$ — is simply the wavelength λ. We can thus picture the wave as being made out of infinite planes of constant value that are moving in the direction of their surface normal (which corresponds to the wave vector). These features can be readily demonstrated by considering Eq. (5.13), and is left as an exercise.

The vector amplitude $\vec{\mathbf{u}}_0$ describes the amplitude and direction of the wave oscillation. The direction of oscillation, which strictly speaking corresponds to the unit vector $\hat{\mathbf{u}}_0 = \vec{\mathbf{u}}_0/||\vec{\mathbf{u}}_0||$, is often referred to as the *polarization* of the wave.

5.3 Helmholtz equation

In the preceding subsection, we have described the wave equation in three spatial dimensions as well as its simplest solutions: the plane wave. It is important to note that plane waves do not really exist in real world. Indeed, a wave whose "planes of constant value" extend to infinity must carry infinite energy, which is unphysical. Nevertheless, it turns out that plane waves are useful: they offer very simple insights into the behaviour of real waves, and arbitrary waves can in fact be expressed as superpositions of plane waves (via Fourier integrals; more of that later).

In the next subsection, we discuss another important solution to the three-dimensional wave equation, namely that of a spherical wave. They do not really exist in real world either, but they get closer to reality than plane waves do. Before we describe the plane wave solutions, we shall introduce a time-independent form of the wave equation, namely the Helmholtz equation.

The wave equation describes how a given disturbance evolves both in space and time. Often, we are, however, interested in the spatial behaviour of waves that exhibit simple sinusoidal oscillations. Considering a scalar wave equation for the sake of simplicity, we may write the wave as

$$u(\vec{\mathbf{r}}, t) = \psi(\vec{\mathbf{r}})e^{i\omega t}. \tag{5.15}$$

Substituting this expression into the wave equation yields the *Helmholtz* equation:

$$\boxed{\nabla^2\psi(\vec{\mathbf{r}}) + k^2\psi(\vec{\mathbf{r}}) = 0.} \tag{5.16}$$

One should note the analogy between the Helmholtz equation and the time-independent Schrödinger equation.

The applicability of the Helmholtz equation is not limited to simple sinusoidal time-dependencies. It can also be derived by a more general separation of variables viz. $u(\vec{\mathbf{r}}, t) = \psi(\vec{\mathbf{r}})s(t)$, and moreover, it arises when one takes the Fourier transform (against the time variable) on both sides of the wave equation [exercise]. It should therefore be clear that the Helmholtz equation governs the spatial-dependency of all the different spectral components $\tilde{u}(\vec{\mathbf{r}}, \omega)$ that make up a given wave. This means that the full spatio-temporal wave behaviour can be investigated by using the Helmholtz equation to analyze the spatial evolution of different spectral components,

and by subsequently taking the inverse Fourier transform to reconstruct the time-domain waveform. In what follows, we assume however a simple sinusoidal wave.

5.4 Spherical waves

It is easy to see that our plane wave with the spatial dependency $\exp(-i\vec{\mathbf{k}} \cdot \vec{\mathbf{r}})$ satisfies the Helmholtz equation. To derive the spherical wave solutions, we first transform the Helmholtz equation into spherical coordinates:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial \psi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial \psi}{\partial \theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\psi}{\partial \phi^2} + k^2\psi = 0. \tag{5.17}$$

The equation can be simplified considerably by assuming the wave to be spherically symmetric, i.e., $\psi(r,\theta,\phi) = \psi(r)$. This of course a fairly strong assumption, but one that is not wholly unrealistic. Indeed, waves generated by point sources are (almost) spherical; consider for example a water wave generated by a fallen rock on a pond.

Under the assumption of spherical symmetry, Eq. (5.17) reduces to

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial \psi}{\partial r}\right) + k^2\psi = 0, \tag{5.18}$$

$$\frac{\partial^2\psi}{\partial r^2} + \frac{2}{r}\frac{\partial \psi}{\partial r} + k^2\psi = 0, \tag{5.19}$$

$$\frac{1}{r}\frac{\partial^2(r\psi)}{\partial r^2} + k^2\psi = 0. \tag{5.20}$$

Multiplying by r and introducing the transformation $g(r) = r\psi(r)$ allows us to write

$$\frac{\partial^2 g}{\partial r^2} + k^2 g = 0. \tag{5.21}$$

This corresponds to the standard Helmholtz equation in one dimension, and the solutions will hence be plane waves: $u(r) = \psi_0 \exp(\pm ikr)$. Accordingly, a solution to our original spherically symmetric Helmholtz equation, described by our original function $\psi(r) = u(r)/r$, reads

$$\boxed{\psi(r) = \frac{\psi_0}{r}e^{\pm ikr}.} \tag{5.22}$$

This is the (spatial-dependence) of the simple spherical wave solution to the 3D wave equation [time-dependence is simple harmonic, as separated above].

It is easy to see that all points the same distance r away from the origin exhibit the same value for the wave $\psi(r)$. Accordingly, the points of equal wave value in space form spherical surfaces (hence the name). It is also easy to see that the separation between two different surfaces associated with the same wave value is equal to the wavelength λ (or an integer multiple thereof). Including the time-dependence $\exp(i\omega t)$ in the solution, one can finally see that the the surfaces of constant value are moving radially inwards or outwards depending on the algebraic sign in the exponential in Eq. (5.22), as shown in Fig. 13.

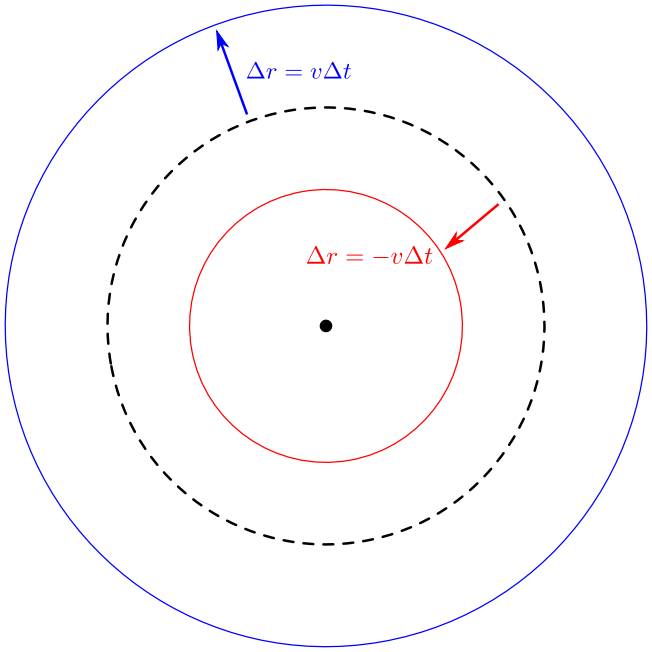

Figure 13: The wavefronts (points along the wave with constant phase) of a spherical wave move radially inwards or outwards depending on the combination of algebraic signs in the exponential.

Beyond spherical symmetry

Spherical waves correspond to spherically symmetrical solutions of the Helmholtz equation. Solutions that are not spherically symmetric can be found by using the method of separation of variables. Specifically, one assumes the solution to be of the form $\psi(r, \theta, \phi) = R(r)\Theta(\theta)\Phi(\phi)$, and substitutes this into Eq. (5.17). Separating equations for the three variables results in three differential equations whose solutions correspond to *special* functions, namely so-called spherical harmonics and Bessel functions. The entire derivation is a good exercise and is left as such.

5.5 Normal modes revisited

In the preceding subsections, we have considered solutions to the wave equation in different dimensions and coordinate systems. We have found that the wave equation can admit infinitely many different solutions with different frequencies and corresponding spatial behaviours. However, we know from past experience that many physical systems — such as guitar strings, drum membranes, laser resonators, or coupled oscillators — only admit waves that oscillate (steadily) at certain specific frequencies. In fact, these specific frequencies (and associated spatial oscillations patterns) correspond to the *normal modes* that were briefly introduced in section 4.2. In this subsection, we will take a deeper look into normal modes and in particular analyze the normal modes for a variety of systems from the vantage of the Helmholtz equation. It is important to emphasize that the wave equation is only half the story, and alone cannot describe normal modes. To capture this behaviour, we must additionally consider the *boundary conditions* of the specific system under study: natural modes will be those

that satisfy both the wave equation and the boundary conditions.

Throughout this subsection, we assume the wave to be monochromatic and scalar such that the motion of the system is described by the Helmholtz equation

$$\nabla^2\psi(\vec{\mathbf{r}}) + k^2\psi(\vec{\mathbf{r}}) = 0. \tag{5.23}$$

5.5.1 Normal modes of a string

Let us start from a very simple and familiar example, namely that of a string that is described by the one-dimensional ($\nabla^2 \to \partial^2/\partial x^2$) wave Eq. (5.1). Assume the length of the string to be L, and that the endpoints of the string are fixed, such that $\psi(0) = \psi(L) = 0$. It should be immediately clear that, while a simple exponential $\psi = \exp(ikx)$ solves the wave equation, it does not satisfy the boundary conditions. In fact, the simplest expression that can satisfy both is given by

$$\psi(x) = \frac{\psi_0}{2i}\left[e^{ikx} - e^{-ikx}\right] = \psi_0 \sin(kx), \tag{5.24}$$

with the boundary conditions demanding that the wavenumber satisfies

$$\boxed{k = \frac{n\pi}{L},} \tag{5.25}$$

where n is an integer. We thus see that only certain specific wave numbers are "allowed" which implies (through the dispersion relation) that the string can only oscillate at certain specific frequencies:

$$\boxed{\omega_n = 2\pi f_n = \frac{n\pi v}{L}.} \tag{5.26}$$

These frequencies [and corresponding spatial patterns $\sin(kx)$] represent the normal modes of the string, i.e., the natural oscillation patterns of the system. Figure 14 shows selected examples of different spatial oscillation patterns.

The above result should not be particularly surprising. Indeed you should have encountered *standing wave* patterns in your earlier studies — whether in the context of string instruments or laser resonators — and should therefore already be cognizant of the fact that a one-dimensional wave-supporting systems can only oscillate at specific frequencies. In the context of musical instruments, only oscillations with small integer values of n are typically excited; the oscillation of the lowest frequency [$n = 1$ in Eq. (5.26)] is known as the *fundamental mode*, while oscillations with $n > 1$ are known as *overtones*. When a string is struck, both the fundamental mode as well as a number of overtones are excited, and it is the combination of the different frequency components that gives the instrument its characteristic sound. It is important note that the spacing of allowed frequencies is constant: $\Delta\omega = \omega_{n+1} - \omega_n = \pi v/L$. This implies that the frequencies of the overtones are integer multiples of the frequency of the fundamental mode. In the context of optical (laser) cavities (where the oscillations correspond to electromagnetic waves), the spacing between allowed "resonance" frequencies is known as the *free-spectral range* of the cavity.

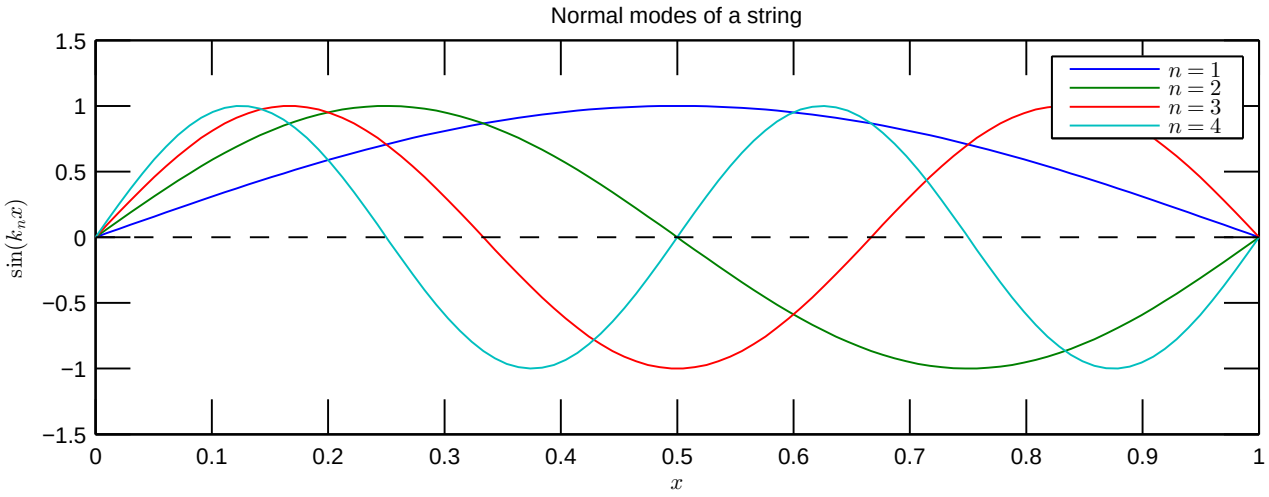

Figure 14: Examples of normal modes for a string with length $L = 1$ m. When including the time-dependence $\exp(i\omega t)$ and taking the real part, we observe that the patterns oscillate uniformly.

5.5.2 Normal modes of a drum

Let us now consider a two-dimensional system: a drum membrane held taught with some tension (e.g. a drum). Such a system can be shown to obey a two-dimensional wave equation [exercice]. Here we consider a circular membrane with radius R, and we assume that the membrane is held fixed at the edge. In cylindrical coordinates, $\psi(r = R, \phi) = 0$, where $\psi(r, \phi)$ describes the displacement of the drum membrane in the z-direction. (Note that the displacement does not depend on the coordinate z; it is rather directed along that direction.) In cylindrical coordinates (with $\partial\psi/\partial z = 0$), the Helmholtz equation reads

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\psi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2\psi}{\partial\phi^2} + k^2\psi = 0. \tag{5.27}$$

We solve the equation by using the method of separation of variables. Specifically, we write $\psi(r, \phi) = F(r)G(\phi)$ and substitute this expression into Eq. (5.27) above. This yields

$$\frac{r}{F(r)}\frac{\partial}{\partial r}\left(r\frac{\partial F(r)}{\partial r}\right) + k^2r^2 = -\frac{1}{G(\phi)}\frac{\partial^2 G(\phi)}{\partial\phi^2} \tag{5.28}$$

The left-hand side of the equation depends only on r, while the right-hand side depends only on ϕ. This implies that they must both be equal to a constant. Indeed, consider changing r while keeping ϕ contant: if the left-hand side should change, then the equality would no longer hold (since the right-hand side would not change). The only way for the equality to hold for all r and ϕ is for both sides of the equation to be equal to a constant. Assuming this constant is equal to m^2, we have

$$\frac{r}{F(r)}\frac{\partial}{\partial r}\left(r\frac{\partial F(r)}{\partial r}\right) + k^2r^2 = m^2, \tag{5.29}$$

$$-\frac{1}{G(\phi)}\frac{\partial^2 G(\phi)}{\partial\phi^2} = m^2. \tag{5.30}$$

The second equation can be rearranged to yield

$$\frac{\partial^2 G(\phi)}{\partial \phi^2} = -m^2 G(\phi). \tag{5.31}$$

The solution is trivial:

$$\boxed{G(\phi) = e^{\pm im\phi}.} \tag{5.32}$$

Because of the symmetry of the system, the constant m cannot take arbitrary values. Indeed, the point (r, ϕ) is physically equal to the point $(r, \phi+2\pi)$, such that $\psi(r, \phi) = \psi(r, \phi+2\pi)$. This implies that $G(\phi) = G(\phi+2\pi)$, which is possible only if m is an integer.

We have found that the angular part of the solution reads $G(\phi) = \exp(\pm im\phi)$, where m is an integer. Let us now look at the radial part. We first rearrange Eq. (5.29) into

$$r^2 \frac{\partial^2 F}{\partial r^2} + r \frac{\partial F}{\partial r} + \left[k^2 r^2 - m^2\right] F = 0. \tag{5.33}$$

This equation has the form of the *Bessel equation* — an important differential equation that arises in many different physical contexts[18]. Its solution can be written in terms of so-called Bessel functions

$$F(r) = AJ_m(rk) + BY_m(rk), \tag{5.34}$$

where A and B are integration constants and J_m and Y_m are Bessel functions of the first and the second kind, respectively. Figure 15 shows selected examples of $J_m(kr)$ and $Y_m(kr)$. As can be seen, $Y_m(kr)$ diverges at

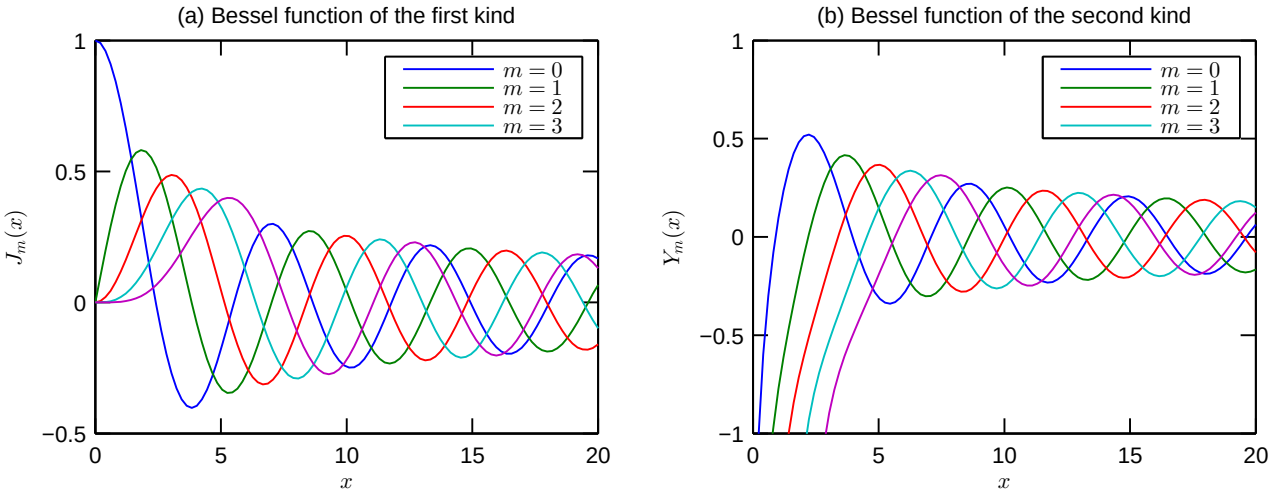

Figure 15: Bessel functions of the (a) first (J_m) and (b) second (Y_m) kind for orders $m = 0, 1, 2, 3$ as indicated. Note that the functions are symmetric with $x = 0$ and that the Bessel function of the second kind diverges at $x = 0$.

[18] Such as e.g. electromagnetic wave propagation in optical fibres, heat conduction in a cylindrical object, solutions to the radial Schrödinger equation for a free particle etc.

$r = 0$, and a physical solution therefore requires $B = 0$. Moreover, boundary conditions require that the wave number must be such that $J_m(kR) = 0$. There are no known analytical formulae for the zeros of the Bessel functions, but they can be found from numerical tables or from common mathematical software (e.g. Matlab or Scipy of Python). There are infinitely many of them for each value of the angular order m, and we can therefore label the allowed wavenumbers as $k_n^{(m)}$ with n an integer that identifies the order of the zero of the Bessel function ($n = 1$ corresponds to the smallest root).

Putting everything together, the normal modes of a drum can be written as

$$\psi_{nm}(r, \phi) = J_m \left(k_n^{(m)} r \right) e^{im\phi}, \tag{5.35}$$

where m is an integer and the wavenumber satisfies

$$J_m \left(k_n^{(m)} R \right) = 0. \tag{5.36}$$

It is important to note that, to obtain the physical oscillation profiles, we must take the real part $\mathrm{Re}[\psi_{nm}(r, \phi)]$. Figure 16 shows examples of the two-dimensional spatial profiles of a drum with $R = 30$ cm.

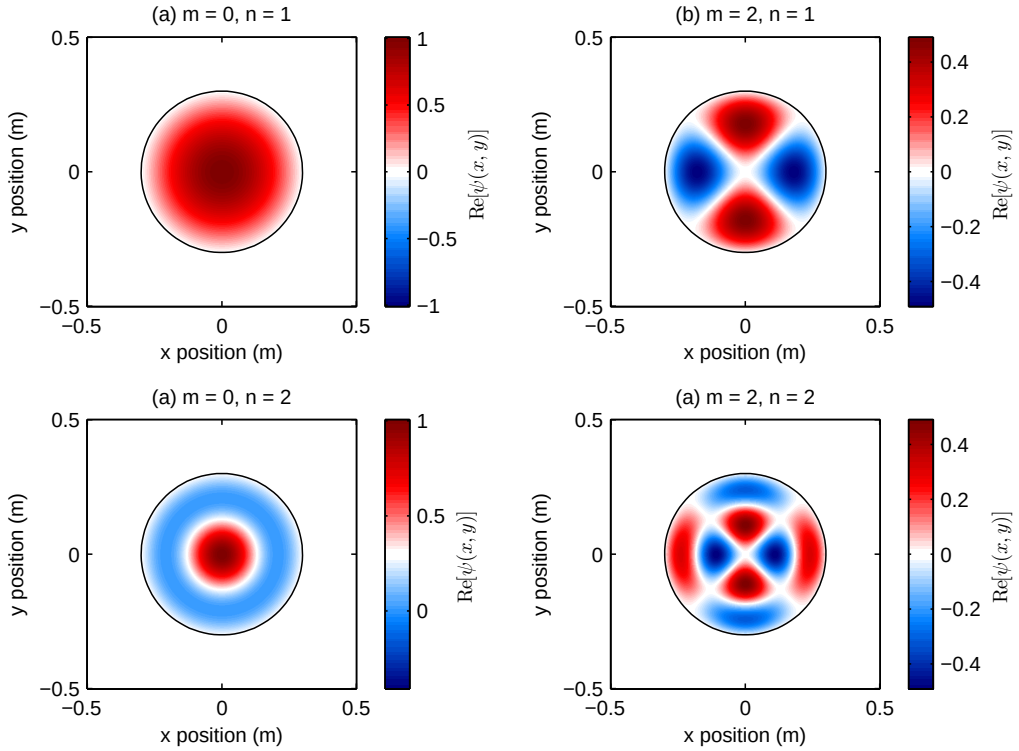

Figure 16: Examples of different spatial profiles corresponding to different natural modes of a drum with $R = 30$ cm (circumference highlighted with black solid curves). All the different parts of the drum would oscillate at the frequency $\omega_n^{(m)}$ in time, and so the minima would periodically turn to maxima and vice versa.

It is interesting note that the eigenfrequencies ($\omega_n^{(m)} = k_n^{(m)} v$) of a drum are not equally spaced, which is in stark contrast with the case of a string. This is simply because the zeros of the Bessel function are not equally spaced. Since the allowed wave numbers $k_n^{(m)} = \alpha_n^{(m)}/R$, the spacing between two eigenfrequencies $\Delta\omega = \omega_{n+1}^{(m)} - \omega_n^{(m)} = v/R(\alpha_{n+1}^{(m)} - \alpha_n^{(m)})$ will depend on the index n. This implies then that the overtones of a drum are *not* integer harmonics of the fundamental mode ($n = 1$).

5.5.3 Normal modes of a sphere

As a final example, we consider the normal modes of a spherical surface with radius R. This example is relevant to a particularly important physical system: a planet such as the Earth. The normal modes of a planet would be excited e.g. at the onset of an earthquake, and they will describe how the planet is deformed as the seismic waves traverse through it. Here we only consider waves that propagate along the surface (i.e., surface waves), and leave wave motion in the *interior* of the sphere (i.e., body waves) as an exercise.

Our starting point is the spherically symmetric Helmholtz Eq. (5.17). Since we are interested in surface waves, we set $\partial\psi/\partial r = 0$ to obtain:

$$\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2} + k^2R^2\psi = 0. \tag{5.37}$$

We again use separation of variables to solve the equation, and specifically assume $\psi(\theta,\phi) = F(\theta)G(\phi)$. Substituting this expression into the equation above yields after wee rearrangement

$$\frac{\sin\theta}{F(\theta)}\frac{d}{d\theta}\left(\sin\theta\frac{dF(\theta)}{d\theta}\right) + k^2R^2\sin^2\theta = -\frac{1}{G(\phi)}\frac{d^2G(\phi)}{d\phi^2}. \tag{5.38}$$

The left-hand side depends exclusively on θ while the right-hand side depends exclusively on ϕ; hence they must be equal to the same constant m^2. The right-hand side is the same as Eq. (5.31), and so we have

$$G(\phi) = e^{im\phi}. \tag{5.39}$$

The equation for $F(\theta)$ can be written as

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{dF(\theta)}{d\theta}\right) + \left(k^2R^2 - \frac{m^2}{\sin^2\theta}\right)F(\theta) = 0. \tag{5.40}$$

Introducing a new variable $x = \cos\theta$, this equation can be written as

$$\boxed{\frac{d}{dx}\left((1-x^2)\frac{dF(x)}{dx}\right) + \left(k^2R^2 - \frac{m^2}{1-x^2}\right)F(x) = 0.} \tag{5.41}$$

This equation has the form of the so-called *associated Legendre differential equation*. It has a non-finite solution only when (i) $k^2R^2 = l(l+1)$, where l is a positive integer, and when (ii) $|m| \leq l$. **Note that these conditions determine the allowed wavenumbers, and hence the allowed oscillation frequencies.** There can again be an infinite amount of different natural modes associated with different wavenumbers and frequencies.

The solutions to Eq. (5.41) are known as the *associated Legendre polynomial*, and they are denoted $P_l^m(x)$. With this in mind, we can write the solution to Eq. (5.40) as

$$\boxed{F(\theta) = P_l^m(\cos\theta).} \tag{5.42}$$

And finally, the full normal modes of a spherical surface read

$$\boxed{\psi(\theta,\phi) = P_l^m(\cos\theta)e^{im\phi} = Y_{lm}(\theta,\phi).} \tag{5.43}$$

Here we introduced a set of new functions $Y_{lm}(\theta,\phi)$ which are known as the *spherical harmonics*.

Spherical harmonics

The spherical harmonics are special functions defined on the surface of a sphere. They form a complet set of orthogonal functions on the sphere, and can therefore be used to represent functions on a circle; just like we use Fourier transforms to represent functions as a superposition of different frequency components. They arise in many physical problems ranging from the computation of atomic electron configurations to thr representation of gravitational and magnetic fields of a planetary body.

5.6 Dispersive wave equations

The wave equations we have discussed up to now [e.g. Eqs. (5.1) and (5.11)] are *non-dispersive*. That is to say, the wave velocity v is a constant that does not depend on the wavelength or frequency. But of course, common wisdom tells us that this is not true in general; EM waves for example propagate in media at the velocity $v = c/n$, where the refractive index $n \equiv n(\omega)$ depends on frequency[19].

Effects arising from a frequency-dependent wave velocity can be captured by using a generalised form of the wave equation. Restricting our attention to the one-dimensional scalar case for simplicity, the generalisation can be written as

$$\boxed{\frac{\partial^2 u(x,t)}{\partial x^2} = f(t) * \frac{\partial^2 u(x,t)}{\partial t^2},} \tag{5.44}$$

where $f(t)$ is a time-domain response function whose physical significance will be elaborated shortly, and $*$ denotes convolution. To illustrate how convolution with the response $f(t)$ gives rise to a frequency-dependent wave velocity, we substitute the plane wave ansatz $u(x,t) = \exp(i\omega t - ikx)$ into Eq. (5.44). This yields

$$-k^2 u(x,t) = -\omega^2 \int_{-\infty}^{\infty} f(\tau) e^{i[\omega(t-\tau)-kx]}\,\mathrm{d}\tau \tag{5.45}$$

$$= -\omega^2 e^{i(\omega t - kx)} \int_{-\infty}^{\infty} f(\tau) e^{-i\omega\tau}\,\mathrm{d}\tau \tag{5.46}$$

$$= -\omega^2 \tilde{F}(\omega) u(x,t), \tag{5.47}$$

[19] How else would Newton have been able to separate white light into its constituent colours using a prism?

where $\tilde{F}(\omega)$ is the Fourier transform of the time-domain response function $f(t)$. From the above expression, we obtain the general dispersion relation

$$k(\omega) = \sqrt{\tilde{F}(\omega)}\omega, \tag{5.48}$$

and corresponding wave velocity

$$v = \frac{\omega}{k} = \frac{1}{\sqrt{\tilde{F}(\omega)}}. \tag{5.49}$$

And so we see that a convolution in the wave equation yields a frequency-dependent wave velocity. The fact that you may not have encountered such convolutions before, in the simple wave equations that you have considered, is simply because of unrealistic approximations made in the derivations of said equations. As a concrete example, when deriving the EM wave equation, one typically makes the approximation that bound electrons of a material react instantaneously to an applied electric field, which results in a simple wave equation that cannot predict or explain chromatic dispersion. In contrast, if one models the light-matter interaction more accurately, taking into account the fact that electrons do not react instantaneously, a wave equation with a convolution (and corresponding chromatic dispersion) arises. This issue is explored in Exercise 5.2.

We should stress that the convolution can also arise in the spatial domain. Specifically, another form of a generalised wave equation could read:

$$\frac{\partial^2 u(x,t)}{\partial x^2} = g(x) * \frac{\partial^2 u(x,t)}{\partial t^2}, \tag{5.50}$$

where $g(x)$ is a spatial response function that physically accounts for some non-local spatial coupling in the material[20], and the convolution should be evaluated with respect to x. Repeating the analysis above would now reveal the dispersion relation

$$\omega(k) = \frac{k}{\sqrt{\tilde{G}(k)}}. \tag{5.51}$$

It is quite common to find dispersion relations quoted in the latter form, i.e., frequency as a function of wavenumber. Of course, the wavenumber (wavelength) and the frequency are mathematically related, and so the two different approaches to writing the dispersion relation are ultimately equivalent.

5.6.1 Phase and group velocity

The wave velocity discussed above corresponds to the *phase velocity* of the wave. It is the velocity at which a given point (or phase) of a purely monochromatic, sinusoidal wave propagates. In contrast, in dispersive media,

[20] For example, think about a situation where coupled oscillators are affected by the motion of oscillators other then their nearest neighbours.

a wave packet that is made out of several different frequency components [see e.g. Fig. 1(b)] propagates at a different velocity, known as the *group velocity*. To see this, we write the wave packet as a superposition of harmonic waves:

$$u(x,t) = \int_{-\infty}^{\infty} \tilde{u}(\omega) e^{i(\omega t - k(\omega)x)} \, \mathrm{d}\omega, \tag{5.52}$$

where the dispersion relation is expressed as $k(\omega)$. We then assume that the spectrum of the wave packet is narrow and peaks around some frequency ω_0. We can then expand the wavenumber as a Taylor series and only retain the leading two terms:

$$k(\omega) \approx k_0 + k'(\omega - \omega_0), \tag{5.53}$$

where $k' = \mathrm{d}k/\mathrm{d}\omega|_{\omega_0}$. Substituting this expansion into Eq. (5.52) yields

$$u(x,t) = e^{i(k'\omega_0 x - k_0 x)} \int_{-\infty}^{\infty} \tilde{u}(\omega) e^{i\omega(t-k'x)} \, \mathrm{d}\omega, \tag{5.54}$$

The multiplicative factor in the front is a phase shift that does not affect the "envelope" of the wave packet. Focussing then on the integrand, consider an arbitrary point along the wave packet at $t = 0$, say x_0. At a later time t, that point shifts in space to a new position x so as to conserve the integrand, i.e.,

$$\int_{-\infty}^{\infty} \tilde{u}(\omega) e^{-i\omega k' x_0} \, \mathrm{d}\omega = \int_{-\infty}^{\infty} \tilde{u}(\omega) e^{i\omega(t-k'x)} \, \mathrm{d}\omega. \tag{5.55}$$

From this expression we can readily see that the arbitrary point we picked, and hence the entire wave packet, has shifted to $x = x_0 + t/k'$. The group velocity is therefore

$$\boxed{v_\mathrm{g} = \frac{x - x_0}{t} = \frac{1}{k'} = \frac{\mathrm{d}\omega}{\mathrm{d}k}.} \tag{5.56}$$

In a non-dispersive medium, we have $\omega = kv$, and so the group velocity is equal to the phase velocity, $v_\mathrm{g} = v$. In contrast, in the presence of dispersion, the two are clearly different. This should not of course be surprising; after all, the wave packet is made out of different frequency components, each of which propagates at a different group velocity.

It is interesting to note that the above analysis predicts the envelope of the wave packet to maintain constant shape: the integrand is constant except for a shift in space. However, this turns out to be strictly true only when the first-order Taylor series expansion of $k(\omega)$ is exact, i.e., $k = k_0 + k'(\omega - \omega_0)$. In the presence of higher-order dispersion terms, the envelope of the wave packet becomes distorted. For example, it can become wider or narrower or undergo various sorts of other changes.

Problems

5.1 Consider a homogeneous, two-dimensional membrane with constant mass per unit area. The membrane is stretched and then fixed along its entire boundary in the xy plane. (We're talking about a drum here.) The tension per unit length T caused by stretching the membrane is the same at all points and in all directions and does not depend on time. Show that the deflection $u(x, y, t)$ obeys a two-dimensional wave equation. You may assume the deflection is small compared to the size of the membrane and that all angles of inclination are small.

5.2 Maxwell's equations govern all (classical) phenomena related to electricity and magnetism. In a dielectric medium with no free currents or charges, the equations can be written as:

$$\nabla \cdot \vec{\mathbf{D}} = 0 \qquad \nabla \cdot \vec{\mathbf{B}} = 0$$

$$\nabla \times \vec{\mathbf{E}} = -\frac{\partial \vec{\mathbf{B}}}{\partial t} \qquad \nabla \times \vec{\mathbf{H}} = \frac{\partial \vec{\mathbf{D}}}{\partial t}.$$

Here, $\vec{\mathbf{D}}$ is the electric dispacement field, $\vec{\mathbf{B}}$ is the magnetic field, $\vec{\mathbf{E}}$ is the electric field, and $\vec{\mathbf{H}}$ is the magnetic displacement field. For a material that is non-magnetic, isotropic, and dielectric, we have $\vec{\mathbf{B}} = \mu_0 \vec{\mathbf{H}}$ (μ_0 is the vacuum permeability) and $\vec{\mathbf{D}} = \varepsilon_0 \vec{\mathbf{E}} + \vec{\mathbf{P}}$, where the material polarization $\vec{\mathbf{P}} = \varepsilon_0 \chi \vec{\mathbf{E}}$ with ε_0 the vacuum permittivity and χ a material parameter known as the electric susceptibility. Show that Maxwell's equations imply that the electric and magnetic fields satisfy a three-dimensional wave equation. Based on your derivation, answer the following questions.

(a) What is the speed of light in vacuum ($\chi = 0$)?

(b) How does the material refractive index depend on the parameter χ?

(c) The relationship $\vec{\mathbf{P}} = \varepsilon_0 \chi \vec{\mathbf{E}}$ arises from the assumption that the bound electrons of the material react instantaneously to the applied electric field[21]. A more realistic relationship reads

$$\vec{\mathbf{P}}(\vec{\mathbf{r}}, t) = \varepsilon_0 \int_{-\infty}^{t} \chi(t - \tau) \vec{\mathbf{E}}(\vec{\mathbf{r}}, \tau) \, \mathrm{d}\tau,$$

where $\chi(t)$ is a time-domain response function. How does the refractive index predicted by this, more realistic model of the dielectric polarization $\vec{\mathbf{P}}(\vec{\mathbf{r}}, t)$, compare with the one you found in part (b)?

5.3 Consider the "plane wave" solution to the 3D wave equation:

$$\vec{\mathbf{u}}(\vec{\mathbf{r}}, t) = \vec{\mathbf{u}}_0 e^{i(\omega t - \vec{\mathbf{k}} \cdot \vec{\mathbf{r}})}. \tag{5.57}$$

(a) Show that the wavefronts, i.e., the points $\vec{\mathbf{r}}$ where the wave attains the same value (up to a phase shift of 2π), of this solution correspond to two-dimensional planes that are perpendicular to the wave vector $\vec{\mathbf{k}}$.

(b) Show that the wavelength, i.e., the separation between two wavefronts with the same value, is given by $\lambda = 2\pi/k$, where the wave number $k = ||\vec{\mathbf{k}}||$.

[21] Note that the macroscopic polarization $\vec{\mathbf{P}}$ arises physically from the charge separation of electrons from the positive nuclei.

(c) Show that the wavefronts (i.e., planes of constant wave value) propagate at the speed v in the direction of the wave vector.

5.4 Write the Helmholtz equation in cylindrical coordinates, and derive expression for so-called *cylindrical waves*. Describe their salient characteristics.

5.5 Show that a general separation of variables applied on a three-dimensional, scalar wave equation results in the Helmholtz equation.

5.6 Derive the Helmholtz equation by taking the Fourier transform of the three-dimensional wave equation against the time variable.

5.7 Consider a radial wave with a spatial dependence given by Eq. (5.22). Show that the wavefronts are moving radially inwards or outwards depending on the algebraic sign $\exp[\pm ikr]$.

5.8 In the complex notation, the quantity $|u(\vec{\mathbf{r}}, t)|^2$ is proportional to the power transmitted by the wave per unit area. Show that the energy transmitted by a radial wave through a spherical surface of radius R is constant (and does not depend on R).

5.9 Repeat the derivations for the natural modes of a string, drum, and a spherical surface outlined in subsection 5.5.

5.10 Consider a 24-inch bass drum with a fundamental ($n = 1, m = 0$) resonance frequency of 74 Hz. What is the speed of waves on the drum membrane? What are the frequencies of the modes ($n = 1, m = 1$) and ($n = 2, m = 0$)?

5.11 Derive the normal modes for a solid sphere assuming the oscillation amplitude is zero at the surface of the sphere (boundary condition). These modes (and hence the derivation problem) are encountered in many different branches of physics. For example, they correspond to the natural oscillation patterns of planets, and are hence excited by earthquakes; seismic waves are found to predominantly oscillate at certain frequencies that coincide with the normal modes of the planet. Moreover, in optics, the normal modes of a sphere correspond to so-called whispering gallery modes that can be excited with coherent laser light in millimetre size dielectric spheres.

6 Fourier wave propagation

In the previous Section, we have discussed wave equations and their general solutions. If we know the full functional form $u(\vec{\mathbf{r}}, t)$ of a particular wave at all spatial points $\vec{\mathbf{r}}$ and times t, we know everything about the wave. In practice, we rarely have such complete knowledge a priori. We may rather know for example what the wave looks like at some spatial plane of observation at some given time, and from this information, we may need to deduce how the wave will propagate to some subsequent plane. Consider for example the diffraction of a wave at an obstacle. If we know the profile of the wave at the obstacle, how can we predict the wave profile at some distant observation plane? In this Section, we show that Fourier analysis provides an elegant solution to the conundrum. In the context of optics, the methods are generally referred to as "Fourier optics", and they play a key role in the design and analysis of optical systems.

Throughout this Section, we consider a scalar, *monochromatic*[22] wave $u(\vec{\mathbf{r}}, t) = \psi(\vec{\mathbf{r}}) \exp(-i\omega t)$ that satisfies the Helmholtz Eq. (5.16). Our goal is to deduce how a given wave, whose spatial profile $\psi(\vec{\mathbf{r}})$ is known at some initial plane, propagates in space. To achieve this, we need to find a suitable solution to the Helmholtz equation that satisfies our initial condition. The simplest solutions to the Helmholtz equation are harmonic functions of the form $\exp(i\vec{\mathbf{k}} \cdot \vec{\mathbf{r}})$. There are infinitely many solutions of this form, each associated with a different wave vector $\vec{\mathbf{k}}$: the only constraint is that the magnitude of the wave vector satisfies the dispersion relation, i.e., $k = ||\vec{\mathbf{k}}|| = \omega/v$. Because the Helmholtz equation is linear, superpositions of simple harmonic functions are also solutions. In fact, recalling that harmonic functions form a simple, orthogonal set of basis functions, we can express arbitrary solutions as superpositions (integrals) of harmonic functions.

Based on the discussion above, we may conclude that a general solution to the Helmholtz equation can be expressed as a triple integral over the different cartesian vector components of the wave vector $\vec{\mathbf{k}} = [k_x, k_y, k_z]^{\mathrm{T}}$, with the integration limits chosen such that that the dispersion relation holds. This can be expressed in a much simpler form by noting that the dispersion relation in fact stipulates that only two of the three components of the wave vector are independent. Arbitrarily choosing the independent components to be k_x and k_y, we have

$$\boxed{k_z = \pm\sqrt{k^2 - k_x^2 - k_y^2},} \tag{6.1}$$

where the $+$ and $-$ signs correspond to forwards and backwards propagating waves (relative to the $+z$ direction)[23], respectively. This observation allows us to write an arbitrary solution of Eq. (5.16) as an integral over different wave vector components k_x and k_y:

$$\boxed{\psi(x, y, z) = \frac{1}{(2\pi)^2} \iint_{-\infty}^{\infty} I(k_x, k_y) e^{i(k_x x + k_y y)} e^{\pm iz\sqrt{k^2 - k_x^2 - k_y^2}}\, dk_x\, dk_y,} \tag{6.2}$$

where the function $I(k_x, k_y)$ is known as the *plane wave spectrum* of the wave (for reasons that will soon be apparent). Note that the integration limits in Eq. (6.2) are allowed to extend to infinity, as Eq. (6.1) ensures that the dispersion relation holds.

[22] The propagation of waves that are not monochromatic can be constructed as a Fourier integral over monochromatic frequency components analysed in this section. Note also that we will be using the convention $\exp(i\vec{\mathbf{k}} \cdot \vec{\mathbf{r}} - i\omega t)$ to denote a plane wave; this convention is different (complex conjugate) of the convention used in earlier sections. While the physics remains the same, it turns out some of the problems are easier to analyse using this convention.

[23] Recall that the time dependence is $\exp(-i\omega t)$.

To facilitate interpretation, let us consider the (x, y) plane where $z = 0$. Here we have

$$\psi(x, y, z = 0) = \frac{1}{(2\pi)^2} \iint_{-\infty}^{\infty} I(k_x, k_y) e^{i(k_x x + k_y y)} \, dk_x \, dk_y. \tag{6.3}$$

We can recognise this relationship as a two-dimensional (inverse) Fourier transform, with k_x and k_y representing *spatial (angular) frequencies* that describe how often sinusoidal components of the wave repeat per unit distance. In analogy with the more conventional one-dimensional time-frequency Fourier transform, Eq. (6.3) shows that an arbitrary 2D *spatial profile* of $\psi(x, y, z = 0)$ can be obtained as an (infinite) sum of sinusoids with different *spatial* frequencies k_x and k_y [see Fig. 17]. The plane wave spectrum $I(k_x, k_y)$ describes the amplitudes and phases with which the different spatial sinusoids contribute to the original function. It can be

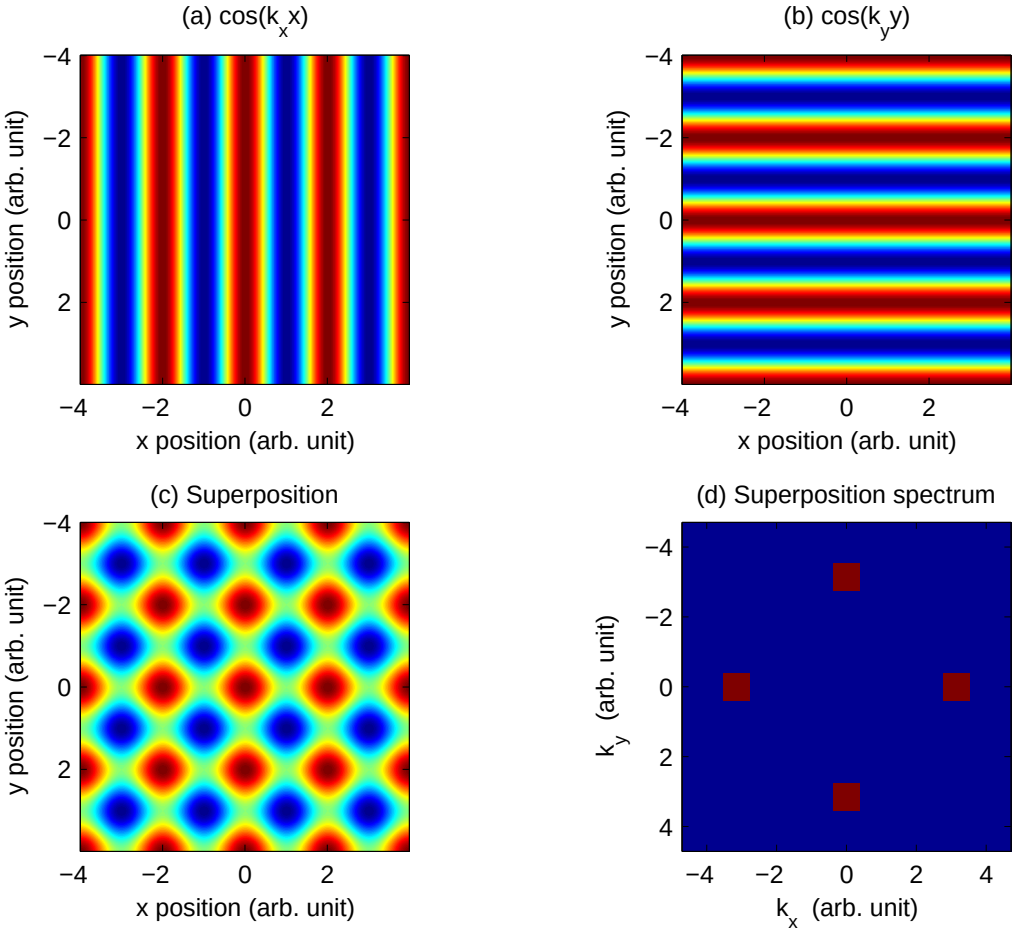

Figure 17: Illustration of plane wave decomposition of a 2D pattern. The superposition of co-sinusoidal oscillations in the (a) x and (b) y directions result in (c) a pattern that varies both along x and y. The spectrum of the superposition, obtained via 2D Fourier transform, shows four components as expected. In this example, the wavelengths of the patterns are equal to $\lambda_x = \lambda_y = 2$ arb. unit, such that the spectral components manifest themselves at $k_{x,y} = \pm 2\pi/\lambda_{x,y} = \pm\pi$.

obtained as the 2D Fourier transform of the original function $\psi(x, y, z = 0)$, defined as

$$I(k_x, k_y) = \iint_{-\infty}^{\infty} \psi(x, y, z = 0) e^{-i(k_x x + k_y y)}\, dx\, dy. \tag{6.4}$$

We leave it as an exercise to show that Eqs. (6.3) and (6.4) indeed form Fourier transform pairs.

Notes about spatial Fourier transform

Spatial frequencies and spatial Fourier transforms may seem somewhat obscure at first, but they are not all that scary. Consider a two-dimensional function $u(x, y) = \cos(k_x x)$. The function is clearly constant in the y direction and periodic in the x direction, with the periodicity $\lambda_x = 2\pi/k_x$. The number of oscillations the function exhibits in the x direction over some distance L is $N = L/\lambda_x$. Accordingly, in analogy with ordinary frequency, the spatial frequency of the oscillations in the x direction is $N/L = 1/\lambda_x$. This corresponds to the spatial frequency of the function, while $k_x = 2\pi/\lambda_x$ is the corresponding spatial angular frequency. As illustrated in Fig. 17, complex 2D functions can be expressed as a superpositions of harmonic functions with different spatial (angular) frequencies k_x and k_y.

6.1 Propagation and evanescent waves

The discussion above shows that, at $z = 0$, the wave $\psi(x, y, z = 0)$ can be described as a superposition of different harmonic waves with different transverse wave vectors k_x and k_y. On the other hand, we recall that the wave vector represents the direction of wave propagation, implying that the different harmonic components of the wave will travel in different directions as a function of z. Because the original wave is comprised of plane waves that are propagating in different directions, it should not be surprising that the transverse field profile of the wave itself changes as we move from the plane $z = 0$ to some other plane $z = d$.

Let us now consider the situation where we know the field profile $\psi(x, y, z = 0)$ at some *input plane* where $z = 0$ and we wish to know the field profile $\psi(x, y, z = d)$ at some *output plane* where $z = d$ [see Fig. 18]. The transformation is described by Eq. (6.2). Assuming forward-propagating waves, we can write this equation entirely in the spatial Fourier domain as

$$O(k_x, k_y) = I(k_x, k_y) e^{id\sqrt{k^2 - k_x^2 - k_y^2}} = I(k_x, k_y) H(k_x, k_y, d), \tag{6.5}$$

where $O(k_x, k_y) = \mathcal{F}_{2D}[\psi(x, y, z = d)]$ denotes the 2D Fourier transform of the field profile at the *output* plane where $z = d$ and we defined the *transfer* function

$$H(k_x, k_y, d) = e^{id\sqrt{k^2 - k_x^2 - k_y^2}}. \tag{6.6}$$

Equation (6.5) shows that the plane wave spectrum at the output plane is equal to the spectrum at the input plane multiplied with the transfer function $H(k_x, k_y, d)$. Physically, each plane wave component of the wave whose spatial frequencies satisfy $k_x^2 + k_y^2 < k^2$ acquires a pure phase shift as it propagates from the input plane to the output plane. At higher spatial frequencies $k_x^2 + k_y^2 > k^2$, the quantity under the square root in Eq. (6.6)

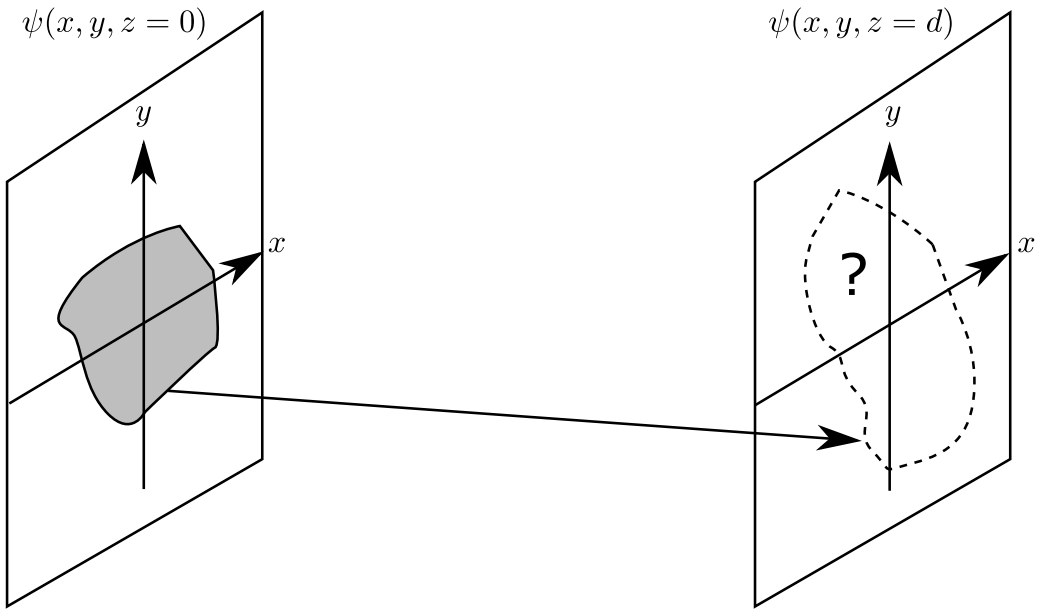

Figure 18: Schematic illustration of the transformation of the wave spatial profile $\psi(x, y, z)$ as it propagates from plane $x = 0$ to some other plane $z = d$.

is negative such that the exponential function is real and the transfer function represents an *attenuation* factor[24]. Such high-frequency components are known as *evanescent waves*, and they are unable to propagate into the far-field as they are exponentially attenuated.

In practical terms, the propagation of a wave whose input profile is known can be analysed by following the block diagram shown in Fig. 19. First take the 2D Fourier transform[25] of the input $\psi(x, y, z = 0)$ to obtain the plane wave spectrum $I(k_x, k_y)$; then multiply each spectral component of the plane wave spectrum with the transfer function $H(k_x, k_y, d)$ to obtain the spectrum $O(k_x, k_y)$ at plane $z = d$; finally take the inverse 2D Fourier transform of the product to obtain the output profile $\psi(x, y, z = d)$. As an example, let us consider an

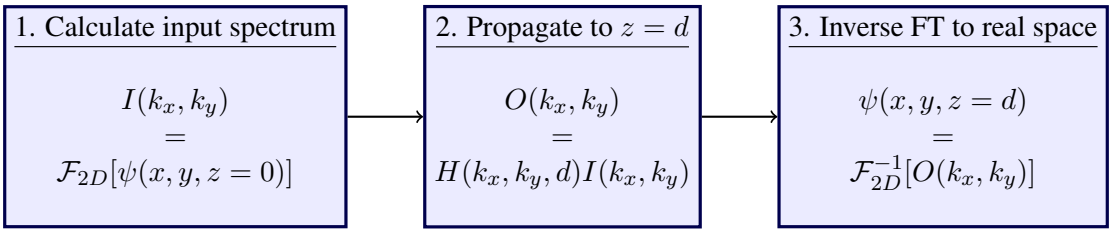

Figure 19: Block diagram illustrating the propagation of a wave whose initial profile is known at some plane $z = 0$.

[24]Note: had we used the convention $\exp(i\omega t - i\vec{\mathbf{k}} \cdot \vec{\mathbf{r}})$, we would have to choose different signs of the exponent for cases where $k_x^2 + k_y^2 < k^2$ and $k_x^2 + k_y^2 > k^2$, respectively. This is a little bit cumbersome, which is why we chose to go with the alternate convention.

[25]Software such as Python or Matlab have algorithms for calculating 2D Fourier transforms. For example, the numpy library for Python has the function fft2 for this purpose.

electromagnetic wave that has a Gaussian transverse profile at $z = 0$ and a wavelength of 800 nm propagating in free-space. Assuming the beam is propagating along the z-axis, we have

$$\psi(x, y, z = 0) = Ae^{-\frac{x^2 + y^2}{w^2}}, \tag{6.7}$$

where w describes the spot size of the beam. Figure 20(a) shows the input transverse profile of the wave whilst Fig. 20(b) shows the corresponding plane wave spectrum. Here the spot size was chosen to be sufficiently small such that the plane wave spectrum contains high-frequency components for which $k_x^2 + k_y^2 > k^2$ (see caption). Figures 20(c) and (d) show corresponding profiles at the output plane after propagation of just one wavelength ($d = \lambda$). We can see that the beam profile has grown significantly bigger. The reason for this is that the high-frequency components corresponding to evanescent waves are attenuated [c.f. Fig. 20(d)], which

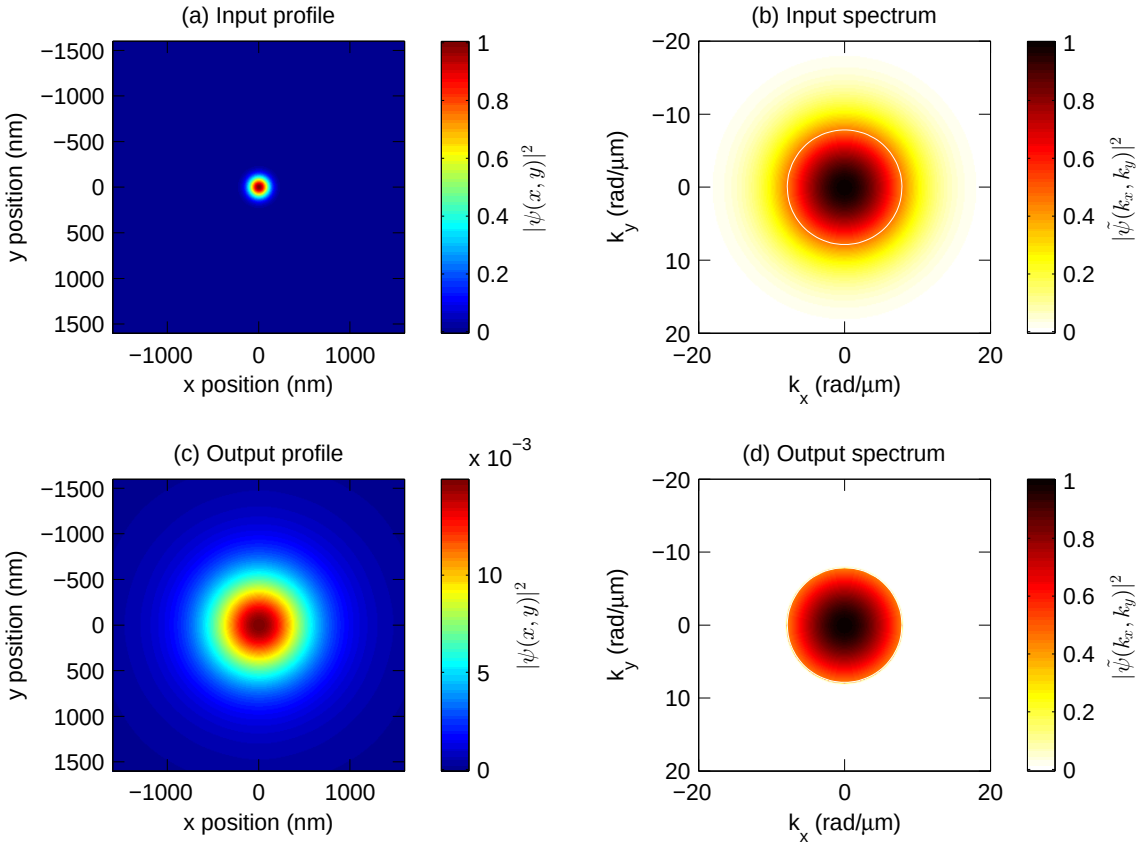

Figure 20: Example of the low-pass frequency filtering of free-space wave propagation. (a) Spatial profile of a Gaussian beam with width $w = \lambda/4$ where $\lambda = 800$ nm is the wavelength of the wave. (b) Plane wave spectrum corresponding to (a). (c) Spatial profile after the beam has propagated for one wavelength, i.e., at $z = \lambda$. (d) Plane wave spectrum corresponding to (c). The white circles in (b) and (d) have a radius of $k = 2\pi/\lambda \sim 8\ \mathrm{rad}/\mu\mathrm{m}$, highlighting how spectral components with $k_x^2 + k_y^2 > k^2$ are exponentially attenuated.

reduces the width of the plane wave spectrum and hence increases the profile size[26].

Even if the plane wave spectrum only contains low-frequency components ($k_x^2 + k_y^2 < k^2$), the beam profile will expand in size (albeit much more slowly). This is simply a manifestation of the well-known phenomenon of *diffraction*. Physically, we can picture the input beam profile as being made of plane waves that travel in different directions, hence resulting in expansion of the transverse profile. Mathematically, broadening arises from the different phase shifts imparted by the transfer function $H(k_x, k_y)$ on the different components of the spectrum. The narrower the input beam, the faster the broadening, as the plane wave spectrum contains a wider range of frequency components. An example is shown in Fig. 21; here, the initial beam width $w = 10\ \mu\mathrm{m}$, and we see that the beam doubles in size after roughly $d = 1.2\ \mathrm{mm}$ of propagation.

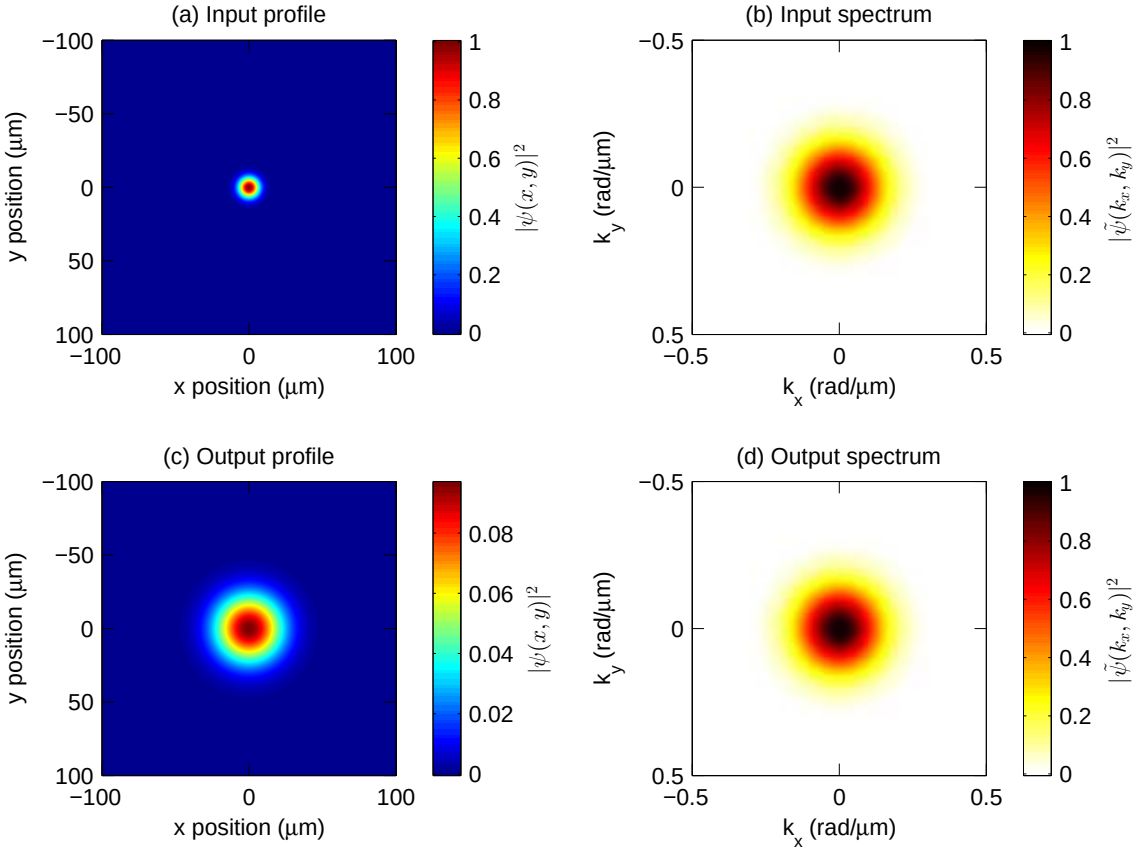

Figure 21: Example of diffraction of a Gaussian beam with width $w = 10\ \mu\mathrm{m}$ and wavelength $\lambda = 800\ \mathrm{nm}$. (a) Input spatial profile at $z = 0$ and (b) corresponding plane wave spectrum. (c) Output spatial profile at $z = 1.2\ \mathrm{mm}$ and (d) corresponding plane wave spectrum. Note that $k \sim 8\ \mathrm{rad}/\mu\mathrm{m}$, and so all the spectral components satisfy $k_x^2 + k_y^2 < k^2$.

[26]Recall from Heisenberg's uncertainty principle or Fourier bandwidth limits that narrow features require broad spectral content.

6.2 Paraxial (Fresnel) approximation

Wave propagation described by the full transfer function $H(k_x, k_y)$ given by Eq. (6.6) is not amenable to analytical treatment. The problem can, however, be simplified when the wave is predominantly travelling along a single direction. In this subsection, we develop a *paraxial* approximation for wave propagation.

Our starting point is the Helmholtz Eq. (5.16), and we assume that the wave is predominantly propagating in the z direction. We write the spatial dependence of the wave as the product of a *carrier* term $\exp(ikz)$ and an *envelope* function $f(x, y, z)$:

$$\boxed{\psi(x, y, z) = f(x, y, z)e^{ikz}.} \tag{6.8}$$

Substituting this expression into the Helmholtz equation yields

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} + 2ik\frac{\partial f}{\partial z} = 0. \tag{6.9}$$

We now make the so-called *slowly-varying envelope* approximation. Specifically, we assume that the envelope function $f(x, y, z)$ changes slowly with z. [Indeed, the main evolution along z has been factored out by the carrier term $\exp(ikz)$.] The meaning of "slow" change is that, over a distance of one wavelength, the envelope does not change much. In particular, we have:

$$|\delta f| = \left|\frac{\partial f}{\partial z}\right| \lambda \ll |f|, \tag{6.10}$$

which implies that

$$\left|\frac{\partial^2 f}{\partial z^2}\right| \ll \frac{1}{\lambda}\left|\frac{\partial f}{\partial z}\right| \sim 2k\left|\frac{\partial f}{\partial z}\right|. \tag{6.11}$$

Thus, $|\partial^2 f/\partial z^2| \ll 2k|\partial f/\partial z|$, allowing us to drop the third term in Eq. (6.9). This yields the so-called **paraxial wave equation**:

$$\boxed{\frac{\partial f}{\partial z} = \frac{i}{2k}\left[\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}\right].} \tag{6.12}$$

Analogies

Equations with forms similar to the paraxial wave equation [Eq. (6.12)] appear in numerous different contexts. For example, the heat equation, which describes the flow of heat through a material body, is written in three dimensions as

$$\frac{\partial u}{\partial t} = \alpha\left[\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right]. \tag{6.13}$$

This equation has the exact same form as Eq. (6.12), with the propagation direction z playing the role of time t of the heat equation (and the wave equation being 2D in space). Similarly, the diffusion equation, which describes the behavior of the collective motion of micro-particles in a material resulting from the random movement of each micro-particle, is equal to the heat equation (and hence the paraxial wave equation) when the diffusion coefficient is constant. Other examples include the propagation of light

pulses in optical fibres (relevant to e.g. telecommunications), time-dependent Schrödinger's equation, and even the Black-Scholes equation used to price financial instruments. Accordingly, knowing how to solve the paraxial wave equation can prove very handy!

6.2.1 Solving the paraxial wave equation

The paraxial wave equation can be easily solved using Fourier analysis[27]. Specifically, taking the 2D Fourier transform with respect to x and y of Eq. (6.12) we obtain

$$\frac{\partial \tilde{F}}{\partial z} = -\frac{i}{2k}\left[k_x^2 \tilde{F} + k_y^2 \tilde{F}\right], \tag{6.14}$$

where $\tilde{F} = \tilde{F}(k_x, k_y, z)$ is the Fourier transform of $f(x, y, z)$. The solution is trivial:

$$\tilde{F}(k_x, k_y, z) = \tilde{F}(k_x, k_y, 0) \exp\left[-\frac{i}{2k}\left(k_x^2 + k_y^2\right) z\right]. \tag{6.15}$$

We see that, as the wave propagates, different spatial frequency components accumulate a quadratic phase; the inverse Fourier transform then yields the profile $f(x, y, z)$ in real space. Recalling that the carrier-envelope decomposition in Eq. (6.8) also contains the $\exp(ikz)$ factor, we can write the total wave evolution in the Fourier domain as

$$\tilde{\psi}(k_x, k_y, z) = H_{\mathrm{p}}(k_z, k_y, z)\tilde{\psi}(k_x, k_y, z = 0), \tag{6.16}$$

where the paraxial transfer function

$$H_{\mathrm{p}}(k_x, k_y, z) = \exp[ikz] \exp\left[-\frac{i}{2k}\left(k_x^2 + k_y^2\right) z\right] \tag{6.17}$$

It is instructive to note that the transfer function in the paraxial approximation [Eq. (6.17)] can be obtained from the approximation-free transfer equation [Eq. (6.6)] with the approximation

$$k_x^2 + k_y^2 \ll k^2. \tag{6.18}$$

[27] Okay, we should be a little bit more precise here. In general, the solutions of the paraxial wave equation (and other analogous equations) depend on the initial and boundary conditions. The general method of solving these equations is to use separation of variables, which readily allows one to deal with common types of boundary conditions, such as e.g. Dirilecht (Neumann) conditions where the value (derivative) of the solution is fixed at the boundary. However, the approach is somewhat cumbersome, especially when we want to just quickly simulate a particular problem. When we use Fourier transform to solve the equations, we are essentially assuming the boundaries to be at infinity, and since the solution must be localized for the Fourier transform to exist, we do not need to enforce any additional boundary conditions. A peculiar situation arises, however, when the Fourier transforms are calculated on a computer using DFT. In this case, the computational domain is always finite, and so the boundaries cannot be at infinity. It turns out that the DFT automatically enforces *periodic* boundary conditions, in line with the assumption of periodicity made when deriving DFT. This means that, in computer simulations based on DFT, when a portion of a wave exits from one edge, it will reappear from the opposite edge.

This observation shows that the z component of the wave vector is the dominant, such that paraxial approximation indeed describes waves that are paraxial, i.e., waves that are propagation *almost* parallel to the z axis.

The paraxial wave equation admits analytical solutions in the form of waves with Gaussian transverse profiles[28] Such solutions describe very well for example laser beams, and they are of significant value in the design and analysis of optical systems. However, in general the evolution of an arbitrary transverse profile must be solved numerically even in the paraxial approximation (following a block diagram similar to that in Fig. 19 but with a different transfer function). A considerable simplification can be obtained in the "far field" limit, where z is very large. To develop this approximation, we explicitly write the slowly-varying transverse profile $f(x, y, z)$ as the inverse Fourier transform

$$f(x,y,z) = \frac{1}{(2\pi)^2} \iint_{-\infty}^{\infty} \tilde{F}(k_x, k_y, 0) e^{-\frac{i}{2k}(k_x^2+k_y^2)z} e^{i(k_x x + k_y y)}\, dk_x\, dk_y. \tag{6.19}$$

Combining the exponentials and completing squares we can rewrite

$$f(x,y,z) = \frac{e^{i\frac{k}{2z}(x^2+y^2)}}{(2\pi)^2} \iint_{-\infty}^{\infty} \tilde{F}(k_x, k_y, 0) e^{-\frac{i}{2k}z(k_x-\frac{k}{z}x)^2} e^{-\frac{i}{2k}z(k_y-\frac{k}{z}y)^2}\, dk_x\, dk_y. \tag{6.20}$$

In the far-field, we have $z \gg \lambda$, such that $z/k \gg 1$. In this case, the integrals can be evaluated using the *stationary phase approximation*. The basic idea is that only the points where the arguments of the exponentials are zero contribute to the integral. When the argument is not zero, the oscillations are very rapid (since $z/k \gg 1$), such that the sinusoids with different frequencies cancel out.

Since the arguments of the exponentials are zero at $k_x = kx/z$ and $k_y = ky/z$, we may write the result as

$$\boxed{f(x,y,z) \propto \tilde{F}\left(\frac{k}{z}x, \frac{k}{z}y, z=0\right).} \tag{6.21}$$

We have omitted the proportionality coefficient[29] to facilitate the interpretation. Specifically, Eq. (6.21) shows that the transverse wave profile in the far field is proportional to the 2D spatial Fourier transform of the profile at $z = 0$! In other words, **if a (paraxial) wave is allowed to propagate over a sufficiently large distance, its profile will shape into its Fourier transform.** When Eq. (6.21) is valid, we say we are operating in the *Fraunhofer* approximation.

Real time spectrometers

As mentioned before, propagation of light pulses in optical fibres obeys an equation analogous to the paraxial wave equation. This equation reads

$$\frac{\partial A}{\partial z} = -i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial t^2}. \tag{6.22}$$

[28] For further details, see lecture notes for "Physics 333 – Lasers and electromagnetic waves" by ME.

[29] It would be $-k/(2\pi z)\exp[-ik/2z(x^2+y^2)]$ or something along these lines. It should be highlighted that, for many classes of waves (such as light), the physically observable quantity is the absolute value squared of the wave: $|\psi(x,y,z)| = |f(x,y,z)|$. In this case, the complex prefactor disappears altogether, and the profile observable in the far field corresponds to the absolute value squared of the Fourier spectrum.

Here $A(z,t)$ is the slowly-varying envelope of the electric field, z is the propagation coordinate along the fibre, t is time expressed in a reference frame that is moving with the pulse (at the speed of light), and β_2 is known as the group-velocity dispersion coefficient. If the fibre length L is sufficiently large, we can reach the Fraunhofer limit where the temporal profile $A(t,L)$ mimics the spectrum at $z=0$:

$$A(t,L) \propto \tilde{A}\left(\omega = \frac{t}{\beta_2 L}, z=0\right). \tag{6.23}$$

This implies that we can measure the spectrum of a light pulse by propagating it through a long segment of fibre and recording the temporal pulse profile. This technique is known as the dispersive Fourier transform.

The strength of the dispersive Fourier transform is that it enables extremely fast measurements in real time. Conventional techniques (e.g. grating-based spectrometers) are typically quite slow, and they are unable to resolve the spectra of individual pulses emitted by ultrafast lasers. Such lasers can emit millions of pulses per second, and a conventional spectrometer will only be able to register their ensemble average spectral features. Thanks to the availability of high-speed oscilloscopes and photodetectors, it is straightforward to capture the individual pulses in the time domain however. Accordingly, by propagating the pulses through a long segment of optical fibre (such that their temporal profile shapes into their spectrum), it is possible to record the shot-to-shot spectra of an ultrafast pulse train.

6.3 Diffraction

Diffraction, which refers to a wide variety of phenomena that occur when a wave encounters an obstacle or a slit, is a fundamental property shared by all waves. In this subsection, we will see how the Fourier methods described above allow us to straightforwardly predict the diffracted profile of a wave in the far-field.

Consider a wave $\psi_{\mathrm{b}}(x,y,z)$ that encounters an obstacle at $z=0$ that has some finite aperture (i.e., region that transmits the wave). The simplest theory of diffraction is based on the assumption that the incident wave is transmitted without change at points within the aperture and not at all at points that are outside the aperture (and hence blocked by the obstacle). If $\psi_{\mathrm{b}}(x,y,z=0)$ and $\psi(x,y,z=0)$ represent the complex amplitudes of the wave right before and after the obstacle, respectively, we may write

$$\boxed{\psi(x,y,0) = \psi_{\mathrm{b}}(x,y,0)p(x,y),} \tag{6.24}$$

where $p(x,y)$ is called the *aperture* function. In accordance with our assumption above, we have

$$p(x,y) = \begin{cases} 1, & \text{inside the aperture} \\ 0, & \text{outside the aperture.} \end{cases} \tag{6.25}$$

With the transverse profile right after the obstacle known [$\psi(x,y,0)$], we can calculate the transverse profile in the far-field using the analyses presented above (assuming that the pertinent approximations hold). The diffraction pattern in the far-field is known as **Fraunhofer diffraction** or **Fresnel diffraction** depending on the

used approximations. Note that both approximations refer to "far-field", as both neglect the evanescent waves that would be present in the (really) *near-field*. Of course, based on our discussion, you should recognise that the Fraunhofer regime is "more far-field" than the Fresnel regime. Computation of Fresnel diffraction patterns generally requires numerics, while Fraunhofer diffraction offers a very straightforward analytical analysis.

6.3.1 Fraunhofer diffraction

Assume Fraunhofer approximation holds, such that the far-field transverse wave profile is given by Eq. (6.21). Moreover, we assume that the incident wave is a plane wave travelling in the z direction so that $\psi_{\mathrm{b}}(x, y, 0) = A$, where A is a constant. At $z = d$, the far-field diffraction pattern

$$\psi(x, y, d) \propto \tilde{P}\left(\frac{k}{d}x, \frac{k}{d}y\right), \tag{6.26}$$

where $\tilde{P}(k_x, k_y)$ is the Fourier transform of the aperture function $p(x, y)$. And so we see that **the far-field Fraunhofer diffraction pattern corresponds to the Fourier transform of the aperture function $p(x, y)$.**

As an example, let us consider a rectangular aperture with height D_y and width D_x. In this case,

$$p(x, y) = \begin{cases} 1, & |x| \leq D_x/2 \text{ and } |y| \leq D_y/2 \\ 0, & \text{elsewhere.} \end{cases} \tag{6.27}$$

The Fourier transform of the aperture function reads

$$\tilde{P}(k_x, k_y) = \int_{-D_x/2}^{D_x/2} \int_{-D_y/2}^{D_y/2} e^{-i(k_x x + k_y y)}\, dx\, dy \tag{6.28}$$

$$= D_x D_y \mathrm{sinc}\left(\frac{k_x D_x}{2}\right) \mathrm{sinc}\left(\frac{k_y D_y}{2}\right), \tag{6.29}$$

where the sinc-function is defined as $\mathrm{sinc}(x) = \sin(x)/x$. Finally using Eq. (6.26), the diffraction pattern is given by

$$f(x, y, d) \propto \mathrm{sinc}\left(\frac{\pi D_x}{\lambda d}x\right) \mathrm{sinc}\left(\frac{\pi D_y}{\lambda d}y\right). \tag{6.30}$$

Figure 22(a) shows an example of the far-field diffraction pattern for a 2 mm slit, observed 2 m from the aperture, as predicted based on Fraunhofer diffraction. Figure 22(b) shows the corresponding Fresnel diffraction pattern obtained by numerically solving the paraxial wave equation, and we see that the result is identical with Fig. 22(a).

6.4 Huygens-Fresnel principle

You have likely encountered the Huygens-Fresnel principle in your past studies. The principle states that every point on a wavefront is itself the source of spherical wavelets, and the sum of these spherical wavelets forms

the wavefront. It predicts that the waveform at a plane $z = d$ can be understood to arise as a superposition of all the spherical wavelets generated by a waveform at an earlier plane $z = 0$. How can we relate this principle with our analysis above?

Consider the frequency-domain description of wave propagation in the paraxial approximation [Eq. (6.16)]. Taking the inverse Fourier transformation, we can write this expression in real space as a convolution

$$\psi(x, y, z) = [\psi(x, y, 0) * h_{\mathrm{p}}(x, y, z)](x, y), \tag{6.31}$$

where $h_{\mathrm{p}}(x, y, z) = \mathcal{F}_{2D}^{-1}[H_{\mathrm{p}}(k_x, k_y, z)]$ is the impulse response of the system. By explicitly calculating the inverse Fourier transform [exercise], we obtain

$$h_{\mathrm{p}}(x, y, z) = \frac{1}{i\lambda z} e^{ikz} e^{i\frac{k}{2z}(x^2+y^2)}. \tag{6.32}$$

In the paraxial limit, we have $z^2 \gg x^2 + y^2$, allowing us to approximate [exercise]

$$h_{\mathrm{p}}(x, y, z) \approx \frac{1}{i\lambda z} e^{ik\sqrt{x^2+y^2+z^2}}, \tag{6.33}$$

Let us substitute this approximation in Eq. (6.31) and write the full form of the convolution:

$$\psi(x, y, z) \approx \iint_{-\infty}^{\infty} \psi(x_0, y_0, 0) \frac{1}{i\lambda z} e^{ikr} \mathrm{d}x_0 \mathrm{d}y_0, \tag{6.34}$$

where $r = \sqrt{(x - x_0)^2 + (y - y_0)^2 + z^2}$ represents the distance between points $(x_0, y_0, 0)$ and (x, y, z). By further approximating $z \approx r$ in the denominator, we realise that the term inside the integral is exactly identical

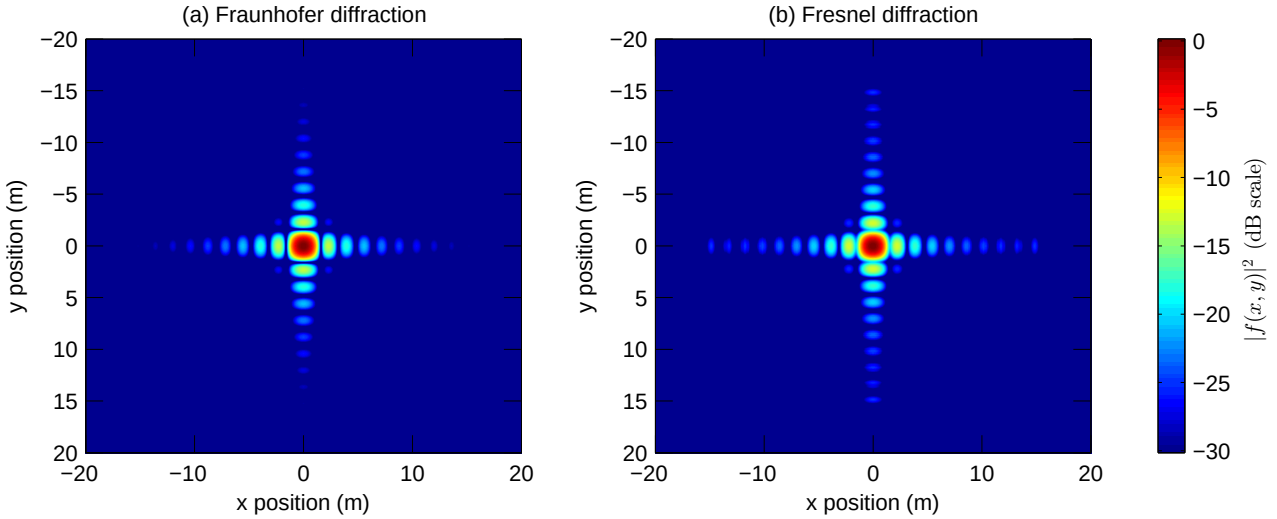

Figure 22: Diffraction pattern from a square-shaped aperture with 2 mm width and height, observed 2 m from the aperture. (a) and (b) show predictions from Fraunhofer and Fresnel theory, respectively. Note that the full wave theory (without paraxial approximation) yields a result identical to that in (b).

to the spherical wave derived in Eq. (5.22). Thus, Eq. (6.34) shows that the wave at an arbitrary plane z can be understood as a superposition of spherical waves generated at each point along the input wave profile at $z = 0$. This is the Huygens-Fresnel principle.

Note that our analysis above may feel a bit strange in that we are approximating the paraxial wave result (which in itself is an approximation) to get a form that resembles the Huygens-Fresnel principle. More rigorous analysis would require us to forego the paraxial approximation, and compute the inverse Fourier transform of the full transfer function given by Eq. (6.6). Does this yield an exact integral over spherical waves in accordance with Huygens-Fresnel? You tell me [exercise].

Problems

6.1 Consider a laser beam with wavelength $\lambda = 1064\ \mathrm{nm}$ that assumes a Gaussian profile [Eq. (6.7)] at some initial plane $z = 0$. An analytical solution based on the paraxial approximation predicts that the beam stays Gaussian as it propagates, but its width evolves as $w(z) = w_0\sqrt{1 + z^2/z_\mathrm{R}^2}$, where $w_0 = w(0)$ and $z_\mathrm{R} = w_0^2\pi/\lambda$. In particular, assuming unity amplitude (at $z = 0$), the paraxial approximation predicts that the beam profile satisfies

$$|\psi(x, y, z)| = \frac{w_0}{w(z)} e^{-\frac{x^2+y^2}{w(z)}}$$

for all z. Test this prediction by writing a Matlab / Python code that allows you to compute the transverse profile $\psi(x, y, z = d)$ without the use of the paraxial approximation. Considering the scenarios listed below, compare the outputs from your code and the analytical prediction to gauge the validity of the paraxial approximation.

(a) Initial width $w = 4.0\ \mu\mathrm{m}$, propagation distance $d = 2\ \mathrm{mm}$.

(b) Initial width $w = 1.0\ \mu\mathrm{m}$, propagation distance $d = 0.2\ \mathrm{mm}$.

(c) Initial width $w = 0.5\ \mu\mathrm{m}$, propagation distance $d = 0.02\ \mathrm{mm}$.

(d) Initial width $w = 0.1\ \mu\mathrm{m}$, propagation distance $d = 0.002\ \mathrm{mm}$.

Note that a good way to compare the profiles is to plot 2D slices e.g. along $y = 0$.

6.2 Consider the Gaussian transverse profile given by Eq. (6.7).

(a) Use our definition of the 2D Fourier transform [Eq. (6.4)] to show that

$$\mathcal{F}_\mathrm{2D}[\psi(x, y, z = 0)] \propto e^{-\frac{k_x^2 + k_y^2}{4} w^2}. \tag{6.35}$$

(b) Derive an estimate for the diffraction limit, i.e., the minimum spot size w that a beam can have while still propagating to the far field without significant attenuation.

6.3 By comparing the transfer functions given by Eq. (6.6) and Eq. (6.17), derive the following condition of validity for the paraxial (Fresnel) approximation:

$$w \left(\frac{8\pi}{d}\right)^{1/4} k^{3/4} \gg 1, \tag{6.36}$$

where w is the width of the wave profile in real space.

6.4 Consider the convolutional relationship given by Eq. (6.31). Assume $\psi(x, y, 0)$ is confined to a small circular area of radius b. Show that the condition

$$\frac{kb^2}{2z} \ll \pi \tag{6.37}$$

results in the Fraunhofer approximation.

6.5 Show that a rectangular slit gives rise to the far-field diffraction pattern given by Eq. (6.30).

6.6 Compute the Fraunhofer diffraction pattern from a circular aperture of diameter d illuminated by a plane wave along the z axis. Find where the first zero of the diffraction pattern occurs. You may find the following identities useful

$$J_0(z) = \frac{1}{\pi}\int_0^\pi \cos(z\cos(\theta))\,\mathrm{d}\theta$$

$$zJ_1(z) = \int_0^z tJ_0(t)\,\mathrm{d}t.$$

6.7 Repeat the derivation in subsection 6.4. In particular, show that the paraxial wave equation approximately predicts the same thing as the Huygens-Fresnel principle. If you are brave enough, also analyse the more general case that is not subject to the paraxial approximation (no analytical answer guaranteed).

7 Wave reflection and transmission at interfaces

So far, we have considered waves propagating in uniform, homogeneous media. In this Section, we consider what happens when a wave is incident on the interface between two different media. Referring to the wave equation, the only parameter that can be used to distinguish two materials from each other is the wave velocity v. Physically, a difference in wave velocity arises from differences in more fundamental material properties pertinent to the system under study, such as:

- refractive index n for electromagnetic waves,
- tension and mass for waves on a string,
- density and elasticity for seismic waves.

The question we wish to address is what happens when a wave meets the interface? We already know the answer based on our prior experience: some part of the wave will be transmitted and some part of the wave will be reflected. But how much is transmitted and how much is reflected? And in what directions if we have a 2D or 3D system?

The behaviour of waves at an interface cannot be analysed using the wave equation alone. The physics of the problem must rather provide additional information in the form of *interface conditions* that describe how the wave disturbance reacts to a discontinuous change in material properties. For example, in electromagnetism, one can show starting from Faraday's law that the tangential component of the electric field vector $\vec{\mathbf{E}}$ must be continuous across the interface; for simple waves propagating along a string, the displacement of the string must be continuous as otherwise the string would have to be broken.

Since the interface conditions are tied to the particular physical system under study, it is not straightforward to present a fully general analysis of wave behaviour at an interface. However, once the interface conditions are known, the analysis of the problem always follows (almost) identical ideas. In this section, we go through these ideas by examining selected physically relevant situations. We first analyse a simple 1D scalar system where a wave propagates through two different strings that are tied together. Subsequently, we consider a generic 3D system and show that the familiar laws of reflection and refraction emerge from an interface condition that enforces the wave to be continuous across the interface. We close this Section by briefly summarizing the main results relevant to a more complex vectorial 3D example, namely reflection and transmission of light (or electromagnetic waves) incident on a dielectric interface.

7.1 General idea

The salient question we wish to answer is: what happens when a given incident wave meets a boundary between two media with different wave velocities? Prior knowledge already tells us that some part will be reflected while the rest will be transmitted; we wish to dig deeper and deduce what exactly do the transmitted and reflected waves look like. The problem can be analysed as follows. We first separate the entire wave function into two parts on either side of the interface:

$$\vec{\mathbf{u}}(\vec{\mathbf{r}},t) = \begin{cases} \vec{\mathbf{u}}_{\mathrm{L}}(\vec{\mathbf{r}},t), & \vec{\mathbf{r}} \text{ on the one side of the interface} \\ \vec{\mathbf{u}}_{\mathrm{R}}(\vec{\mathbf{r}},t), & \vec{\mathbf{r}} \text{ on the other side of the interface.} \end{cases} \tag{7.1}$$

Here subscripts L and R refer to left and right, though in general the directions are of course arbitrary. The partial waves in the two different regions (L and R) must satisfy two different wave equations associated with

different velocities. Considering the case of a single wave incident on the interface from the L side, the wave in that region will consist of a superposition of the incident wave moving towards the interface and a reflected wave moving away from the interface:

$$\vec{\mathbf{u}}_{\mathrm{L}}(\vec{\mathbf{r}},t) = \vec{\mathbf{u}}_{\mathrm{i}}(\vec{\mathbf{r}},t) + \vec{\mathbf{u}}_{\mathrm{r}}(\vec{\mathbf{r}},t). \tag{7.2}$$

On the other hand, on the R side of the interface, we only have a single transmitted wave that is moving away from the boundary:

$$\vec{\mathbf{u}}_{\mathrm{R}}(\vec{\mathbf{r}},t) = \vec{\mathbf{u}}_{\mathrm{t}}(\vec{\mathbf{r}},t). \tag{7.3}$$

Let us assume that the incident wave function $\vec{\mathbf{u}}_{\mathrm{i}}(\vec{\mathbf{r}},t)$ is known. The reflected and transmitted waves represent two unknowns; to find them, we need **two** independent equations. These **two** equations must come from the interface conditions specific to the particular system under study. Once the interface conditions are known, we can solve for the reflected and transmitted wave functions, and then reconstruct the entire wave at all $\vec{\mathbf{r}}$. In the following subsection, we demonstrate these ideas by considering a simple example of wave propagation on a string.

7.2 1D scalar example: waves on a string

We begin with a brief recap of the derivation of the wave equation for a mechanical string. To this end, we consider a string with length L and mass m held taught with a tension T. We can imagine the string to be composed of small mass elements with mass $\mu\Delta x$, where $\mu = L/m$ is the linear mass density of the string, connected to each other with massless strings of length Δx [c.f. Fig. 23]. Neglecting gravity, the n^{th} mass element is subject to two forces with equal magnitude T but (possibly) different directions: one towards the $(n-1)^{\mathrm{th}}$ and another towards the $(n+1)^{\mathrm{th}}$ mass element. The equation of motion for the vertical displacement $u_n(t)$ of the n^{th} element reads [see Fig.23]:

$$\begin{aligned}
\mu\Delta x\frac{d^2u_n}{dt^2} &= F_{\mathrm{net,v}} && (7.4)\\
&= T\sin\theta_2 + T\sin\theta_1 && (7.5)\\
&= T\left[\frac{u_{n+1}-u_n}{\Delta x} + \frac{u_{n-1}-u_n}{\Delta x}\right] && (7.6)\\
&= T\Delta x\left[\frac{u_{n+1}-2u_n+u_{n-1}}{\Delta x^2}\right]. && (7.7)
\end{aligned}$$

In the continuum limit ($\Delta x \to 0$), the part in the brackets of the right-hand side of the final expression becomes a second-order partial derivative. And so we get

$$\boxed{\frac{\partial^2 u(x,t)}{\partial t^2} = \frac{T}{\mu}\frac{\partial^2 u(x,t)}{\partial x^2},} \tag{7.8}$$

which is of course the wave equation with the velocity $v = \sqrt{T/\mu}$.

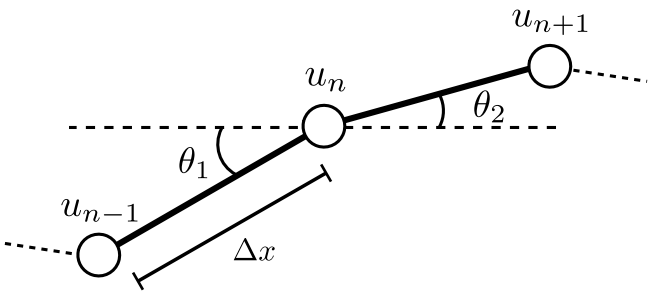

Figure 23: Schematic illustration of a discrete system that in the continuum limit becomes a continuous string.

7.2.1 Separation into regions

Let us now consider a string that is composed of two sections with varying tension T_1 and T_2 and correspondingly different wave velocities v_1 and v_2, respectively. As shown in Fig. 24, we assume the sections are connected at $x = 0$, with the $x < 0$ ($x > 0$) region associated with tension T_1 (T_2). As above, we write the vertical displacement of the string as $u_{\rm L}(x,t)$ left to the boundary and $u_{\rm R}(x,t)$ to the right of the boundary. The overall displacement, or wave function, is then given by

$$u(x,t) = \begin{cases} u_{\rm L}(x,t), & x < 0 \\ u_{\rm R}(x,t), & x \geq 0. \end{cases} \tag{7.9}$$

We can write this expression in a single equation by using the Heaviside step function $H(x)$:

$$\boxed{u(x,t) = u_{\rm L}(x,t)H(-x) + u_{\rm R}(x,t)H(x).} \tag{7.10}$$

The wave functions $u_{\rm L}(x,t)$ and $u_{\rm R}(x,t)$ satisfy the wave equation (7.8) with string tensions T_1 and T_2 (or equivalently, velocities v_1 and v_2), respectively. Assume the incident wave approaches the boundary from the left, i.e., from $x < 0$. The total wave in that region consists of a superposition of the incident wave moving towards the boundary and a reflected wave moving away from the boundary. Recalling the general solution to the 1D wave equation [see Eq. (5.10)], we may write

$$u_{\rm L}(x,t) = u_{\rm i}(t - x/v_1) + u_{\rm r}(t + x/v_1). \tag{7.11}$$

Here $u_{\rm i}(t - x/v_1)$ and $u_{\rm r}(t + x/v_1)$ correspond to the incident and reflected waves, respectively; they are written in the general form of arbitrary (for now) profiles moving to the right or left with velocity v_1. In the region $x > 0$, we only have a single transmitted wave that is propagating to the right at speed v_2:

$$u_{\rm R}(x,t) = u_{\rm t}(t - x/v_2). \tag{7.12}$$

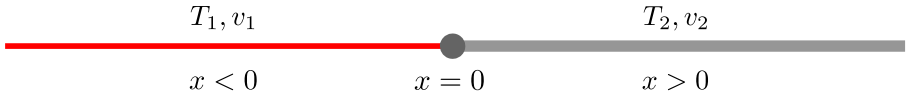

Figure 24: Schematic illustration of a composite system consisting of two strings held at different tensions T_1 and T_2. Because of the different tensions, the two different regions exhibit different wave velocities v_1 and v_2 as shown.

Note that the transmitted wave propagates with speed v_2, whilst the reflected and incident waves propagate at speed v_1. To find the reflected and transmitted wave functions $u_\mathrm{r}(t + x/v_2)$ and $u_\mathrm{t}(t - x/v_2)$, respectively, we need two independent equations. To find these equations, we must consider the interface conditions of the system.

7.2.2 Interface conditions

The first interface condition is easy to deduce: the wave must be continuous across the interface. Were this not the case, there would have to be a gap in the string and then no wave could propagate anyway. Since the interface is at $x = 0$, we have:

$$\boxed{u_\mathrm{i}(t) + u_\mathrm{r}(t) = u_\mathrm{t}(t).} \tag{7.13}$$

To obtain the second interface condition, we consider the mass element at the boundary with vertical displacement u_n. Since the element is attached to its neighbouring elements with differing tensions, the equation of motion reads

$$\mu \Delta x \frac{d^2 u_n}{dt^2} = T_2 \frac{u_{n+1} - u_n}{\Delta x} + T_1 \frac{u_{n-1} - u_n}{\Delta x}. \tag{7.14}$$

In the continuum limit, we can identify the terms on the right-hand-side of this expression to represent first order spatial derivatives [$df/dx \approx (f(x + h) - f(x))/h$]:

$$\mu \Delta x \frac{d^2 u}{dt^2} = T_2 \left.\frac{\partial u_\mathrm{R}}{\partial x}\right|_{x=0} - T_1 \left.\frac{\partial u_\mathrm{L}}{\partial x}\right|_{x=0}. \tag{7.15}$$

However, in the continuum limit, we also have $\Delta x \to 0$, which implies that the left-hand-side goes to zero! We can thus obtain our second interface condition by requiring that also the right-hand side is zero. Substituting the expressions for u_R and u_L, and evaluating the derivatives at the interface ($x = 0$) using the chain rule, we obtain

$$T_2 \left[-\frac{1}{v_2}\frac{du_\mathrm{t}}{dt}\right] = T_1 \left[-\frac{1}{v_1}\frac{du_\mathrm{i}}{dt} + \frac{1}{v_1}\frac{du_\mathrm{r}}{dt}\right] \tag{7.16}$$

$$\frac{d}{dt}\left[\frac{T_2}{v_2}u_\mathrm{t} + \frac{T_1}{v_1}u_\mathrm{r} - \frac{T_1}{v_1}u_\mathrm{i}\right] = 0. \tag{7.17}$$

By noting that the term inside the brackets must be constant for the derivative to be zero, we obtain our second interface condition [at $x = 0$]:

$$\boxed{\frac{T_2}{v_2}u_\mathrm{t}(t) + \text{cnst} = \frac{T_1}{v_1}\left[u_\mathrm{i}(t) - u_\mathrm{r}(t)\right].} \tag{7.18}$$

7.2.3 Reflection and transmission

We can set the integration constant in Eq. (7.18) to be zero without loss of generality, as this only represents an overall offset of the entire T_2 string segment. Then substituting Eq. (7.13) into Eq. (7.18), we obtain

$$\frac{T_2}{v_2}\left[u_\mathrm{i}(t) + u_\mathrm{r}(t)\right] = \frac{T_1}{v_1}\left[u_\mathrm{i}(t) - u_\mathrm{r}(t)\right]. \tag{7.19}$$

Solving for the reflected wave yields

$$u_{\mathrm{r}}(t) = \left(\frac{Z_1 - Z_2}{Z_1 + Z_2}\right) u_{\mathrm{i}}(t), \tag{7.20}$$

where $Z_n = T_n/v_n$ is known as the *impedance* of the string segment. For the transmitted wave, we have

$$u_{\mathrm{t}}(t) = \left(\frac{2Z_1}{Z_1 + Z_2}\right) u_{\mathrm{i}}(t). \tag{7.21}$$

The equations above show that the profile of the reflected and transmitted waves are identical to the incident wave, but their amplitudes are decreased. It is common to define *reflection* and *transmission* coefficients as

$$r = \frac{Z_1 - Z_2}{Z_1 + Z_2} \tag{7.22}$$

$$t = \frac{2Z_1}{Z_1 + Z_2}, \tag{7.23}$$

such that $u_{\mathrm{r}}(t) = r u_{\mathrm{i}}(t)$ and $u_{\mathrm{t}}(t) = t u_{\mathrm{i}}(t)$. It is noteworthy that the reflection coefficient can be either positive or negative depending on the impedances Z_1 and Z_2. The physical implication is that the reflected wave may exhibit a 180 degree phase shift upon reflection, as illustrated in Fig. 25.

With the reflected and transmitted wave functions known at $x = 0$, we can reconstruct the entire wave at all x and t using Eq. (7.10). In fact, this is precisely how Fig. 25 was constructed [for parameters, see figure caption]. It is worth noting that the reflection and transmission coefficients do not add up to unity, which may seem strange in light of energy conservation. However, we must emphasize that the reflection and transmission coefficients are only amplitude coefficients: careful analysis shows that the power propagating in a string is in fact proportional to the product of the string impedance and the oscillation amplitude squared. When all things are considered, energy is found to be conserved during reflection and transmission [exercise].

Impedance and impedance matching

Impedance is a concept that is used widely in different contexts of physics. In mechanical systems such as the one described above, impedance is a measure of how much a structure resists motion when subjected to a harmonic force. More specifically, mechanical impedance relates forces with velocities acting on a mechanical system as $Z(\omega) = F/v$. If a large force is required to induce a small velocity v, the impedance is large (and vice versa). In acoustics, impedance similarly describes the opposition that a system presents to the acoustic flow resulting from an acoustic pressure applied to the system. Impedance is also (more frequently) encountered in electronics, where it represent the opposition that a circuit presents to a current subject to time-varying voltage. It can therefore be understood as the generalization of resistance for alternating currents.

It is interesting to note that, regardless of the system, the reflection and transmission coefficients

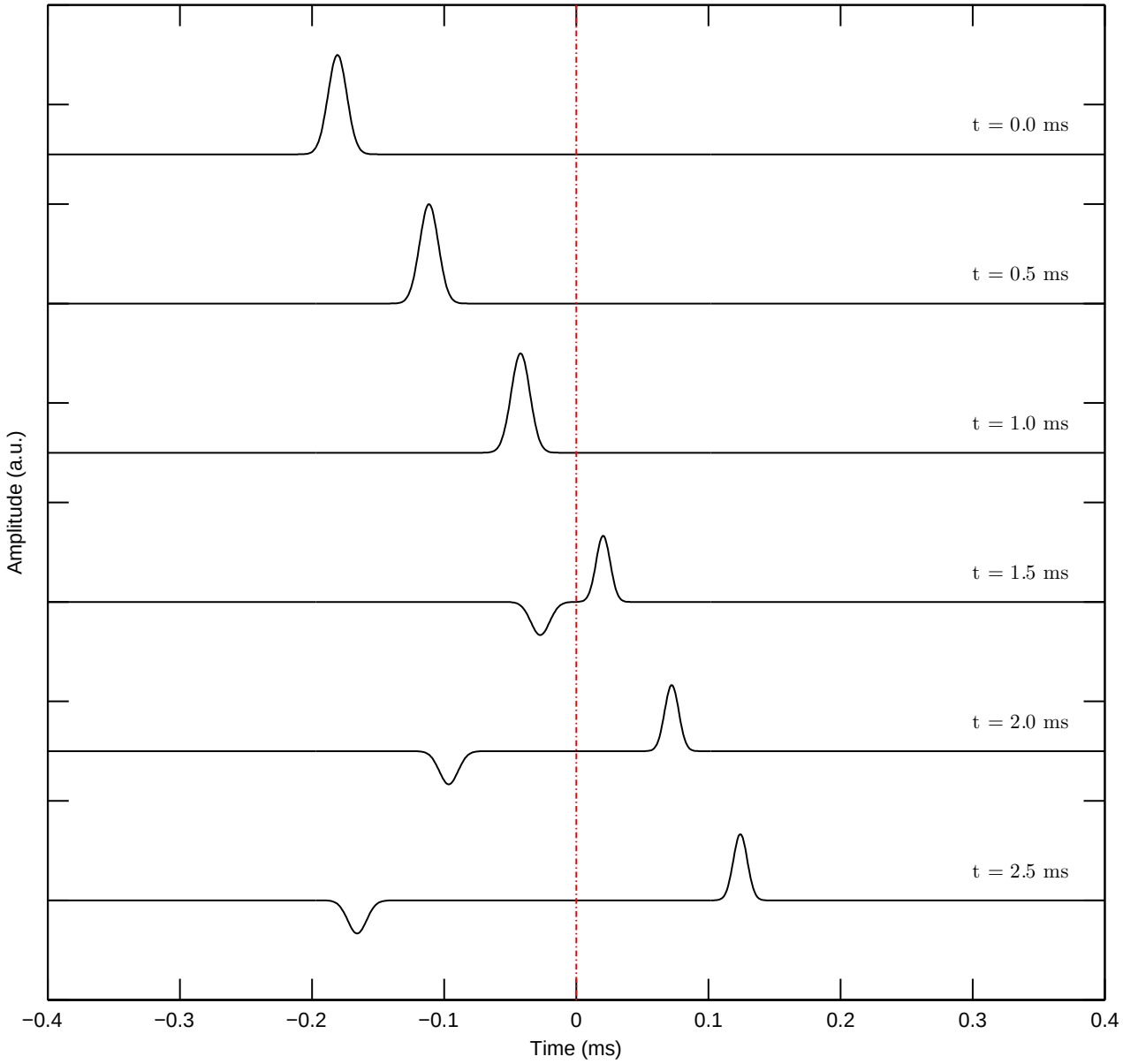

Figure 25: Reflection and transmission of a Gaussian pulse from the boundary between two strings held at different tension. The impedances $Z_1 = 0.5$ and $Z_2 = 1$ while the wave velocities $v_1 = 140\ \mathrm{m/s}$ and $v_2 = 100\ \mathrm{m/s}$

for a wave moving from medium 1 to medium 2 can very often be written as a function of the material impedances as in Eqs. (7.22) and (7.23) respectively. In addition to waves on a string, this is also the case for electrical transmission lines as well as light reflecting (and transmitting) from dielectric boundaries at normal incidence. An important observation is that the reflection coefficient is zero if the

two media have the same impedance. For the case of two strings held at different tension, this would require that the linear mass density of the strings are different. Tailoring the impedances to be identical so as to avoid reflections is known as impedance matching, and is widely used in many contexts, in electronics in particular.

7.3 2D and 3D: Law of reflection and refraction

The example above encompasses only a single spatial dimension x, and so there is no question about the directions of propagation of the waves. Let us now consider a more general 3D situation where the incident wave can strike the interface from an arbitrary direction, such that it makes an arbitrary angle with the surface that forms the interface. The question we wish to answer is: what are the directions of the reflected and transmitted waves? We consider a scalar wave $u(\vec{\mathbf{r}}, t)$ **and assume the wave to be continuous across the interface.** We must emphasize that the continuity of the wave is a very common interface condition.

Figure 26 shows a schematic illustration of the scenario under study. A wave $u_{\rm i}(\vec{\mathbf{r}}, t)$ is incident on an interface of two media associated with initial and final wave velocities v_1 and v_2, respectively. The incident wave will give rise to a reflected wave $u_{\rm r}(\vec{\mathbf{r}}, t)$ in the initial medium and a transmitted wave $u_{\rm t}(\vec{\mathbf{r}}, t)$ in the final medium. The continuity condition enforces that, at the interface, the wave amplitudes at both sides are equal. Since the incident region consists of two waves (incident and reflected), we have at all points $\vec{\mathbf{r}}$ on the interface

$$u_{\rm i}(\vec{\mathbf{r}}, t) + u_{\rm r}(\vec{\mathbf{r}}, t) = u_{\rm t}(\vec{\mathbf{r}}, t), \tag{7.24}$$

Waves on each side of the interface independently satisfy the respective wave equation. As we are interested in wave directions, we consider them to be plane waves whose direction is unambiguous. The equation above then reads

$$A_{\rm i} e^{i(\omega_{\rm i} t - \vec{\mathbf{k}}_{\rm i}\cdot\vec{\mathbf{r}})} + A_{\rm r} e^{i(\omega_{\rm r} t - \vec{\mathbf{k}}_{\rm r}\cdot\vec{\mathbf{r}})} = A_{\rm t} e^{i(\omega_{\rm t} t - \vec{\mathbf{k}}_{\rm t}\cdot\vec{\mathbf{r}})}, \tag{7.25}$$

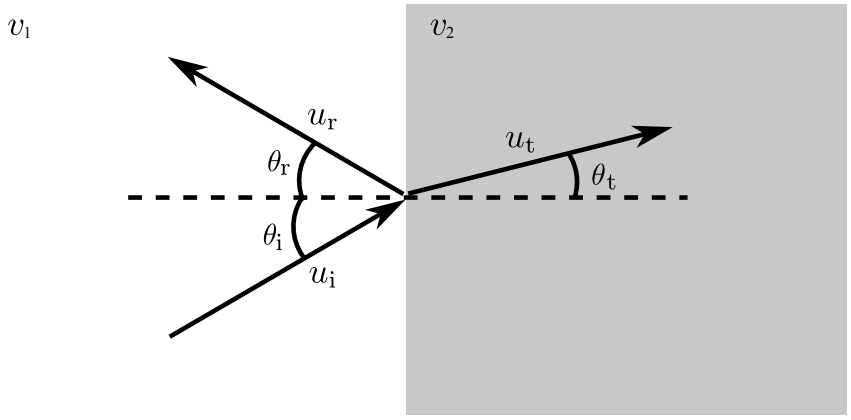

Figure 26: Schematic illustration of wave reflection and transmission at the interface between two media associated with different wave velocities v_1 and v_2. Note that, for simplicity, the figure assumes all the wave directions to lie on the same plane, which also contains the surface normal. As described in the main text, continuity of the field across the interface happens to instil this condition.

where ω_n and $\vec{\mathbf{k}}_n$ are the (angular) frequencies and wave vectors of the different waves. Equation (7.25) must hold at all times t and positions $\vec{\mathbf{r}}$ (on the interface). This implies that the time and space dependencies of the different terms are all equal[30]. In other words,

$$\boxed{\omega_\mathrm{i} t - \vec{\mathbf{k}}_\mathrm{i} \cdot \vec{\mathbf{r}} = \omega_\mathrm{r} t - \vec{\mathbf{k}}_\mathrm{r} \cdot \vec{\mathbf{r}} = \omega_\mathrm{t} t - \vec{\mathbf{k}}_\mathrm{t} \cdot \vec{\mathbf{r}}} \tag{7.26}$$

For the time part, the condition above is satisfied iff

$$\omega_\mathrm{i} = \omega_\mathrm{r} = \omega_\mathrm{t}. \tag{7.27}$$

In other words, the frequency of a wave is unchanged during reflection and transmission.

Looking now at the relationship between the incident and reflected wave vectors, we have

$$\vec{\mathbf{k}}_\mathrm{i} \cdot \vec{\mathbf{r}} = \vec{\mathbf{k}}_\mathrm{r} \cdot \vec{\mathbf{r}} \tag{7.28}$$

$$(\vec{\mathbf{k}}_\mathrm{i} - \vec{\mathbf{k}}_\mathrm{r}) \cdot \vec{\mathbf{r}} = 0. \tag{7.29}$$

Now consider two points $\vec{\mathbf{r}}_1$ and $\vec{\mathbf{r}}_2$ on the plane that represents the interface. Both points must satisfy the equation above, and so we have

$$(\vec{\mathbf{k}}_\mathrm{i} - \vec{\mathbf{k}}_\mathrm{r}) \cdot (\vec{\mathbf{r}}_1 - \vec{\mathbf{r}}_2) = 0. \tag{7.30}$$

Since the positions $\vec{\mathbf{r}}_1$ and $\vec{\mathbf{r}}_1$ are on the plane, the vector $(\vec{\mathbf{r}}_1 - \vec{\mathbf{r}}_2)$ is parallel to the plane and the null dot product shows that the vector $(\vec{\mathbf{k}}_\mathrm{i} - \vec{\mathbf{k}}_\mathrm{r})$ must be perpendicular to the plane. In other words, $(\vec{\mathbf{k}}_\mathrm{i} - \vec{\mathbf{k}}_\mathrm{r})$ is parallel to the surface normal $\hat{\mathbf{n}}$. Accordingly, we may write $\hat{\mathbf{n}} = a\vec{\mathbf{k}}_\mathrm{i} - a\vec{\mathbf{k}}_\mathrm{r}$, where a is a constant scalar. This form reveals that $\hat{\mathbf{n}}$ lies on a plane defined by the incident and reflected wave vectors. Extending the analysis to the transmitted wave vector allows us to conclude that **all the wave vectors and the surface normal lie on the same plane, the so-called plane of incidence which contains the surface normal and the wave vector of the incident wave [see Fig. 26].** The only question then is: what are the angles of the wave vectors relative to the surface normal.

To obtain the angle of reflection, we recall that the cross product of two parallel vectors is zero, so we can write

$$\hat{\mathbf{n}} \times (\vec{\mathbf{k}}_\mathrm{i} - \vec{\mathbf{k}}_\mathrm{r}) = 0 \tag{7.31}$$

$$\hat{\mathbf{n}} \times \vec{\mathbf{k}}_\mathrm{i} = \hat{\mathbf{n}} \times \vec{\mathbf{k}}_\mathrm{r} \tag{7.32}$$

$$k_\mathrm{i} \sin\theta_\mathrm{i} = k_\mathrm{r} \sin\theta_\mathrm{r}, \tag{7.33}$$

where $k_\mathrm{i} = ||\vec{\mathbf{k}}_\mathrm{i}||$ and $k_\mathrm{r} = ||\vec{\mathbf{k}}_\mathrm{r}||$ are the incident and reflected wave numbers, respectively, and θ_i and θ_r are the angles that the wave vectors make with respect to the surface normal. Since both the incident and the reflected waves propagate in the medium where the wave velocity is equal to v_1, and since the frequencies of the waves are identical, we have $k_\mathrm{i} = k_\mathrm{r} = \omega/v_1$. Thus, we obtain the **law of reflection**:

$$\boxed{\theta_\mathrm{i} = \theta_\mathrm{r}.} \tag{7.34}$$

[30] Otherwise a change in $\vec{\mathbf{r}}$ or t would change the terms unequally, which would destroy the equality.

To put the result in words: **the angle that the incident wave makes with the surface normal is equal to the angle that the reflected wave makes with the surface normal.** No surprises there.

We can repeat the steps above to find the relationship between the incident and the transmitted waves, obtaining

$$k_\mathrm{i} \sin\theta_\mathrm{i} = k_\mathrm{t} \sin\theta_\mathrm{t}. \tag{7.35}$$

In contrast to the case of reflection, the incident and transmitted wave numbers are not the same, as the wave velocities are different in the two different media. Substituting the wave numbers $k_\mathrm{i} = \omega/v_1$ and $k_\mathrm{t} = \omega/v_2$ we obtain **the law of refraction**:

$$\boxed{\frac{1}{v_1}\sin\theta_\mathrm{i} = \frac{1}{v_2}\sin\theta_\mathrm{t}.} \tag{7.36}$$

It is easy to see that, if $v_2 > v_1$, we have $\sin\theta_\mathrm{t} > \sin\theta_\mathrm{i}$, implying $\theta_\mathrm{t} > \theta_\mathrm{i}$ (and vice versa). In words, a wave transmitted into a medium with larger wave speed will be refracted away from the surface normal (and vice versa).

Equation (7.36) should be very familiar: it corresponds to the well-known Snell's law. It is applicable in all systems where the wave (or one of its vector component) is continuous across the interface. In the context of electromagnetic waves (e.g. light), the wave speed in a medium is given by $v = c/n$, where c is the speed of light in vacuum and n is the refractive index of the medium. Equation (7.36) can then be written as

$$n_1 \sin\theta_\mathrm{i} = n_2 \sin\theta_\mathrm{t}. \tag{7.37}$$

This is the common form of Snell's law encountered in optics.

7.4 Reflection and transmission of EM waves: Fresnel equations

As a final example, we consider a full 3D vectorial situation of particular importance: reflection and transmission of EM waves (e.g. light) at dielectric interfaces, such as air-glass or air-water boundaries. Only the main steps and results are presented, with the full analysis left as exercise.

The physics of EM waves is governed by Maxwell's equations, from which one can readily derive the following wave equations for the electric and magnetic fields $\vec{\mathbf{E}}(\vec{\mathbf{r}}, t)$ and $\vec{\mathbf{B}}(\vec{\mathbf{r}}, t)$, respectively:

$$\boxed{\begin{aligned} \frac{\partial^2 \vec{\mathbf{E}}(\vec{\mathbf{r}}, t)}{\partial t^2} &= \frac{c^2}{n^2}\nabla^2 \vec{\mathbf{E}}(\vec{\mathbf{r}}, t), \qquad (7.38)\\ \frac{\partial^2 \vec{\mathbf{B}}(\vec{\mathbf{r}}, t)}{\partial t^2} &= \frac{c^2}{n^2}\nabla^2 \vec{\mathbf{B}}(\vec{\mathbf{r}}, t), \qquad (7.39) \end{aligned}}$$

where c is the speed of light in vacuum and n is the refractive index of the medium under study. From Maxwell's equations, one can also derive the interface conditions that must be met at the boundary of two dielectrics:

- The vector component of $\vec{\mathbf{E}}$ that is tangent to the interface must be continuous.
- The vector component of $\vec{\mathbf{B}}/\mu$, where μ is the magnetic permeability, that is tangent to the interface must be continuous.

The plane wave solutions to the EM wave equations above read

$$\vec{\mathbf{E}}(\vec{\mathbf{r}},t) = \vec{\mathbf{E}}_0 e^{i(\omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}})}, \tag{7.40}$$

$$\vec{\mathbf{B}}(\vec{\mathbf{r}},t) = \vec{\mathbf{B}}_0 e^{i(\omega t - \vec{\mathbf{k}}\cdot\vec{\mathbf{r}})}, \tag{7.41}$$

where $\vec{\mathbf{E}}_0$ and $\vec{\mathbf{B}}_0$ are the vector amplitudes of the electric and magnetic fields, respectively, and the wave vector satisfies the dispersion relation $k = ||\vec{\mathbf{k}}|| = n\omega/c$. The full Maxwell's equations instil additional requirements that the electric and magnetic field amplitudes $\vec{\mathbf{E}}_0$ and $\vec{\mathbf{B}}_0$ must satisfy[31]. First, the electric and magnetic fields amplitudes $\vec{\mathbf{E}}_0$ and $\vec{\mathbf{B}}_0$ are perpendicular to the wave vector $\vec{\mathbf{k}}$ and each other. Second, the directions of the wave vector and the field amplitudes follow from the right-hand rule, and the magnitudes of the field amplitudes satisfy $E_0 = ||\vec{\mathbf{E}}_0|| = c/n||\vec{\mathbf{B}}_0|| = B_0$. Mathematically

$$\vec{\mathbf{k}} \times \vec{\mathbf{E}}_0 = \omega \vec{\mathbf{B}}_0. \tag{7.42}$$

Depending on the polarization of the EM wave (i.e., the direction of $\vec{\mathbf{E}}_0$), there are two different cases to consider. Specifically, as shown in Fig. 27, we can either have the electric field polarized perpendicular [Fig. 27(a)] or parallel [Fig. 27(b)] to the plane of incidence (and vice versa for the magnetic field). The polarization in the former situation is referred to as s polarization and in the latter as p polarization. They are also referred to as *transverse electric (TE)* and *transverse magnetic (TM)* polarizations.

7.4.1 Case 1: s polarization

For s polarization [see Fig. 27(a)], the electric field is parallel to the interface[32]. In other words, the electric field is entirely tangent to the interface, and our first interface condition can therefore be written as

$$\vec{\mathbf{E}}_{\rm i}(\vec{\mathbf{r}},t) + \vec{\mathbf{E}}_{\rm r}(\vec{\mathbf{r}},t) = \vec{\mathbf{E}}_{\rm t}(\vec{\mathbf{r}},t), \tag{7.43}$$

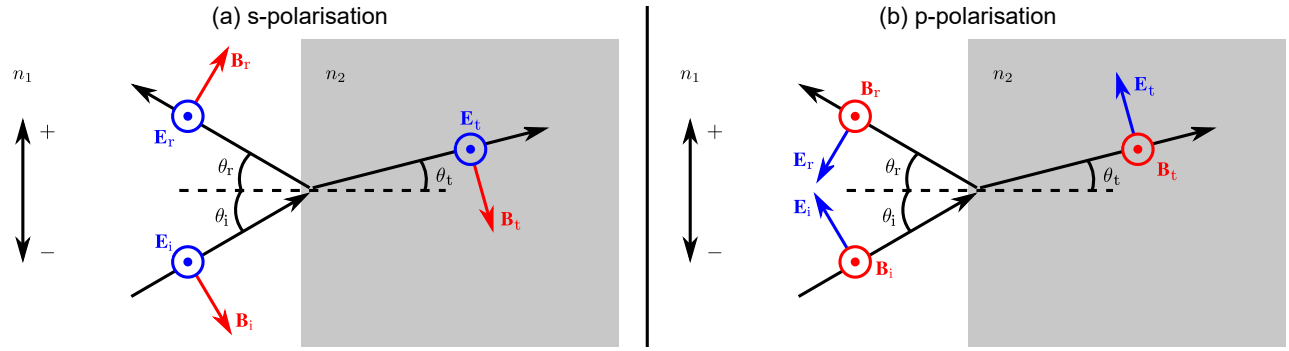

Figure 27: Schematic illustration of reflection and transmission of an electromagnetic wave at the interface of two dielectric media with refractive indices n_1 and n_2.

[31] See lecture notes for "Physics 333 – Lasers and electromagnetic waves" by ME.

[32] Recall that the plane of incidence contains the surface normal, and so being transverse to the plane of incidence implies being perpendicular to the surface normal, hence parallel to the surface itself.

where the subscripts refer to incident, reflected, and transmitted waves. Following the analysis in subsection 7.3, the exponential terms $\exp(i\omega t - i\vec{\mathbf{k}} \cdot \vec{\mathbf{r}})$ cancel out, yielding a condition for the vector amplitudes:

$$\vec{\mathbf{E}}_{0\mathrm{i}} + \vec{\mathbf{E}}_{0\mathrm{r}} = \vec{\mathbf{E}}_{0\mathrm{t}}. \tag{7.44}$$

It can be shown that both the reflected and transmitted waves are also s polarized, and so we can write the equation in scalar form:

$$\boxed{E_{0i} + E_{0r} = E_{0t},} \tag{7.45}$$

where $E_{0n} = ||\vec{\mathbf{E}}_{0n}||$.

To proceed, we invoke our second interface condition: continuity of the tangential component of $\vec{\mathbf{B}}/\mu$. Referring to Fig. 27(a), we have

$$\boxed{-\cos\theta_{\mathrm{i}} \frac{B_{0\mathrm{i}}}{\mu_{\mathrm{i}}} + \cos\theta_{\mathrm{r}} \frac{B_{0\mathrm{r}}}{\mu_{\mathrm{r}}} = -\cos\theta_{\mathrm{t}} \frac{B_{0\mathrm{t}}}{\mu_{\mathrm{t}}}.} \tag{7.46}$$

Note that the negative signs come from the convention used in Fig. 27(a), where we have chosen down to be negative, and we have made a *guess* for the direction of the electric fields of the reflected and transmitted waves (there are two options); if our guess is wrong, the sign of the reflection/transmission coefficient will fix it. Assuming the materials to be non-magnetic, such that the permeability of both materials is equal to the free-space permeability μ_0, and recalling that $\theta_{\mathrm{r}} = \theta_{\mathrm{i}}$, we obtain

$$(B_{0\mathrm{i}} - B_{0\mathrm{r}}) \cos\theta_{\mathrm{i}} = B_{0\mathrm{t}} \cos\theta_{\mathrm{t}} \tag{7.47}$$

Finally using $B_0 = E_0 n/c$ and the first interface condition [Eq. (7.45)], we obtain the following amplitude reflection and transmission coefficients:

$$\boxed{r_s = \left(\frac{E_{0\mathrm{r}}}{E_{0\mathrm{i}}}\right)_s = \frac{n_1 \cos\theta_{\mathrm{i}} - n_2 \cos\theta_{\mathrm{t}}}{n_1 \cos\theta_{\mathrm{i}} + n_2 \cos\theta_{\mathrm{t}}},} \tag{7.48}$$

$$\boxed{t_s = \left(\frac{E_{0\mathrm{t}}}{E_{0\mathrm{i}}}\right)_s = \frac{2\, n_1 \cos\theta_{\mathrm{i}}}{n_1 \cos\theta_{\mathrm{i}} + n_2 \cos\theta_{\mathrm{t}}},} \tag{7.49}$$

where n_1 and n_2 are the refractive indices of the incident and the final media, respectively. These equations, in conjunction with Snell's law, provide the amplitude reflection and transmission coefficients for s-polarized light incident on dielectric boundaries, allowing us to deduce the amplitudes of the electric fields corresponding to the reflected and transmitted waves. It is worth noting that the coefficients can be negative — or even complex — which simply imply phase shifts for the EM waves. It is also worth noting that equations very similar to those in Eqs. (7.48) and (7.49) can be derived for the reflection of acoustic waves.

7.4.2 Case 2: p polarization

Let us now consider the case where the electric field is polarized along the plane of incidence [see Fig. 27(b)]. The continuity condition of the tangential component of the electric field reads

$$\boxed{E_{0\mathrm{i}} \cos\theta_{\mathrm{i}} - E_{0\mathrm{r}} \cos\theta_{\mathrm{r}} = E_{0\mathrm{t}} \cos\theta_{\mathrm{t}}.} \tag{7.50}$$

The magnetic field $\vec{\mathbf{B}}_0$ is now fully tangential, and so the second interface condition reads

$$\frac{B_{0\mathrm{i}}}{\mu_\mathrm{i}} + \frac{B_{0\mathrm{r}}}{\mu_\mathrm{r}} = \frac{B_{0\mathrm{t}}}{\mu_\mathrm{t}}. \tag{7.51}$$

Again assuming the materials to be non-magnetic and using $B_0 = E_0 n/c$, we can combine the two interface conditions to yield the following amplitude reflection and transmission coefficients:

$$r_\mathrm{p} = \left(\frac{E_{0\mathrm{r}}}{E_{0\mathrm{i}}}\right)_p = \frac{n_2 \cos\theta_\mathrm{i} - n_1 \cos\theta_\mathrm{t}}{n_1 \cos\theta_\mathrm{t} + n_2 \cos\theta_\mathrm{i}}, \tag{7.52}$$

$$t_\mathrm{p} = \left(\frac{E_{0\mathrm{t}}}{E_{0\mathrm{i}}}\right)_p = \frac{2\, n_1 \cos\theta_\mathrm{i}}{n_1 \cos\theta_\mathrm{t} + n_2 \cos\theta_\mathrm{i}}, \tag{7.53}$$

Fresnel equations

Equations for the reflection and transmission amplitude coefficients $r_{s,p}$ and $t_{s,p}$ presented above are known as the Fresnel equations. They can be used (together with Snell's law) to deduce the amount of light reflected (and transmitted) from a given dielectric boundary. It is worth noting that arbitrary polarization states can be expressed as a liner superposition of s and p polarizations, and so the Fresnel equations can be used to evaluate the reflection and transmission of light with arbitrary polarization.

As an example, consider light incident from air ($n_1 = 1$) to glass ($n = 1.45$) at normal incidence $\theta_\mathrm{i} = 90°$. The magnitude of the reflection coefficient for both the s and p polarizations is

$$|r| = \left|\frac{n_2 - n_1}{n_1 + n_2}\right| \approx 0.184. \tag{7.54}$$

7.4.3 Energy conservation

It is straightforward to see that the Fresnel reflection and transmission coefficients do not add up to unity. As with our earlier example of reflection and transmission of waves at the boundary of two strings, this observation does not imply that energy conservation is violated.

It can be shown that a plane EM wave transfers energy in the direction of its propagation (i.e., along the direction of its wave vector $\vec{\mathbf{k}}$). The rate of energy transfer (power) per unit area is described by the intensity of the wave[33]. For a plane EM wave with complex amplitude E, the intensity is given by

$$I = \frac{n}{c\mu_0}|E|^2. \tag{7.55}$$

[33] More technically, instantaneous energy transfer of an EM wave is described by the so-called Poynting vector. This vector is parallel to the wave vector (i.e., energy is transferred along the wave vector), while its time-averaged magnitude corresponds to intensity.

Importantly, this quantity describes the rate at which energy passes through a surface that is *perpendicular* to the wave propagation direction. If we wish to know the rate of energy transfer per unit area through some other surface that is not perpendicular to the wave, we need to project the wave vector on the normal of the surface, i.e.,

$$I_{\rm n} = I\,\hat{\mathbf{k}}\cdot\hat{\mathbf{n}} = I\cos\theta, \tag{7.56}$$

where $\hat{\mathbf{n}}$ is the unit vector along the surface normal and θ is the angle between the wave vector and that surface normal.

In light of the above discussion, the intensity incident *on* the surface can be written as

$$I_{\rm i,n} = \frac{n_1}{c\mu_0}|E_{0\rm i}|^2\cos\theta_{\rm i}. \tag{7.57}$$

Similarly, the intensities of the reflected and transmitted waves projected along the surface normal can be written as

$$I_{\rm r,n} = \frac{n_1}{c\mu_0}|r|^2|E_{0\rm i}|^2\cos\theta_{\rm i} = RI_{\rm i,n}, \tag{7.58}$$

$$I_{\rm t,n} = \frac{n_2}{c\mu_0}|t|^2|E_{0\rm i}|^2\cos\theta_{\rm t} = TI_{\rm i,n}. \tag{7.59}$$

Here we have introduced the *intensity* reflection and transmission coefficients:

$$\boxed{\begin{aligned} R &= |r|^2, \qquad (7.60)\\ T &= \frac{n_2\cos\theta_{\rm t}}{n_1\cos\theta_{\rm i}}|t|^2. \qquad (7.61)\end{aligned}}$$

It is straightforward to show that the intensity reflection and transmission coefficients do add up to unity [exercise], as expected on the basis of energy conservation.

7.4.4 Angular dependence

The Fresnel equations show that the reflection and transmission coefficients depend on the incident angle $\theta_{\rm i}$. In Fig. 28(a) and (b), we respectively plot the magnitude and phase of the reflection coefficients as a function of the incident angle for the case where $n_1 < n_2$. For s polarized waves, a π phase shift occurs for all incident angles (i.e., $r_{\rm s} < 0$), while the magnitude of the reflection coefficient increases monotonically, reaching $|r_{\rm s}| = 1$ at $\theta_{\rm i} = 90^\circ$. In contrast, for the p polarization, we see that the magnitude of the reflection coefficient initially decreases, and becomes zero before starting to increase. The incident angle for which the reflection coefficient of the p polarization is zero is known as **Brewster's angle** and denoted $\theta_{\rm B}$. It can be shown that the Brewster's angle corresponds to the incident angle for which the sum of the incident and transmitted angles is 90 degrees. Mathematically,

$$\boxed{\theta_{\rm B} + \theta_{\rm t} = 90^\circ.} \tag{7.62}$$

At the Brewster's angle, a p polarised EM wave is totally reflected (zero transmission), and the reflected EM wave undergoes a π phase-shift.

In Fig. 28(d) and (c) we show the magnitude and phase of the reflection coefficients for the case where $n_1 > n_2$. There is one significant difference compared to the $n_1 < n_2$ case: the magnitudes of both reflection coefficients are equal to unity above a certain critical angle $\theta_c < 90°$. This corresponds to the familiar phenomenon of *total internal reflection*, where all light incident on the surface is reflected. The critical angle corresponds to the incidence angle for which the transmitted angle is equal to $90°$. Using Snell's law, we have

$$\theta_c = \sin^{-1}\left[\frac{n_2}{n_1}\right]. \tag{7.63}$$

It is left as an exercise to show that, for all angles $\theta_i > \theta_c$, the amplitude reflection and transmission coefficients are complex numbers with unity magnitude. The phases of the coefficients are not zero [c.f. Fig. 28(d)], which implies that the reflected waves acquire a non-trivial phase-shift during total internal reflection.

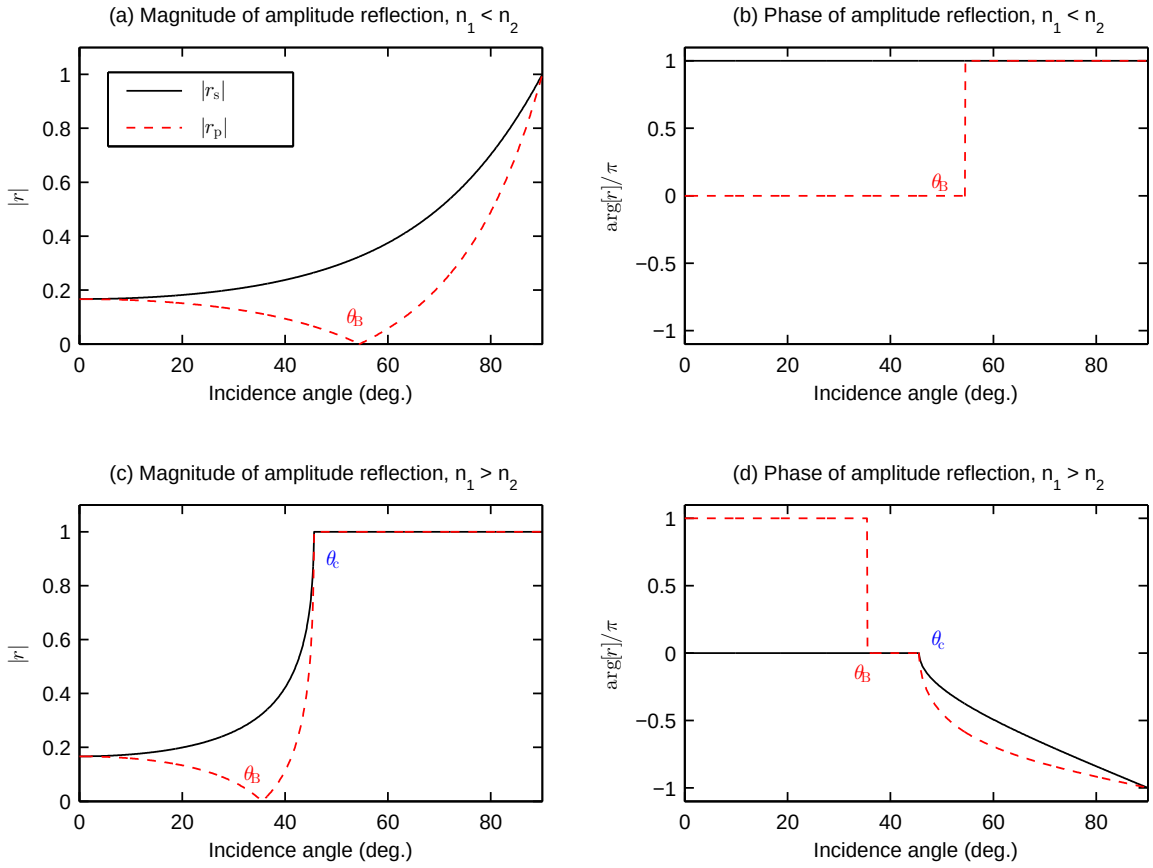

Figure 28: Magnitude (a, c) and phase (b, d) of the amplitude reflection coefficient r for s-polarised (solid black curve) and p-polarised (dashed red curve) EM waves. (a, b) corresponds to the situation $n_1 < n_2$ while (c, d) corresponds to the situation $n_1 > n_2$. The indices used are $n = 1$ and $n = 1.4$ which correspond to air and glass, respectively. θ_B and θ_c highlight the Brewster's angle and the critical angle of total internal reflection.

Problems

7.1 All waves carry energy. In this exercise, we will consider a mechanical wave $u(x,t) = A\cos(\omega t - kx)$ propagating along a string with linear mass density μ, and we will investigate the energy carried by the wave. The end goal is to show that energy is conserved when a mechanical wave is incident on a boundary between two strings with different tensions.

(a) Because the wave is associated with (transverse) motion of the string, it possesses kinetic energy. Starting from an infinitesimally small segment of the string, show that the kinetic energy contained within one wavelength λ is equal to

$$K_\lambda = \frac{1}{4}\mu A^2\omega^2\lambda. \tag{7.64}$$

(b) Because displaced segments of the string experience a harmonic force towards their equilibrium state, the wave also possesses elastic potential energy. A small segment on the string exhibits harmonic oscillation in time akin to a mass attached to a spring with spring constant $k_s = \omega^2\Delta m$, where Δm is the mass of the segment. Show that the potential energy contained within one wavelength is equal to

$$U_\lambda = \frac{1}{4}\mu A^2\omega^2\lambda = K_\lambda. \tag{7.65}$$

(c) Show that the power transferred by the wave is equal to

$$P = \frac{1}{2}ZA^2\omega^2, \tag{7.66}$$

where Z is the impedance.

(d) Use the analysis above together with Eqs. (7.20) and (7.21) to show that energy is conserved when a mechanical wave is incident on a boundary between two different strings.

7.2 Starting from Maxwell's equations, derive the interface conditions used in subsection 7.4.

7.3 Repeat the derivations of the Fresnel reflection and transmission coefficients summarised in subsection 7.4.

7.4 The impedance for electromagnetic waves is defined as $Z = c\mu_0/n$, where c is the speed of light, μ_0 is the vacuum permeability, and n is the refractive index of the medium. Consider the reflection of p-polarised electromagnetic waves at normal incidence $\theta_\mathrm{i} = 0$. How does the amplitude reflection coefficient compare with that of two strings tied together [i.e., Eq. (7.22)]? What about s-polarised waves?

7.5 Show that Fresnel's intensity reflection and transmission coefficients given by Eqs. (7.60) and (7.61), respectively, add up to unity for both s- and p-polarized waves.

7.6 Write a Matlab/Python code that plots the Fresnel intensity reflection and transmission coefficients as a function of the incident angle θ for an EM wave that is incident from air ($n = 1$) to glass ($n = 1.45$).

7.7 Show that the Brewster's angle can be written as

$$\theta_{\mathrm{B}} = \tan^{-1}\left[\frac{n_2}{n_1}\right]. \tag{7.67}$$

Further show that the Brewster's angle satisfies $\theta_{\mathrm{B}} + \theta_{\mathrm{t}} = 90^{\circ}$.

7.8 Consider an s-polarised EM wave that is incident on the boundary between dielectric medium with $n_1 = 1.5$ and air $n_2 = 1$. Derive an expression for the amplitude reflection coefficient that depends only on $n = n_2/n_1$. Show that, for incident angles larger than the critical angle of total internal reflection ($\theta_i > \theta_c$), the amplitude reflection coefficient can be written as

$$r = \frac{\cos\theta_i - i\sqrt{\sin^2\theta_i - n^2}}{\cos\theta_i + i\sqrt{\sin^2\theta_i - n^2}}, \tag{7.68}$$

where the quantity inside the square roots is real and positive. What is the value of the intensity reflection coefficient $|r|^2$ for incident angles $\theta_i > \theta_c$?

8 Waveguides

In the preceding Section, we have considered the reflection and transmission of waves at interfaces. We have seen – in the context of EM waves – that a wave incident from a medium with lower wave velocity (large refractive index n_1) into a medium with higher wave velocity (small refractive index n_2) can undergo total internal reflection. Although our example focussed on EM waves, the phenomenon itself is more ubiquitous. This can be appreciated from the fact that total internal reflection can be qualitatively inferred already from Snell's law, which applies to numerous systems subject to continuity of the wave (or its vector component) at an interface, such as e.g. seismic waves and acoustic waves.

Total internal reflection gives rise to an interesting prospect. Specifically, consider a medium of low wave velocity sandwiched between two media with larger wave velocity [Fig. 29]. The wave can be envisaged to undergo total internal reflection at both surfaces, and hence remain trapped inside the medium of low wave velocity. The resulting structure is known as a *waveguide*, since it allows the wave to remain confined (in one or more dimensions) and so be guided along a preferred direction.

Waveguides manifest themselves in numerous physical contexts. Arguably of greatest practical significance are the cylindrical glass optical fibres that allow light to be guided from place A to place B. Such optical fibres form the backbone of today's telecommunication systems, underpinning everyday applications such as the internet and all other forms of long-distance communication.

Waveguides come in many different geometries, and the precise physics that underpins wave behaviour within them can vary between different physical systems (e.g. due to different interface conditions). However, in all situations, the basic approach to analysing waveguides is essentially the same. In this Section we first cover the general idea, and subsequently consider what is arguably the simplest possible waveguide for the sake of simplicity: *a symmetrical infinite slab waveguide* like the one shown in Fig. 29. We assume (i) the wave velocity v_1 in the "core" to be smaller than the wave velocity v_2 in the "cladding", and (ii) the slab to be infinite in extent and homogeneous in the yz-plane. The slab waveguide is clearly an idealization of real waveguides, which are not infinite in width. However, the methods needed to solve real waveguides are essentially the same

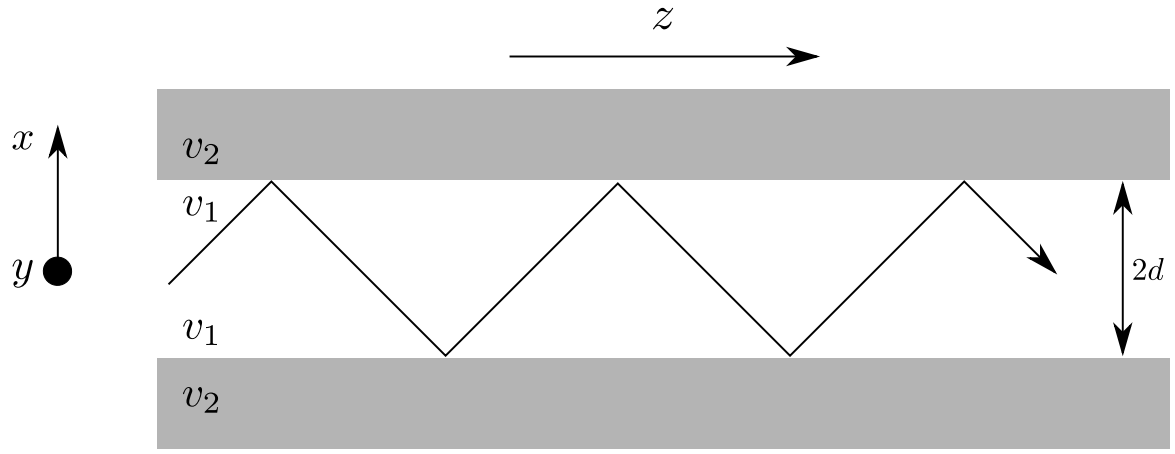

Figure 29: Schematic illustration of an infinite slab waveguide consisting of a medium of low wave velocity v_1 sandwiched between two regions of larger wave velocity v_2. Thanks to multiple total internal reflections, a wave can propagate in the low-velocity layer without escaping.

as those used to analyse the slab waveguide (just a bit more complicated), so the latter forms a good starting point. Indeed, we will end this Section by putting together a variety of the topics covered in these lecture notes and analyse the wave behaviour of light in a cylindrical optical fibre.

8.1 General idea

We assume that our wave propagates along the z-axis of the waveguide. Our aim is to find the **modes of the waveguide**: waves that maintain constant transverse spatial distribution at all distances along the waveguide axis. Mathematically, we look for solutions to the entire waveguide system that have the form

$$\vec{\mathbf{u}}(\vec{\mathbf{r}},t) = \vec{\mathbf{g}}(x,y)e^{i\omega t - i\beta z}, \tag{8.1}$$

where $\vec{\mathbf{g}}(x,y)$ is the transverse distribution of the mode, ω is the angular frequency of the wave, and β is known as the *propagation constant* of the mode. As this equation shows, the modes of the waveguides only accumulate a phase shift as they propagate, but their transverse distribution remains constant: what goes in comes out. We note that the concept of a waveguide mode is reminiscent of the normal modes covered in Section 5. The main difference is that the normal modes studied in Section 5 are non-propagating in all dimensions in the sense that the spatial pattern only oscillates with time; in other words, normal modes correspond to standing waves. Waveguide modes, in contrast, are propagating in one dimension (z), and only "maintain" their spatial structure in the two other dimensions (x and y). In a sense, waveguide modes can be understood as 2D standing waves that propagate, with the propagation constant β corresponding to an "effective" wave number.

To find the modes of a particular waveguide, we first note that the entire wave $\vec{\mathbf{u}}(\vec{\mathbf{r}},t)$ must simultaneously satisfy the wave equation both in the core and in the cladding. Referring to Fig. 29, the two wave equations read

$$\nabla^2\vec{\mathbf{u}}(\vec{\mathbf{r}},t) = \frac{1}{v_1^2}\frac{\partial^2\vec{\mathbf{u}}(\vec{\mathbf{r}},t)}{\partial t^2}, \quad \text{in the core} \tag{8.2}$$

$$\nabla^2\vec{\mathbf{u}}(\vec{\mathbf{r}},t) = \frac{1}{v_2^2}\frac{\partial^2\vec{\mathbf{u}}(\vec{\mathbf{r}},t)}{\partial t^2}, \quad \text{in the cladding.} \tag{8.3}$$

Substituting $\vec{\mathbf{u}}(\vec{\mathbf{r}},t)$ from Eq. (8.1) yields two Helmholtz-like equations

$$\nabla_\perp^2\vec{\mathbf{g}}(x,y) = \left[\beta^2 - k_1^2\right]\vec{\mathbf{g}}(x,y), \quad \text{in the core} \tag{8.4}$$

$$\nabla_\perp^2\vec{\mathbf{g}}(x,y) = \left[\beta^2 - k_2^2\right]\vec{\mathbf{g}}(x,y), \quad \text{in the cladding.} \tag{8.5}$$

Here $\nabla_\perp^2$ is a transverse Laplacian, defined as $\nabla_\perp^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$, while $k_{1,2} = \omega/v_{1,2}$ are the wave numbers of plane waves in the different media. To find the modes, we:

1. Solve the Helmholtz-like equations above to obtain the possible transverse distributions $\vec{\mathbf{g}}(x,y)$ in the core and in the cladding, and

2. pick those solutions that satisfy the interface conditions relevant to the physical system under study.

Waveguide modes correspond to those field distributions that simultaneously satisfy (i) wave equations in both media and (ii) the interface conditions at the boundary. One typically finds that only a discrete set of propagation constants β permit these conditions to be simultaneously satisfied; each waveguide mode is associated

with a unique propagation constant, and an essential step in waveguide analysis is to actually find what these propagation constants are. The analysis typically results in a transcendental eigenvalue equation from which the propagation constants (and hence the field profiles) can be obtained. We again note the similarity with the normal mode analysis presented in Section 5. The main difference is that now we are dealing with two Helmholtz-like equations in two different regions, and the "boundary" conditions will be such that solutions in the different domains "merge" appropriately at the interface.

In practice, solving the Helmholtz-like equations above requires consideration for the geometry of the waveguide. For example, for cylindrical optical fibres, the equations are convenient to be solved in cylindrical coordinates. For more complex geometries, analytical solutions may not exist, and the computations must be carried out numerically. Moreover, it should be evident that the interface conditions play an important role in determining the waveguide modes. In the next subsection, we present an example that illustrates the general ideas by considering a particularly simple waveguide geometry and interface conditions that are physically relevant.

8.2 Modes of a slab waveguide

We consider a *symmetrical slab waveguide*[34] like the one shown in Fig. 29, consisting of a medium of width $2d$ and wave velocity v_1 sandwiched between two layers of a medium with wave velocity v_2. We assume the wave to be polarized along the y-axis such that $\vec{\mathbf{g}}(x,y) = \psi(x,y)\hat{\mathbf{y}}$, and that the interface conditions of the system assert that both $\psi(x,y)$ and $\partial\psi(x,y)/\partial x$ are continuous across the boundary. It is worth noting that these interface conditions arise e.g. if $\vec{\mathbf{g}}(x,y)$ is interpreted as the electric field of an EM wave [exercise]; the resulting modes are known, in the context of optics, as the transverse electric (TE) modes of the slab waveguide.

Since the waveguide extends to infinity in the y-direction, it is reasonable to assume that the waveguide modes do not depend on y, i.e., $\psi(x,y) = \psi(x)$. Writing $p^2 = k_1^2 - \beta^2$ and $q^2 = \beta^2 - k_2^2$, we can rewrite Eqs. (8.4) and (8.5) as[35]

$$\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} = -p^2\psi, \quad |x| < d, \tag{8.6}$$

$$\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} = q^2\psi, \quad |x| \geq d. \tag{8.7}$$

The nature of the solutions of these equations depends on whether p and q are real or imaginary; it can be shown [exercise] that, for guided modes, $k_1 > \beta > k_2$ such that $p, q \in \mathbb{R}$. Oscillatory functions solve Eq. (8.6), and the general solution reads

$$\psi(x) = A\cos(px) + B\sin(px). \tag{8.8}$$

It can be shown [exercise] that sine and cosine functions cannot simultaneously fulfil the interface conditions. Hence, we have two separate sets of waveguide modes: even and odd (or symmetric and anti-symmetric). In the core, the profiles read

$$\psi_{\mathrm{core}}(x) = \begin{cases} A\cos(px) & \text{even modes} \\ B\sin(px) & \text{odd modes} \end{cases} \tag{8.9}$$

[34] The waveguide is "symmetrical" since both cladding layers are assumed to have identical wave velocities. The more general case involving three different media is left as an exercise.

[35] We assume $x = 0$ in the middle of the core.

In the cladding region, the wave profiles are given by exponential functions; the general solution of Eq. (8.7) reads

$$\psi(x) = Ce^{qx} + De^{-qx}. \tag{8.10}$$

Solutions that diverge to infinity as $|x| \to \infty$ can be ignored as unphysical, and so the profile in the cladding reads

$$\boxed{\psi_{\text{cladding}}(x) = \begin{cases} Ce^{-qx} & x \geq d \\ De^{qx} & x \leq -d. \end{cases}} \tag{8.11}$$

Next, we use interface conditions to find the values of the different coefficients. For the even modes, we have $\psi_{\text{core}}(d) = \psi_{\text{core}}(-d)$, and the condition of continuity of $\psi(x)$ and $\mathrm{d}\psi/\mathrm{d}x$ yields

$$\boxed{\begin{aligned} A\cos(pd) &= Ce^{-qd}, \\ Ap\sin(pd) &= Cqe^{-qd}. \end{aligned}} \tag{8.12}$$ (8.13)

By dividing the two equations from one another, we obtain

$$\boxed{\tan(pd) = \frac{q}{p}.} \tag{8.14}$$

For the odd modes, the interface conditions yield [exercise]

$$\boxed{\cot(pd) = -\frac{q}{p},} \tag{8.15}$$

where $\cot(pd) = 1/\tan(pd)$. These are the *characteristic equations* from which the mode propagation constants β can be obtained for even [Eq. (8.14)] and odd [Eq. (8.15)] modes. Indeed, we recall that $p = \sqrt{k_1^2 - \beta^2}$ and $q = \sqrt{\beta^2 - k_2^2}$, and so the only unknown in Eqs. (8.14) and (8.15) is the mode propagation constant β. The equations are transcendental, requiring the solutions to be found graphically and/or numerically. As an illustrative example, in Fig. 30(a) and (b) we plot the left- and right-hand sides of the two equations as a function of β for parameters listed in the figure caption. As can be seen, the curves cross at discrete and finite values of β, with each crossing corresponding to a distinct waveguide mode.

Once the mode propagation constant β is known, we may evaluate the coefficients p and q and then finally construct the whole spatial wave profile $\psi(x)$ using Eq. (8.9) and Eq. (8.11). Note in this context that the mode amplitudes (A for even modes and B for odd modes) can be chosen freely; the continuity conditions at the interface however enforce values for the coefficients C and D appearing in Eq. (8.11). In Figs. 31(a) and (b), we plot spatial profiles of selected even and odd modes whose propagation constants can be inferred from Figs. 30(a) and (b), respectively. These modes specifically correspond to the six modes with the largest propagation constant.

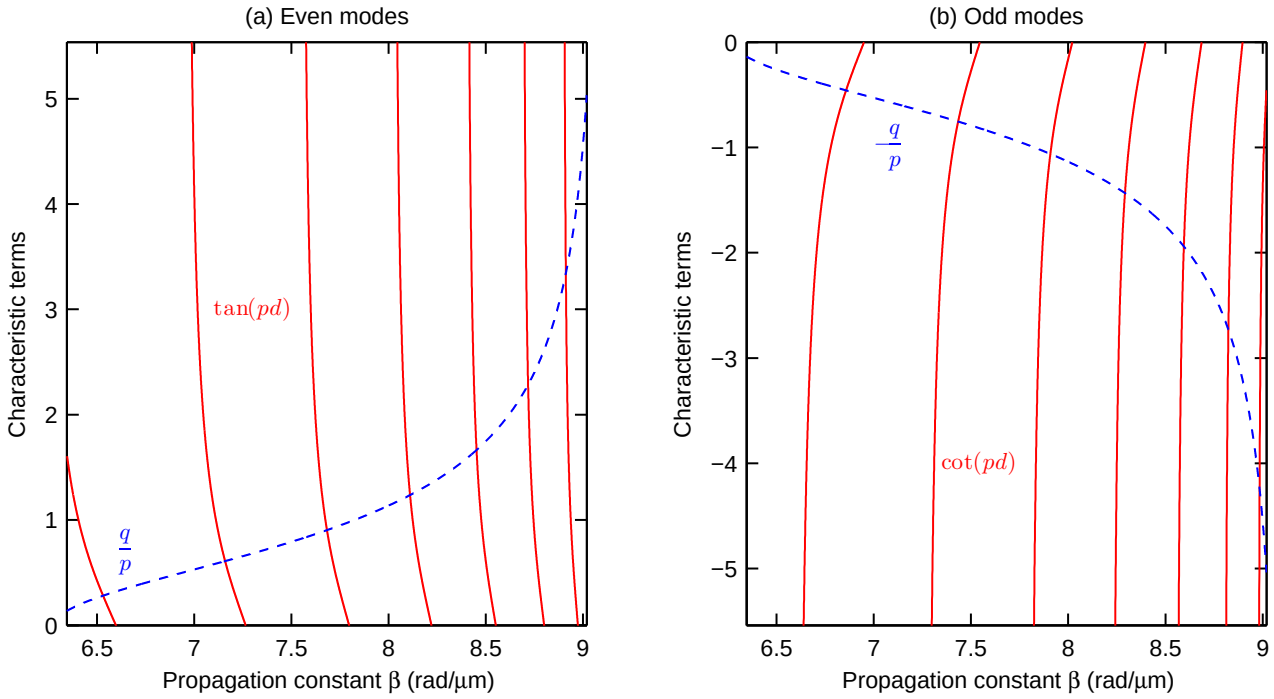

Figure 30: Visual representation of the characteristic equation for (a) even and (b) odd modes of a symmetric slab waveguide. Propagation constants for guided modes can be found from the intersection points of the different curves. The parameters of the waveguide structure correspond to a dielectric optical waveguide with core half-width $d = 4$ μm and wave velocities $v_1 = c/1.45$ and $v_2 = c$ in the core and in the cladding, respectively, with c the speed of light in vacuum. The mode has a vacuum wavelength of $\lambda_0 = 1$ μm.

8.2.1 Number of modes

A given waveguide only supports a finite number of modes. As an example, let us consider the even modes in our slab waveguide. We first write Eq. (8.14) as:

$$\tan(pd) = \sqrt{\frac{k_1^2 - k_2^2 - p^2}{p^2}}. \tag{8.16}$$

Since propagation constants of guided modes must lie on the interval $\beta \in (k_2, k_1)$, we have $p \in (0, \sqrt{k_1^2 - k_2^2})$. In this interval, the right-hand side of Eq. (8.16) is a monotonously decreasing, positive-valued function, which assumes the values of ∞ and 0 at the two ends of the interval [see Fig. 32]. On the other hand, the left-hand side of Eq. (8.14) repeats periodically with a period of $\Delta p = \pi/d$. Since the tangent function traces through all positive (and negative) numbers over a single period, the right- and left-hand sides of Eq. (8.16) must coincide once each period. Thus, the number of even modes is

$$N_{\text{even}} = \left\lceil d\frac{\sqrt{k_1^2 - k_2^2}}{\pi} \right\rceil. \tag{8.17}$$

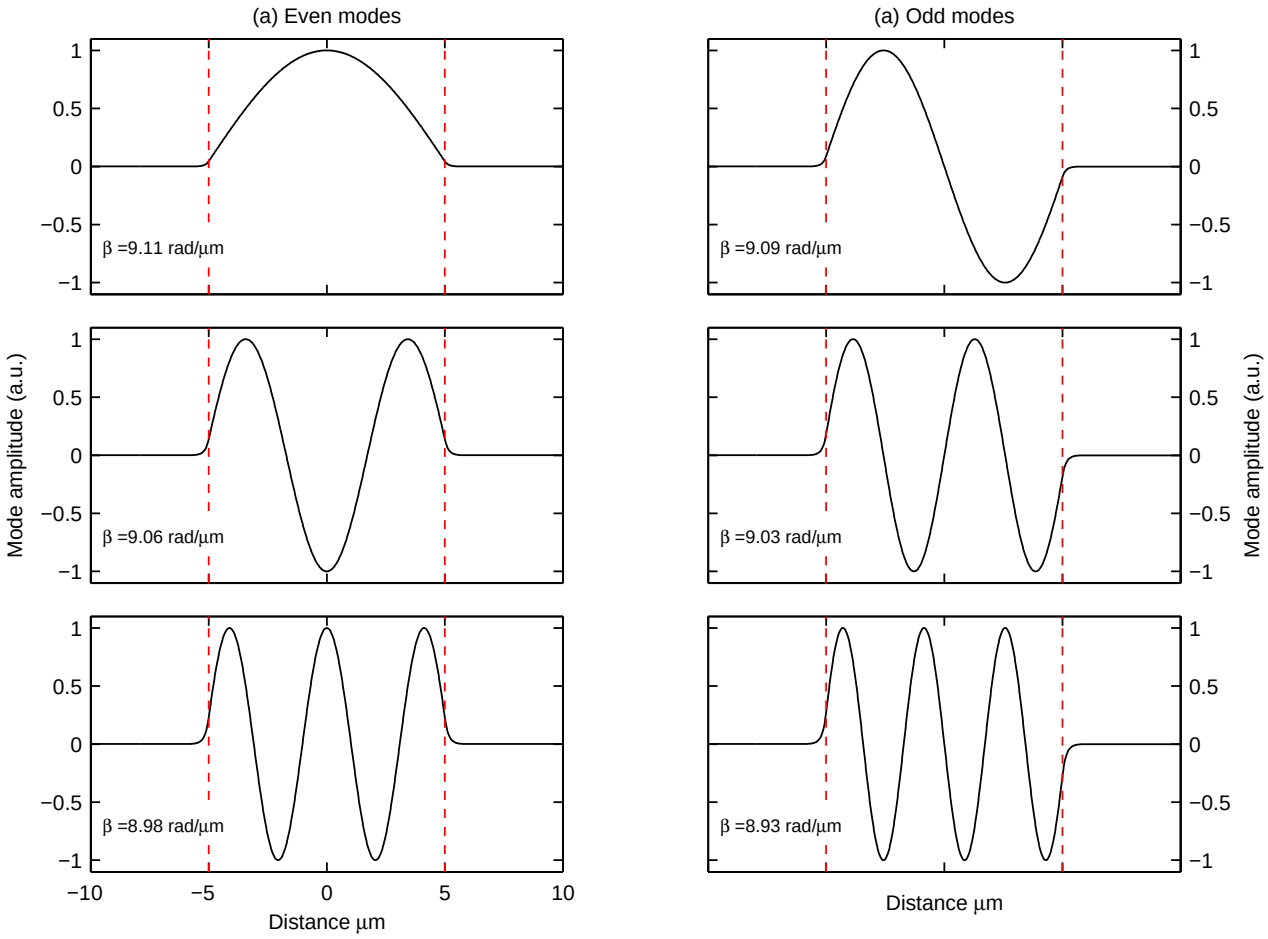

Figure 31: Spatial mode profiles for three (a) even and (b) odd modes of a symmetric slab waveguide. The parameters are the same as in Fig. 30, and the modes shown here correspond to the six modes with the largest propagation constants. The dashed vertical lines indicate the core-cladding boundaries at $d = \pm 5\ \mu\mathrm{m}$.

The necessity of the ceiling function can be understood by inspection of Fig. 32. It is straightforward to show [exercise] that the number off odd modes is the same, except that the ceiling function becomes the floor function. The total number of modes is then

$$N = \left\lceil 2d \frac{\sqrt{k_1^2 - k_2^2}}{\pi} \right\rceil . \tag{8.18}$$

Recalling that $k_n = \omega / v_n = 2\pi / \lambda_n$, the number of modes can be written as

$$N = \left\lceil 4d \sqrt{\frac{1}{\lambda_1^2} - \frac{1}{\lambda_2^2}} \right\rceil . \tag{8.19}$$

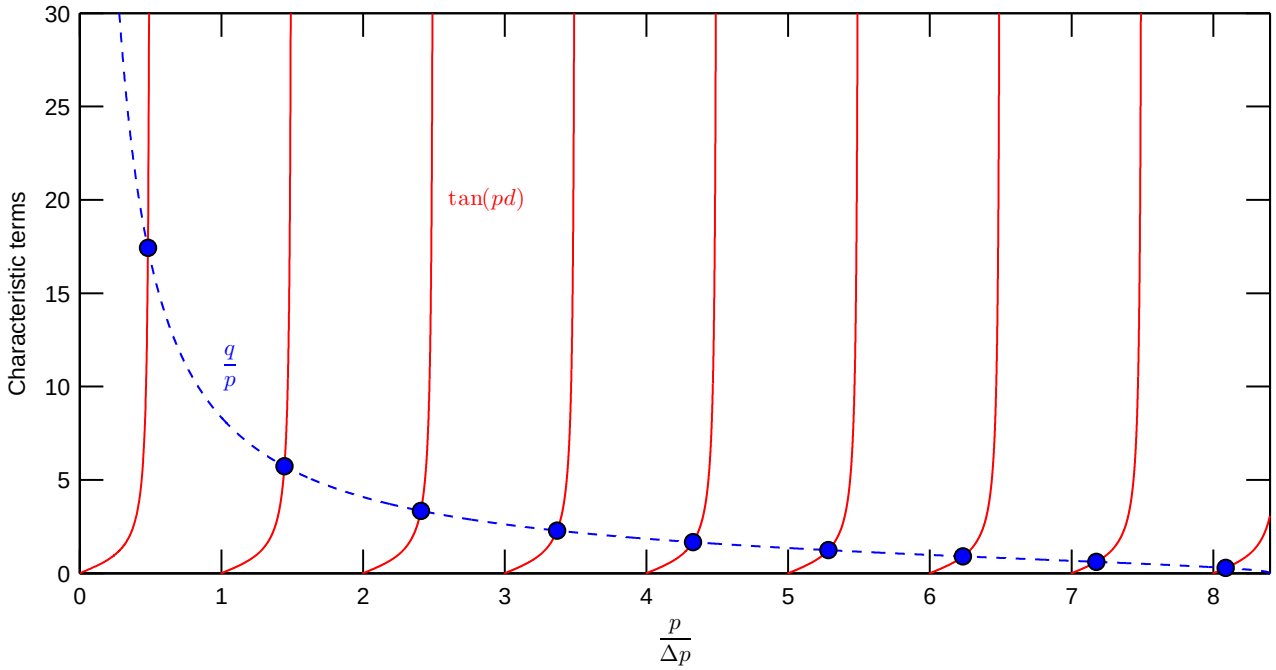

Figure 32: The characteristic equation for even modes of a symmetric slab waveguide plotted as a function of $p/\Delta p$, where $\Delta p = \pi/d$ is the periodicity of the tangent function. Points where the two curves intersect (highlighted with solid circles) correspond to guided modes. Each interval of width Δp contains one mode.

This expression shows that the smaller the wavelength λ_1 in the core, the larger the number of modes. This makes sense, as the wavelength of the wave can be loosely understood as a measure of how tightly the wave can be transversely confined[36]. If the wavelength is too large, the mode profile can be too big to fit in the core and hence cannot be guided.

Optical waveguides

The results above apply to all physical systems that are subject to the particular interface conditions. Notably, they apply to electromagnetic waves that are polarized along the surface (i.e., along the y-direction). The resulting waveguide modes are known as *transverse electric (TE)* modes to highlight the fact the electric field is transverse to the plane of incidence; it should be clear that the TE polarization can be identified as the s polarization described in the context of wave reflection and transmission at interfaces. For the p polarization, the interface conditions are somewhat different, leading to another set of modes known as *transverse magnetic* (TM) modes.

For electromagnetic waves, we have $\lambda_i = \lambda_0/n_i$, where n_i is the refractive index of medium i

[36] Consider for example the familiar diffraction-limited spot size for electromagnetic waves, ie., the smallest spot size to which light can be focussed: $d = \lambda/(2\,\mathrm{NA})$, where NA is known as the numerical aperture.

and λ_0 is the vacuum wavelength. The number of modes can now be written as

$$N = \left\lceil \frac{4d}{\lambda_0} \sqrt{n_1^2 - n_2^2} \right\rceil . \tag{8.20}$$

The factor $\sqrt{n_1^2 - n_2^2}$ is known as the *numerical aperture (NA)* of the waveguide, and it describes the range of angles over which the system can accept or emit light. The numerical aperture is typically of the order of unity (or smaller), which makes clear that the number of modes is indeed governed by the size of the core relative to the wavelength.

8.2.2 Fundamental mode and cutoff wavelength

Inspection of Fig. 32 reveals that, regardless of parameters, there is always at least one root for Eq. (8.14). This implies that our slab waveguide will always support at least one mode. This mode is known as the *fundamental mode*, and it can be identified as the mode with the largest propagation constant β. It is straightforward to see that the fundamental mode is an even mode; indeed, one can readily envisage a set of parameters for which Eq. (8.14) has no roots. The top even mode shown in Fig. 31 corresponds to the fundamental mode of that waveguide. When a waveguide is only supporting a single mode, it is said to be a *single-mode* waveguide.

In the context of optics, it is customary to define a *cut-off wavelength* λ_c: the vacuum wavelength above which the waveguide only guides the fundamental mode. Using Eq. (8.20), we have:

$$\lambda_c = 4d\sqrt{n_1^2 - n_2^2}. \tag{8.21}$$

As can be seen, a small waveguide width d implies a small cutoff wavelength λ_c and vice versa.

8.3 Quantum mechanical analogy

In the subsection above, we have solved the modes of an infinite slab waveguide by first finding the transverse wave profiles in the core and in the cladding and then joining them at the interface using boundary conditions. Under certain conditions, the physics of the problem may permit the waveguide problem to be analysed from a somewhat different perspective. Here, one expresses the problem as a single wave equation that incorporates the spatial-dependence of the wave velocity. Considering a one-dimensional scalar situation (e.g. a slab waveguide with velocity variation along x), we have

$$\frac{\partial^2 u(\vec{\mathbf{r}}, t)}{\partial t^2} = v^2(x) \nabla^2 u(\vec{\mathbf{r}}, t). \tag{8.22}$$

Note the explicit dependence of the wave velocity on the position x. We must emphasize that this equation is not guaranteed to be a universally valid description of the physical system under study. For example, in electromagnetism, Eq. (8.22) can be strictly derived [exercise] only in the case where the electric field is polarized parallel to the interface (i.e., TE waves)[37]. More generally, Eq. (8.22) implicitly implies that the wave function

[37] Manipulation of Maxwell's equations results in a wave equation for the electric field that contains a term that disappears when the wave is polarised along the interface, but that does not disappear if the field exhibits a component that is perpendicular to the interface. However, if the velocity differences are very small, this term can be neglected even for the latter case.

is twice-differentiable at all spatial positions, which requires the continuity of the wave function and its first derivative. Thus, Eq. (8.22) can be considered valid whenever the interface conditions require continuity of the wave function and its first derivative[38].

The form of Eq. (8.22) opens up an interesting analogy. To see this, let us again assume that the transverse profile is constant in the y-direction and write $u(\vec{\mathbf{r}}, t) = \psi(x)\exp(i\omega t - i\beta z)$. This yields the following Helmholtz-like equation:

$$\boxed{\frac{\partial^2 \psi}{\partial x^2} = \left[\beta^2 - k^2(x)\right]\psi,} \tag{8.23}$$

where $k(x) = \omega/v(x)$. Equation (8.23) is formally identical with the 1D time-independent Schrödinger's equation:

$$\boxed{\frac{\partial^2 \psi}{\partial x^2} = \frac{2m}{\hbar^2}\left[-E + V(x)\right]\psi.} \tag{8.24}$$

Here E is the total energy of the particle and $V(x)$ is the potential energy at each point x. The similarity of the two equations reveals that the problem of finding waveguide modes is in fact analogous to finding the bound states for a particle in a finite potential well. It is worth noting that the interface conditions in quantum mechanics instil the continuity of the wave function ψ and its first derivative, which are captured by the requirement of twice-differentiability manifest in Eq. (8.24).

In the quantum mechanical problem, and considering a rectangular potential well (analogous to a slab waveguide) that assumes the values V_1 in the well and V_2 outside the well, the energies of bound states lie on the interval $E \in (V_1, V_2)$. Indeed, if $E < V_1$, the kinetic energy of the particle would be negative everywhere, which is unphysical. On the other hand, if $E > V_2$, the kinetic energy outside of the well would be positive, permitting the particle to enter that region. It is only for $E \in (V_1, V_2)$ that the kinetic energy is positive in the well and negative elsewhere, allowing the wave to remain localized in the well.

Using the analogy, we may identify $\beta^2 \equiv -E$ and $k^2(x) \equiv -V(x)$. Since bound states require that $E \in (V_1, V_2)$, our analogy now reveals immediately that guided modes similarly require $\beta \in (k_2, k_1)$.

8.4 Mode dynamics

Our analysis above shows that a waveguide can sustain a finite number of eigenmodes that maintain constant transverse shape as they propagate. But what if we launch a profile that does not match any of the eigenmodes into a waveguide? What will happen? Short answer is that some part of the launched energy will reshape into one or more of the eigenmodes, while some other part will escape into the cladding and not be guided.

To get an intuitive feel for this process, we consider a slab waveguide described by Eq. (8.22). We write the wave in the form $u(\vec{\mathbf{r}}, t) = \psi(x, z)\exp(i\omega t - ik_1 z)$ and use the paraxial approximation [see Section 6.2] to derive [exercise]

$$\boxed{\frac{\partial \psi}{\partial z} = -\frac{i}{2k_1}\frac{\partial^2 \psi}{\partial x^2} - \frac{i}{2k_1}\left[k^2(x) - k_1^2\right]\psi.} \tag{8.25}$$

[38] These conditions are satisfied by TE-polarized EM waves, but they are not satisfied by TM-polarized waves.

Here we have a differential equation that tells us how a given incident wave, described by the initial condition $\psi(x, z = 0)$, will propagate along the waveguide. The equation cannot be solved analytically. Accordingly, we solve it numerically, using the so-called split-step Fourier algorithm.

Split-step Fourier

If the right-hand side of Eq. (8.25) would only contain the first term, the differential equation could be trivially solved in the Fourier domain. In fact, this would simply correspond to the paraxial beam propagation Eq. (6.12) which we solved in Section 6.2.1. On the other hand, if the equation would only contain the second term, we could trivially solve it in the real space domain. Unfortunately, because the equation contains both terms, we must solve it numerically.

The split-step Fourier method is a scheme that allows us to find numerical solutions to equations like Eq. (8.25). The idea is that, over a small interval h in z, we can assume that the two terms on the right-hand side of Eq. (8.25) act independently. More specifically, propagation from z to $z + h$ is carried out in two steps. In the first step, the first term acts alone and the second term is zero. In the second step, the first term is zero and the second term acts alone. After the first step, we have

$$\psi_1(x, z + h) = \mathcal{F}^{-1}\left\{\tilde{\psi}(k_x, z) \exp\left[i\frac{k_x^2}{2k_1}h\right]\right\}, \tag{8.26}$$

where $\tilde{\psi}(k_x, z)$ is the Fourier transform of $\psi(x, z)$, k_x is the (angular) frequency variable of the Fourier transform, and $\mathcal{F}^{-1}$ indicates the inverse Fourier transform. In the second step, we use $\psi_1(x, z + h)$ as the initial condition to obtain:

$$\psi(x, z + h) = \psi_1(x, z + h) \exp\left[-i\frac{k^2(x) - k_1^2}{2k_1}h\right]. \tag{8.27}$$

The two steps above can be used to give us the wave profile $\psi(x, z + h)$ provided the profile $\psi(x, z)$ is known. Typically we know the initial condition $\psi(x, 0)$, i.e., the wave profile at the input facet of the waveguide. To obtain the profile at an arbitray z, we must apply the two steps iteratively: $0 \to h \to 2h \to 3h...$ until the desired point is reached.

We must emphasize that the split-step Fourier scheme is not exact: the two terms on the right-hand side of Eq. (8.25) do not act independently in reality. The assumption that they do act independently is an approximation that is only valid for sufficiently small step size h. Accordingly, the accuracy of the whole scheme depends on the step size h, with smaller step sizes giving better result.

8.4.1 Illustrative simulations

Let us now use the split-step Fourier scheme to examine how a given initial condition evolves as it propagates along a waveguide. Figure 33(a) shows results from illustrative numerical simulations [for parameters, see caption]. Here we consider a waveguide that supports one even mode and one odd mode, and assume the initial

condition to correspond to a Gaussian function with a width two times larger than the waveguide core. As can be seen, a considerable part of the wave energy is cast away from the core, while the part remaining in the core reshapes into a profile that coincides with that of the fundamental mode of the waveguide [Fig. 33(b)]. It is worth highlighting that, because the initial condition is an even function, only the even mode is excited.

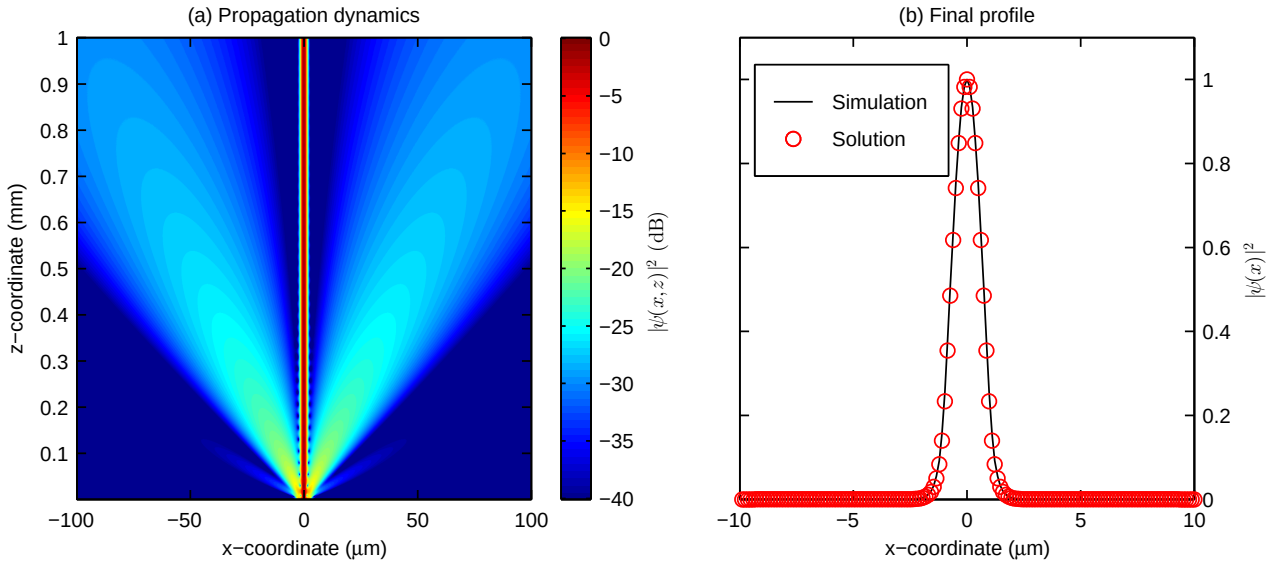

Figure 33: (a) Dynamical evolution of the wave profile $\psi(x, z)$ as it propagates through a symmetric slab waveguide. (b) Solid black curve shows the wave profile at the output of the simulation shown in (a), i.e., at $z = 1$ mm. Open red circles show the fundamental waveguide mode constructed using the equations presented in Section 8.2. The waveguide parameters mimic a dielectric optical waveguide with $d = 1$ μm and wave velocities $v_1 = c/1.45$ and $v_2 = c/1.4$ in the core and in the cladding, respectively, with c the speed of light in vacuum. The mode has a vacuum wavelength of $\lambda_0 = 1$ μm.

Another example is shown in Fig. 34. Here we consider a waveguide that supports three modes (two even modes and one odd mode). We see that a Gaussian initial condition gives rise to a transverse profile in the core that oscillates along propagation. This is because the input excites both even modes of the waveguide[39]. The oscillatory behaviour stems from the fact that the two modes are associated with different propagation constants. Specifically, the spatial dependence of a wave comprised of two modes can be written as

$$u(x, z) = u_1(x)e^{-i\beta_1 z} + u_2(x)e^{-i\beta_2 z}, \tag{8.28}$$

where $u_n(x)$ and β_n is the transverse profile and propagation constant of the n^{th} even mode, respectively. The absolute value squared [visualized in Fig. 34] reads

$$|u(x, z)|^2 = |u_1(x)|^2 + |u_2(x)|^2 + 2|u_1(x)||u_2(x)| \cos(\Delta\beta z + \phi), \tag{8.29}$$

where $\Delta\beta = \beta_2 - \beta_1$ and ϕ is a phase factor. Thus, we indeed expect oscillatory behaviour with a period of $z_{\text{per}} = 2\pi/\Delta\beta$. With the parameters used in the simulation shown in Fig. 33, we have $\beta_1 \approx 9.05$ rad/μm and $\beta_2 \approx 8.82$ rad/μm such that $\Delta\beta = 0.25$ rad/μm. We thus predict the periodicity to be $z_{\text{per}} \approx 25$ μm. This value is in close agreement with the periodicity seen in the simulation [see Fig. 34(b)].

[39]The odd mode is again not excited because the initial condition is even.

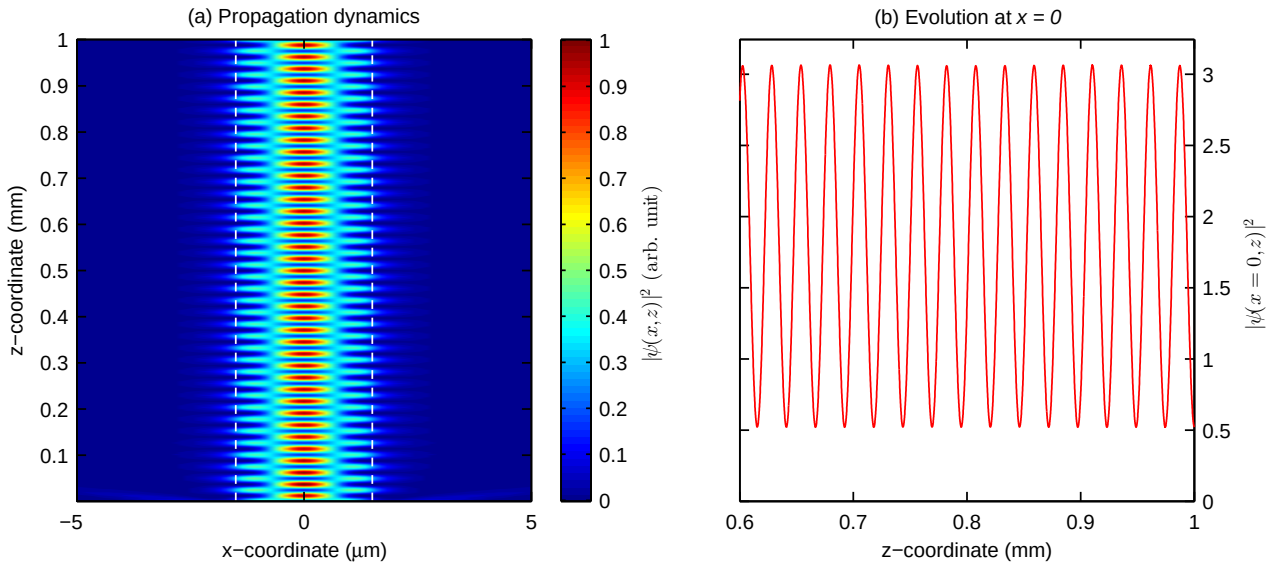

Figure 34: (a) Dynamical evolution of the wave profile $\psi(x, z)$ as it propagates through a symmetric slab waveguide. The parameters are the same as in Fig. 33 except for the waveguide width which is now set at $d = 1.5$ μm so as to permit two even modes (dashed vertical lines in (a) highlight the core-cladding boundaries). (b) Evolution of the on-axis wave intensity, i.e., $|\psi(x = 0, z)|^2$, highlighting the periodicity of the wave evolution.

8.5 Optical fibres

Above we have considered the modes of an infinite slab waveguide. Such waveguides do not really exist in practice, but they offer arguably the simplest geometry for trying out the general approach for solving problems related to waveguides. Let us now consider a geometry that is of far greater practical significance: a cylinder. Optical fibres that underpin modern telecommunications represent an example of a cylindrical waveguide (for light); motivated by their utility, we will base our discussion on EM waves, yet emphasise the generality of our analysis.

A typical optical fibre consists of a cylindrical core with refractive index n_1 and radius a surrounded by a cladding with refractive index n_2 [see Fig. 35]. To achieve total internal reflection required for waveguiding, we have $n_1 > n_2$, such that the wave velocity in the core is smaller than that in the cladding ($v = c/n$). Both in the core and in the cladding, the electric and magnetic fields satisfy the EM wave equations [see Eqs. (7.38) and (7.39)] derived from Maxwell's equations. Following the general approach outlined above, we substitute Eq. (8.1) into the EM wave equations, and obtain the two Helmholtz-like equations given by Eqs. (8.4) and (8.5), where the vector field $\vec{\mathbf{g}}(x, y)$ now represents either the electric or magnetic field. Nothing new so far.

The electric and magnetic fields can contain six vector components in total. It should be evident from Eqs. (8.4) and (8.5) that each of the vector components satisfy the scalar equations

$$\nabla_\perp^2 \psi(x,y) = \left[\beta^2 - k_1^2\right] \psi(x,y), \quad \text{in the core} \tag{8.30}$$

$$\nabla_\perp^2 \psi(x,y) = \left[\beta^2 - k_2^2\right] \psi(x,y), \quad \text{in the cladding.} \tag{8.31}$$

where $\psi(x, y)$ now represents one of the six vector components of the EM field, while $k_1 = n_1\omega/c$ and

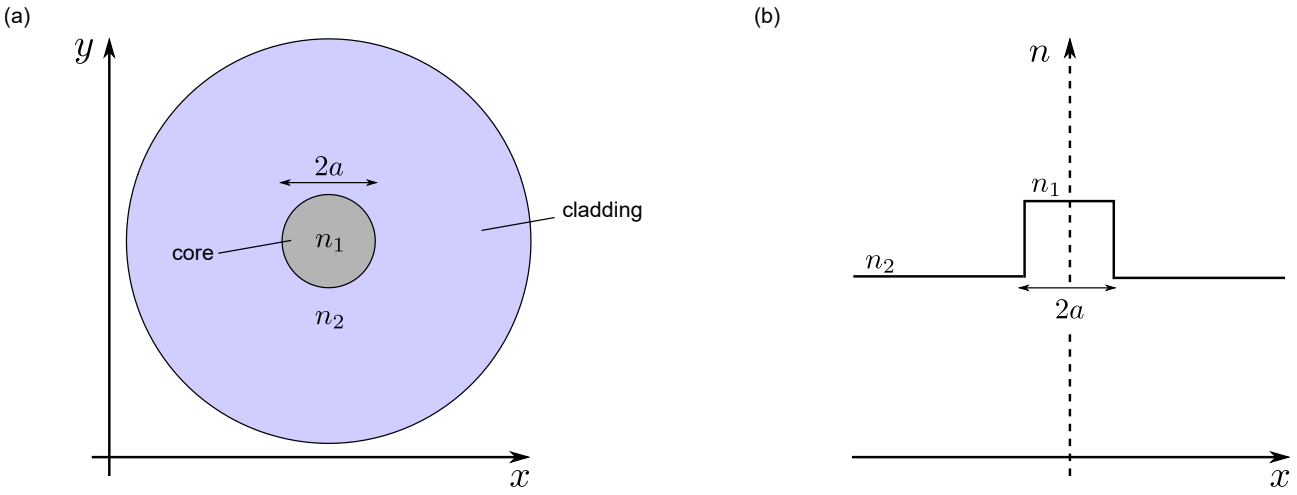

Figure 35: Schematic illustration of a cylindrical optical fibre waveguide. (a) Cross-sectional profile showing a core with refractive index n_1 surrounded by a cladding with refractive index n_2. (b) Cross-sectional view of the refractive index.

$k_2 = n_2\omega/c$ are the wavenumbers in the core and in the cladding, respectively. As above, guided modes satisfy $k_1 > \beta > k_2$.

Because of the cylindrical geometry, it is natural to write the equations in cylindrical coordinates. Let us consider first the case where $r = \sqrt{x^2 + y^2} < a$, i.e., the field within the core. Equation (8.30) can be written in cylindrical coordinates as

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \psi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 \psi}{\partial \phi^2} + (k_1^2 - \beta^2)\psi = 0. \tag{8.32}$$

But hang on, we have already seen this equation when analysing the normal modes of a drum [see Eq. (5.27)]. Briefly, using separation of variables to write $\psi(r, \phi) = F(r)G(\phi)$, we find that $G(\phi) = \exp(\pm im\phi)$, where m is an integer, whilst the radial part solves the Bessel equation:

$$r^2\frac{\partial^2 F}{\partial r^2} + r\frac{\partial F}{\partial r} + \left[r^2(k_1^2 - \beta^2) - m^2\right] F = 0. \tag{8.33}$$

Defining $p = \sqrt{k_1^2 - \beta^2}$, the solution can be written in terms of Bessel functions:

$$F(r) = AJ_m(rp) + BY_m(rp), \tag{8.34}$$

where A and B are integration constants and J_m and Y_m are Bessel functions of the first and the second kind, respectively. Since Y_m diverges at $r = 0$, it cannot contribute to a physical solution, and so we obtain

$$\boxed{\psi(r, \phi) = AJ_m\left(rp\right) e^{im\phi}, \quad \text{in the core.}} \tag{8.35}$$

So far the modes look just like the normal modes of a drum. The difference comes from outside of the core. For a drum, we (sensibly) assumed $\psi(r, \phi) = 0$ for $r > a$, but this is not the case for a cylindrical waveguide.

To find the transverse wave profile in the cladding, we simply repeat the analysis above with $k_1 \to k_2$. A complication arises from the fact that $\beta > k_2$, such that $\sqrt{k_2^2 - \beta^2}$ is purely imaginary. To circumvent the issue, we write the radial equation as:

$$r^2 \frac{\partial^2 F}{\partial r^2} + r \frac{\partial F}{\partial r} - \left[r^2(\beta^2 - k_2^2) + m^2 \right] F = 0. \tag{8.36}$$

One notes how the third term has inverted signs compared to Eq. (8.33). Equation (8.36) has the form of the so-called *modified Bessel equation*. Defining $q = \sqrt{\beta^2 - k_2^2}$, the solutions of the equation can be written as

$$F(r) = CI_m(rq) + DK_m(rq), \tag{8.37}$$

where C and D are integration constants and I_m and K_m are the *modified* Bessel functions of the first and the second kind, respectively[40]. In contrast to the standard Bessel functions, which exhibit oscillatory behaviour, the modified Bessel functions grow (I_m) or decay (K_m) exponentially as their argument approaches infinity [see Fig. 36]. Since I_m diverges as $r \to \infty$, it cannot represent a physically sensible cladding profile of a guided mode. As a consequence, we may write the entire transverse profile of the fibre mode as

$$\psi(r, \phi) = \begin{cases} AJ_m(rp)\, e^{im\phi} & r \leq a, \\ DK_m(rq)\, e^{im\phi} & r > a \end{cases} \tag{8.38}$$

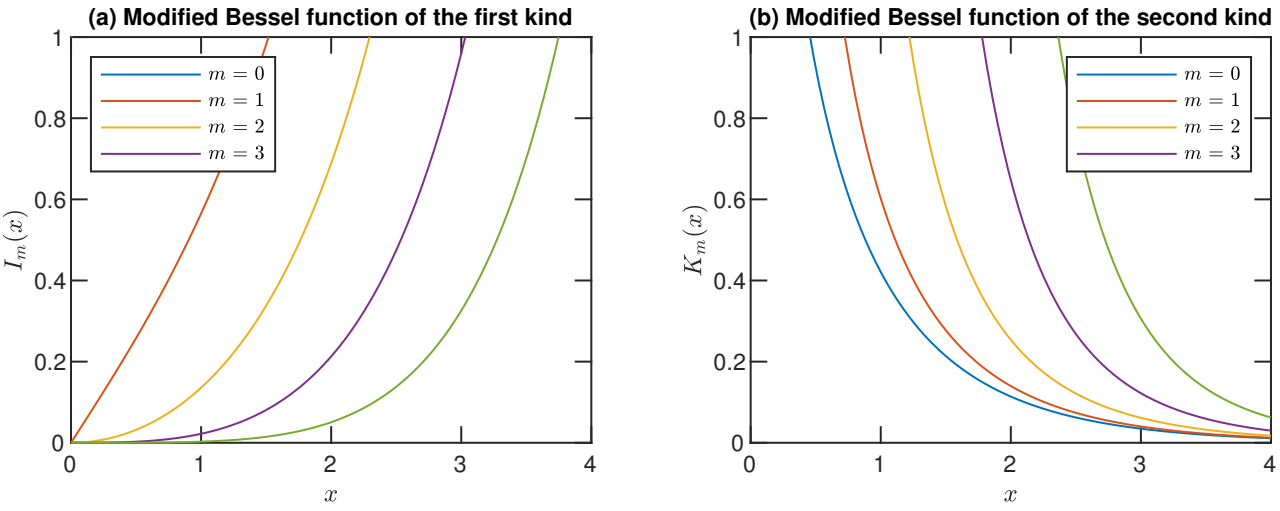

Figure 36: Modified Bessel functions of the (a) first (I_m) and (b) second (K_m) kind for orders $m = 0, 1, 2, 3$ as indicated.

8.5.1 Weakly-guiding approximation

Equation (8.38) allows us to construct the spatial profile of a guided mode in an optical fibre. Following the general idea outlined above, the next step is then to apply the interface conditions to deduce the propagation

[40] You should be able to convince yourself that the modified Bessel functions can be obtained from the standard Bessel functions by using a purely imaginary argument, e.g., $I_m(x) = J_m(ix)$.

and integration constants. This procedure turns out to be rather tedious in general, as the EM wave can have up to six nonzero vector components, each of which obeys Eq. (8.38) but with different integration constants, and each of which can contribute to the interface conditions. Luckily, a significant simplification can be made in the *weakly-guiding limit* that is almost always valid in real optical fibres.

In weakly-guiding optical fibres, the refractive index difference is very small, i.e., $n_1 = n_2 + \epsilon$, where $\epsilon \ll 1$. It can be shown that, in this case, the fields are approximately transverse, such that the longitudinal components E_z and B_z are zero. Importantly, this implies that the interface conditions are approximately equivalent to the continuity of the scalar function $\psi(r, \phi)$ in Eq. (8.38) – which describes now the transverse components of the fields – and its first derivative. It is straightforward to see that these continuity conditions imply the following characteristic equation:

$$\boxed{\frac{pJ'_m(pa)}{J_m(pa)} = \frac{qK'_m(qa)}{K_m(qa)},} \tag{8.39}$$

where the apostrophes refer to differentiation with respect to the argument, e.g.,

$$J'_m(pa) = \frac{\mathrm{d}J_m(pa)}{\mathrm{d}(pa)}. \tag{8.40}$$

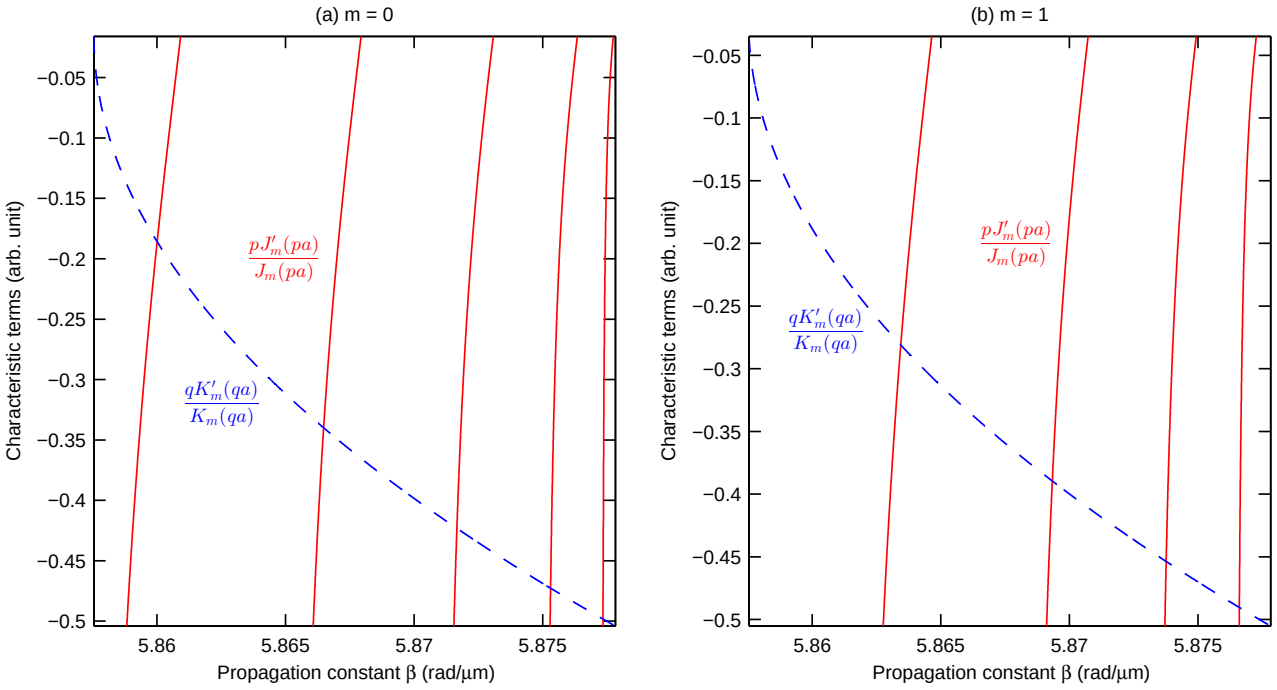

Figure 37: Visual representation of the characteristic equation for a cylindrical (optical fibre) waveguide. Propagation constants for guided modes can be found from the intersection points of the different curves. The parameters of the waveguide structure correspond to an optical fibre with core radius $a = 30\ \mu\mathrm{m}$ and refractive indices $n_1 = 1.45$ and $n_2 = 1.445$ in the core and in the cladding, respectively. The mode has a vacuum wavelength of $\lambda_0 = 1.55\ \mu\mathrm{m}$.

Note that, in practice, the derivatives of Bessel functions are easiest to evaluate using the following identities:

$$J'_m(x) = \frac{J_{m-1}(x) - J_{m+1}(x)}{2}, \tag{8.41}$$

$$K'_m(x) = -\frac{K_{m-1}(x) + K_{m+1}(x)}{2}. \tag{8.42}$$

We can graphically or numerically solve Eq. (8.39) to obtain the propagation constant β. Figure 37 shows a typical graph of the left- and right-hand side of the equations [for parameters see figure caption]. As was the case for the slab waveguide considered above, we see that the two curves cross multiple times; each crossing corresponds to a distinct guided mode. It is conventional[41] to label the propagation constants of the different modes as β_{ml}, where m refers to the order of the Bessel function and l to the order of the root of the characteristic equation. The modes themselves are commonly referred to as the LP_{ml} modes, with LP referring to "linearly polarised".

With the propagation constant known, we can construct the full spatial profile of fibre LP modes using Eq. (8.38). Note that one of the integration constants (A or D) can be chosen freely, whilst the other one follows from the continuity of the field at the interface. Figure 38 shows an assortment of the different modes supported by a fibre whose parameters were quoted in the caption of Fig. 37. Inside the core, the mode profiles are identical to the natural modes of a drum [see Fig. 16], but in contrast to the drum, the fields in a waveguide extend slightly into the cladding.

8.5.2 Number of modes

As seen above, an optical fibre can in general support a number of modes. The number of modes supported by a given fibre can be approximated using the so-called normalized frequency defined as

$$\boxed{V = a\sqrt{k_1^2 - k_2^2}.} \tag{8.43}$$

It can be shown that, in the limit where a given fibre supports a very large number of modes, that number is approximately given by

$$M \approx \frac{4}{\pi} V^2. \tag{8.44}$$

In contrast, if $V < 2.405$, the fibre supports only a single mode that corresponds to the LP_{01} mode shown in Fig. 38(a). This mode is known as the fundamental mode of the fibre, and it has an approximately Gaussian shape. It is worth noting that $x = 2.405$ corresponds to the first root of the Bessel function $J_0(x)$; the author kindly leaves it for you to figure out why the V parameter has to be lower than this value to ensure single-mode operation [see Exercise 8.12].

[41] Well actually, most sources would probably swap the indices m and l, but since the author here already used m everywhere during separation of variables, we will stick to our notation.

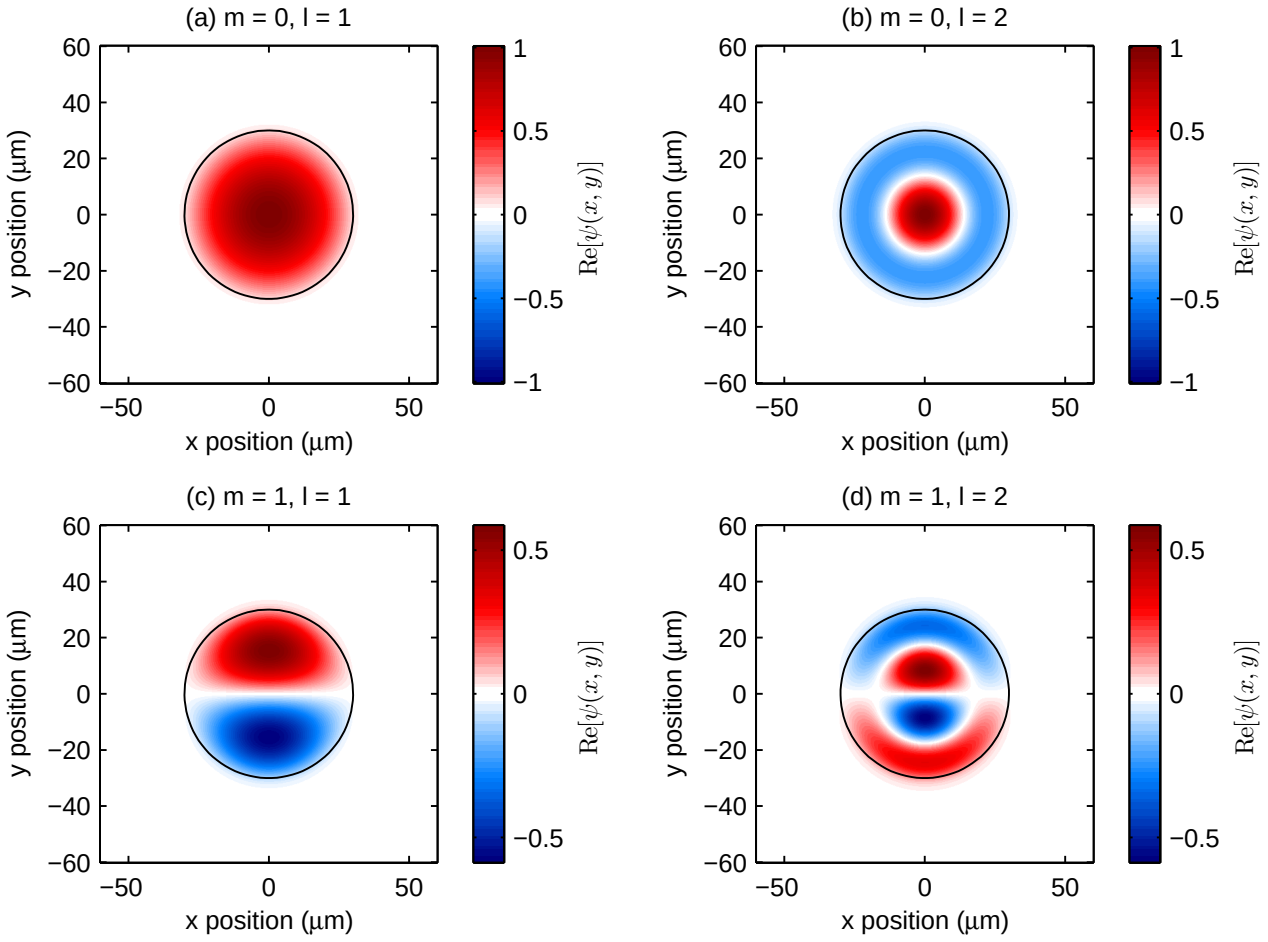

Figure 38: Transverse spatial profiles of selected step-index fibre modes as indicated. The solid black circle indicates the core circumference.

Problems

8.1 Derive the modes for an *asymmetric* slab waveguide that consists of a medium with wave velocity v_1 sandwiched between layers of two different media associated with wave velocities v_2 and v_3 with $v_1 < v_2 < v_3$. Assume the same interface conditions as in Section 8.2.

8.2 A dielectric, non-magnetic medium with refractive index n_1 sandwiched between two layers of a medium with refractive index $n_2 < n_1$ can act as a waveguide for EM waves. Assume Eq. (8.1) represents the electric field of the EM wave, and that the refractive index variation is along the x-direction [see Fig. 29]. Further assume that the EM wave is polarised along the y-direction such that the vector amplitude $\vec{\mathbf{g}}(x, y) = \psi(x)\hat{\mathbf{y}}$. Show that the EM interface conditions outlined in Section 7 imply the interface conditions used in Section 8.2, i.e., the continuity of $\psi(x)$ and $d\psi(x)/dx$.

8.3 Consider a slab waveguide for EM waves as in Question 8.2. Now assume that Eq. (8.1) represents the

magnetic field $\vec{\mathbf{B}}$ of the wave, and that the wave is polarised such that the magnetic field points in the y-direction, i.e., $\vec{\mathbf{g}}(x, y) = \psi(x)\hat{\mathbf{y}}$. Answer the following questions.

(a) What are the interface conditions for this problem?

(b) Solve for the waveguide modes in this situation. The resulting modes are known as *transverse magnetic* (TM), and should be contrasted with the *transverse electric* (TE) modes that are found with the interface conditions considered in Section 8.2.

8.4 Consider Eqs. (8.6) and (8.7), and show that physically sensible guided modes only exist for $k_1 > \beta > k_2$.

8.5 Consider the general solution to the Helmholtz-like equation in the core of a symmetric slab waveguide [Eq. (8.8)]. Show that sine and cosine functions cannot simultaneously (i.e., for the same β) satisfy the interface conditions.

8.6 Show that Eq. (8.15) is the characteristic equation for the odd modes of a symmetric slab waveguide (with interface conditions that stipulate the continuity of the wave function and its first derivative).

8.7 Show that the number of odd modes supported by a symmetric slab waveguide is given by

$$N_{\text{odd}} = \left\lfloor d\frac{\sqrt{k_1^2 - k_2^2}}{\pi} \right\rfloor . \tag{8.45}$$

8.8 Show that a wave equation of the form given by Eq. (8.22) can be derived from Maxwell's equations for an electric field that is polarised along the interface (i.e., the y-direction).

8.9 Starting from Eq. (8.22), derive Eq. (8.25).

8.10 Consider a symmetric slab waveguide for EM waves characterised by width $d = 5\ \mu\text{m}$ and wave velocities $v_1 = c/1.5$ and $v_2 = c$ in the core and cladding, respectively. Further assume an EM wave with vacuum wavelength $\lambda_0 = 1\ \mu\text{m}$.

(a) Write a Matlab/Python code that allows you to find the propagation constants for the even and odd modes of the waveguide.

(b) Use your code to plot the transverse spatial profiles of the three even modes with the largest propagation constants.

8.11 Consider a symmetric slab waveguide for EM waves characterised by width $d = 0.8\ \mu\text{m}$ and wave velocities $v_1 = c/1.5$ and $v_2 = c/1.45$ in the core and cladding, respectively. Further assume an EM wave with vacuum wavelength $\lambda_0 = 1\ \mu\text{m}$.

(a) How many even and odd modes does the waveguide support?

(b) Write a Matlab/Python code that uses the split-step Fourier algorithm to simulate the evolution of a given initial condition.

(c) Show that a given initial condition reshapes into one of the waveguide modes. Try to excite both even and odd modes (if they exist!).

8.12 Using the characteristic equation (8.39), show that $V < 2.405$ corresponds to the single-mode condition for a weakly-guiding optical fibre.