1 The essence of wave motion

1.1 Introduction

The physics of waves is too often presented only in a few rather straightforward and sometimes uninspiring contexts: the motion of strings, sound, light and so on. Students may be led to regard the topic with disdain; and they may be left with some crucial misconceptions, such as that all waves are sinusoidal. Wave physics may hence be considered an old-fashioned field with little relevance to the more modern, exciting areas of quantum physics, nanotechnology and cosmology. Yet there is plenty to find interesting just in classical and modern optics and the physics of musical instruments; and wave phenomena prove to be central to most of the fascinating and newly emerging branches of both fundamental and applied physics.

Most aspects of physics may be viewed from two perspectives: one, named after Lagrange, addresses particles, while the other, due to Euler, considers fields. We may, for example, establish the electromagnetic properties of matter by considering the Coulomb forces among all the constituent charged particles; or we may describe the material's bulk response to a field and tackle the problem that way. This duality pervades most areas of physics and, while one of the alternatives often proves vastly more convenient than the other, the two are ultimately quite consistent, equivalent viewpoints.

When we extend our analysis to dynamic systems, the particle approach becomes a form of 'kinetics' or ballistics, and changes in the field description are manifest as waves. So, when we consider the physics of waves, we're really addressing the general subject of time-dependent field theory. Stringed and wind instruments and so on are merely particularly easy examples to visualize.

In this book, we address the physics of waves in a fairly thorough fashion, but illustrate each fundamental concept with practical examples such as sonar, imaging optics, water ripples, radar and so on. As we proceed, we develop a range of techniques that have much broader application, and allow us to acquire the concepts and methods behind many other areas – the most important of which is probably that of quantum mechanics, where particle–wave duality is clearly, impressively and sometimes confusingly apparent.

1.2 A local view of wave propagation

You have doubtless already met countless examples of what are termed *waves*, from the surface waves in ripple tanks to the vibrations of a guitar string or the air column in an organ pipe, to radio waves and light. You may have been led to believe that each of these is a periodic, and perhaps even sinusoidal, travelling disturbance. But what of other, more commonplace examples? Should physicists consider ocean waves, tidal waves, bow waves and shock waves? Are journalists justified in writing of a *wave of fear* or *wave of protest*?

We shall consider many examples of wave propagation, and manifestations of wave phenomena, in the following chapters; and our understanding will stem as much from these accumulated examples as from succinct summarizing statements. We shall see for example that waves need not be periodic, that the propagated properties may sometimes be neither transverse nor longitudinal, and that it is not only for simple physical systems that a wave-propagation description can have some validity. Common principles of propagation allow very different systems to show the same characteristic phenomena of refraction, reflection, dispersion, superposition, interference and diffraction. We shall thus come to understand features of wave motion both in an abstract, generic sense and through the specific manifestations in particular examples.

It is nonetheless helpful to begin by defining what we mean by a wave, and we shall in this chapter consider three aspects. First, we shall consider what happens at a local level to allow a wave to propagate through a given region, and we shall see that important concepts are the point-to-point propagation of a physical effect, and the time lag as the effect travels from one point to another. Secondly, we shall consider the forces of interaction between two charges or masses, and shall see that waves occur when, for whatever reason, there is a finite speed of propagation between them. Finally, we shall briefly consider the nature of the disturbance propagated by the wave.

Numerous examples will illustrate these characteristics throughout this book, and a couple of specific examples of the origins of wave behaviour are considered in this chapter. To identify the defining characteristics of a wave, and illustrate the mechanisms by which it propagates, however, we start with a rather everyday, and somewhat unscientific, example.


Suggested reading

Suggestions for further reading are indicated throughout the following chapters by the sketch of an open book in the margin. For recent editions of common textbooks, abbreviated to the author's surname, specific pages or sections are given. Full details of all cited texts may be found towards the end of this book.

Fig. 1.1 The Mexican wave *la ola*, reportedly seen in North America since the 1960s, gained international popularity after the World Cup in Mexico in 1986. © Walter Spaeth, ARTside.de

1.2.1 *La ola*

We begin in Mexico City, where the Azteca Stadium was the centre of the 1968 Olympic Games and, in the lulls between events, television viewers across the world were delighted to watch a new phenomenon. Restless spectators joined in as their playful neighbours stood and waved briefly in synchronism, and the original motion spread as a ripple around the crowd. The *Mexican wave* (Fig. 1.1) had (reportedly, at least) made its international debut.

I. Farkas *et al.* [20].

The Mexican wave is found typically to travel at 12 m s^{-1} and have a width of around 15 seats. Only 30 or so spectators are needed to start a full wave.

A crowd of human spectators is, of course, a poorly defined, wildly nonlinear and inhomogeneous medium, yet the Mexican wave clearly illustrates the crucial characteristics of wave motion. The **disturbance** at any point (in this case, the vertical displacement of a particular spectator's hands, say) is a response to the action of a neighbour in raising or lowering his/her own hands; and the response is slightly delayed, through either inertia or simply the time lag in perception. These properties prove to be completely general.

1.2.2 Microscopic definition of a wave

A **wave**, then, is

> a collective bulk disturbance in which what happens at any given position is a delayed response to the disturbance at adjacent points.

The progress of a disturbance from one point to its neighbours, and thence to their neighbours and so on, is known as **propagation**; and the wave propagates through a **medium**, which may, for example, be water, air or glass, a guitar string or drum skin, an electrical cable, a crowd of spectators, or pure vacuum. The medium showing the wave motion need not necessarily be linear or homogeneous, and the wave does not need to be sinusoidal or even periodic.

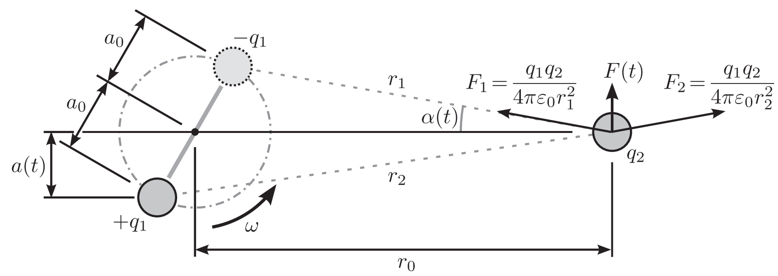


Fig. 1.2 Coulomb interaction between the charges $+q_1$ and $-q_1$ of a rotating dipole and a test charge q_2.

We shall see that it is often helpful to consider such cases, but they are specific examples rather than the only forms allowed.

Positions within the medium are described by *coordinates* (usually in one, two or three dimensions, depending upon whether the medium is a string, drum skin or ocean), and at each position the wave is described by one or more further variables – such as the fluid pressure, electric field strength or elevation of a spectator's arms – that quantify the disturbance. These variables, to which we shall generally refer explicitly, are all **wavefunctions**. Mathematically, each wavefunction depends upon, or is defined by, the coordinates of the position at which it is determined, together with a further variable: *time*.

1.3 Cause and effect

One of the most common examples of waves is the electromagnetic wave, manifest as light, radio waves and so on. We're often tempted to adopt quite a local picture of electromagnetic wave propagation, and consider a lonely photon traversing the Universe, or some region of space in which electric and magnetic fields oscillate in synchronism for no obvious reason. But it can be helpful to consider as well the sources of waves, and the effects that they subsequently have as they are detected: for electromagnetic waves, the overall process is simply that moving charges influence other charges through a retarded Coulomb interaction. All waves have a cause and effect, and their explicit inclusion can help to clarify otherwise mysterious phenomena.

1.3.1 Electromagnetic waves

Consider the situation in Fig. 1.2, in which two equal but opposite charges $\pm q_1$, separated by a distance $2a_0$, form a *dipole* that rotates slowly, with angular frequency ω, about its centre. This is our 'transmitter' or wave source. Some distance away, our 'receiver' (or detector) is a single charge q_2. Both charges of the dipole will exert a Coulomb force on q_2; in each case, the magnitude will be $q_1 q_2/(4\pi\varepsilon_0 r^2)$, where $r = r_1, r_2$ is the distance between the test charge and the dipole charge; and the direction, whether attractive or repulsive, will be along the line joining the charge centres. If $a_0 \ll r_0$, where r_0 is the distance between the test charge and the dipole centre, then the horizontal component of the force due to $+q_1$ will cancel out that due to $-q_1$. We're therefore left with just the vertical components, which will add. We shall work out shortly the magnitude of this force, but for now it suffices to understand how it arises.

C.-A. de Coulomb [14]

As the dipole rotates slowly about its centre, the vertical – or *transverse* – force on q_2 will grow and fade, and the oscillatory motion of the charges in the dipole is thereby transmitted to the other charge. The rotating dipole here represents the flow of charges in the antenna of a radio transmitter and, at a different scale, the variation of the charge distribution in the electron 'cloud' of a radiating atom. So far, then, we've considered nothing more than the electrostatic Coulomb interaction between separated charges.

For the rotating dipole of Fig. 1.2 $a(t)$ may be written explicitly in the form $a(t) = a_0 \cos(\omega t + \varphi)$, where a_0, ω and φ characterize the specific example. But we do not need to be so specific, and therefore write in terms of a general *function* $a(t)$ that indicates only that the vertical distance a depends upon the time t.
The function $a(t)$ should not be confused with the simple product at.

The transverse force component is found by multiplying the individual Coulomb forces F_1 and F_2 by $\sin\alpha(t)$, where $\sin\alpha(t) \approx a(t)/r_0$ and $a(t)$ is the vertical component of a_0 at time t, and adding them to yield the net transverse force $F(t)$ experienced by q_2. With a little working, and assuming $a_0 \ll r_{1,2}$ and $r_{1,2} \approx r_0$, this thus turns out to be

$$F(t) = \frac{2q_1 q_2}{4\pi\varepsilon_0 r_0^3} a(t). \tag{1.1}$$

Rather than characterize the effect of the dipole by the *force* it exerts upon a specific charge q_2, we consider the strength of the electric **field**, which is simply the force exerted upon the test charge divided by the value of the test charge q_2 itself. We may therefore write

$$E(t) = \frac{F(t)}{q_2} = \frac{2q_1}{4\pi\varepsilon_0 r_0^3} a(t). \tag{1.2}$$

We now increase the speed at which the 'transmitter' dipole rotates, and the theory of relativity imposes a time lag or *retardation* on the force caused by the dipole: any force caused by the charges of the dipole is delayed by a time r_0/c, so that q_2 sees the dipole as it was a time r_0/c earlier. When we take this into account, our expression for the electric field experienced by q_2 at a time t becomes

$$E(t) = \frac{2q_1}{4\pi\varepsilon_0 r_0^3} a\left(t - \frac{r_0}{c}\right) \tag{1.3}$$

and we thus have our wave: a function that depends upon what was happening at another position some time previously.

Feynman [22] Chapters 1 and 2

This is how all waves propagate. We can add a medium – other charges between q_1 and q_2 – so that the disturbance can also travel from q_1 to q_3 (say) and then from q_3 to q_2, and this turns out to be the origin of the refractive index of a material and so on. But the basic principle is that what happens at one point in space has an influence upon what will happen elsewhere, with a time lag because of the finite speed of propagation. Waves, then, are what happens to a system of forces when we move things around quickly enough for delays to be significant.

You may have spotted that the description here of electromagnetic radiation isn't complete: it gives an oscillating electric field, but no accompanying magnetic field; and the electric field decreases with the cube of the distance from the dipole, which, although correct at small distances, does not account for the generally more important field that varies as r_0^{-1} and allows the radiation of energy. This is because, while we introduced the time lag required by the theory of relativity, we omitted to perform the Lorentz transformations that take into account the speeds of the moving charges, and to include the variation in retardation as the individual dipole charges move towards and away from the test charge.

We shall see in Chapter 19 that a full calculation introduces important corrections to the electric field. Some can be simply written in terms of the motions of the dipole charges, and correspond to radiated electric fields that vary with the inverse or inverse square of the distance. Other terms turn out to be non-zero only when q_2 is also moving; these yield the radiated magnetic field components. Magnetism, it turns out, is simply the relativistic correction to the electrostatic Coulomb force.

Billingham & King [7]
pp. 210–212
D. H. Frisch & L. Wilets [30]

1.3.2 Macroscopic definition of a wave

The example of electric dipole radiation allows us to offer a second definition of a wave, as

> a time-dependent feature in the field of an interacting body, due to the finite speed of propagation of a causal effect.

While this is more of a *description* of a wave than an indication of how it arises, it is both a useful summary of how waves are manifest and a reminder that they always ultimately emanate from some form of source.

At a more abstract level, waves may be regarded as dynamic solutions to a time-dependent field theory. Just as simple systems of interacting particles, which in static equilibrium would be stationary, more generally show oscillatory motions, so the steady-state fields of stationary charges or masses turn into waves in the more general dynamic case.

1.3.3 Gravitational waves

Just as electromagnetic waves result from the retardation of an oscillating electrostatic force, so variations in a gravitational force can be manifest in gravitational waves. The distances involved are, naturally, cosmological.

Figure 1.3 shows a rather hypothetical scenario in which we have the ability to move planets. One planet, of mass m_1, can be moved in an oscillatory or rotary motion of amplitude a_0, and we are interested in the effect of this

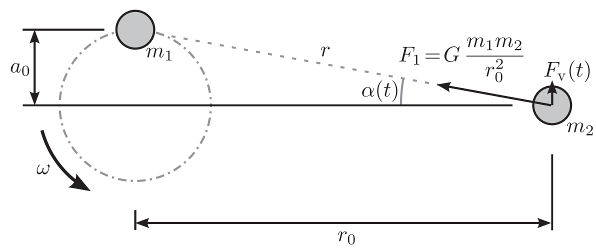
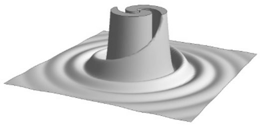


Fig. 1.3 Gravitational interaction between a moving mass m_1 and a test mass m_2.

A much-considered source of gravitational waves is a coalescing binary star. In a typical example, neutron stars, each of 1.4 solar masses, are separated by a few tens of kilometres, and rotate about their common centre of mass several times a second. Coalescence occurs within a few minutes because the stars lose energy and angular momentum through gravitational radiation with spiral wavefronts as illustrated below:



Such mind-boggling systems are the favoured sources to be observed by extremely sensitive (and expensive) gravitational-wave detectors currently being built at various sites world-wide and planned for space.

For details of laser interferometric gravitational-wave detection see [89–91].

motion upon a test planet of mass m_2. We shall concern ourselves only with the vertical component of the force experienced by m_2, for the large horizontal component will be largely constant. As shown, the gravitational force is given by the usual formula to be inversely proportional to the square of the distance r_0 between the masses. We shall assume that $a_0 \ll r_0$.

The vertical component of the gravitational force is given by

$$F_v(t) = F \sin\alpha(t) = F\frac{a(t)}{r_0} \tag{1.4}$$

and leads to an apparent gravitational field at m_2 with a vertical component g_v given by $m_2 g_v = F_v(t)$, and hence

$$g_v = G\frac{m_1}{r_0^3}a(t). \tag{1.5}$$

Because of the finite speed of gravitational propagation, however, the effect is retarded, and depends not on the current position of planet m_1 but upon that seen by an observer on planet m_2 – just as with the electromagnetic wave:

$$g_v(r,t) = G\frac{m_1}{r_0^3}a\left(t - \frac{r_0}{c}\right). \tag{1.6}$$

In reality, a motion such as that of m_1 must be due to rotation about another heavenly body, as shown in Fig. 1.4. The second mass has the unfortunate effect of cancelling out, to first order, the gravitational field oscillations produced by the first. The result is that only higher-order, *quadrupole* radiation components may be detected. Their calculation requires a slightly more careful evaluation.

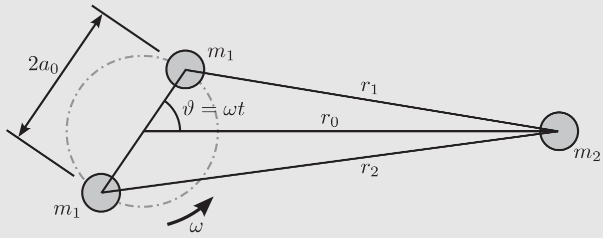


Fig. 1.4 Gravitational interaction between a rotating pair of masses m_1 and a test mass m_2.

By application of the cosine rule, $r_{1,2}^2 = r_0^2 + a_0^2 \pm 2a_0 r_0 \cos\vartheta$, where ϑ defines the orientation of the pair of masses with respect to the test mass, the net vertical component of the gravitational field is found to be

$$g_{\text{v}} = G\frac{m_1 a_0 \sin\vartheta}{r_0^3}\left[\left(1+\frac{a_0^2}{r_0^2} - 2\frac{a_0}{r_0}\cos\vartheta\right)^{-1} - \left(1+\frac{a_0^2}{r_0^2} + 2\frac{a_0}{r_0}\cos\vartheta\right)^{-1}\right]$$
$$= G\frac{2m_1 a_0^2}{r_0^4}\sin(2\omega t). \tag{1.7}$$

With the caveat that, as for electromagnetic waves, we have here calculated only the short-range component, it is this field that gravitational-wave observatories are intended to detect.

1.3.4 The æther

In describing the propagation of electromagnetic and gravitational forces, we have tiptoed straight through the middle of one of the greatest controversies of the nineteenth century: how can objects influence each other through a vacuum, unless there is an *æther* – an intangible fluid – that connects them? By the early twentieth century, the existence of an æther had been all but disproved: Maxwell and Einstein had formulated elegant descriptions of the propagation of electromagnetic and gravitational fields and waves that made no reference to the presence or properties of any æther; and the results of various experiments – including those of Michelson and Morley [63, 64] – made any theory of an æther increasingly untenable. In the twenty-first century, then, we are generally comfortable that static forces may be transmitted through a vacuum.

But what of waves? The answer here must be the same, yet there are grounds for disquiet that we should mention, even though they will then be dismissed. The first is the concept of point-to-point propagation, which will become the basis of the Huygens description of wave motion: we tend to imagine photons travelling in straight lines through space, yet diffraction suggests that they can take other routes even when there is nothing to deflect them. The second is the question of radiation. The motions of charges and masses can be damped by transferring their energy to other charges or masses, in a process that is extremely clear under Newtonian mechanics; yet the process of radiation means that the energy must leave the motion before it arrives at the destination charges or masses – and, indeed, without apparently even requiring their existence!

We shall not offer here any answer to this conundrum, but there are some comments that can be made. First, the problem lies not with the wave treatment but with the propagation of the forces in themselves – so the question goes beyond the scope of this book. Secondly, the problem is one of our imagination: physics has well-tested models that account very accurately for how the forces propagate; our problem is the more philosophical 'why?' – so, arguably, it's not even a question for physics! Summarized another way: we have very good descriptions of how electromagnetic and gravitational forces work, and the propagation of electromagnetic and gravitational waves is simply a consequence that we derive from those initial descriptions. Neither the forces nor their wave propagation has yet presented any inconsistency, either with itself or with experimental observation. So, if we find it hard to imagine, that's a defect with our imagination. But it shouldn't stop us asking interesting questions!

1.4 Examples of wave disturbance

In the examples above, we have considered two wave motions in which the propagated property or **disturbance** is a force or displacement *perpendicular* to the direction of wave propagation. It is common to refer to these as *transverse* waves, and to distinguish them from *longitudinal* motions that are *parallel* to the direction of wave propagation. To end this chapter we consider the validity of such categorizations for some common examples of wave motion.

Transverse wave motions, including electromagnetic and gravitational waves, are characterized by the propagation or measurement of a transverse physical displacement; the disturbance, in other words, is a vector property directed perpendicular to the direction of wave propagation. As well as transverse waves in vacuum, there are many examples in what may be considered continuous media. The transverse motion of a guitar string, for example, is considered to result from the curvature of the taut string, so that the tension does not pull in exactly opposite directions on opposite sides of any given point. In this category, we might also consider surface water waves, due to the transverse components of surface tension or gravitational forces, and mountain lee waves in the atmosphere. The skin of a drum is a two-dimensional version of the guitar string; and surface acoustic waves used in optoelectronic devices resemble ocean waves with the bulk elasticity of the medium playing the role of gravity.

Longitudinal waves are those in which the physical displacement or disturbance is once again a vector quantity, but is parallel to the direction of propagation. The classic example here is sound, and the motion results because the pressures either side of a given region are not equal. The pulse of our blood flow may be considered a form of sound wave, as may the pulses observed in the exhausts of jet engines.

There is no reason why a vector disturbance should be aligned parallel or perpendicular to the propagation direction, however: it is possible to generate longitudinal electromagnetic and gravitational waves, and when magnetization propagates as a *spin wave* it can have transverse and longitudinal components.

Terrestrial and solar seismology similarly involve manifestations of both longitudinal and transverse wave motion.

The wave disturbance need not be a physical force or displacement; it could be a scalar quantity, or a vector unrelated to the propagation coordinates. Temperature, for example, may propagate as thermal waves; and the spatially dependent concentrations of chemical reactants and products can change as chemical waves travel through a *reaction–diffusion* system such as the human heart, or as a flame front travels through a flammable material. Perhaps the most fascinating example of a propagated property that is not a displacement is the quantum wavefunction – the mysterious quantity that, according to quantum mechanics, contains everything we can know about a particle. Quite what the quantum wavefunction is, and what it means, are profound questions that must be left to other texts; but the nature of its propagation is identical to that of the other wave manifestations that we'll study here.

One could reason that thermal waves propagate by the *longitudinal* flow of heat, or argue that sound waves may be described by the scalar property of pressure. The simple characterization into transverse, longitudinal and scalar waves is therefore not always straightforward. Fortunately, such labels are really only for convenience; there are few physical consequences, and we shall see that in specific cases an examination of the propagation mechanisms resolves any doubt as to the phenomena observed.

So what of the waves of emotion and protest mentioned earlier? A key aspect of wave propagation is that the disturbance should propagate from point to point in a causal fashion, rather than simply reflect the staggered arrival times at adjacent points via independent routes. So, provided that the fear or protest of each person is inspired by the fear or protest of a neighbour – and social scientists can be reasonably clear in identifying such mechanisms – it may indeed be valid to regard the propagation of such properties as a wave. That the medium through which the wave propagates is composed of people who are discrete, nonlinear and to some extent irreproducible does not differ fundamentally from many granular, nonlinear and noise-ridden examples of more classical physical systems. The wave description may not only serve as a useful shorthand to imply the neighbour-to-neighbour-mediated propagation: subject to the natural imprecision of such systems, it may even allow their mathematical simulation and prediction.

Exercises

The following exercises do not directly concern wave propagation but address the mathematical techniques used in subsequent chapters. Substitute a few arbitrary values for a simple numerical check.