The University Of Sheffield.

Department of Physics and Astronomy

PHY221

Classical Physics

Dr Rhoda Hawkins

Autumn Semester

Contents

Preamble

Acknowledgements

These lecture notes are heavily based on lecture notes developed by Dr Martin Grell who taught this course before me. I am very grateful to him for sharing his material with me.

Your input

Please report any mistakes you find and feedback any comments you have to me so I can improved these notes. Thanks!

Recommended reading

- Fowles “Analytical Mechanics”, or the newer edition, Fowles and Cassiday, “Analytical Mechanics”
- Goldstein “Classical Mechanics”
- Pain “The Physics of Vibrations and Waves”
- Boas “Mathematical Methods in the Physical Sciences”

Introduction

What is "Classical Physics"?

The easy answer is "anything that's not quantum!". In this course we also exclude relativity and therefore this course could be called pre-20th century physics. Relativity is included in some definitions of classical physics although it is not classical mechanics (Newtonian mechanics). We exclude it in this course along with optics and thermodynamics, since these are covered in other courses.

It is worth starting by stating the assumptions on which this course is based:

- 3-dimensional space with no curvature.
- Time, t, is absolute i.e. it ticks away in the same way for everyone.
- Space is where physical events occur and, separately, time is when they occur but time and space are not themselves involved in the events.
- There is no limit to the velocity a body can have, and its mass is independent of its velocity.
- Both the position and momentum of all bodies are simultaneously precisely defined.
- Bodies interact only through forces.

Exercise. *What quantum mechanical interactions between particles are NOT forces?*

When are these assumptions justified?

- Bodies (mass and size) $\gg$ elementary particles

- Velocities $\ll$ speed of light

In practice they are valid for much of our everyday world e.g. all of mechanical engineering.

Topic 1

Harmonic oscillators

1.1 Simple harmonic oscillator

A simple harmonic oscillator is sometimes called a linear harmonic oscillator. Oscillations are common in both classical and quantum mechanical systems. Here we will discuss mechanical oscillations, but the concepts developed can be generalised to e.g. electrical oscillators. All oscillators contain elements that can store and release energy, e.g. springs (stores potential energy) and masses (stores kinetic energy) or capacitors (store electrical energy) and coils (store magnetic energy). Simple harmonic oscillators will go on forever. Realistically, however, oscillators will also dissipate energy, something we will discuss later (section 1.2)

Let us consider the simple example of an oscillator of a body of mass m attached to an ideal spring, with the mass being able to move only in the direction of the spring's long axis. We choose to call this axis the x-axis. Such an oscillator is called a simple or linear harmonic oscillator. An "ideal" spring is a spring that has zero mass of its own, and responds to stretching or compression away from its equilibrium length with a restoring Force, $F_{\text{res}} = -k(x - x_0)$ pointing back towards the equilibrium point, x_0. This is Hooke's law. k is known as the "spring constant", and is a characteristic of the spring.

We may choose to put the origin of the coordinate system we are using at any point we find convenient — say at x_0. (Note we do NOT have to choose the point where the spring is anchored as the origin!). We can therefore always, without loss of generality, say $x_0 = 0$, and $F_{\text{res}} = -kx$.

Note that the "spring" is a model for many kinds of restoring forces. Why do we assume restoring forces are linear? i.e. why $F_{\text{res}} = -kx$ and not another function of x? To answer this question let us take a mass m

that is initially at rest at x_0 at a local minimum of a general potential energy $V(x)$. As above, we can choose the origin of our coordinate system so that $x_0 = 0$. We may also set the potential energy so that $V(x_0 = 0) = 0$. This is because force is $F = -\frac{\mathrm{dV}}{\mathrm{dx}}$ so adding or subtracting any constant V_0 to the potential will not change the force. Our question is now, what restoring force will the mass experience when it is displaced for a small distance from $x = x_0 = 0$? To answer that, we use the first few terms of a Taylor expansion of the potential energy around its local minimum at $x_0 = 0$ as an approximation when the displacement x is small:

$$V(x) = V(0) + x\frac{\mathrm{dV}}{\mathrm{dx}}\bigg|_{x=0} + \frac{1}{2}x^2\frac{\mathrm{d^2V}}{\mathrm{dx^2}}\bigg|_{x=0} + \dots \tag{1.1}$$

where $V(0)$ is a constant so we can set it to zero as described above. $\frac{\mathrm{dV}}{\mathrm{dx}}\Big|_{x=0}$ is zero since $x = x_0 = 0$ is a minimum so the first derivative is zero there. Therefore, for small displacements x, any potential energy is quadratic in x. A potential energy that scales quadratically with deflection is called "harmonic". For small amplitudes all oscillations are harmonic. The potential V and force F are given by:

$$V = \frac{1}{2}x^2\frac{\mathrm{d^2V}}{\mathrm{dx^2}}\bigg|_{x=0}$$
$$F = -\frac{\mathrm{dV}}{\mathrm{dx}} = -kx$$

where the spring constant k is given by the second derivative of the potential at the minimum, $k = \frac{\mathrm{d^2V}}{\mathrm{dx^2}}\Big|_{x=0}$.

Now, we apply Newton's second law $F = ma$ to get the equation of motion for our system giving:

$$m\frac{\mathrm{d^2x}}{\mathrm{dt^2}} = -kx$$
$$\frac{\mathrm{d^2x}}{\mathrm{dt^2}} + \frac{k}{m}x = 0 \tag{1.2}$$

This is a second order, linear, homogeneous equation. Since it is second order there will be 2 linearly independent particular solutions and the general solution will be a linear combination of these 2 solutions.

Exercise. *Show that $x(t) = A\cos(\omega_0 t)$, and $x(t) = B\sin(\omega_0 t)$ are solutions of equation (1.2). What are the dimensions of A and B?*

ω_0 is the angular frequency of the harmonic oscillator and it is given by

$$\boxed{\omega_0 = \sqrt{\frac{k}{m}}}$$

Exercise. *Use dimensional analysis to derive $\omega_0 = \sqrt{(k/m)}$. Initially, consider the possibility that ω_0 may depend on amplitude $x_{\max}$ as well as k and m — dimensional analysis will prove that it does not.*

Now we have 2 particular solutions, we can write the general solution of the harmonic oscillator:

$$x(t) = A\cos\omega_0 t + B\sin\omega_0 t \tag{1.3}$$

which is mathematically equivalent to

$$x(t) = x_{\max}\cos\left(\omega_0 t + \phi\right) \tag{1.4}$$

where ϕ is called the phase.

Exercise. *Show that the above equations (1.3) and (1.4) are equivalent, and show how A and B are related to $x_{\max}$ and ϕ.*

Note that equation (1.4) is more useful, as it allows us to directly read off the amplitude $x_{\max}$ of the oscillation. The amplitude $x_{\max}$ depends on how much energy the oscillator has when it starts. The simple harmonic oscillator then continues to oscillate forever with this amplitude $x_{\max}$ and phase ϕ. The oscillations repeat with frequency $f = \frac{\omega_0}{2\pi}$ and period $T = 1/f$. Oscillators are clocks. During the oscillation, energy is converted backwards and forwards between potential and kinetic energy, with maximum potential energy and zero kinetic energy when $x = x_{\max}$, and maximum kinetic energy and zero potential energy at $x = 0$. The maximum velocity $v_{\max}$, maximum acceleration $a_{\max}$, maximum kinetic energy $T_{\max}$ and maximum potential energy $V_{\max}$ are given by:

$$v_{\max} = \omega_0 x_{\max} \tag{1.5}$$

$$a_{\max} = \omega_0^2 x_{\max} \tag{1.6}$$

$$T_{\max} = \frac{1}{2}mv_{\max}^2 = V_{\max} = \frac{1}{2}kx_{\max}^2 \tag{1.7}$$

Exercise. *Derive (1.7) from (1.4).*

Note that the phase angle, ϕ, is absent from (1.7). In fact, since we can choose any point in time as the beginning $t = 0$, we can choose $x = x_{\max}$ at $t = 0$ and then $\phi = 0$. i.e. if we start by pulling to maximum amplitude and then let go the phase $\phi = 0$. For a single oscillator the phase ϕ is a rather meaningless concept. It becomes meaningful only when we compare two oscillators, which may or may not be in step (in phase) with each other.

Another common example of simple harmonic motion is that of a pendulum. For a pendulum with a mass m and length of string l oscillating at small angles, Newton's second law gives us:

$$\begin{aligned} ml\frac{\mathrm{d}^2\theta}{\mathrm{dt}^2} &= -mg\sin\theta \\ \frac{\mathrm{d}^2\theta}{\mathrm{dt}^2} + \frac{g}{l}\theta &= 0 \end{aligned} \tag{1.8}$$

where we have used the small angle approximation $\sin\theta \approx \theta$. In this case the angular frequency is given by

$$\boxed{\omega_0 = \sqrt{\frac{g}{l}}}$$

1.2 Damped harmonic oscillator

Real oscillations don't go on forever like the simple harmonic oscillator does. Real oscillators dissipate energy due to damping. Damping causes energy loss due to friction. Note damping is often a good thing (e.g. in a closing door, shock absorbers, suspensions). We model damping with a dashpot but this can represent different types of damping e.g. air resistance or electrical resistance in different types of oscillators. The friction force F_f is assumed to point in the opposite direction to the velocity v and to have the form:

$$F_f = -cv \tag{1.9}$$

where c is a constant (NB it is not a velocity!). Unlike for the restoring force ($F_{\text{res}} = -kx$) where there was a good justification, $F_f = -cv$ is not well justified. In general we should write $F_f = -cv^n$. There are different types of friction known, with different laws of friction with different powers n. Stokes friction (slow movement in highly viscous medium) does have $n = 1$, but there are other friction laws. Coulomb friction (friction of a dry body on dry surface) has, $n = 0$ as you may know from the friction law $F_f \leq \mu F_N$ where μ is the coefficient of friction and F_N is the normal force.

Newton friction (fast movement in low viscosity medium, $n = 2$) and a final example is that of Reynolds friction (between lubricated solid bodies, $n = 1/2$).

For now, we will stick with $F_f = -cv$ for our further discussion. We have to add this force due to friction into the oscillator's equation of motion, leading to the following differential equation of motion for the damped harmonic oscillator:

$$m\frac{\mathrm{d}^2\mathrm{x}}{\mathrm{dt}^2} = -kx - c\frac{\mathrm{dx}}{\mathrm{dt}}$$
$$\frac{\mathrm{d}^2\mathrm{x}}{\mathrm{dt}^2} + 2\gamma\frac{\mathrm{dx}}{\mathrm{dt}} + \omega_0^2 x = 0 \tag{1.10}$$

where we define the damping factor $\gamma = \frac{c}{2m}$, and as before, $\omega_0^2 = k/m$. It looks a bit odd at first to define $\gamma = \frac{c}{2m}$, and then have 2γ in the equation, but you'll see why we do this later.

Exercise. *Looking at equation (1.10), why did we insist in the assumption $n = 1$ even though this is not true for all types of friction?*

As long as $n = 1$, equation (1.10) is a linear differential equation. For $n \neq 1$ the equation becomes nonlinear and it may not be possible to solve it analytically. $\sin$ and $\cos$ are not solutions to equation (1.10) so we solve the equation using the general method for solving linear homogeneous differential equations, which is to guess a trial solution (an "ansatz", which means an educated guess) of the form

$$x(t) = Ae^{\alpha t} \tag{1.11}$$

Exercise. *Can you remember how exponentials relate to sin/cos?*

We put the trial solution into equation (1.10) to find what is known as the characteristic equation:

$$\alpha^2 + 2\gamma\alpha + \omega_0^2 = 0 \tag{1.12}$$

Note that the characteristic equation is no longer a differential equation, but a conventional ("algebraic") equation. The characteristic equation is always of the same order as the differential equation was, here the characteristic equation is quadratic because the differential equation is 2nd order. Equation (1.12) can be solved by the standard method for quadratic equations, giving the 2 solutions:

$$\alpha = -\gamma \pm \sqrt{\gamma^2 - \omega_0^2} \tag{1.13}$$

By looking at (1.13) we can see that the α's may be complex (if $\omega_0 > \gamma$) and in which case the 2 solutions will be complex conjugates of each other. The solutions (1.13) of equation (1.10) may therefore be of different types depending on whether γ^2 is larger, equal to or less than ω_0^2. Since $\gamma = \frac{c}{2m}$ and $\omega = \sqrt{k/m}$ this is equivalent to saying whether c^2 is larger, equal to or less than $4mk$.

1.2.1 Overdamping $\gamma^2 > \omega_0^2$

The first case we will consider is that of $\gamma^2 > \omega_0^2$, or equivalently $c^2 > 4km$. In this case friction/damping is greater than the energy stored in the oscillations (characterised by k and m). In this case both roots (1.13) of the quadratic equation (1.12) are real and they are both negative. This case is known as "overdamping". Damping is so strong that the "oscillator" no longer oscillates, as you will see from entering (1.13) into (1.11) leading to:

$$x(t) = A_1 e^{\alpha_1 t} + A_2 e^{\alpha_2 t} \tag{1.14}$$

where α_1 and α_2 are the 2 solutions given in (1.13). Since both $\alpha_1 < 0$ and $\alpha_2 < 0$, the exponentials decay to zero for large times. The only unknowns in the general solution (1.14) are A_1 and A_2, which have to be found from the specific initial conditions of the system.

Exercise. *Show that for the initial conditions $x(0) = x_0$, and $v(0) = 0$ A_1, and A_2 are given by $A_1 = \frac{\alpha_2 x_0}{(\alpha_2 - \alpha_1)}$, and $A_2 = \frac{\alpha_2 x_0}{(\alpha_2 - \alpha_1)}$.*

Note that if the initial velocity $v(0) = 0$, then $x(t)$ never changes sign: $x(0)$ is the largest deflection in modulus the system will ever have, from then on, it decays, but it never changes sign. If $v(0) \neq 0$, $x(t)$ may change sign once, but no more than once. The overdamped "oscillator" does not oscillate. The system "creeps" to zero. This is the case most different from the original, undamped oscillator we had discussed first. Friction, quantified by c, dominates over energy storage, quantified by k and m.

1.2.2 Weak damping $\gamma^2 < \omega_0^2$

The second case we'll consider is that of $\gamma^2 < \omega_0^2$, or equivalently $c^2 < 4km$. In this case damping is weak compared to the energy of the oscillator. Now, there is a negative number under the square root in (1.13). Consequently, the 2 solution α_1 and α_2 are complex conjugates ($\alpha_2 = \alpha_1^*$) with

equal real parts $\mathrm{Re}[\alpha_1] = \mathrm{Re}[\alpha_2] = -\gamma$. We write these solutions as

$$\alpha = -\gamma \pm i\omega_d$$

where we introduce the quantity ω_d as the angular frequency of the damped harmonic oscillator with the subscript d for "damped":

$$\boxed{\omega_d = \sqrt{\omega_0^2 - \gamma^2}} \tag{1.15}$$

and therefore the general solution is

$$x(t) = e^{-\gamma t}(A_+ e^{i\omega_d t} + A_- e^{-i\omega_d t}) \tag{1.16}$$

$x(t)$ has to be real but this is true as long as $A_- = A_+^*$ i.e. A_+ and A_- are complex conjugates. There are then only 2 independent unknowns in A_+ and A_- so 2 initial conditions will be sufficient to determine them. It is clear that (1.16) is real from its mathematically equivalent form:

$$x(t) = x_{\max} e^{-\gamma t} \cos(\omega_d t + \phi) \tag{1.17}$$

where $x_{\max}$ and ϕ can be related to A_+ and A_-.

Exercise. *Derive the relationship between $x_{\max}$ and ϕ and A_+ and A_-.*

(1.17) describes an oscillation, similar to the undamped oscillator (1.4). The angular frequency of this oscillation is $\omega_d < \omega_0$, smaller than the angular frequency of the corresponding undamped oscillator ω_0. For very weak damping, $\gamma \ll \omega_0$, ω_d is very close to ω_0. Again, if we assume oscillation starts at maximum amplitude and zero velocity, then the phase $\phi = 0$. The main difference between (1.4) and (1.17) is that in (1.4), the oscillator amplitude always remains at $x_{\max}$, because in the absence of damping, the oscillator doesn't lose energy. The amplitude of the damped oscillator, described by (1.17), decays over time with a time constant $\tau = 1/\gamma$, i.e. $x_{\max}(t) = x_{\max}(0)\exp(-t/\tau)$, because the oscillator loses energy due to damping. We have an oscillation with somewhat lower angular frequency, oscillating within an exponential decay "envelope". This is quite different from the decay observed in the overdamped case where there were no oscillations. When just referring to a "damped oscillator", we usually mean weak damping described by (1.17).

1.2.3 Critical damping $\gamma^2 = \omega_0^2$

There is a 3rd case, which we call critical damping for which $\gamma^2 = \omega_0^2$ and $c^2 = 4mk$. In this case the dissipation (c^2) and the stored energy ($4mk$) are exactly balanced. Now there is only one root in (1.13), which is real and negative, $\alpha_1 = \alpha_2 = -\gamma = \omega_0$. This is sometimes called **degeneracy**. The characteristic equation give only one particular solution, but we need two particular solutions to make up the general solution. In this case the general solution is given by:

$$x(t) = (At + B)e^{-\gamma t} \tag{1.18}$$

Again, there are no oscillations, only decay towards zero. The time it takes to reach zero is as short as possible without oscillations. The time constant $\tau = 1/\gamma = 1/\omega_0$. Critical damping is the fastest return to equilibrium possible without oscillations. Critical or near-critical damping is often useful, e.g. vehicle suspension systems. Wheels are linked to springs to soften the blows from potholes etc. However, with springs alone, your car/bike would soon hop along the road like a bouncy ball. Therefore, in parallel to the springs, vehicles have shock absorbers, that is dashpots with damping. Shock absorbers should be large enough (c large enough) to stop your vehicle oscillating but the vehicle should creep back to the equilibrium position quickly. Ideally, therefore, you should be precisely at critical damping. Since the mass of the vehicle may change, engineers tend to err on the safe side and somewhat overdamp the suspension, but when you overload the vehicle you may cross the critical boundary: Overloaded cars tend to swing a few times after going over a pothole.

What about the other friction laws we talked about? If the friction is not linear in the velocity but has a power $n \neq 1$ of velocity the resulting differential equation is no longer linear, and it it difficult to solve. In the case of weak damping, approximate solutions to the nonlinearly damped oscillator can be found. Such an oscillator will still undergo decaying harmonic oscillations, but the decay envelope is not exponential. Table 1.1 shows a few examples.

Whatever the decay law, every practical oscillator is somewhat damped, and will not go on forever but will eventually decay to zero.

1.3 Driven harmonic oscillator

In practice all oscillators are damped and will not oscillate forever, unless we “drive” them. In this section we consider a driven (or “forced”) harmonic

n	**Shape of envelope**
0	Linear
$\frac{1}{2}$	Parabolic
1	Exponential
2	Hyperbolic

Table 1.1: Table showing the different shaped decay envelopes of damped harmonic oscillators for damping forces proportional to v^n, where v is the velocity.

oscillator. To drive the oscillator we need to add an external force F_{ext} to the equation of motion (1.10) giving:

$$\frac{\mathrm{d}^2\mathrm{x}}{\mathrm{dt}^2} + 2\gamma\frac{\mathrm{dx}}{\mathrm{dt}} + \omega_0^2 x = \frac{F_{\text{ext}}(t)}{m} \tag{1.19}$$

where we have divided F_{ext} on the right hand side by the mass m because we have written the left hand side in terms of accelerations not forces. Now we have a function of t (time) on the right hand side the equation is no longer homogeneous but is now inhomogeneous.

What function should $F_{\text{ext}}(t)$ be to drive the oscillator? What if $F_{\text{ext}}(t) = F_0 =$ constant? This turns out to be the same as a free (unforced/undriven) oscillator but now the oscillations are about a new equilibrium point $x_0 = F_0/k$ instead of $x_0 = 0$.

Exercise. *Show that for a constant external force, $F_{\text{ext}}(t) = F_0$, the solution of equation (1.19) is exactly the same as that for the damped free oscillator (equation 1.10), apart from the fact that the mass no longer oscillates around the origin (the equilibrium of the spring under no force), but around $x_0 = F_0/k$, the equilibrium of the spring stretched by the force F_0.*

Therefore a constant force will not force the oscillator to keep oscillating forever. What about if the force decreases with time? In this case at long times it will tend to a constant/zero force and therefore be just like a free damped oscillator at long times. What if the force increases with time? If the force increases with time it will eventually break the spring. So we need a force that is periodic in time to drive the oscillator. e.g.

$$F_{\text{ext}}(t) = F_0 e^{i\omega t} \tag{1.20}$$

This is a harmonic driving force with an arbitrary driving frequency, ω . NB ω is not the same as the free undamped oscillator frequency, ω_0 or that of the free damped oscillator ω_d. We choose the driving frequency ω, unlike

ω_0 and ω_d which are determined by the properties of the oscillator (i.e. by k, m, c). If you drive your oscillator by an external motor, you may choose ω, completely independently of k, m, c.

We choose harmonic periodic forces, as they are the most basic periodic function. The free oscillator undergoes harmonic motion, so it appears natural to drive it by one. In fact, the driver may be another oscillator. From Fourier's theorem we can write any periodic function as a superposition of harmonic oscillations of different angular frequency, amplitude, and phase. So, if we find a general solution for the driven oscillator equation in response to a harmonic external force, we can find the solution of the general problem for an oscillator driven by any periodic external force just by finding and adding the response of each harmonic component. Later, we'll see that due to resonance we can usually ignore most of the Fourier components apart from those near the resonance frequency. So, we describe the driven oscillator by the inhomogeneous, linear differential equation:

$$\frac{\mathrm{d}^2\mathrm{x}}{\mathrm{dt}^2} + 2\gamma\frac{\mathrm{dx}}{\mathrm{dt}} + \omega_0^2 x = \frac{F_0}{m}e^{i\omega t} \tag{1.21}$$

From your maths course you know that the general solution of an inhomogeneous differential equation is given by $x = x_c + x_p$ where x_c is the complementary solution which is the solution of the corresponding homogeneous equation (in this case that of the free damped oscillator, equation (1.10) given by (1.11) and (1.13)) and x_p is a particular solution of the full inhomogeneous equation (1.21). Physically, this means that whatever particular solution of the inhomogeneous equation we find, the harmonic oscillations of the free damped oscillator (with frequency ω_d) will be superimposed on it. However since the free damped oscillations are decaying with time, after long enough times ($t \gg \tau$ where $\tau = 1/\gamma$) these free oscillations will have died away, i.e. these free oscillations are transient. In the following we will assume our driven oscillator has been driven for a long time, $t \gg 1/\gamma$, so that all free oscillations have died away. Keep in mind, however, that shortly after switching the driver on, your oscillator may behave differently — we should let it settle down into its "steady state" first and then we simply have $x = x_p$ at steady state. To find x_p we use the ansatz (trial solution)

$$x(t) = A(\omega)e^{i(\omega t - \phi(\omega))} \tag{1.22}$$

i.e., we assume that the response of the harmonic oscillator to a harmonic driving force is a harmonic oscillation. Also we assume the resulting oscillator amplitude may depend on ω (in a way we'll work out below) and that there may be a phase shift ϕ between the driving force and the oscillator response. Unlike for the single free oscillator, for a driven oscillator a

nonzero phase ϕ is meaningful. Why do we assume the angular frequency of the response is equal to that of the driver (ω)? This must be the case if the driver and oscillator are connected by a stiff connection or else the rod connecting the driver to the oscillator would be stretched/broken.

We substitute the ansatz (1.22) into the equation of motion (1.21) giving

$$-A(\omega)\omega^2 e^{i(\omega t-\phi(\omega))} + 2\gamma A(\omega) i\omega e^{i(\omega t-\phi(\omega))} + \omega_0^2 A(\omega) e^{i(\omega t-\phi(\omega))} = \frac{F_0}{m} e^{i\omega t}$$

$$(-\omega^2 + 2\gamma i\omega + \omega_0^2) A(\omega) = \frac{F_0}{m} e^{i\phi(\omega)} \tag{1.23}$$

In the algebraic equation (1.23) there are 2 unknowns (A and ϕ) so how can we find them both from this one equation? By taking the real and imaginary parts separately (remember $e^{i\phi} = \cos\phi + i\sin\phi$).

$$\begin{aligned} \text{Re}: \quad & (-\omega^2 + \omega_0^2) A(\omega) = \frac{F_0}{m}\cos\phi(\omega) \\ \text{Im}: \quad & 2\gamma\omega A(\omega) = \frac{F_0}{m}\sin\phi(\omega) \end{aligned} \tag{1.24}$$

Dividing the imaginary parts by the real parts gives

$$\tan\phi = \frac{2\gamma\omega}{\omega_0^2 - \omega^2} \tag{1.25}$$

Remembering that $\cos^2\phi + \sin^2\phi = 1$ helps us find

$$A(\omega) = \frac{F_0/m}{\sqrt{(\omega_0^2 - \omega^2)^2 + 4\gamma^2\omega^2}} \tag{1.26}$$

Exercise. *In the limit $\omega \to 0$ the driving force becomes a constant, F_0. Show that (1.26) reproduces the previous result for constant force, i.e. $x_0 = A(0) = F_0/k$.*

At what ω is $A(\omega)$ maximum?

Exercise. *Find the maximum of $A(\omega)$ given in (1.26) by setting $\frac{\mathrm{dA}}{\mathrm{d}\omega} = 0$.*

The amplitude A is maximum when $\omega = \omega_r$ where:

$$\boxed{\omega_r = \sqrt{\omega_d^2 - \gamma^2}} \tag{1.27}$$

where we have used $\omega_d = \sqrt{\omega_0^2 - \gamma^2}$ from equation (1.15). The fact that $A(\omega)$ shows a peak is called **resonance** and the frequency at the maximum, ω_r, is called the **resonance frequency** hence the subscript "r" for

resonance. For weak damping, it is close to, but slightly smaller than, ω_d, the frequency of the free damped oscillator. Resonance is the key phenomenon in the physics of the driven oscillator, and therefore the driven oscillator is sometimes called a "resonator". Don't get confused between all the different ω's:

$$\begin{aligned}\omega_0 &= \text{frequency of free SHO no damping}\\ \omega_d &= \text{frequency of free damped oscillator}\\ \omega_r &= \text{resonance frequency}\\ \omega &= \text{driving frequency}\end{aligned}$$

To calculate the height of the amplitude peak, $A_{\max} = A(\omega_r)$ we just put (1.27) into (1.26) to give;

$$A_{\max} = A(\omega_r) = \frac{F_0}{2m\gamma\omega_d} \tag{1.28}$$

1.3.1 Quality factor, Q

A useful measure of how peaked the resonance peak is the quality factor Q. NB it is a property of the oscillator not the driver (and can therefore be calculated for a free damped oscillator too). The quality, Q, of an oscillator has several equivalent definitions. One definition is the ratio of the height of the amplitude peak at resonance compared to the amplitude at zero frequency (i.e. constant force):

$$Q = \frac{A(\omega_r)}{A(0)} = \frac{\omega_0^2}{2\gamma\omega_d} \approx \frac{\omega_0}{2\gamma} \tag{1.29}$$

where the approximation $\approx$ is valid for weakly damped oscillators, $\gamma \ll \omega_0$ for which $\omega_d \approx \omega_0$. At resonance the amplitude is Q times higher than when the same oscillator is subject to a constant force of the same magnitude. Since energy is proportional to amplitude squared, an oscillator at resonance has Q^2 times more energy than at zero frequency $\rightarrow$ the destructive power of resonance.

From definition (1.29), $Q \approx \frac{\omega_0}{2\gamma} = \frac{\sqrt{km}}{c}$ for a mass-spring system showing it is the ratio of the energy storing elements m and k to the energy dissipation c. However Q is a key general concept for all oscillators and allows comparisons between say mechanical and electrical oscillators. Another, general, definition of Q is

$$Q = \frac{2\pi \text{ energy stored}}{\text{energy lost in period } T} \approx \frac{2\pi\tau_{\text{energy}}}{T} \tag{1.30}$$

where τ_{energy} is the time constant for the decay in energy and the approximation is valid for weak damping. Since energy is proportional to amplitude squared $\tau_{\text{energy}} = \tau_{\text{ampl}}/2 = \frac{1}{2\gamma}$ so $Q = \frac{\omega_d}{2\gamma} \approx \frac{\omega_0}{2\gamma}$ for weak damping as in definition (1.29).

Since Q is a property of the oscillator not the driver it can be also calculated for a free damped oscillator from the properties ω_0 and γ given by k, m and c. However definition (1.30) leads to a way of calculating Q from freely decaying oscillations without knowing the properties of the oscillator (such as k, m and c):

$$Q \approx \frac{2\pi\tau_{\text{energy}}}{T} = \frac{\pi\tau_{\text{ampl}}}{T} = \pi N \tag{1.31}$$

where N is the number of oscillations until the amplitude decays to $1/e$ of the original amplitude. i.e. there are N oscillations until the transient oscillations die away and the oscillator reaches its steady state (zero if not driven). So Q can be counted for an electrical oscillator by counting N on an oscilloscope screen without needing to know the electrical properties (capacitors, inductances, and resistors) of the oscillator. In musical instruments like the guitar Q is know as the "sustain" - a high sustain corresponds to a long ringing after a string has been plucked.

1.3.2 Sharpness of the resonance peak

How sharp is the resonance? i.e. how wide is the peak? This tells us what range of frequencies around the resonance frequency will still give a large response. Quantitatively this is measured by the "Full Width at Half Maximum" (FWHM), $\Delta\omega_{\text{FWHM}}$. We will introduce this here in terms of the intensity peak rather than the amplitude since it leads to a nicer expression for the quality Q. The intensity, I, is proportional to the amplitude squared. We define the full width at half maximum frequencies, ω_{FWHM}, by

$$I(\omega_{\text{FWHM}}) = \frac{1}{2}I(\omega_r)$$
$$A^2(\omega_{\text{FWHM}}) = \frac{1}{2}A^2(\omega_r)$$

i.e. at ω_{FWHM} the intensity is half that at resonance. Substituting in (1.26) and (1.28) leads to the following solutions (after some tedious algebra and assuming weak damping so $\omega_r \approx \omega_d \approx \omega_0$)

$$\omega_{\text{FWHM}} \approx \omega_0 \pm \gamma \tag{1.32}$$

Oscillator	Q
Piano string	3000
Microwave cavity	10^4
Electron shell of atom	10^7
Nuclear γ-transition	$10^{12} \ldots 10^{13}$

Table 1.2: Table showing the quality Q for some typical oscillators.

and so the FWHM $\Delta\omega_{\mathrm{FWHM}}$ is given by

$$\Delta\omega_{\mathrm{FWHM}} \approx 2\gamma \tag{1.33}$$

where again the approximation is for weak damping. The definition of the quality factor in terms of the FWHM is

$$Q = \frac{\omega_r}{\Delta\omega_{\mathrm{FWHM}}} \approx \frac{\omega_0}{2\gamma} \tag{1.34}$$

i.e. a large quality Q corresponds to a narrow peak. Note that is we had defined the FWHM in terms of amplitudes instead of intensity (i.e. $A(\omega_{\mathrm{FWHMampl}}) = \frac{1}{2}A(\omega_r)$ this leads to $\omega_{\mathrm{FWHMampl}} \approx \omega_0 \pm \sqrt{3}\gamma$ and $Q = \frac{\sqrt{3}\omega_r}{\Delta\omega_{\mathrm{FWHMampl}}} \approx \frac{\omega_0}{2\gamma}$

Q is a universal and dimensionless measure that allows us to compare different types of oscillators e.g. mechanical oscillators, electrical oscillators, atomic spectra.

Can we force an oscillator to go at our chosen driving frequency ω rather than at its own natural frequency ω_d? Yes but for large Q the oscillator will only respond with a large amplitude if ω is near to ω_r i.e. near to ω_d assuming weak damping. The larger Q the closer you have to match the driving frequency to resonance to get a large amplitude. In practice Q is often very large. Table 1.2 shows a few typical Qs.

The extremely high Q (extremely narrow spectral lines) of nuclear γ-transitions are the basis of Mössbauer spectroscopy (Nobel price 1961), which is the most sensitive spectroscopy known. Because Q is so high for γ-transitions, Mössbauer spectroscopy can measure the tiny frequency shift from the energy loss of a γ-photon going against the Earth's gravity, thus confirming one of the predictions of general relativity. It can also measure the very small energy shifts that occur in the energy levels inside γ-active atomic nuclei resulting from different chemical bonds the respective atom may be engaged in.

1.3.3 Feedback

In a feedback loop, an oscillator is driven by a driver, as before, only now, the driving frequency is set by the oscillator itself not externally. It is therefore synchronised, and the system stabilises at the oscillator's resonance frequency ω_r with constant amplitude until the driver runs out of power. (NB Frequency stabilisation is not perfect, the tolerance is proportional to the width of the resonance curve, i.e. to $1/Q$). This is how clocks work.

An example is the "Accutron", the first electronic clock (1960s). A tuning fork that resonates at 360 Hz drives a mechanical clockwork that turns the hands of the clock. The fork is made of magnetic material and is itself driven by the AC magnetic field of two induction coils, which periodically receive a current pulse from a drive transistor. Feedback is facilitated by a third coil that acts as pickup. The vibrating fork induces an AC current in the pickup coil, which is driven into the transistor's base. In this way, the oscillator synchronises its driver.

Modern electronic clocks use an oscillating quartz crystal instead of a tuning fork. Via a phenomenon known as piezoelectricity, the quartz couples mechanical oscillations to an electric oscillator circuit, and stabilises the electric oscillator at the resonant frequency of the mechanical oscillations of the quartz. In terms of electronics, the drive and pickup of a quartz oscillator is capacitative (via electric fields), while for the Accutron, it is inductive (via magnetic fields).

Feedback is an extremely important phenomenon both in electrical engineering, and nature (biological clocks), and can get extremely unpredictable or chaotic when feedback is time-delayed (e.g. population dynamics in predator/prey ecosystems).

1.4 Coupled oscillators

Often, oscillators interact with each other, i.e. they are coupled e.g. atoms in a crystal. Atoms have clearly defined equilibrium positions, but due to thermal motion, will not sit still at these positions, but oscillate around them. A first approximation to thermal oscillations might be that atoms are bound to their equilibrium positions by harmonic forces, and indeed an early theory of heat capacity in crystals assumed just that (Einstein theory of heat capacity). However, if you think about it, atoms will rather be bound by harmonic forces to their neighbours, not their equilibrium positions – the "spring" will be a chemical bond, and the bond is between atom and atom, not between atom and lattice site. Hence, their oscillations will be

coupled. That led to the Debye theory of heat capacity based on coupled oscillations, which gives much better agreement with measured data.

Let us consider a very simple example of coupled oscillators: 2 harmonic oscillators with the same spring constant k, and mass m, which are connected together by a 3rd spring of spring constant k', parallel to the first two, linking the two masses. k' couples the two previously independent oscillators. We call x_1 and x_2 the displacements of each mass from their equilibrium positions (note we use different origins for x_1 and x_2 but call the same direction "positive". We assume one-dimensional motion along the springs only, and neglect damping. The motion of the coupled oscillators is described by the pair of coupled differential equations:

$$\begin{aligned} m\ddot{x}_1 &= -kx_1 + k'(x_2 - x_1) \\ m\ddot{x}_2 &= -kx_2 - k'(x_2 - x_1) \end{aligned} \tag{1.35}$$

where $\ddot{x}_{1,2} = \frac{\mathrm{d}^2 \mathrm{x}_{1,2}}{\mathrm{dt}^2}$. If $k' = 0$ (i.e. there is no middle spring) each equation is that of an independent oscillator. When the force due to the middle spring k' is added the equations become coupled (that is x_1 and x_2 both appear in each equation). Note it can be tricky to get the right signs here. To check the signs in the first equation, think about the effect of a positive x_1 displacement. This compresses the middle spring and the resulting restoring force will act to decrease x_1 (hence $-k'x_1$ on the right hand side) and increase x_2 (hence $+k'x_2$ on the right hand side). Similarly to check the signs in the second equation, think about the effect of a positive x_2 displacement. This will stretch the middle spring and the resulting restoring force will act to decrease x_2 and increase x_1 hence the sign of the k' term is the opposite from that in the first equation.

Although coupled, equations (1.35) are linear and can be solved by assuming the solutions are harmonic oscillators, i.e. the ansatz

$$x_{1,2} = A_{1,2}e^{i\omega t}$$

where the amplitudes A_1 and A_2 may be complex if there is a phase shift between the oscillators. We now substitute this ansatz into equations (1.35) to work out what ω is.

$$\begin{aligned} -\omega^2 m A_1 + kA_1 - k'A_2 + k'A_1 = 0 \\ -\omega^2 m a_2 + kA_2 + k'A_2 - k'A_1) = 0 \end{aligned}$$

Putting this into matrix form gives;

$$\begin{pmatrix} -m\omega^2 + k + k' & -k' \\ -k' & -m\omega^2 + k + k' \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{1.36}$$

The trivial solution of zero amplitude $A_1 = A_2 = 0$ is not interesting since this just means the oscillator is at rest. To find the interesting solutions we need to find when the determinant of the matrix is zero. i.e.

$$(-m\omega^2 + k + k')^2 - (k')^2 = 0 \tag{1.37}$$

which is a characteristic equation that is quadratic in ω^2 and since we know ω is positive we take the positive square root to find ω. This leads to $\omega^2 = (k + k' \pm k')/m$ and the 2 solutions are

$$\begin{aligned} \omega_S &= \sqrt{\frac{k}{m}} \\ \omega_A &= \sqrt{\frac{k + 2k'}{m}} \end{aligned} \tag{1.38}$$

The first solution, ω_S, is the same as for the uncoupled oscillator. We can substitute this solution back into equation (1.36) to work out that $A_1 = A_2$ but we can also see this physically. Since ω_S does not contain k' it is not being stretched or compressed and the 2 masses are oscillating in phase with the same amplitude without stretching the middle spring. We will call this solution the symmetric solution, hence the subscript S. The other solution, ω_A, contains k' so the middle spring will be stretch (twice as much as the k springs in fact due to the factor of 2). Substituting this into equation (1.36) gives $A_1 = -A_2$ confirming that the 2 masses oscillate $180°$ out of phase. This is called the antisymmetric solution, hence the subscript A. The frequency of the antisymmetric solution is higher, because now, the coupling spring contributes to the restoring force.

Together, the two solutions we have found for the coupled oscillator are known as **normal modes**. Remember that superpositions of solutions of linear differential equations are also solutions. Hence, the coupled oscillator can undergo a motion that is neither purely symmetric, nor purely antisymmetric but a combination of the two. But every motion of the oscillator can be broken down into these 2 components (the 2 normal modes). Normal modes are a bit like the unit vectors (or basis vectors) of a vector space - not every vector is a unit vector, but every vector can be expressed as a linear combination of the unit vectors. Normal modes are linearly independent (like basis vectors) and energy is not exchanged between normal modes - the normal modes never mix. If a motion of the coupled oscillator can be broken down at any point in time into, say, 70% symmetric and 30% antisymmetric, then it will always remain like that. Energy may be exchanged between the masses though.

Note that there may be many normal modes in the system (not just 2 like in our simple example). Examples include vibrations of molecules or crystals . In thermodynamics normal modes are called degrees of freedom. In quantum mechanics there are quantised normal modes called phonons. More complex examples include vibrations of polymers or biological molecules like proteins.

Symmetric modes are sometimes called “acoustic” modes and antisymmetric modes are sometimes called “optical” modes. This is because antisymmetric modes can have dipoles and therefore can absorb/emit radiation (usually infrared IR) but symmetric modes can’t. For example nitrogen N_2 only has a symmetric mode so cannot absorb radiation but carbon dioxide CO_2 has symmetric and antisymmetric modes so can absorb IR and cause global warming.

Topic 2

Waves

2.1 Introduction and definitions

What's the difference between an oscillation and a wave? An oscillation is a deflection/displacement that is periodic in time (oscillates in time). A wave is also a deflection that changes over time but unlike an oscillation it travels in space through the medium. A wave is a function of both time and space, i.e. $y = f(x, t)$. A "standing wave" is like a cross between the two as we will see later. If the deflection is parallel to the direction the wave travels (propagates) in, it is called a **longitudinal** wave. If the deflection is perpendicular (orthogonal) to the direction the wave travels, it is called a **transverse** wave. Note there are two directions orthogonal to any given direction of propagation, so there are two possible directions the distortion can be in. Which of these directions the distortion is in is known as the polarisation of the wave. (Note polarisation has more than one meaning even within physics. Here we are not talking about polarisation of electrical charges as in polarisation of a dielectric).

Let's illustrate a travelling distortion. At first, we assume that the shape of the distortion does not change as it travels, it only changes location. If the peak of the deflection is at $x = 0$ at $t = 0$, at some time later $t = \Delta t$ it is at a new location, Δx, but its shape is the same. The function $f(x, t)$ describing this deflection must be of either of the two forms:

$$\begin{aligned} y &= f(x - ct) \qquad \text{direction of travel positive } \rightarrow \\ y &= f(x + ct) \qquad \text{direction of travel negative } \leftarrow \end{aligned} \tag{2.1}$$

where c is the speed of the wave. i.e. f (called the wave function) is a function of x and t only in their combination $x \pm ct$. To see which sign designates which direction of travel, think about when you have $f(0)$. For

$f(x-ct)$ we get $f(0)$ at $x=t=0$ and at $x=ct$. So after time t the initial deflection $f(0)$ has moved to $x=ct$ so it has moved in the positive x direction (left to right). For $f(x+ct)$ however we get $f(0)$ at $x=t=0$ and at $x=-ct$. So in this case, after time t the initial deflection $f(0)$ has moved to $x=-ct$ so it has moved in the negative x direction (right to left).

Exercise. *Which one of the following describes a wave:* $f_1 = A/(x+ct^2)$, $f_2 = B/(x+ct)^2$, $f_3 = C\exp(-at)\sin(x+ct)$*?*

Mathematically, f is a function of only one variable ($x \pm ct$), not two (x and t). Physically, this means the wave travels in space and time. To observe a wave, we can take a snapshot at a fixed time of the wave over all space ($-\infty < x < \infty$). Or we can observe it at a fixed point in space, and let the wave wash over us for a long time ($0 < t < \infty$). Both of these observations give us the full range of $x \pm ct$ and therefore the full picture of the wave.

What are the dimensions of this combined variable $(x \pm ct)$? Since it has dimensions of length we cannot put it directly into a function like sine. So we have to multiply it by a constant, which we call the **wave number**, k, with dimensions L^{-1}. We also introduce the definition:

$$\boxed{\omega = ck} \tag{2.2}$$

ω is the angular frequency, like in oscillations. Actually, as we will discuss next week, usually the speed c is a function of k and therefore $\omega(k) = c(k)k$.

Now, we can write our wave function in the form $f(kx \pm \omega t)$, with the dimensionless group $kx \pm \omega t$. We call this dimensionless variable the **phase** $\phi = kx \pm \omega t$. Points in space that have equal phase are called **wave fronts**. The wave function has the same value along a wave front, because it has the same argument ϕ. Note, wave fronts are a meaningful concept in 2 or 3 spatial dimensions, not in 1D.

We now consider an example of a specific form of the wave function

$$f(kx - \omega t) = A\sin(kx - \omega t) \tag{2.3}$$

Waves don't have to be sinusoidal, in fact the deflection we sketched at the beginning of the lecture was a 'hump' with a single peak, unlike a sine wave. So why do we consider a sine wave here? As you are learning in your maths module, any function can be written as a sum (or 'superposition') of sine/cosine waves of different angular frequencies ω and amplitudes (Fourier). So if we can describe the propagation of sine/cosine waves, we can describe the propagation of any wave function by first decomposing it into its sine/cosine components (by Fourier analysis). Note

that if the components of the wave with different ω propagate with different velocities c this will lead to dispersion, which we will discuss next week.

Our example wave function (2.3) is periodic in 2π. Physically sine waves are periodic in space with a wavelength, λ, and periodic in time with a period, T. The spatial periodicity gives

$$k(x+\lambda) - \omega t = \phi + 2\pi = kx - \omega t + 2\pi \tag{2.4}$$

i.e. a phase shift of 2π corresponds to moving λ in space. The above equation gives the wave number k:

$$\boxed{k = \frac{2\pi}{\lambda}}$$

Note, despite being called a "number" the wave number has units! In 2D/3D the phase is given by $\phi = \boldsymbol{k}\cdot\boldsymbol{r} - \omega t$ where $\boldsymbol{k}$ is a vector called the wave vector. Its modulus is $|\boldsymbol{k}| = \frac{2\pi}{\lambda}$ and its direction is the direction of the wave propagation.

The periodicity in time gives

$$kx - \omega(t+T) = \phi \pm 2\pi = kx - \omega t - 2\pi \tag{2.5}$$

giving

$$\omega = \frac{2\pi}{T}$$

From definition (2.2), $\omega = ck$, giving $c = \omega/k = \lambda/T$ and therefore

$$\boxed{c = \lambda f} \tag{2.6}$$

where $f = 1/T$ is the frequency of the wave (not the wave function $f(kx \pm \omega t)$). c is called the phase velocity of the wave. The phase velocity c is how fast the phase ϕ changes with time. In general this is not how fast the wave travels (called the group velocity, see section 2.5.2). However for sine waves the group velocity is equal to the phase velocity.

2.2 Wave equation

The general form of the wave equation is the following partial differential equation:

$$\frac{\partial^2 f}{\partial t^2} = c^2 \frac{\partial^2 f}{\partial x^2} \tag{2.7}$$

Any function $f(\phi)$ is a solution to this equation as long as $\phi = kx \pm \omega t$. To verify this claim we use the chain rule:

$$\frac{\partial f}{\partial t} = \frac{\mathrm{df}}{\mathrm{d}\phi}\frac{\partial \phi}{\partial t} = \pm\omega\frac{\mathrm{df}}{\mathrm{d}\phi}$$
$$\frac{\partial^2 f}{\partial t^2} = \omega^2\frac{\mathrm{d^2f}}{\mathrm{d}\phi^2}$$
$$\frac{\partial f}{\partial x} = \frac{\mathrm{df}}{\mathrm{d}\phi}\frac{\partial \phi}{\partial x} = k\frac{\mathrm{df}}{\mathrm{d}\phi}$$
$$\frac{\partial^2 f}{\partial x^2} = k^2\frac{\mathrm{d^2f}}{\mathrm{d}\phi^2}$$

Substituting the second derivatives into the wave equation (2.8) gives

$$\omega^2\frac{\mathrm{d^2f}}{\mathrm{d}\phi^2} = c^2k^2\frac{\mathrm{d^2f}}{\mathrm{d}\phi^2}$$

We see that the function f cancels and we are left with $c = \omega/k$, which is consistent with equation (2.2). In three dimensions, the wave equation is extended to

$$\frac{\partial^2 f}{\partial t^2} = c^2\nabla^2 f \tag{2.8}$$

where $\nabla^2 = \triangle = (\frac{\partial^2}{\partial^2 x} + \frac{\partial^2}{\partial^2 y} + \frac{\partial^2}{\partial^2 z})$ is the Laplace operator.

Note that real waves are often damped, which we neglect here. Some wave equations are more complex and very difficult to solve since they are nonlinear e.g. the Korteweg de Vries equation for shallow water waves (sine waves are a small amplitude approximation). However, the terms and concepts we develop in this course (ω, k, phase velocity, group velocity, dispersion) can be applied to water (and other) waves.

2.2.1 Boundary conditions and standing waves

To completely solve the wave equation (2.8) we need initial conditions $f(x, t = 0)$/boundary conditions $f(x = \text{boundary}, t)$. So far we have discussed waves in infinite media with no boundaries. In a finite medium we need boundary conditions. Boundary conditions can completely change the character of the solutions of the wave equation (2.8).

In this section we consider a system with boundaries that leads to standing waves. Standing waves do not travel in the same way as normal waves because they are confined within the boundaries. This is why they are called “standing”. In our example there are walls at $x = 0$ and

$x = L$ with fixed boundary conditions $f(0,t) = f(L,t) = 0$. We are also given the initial condition $f(x, t = 0) \neq 0$ is a sinusoidal oscillation.

We can solve the partial differential equation (2.8) by separating the variables. We assume we can write the solution in the form $f(x,t) = X(x)\Theta(t)$ and substitute this into equation (2.8) giving:

$$X(x)\frac{\mathrm{d}^2\Theta}{\mathrm{dt}^2} = c^2\Theta(t)\frac{\mathrm{d}^2\mathrm{X}}{\mathrm{dx}^2}$$
$$\frac{1}{\Theta(t)}\frac{\mathrm{d}^2\Theta}{\mathrm{dt}^2} = \frac{c^2}{X(x)}\frac{\mathrm{d}^2\mathrm{X}}{\mathrm{dx}^2} = \alpha$$

where α is a constant. This gives

$$\frac{\mathrm{d}^2\Theta}{\mathrm{dt}^2} = \alpha\Theta(t)$$
$$\frac{\mathrm{d}^2\mathrm{X}}{\mathrm{dx}^2} = \frac{\alpha}{c^2}X(x)$$

Since we expect sinusoidal oscillations, α must be negative and we let $\alpha = -\omega^2$. This gives $X(x) = A\sin\frac{\omega x}{c}$ where we know from the boundary condition $f(0,t) = 0$ that the coefficient of the cosine bits is zero and we are left with just the sine bits with the coefficient $A \neq 0$. The second equation gives $\Theta(t) = B\cos\omega t$ where we know from the initial condition that $B \neq 0$. We combine these solutions together to give

$$f(x,t) = A'\cos\omega t\sin kx$$

where $k = \omega/c$. From the second boundary condition $f(L,t) = 0$ we find $f(L,t) = A'\cos\omega t\sin kL = 0$ and therefore $kL = n\pi$ so $k = n\pi/L$. Since we know that $k = 2\pi/\lambda$ where λ is the wavelength, we can write $\lambda = 2L/n$ enabling us to sketch the standing wave solutions. From the solution above we can see that one way of viewing a standing wave is a simple harmonic oscillation ($\cos\omega t$) with amplitude $A'\sin kx$. Another way of viewing a standing wave is the superposition of 2 waves, one travelling right and one travelling left since

$$f(x,t) = \frac{A'}{2}\left(\sin\left(kx - \omega t\right) + \sin\left(kx + \omega t\right)\right)$$

Exercise. *Check for yourself using trigonometric identities that the 2 expressions for $f(x,t)$ given above are equivalent.*

Note that the equivalence of these 2 ways of writing the standing wave solution $f(x,t)$ explains why we can assume $f(x,t) = X(x)\Theta(t)$ to use the method of separation of variables and $f(x,t)$ is still a function only of the phase $\phi = kx \pm \omega t$.

2.3 Examples of waves

2.3.1 Waves on strings

Waves on strings are transverse waves. One application is in stringed musical instruments. Here we will derived the wave equation for a wave on a string.

Consider a piece of string held under tension with a force F_0 along the $\pm x$ direction. The string has mass per unit length of μ and is plucked at time $t = 0$ giving an initial deflection $y(x, t = 0) = f(x)$. To derive the wave equation we consider newton's law of motion ($F = ma$) in the y direction of a small piece of string between points x and $x + \Delta x$. The equation of motion is then

$$F_y(x + \Delta x) - F_y(x) = \mu \Delta x \frac{\partial^2 y}{\partial t^2}$$

where $F_y(x + \Delta x)$ is the magnitude of the component of the tension force F_0 in the y direction at point $(x + \Delta x)$ and $F_y(x)$ is the magnitude of the component of F_0 in the y direction at point x. If we call the component of the tension force along the string $F_T(x)$, the y component is $F_y(x) = F_T(x) \sin\theta(x)$ where $\theta(x)$ is the angle from the x axis to the string at point x. We calculate $F_T(x)$ from the x component which we know is F_0 i.e. $F_x(x) = F_T(x)\cos\theta = F_0$ therefore $F_y(x) = F_0 \frac{\sin\theta(x)}{\cos\theta(x)} = F_0 \tan\theta(x)$. For small Δx we use $\tan\theta(x) = \frac{\partial y}{\partial x}\big|_x$ and therefore $F_y(x) = F_0 \frac{\partial y}{\partial x}\big|_x$ and similarly $F_y(x + \Delta x) = F_0 \frac{\partial y}{\partial x}\big|_{x+\Delta x}$ which we put into the equation of motion to give:

$$\begin{aligned} F_0 \left(\frac{\partial y}{\partial x}\Big|_{x+\Delta x} - \frac{\partial y}{\partial x}\Big|_x \right) &= \mu \Delta x \frac{\partial^2 y}{\partial t^2} \\ F_0 \frac{\Delta \frac{\partial y}{\partial x}}{\Delta x} &= \mu \frac{\partial^2 y}{\partial t^2} \\ F_0 \frac{\partial^2 y}{\partial x^2} &= \mu \frac{\partial^2 y}{\partial t^2} \\ \frac{\partial^2 y}{\partial t^2} &= \frac{F_0}{\mu} \frac{\partial^2 y}{\partial x^2} \end{aligned}$$

which clearly has the form of the wave equation with phase velocity, c given by

$$c = \sqrt{\frac{F_0}{\mu}} \tag{2.9}$$

Exercise. *Why does a double bass (the biggest stringed instrument) play lower notes than a violin? Why are such stringed instruments tuned by*

tightening/loosening the strings? (remember $c = \lambda f$) If you shorten the length L of the string explain how will this affect the frequency.

2.3.2 Sound (Pressure waves in gases and liquids)

In a fluid (gas or liquid), sound waves are always longitudinal — the fluid is compressed/expanded along the direction of the wave's propagation. In this section we will derive the wave equation for sound in a fluid (e.g. air).

Assume a cylinder of air of cross section A in which the air may move forwards or backwards with velocity, v. We will consider a small volume $V = Al$ of fluid between locations x, and $x + l$. At position, x, the fluid has pressure, $P(x)$, and moves with velocity $v(x)$ along the tube (i.e, longitudinally). At $x + l$, the pressure is $P(x + l)$ and velocity $v(x + l)$. Since the pressure at x and $x+l$ may be different, an overall force F acts on V given by

$$F = -A\Delta P = -A(P(x + l) - P(x)) \approx -Al\frac{\partial p}{\partial x} = -V\frac{\partial p}{\partial x},$$

where we have assumed $l \ll \lambda$ where λ is the length scale of the total pressure change, so that we can make the approximation $\Delta P \approx l\frac{d\partial P}{d\partial x}$. Since in the end l will cancel we can always choose it to be small enough. We now put this force into Newton's equation of motion $F = ma$

$$-V\frac{\partial p}{\partial x} = \rho V\frac{\partial v}{\partial t} \tag{2.10}$$

where ρ is the density of the air. The change in volume depends on the compressibility, $\kappa = -\frac{1}{V}\frac{\partial V}{\partial P}$, so the change in volume is $\Delta V = -\kappa V\Delta P$. The change in volume $\Delta V = A\Delta l$ is given by the change in l over a time Δt given by the velocities at the points x and $x + l$ i.e. $\Delta V = A\Delta l = A(v(x+l)-v(x))\Delta t \approx Al\frac{\partial v}{\partial x}\Delta t$. Equating these expressions for the change in volume we get $-\kappa V\Delta P = Al\frac{\partial v}{\partial x}\Delta t$ leading to

$$\frac{\partial v}{\partial x} = -\kappa\frac{\partial P}{\partial t} \tag{2.11}$$

where we have assumed small Δt and ΔP to make the approximation $\frac{\Delta P}{\Delta t} \approx \frac{\partial P}{\partial t}$. We take the derivative with respect to x of equation (2.10) and the time derivative of equation (2.11) giving:

$$\frac{\partial^2 v}{\partial x\partial t} = -\frac{1}{\rho}\frac{\partial^2 P}{\partial x^2} = -\kappa\frac{\partial^2 P}{\partial t^2}$$

leading to the wave equation for sound in a fluid:

$$\frac{\partial^2 P}{\partial t^2} = \frac{1}{\kappa\rho}\frac{\partial^2 P}{\partial x^2} \tag{2.12}$$

and therefore the phase velocity of sound in a fluid is

$$c = \frac{1}{\sqrt{\kappa\rho}}$$

2.3.3 Mechanical waves in solids

Mechanical wave in solids are sometimes called elastic waves. Solids can sustain longitudinal waves and transverse waves.

Longitudinal (sound in solids)

Longitudinal mechanical waves in solids are sound (or acoustic) waves. We can derive the wave equation in a similar way to sound waves in fluids but now the stress, $\sigma = F/A$ plays the role pressure played before. Instead of the compressibility, the relevant mechanical property of the material we need here is the Young's modulus (also called the elastic modulus) E defined as

$$E = \frac{\text{stress}}{\text{strain}} = \frac{\sigma}{\frac{\delta l}{l}} \tag{2.13}$$

where δl is the amount the material is stretched by. We calculate the force exerted to stretch/compress block of material with crosssectional area A and length l between points x and $x + \Delta x$ as:

$$F = A(\sigma(x + \Delta x) - \sigma(x)) = AE\left(\frac{\delta l}{l}(x + \Delta x) - \frac{\delta l}{l}(x)\right)$$

where $\frac{\delta l}{l}(x)$ is the strain at point x. If $\delta l \ll l$ we can approximate the strain at point x to be the derivative i.e. $\frac{\delta l}{l}(x) \approx \frac{\partial l}{\partial x}\big|_x$ we can then write

$$F = AE\left(\frac{\partial l}{\partial x}\big|_{x+\Delta x} - \frac{\partial l}{\partial x}\big|_x\right) = AE\Delta\frac{\partial l}{\partial x} = AE\frac{\partial^2 l}{\partial x^2}\Delta x$$

We put this force into Newton's equation $F = ma$;

$$AE\frac{\partial^2 l}{\partial x^2}\Delta x = \rho A\Delta x\frac{\partial^2 l}{\partial t^2}$$

giving the wave equation for sound in a solid:

$$\frac{\partial^2 l}{\partial t^2} = \frac{E}{\rho}\frac{\partial^2 l}{\partial x^2}$$

so the phase velocity of sound in solids is given by

$$c = \sqrt{\frac{E}{\rho}} \tag{2.14}$$

Transverse (shear waves)

Solids can also sustain another type of waves, known as shear waves. Shear waves are transverse rather than longitudinal waves, and therefore can be polarised. Shear is defined as a sideways deformation (shear) of a body, rather than expansion or compression. A force F is applied along the cross section surface, of area A, of a body, not normal to it. The shear stress $\tau = F/A$ shears the body by an angle θ. How much the body shears depends on the shear modulus G defined as

$$G = \frac{\tau}{\theta}. \tag{2.15}$$

Analogous to the situation for longitudinal waves in solids, we can find the wave equation for shear waves in solids:

$$\frac{\partial^2 l}{\partial t^2} = \frac{G}{\rho}\frac{\partial^2 l}{\partial x^2}$$

and therefore shear waves propagate with phase velocity given by:

$$c = \sqrt{\frac{G}{\rho}} \tag{2.16}$$

Some materials may display anisotropic shear modulus, i.e. shear modulus is different in different directions and so different for the two possible polarisations of shear waves. Most materials when stretched in one direction, compress in the other, however there are some materials (negative Poisson ratio) that when stretched in one direction, extend in the other direction! However, for almost all materials, the shear modulus is smaller than elastic modulus, typically $2G < E < 3G$. This means it is harder to expand/compress the material than to shear it. Consequently, shear waves have slower phase velocity than longitudinal mechanical waves. Shear waves are unique to solids, fluids (liquids or gases) have zero shear modulus and therefore, cannot sustain shear waves.

Seismic waves

An important example of mechanical waves in solids are seismic waves — earthquakes. A seismic event in the Earth's mantle generates both longitudinal and shear waves. Seismic wave propagation can be complicated since the properties of the medium (rock) depend on the depth below ground and waves can travel along the surface as well as through the bulk. Often an earthquake is felt twice. The first wave (called the primary or P wave) is longitudinal and the second wave felt (called the secondary or S wave) is a shear wave.

Exercise. *Why does the longitudinal wave arrive before the transverse wave? Hint compare the speeds calculated in the previous section)*

The S-wave usually is more destructive. Seismic waves hitting the sea bed from below can cause tsunamis.

Exercise. *Is it the P wave or the S wave that causes a tsunami?*

2.4 Transmission/Reflection

We will now study the behaviour of waves that encounter a boundary between 2 media. e.g. light travelling in air hitting glass - how much light (what fraction of the intensity) will be transmitted and how much reflected?

2.4.1 Intensity

The intensity, I, of a wave is the energy it transports per unit area per unit time and is given by how fast the energy density propagates, i.e.

$$\begin{aligned} I &= \frac{\text{energy transported}}{\text{area} \times \text{time}} \\ &= \text{energy density} \times \text{velocity} \\ I &= wc \end{aligned} \tag{2.17}$$

where w is the energy density i.e. energy/volume. Waves, (similar to an oscillator), can store energy in two way: as kinetic energy (e.g., a moving mass of gas, or piece of string), or, as potential energy (e.g., as a compressed gas, or a stretched piece of string). As the wave propagates, it continuously converts energy form one form to the other but the sum of the two is always the same. So when you have maximum kinetic energy

you have minimum potential energy and vice versa. e.g. for sound waves in air this corresponds to maximum velocity when you have minimum pressure and vice versa. The energy density w can therefore be calculated in two ways, either from the amplitude of the velocity of the gas, v_0, (all the energy is kinetic), or the pressure amplitude, ΔP_0, (all the energy is potential). So for sound waves,

$$w = \frac{1}{2}\rho v_0^2 \tag{2.18}$$

$$w = \frac{1}{2}\kappa \Delta P_0^2 \tag{2.19}$$

This gives us various expressions for the intensity $I = \frac{1}{2}\rho v_0^2 c = \frac{1}{2}\kappa \Delta P_0^2 c$ and different expressions for different types of waves (e.g. light would be in terms of electric/magnetic energy). What we want to know is how to work out how much of this intensity is reflected at a boundary between media and how much is transmitted to the second medium. To do this we need to define a new quantity - Impedance.

2.4.2 Impedance

The impedance Z is defined as the ratio of the amplitudes of the quantities characterising the different types of energy involved in the wave. Sticking with our example of sound waves, conservation of energy means we can equate the 2 types of energy, equation (2.18) and (2.19) to find the ratio of the amplitudes:

$$\frac{1}{2}\rho v_0^2 = \frac{1}{2}\kappa \Delta P_0^2$$

$$Z = \frac{\Delta P_0}{v_0} = \sqrt{\frac{\rho}{\kappa}}$$

from last week we know that the speed of sound in air is $c = 1/\sqrt{\kappa\rho}$ giving the impedance for sound waves:

$$\boxed{Z = c\rho} \tag{2.20}$$

You should remember this expression for the impedance. We can now express the intensity in terms of the impedance $Z = c\rho$. From equation (2.17) and (2.18)

$$I = wc = \frac{1}{2}\rho v_0^2 c = \frac{1}{2}v_0^2 Z \tag{2.21}$$

For other types of waves the amplitudes defining the impedance will be different e.g. for electromagnetic waves, energy is stored either as electric or as magnetic. The ratio between the amplitudes of the electric field E_0 and magnetic field H_0 is given by:

$$Z = \frac{E_0}{H_0} = \sqrt{\frac{\mu_r \mu_0}{\epsilon_r \epsilon_0}} \tag{2.22}$$

You will learn more about this in your electromagnetism course next semester. Note the units for Z are different for different types of waves!

Impedance is a property of the medium not the wave. Impedance

2.4.3 Transmission and reflection coefficients

Impedance is the general quantity that controls the transmission and reflection of waves that encounter a boundary between two different media.

Imagine a wave incident on a boundary between two media with different impedance Z_1 and Z_2. Some of the intensity of the wave may transmit from the first media into the second, whereas some may be reflected. Two extremes of this are that of a free end (e.g. a loose end of a string) for which $Z_2 = 0$ and a fixed end (e.g. a string attached to a wall) for which $Z_2 \to \infty$. An in between situation would be e.g. a thin string attached to a thicker rope.

Here we will derive expressions for the transmission coefficient $t = \frac{I_t}{I_i}$, which is the ratio of the intensity transmitted I_t to the incident intensity I_i and the reflection coefficient $r = \frac{I_r}{I_i}$, which is the ratio of the intensity reflected I_r to the incident intensity I_i. We call the velocity amplitude of the incident, reflected and transmitted waves, v_i, v_r and v_t respectively.

From conservation of energy and using equation (2.21) giving the intensity in terms of the impedance we obtain:

$$\begin{aligned} I_i &= I_r + I_t \\ \frac{1}{2} Z_1 v_i^2 &= \frac{1}{2} Z_1 v_r^2 + \frac{1}{2} Z_2 v_t^2 \\ Z_1(v_i^2 - v_r^2) &= Z_2 v_t^2 \end{aligned} \tag{2.23}$$

The deflection must be continuous at the boundary (otherwise the medium would be broken!) so the velocity amplitudes must be continuous, i.e.

$$v_i + v_r = v_t \tag{2.24}$$

Dividing equation (2.23) by equation (2.24) gives

$$\frac{Z_1(v_i^2 - v_r^2)}{v_i + v_r} = \frac{Z_2 v_t^2}{v_t}$$
$$\frac{Z_1(v_i + v_r)(v_i - v_r)}{v_i + v_r} = Z_2 v_t$$
$$Z_1(v_i - v_r) = Z_2 v_t$$

Substituting in $v_r = v_t - v_i$ from equation (2.24) gives

$$v_t = \frac{2Z_1 v_i}{Z_1 + Z_2}$$
$$v_r = \frac{(Z_1 - Z_2)v_i}{Z_1 + Z_2}$$

Substituting this into the condition for energy conservation (2.23) expressed as fractions of the incident intensity I_i, we find:

$$\frac{I_r}{I_i} + \frac{I_t}{I_i} = 1$$
$$\frac{Z_1 v_r^2}{Z_1 v_i^2} + \frac{Z_2 v_t^2}{Z_1 v_i^2} = 1$$
$$\frac{(Z_1 - Z_2)^2}{(Z_1 + Z_2)^2} + \frac{4Z_1 Z_2}{(Z_1 + Z_2)^2} = 1$$
$$r + t = 1$$

i.e.

$$\text{transmission coefficient} = t = \frac{I_t}{I_i} = \frac{4Z_1 Z_2}{(Z_1 + Z_2)^2} \tag{2.25}$$

$$\text{reflection coefficient} = r = \frac{I_r}{I_i} = \frac{(Z_1 - Z_2)^2}{(Z_1 + Z_2)^2} \tag{2.26}$$

The transmission and reflection coefficients (t and r) describing the fractions of intensity transmitted and reflected are given in terms of the impedances of the media. This is why the concept of impedance is important.

We can now see that the for a free end ($Z_2 = 0$), $t = 0$ and $r = 1$ so all the wave is reflected. Similarly for a fixed end ($Z_2 \to \infty$), $t = 0$ and therefore since $r + t = 1$ it must be that $r = 1$ (note this is not clear from just putting $Z_2 \to \infty$ into equation (2.26). For the fixed end $v_t = 0$ i.e. it is a node and so from equation (2.24) it follows that $v_r = -v_i$ and the reflected wave is π ($180°$) out of phase with the incident wave. However for a free end there is no phase jump because the velocity amplitude doesn't have to be zero ($I_t = 0$ because $Z_2 = 0$ regardless of the value of v_t).

Exercise. *If Z_1 and Z_2 are swapped. i.e. the wave approaches from the other medium, what happens to t and r?*

Now let us consider 2 successive junctions. To get the final transmitted intensity we multiply the transmission coefficients for each junction, i.e. $I_t = I_i t_1 t_2$.

Note that for light Z in equations (2.25) and (2.26) can be replaced by the refractive index n (as long as the light is incident normal perpendicular to the interface).

2.5 Dispersion of waves

Remember the distortion/deflection peak (colloquially a "hump", scientifically a "wave packet" - see below) we drew when we first introduced waves (section 2.1)? We said a wave is a distortion that travels in space. There we assumed that the shape of this distortion did not change, only its position in space. In reality this assumption is not always valid. Imagine for example ripples on the surface of water. Over time, as the ripples travel in space they also spread out decreasing in amplitude. This is also the case for sound waves in the example of thunder whose intensity decreases with distance (as $I \sim 1/r^2$) but also spreads out such that we hear a longer rumble.

This kind of spreading out behaviour, called dispersion, does not happen for a pure sine wave. A hump like the one we drew is actually made up of a sum of sine waves of different wave numbers k, as you know from your work on Fourier series. Dispersion only happens if the phase velocity is not a constant but a function of wavenumber $c(k)$. If this is the case, different Fourier components will have different phase velocities so some components will travel faster than others and the hump shape will spread out. Obviously for a pure sine wave there is only one component so only one phase velocity, c, so there cannot be any dispersion. In reality there is almost always dispersion. The only scenario showing no dispersion is light in a vacuum.

2.5.1 Wave packet

Information is transmitted in "humps" in what we call wave packets. A pure sine wave contains no information because it continues in the same way indefinitely. To transmit a message however we have to transmit pulses or peaks, i.e. wave packets. A wave packet is a superposition of different

sin/cos waves with different wavenumbers k. Note this is different from interference, which is superposition of waves of the same k but different origin/direction. The simplest example of a wave packet is one containing just 2 components. e.g. a superposition of 2 waves with slightly different k leads to a wave shape known as "beats" (see Figure 2.1).

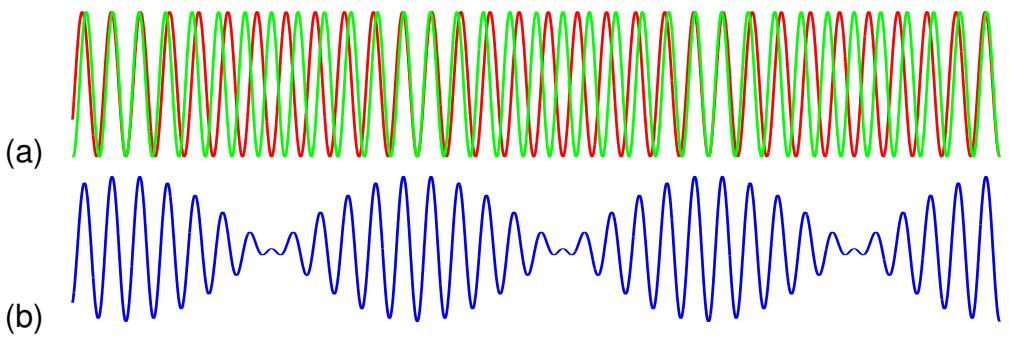

Figure 2.1: (a) Two sinusoidal waves with wavenumbers differing by $\frac{\Delta k}{k} = 0.1$. (b) The two waves from (a) added together forming a pattern called "beats".

2.5.2 Group velocity

How fast does information travel? Information travels with the velocity the wave packet travels with. How fast does a wave packet move? The wave packet is a superposition of sine/cos waves of different k and each component may travel with a different phase velocity. Therefore the velocity of the wave packet is not the same as the phase velocity of one of its components. We define the velocity of the wave packet as the velocity of its peak. We call this the "group velocity" (as it is the velocity of the group of components making up the wave packet). It is the group velocity that is the velocity of the physical object or information not the phase velocity.

Let us derive the expression for the group velocity by considering the simplest possible group (wave packet) – a superposition of 2 sine waves with angular frequencies ω and $\omega + \Delta\omega$ and wavenumbers k and $k + \Delta k$. We assume that $\Delta\omega \ll \omega$ and $\Delta k \ll k$. The peak of the wave packet is where the 2 waves are in phase, located at a point we choose to call $x_p(t)$. At the point the phase ϕ of each wave is equal. Remember the definition

of the phase, $\phi = kx - \omega t$. For our 2 waves we can then write:

$$\begin{aligned} \phi_1 &= \phi_2 \\ kx_p - \omega t &= (k + \Delta k)x_p - (\omega + \Delta\omega)t \\ \Delta k\, x_p &= \Delta\omega\, t \\ x_p &= \frac{\Delta\omega}{\Delta k}\, t \end{aligned}$$

and therefore the peak moves with velocity $\frac{\mathrm{dx_p}}{\mathrm{dt}} = \frac{\Delta\omega}{\Delta k}$. In the limit of infinitesimally small $\Delta\omega$ and Δk this becomes $\frac{\Delta\omega}{\Delta k} \rightarrow \frac{\mathrm{d}\omega}{\mathrm{dk}}$ The definition of the group velocity, v_G is then

$$\boxed{v_{_G} = \frac{\mathrm{d}\omega}{\mathrm{dk}}} \tag{2.27}$$

This is important to remember. Note the group velocity is the derivative of ω with respect to k whereas the phase velocity $c = \frac{\omega}{k}$. To find the group velocity we have to first find the function $\omega(k)$ which is called the **dispersion relation**

$$\omega(k) = c(k)k \tag{2.28}$$

If the phase velocity is constant or there is only one component the $c(k) = c =$constant and therefore $v_G = c$ i.e. the group velocity is equal to the phase velocity. This is the case of the pure sine waves we considered earlier in this topic.

However, for large k (short wavelength λ) all media show dispersion (apart from light in a vacuum). e.g. for mechanical waves in solids the expression $c = \sqrt{\frac{E}{\rho}}$ (equation 2.14) breaks down for small wavelengths of the order of the atomic spacing of the solid. In general, all media show dispersion, and $v_G \neq c$. Establishing the dispersion relation is an important and non-trivial exercise in the description of a medium. Note that some dispersion relations contain a point where $v_G = 0$ so a wave can stand still after all, without being confined by boundaries!

2.5.3 Dispersion of light in matter

Light is an electromagnetic wave so is described by Maxwell's equations (see your electromagnetism course for more details):

$$\frac{\partial B}{\partial t} = -\frac{\partial E}{\partial x} \tag{2.29}$$

$$\frac{\partial E}{\partial t} = -\frac{1}{\epsilon_r \epsilon_0 \mu_r \mu_0}\frac{\partial B}{\partial x} \tag{2.30}$$

where E is the electric field, B is the magnetic field, $\epsilon = \epsilon_r \epsilon_0$ is the dielectric constant (ϵ_0 being that of free space and ϵ_r the relative dielectric permittivity), $\mu = \mu_r = \mu_0$ is the magnetic permeability (μ_0 being that of free space and μ_r the relative magnetic permeability). Differentiating equation 2.29 by x and equation 2.30 by t leads to the wave equation for electromagnetic waves:

$$\frac{\partial^2 E}{\partial t^2} = \frac{1}{\epsilon\mu}\frac{\partial^2 E}{\partial x^2} \tag{2.31}$$

so the phase velocity of light is given by

$$c = \frac{1}{\sqrt{\epsilon\mu}} = \frac{1}{\sqrt{\epsilon_r\epsilon_0\mu_r\mu_0}} = \frac{c_0}{\sqrt{\epsilon_r\mu_r}} = \frac{c_0}{n} \tag{2.32}$$

where $c_0 = (\epsilon_0\mu_0)^{-1/2}$ is the phase velocity of light in a vacuum (i.e. the speed of light $c_0 = 3 \times 10^8\,\mathrm{ms}^{-1}$) and n is the refractive index given by

$$n = \sqrt{\epsilon_r\mu_r}$$

The dispersion relation $\omega(k)$ of light is usually given in terms of $n(\lambda)$. As you know from optics, Snell's law ($\frac{n_2}{n_1} = \frac{c_1}{c_2} = \frac{\sin\theta_1}{\sin\theta_2}$) tells you about how light incident from air bends as it enters glass. Since the refractive index $n(\lambda)$ depends on the wavelength (colour) of light, different colours of light are bent by different amounts by a prism leading to a rainbow separation of colours. It is dispersion that leads to this separation of colours.

In lens optics dispersion is a problem because different colours have slightly different focal points. White light will not be focused to a single point but instead separation of colours will cause the image to appear blurred. This is called chromatic aberration. It was corrected for by Abbé in the 19th century who made a achromatic lens (an apochromat). His lens is a combination of a strongly focusing converging lens made of glass with weak dispersion plus a diverging lens made of glass with large dispersion. This was a revolution in microscope technology and consequently biology and medicine.

Topic 3

Fictitious forces

3.1 Coordinate systems

A coordinate system mathematically describes the location of points in space. Coordinate systems are not part of physical reality, they are just our description. Changing the coordinate system does not change the physical system, only the way we describe where the points are mathematically. A coordinate system is a specific way of locating a given point on space. In 3D we need 3 coordinates to specify a point (and n coordinates for a point in n dimensional space). Every coordinate system has an origin, which can be chosen arbitrarily. We are free to choose any origin we like but it will be a good idea to pick the most convenient one e.g. the potential energy of a spring being deflected about its equilibrium length x_0 is given by $V = \frac{1}{2}k(x - x_0)^2$ but if we set the origin to be at the equilibrium point $x_0 = 0$ the expression simplifies to $V = \frac{1}{2}kx^2$ which is easier to work with.

In comparison, a frame of reference refers to the state of motion of the origin of a coordinate system and or the rotation of the coordinate system. We can use different coordinate systems in the same frame of reference and we can use the same coordinate system in different frames of reference.

In the following we discuss the most common coordinate systems.

3.1.1 Cartesian coordinates

Cartesian coordinates, shown in figure 3.1, are the ones you have grown up with and are most comfortable with. In 3D the 3 basis vectors (unit vectors) are along the x, y and z axes respectively and are referred to by

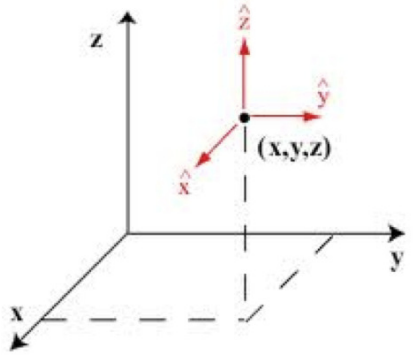

Figure 3.1: Cartesian coordinates

various notations: $\boldsymbol{i}$, $\boldsymbol{j}$, $\boldsymbol{k}$ or $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$, $\hat{\boldsymbol{z}}$ or $\boldsymbol{e_x}$, $\boldsymbol{e_y}$, $\boldsymbol{e_z}$ or even $\boldsymbol{e_1}$, $\boldsymbol{e_2}$, $\boldsymbol{e_3}$. The position vector $\boldsymbol{r}$ of any point can then be written as

$$\begin{aligned} \boldsymbol{r} = (x, y, z) &= x\,\boldsymbol{i} + y\,\boldsymbol{j} + z\,\boldsymbol{k} \\ & x\,\hat{\boldsymbol{x}} + y\,\hat{\boldsymbol{y}} + z\,\hat{\boldsymbol{z}} \\ & x\,\boldsymbol{e_x} + y\,\boldsymbol{e_y} + x\,\boldsymbol{e_z} \end{aligned}$$

By definition these basis vectors have modulus 1 (they are unit vectors) and they are mutually perpendicular (orthogonal/normal). These two properties can be neatly written mathematically:

$$\boldsymbol{e_i} \cdot \boldsymbol{e_j} = \delta_{ij}$$

where the dot designates the dot product (also known as the inner or scalar product) and δ_{ij} is the Kronecker delta function defined as:

$$\delta_{ij} = \begin{matrix} 0 & i \neq j \\ 1 & i = j \end{matrix}$$

Mathematically we say the basis vectors form a "complete orthonormal set" since they are all normal to each other and completely specify all points.

To set up a Cartesian coordinate system we first choose an origin (any point we like, as discussed above). The direction of the first unit vector (say $\boldsymbol{e_x}$) is totally arbitrary so we pick any direction we want for convenience and call this the x direction. We then pick any direction perpendicular to this and call it the y direction. Once the x and y directions have been chosen the z direction is fixed — it has to be perpendicular to both x and y and it's sign (up/down) is fixed by the convention of using a right handed

coordinate system. This is the first right hand rule: x is your thumb, y your first finger and z your middle finger.

3.1.2 Spherical coordinates

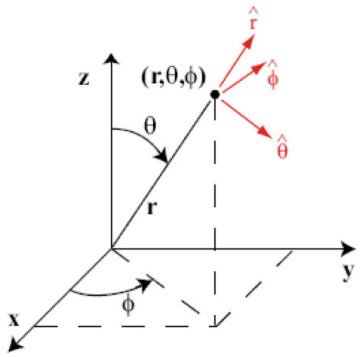

Figure 3.2: Spherical coordinates

If our system has spherical symmetry it may be more convenient to use spherical coordinates, as drawn in figure 3.2. To set up a spherical coordinate system we first choose our origin and then define a convenient direction as the z axis e.g. in a rotating problem z would be the axis of rotation. The positive direction of z is defined by the second right hand rule: if your thumb is z and you rotate your fingers in so they point in the sense of the rotation then your thumb is pointing in the positive z direction. Since many people are right handed most screws are threaded right handed so to tighten a screw remember "righty tighty, lefty loosey".

The 3 coordinates specifying a point in spherical coordinates are the distance r from the origin, the angle θ down from the z axis and the angle ϕ rotated about the z axis in a positive sense as defined by the second right hand rule. Note that the z axis is necessary for the correct definition of the angles θ and ϕ but it is not a coordinate itself. The direction in which $\phi = 0$ is arbitrary and we can choose it as we like. Unlike Cartesian coordinates which each have an infinite range, spherical coordinates do not have the same ranges. The ranges are as follows:

$$0 < r < \infty$$
$$0 < \phi < 2\pi$$
$$0 < \theta < \pi$$

Note that ϕ has a range of 2π but θ only π.

The system of latitude and longitude used to specify points on the curved surface of the earth is almost the same as spherical coordinates with $r = R_E$ fixed as the radius of the Earth, longitude ϕ with $\phi = 0$ defined as at Greenwich however latitude is not quite the same as θ in spherical coordinates since latitude is defined as zero at the equator so the range is $-90° <$ latitude $< 90°$ whereas spherical coordinates have $0 < \theta < 180°$ with $\theta = 0$ being along the z axis (equivalent to the North pole).

To convert from Cartesian to spherical coordinates and vice versa:

$$r = \sqrt{x^2 + y^2 + z^2} \qquad x = r\sin\theta\cos\phi$$
$$\theta = \cos^{-1}(\frac{z}{r}) \qquad y = r\sin\theta\sin\phi$$
$$\phi = \tan^{-1}(\frac{y}{x}) \qquad z = r\cos\theta$$

3.1.3 Cylindrical coordinates

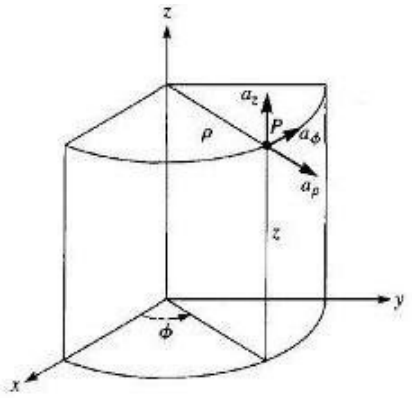

Figure 3.3: Cylindrical coordinates

For systems with cylindrical symmetry we use cylindrical coordinates as shown in figure 3.3. To set up cylindrical coordinate system we choose the direction we want to be the z axis. The coordinates are then the position z along this axis (like in Cartesian coordinates), the distance from the z axis which we call ρ (note this is not the same as the distance r from the origin in spherical coordinates) and the angle ϕ of rotation around the z axis (like spherical coordinates). The distance ρ from the z axis is sometimes called

r or R but, be careful, it is not the same as r in spherical coordinates. Here I will refer to it as ρ. The ranges are as follows:

$$-\infty < z < \infty$$
$$0 < \rho < \infty$$
$$0 < \phi < 2\pi$$

If z is fixed at a constant value we obtain 2D polar coordinates.
To convert form Cartesian to cylindrical coordinates and vice versa:

$$z = z \qquad\qquad z = z$$
$$\rho = \sqrt{x^2 + y^2} \qquad\qquad x = \rho\cos\phi$$
$$\phi = \sin^{-1}(\frac{y}{\rho}) \qquad\qquad y = \rho\sin\phi$$

General comments

We can use any coordinate system that uniquely specifies points in space. So why do we usually use Cartesian? The Cartesian coordinate system is the only one in which the unity vectors point in the same directions at every point in space. The spherical and cylindrical unit vectors point in different directions depending on where on the sphere/cylinder your point is. This property of Cartesian unit vectors makes it much easier to add/multiply vectors in Cartesian coordinates. In other words, calculations involving vectors (e.g. forces and acceleration in Newton's laws) will be much easier in Cartesian coordinates.

If this is the case, then why wouldn't we always use Cartesian coordinates? Spherical/cylindrical coordinates are very useful for systems with spherical/cylindrical symmetry e.g. electrical/gravitational fields are spherically symmetric and the force depends on the distance r from the origin.

So we are faced with a dilemma: to calculate forces we want Cartesian but the symmetry of many systems is spherical/cylindrical. This is why Lagrangian mechanics was developed - see the final topic in this course.

3.1.4 Cross product

It is worth revising the cross product (also called the outer product):

$$\boldsymbol{a} \times \boldsymbol{b} = \boldsymbol{c} \tag{3.1}$$

The vector $\boldsymbol{c}$ is perpendicular to both $\boldsymbol{a}$ and $\boldsymbol{b}$. It is perpendicular to the plane defined by $\boldsymbol{a}$ and $\boldsymbol{b}$. The positive/negative direction of $\boldsymbol{c}$ is specified by the 1st right hand rule. If $\boldsymbol{a}$ and $\boldsymbol{b}$ are parallel the cross product is zero.

The cross product is anti-commutative $\boldsymbol{a} \times \boldsymbol{b} = -\boldsymbol{b} \times \boldsymbol{a}$ so the order matters unlike adding vectors or taking the dot product ($\boldsymbol{a} + \boldsymbol{b} = \boldsymbol{b} + \boldsymbol{a}$).

Note that strictly speaking in two dimensions (2D) or 4D the cross product has no sense — the cross product only works in 3D vector spaces.

To calculate the cross product:

$$\begin{aligned} \boldsymbol{a} \times \boldsymbol{b} &= \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \times \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \\ &= \begin{Vmatrix} \boldsymbol{e_x} & \boldsymbol{e_y} & \boldsymbol{e_z} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{Vmatrix} = \begin{pmatrix} a_2 b_3 - a_3 b_2 \\ -a_1 b_3 + a_3 b_1 \\ a_1 b_2 - a_2 b_1 \end{pmatrix} \end{aligned}$$

Pseudovectors

As a technical aside it is interesting to mention the concept of pseudovectors. By definition vectors are unchanged when the coordinate system is rotated but flip sign when the coordinate system is inverted. i.e. if $\boldsymbol{e_x}, \boldsymbol{e_y}, \boldsymbol{e_z} \rightarrow -\boldsymbol{e_x}, -\boldsymbol{e_y}, -\boldsymbol{e_z}$ then $\boldsymbol{a} \rightarrow -\boldsymbol{a}$. Note that it is the representation of the vector in our coordinates that changes not any physical object it represents because all we have done is change coordinates, which aren't physical they're just the description.

What happens to the sign of $\boldsymbol{c} = \boldsymbol{a} \times \boldsymbol{b}$ under coordinate inversion? If $\boldsymbol{a} \rightarrow -\boldsymbol{a}$ and $\boldsymbol{b} \rightarrow -\boldsymbol{b}$ the $\boldsymbol{c}$ doesn't change so $\boldsymbol{c}$ is not a proper vector! We therefore call it a pseudovector.

Exercise. *Is the cross product of 2 pseudovectors a vector or a pseudovector? Is the cross product of a vector and a pseudovectors a vector or a pseudovector?*

The arrow of a vector defines a direction, therefore flipping the sign changes the direction e.g. up/down or left/right. The arrow of a pseudovector on the other hand defines a sense of rotation, therefore flipping the sign changes the sense of rotation (clockwise/anticlockwise). Examples of pseudovectors include angular velocity $\boldsymbol{\omega}$, angular momentum $\boldsymbol{L}$, torque $\boldsymbol{T}$, magnetic induction $\boldsymbol{B}$ (loops of flux lines)

3.2 Frames of reference

A frame of reference is defined by the state of motion of the origin of a coordinate system &/ a rotation of the coordinate system. There are 2 types of frames of reference:

- inertial (not accelerated and therefore no rotation either)
- non inertial (origin accelerates/rotation)

Remember that inertia means an object resists changes to its motion. Consider an example of 2 different inertial frames of reference. Imagine one person on a train travelling at a constant velocity past a small station and another person standing on the station platform. The origin of frame of reference A is stationary and the origin of frame of reference B is moving with constant velocity $\boldsymbol{v}$ with respect to frame A. In Cartesian coordinate we call the axes of frame A x, y, z and the axes of frame B x', y', z'. We choose the axes of frame B to be parallel to those of frame A (x is parallel to x' etc) Let us say the velocity $\boldsymbol{v}$ is along x and x'. The observer in frame A thinks "I'm at rest and B is moving with velocity $+v$ in the x direction". The observer in frame B thinks "I'm at rest and A is moving with velocity $-v$ in the x' direction. No experiment can show who is "right" according to the laws of mechanics. The **Galilean principle of relativity** states that *the laws of mechanics must be the same in all inertial frames* i.e. acceleration is absolute but velocity is relative. Einstein went a step further by claiming all laws of physics (not just mechanics) are the same in all inertial frames and this lead to special relativity.

Within an inertial frame of reference, all forces are "real" forces i.e. all forces are the result of interactions between bodies (including force fields created by other bodies). Every force is generated by another body (object). In non inertial frames on the other hand we can get apparent/fictitious forces i.e. forces that are not caused by another body. e.g. consider a non inertial frame of reference C in which the origin is accelerating with acceleration $\boldsymbol{a}$ along the x axis. A mass initially at rest at the origin would get left behind in the $-x$ direction experiencing an acceleration a in the $-x$ direction. Therefore from the perspective of an observer in C, Newton's second law $F = ma$ says there must be a force acting on the mass causing it to accelerate in the $-x$ direction. But this "force" has no physical origin and there is no reaction force (Newton's 1st law action=reaction). Therefore we call this apparent force a "fictitious force". It comes from using a non inertial frame of reference and is due to the acceleration of that frame of reference and not to any "real" physical force.

3.3 Fictitious forces derivation

Here we will calculate all the forces including fictitious forces by calculating the acceleration in the non inertial frame, i.e. $\boldsymbol{F'} = m\boldsymbol{a'}$ where the primes indicate quantities measured in the non inertial frame. Note you will not be expected to reproduce this derivation but I am covering it so that you may understand where these fictitious forces come from.

Consider an inertial frame called frame A with coordinates x, y, z and a non inertial frame B with coordinates x', y', z'. Initially we will assume the inertial frame is rotating but its origin is not accelerating (we will add the acceleration of the origin at the end). Let us choose the origin of frame A and frame B to be in the same place. Then the position vector $\boldsymbol{r} = \boldsymbol{r'}$ i.e.

$$\boldsymbol{r} = \boldsymbol{r'} = x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k} = x'\boldsymbol{i'} + y'\boldsymbol{j'} + z'\boldsymbol{k'}$$

we differentiate this to get the velocity:

$$\boldsymbol{v} = \frac{\mathrm{dx}}{\mathrm{dt}}\boldsymbol{i} + \frac{\mathrm{dy}}{\mathrm{dt}}\boldsymbol{j} + \frac{\mathrm{dz}}{\mathrm{dt}}\boldsymbol{k} = \frac{\mathrm{dx'}}{\mathrm{dt}}\boldsymbol{i'} + \frac{\mathrm{dy'}}{\mathrm{dt}}\boldsymbol{j'} + \frac{\mathrm{dz'}}{\mathrm{dt}}\boldsymbol{k'} + x'\frac{\mathrm{d}\boldsymbol{i'}}{\mathrm{dt}} + y'\frac{\mathrm{d}\boldsymbol{j'}}{\mathrm{dt}} + z'\frac{\mathrm{d}\boldsymbol{k'}}{\mathrm{dt}}$$

$$\boldsymbol{v} = \boldsymbol{v'} + x'\frac{\mathrm{d}\boldsymbol{i'}}{\mathrm{dt}} + y'\frac{\mathrm{d}\boldsymbol{j'}}{\mathrm{dt}} + z'\frac{\mathrm{d}\boldsymbol{k'}}{\mathrm{dt}} \tag{3.2}$$

where because the axes $\boldsymbol{i'}$, $\boldsymbol{j'}$, $\boldsymbol{k'}$ are changing in time (since the frame is rotating) we have to differentiate these as well as the coordinates x', y', z'.

To work out what $\frac{\mathrm{d}\boldsymbol{i'}}{\mathrm{dt}}$ is, we consider the change $\Delta\boldsymbol{i'}$ in time Δt. By considering the geometry and assuming $\Delta\boldsymbol{i'}$ is an arc length we can write

$$|\Delta\boldsymbol{i'}| \approx \sin\phi\Delta\theta \tag{3.3}$$

where $\Delta\theta$ is the angle rotated in time Δt and ϕ is the angle between $\boldsymbol{i'}$ and the axis of rotation $\boldsymbol{\omega}$ so $\sin\phi$ is the radius of the circle traced out by the rotating axis $\boldsymbol{i'}$ perpendicular to the axis of rotation $\boldsymbol{\omega}$. If we assume the axis of rotation $\boldsymbol{\omega}$ is about the z' direction then $\phi = \frac{\pi}{2}$ and $\sin\phi = 1$. We now divide equation (3.3) by Δt and take the infinitesimal limit of the small quantities so $\Delta t \to \mathrm{dt}$, $\Delta\boldsymbol{i'} \to \mathrm{d}\boldsymbol{i'}$, $\Delta\theta \to \mathrm{d}\theta$. This gives:

$$\frac{\mathrm{d}|\boldsymbol{i'}|}{\mathrm{dt}} = \frac{\mathrm{d}\theta}{\mathrm{dt}}\sin\phi = \omega\sin\phi$$

We know from the geometry that the direction of $\Delta\boldsymbol{i'}$ is perpendicular to both $\boldsymbol{\omega}$ and $\boldsymbol{i'}$ and therefore the expression above in vector form is just the cross product:

$$\frac{\mathrm{d}\boldsymbol{i'}}{\mathrm{dt}} = \boldsymbol{\omega} \times \boldsymbol{i'}$$

similarly $\frac{\mathrm{d}\boldsymbol{j'}}{\mathrm{dt}} = \boldsymbol{\omega} \times \boldsymbol{j'}$ and $\frac{\mathrm{d}\boldsymbol{k'}}{\mathrm{dt}} = \boldsymbol{\omega} \times \boldsymbol{k'}$. We put this into equation (3.2) which gives:

$$\begin{aligned}
\boldsymbol{v} &= \boldsymbol{v'} + x'\boldsymbol{\omega} \times \boldsymbol{i'} + y'\boldsymbol{\omega} \times \boldsymbol{j'} + z'\boldsymbol{\omega} \times \boldsymbol{k'} \\
\boldsymbol{v} &= \boldsymbol{v'} + \boldsymbol{\omega} \times (x'\boldsymbol{i'} + y'\boldsymbol{j'} + z'\boldsymbol{k'}) \\
\boldsymbol{v} &= \boldsymbol{v'} + \boldsymbol{\omega} \times \boldsymbol{r'} \qquad (3.4)
\end{aligned}$$

Equation (3.4) can be written as $(\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{dt}})_{\text{fixed}} = (\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{dt}})_{\text{rot}} + \omega \times \boldsymbol{r}$ using $\boldsymbol{r} = \boldsymbol{r'}$ as we stated at the beginning. This transformation is true more generally for any vector $\boldsymbol{Q}$:

$$\left(\frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{dt}}\right)_{\text{fixed}} = \left(\frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{dt}}\right)_{\text{rot}} + \boldsymbol{\omega} \times \boldsymbol{Q} \qquad (3.5)$$

so for the velocity vector $\boldsymbol{v}$ we have

$$\left(\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{dt}}\right)_{\text{fixed}} = \left(\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{dt}}\right)_{\text{rot}} + \boldsymbol{\omega} \times \boldsymbol{v}$$

Substituting equation (3.4) into the right hand side gives:

$$\begin{aligned}
\left(\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{dt}}\right)_{\text{fixed}} &= \left(\frac{\mathrm{d}}{\mathrm{dt}}\right)_{\text{rot}} (\boldsymbol{v'} + \boldsymbol{\omega} \times \boldsymbol{r'}) + \boldsymbol{\omega} \times (\boldsymbol{v'} + \boldsymbol{\omega} \times \boldsymbol{r'}) \\
\boldsymbol{a} &= \left(\frac{\mathrm{d}\boldsymbol{v'}}{\mathrm{dt}}\right)_{\text{rot}} + \left(\frac{\mathrm{d}}{\mathrm{dt}}(\boldsymbol{\omega} \times \boldsymbol{r'})\right)_{\text{rot}} + \boldsymbol{\omega} \times \boldsymbol{v'} + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \boldsymbol{r'}) \\
\boldsymbol{a} &= \boldsymbol{a'} + \left(\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{dt}}\right)_{\text{rot}} \times \boldsymbol{r'} + \boldsymbol{\omega} \times \left(\frac{\mathrm{d}\boldsymbol{r'}}{\mathrm{dt}}\right)_{\text{rot}} + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \boldsymbol{r'}) \\
\boldsymbol{a} &= \boldsymbol{a'} + \dot{\boldsymbol{\omega}} \times \boldsymbol{r'} + 2\omega \times \boldsymbol{v'} + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \boldsymbol{r'})
\end{aligned}$$

where we have identified the acceleration $\boldsymbol{a} = (\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{dt}})_{\text{fixed}}$ in the inertial frame and $\boldsymbol{a} = (\frac{\mathrm{d}\boldsymbol{v'}}{\mathrm{dt}})_{\text{rot}}$ in the non inertial frame. We have also used equation (3.5) for $\boldsymbol{\omega}$ to give

$$\left(\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{dt}}\right)_{\text{fixed}} = \left(\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{dt}}\right)_{\text{rot}} + \boldsymbol{\omega} \times \boldsymbol{\omega} = \left(\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{dt}}\right)_{\text{rot}} = \dot{\boldsymbol{\omega}}$$

Finally if we also allow an acceleration $\boldsymbol{a_0}$ of the origin we obtain:

$$\boldsymbol{a} = \boldsymbol{a'} + \dot{\boldsymbol{\omega}} \times \boldsymbol{r'} + 2\omega \times \boldsymbol{v'} + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \boldsymbol{r'}) + \boldsymbol{a_0}$$

We can now write down the forces in the non inertial frame from $\boldsymbol{F'} = m\boldsymbol{a'}$ giving

$$\begin{aligned}
\boldsymbol{F'} = m\boldsymbol{a'} =& m\boldsymbol{a} - m\dot{\boldsymbol{\omega}} \times \boldsymbol{r'} - 2m\omega \times \boldsymbol{v'} - m\boldsymbol{\omega} \times (\boldsymbol{\omega} \times \boldsymbol{r'}) - m\boldsymbol{a_0} \qquad (3.6) \\
\boldsymbol{F'} =& \boldsymbol{F} + \boldsymbol{F'}_{\text{transverse}} + \boldsymbol{F'}_{\text{coriolis}} + \boldsymbol{F'}_{\text{centrifugal}} - m\boldsymbol{a_0}
\end{aligned}$$

where $\boldsymbol{F} = m\boldsymbol{a}$ is the real physical force and the only one seen in the inertial frame. $\boldsymbol{F'}_{\text{trans}} = -m\dot{\boldsymbol{\omega}} \times \boldsymbol{r'}$ is called the transverse force since it is perpendicular (transverse) to $\boldsymbol{r'}$. The transverse force is only experienced if the rotation is accelerating; this is an unusual situation and will not occur in any of the problems we consider in this course. $\boldsymbol{F'}_{\text{cor}} = -2m\boldsymbol{\omega} \times \boldsymbol{v'}$ is the Coriolis force and $\boldsymbol{F'}_{\text{cf}} = -m\boldsymbol{\omega} \times (\boldsymbol{\omega} \times \boldsymbol{r'})$ is the centrifugal force. These 2 forces are the ones we will consider in more detail in the following. The final term $-m\boldsymbol{a_0}$ is due to the translational acceleration of the origin.

3.4 Centrifugal force

The centrifugal force $\boldsymbol{F'}_{\text{cf}}$ is:

$$\boxed{\boldsymbol{F'}_{\text{cf}} = -m\boldsymbol{\omega} \times (\boldsymbol{\omega} \times \boldsymbol{r'})} \tag{3.7}$$

Its magnitude is $|\boldsymbol{F'}_{\text{cf}}| = m\omega^2 r$ and its direction is outwards, perpendicular from the axis of rotation. Note that r is the distance from the axis of rotation (like ρ in cylindrical coordinates).

Imagine me standing at the origin and swinging a mass around me in a horizontal circle (e.g. a ball on the end of a string). From our perspective in the inertial frame of reference A, is the mass accelerating? Yes you know it is because the mass is changing direction. You know the acceleration of the mass is towards the centre of the circle. You should know also that the tangential velocity of the mass is $\boldsymbol{v} = \boldsymbol{\omega} \times \boldsymbol{r}$ where $\boldsymbol{r}$ is the vector from the axis of rotation (or maybe you know it in the form of the magnitudes $v = \omega R$). The angular velocity vector is given by $\boldsymbol{\omega} = \frac{\boldsymbol{r} \times \boldsymbol{v}}{|\boldsymbol{r}|^2}$ and the angular momentum is $\boldsymbol{L} = m\boldsymbol{r} \times \boldsymbol{v}$.

The acceleration of the mass towards the centre is due to the inward force I am exerting on the mass. This force towards the centre of the circle is a real physical force called the **centripetal force**. Newton's first law of motion (action=reaction) tells us that there must be an equal and opposite force applied on me by the mass. This force outwards is called the **centrifugal force**. It is not a real physical force but it is due to the "inertia" of the mass in the rotating frame of reference i.e. the reluctance of the mass to change its speed/direction.

Let us consider the system from the perspective of the mass, i.e. in the non inertial rotating frame of reference B. In the rotating frame B objects move as if there is a force outwards. Consider a second (green) ball loosely attached to the first (black) ball. When I swing the balls around in a circle the green ball moves away from the black ball due to what it thinks is

a force acting outwards. But we know this apparent force is just due to the rotation of the frame of reference. The the green ball is not attached to me so does not experience the centripetal force like the black ball and therefore flies outwards as it wants to continue in a straight line. One way to think about this fictitious force therefore is to say that the *centrifugal force = lack of centripetal force*.

Another example is that of a CD on the dashboard of a car as the car is driven around a corner (try this out next time you're in the passenger seat of a car). As the car goes round a corner to the right the CD slides to the left due to the centrifugal force (it tries to carry on in a straight line whilst the car moves in a circle).

We live in a rotating frame of reference since the Earth is rotating. The centrifugal force we experience due to the rotation of the earth will be different in different places since $\boldsymbol{r'}$ is different in different places on the earth. At the North and South poles the distance to the axis of rotation is zero so the centrifugal force is zero. At the equator the centrifugal force is maximum and acts in the direction opposite to gravity. At other points on the globe the centrifugal force is not parallel to gravity since the centrifugal force is outwards and perpendicular to the axis of rotation whereas gravity acts towards the centre of the earth.

You may have experienced the centrifugal force in rollor coaster rides. The radius of curvature of a curved path of a ride gives you $\boldsymbol{r'}$, which, along with the angular velocity $\boldsymbol{\omega}$, enables you to calculate the centrifugal force. Note that the magnitude of the centrifugal force may change as you go along the ride if the radius of curvature changes.

3.5 Coriolis Force

The coriolis force $\boldsymbol{F'}_{\text{cor}}$ is:

$$\boxed{\boldsymbol{F'}_{\text{cor}} = -2m\boldsymbol{\omega} \times \boldsymbol{v'} = 2m\boldsymbol{v'} \times \boldsymbol{\omega}} \tag{3.8}$$

where the order matters because cross products are anti-commutative. Note that the coriolis force only acts on objects **moving** in the rotating frame i.e. for $\boldsymbol{v'} \neq 0$. The direction of $\boldsymbol{F'}_{\text{cor}}$ is perpendicular to the axis of rotation $\boldsymbol{\omega}$ and the direction of motion $\boldsymbol{v'}$ in the rotating frame. Therefore the coriolis force acting on an object moving parallel to the axis of rotation is zero.

See http://youtu.be/_36MiCUS1ro for a film clip showing an example of the coriolis force acting on a ball being rolled between people sitting on

a roundabout. Consider 2 different frames of reference: inertial frame A looking at the roundabout from above and non inertial (rotating) frame B on the roundabout. The goal (friend) the ball is aimed at is fixed on the edge of the rotating disc so this goal is fixed for an observer in B but rotating for an observer in A. Observer B aims straight for the goal but misses because the path the ball follows curves away from the goal! Observer A sees the ball rolling in a straight line but the goal rotates away. From B's perspective there's a mysterious force pulling the ball away. This is the coriolis force. Note it is not a real physical force (it is just due to the rotation) and is therefore sometimes referred to as the "coriolis effect" rather than coriolis force.

Objects moving on the surface of the Earth will experience the coriolis effect because the Earth is rotating and therefore we live in a rotating frame of reference. The rotation of the Earth has an angular velocity $\boldsymbol{\omega}$ where the vector points North along the axis of rotation (the axis from the South to North pole). The magnitude $|\boldsymbol{\omega}| = \omega = \frac{2\pi}{23\,\mathrm{hrs}56\,\mathrm{mins}}$ is related to the time period $T = 1\,\mathrm{day}$. e.g. the direction of the coriolis force can be worked out by using the right hand rule. The coriolis force on an object moving East along the equator is radially up out from the earth. $\boldsymbol{F'}_{\mathrm{cor}}$ on an object moving North to South crossing the equator is zero since this is parallel to the axis of rotation. $\boldsymbol{F'}_{\mathrm{cor}}$ of an object moving along the surface across the North pole is along the surface perpendicular to the direction of motion. If you jump upwards at the equator the $\boldsymbol{F'}_{\mathrm{cor}}$ experienced will be west.

3.5.1 Cyclones

An important example of the coriolis effect is found within meteorology. Air flows from high to low pressure but in the non inertial rotating frame of the Earth the coriolis force acts on the moving air deflecting it. This results in air moving in a circle/spiral and is the basis of cyclones. Once convection is added to this a cyclone can turn into a tropical cyclone/hurricane with devastating results.

A cyclone is a low pressure system with low pressure in the centre surrounded by high pressure. In the Northern hemisphere the coriolis force causes the air to rotate anticlockwise whereas in the Southern hemisphere the rotation is clockwise.

A high pressure system (high pressure in the centre surrounded by low pressure) is known as an anticyclone and the direction of rotation is opposite. i.e. clockwise in the Northern hemisphere and anticlockwise in the Southern hemisphere.

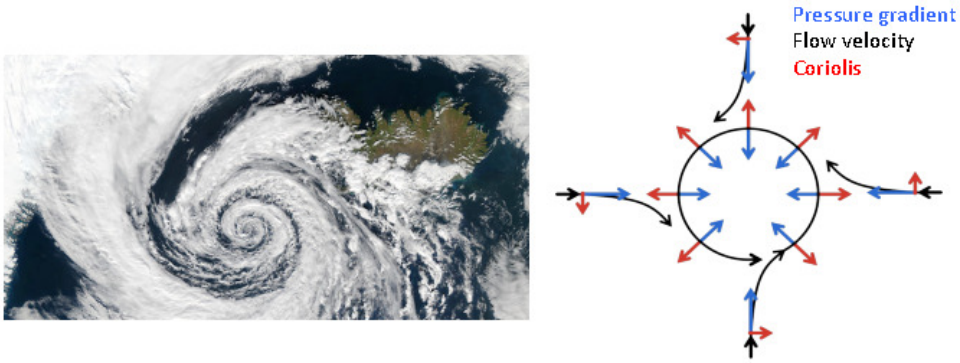

Figure 3.4: Low pressure system (cyclone) over Iceland (Northern Hemisphere). The blue arrows indicate the pressure gradient, black arrows the air flow velocity and red the coriolis force.

Topic 4

Lagrangian mechanics

4.1 What's difficult about Newton?

4.1.1 Many body problems

You're all very familiar with Newton's second law, $F = ma$ which of course should be written as $\boldsymbol{F} = m\boldsymbol{a} = m\ddot{\boldsymbol{r}}$ in vector notation for systems in more than one dimension. In 3D this vector equation is actually a short way of writing 3 scalar equations, one for each dimension. But if there is more than one body (object) in the system then Newton's law is even more complicated:

$$\boldsymbol{F}_i = m_i\boldsymbol{a}_i = m_i\ddot{\boldsymbol{r}}_i \tag{4.1}$$

for body i. So if there are n bodies in 3D there will be $3n$ equations.

But what is $\boldsymbol{F}_i$? In general it depends on the distance of body i from all other bodies in the system and their masses (and maybe their charges too). ie.e $\boldsymbol{F}_i = \boldsymbol{F}_i(\boldsymbol{r}_1, \boldsymbol{r}_2 \ldots \boldsymbol{r}_n)$. To solve Newton's law (integrate it) we need to know the position $\boldsymbol{r}_i$ and mass m_i of all the bodies and the initial conditions (positions and velocities at $t = 0$). $\boldsymbol{F}(\boldsymbol{r})$ but since $\boldsymbol{F} = m\boldsymbol{a}$, $\boldsymbol{F}$ changes the positions $\boldsymbol{r}_i$ and from Newton's first law action=reaction, we end up with a highly coupled set of equations which are therefore complex. Imagine one particle moving through other particles generating force to accelerate our particle of interest. As our particle accelerates the other particles will experience equal and opposite forces and therefore also accelerate. The simple looking equation $F = ma$ is a snapshot at time t but at time $\mathrm{d}t$ later everything will have changed.

If the number of bodies $n = 1$ or $n = 2$ Newton's equations can be integrated analytically (for central forces like gravitational/electromagnetic forces). If $n > 2$ e.g. the 3 body problem in general there is no analytical

solution (unless specific assumptions are made e.g. all masses equal or some masses negligible). For many body problems it becomes impossible to track the motion of every body therefore new concepts are needed i.e. thermodynamics and statistical physics as you are learning in PHY250.

What's difficult about forces?

$\boldsymbol{F} = m\boldsymbol{a}$ is a vector equation and is therefore best done in Cartesian coordinates, but many forces (e.g. gravitational/electrostatic) depend on distance only and are therefore best done in spherical coordinates. This is why Lagrange reformulated Newton's equations of motion in terms of energies instead of forces. Why are energies easier than forces? Because energies are scalars not vectors so we don't have to use Cartesian coordinates, we can use whatever coordinates we like to make the problem simpler. We'll find that sometimes we only need 1 or 2 coordinates to describe the system.

4.2 A simple example using energies

Let us consider a simple example that we know how to do in Newtonian mechanics using forces and solve it using energies instead. Consider a free falling body due to gravity. Let s be the distance fallen and the initial conditions be $s(0) = v(0) = 0$. First let's solve it the way we know using forces:

$$\begin{aligned} F &= ma \\ mg &= m\ddot{s} \\ \frac{\mathrm{d}^2\mathrm{s}}{\mathrm{dt}^2} &= g \\ v &= \frac{\mathrm{ds}}{\mathrm{dt}} = gt + A \\ s &= \frac{1}{2}gt^2 + B \\ s &= \frac{1}{2}gt^2 \end{aligned} \tag{4.2}$$

where the constants A and B are zero due to the initial conditions. Now let us solve this problem using energies. The energies involved are potential energy $V = mgh = -mgs$ and kinetic energy $T = \frac{1}{2}mv^2$. Conservation of

energy means that $T + V$ =constant. The initial conditions tell us that at $t = 0\ T(0) = V(0) = 0$ and therefore $T + V = 0$ so:

$$\begin{aligned} T &= -V \\ \frac{1}{2}mv^2 &= mgs \\ v &= \frac{\mathrm{ds}}{\mathrm{dt}} = \sqrt{2gs} \\ \int \frac{1}{\sqrt{s}}\mathrm{ds} &= \int \sqrt{2g}\mathrm{dt} \\ 2s^{1/2} &= \sqrt{2g}t \\ 4s &= 2gt^2 \\ s &= \frac{1}{2}gt^2 \end{aligned}$$

giving the same answer as before (4.2). So we can get the right answer without thinking about forces at all!

Lagrangian mechanics is a generalised formalism for treating mechanical problems in terms of energies instead of forces. The key principle it is based on is conservation of energy. In this course we will only consider conservative forces, i.e. forces that conserve mechanical energy i.e. kinetic energy + potential energy is constant. Non conservative forces e.g. friction are harder and we won't deal with them.

4.3 The Lagrangian

The Lagrangian L is given by:

$$\boxed{L = T - V}$$

Where T is the kinetic energy and V is the potential energy. Note L is not the total energy. The total energy is called the Hamiltonian $H = T + V$. The Hamiltonian you use in quantum mechanics comes from the classical mechanics Hamiltonian.

4.4 Degrees of freedom

Lagrangian mechanics works with any set of coordinates as long as they are sufficient to completely specify the state of the mechanical system (i.e.

the position and orientation of every body). In the absence of constraints there are 3 coordinates per point like body (and more if not point like to specify orientation). With constraints fewer coordinates are needed. The minimum number of coordinates required to completely describe the state of the system is equal to the number of degrees of freedom N. The number of degrees of freedom is independent of the coordinate system we choose. N may be surprisingly small e.g. a pendulum, a cylinder rolling down a slope and an Atwood machine (a pulley with 2 masses) all only have one degree od freedom.

4.5 Generalised coordinates

Because we don't want to constrain which coordinate system we use we talk about "generalised coordinates" q_i and generalised velocities $\dot{q}_i = \frac{\mathrm{dq_i}}{\mathrm{dt}}$. For a system with N degrees of freedom there will be N generalised coordinates to completely describe the state.

Often there's other sets of coordinates describing the system using more then N coordinates. If this is so we can always eliminate one or more coordinates to reduce the number of coordinates to N. e.g. a pendulum could be described with the Cartesian coordinates x, y of the position of the mass but these can be reduced to one generalised coordinate such as the angle ϕ where $\tan\phi = \frac{x}{l-y}$, $x = l\sin\phi$ and $y = l(1-\cos\phi)$.

Note generalised coordinates do not necessarily have dimensions of length L. The could have any or no dimensions (e.g. an angle).

The set of N generalised coordinates $q_1, q_2 \ldots q_N$ completely describes the state of the mechanical system where N is the number of degres of freedom of the system. Mathematically such a set is called a holonomic set. There may be more than one such set of generalised coordinates (e.g. for a free body in space Cartesian, spherical or cylindrical). We should pick the one appropriate for the symmetry of our system.

In the following we will formulate the potential energy and kinetic energy in terms of these generalised coordinates Q_i.

4.6 Potential energy V

The potential energy V can be expressed in terms of the generalised coordinates (of all bodies in space) and time i.e.

$$V = V(t, q_1 \ldots q_N)$$

Potential energy may (e.g. for charged particles in an AC (alternating current) field) or may not (e.g. gravity) depend explicitly on time. Potential energy will usually depend on time implicitly even if not explicitly i.e. since some generalised coordinates depend on time. Potential energy does not depend on velocities.

Note the dimensions of V are that of energy i.e. $\mathsf{ML^2T^{-2}}$ units J. This is the case whatever weird dimensions the q_i may have.

We often choose the q_i to make V as simple as possible (e.g. for a potential that depends only on distance r use spherical coordinates and $V(r)$ doesn't depend on θ, ϕ). But this might make the kinetic energy more complicated so sometimes we choose q_i to make the kinetic energy simpler at the expense of a more complicated potential energy.

Potential energy is always specified up to an arbitrary constant i.e. we can always add a constant V_0 i.e. we can choose where we want $V = 0$. This is because adding a constant V_0 to the potential energy leads to the same equations of motion. The most theoretically satisfying choice would be to have $V = 0$ at infinity since there is no interaction when objects are infinitely far apart. In this case then at any finite energy $V < 0$. But we don't always want this and sometimes it's not even possible (e.g. there is not infinite distance for a pendulum) so it may be more convenient to use other conventions e.g. $V = 0$ at the lowest point of the pendulum or $V = 0$ at time $t = 0$.

Be **careful about the sign of** V: If $V = 0$ at infinity, any other distance will have V negative. If $V = 0$ at some other point V can be positive or negative. The important thing to remember to get the sign of V right is that V *decreases downhill*. Getting the sign of V wrong is one of the most common mistakes in Lagrangian problems - be careful!

Remember if there's more than one body in the system we have to add the potential energy for each body i.e. $V = V_1 + V_2 + \ldots V_n$ for n bodies.

4.7 Kinetic energy T

Kinetic energy can be expressed in terms of the generalised coordinates q_i and the generalised velocities $\dot{q}_i = \frac{\mathrm{dq_i}}{\mathrm{dt}}$ as:

$$T = T(q_i, \ldots q_N, \dot{q}_1, \ldots \dot{q}_N)$$

Note T never depends on time t explicitly but it does depend on time implicitly because the generalised coordinates depend on time.

As you know, kinetic energy $T = \frac{i}{2}mv^2$. But we have to be careful. The dimensions of generalised velocities are not necessarily that of velocities

LT^{-1}, depending on the dimensions of q_i i.e. generalised velocities are not always 'proper' velocities e.g. if $q_i = \phi$ the angle of a pendulum mass then $\dot{q}_i = \frac{\mathrm{d}\phi}{\mathrm{dt}} = \omega$ the angular velocity not the velocity. So you can't put this $\dot{q}_i$ into $T = \frac{1}{2}mv^2$ in the place of v. This is the other most common mistake in Lagrangian problems. T must end up with dimensions of energy $\mathsf{ML}^2\mathsf{T}^{-2}$ units J. e.g. for the pendulum we need to multiply $\dot{\phi} = \frac{\mathrm{d}\phi}{\mathrm{dt}}$ by the length l of the string to get the velocity v and then $T = \frac{1}{2}m(l\dot{\phi})^2$ i.e. what we've done is:

$$\begin{aligned} T = \frac{1}{2}mv^2 &= \frac{1}{2}m(\dot{x}^2 + \dot{y}^2) \\ &= \frac{1}{2}ml(\cos^2(\phi)\ \dot{\phi}^2 + \sin^2(\phi)\ \dot{\phi}^2) \\ &= \frac{1}{2}ml^2\dot{\phi}^2 \end{aligned}$$

If in doubt work out the velocity v by transforming from Cartesian coordinates as above.

A more complicated example is a spring-pendulum where we replace the pendulum string with a spring. The system has $N = 2$ degrees of freedom so 2 generalised coordinates, let use use the length of the spring, l, and the angle of the pendulum, ϕ. There will be 2 components to the velocity: tangential $v_t = l\dot{\phi}$ and the direction of the extending spring $v_l = \frac{\mathrm{dl}}{\mathrm{dt}} = \dot{l}$. Note the tangential velocity and therefore kinetic energy depends of the generalised coordinate l as well as the generalised velocities. Here v_t is perpendicular to v_l so $v^2 = v_t^2 + v_l^2$ and therefore here we have $T = \frac{1}{2}m(l^2\dot{\phi}^2 = \dot{l}^2)$. Be careful: we can only add the squares of velocity components if they are perpendicular, otherwise we have to convert from Cartesian coordinates to do the dot product $\dot{\boldsymbol{q}} \cdot \dot{\boldsymbol{q}}$ properly.

Note T is always positive in classical physics. If there is more than one body in the system don't forget to add the kinetic energy for each body: $T = T_1 + T_2 + \ldots T_n$ for an n body system. Note that even for one body there may be more than one type of kinetic energy to add e.g translational plus rotational etc.

4.8 Hamilton's (variational) principle/Principle of least action

A mechanical system takes the path in generalised coordinate space $q_i(t_1) \rightarrow q_i(t_2)$ that minimises the action integral I

$$I = \int_{t_1}^{t_2} L(q_i, \dot{q}_i, t)\mathrm{dt} \tag{4.3}$$

where $L = T - V$ is the Lagrangian. I is a *path integral* so its value depends on the path taken. The path for which I is minimum (strictly extremum but usually minimum) is the path the system actually takes. You could think of this as the principle of universal laziness! Any system takes the 'easiest' path.

4.9 Derivation of Lagrange's equations

Here we use the principle of least action (Hamilton's variational principle) to derive Lagrange's equations of motion. Let $\bar{q}(t)$ be the path for which I is minimum and let $\delta q(t)$ be a small variation in this path. Let $\delta q(t_1) = \delta q(t_2) = 0$ so the modified path and the optimal path start and finish at the same place. Then let $q = \bar{q}(t) + \delta q(t)$. The action integral I from equation (4.3)

$$I(q) = \int_{t_1}^{t_2} L(\bar{q} + \delta q, \dot{\bar{q}} + \dot{\delta q}, t)\mathrm{dt}$$

where for convenience we drop the subscript i on the q_i. Using the Taylor series we expand this for small δq:

$$\begin{aligned}
I(q) &= \int_{t_1}^{t_2} \left[L(\bar{q}, \dot{\bar{q}}, t) + \delta q \frac{\partial L}{\partial q} + \dot{\delta q} \frac{\partial L}{\partial \dot{q}} + \dots \right] \\
&= \int_{t_1}^{t_2} L(\bar{q}, \dot{\bar{q}}, t)\mathrm{dt} + \int_{t_1}^{t_2} \delta \mathrm{q} \frac{\partial \mathrm{L}}{\partial \mathrm{q}} \mathrm{dt} + \int_{t_1}^{t_2} \dot{\delta} \mathrm{q} \frac{\partial \mathrm{L}}{\partial \dot{\mathrm{q}}} \mathrm{dt} \\
&= I(\bar{q}) + \int_{t_1}^{t_2} \delta q \frac{\partial L}{\partial q} \mathrm{dt} + \left[\delta \mathrm{q} \frac{\partial \mathrm{L}}{\partial \dot{\mathrm{q}}} \right]_{t_1}^{t_2} - \int_{t_1}^{t_2} \delta \mathrm{q} \frac{\mathrm{d}}{\mathrm{dt}} \frac{\partial \mathrm{L}}{\partial \dot{\mathrm{q}}} \mathrm{dt} \\
I(q) - I(\bar{q}) &= \int_{t_1}^{t_2} \delta q \left[\frac{\partial L}{\partial q} - \frac{\mathrm{d}}{\mathrm{dt}} \left(\frac{\partial L}{\partial \dot{q}} \right) \right] \mathrm{dt}
\end{aligned}$$

where we have used integration by parts and the boundary conditions $\delta q(t_1) = \delta q(t_2) = 0$. To minimise (extremise) we want $I(q) - I(\bar{q}) = 0$

and therefore

$$\int_{t_1}^{t_2} \delta q \left[\frac{\partial L}{\partial q} - \frac{\mathrm{d}}{\mathrm{dt}} \left(\frac{\partial L}{\partial \dot{q}} \right) \right] \mathrm{dt} = 0$$

for all δq therefore:

$$\frac{\partial L}{\partial q} - \frac{\mathrm{d}}{\mathrm{dt}} \left(\frac{\partial L}{\partial \dot{q}} \right) = 0$$

Replacing the subscript i gives Lagrange's equations of motion:

$$\boxed{\frac{\mathrm{d}}{\mathrm{dt}} \left(\frac{\partial L}{\partial \dot{q}} \right) = \frac{\partial L}{\partial q}} \tag{4.4}$$

There is one equation for each i i.e. for each generalised coordinate so there are N equations where N is the number of degrees of freedom.

4.10 Newton's equations from Lagrange's

Langranges's equations (4.4) are equivalent to Newton's equations of motion and here we show this by deriving Newton's equations from Lagrange's equations. Let us consider a system of n particles in a potential $V(r_i)$ where r_i is the distance of particle i. The generalised coordinates are $q_i = r_i$ and generalised velocities $\dot{q}_i = v_i$. The Lagrangian is then

$$L = T - V = \sum_i^n \frac{1}{2} m_i v_i^2 - V(r_i)$$

Putting this into Lagrange's equations (4.4) gives

$$\frac{\mathrm{d}}{\mathrm{dt}} \left(m_i v_i \right) = -\frac{\partial V}{\partial r_i}$$
$$m_i a_i = F_i$$

i.e. Newton's equations of motions $F = ma$. Due to this equivalence we call $\frac{\partial L}{\partial q_i}$ a generalised force and $\frac{\partial L}{\partial \dot{q}_i}$ generalised momentum p_i conjugate to the generalised coordinate q_i.

4.11 Constants of motion & cyclic/ignorable coordinates

In some cases L may not depend on one of the generalised coordinates q_i. If this is the case the missing generalised coordinate is called **ignorable**

or **cyclic** i.e. if $\frac{\partial L}{\partial q_i} = 0$ then q_i is an ignorable/cyclic coordinate. Putting this into Lagrange's equations (4.4) gives $\frac{\mathrm{d}}{\mathrm{dt}}(\frac{\partial L}{\partial \dot{q}_i}) = 0$ i.e. $\frac{\partial L}{\partial \dot{q}_i}$ =constant i.e. the generalised momentum is constant – the generalised momentum is a **constant of motion**. Often this reproduces a well known conservation law e.g. conservation of momentum or conservation of angular momentum.

Appendix A

Dimensional analysis

A.1 Units and dimensions

What's the difference between equations in physics and maths?

One key difference is that physical equations usually relate quantities that have units as well as numbers.

- Mathematically, $3 + 2 = 5$, end of story.
- Physically, e.g. $3\ \text{apples} + 2\ \text{pears} =$?

3 apples plus 2 pears does not give 5 apples nor 5 pears. We cannot sensibly add or equate quantities that have different units (strictly speaking different dimensions as we will see later).

Exercise. *Can we multiply or divide quantities of different units?*

The requirement for quantities being added, subtracted or equated to have the same units/dimensions gives us a consistency check on all physical equations we may have derived, or recalled from memory. If we have an equation that adds, subtracts, or equates quantities of different units we know something must have gone wrong! This consistency check is very powerful and useful, especially in exams. Always keep track of the units as well as the numbers through your calculations and check if the units come out right at the end. If not (e.g. you find the height of the Eiffel tower is 324m^2), you know there must be a mistake somewhere. In fact, often you may get a clue as to what the problem is too e.g. if you try to calculate a length, but the units work out to be m^2, or m^{-1}, you probably forgot to take a square root or an inversion somewhere.

Exercise. *What is the formula for centrifugal acceleration? Let's say you remember it is either* $a = \omega r^2$ *or* $a = \omega^2 r$ *where* ω *is the angular velocity and* r *the distance from the axis of rotation. Which equation is correct?*

A generalisation of the concept of units is that of dimension. NB there are different meanings of the word 'dimension' even within physics. Here we're not talking about dimensions in the sense of 3D space. Here dimension means the type of quantity as is clear in the following example:

Exercise. *Spot the odd one out in the following list: meters, cm, nanometers, kilometres, feet, seconds, inches, miles.*

Obviously apart from seconds they are all are units of length. We therefore say that any quantity that is measured in a unit of length has the 'dimension' length, L, whichever unit that may be (m, km, miles etc.). You may prefer imperial or metric, micrometers or miles, all of these still have something in common, which a second has not. That 'something' is the dimension L. Similarly, we introduce the dimension 'time', T to anything measured in seconds, hours, days, years, or any other unit of time, and the dimension 'mass', M to anything measured in kg, micrograms, etc. (NB be careful with imperial units here, they don't clearly distinguish between mass and weight. A good reason to avoid them altogether!)

From these three dimensions, we can make up the dimensions of less basic quantities such as velocity, or force. Velocity is distance/time, so it has dimensions L/T or LT^{-1}. So km/h, miles/hour or $\mathrm{m\,s}^{-1}$ are all different units of velocity, but all have the same dimension of LT^{-1}. Note this also works for differentials i.e. whether velocity v is defined as $v = \frac{\Delta x}{\Delta t}$ or $v = \dot{v} = \frac{\mathrm{dx}}{\mathrm{dt}}$, it still has units m/s (or km/h etc.), and dimensions LT^{-1}. Acceleration has dimensions LT^{-2}, Force is given by $F = ma$, so we multiply the dimensions of mass and acceleration to get the dimensions of force. Similarly work (energy) is $W = Fs$ (force times distance) so we can work out its dimensions. Whenever you know either a defining equation, or the units of a quantity, you can work out the dimensions.

Exercise. *What are the dimensions of acceleration, force, energy, density?*

NB More dimensions are needed to deal with some situations such as electrical phenomena e.g. electrical charge Q, but we won't worry about this here.

Note that an equation equating two physical quantities with different units can still be correct, as long as it equates quantities with the same dimensions e.g. a speed limit $50\,\mathrm{km/h} = 31\,\mathrm{miles/hour}$ is a **dimensionally consistent** equation.

A.2 Dimensional analysis

Sometimes we can use the fact that equations must be dimensionally consistent, not just to check whether an equation is correct but, to derive an equation from dimensional considerations alone. This is dimensional analysis. Let's try this for the example of the angular frequency ω of a pendulum. The 'proper' method of finding this of course is to write its equation of motion and solve it (assuming small amplitude oscillations). Instead of doing this, let's guess what ω will depend on: length, l, of the pendulum, mass, m, of the pendulum bob and acceleration due to gravity g. Now, we assume we can make up the correct equation for ω by simply multiplying these factors together, each taken to an unknown power:

$$\omega = l^a m^b g^c \tag{A.1}$$

We now find the powers a, b and c by making the equation dimensionally consistent. To do this we write a **dimensional equation** by replacing the quantities ω, l, m and g in equation (A.1) by their dimensions giving:

$$\mathsf{T}^{-1} = \mathsf{L}^a \mathsf{M}^b \left(\frac{\mathsf{L}}{\mathsf{T}^2}\right)^c \tag{A.2}$$

where we have used the dimension of acceleration, LT^2. Comparing the left hand side (LHS) and right hand side (RHS) of equation (A.2) we obtain a set of equations for a, b and c. No factor L appears on the LHS, which means it has power zero and so it must have power zero on the RHS too. Equating the powers of the other dimensions leads to the following set of equations:

$$\begin{aligned} 0 &= a + c \\ 0 &= b \\ -1 &= -2c \end{aligned} \tag{A.3}$$

This is easily solved giving $b = 0$, $c = \frac{1}{2}$ and $a = -\frac{1}{2}$. Substituting these back into equation (A.1) gives:

$$\omega = l^{-1/2} m^0 g^{1/2} = \sqrt{\frac{g}{l}} \tag{A.4}$$

Recognise this? This is the correct equation, at least in the limit of small amplitudes, found from the 'proper' derivation. In particular, we have shown that the pendulum period does NOT depend on the mass of the bob. And all without bothering with the full derivation - like magic!

Where's the catch? Well there are potential problems with dimensional analysis. What if there's a dimensionless factor e.g. 2 or π? With dimensional analysis we can never derive any factors in equations that have no units (are of dimension 1). In our example the unknown numerical factor happens to be one, phew lucky for us! But we won't be that lucky every time. But what about the amplitude, $\phi_{\max}$ of the pendulum? In theory this might affect ω but this did not come into our dimensional analysis. Why not? $\phi_{\max}$, is an angle. The units of an angle are radians (rad), but if you recall the definition of rad, you'll find it is a length (arc length) divided by another length (radius) and therefore it has no units or dimensions at all. Quantities that have no units are called **dimensionless**. That does not mean that $\phi_{\max}$ cannot affect ω, but it means that dimensional analysis cannot tell us if it does or how it does. This is a more serious problem than that of unknown numerical factors because $\phi_{\max}$ is a variable. We should write equation (A.4) as

$$\omega = f(\phi_{\max})\sqrt{\frac{g}{l}} \tag{A.5}$$

where $f(\phi_{\max})$ is an unknown function, which we cannot determine from dimensional reasoning. All we know is that $f(\phi_{\max})$ has to be dimensionless, just like its argument, $\phi_{\max}$. As it happens here, for small amplitudes, $f(\phi_{\max}) \to 1$, again lucky for us!

In summary, dimensional analysis separates a problem into dimensional, and non-dimensional parts. It then solves (or, as we will see, sometimes only partly solves) the dimensional part but it can't solve the non-dimensional part. Dimensional analysis is therefore also sometimes called non-dimensionalisation.

A.3 Dimensionless quantities

Sometimes called dimensionless groups, dimensionless variables or numbers, dimensional quantities are combinations of variables/quantities that together have an overall dimension of one. i.e. are dimensionless. They are particularly used in situations where we don't yet have a full theory (therefore often in research). This is because the complexity of a problem can be reduced by combining some of the variables into dimensionless groups such that the number of dimensionless variables is less than the number of relevant variables we started with.

A practically very important example of such an application of dimensional analysis is in fluid dynamics (hydrodynamics and aerodynamics).

Predicting the drag force, F, experienced by an object of a known size and shape moving in water or air, or a fluid flowing through a pipe, is extremely difficult. There is, as yet, no general solution. However dimensional analysis can significantly simplify the amount of experimental work required to study drag. Unlike in the example of the angular frequency of the pendulum, we will not even be able to solve the dimensional part of the hydrodynamic drag problem completely by dimensional analysis, but we will be able to reduce the complexity of the problem by combining some of the relevant quantities into dimensionless groups. The number of dimensionless variables will be smaller than the number of relevant quantities.

Hydrodynamics assumes a body that is totally submerged in (or a pipe completely filled by) an incompressible fluid. This assumption of incompressibility is OK for most liquids. The relevant quantities that determine drag force are then the density, ρ, of the fluid, the velocity, v between fluid and object (or pipe), the viscosity, η, of the fluid and the size, l, of the object. For compressible fluids (e.g., air in aerodynamics), the compressibility κ also comes into it, but for simplicity, we will work with the example of an incompressible fluid here.

Exercise. *What are the dimensions of force, density, velocity, viscosity, and size?*

We assume the drag force to be a function of all the relevant quantities in the form:

$$F = Al^a\rho^b\eta^c v^d \tag{A.6}$$

Where A is a dimensionless factor that will depend on the (dimensionless) shape of the object (or pipe). Eq. (A.6 translates into the dimensional equation

$$\mathsf{MLT}^{-2} = \mathsf{L}^a(\mathsf{ML}^{-3})^b(\mathsf{ML}^{-1}\mathsf{T}^{-1})^c(\mathsf{LT}^{-1})^d \tag{A.7}$$

leading to

$$\begin{aligned} 1 &= b + c \\ 1 &= a - 3b - c + d \\ -2 &= -c - d \end{aligned}$$

Can we solve these? No! These are 3 equations for 4 unknowns, so there cannot be a complete solution. However, it is possible to express 3 of them all in terms the 4th, leaving us with a single unknown. Of course, it is somewhat arbitrary which power you leave as the remaining unknown,

and different choices lead to different dimensionless groups. Here we will take the one that is easiest, c, and write the others all in terms of c

$$\begin{aligned} b &= 1 - c \\ d &= 2 - c \\ a &= 2 - c \end{aligned} \tag{A.8}$$

Now, we substitute a, b and d in eq. (A.6):

$$F = Al^{2-c}\rho^{1-c}\eta^{c}v^{2-c}$$

and put all the known powers on the left hand side (LHS) and unknowns together on the right hand side (RHS):

$$\frac{F}{l^2\rho v^2} = A\left(\frac{l\rho v}{\eta}\right)^{-c} \tag{A.9}$$

$(\frac{l\rho v}{\eta})$ should be dimensionless, otherwise there would be trouble raising it to unknown power c as we might end up with strange dimensions (e.g. what if $c = \sqrt{17}$?) If $(\frac{l\rho v}{\eta})$ is dimensionless, then the LHS of eq. (A.9) must also be dimensionless. What we have done is to rewrite eq. (A.6) in simpler terms with the variables lumped together into dimensionless groups, which are sometimes called just numbers. These particular dimensionless quantities have special names:

$$\frac{F}{l^2\rho v^2} = \mathrm{Ne} = \text{Newton number}$$

$$\frac{\rho v l}{\eta} = \mathrm{Re} = \text{Reynold's number} = \frac{\text{inertial force}}{\text{viscous force}}$$

The Reynold's number tells us the ratio of the inertial forces over the viscous forces. A low Reynold's number corresponds to small objects in very viscous fluids e.g. us in honey or bacteria in water. High Reynold's number corresponds to large objects in runny liquids e.g. us in water or a hippo in mud.

Exercise. *Check directly that Ne and Re are both dimensionless.*

Note the point of all this is that, we started with force being a function of four unknown powers, but finished with a number (Ne) being an unknown power of only one variable, Re, i.e. we now have the much simpler equation:

$$\mathrm{Ne} = A\mathrm{Re}^{-c}$$

To be totally general and cover our backs we should say

$$\mathrm{Ne} = Af(\mathrm{Re}) \tag{A.10}$$

where $f(\mathrm{Re})$ is an unknown function which could be a potentially complicated function rather than a simple power (because e.g. $\mathrm{Re}^{-c} + \mathrm{constant}$ is still dimensionless)

Dimensional analysis can't tell us anything about A, which depends on the shape of the object and so has no dimensions. But we do know that A will depend on shape only and is independent of size, velocity, etc. Also, dimensional analysis cannot give us the unknown function, f but we do know it's only a function of one variable, Re. This significantly reduces the effort required if we want to measure f experimentally, because we have reduced the number of independent variables to one. For example if we measure the hydrodynamic behaviour of water in a pipe, then we know the behaviour or oil in a pipe of the same shape. The relationship between Ne and Re will be the same, we just have to calculate Ne and Re with the appropriate different density, viscosity, velocity, size. Similarly the behaviour of a cm sized model submarine in a tank can tell us how the full sized submarine will behave in the sea.

A particularly useful application of dimensional analysis is **scaling**. Dimensional analysis allows us to do experiments on scaled-down models, rather than the full sized objects. As long as the shape of the object is kept the same, dimensionless equations such as eq. (A.10) remain valid — we just use the different length l to calculate Ne and Re. The potential savings in effort and cost are obvious in e.g. ship or aircraft design. Dimensional analysis tells us how to relate experimental conditions and results between the model and the full size version.

It is worth noting that the definition of a dimensionless quantity is somewhat arbitrary. The inverse, square, square root, or another power of a dimensionless quantity will again be dimensionless. Therefore, there is some ambiguity in the definition of a dimensionless group. Instead of taking $\mathrm{Re} = \frac{\rho v l}{\eta}$ we could have taken Re^2 or $1/\mathrm{Re}$ as our dimensionless variable, or even $2\pi\mathrm{Re}$ as these would all be still dimensionless. An unknown power c of Re is also an unknown power of Re^2 or $1/\mathrm{Re}$ (a different one, but still unknown). Some dimensionless variable definitions are conventions due to historic reasons, but often a particular form of a dimensionless group is chosen because it has a clear physical interpretation. An example is the Mach number in aerodynamics, which can be interpreted as the ratio of the speed of an object to the speed of sound. However, Mach number does not immediately emerge in that form from dimensional analysis. Ac-

tually, a different dimensionless quantity emerges and the Mach number is a power of that quantity.

Finally we should note that dimensional analysis can be used in all areas of physics, including quantum mechanics, not just in classical physics. All physical equations must be dimensionally consistent.

Exercise. *Show that the Schrödinger equation is dimensionally consistent.*

Appendix B

Answers to exercises

Coming soon...