1
Introduction

1.1 Newtonian Gravity

No student of physics can fail to notice the similarity between Coulomb's law for the electrostatic force between two point charges q_1 and q_2 a distance r apart,

$$\mathbf{F} = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r^2}\hat{\mathbf{r}}, \tag{1.1}$$

and the gravitational force between two point masses m_1 and m_2 a distance r apart, according to Newton's universal law of gravitation,

$$\mathbf{F} = -\frac{G m_1 m_2}{r^2}\hat{\mathbf{r}}. \tag{1.2}$$

($\hat{\mathbf{r}} = \mathbf{r}/r$ is a unit vector pointing from particle 1 to particle 2, and these are the forces on particle 2 due to particle 1.) Newton published his universal law of gravitation in 1686 (though Robert Hooke appreciated the significance of the inverse square law for planetary motion before Newton and accused Newton of plagiarism), and Coulomb published his law almost exactly 100 years later, in 1785. Coulomb must have been very excited by his discovery; Equations (1.1) and (1.2) appear to imply a very close connection between electrostatics and gravity.

Even in the static case there are obvious differences, of course; the prefactor $\frac{1}{4\pi\epsilon_0}$ in Coulomb's law[1] is replaced with $-G$ in Newton's law, but this is partly just a choice of units. Instead of measuring electric charge in coulombs (the MKSA unit of charge), we could use a different set of units with $\tilde{q} = \frac{q}{\sqrt{4\pi\epsilon_0}}$ and then Coulomb's law reads

$$\mathbf{F} = \frac{\tilde{q}_1 \tilde{q}_2}{r^2}\hat{\mathbf{r}}, \tag{1.3}$$

[1] ϵ_0 goes by the unfortunate name of 'the electric permittivity of the vacuum'.

https://doi.org/10.1017/9781009263689.002 Published online by Cambridge University Press

and instead of measuring mass in kilograms we could define $\widetilde{m} = \sqrt{G}\, m$, in which case Newton's law of gravitation reads

$$\mathbf{F} = -\frac{\widetilde{m}_1 \widetilde{m}_2}{r^2}\hat{\mathbf{r}}. \tag{1.4}$$

($\tilde{q}$ and $\widetilde{m}$ have the same units, kilograms$^{1/2}\times$ metres$^{3/2}$/second.) They look even more similar – apart from that minus sign. This mathematical similarity, however, hides the physical fact that the forces are of very different magnitudes. In MKSA units Newton's universal constant of gravitation is $G = 6.67 \times 10^{-11}$ kg^{-1}m^{3}s^{-2} while the electric permittivity of the vacuum is $\epsilon_0 = 8.85 \times 10^{-12}$C^{2}m^{3}s^{-2}kg^{-1}. The relative strengths of the gravitational to the electrostatic force between two electrons, with $q_1 = q_2 = -1.6 \times 10^{-19}$C and $m_1 = m_2 = 9.1 \times 10^{-31}$kg, is $4\pi\epsilon_0 G \left(\frac{9.1\times10^{-31}}{1.6\times10^{-19}}\right)^2 \approx 10^{-43}$, a dimensionless number which is independent of the units used. Gravity is an extremely weak force, which is why a small magnet can beat the gravitational attraction of the entire Earth and lift a metal pin off a table. Nevertheless, gravity dominates the Universe on large scales, as there is only one sign for the gravitational 'charge', m, while it is a fact from experimental observation that electric charge q can be either positive or negative, like charges repel and unlike charges attract, while m always seems to be positive and, because of that minus sign, all masses attract under Newton's gravitational force. With an equal number of positive and negative charges in the Universe, electrostatic forces cancel out in the large, while gravitational forces are cumulative and dominate on astrophysical scales.

Nevertheless, the mathematical similarity between (1.3) and (1.4) makes it very tempting to think there must be some deep relation between electromagnetism and gravity that remains to be uncovered, if only we could see a little more deeply into the nature of the two forces. Indeed, Einstein himself spent much of his later career trying to find a unified mathematical framework: a unified theory of electricity, magnetism, and gravity. He failed to achieve his dream of a unified theory. It happens that this connection is largely illusory when we start to consider moving charges and masses, particularly with velocities approaching the speed of light: we need to consider time-varying electromagnetic and gravitational fields, and the dynamics of these two fields are very different. (Einstein was well aware of this; he was motivated by deeper considerations.) Coulomb's law and magnetostatics generalise to Maxwell's equations (1865) which unify electricity and magnetism into a single mathematical framework, the theory of electromagnetism, while

Newton's law generalises to Einstein's general theory of relativity (1916), and Einstein's equations are very different from Maxwell's equations. Both predict oscillating waves, electromagnetic waves for electromagnetism (radio waves were discovered by Hertz in 1888) and gravitational waves for general relativity (discovered more recently in 2016); but the physics, and the mathematics, of the two theories turns out to be very different – that minus sign, and the absence of negative masses, are just the tip of the iceberg.[2]

It was Michael Faraday (1791–1867) who abstracted the notion of a *field* from Coulomb's law. He suggested that an electrically charged particle generates an electric field,

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0}\frac{q_1}{r^2}\hat{\mathbf{r}},$$

and a second particle, with charge q, then experiences a force,

$$\mathbf{F} = q\mathbf{E},$$

in the presence of that field. Since the electrostatic force is conservative, we can rephrase this in terms of the electrostatic potential. The electric field $\mathbf{E}$ is the electric force per unit charge, and we define the electrostatic potential φ for a distribution of charges through

$$\mathbf{E} = -\nabla\varphi.$$

If we have N charges $q_1, \ldots, q_N$ at points $\mathbf{r}_1, \ldots, \mathbf{r}_N$, they generate an electrostatic potential at $\mathbf{r}$ given by

$$\varphi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\sum_{i=1}^{N}\frac{q_i}{|\mathbf{r}-\mathbf{r}_i|},$$

and the electrostatic force experienced by a charge q at the point $\mathbf{r}$ is

$$\mathbf{F} = -q\nabla\varphi(\mathbf{r}).$$

The electric field satisfies Gauss' law and, when expressed as a differential equation, this is

$$\nabla.\mathbf{E} = -\nabla^2\varphi = \frac{\rho}{\epsilon_0}, \tag{1.5}$$

where ρ is the charge density.

[2] When quantum mechanics is taken into consideration, the situation is even worse: a relativistic quantum theory of electromagnetism, quantum electrodynamics, or QED, for which Richard Feynman, Sin-Itro Tomonaga, and Julian Schwinger received the Nobel Prize in Physics, was fully developed in the 1940s, but a fully credible quantum theory of gravity still eludes us.

https://doi.org/10.1017/9781009263689.002 Published online by Cambridge University Press

Similar language can be used for Newtonian gravity – think of mass m_1 as creating a gravitational field,

$$\mathbf{g} = -\frac{Gm_1}{r^2}\hat{\mathbf{r}}, \tag{1.6}$$

and then another mass m experiences a force,

$$\mathbf{F} = m\mathbf{g}.$$

The gravitational force is also conservative, and $\mathbf{g}$, the acceleration due to gravity, can be defined in terms of the gravitational potential, Φ, due to a distribution of masses:

$$\mathbf{g} = -\nabla\Phi. \tag{1.7}$$

If we have N point masses $m_1, \ldots, m_N$ at points $\mathbf{r}_1, \ldots, \mathbf{r}_N$, they generate the gravitational potential

$$\Phi(\mathbf{r}) = -G\sum_{i=1}^{N}\frac{m_i}{|\mathbf{r}-\mathbf{r}_i|}, \tag{1.8}$$

and the gravitational force on a mass m at the point $\mathbf{r}$ is

$$\mathbf{F} = -m\nabla\Phi(\mathbf{r}) = m\mathbf{g}. \tag{1.9}$$

The gravitational field $\mathbf{g}$ satisfies the gravitational version of Gauss' law, Poisson's equation,

$$\nabla.\mathbf{g} = -\nabla^2\Phi = -4\pi G\rho, \tag{1.10}$$

where ρ is the mass density. Again, the similarity between (1.5) and (1.10) can be traced to the fact that the Coulomb force and the Newtonian gravitational force are both inverse square laws; the only difference is that $\frac{\rho}{\epsilon_0}$ in (1.5) is replaced with $-4\pi G\rho$ in (1.10).

Forces cause things to accelerate, and here we meet an important difference between electricity and gravity (which will turn out to be crucial): mass plays two completely different roles in Newton's universal law of gravitation and in Newton's second law. In Newton's universal law of gravitation (1.2), the gravitational field is (1.6) and m_1 is like a gravitational charge; it is a 'charge' generating a gravitational field. On the other hand, in Newton's second law, (1.9) written as

$$\mathbf{F} = m_I\mathbf{a},$$

with $\mathbf{a} = \mathbf{g}$, m_I is the inertial mass, a measure of the reluctance, or inertia, of a body to be accelerated. Two bodies with different inertial masses experiencing the same force will undergo different accelerations; the one with the larger inertial mass will accelerate more slowly.

https://doi.org/10.1017/9781009263689.002 Published online by Cambridge University Press

For example, a proton and a singly ionised atom of helium $^4He^+$ both experience the same force in a static electric field $\mathbf{E}$,

$$\mathbf{F} = q\mathbf{E},$$

where $q = 1.60 \times 10^{-19}$C is the charge of a proton. But they will consequently have different accelerations – for the proton,

$$\mathbf{a} = \frac{q}{m_I}\mathbf{E},$$

where $m_I = 1.67 \times 10^{-27}$kg is the inertial mass of a proton, while for the ionised helium,

$$\mathbf{a} = \frac{q}{M_I}\mathbf{E},$$

where M_I is the inertial mass of a helium atom (to a reasonable approximation some four times m_I). Under the *same* force, the helium atom undergoes an acceleration some four times *less* than that of the proton.

However, put the proton and the helium atom in the same gravitational field $\mathbf{g}$ and the proton will experience a force determined by Newton's Universal Law of Gravitation (1.2), analogous to $\mathbf{F} = q\,\mathbf{E}$ in electrostatics,

$$\mathbf{F} = m_G\,\mathbf{g},$$

where m_G here is the gravitational mass of the proton, while the helium atom will feel a different force,

$$\mathbf{F} = M_G\,\mathbf{g},$$

where M_G is the gravitational mass of the helium atom.

The proton then experiences an acceleration,

$$\mathbf{a} = \frac{\mathbf{F}}{m_I} = \frac{m_G}{m_I}\mathbf{g},$$

while for the helium atom,

$$\mathbf{a} = \frac{\mathbf{F}}{M_I} = \frac{M_G}{M_I}\mathbf{g}.$$

Since inertial mass equals gravitational mass,[3] $m_I = m_G$ and $M_I = M_G$, both the proton and the helium undergo *exactly* the same acceleration in the gravitational field,

$$\mathbf{a} = \mathbf{g}.$$

[3] More generally, inertial mass and gravitational mass are proportional to one another, $m_I = km_G$, with k some constant. But we are free to re-scale m_G by changing the units in which we define Newton's constant, and it is a convention, albeit a very natural one, to define G so that $k = 1$ and $m_I = m_G$. From now on we shall not distinguish between gravitational mass and inertial mass.

https://doi.org/10.1017/9781009263689.002 Published online by Cambridge University Press

The equality of gravitational and inertial mass is related to the difference between weight and mass. In elementary physics courses mass is usually defined as force divided by acceleration,

$$m = \frac{F}{a}.$$

This is really the inertial mass. Weight, on the other hand, is the force you feel on your feet when you stand on the floor, as a result of the acceleration due to gravity; the upward force is

$$F = -mg,$$

if g is downwards, and m here is the gravitational mass. In outer space, where $g = 0$, you are weightless, not massless.

This aspect of the gravitational force was understood even before Newton discovered his Universal Law of Gravitation. There is the tale (probably apocryphal) of Galileo dropping two different weights from the top of the leaning tower of Pisa. A more modern demonstration was when the Apollo 15 commander David Scott dropped a geological hammer and a falcon's feather onto the surface of the Moon and the world saw them accelerate downwards at exactly the same rate and hit the lunar surface at exactly the same time, despite the huge disparity in their masses.[4]

This equivalence of inertial mass and gravitational mass has very far reaching consequences. For example, it is impossible to tell the difference between a uniform constant gravitational field and a constant acceleration. Imagine standing in a featureless closed room, or box, with no windows; you feel a force on your feet from the floor pushing against you as the gravitational field of the Earth tries to accelerate you downwards, but the solid floor stops the acceleration with an equal and opposite force. Now imagine the box is in empty space far away from any source of gravity, but it is in a rocket with a thruster that is accelerating it at precisely 9.81 $\mathrm{m\,s^{-2}}$. You cannot tell the difference between the effects of the acceleration due to the rocket's thrusters and the Earth's gravitational field – they are indistinguishable because gravitational mass equals inertial mass (see the two pictures on the right-hand side of Figure 1.1).

Now suppose the box is in a lift shaft and is suspended by a metal cable. You feel the force of the floor on your feet. If the cable snaps, however, you will suddenly be in free fall and the force on your feet

[4] www.youtube.com/watch?v=KDp1tiUsZw8

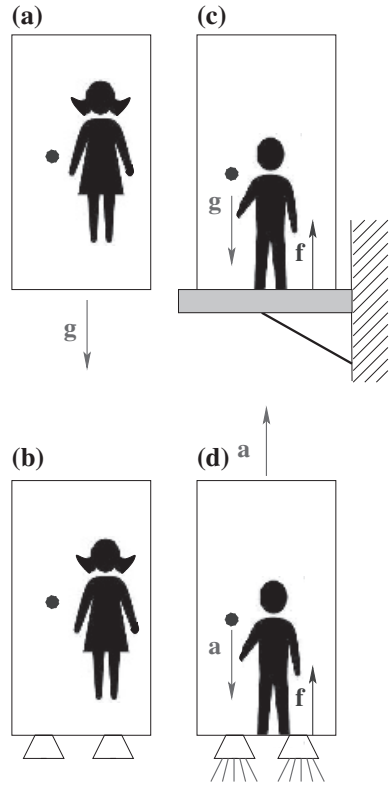


Figure 1.1 **The equivalence of inertial and gravitational mass.** The situation in a freely falling lift (a) is indistinguishable from that in a rocket in empty space far removed from any gravitational field (b); objects just float and experience no forces. An observer in a static box in a gravitational field feels a force on his feet due to his weight $\mathbf{f} = -m\mathbf{g}$ (c) which is indistinguishable from the force due to acceleration ($\mathbf{f} = -m\mathbf{a}$) of an accelerating rocket in empty space (d).

will disappear – as long as the box continues to accelerate downwards, without hitting anything, you will float freely inside the box as if there were no gravitational field at all, just like an astronaut in empty space in a rocket that is not accelerating. (See the two pictures in parts (a) and (b) of Figure 1.1.)

This is only true for uniform gravitational fields. In fact, the Earth's gravitational field is not uniform; it converges, and changes in magnitude, as one moves towards the centre of the Earth. If our experimenter

https://doi.org/10.1017/9781009263689.002 Published online by Cambridge University Press

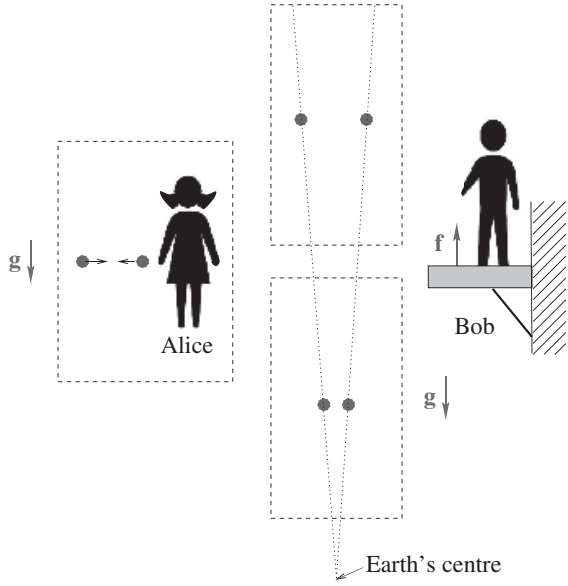


Figure 1.2 **Tidal forces.** Bob, standing and watching a falling lift accelerate past him, feels a force on his feet opposite in direction to the acceleration due to gravity $\mathbf{f} = -m\mathbf{g}$. A non-uniform gravitational field can be detected by someone floating in free fall (Alice). The gravitational field generated by a spherical mass such as the Earth or the Moon is not uniform over large distances; it converges towards a point at the centre of the mass, and this gives rise to tidal forces. A uniform gravitational field would not generate any tidal forces.

who is freely falling in the lift shaft (Alice) takes two objects (such as a hammer and a feather) and releases them together at rest, they will just float in front of her. But if she waits long enough (about 20 minutes for free fall in a tunnel drilled through the Earth), she will see them start to drift towards each other as the box gets nearer to the centre of the Earth (see Figure 1.2). She will then know that there is a non-uniform gravitational field present – this is an example of a tidal force. However, as long as we restrict our considerations to regions of space over which a gravitational field does not vary appreciably in either magnitude or direction, we cannot tell the difference between a gravitational field and an acceleration – this is known as the *Equivalence Principle*.[5]

[5] Most textbooks distinguish between different kinds of Equivalence Principle, the *Weak* Equivalence Principle and the *Strong* Equivalence Principle, depending on how widely it is applied to physical phenomena. We shall not make that distinction here.

https://doi.org/10.1017/9781009263689.002 Published online by Cambridge University Press

1.2 Equivalence Principle

Because inertial mass equals gravitational mass, all massive objects follow the same trajectory in a gravitational field, if they start from the same place with the same velocity. This observation led Einstein to suggest that the trajectory of a falling body is not determined by any properties of the body itself, such as its internal construction or constituent parts, but by the properties of the space (more correctly space-time) in which the body moves – and the trajectories are not straight because space-(time) is curved. In fact, freely falling bodies follow trajectories between two points in space-time that extremise the time it takes to go from one event to the other, as measured by a clock carried by the body (called the body's *proper time*).[6]

A simple example from 3-dimensional geometry illustrates Einstein's thinking. Consider an airplane flying from Delhi to Vancouver. To minimise fuel costs, the pilot wants to follow a route which takes the shortest path between the two cities – this takes her close to the North Pole. Now consider a second pilot flying from Mumbai to San Francisco; to minimise his fuel costs he takes a trajectory that takes him south, as in Figure 1.3. We see from the figure that the planes initially diverge, then their trajectories become parallel, and then they start converging. Each of the pilots is taking the shortest path to their destination, and to each pilot it looks as though the other plane is initially moving away but is being pulled inexorably towards them by some 'force' which depends not on the properties of the planes, but on the curvature of the Earth.

On a curved surface, the angles of a triangle do not necessarily add up to 180°. For example, on the surface of the Earth the angles of the triangle with one vertex at the north pole, one in Quito (the capital of Ecuador, on the equator 80°W of the Greenwich meridian), and the third in Libreville (the capital of Gabon, on the equator 10°E of the Greenwich meridian), the angles add to $90° + 90° + 90° = 270°$.

The German mathematician Gauss, one of the fathers of non-Euclidean geometry, was involved in a land survey in Germany and made many trigonometric measurements between 1818 and 1832. These included measuring the angles of a large triangle with vertices at the tops of three prominent hills near Göttingen in northern Germany. Within

[6] We shall see that, for massive bodies, the proper time is *maximised*. There is a principle in optics, *Fermat's principle of least time*, which states that the path taken by a beam of light in a refractive medium is that which *minimises* the time taken for the light to travel between two fixed points within the medium, but this is not the proper time of the light beam.

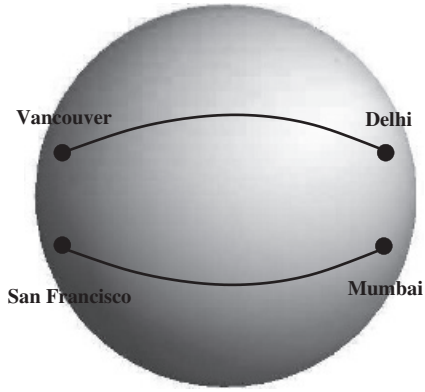


Figure 1.3 **Great circles.** Two airplanes taking paths of shortest length on the surface of the Earth. Technically the two paths are called *great circles*; they are circles whose centre is the centre of the Earth and whose radius is the radius of the Earth.

the accuracy of the measurements, the angles added up to the Euclidean value of 180°, giving no evidence of any curvature on the length scales that were accessible to him.

A modern measurement was made by the Planck satellite in 2015. This set limits on any spatial curvature of our Universe over cosmological distances: if it is non-zero, it must involve length scales at least of order 10^{28}m and possibly greater – some 100 times greater than the size of the observable Universe. The Planck results are compatible with zero spatial curvature.

However, any efforts to understand gravity in terms of the curvature of 3-dimensional space were doomed to failure. The analogy with the airline pilots is misleading, because in a gravitational field bodies with different starting velocities have different trajectories; this is an indication that in gravity time must somehow enter the picture as well as space. Having formulated his special theory of relativity in 1905 (which did not include gravity), Einstein was in a unique position to make progress in developing this idea further by postulating that, not just 3-dimensional space but *4-dimensional space-time* is curved. Einstein's interpretation of gravity is that tidal forces are a manifestation of a curvature of 4-dimensional space-time and all particles starting from the same point at the same time with the same velocity will follow the same trajectory in a curved 4-dimensional space-time. This is Einstein's Equivalence Principle at work. Einstein commented some years after developing the

https://doi.org/10.1017/9781009263689.002 Published online by Cambridge University Press

general theory of relativity that, when this idea occurred to him, it was 'the happiest thought of my life'.[7]

Understanding the consequences of these ideas requires studying the geometry of curved 4-dimensional space-times. This is inevitably rather technical mathematically, but before embarking on that journey let us emphasise the philosophy here. In order to understand physics in a gravitational field, we take the known laws of physics in flat 4-dimensional space-time (called *Minkowski* space-time) and try to write them out for a curved space-time. The complications of general relativity lie mainly in learning how to write the laws of physics in a curved space-time.

Problems

1) Calculate the speed of a planet moving in a circular orbit of radius r around a star of mass M. (Ignore the mass of the planet relative to M.) Show that
$$v^2 = \frac{GM}{r}.$$
2) Evaluate the escape velocity from the surface of a planet of mass M and radius R. What is its value when $R = 2GM/c^2$, where c is the speed of light?
3) **The geometry of an ellipse:** Kepler's first law of planetary motion states that the planets move around the Sun in an ellipse with the Sun at one focus. An ellipse is a very precise shape, and we will need the mathematical formulation of that shape when we discuss the Schwarzschild space-time later.

   In Cartesian coordinates (x', y'), with O' as the origin, the equation of an ellipse is
$$\frac{x'^2}{a^2} + \frac{y'^2}{b^2} = 1.$$
   If $a > b$, a is called the semi-major axis (half the larger diameter) and b is the semi-minor axis (half the smaller diameter).

   An ellipse can be drawn on a piece of paper by tying a piece of string into a loop of length l, fixing two drawing pins into the paper a distance d apart, looping the string around the drawing pins (this requires $l > 2d$) and using the tip of a pencil to pull the string taut

[7] Einstein, A., *Fundamental ideas and methods of the theory of relativity, presented in their development: in Collected papers of Albert Einstein vol. 7: The Berlin years 1918–1921* (English translation), Princeton University Press (2002).

https://doi.org/10.1017/9781009263689.002 Published online by Cambridge University Press

and then move the pencil in an arc around the drawing pins, keeping the string taut.

In the following unnumbered drawing, O' is the symmetric centre of the ellipse, one drawing pin is at O, and the other is at $\widetilde{O}$, a distance $\Delta = \frac{d}{2}$ from O', to its left. The points O and $\widetilde{O}$ are the called the *foci* of the ellipse.

Show that:

a) $l = 2(\Delta + a)$.

b) $r + \tilde{r} = 2a$.

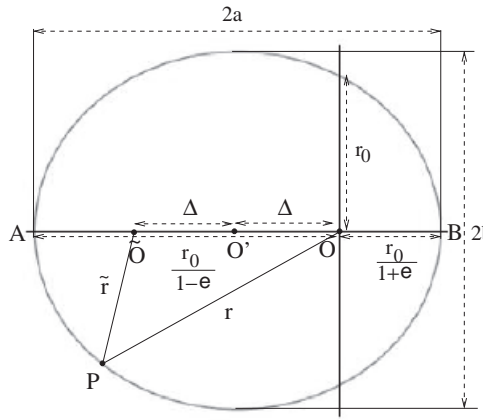
c) Define r_0 and e as shown in the figure: the distance from O to A is $\frac{r_0}{1-e}$ and the distance from O to B is $\frac{r_0}{1+e}$, with $0 \leq e < 1$. Show that:

i) $a = \frac{r_0}{1-e^2}$;

ii) the distance between the centre O' and a focus O is $\Delta = \frac{er_0}{1-e^2}$;

iii) $b = \frac{r_0}{\sqrt{1-e^2}}$;

iv) $\frac{2r_0}{1-e} = l$;



d) the equation of the ellipse in polar coordinates (r', θ') relative to O' is

$$r'^2 = \frac{r_0^2}{(1-e^2)(1-e^2\cos^2\theta')};$$

e) the equation of the ellipse in polar coordinates (r, θ) relative to the focus O is

$$r = \frac{r_0}{(1+e\cos\theta)}.$$

https://doi.org/10.1017/9781009263689.002 Published online by Cambridge University Press

f) the distance between the focus $\widetilde{O}$ and P is

$$\tilde{r} = \frac{r_0(1 + 2e\cos\theta + e^2)}{(1 - e^2)(1 + e\cos\theta)}.$$

g) Calculate the area enclosed by the ellipse.
Note: you may find the following integral useful:

$$\int_0^{2\pi} \frac{d\theta}{(1 + e\cos\theta)^2} = \frac{2\pi}{(1 - e^2)^{3/2}}.$$

4) Treating the Earth as a perfect sphere uniformly covered in water, calculate the height of the tides raised by the Moon (ignoring the Sun) and the height of the tides raised by the Sun (ignoring the Moon). Which is the larger effect?
(The Earth has mass 5.97×10^{24}kg and equatorial radius 6,370 km; the Moon has mass 7.35×10^{22}kg and is 384,000 km from the Earth on average; the Sun has mass 2.0×10^{30}kg and is 150 million km from the Earth on average.)

https://doi.org/10.1017/9781009263689.002 Published online by Cambridge University Press