Chapter 3. Kepler's Laws

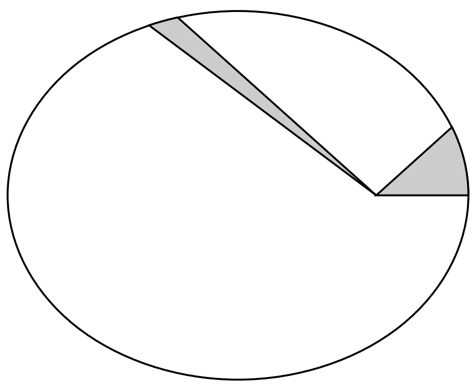
The curves defined by conic sections were apparently first analyzed in generality by the Greeks, particularly Apollonius. The theory only came into its own, however, when it was discovered by Kepler that within any error he could measure, the path of a planet orbiting around the Sun is an ellipse, and that there is a very simple rule for determining at what velocity a planet moved along its elliptical orbit. These discoveries were made by Kepler between 1605 and about 1615. They were far more revolutionary if more technical than the pronouncements of Copernicus, although still closer to the spirit of ancient rather than modern astronomy. About 50 years later, however, they led to Newton's discovery of the basic law of gravity underlying Kepler's laws, a major step in the development of modern science. In this chapter we shall introduce Kepler's Laws and examine some of their consequences. We shall postpone explaining the physics that underlies them, however, and attempt to look at them more in the geometrical spirit in which Kepler himself conceived them.

1. The statement of the Laws

Kepler's discoveries are usually formulated in three simple statements.

- *The planets move in ellipses, with the Sun at one focus.*
- *The motion of a single planet is such that equal areas are swept out in equal times.*
- *The period of a planet is proportional to the* $3/2$ *power of its semi-major axis, with a constant of proportionality valid for the entire Solar system.*

The following figure shows how the second law works, showing how a planet moves in a fixed interval of time at different locations in its orbit.



Towards remembering these laws, the first thing to notice is that the first law is concerned purely with the **geometry** of a planet's motion, while the other two are concerned with its **dynamics**.

Kepler was able to make his discoveries largely because he was able to use Tycho Brahe's observations of planetary motion, which were far more accurate than any which had been made before. Later on, I hope to give you some idea of the scale and difficulty of his achievement, and the exact obstacles which he met in carrying out his work. However, the observations he used were obsolete almost as he was using them, because around 1610 the telescope became a practical tool for observational astronomy, and in a very short time (practically overnight, so to speak) the accuracy of observations grew by an order of magnitude. It is perhaps just as well that he did not or could not avail himself of the new accuracy, because in fact his Laws are only approximations to the truth, and greater accuracy might have postponed for a very long time the essentially simple nature of gravity. The main cause of complication in the phenomena observed, as opposed to the basic laws bringing about these phenomena, is the effect of the other planets on a single planet's motion. This is at any moment hardly measurable, but time in astronomy is of such a scale that noticeable effects accumulate in a noticeable length of time, and most in a man's lifetime. Even in a Solar system of one planet and one sun, however, there is a problem. The true formulation of the first Law in this ideal environment is that he Sun and the planet together follow elliptical orbits around the centre of gravity of the pair of objects. Since the Sun is so large in mass compared to even the largest planet, this was not within the range of the accuracy of Kepler's data. We, too, shall ignore it, at least for now.

Kepler was only concerned with the planets of the Solar system which by definition are those astronomical bodies we see repeatedly. In light of Newton's law of gravitaion we can extend the first rule somewhat to include all cosmic objects:

- *The orbits of isolated objects in the Solar system are conic sections with one focus at the Sun.*

What distinguishes objects that move on ellipses from the others is that they do not have enough energy to escape the system.

The second and third laws can be reformulated in a more mathematical fashion. Suppose an interval of time Δt elapses, and a planet sweeps out in its orbit sweeps out a radial area of ΔA. In the next interval of amount Δt it will move through area ΔA also, according to Kepler's second law. In a total time elapsed of $n\Delta t$ it will move through area $n\Delta A$. This suggests:

- *For each object in the solar system the amount of radial area swept out is proportional to time elapsed. There exists a constant k for each planet such that*

$$\Delta A = k\Delta t \ .$$

The Earth's orbit is close enough to a circle that we can imagine without difficulty an ideal Earth that does move in a circle with the Sun at its centre. It is called the **mean** Earth. For this Earth the radius of the circle is one astronomical unit and the period is one Earth year. In these units, the third law asserts:

- *For a planet with semi-major axis a, its period is $a^{3/2}$.*

From a combination of the second and third laws we can get an even better version of the second law. Since we obtain an ellipse from scaling a circle by constants a and b,a nd since these scale changes scale all areas by the product ab, the area of an ellipse with semi-axes a and b is πab. The period of a planet with this ellipse as orbit is $a^{3/2}$. We therefore have

$$k = \frac{\pi ab}{a^{3/2}} = \frac{\pi b}{\sqrt{a}} \ .$$

For orbits which are not ellipses we can also derive a version of the third law, even though the period is infinite. The point in this version is to use the parameters p and e instead of a and b to characterize an orbit. We can calculate easily that $p = b^2/a$ and that $e = \sqrt{1-(b/a)^2}$. This gives us

$$b = \frac{b}{\sqrt{1-e^2}}, \qquad a = \frac{p}{1-e^2} \ .$$

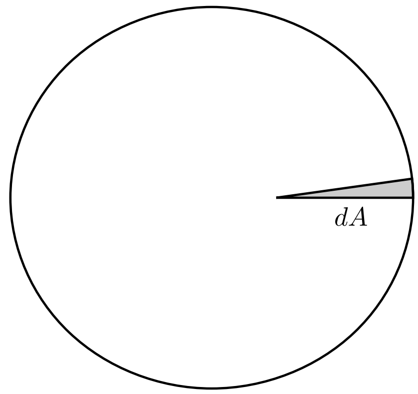
Therefore

$$k = \pi\sqrt{p}\,\sqrt{1-e^2}$$

which makes sense—and is true—for all orbits whether elliptical or not.

2. Calculating velocity

We can use Kepler's equation relatively easily to understand how the velocity of a planet changes as it moves around the Sun. In this section we calculate speed at the four simplest points on the orbit.



Suppose we let a small interval of time dt elapse from perihelion. The area covered will be $dA = k\,dt$. But the distanace from the focus to perihelion is p/e, and the area covered is essentially a triangle, so the area is also

$$(1/2)(p/1+e)\,ds = (1/2)(p/1+e)v\,dt\ .$$

Therefore

$$k\,dt = (1/2)(p/1+e)v\,dt, \qquad v = \frac{2k(1+e)}{p}\ .$$

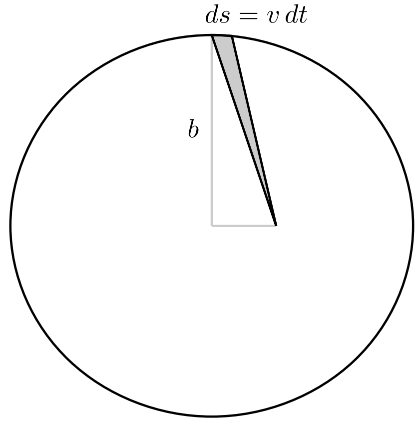
Similarly, at **apihelion**, where the planet ius farthest away from the sun, the velocity is

$$v = \frac{2k(1-e)}{p}$$

The velocity is higher at perihelion than at apihelion, and the ratio of the two speeds is

$$\frac{1+e}{1-e}\ .$$

Now we look at the point at the top of the orbit, where $y = b$. The area this time is $b\,ds/2$, from which we deduce



Here

$$v = \frac{2k\sqrt{1-e^2}}{p}$$

The general idea of these calculatoions is this: Suppose an interval of time dt passes, and the planet moves a distance ds, sweeping out area dA. Then

$$ds = v\,dt, \qquad dA = k\,dt, \qquad v = \frac{k}{dA/ds}\ .$$

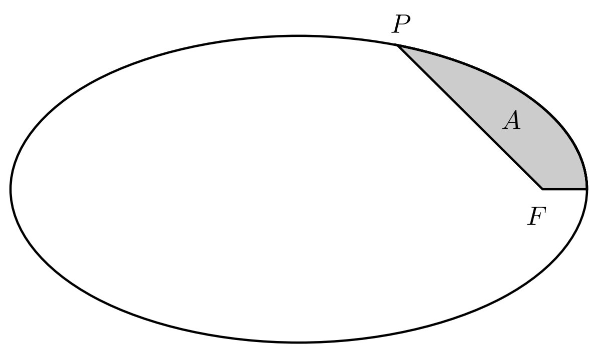
The denominator dA/ds ought to involve a purely geometric calculation. At points other than the ones we have just looked at, we need a tool to simplify the process.

3. Kepler's equation

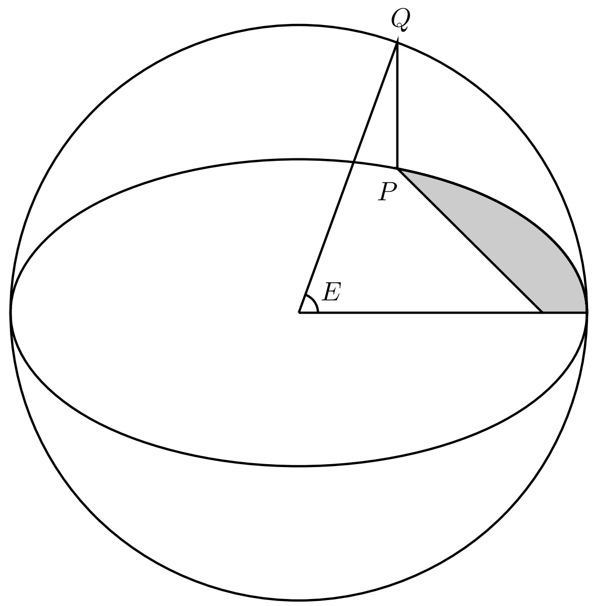
Consider an orbit with semi-major axis a and semi-minor axis b. Put the Sun at the focus $(f, 0)$ with $f > 0$. We know that $f = \sqrt{a^2 - b^2} = ae$. Suppose that at time $t = 0$ the planet is at **perihelion**, that is to say the point in its orbit nearest to the Sun. Suppose that at time t it is at location P on the ellipse. Let A be the area of the region traversed by the radius vector of the planet from the Sun. We know that

$$A = \alpha t$$

for some constant α (which will in general depend on the masses of the Sun and the planet as well as the geometry of the orbit, but is independent of t).

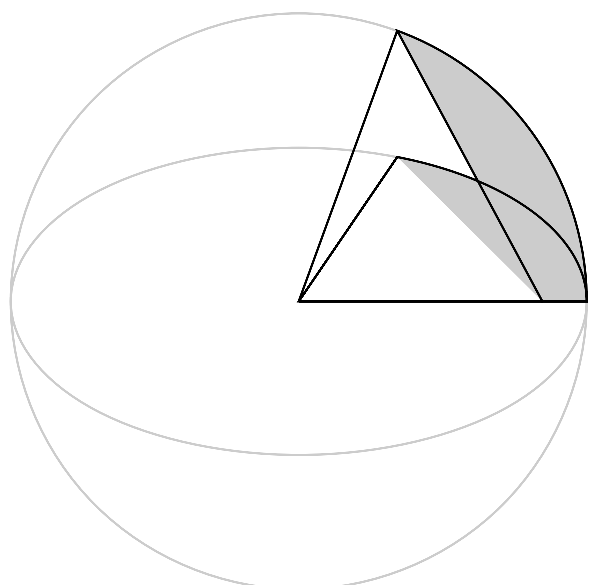


We shall use a coordinate system with origin at the centre of the ellipse, and we shall describe points on the orbit in terms of points in the circle one can inscribe around the ellipse with radius a. This circle is sometimes called the **Kepler circle** of the ellipse. The ellipse is obtained from this circle by compressing along the y axis by a factor b/a. Thus if $Q = (a\cos E, a\sin E)$ is a point on the circle, it corresponds to the point $P = (a\cos E, b\sin E)$ on the ellipse.



We want to relate the area A to the point P. We can easily relate the analogous things on the circle, however, and we obtain the ones on the ellipse by scaling vertically by the factor b/a. The area A is the difference between the area of the elliptic sector with angle E, and the triangle OFP, which has height $b \sin E$ and base $f = \sqrt{a^2 - b^2}$. Therefore

$$A = \frac{b}{a}\pi a^2 \frac{E}{2\pi} - \frac{b}{a}\frac{fa \sin E}{2} = \frac{ab}{2}E - \frac{b\sqrt{a^2-b^2}}{2}\sin E$$



which we can rewrite as

$$M = E - e\sin E$$

where

$$M = \frac{A}{ab/2}, \qquad e = \frac{f}{a} = \frac{\sqrt{a^2 - b^2}}{a}$$

The equation

$$M = E - e\sin E$$

is called **Kepler's equation**. Remember that knowing E is equivalent to knowing P, since

$$P = (a\cos E, b\sin E)$$

while knowing M is equivalent to knowing t, since M is proportional to A and A to t. Thus Kepler's equation asserts directly that *if we know the position P we can tell what t is.* This is usually opposite to what we usually want to know, which is how to determine P in terms of t. *In order to do this we must solve Kepler's equation for E in terms of M.* This is not so simple. In fact, there is no simple formula for E in terms of M (or P in terms of t), and we must solve the equation numerically.

4. The orbital velocity diagram

We shall now look again at the problem of calculating velocity at different parts of the orbit.

Orbital position at time t is

$$P(t) = (a\cos E, b\sin E)$$

where $E = E(t)$ is the unique solution of

$$\frac{2\pi t}{T_{\text{period}}} = E - e\sin E\ .$$

By the chain rule, its velocity at time t is

$$V(t) = P'(t) = E'(t)(-a\sin E, b\cos E)$$

We can determine $E'(t)$ in terms of E since by differentiating Kepler's equation we get

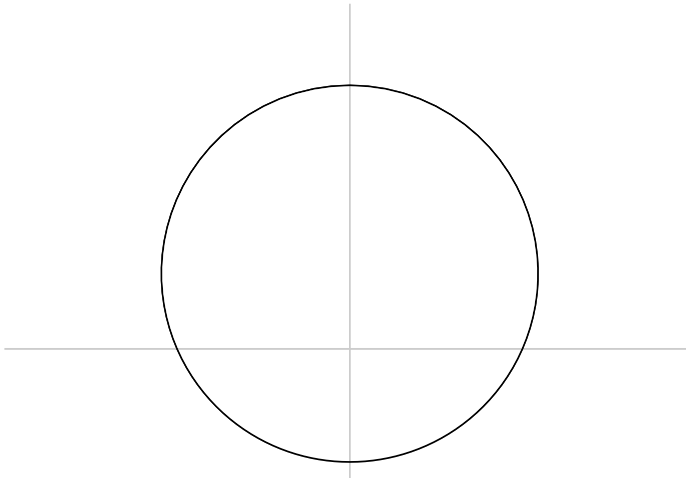
$$\frac{2\pi}{T_{\text{period}}} = E' - eE'\cos E, \quad E' = \frac{2\pi}{T_{\text{period}}(1 - e\cos E)}\ .$$

Therefore

$$V(t) = \frac{2\pi}{T_{\text{period}}}\left(\frac{-a\sin E}{1 - e\cos E}, \frac{b\cos E}{1 - e\cos E}\right)\ .$$

The **velocity diagram** of an orbit is the curve we get by plotting all possible values of this as t varies from 0 to T_{period}, or equivalently as E varies from 0 to 2π.

Here is a typical plot (with $e = 0.4$):



The top of the diagram, where velocity is greatest, is at perihelion, and the bottom at apihelion. Of course this looks suspiciously like a circle.

The velocity diagram is always a circle. But of course its centre is not at the origin. It has to be at the average of the two extreme values, which is (up to a scalar)

$$C = \frac{1}{2}\left(\frac{1}{1-e} - \frac{1}{1+e}\right) = \frac{e}{1-e^2} \ .$$

So to prove this claim, we have to calculate

$$x^2 + y^2$$

where

$$\begin{aligned}(x, y) &= \frac{1}{1-e\cos E}\,\frac{1}{1-e^2}\big(-a\sin E(1-e^2), b(1-e^2) - be(1-e\cos E)\big) \\ &= \frac{1}{1-e\cos E}\,\frac{1}{1-e^2}\big(-a\sin E(1-e^2), b\cos E - be\big) \ .\end{aligned}$$

Recall that

$$b = a\sqrt{1-e^2} \ .$$

Therefore $x^2 + y^2$ is the factor

$$\left(\frac{1}{1-e\cos E}\right)^2 \left(\frac{1}{1-e^2}\right)^2$$

times

$$\begin{aligned}a^2(1-e^2)(1-e^2)\sin^2 E + b^2\cos^2 E + b^2e^2 - 2b^2e\cos E &= b^2 - b^2e^2\sin^2 E + b^2e^2 - 2b^2e\cos E \\ &= b^2 - 2b^2e\cos E + b^2e^2\cos^2 E \\ &= b^2(1-e\cos E)^2 \ .\end{aligned}$$

Therefore

$$x^2 + y^2 = \frac{b^2}{(1-e^2)^2} \ .$$

In other words

- *The centre of the velocity diagram is at*

$$\big(0, e/(1-e^2)\big)$$

and its radius is

$$\frac{1}{1-e^2} \ .$$

Exercise 4.1. *What does the velocity diagram look like for a parabolic orbit? An hyperbolic one?*

Exercise 4.2. *Newton's law of gravity asserts that the force on an orbiting body is inversely proportional to distance, and directed towards the central mass. In mathematical terms, this says that*

$$\frac{dV}{dt} = -K\frac{(\cos\theta, \sin\theta)}{r^2}$$

for some constant $K > 0$, where (r, θ) are polar coordinates with respect to a focus. Kepler's second law asserts that time is proportional to radial area traversed. Since radial area A satisfies $dA = r^2 d\theta/2$, this translates to the equation

$$\frac{d\theta}{dt} = \frac{h}{r^2}$$

for some other constant h. If we combine these we get

$$\frac{dV}{d\theta} = -\frac{K}{h}(\cos\theta, \sin\theta) \ .$$

Why does this imply that the velocity diagram is a circle?

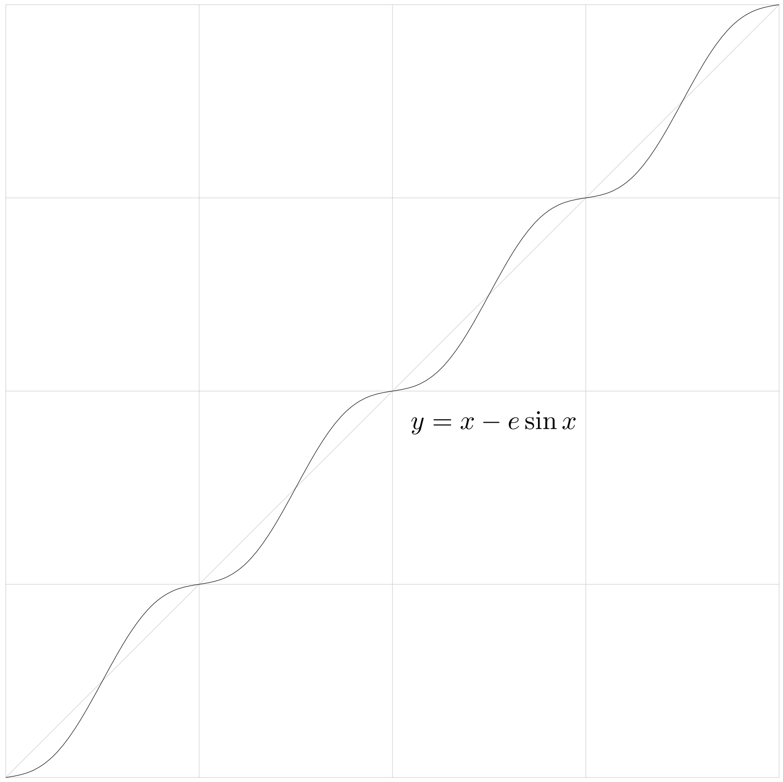
5. A simple way to solving Kepler's equation

Findling location in terms of time involves solving the equation

$$M = E - e\sin E$$

where e is the ellipticity, and M is the ratio $2\pi t/T_{\mathrm{period}}$. The angle M is called the **mean anomaly** because it is proportional to time, E the **eccentric anomaly** because its variation with time is complicated. From E we calculate position to be $(a\cos E, b\sin E)$.

The first thing to do to see what we up against is to get a rough idea of how things should go. This we can do by graphing M as a function of E for a typical value of e.



Here $e = 0.846$, rather arbitrarily chosen. In general, the smaller e is, the less the deviation of this graph from the diagonal line $M = E$. In any event, we can guess that the function taking E to $E - e \sin E$ is always **monotonic increasing**, which means that as E increases so does M. In fact, the proof is simple: we just have to verify that the slopeof the graph is always non-negative. But

$$\frac{dE - e \sin E}{dE} = 1 - e \cos E > 0 \ .$$

since we are assuming $e < 1$. The consequence of monotonicity is that for any given value of M there is a unique value of E for which $M = E - e \sin E$. This remains true as long as $0 \leq e < 1$. or in other words for all elliptical orbits. So the problem we are trying to solve is well defined.

Therefore we now consider directly the question of how to find E given M. If e is small, the following trick works well. We rewrite the equation

$$E = M + e \sin E$$

so that we are looking for a number E taken to itself by the transformation $E \mapsto M + e \sin E$. We start with some initial approximation E_0 for E. As long as e is small, $E_0 = M$ will be close enough, since $M + e \sin M$ will not be too far from M. Then we calculate in succession

$$E_{n+1} = M + e \sin E_n$$

If the successive values of E_n converge, they will certainly converge to the value of E we are trying to find.

Example. Set $e = 0.1$, say, and $M = 1$. We get values

$$\begin{aligned}
E_0 &= 1 \\
E_1 &= 1 + 0.1\ \sin 1 \\
&= 1.084147 \\
E_2 &= 1.088390 \\
E_3 &= 1.088588 \\
E_4 &= 1.088597 \\
E_5 &= 1.088598 \\
E_6 &= 1.088598 \\
&\dots
\end{aligned}$$

so it does in fact converge quickly in this case. In order to calculate planet positions as a function of time it is clearly important to be able to solve Kepler's equation efficiently. Of course in practice this involves a computer.