Geometry and Planetary Motion

Introduction

This short article provides a geometrical description of the mechanics of planetary motion. Geometry is the tool Newton employed in his Principia Mathematica and though his methods are now outdated there is an elegance and directness to them that some readers may find more satisfying than high flying mathematics. The article is based on a presentation given by James Clerk Maxwell in his book 'Matter and Motion', published in 1877 and also on a lecture given by Richard Feynman in 1964 that appeared in the book 'Feynman's Lost Lecture' by D.L. Goodstein. The approach is more accessible than Newton's original, which requires familiarity with the geometry of conic sections as described by Apollonius of Perga, no less. Nevertheless a basic knowledge of Euclidian geometry, at least to high school level, is required. Hopefully everyone reading this will have enough in their background to follow and some of the more obscure points will be explained.

Part 1: The sum of two vectors.

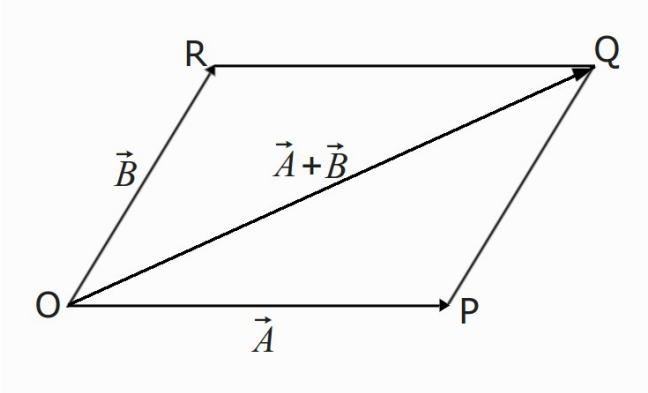

Figure 1.

In Figure 1 $\vec{A}$ and $\vec{B}$ are two vectors. In our case these are velocity vectors, which have both a speed and direction. They are drawn as arrows in the figure. The speed is indicated by the *length* of the arrow and the direction is indicated by the orientation of the arrow on the page. Normally an object is expected to have just one velocity, which shows its direction and speed with respect to (say) the Earth. However, sometimes a second velocity comes into play. For example an aeroplane flying at a certain speed and direction with respect to the air, while the air itself is be moving in a different direction with respect to the ground (due to wind). The question then is what is the velocity of the aeroplane measured from the ground?

Suppose $\vec{A}$ is the velocity of the plane in the air and $\vec{B}$ is the velocity of the wind across the ground. Each of these velocities is physically independent of the other, so they can be considered independently. Let us first look at vector $\vec{A}.$ Taking the velocity $\vec{A}$ in isolation in Figure 1, it would carry the plane from point O a distance A in *one unit of time* to arrive at the location $P.$ Then the wind velocity $\vec{B},$ applied at the point $P,$ would after one unit of time carry the plane a distance B to point $Q.$ Alternatively, consider the alternative route, which allows the wind first to carry the plane from point O a distance B to the point R in one unit of time, and then let allows the plane's velocity carry it from there a distance A in one unit of time to arrive again at $Q.$ These two paths must give the same result since the two

vectors are supposed to be independent of each other. The effect overall is that after one unit of time vector $\vec{A}$ produces a shift of the plane through a distance A and vector $\vec{B}$ a shift of the distance $B.$ The different paths inevitably construct the parallelogram evident in Figure 1. The order of application of each vector is irrelevant, so both can be applied simultaneously. The result is that the plane will then arrive at Q in *one* unit of time, since the movements are concurrent. The distance the plane has travelled is given by the diagonal of the parallelogram and it velocity will be the length of the diagonal divided by one unit of time. The net effect (i.e. the sum) of the two vectors is the diagonal vector indicated.

Part 2: Areal velocity.

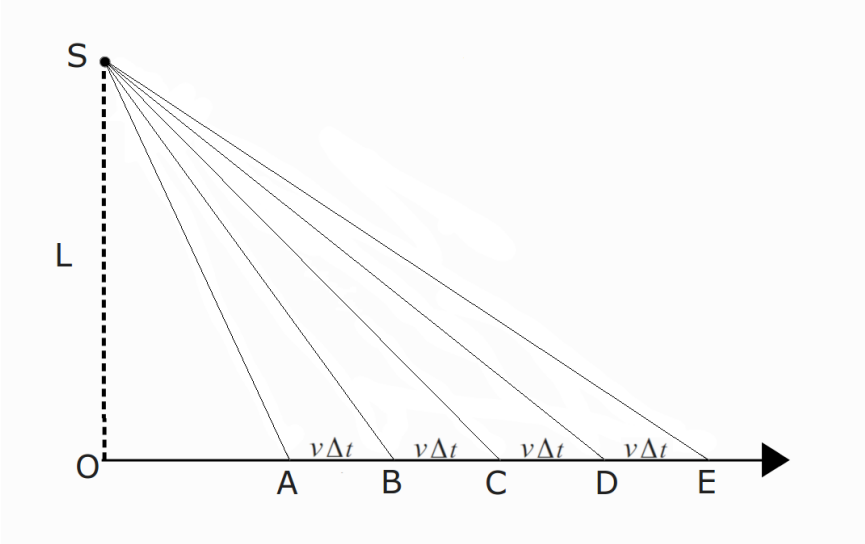

Figure 2.

In Figure 2 it is supposed that a particle is travelling with a velocity $\vec{v}$ along a straight line $AE.$ After point $A,$ at time intervals of $\Delta t,$ the points B,C,D *etc.* are marked out and as a result are separated by equal distances of $v\Delta t,$ where v is the scalar speed of $\vec{v}.$

It is now supposed that there is a point S fixed at some distance from the line $AE,$ and drop the perpendicular from S to meet the extended line at $O.$ The length of this line is the distance $L.$ Next the point S is connected to each of the points $A,\ B,\ C,\ D,\ E$ as shown. It then turns out that: the areas of all the triangles $ASB,\ BSC,\ CSD,\ DSE$ are *equal*. This follows because the area of any triangle is given by half the product of its base length and its perpendicular height and in the case of all the triangles indicated, the base length is $v\Delta t$ and the perpendicular height is the distance $L.$

It is clear from this that, taken over successive time intervals of $\Delta t,$ that a particle moving with a constant velocity $\vec{v},$ past a point S at a perpendicular distance $L,$ has a constant triangular area. Using the symbol A_Δ to represent the area of the triangles, a variable a_v may be defined as

$$a_v=\frac{A_\Delta}{\Delta t}=\frac{1}{2\Delta T}Lv\Delta T=\frac{1}{2}Lv, \qquad (1)$$

which is the *areal velocity* of the particle.

Part 3: Areal velocity of a particle orbiting a fixed point

Figure 3 presents a similar construction to that shown in Figure 2, except that the particle is now orbiting around the fixed point S instead of merely travelling past in a straight line. S is therefore to be regarded as the *centre of force* deflecting the particle from a straight line.

The deflection of the particle's path into the orbit can be described in the following manner.

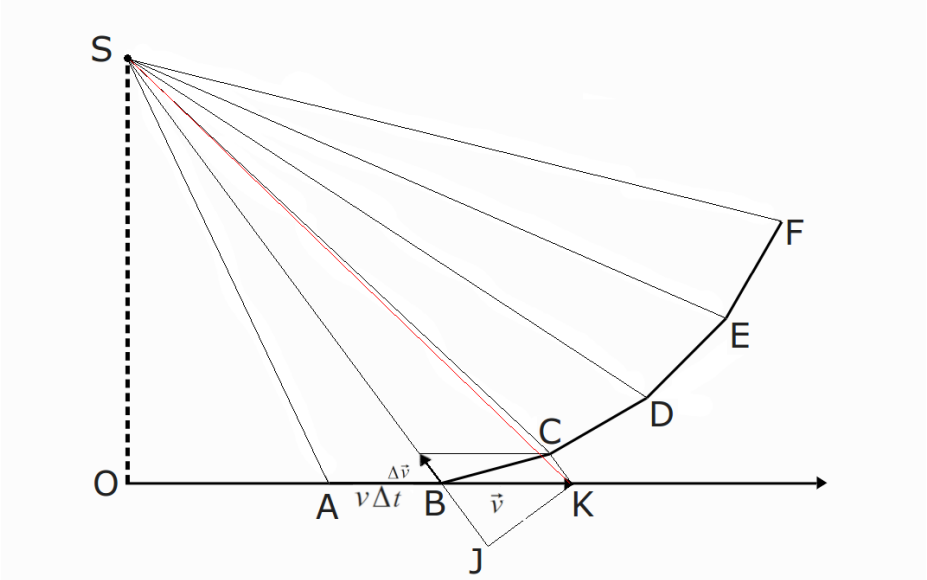

Figure 3.

Between points A and B the particle travels with a velocity $\vec{v}$ for a time interval Δt. The areal velocity of the triangle ASB is thus given by equation (1). On arriving at point B the particle is subjected to an *impulse force*, which is a force that is applied i*nstantaneously*, in the direction of the point S and giving it an additional velocity $\Delta\vec{v}$ in that direction. Meanwhile, since velocities $\vec{v}$ and $\Delta\vec{v}$ are independent of each other, the particle retains its original velocity $\vec{v}$ in the direction of the extended AB line. The velocities $\vec{v}$ and $\Delta\vec{v}$ combine as a vector sum $\vec{v}+\Delta\vec{v}$, (see Part 1), which in the time Δt carries the particle to the point C. It can now be asked what is the areal velocity of the particle for the triangle BSC? To obtain this it is necessary to construct some additional lines:

a) From C construct a line parallel to line SB and $\Delta\vec{v}$ to meet the extended line AB at K. As a consequence of Part 1, this line has length $\Delta t\Delta\vec{v}$.
b) Extend line SB downwards and drop a line perpendicular to this from K to intercept at J.
c) Draw the line SK.

Now, as a consequence of Part 1, the distance BK equals $v\Delta t$, which is the same as the distance AB. The triangles ASB and BSK must therefore be equal in area, because they have the same perpendicular height OS and the same base length. Also triangles BSK and BSC have the same area, because they have a common base BS and the same perpendicular height, which is the distance JK. (This follows because lines CK and BJ are parallel by construction and JK is therefore perpendicular to both lines.) It follows that the areas of triangles ASB and BSC are equal. Just as they are in Figure 2.

An impulse force could now be applied to the particle at C in the direction of S and the same argument used to show that, after the interval $\Delta t,$ the triangles BSC and CSD are also equal in area. So it goes for all subsequent triangles obtained in successive intervals of $\Delta t.$ Therefore as the particle orbits the point S in discrete steps of $\Delta t,$ the areas of all the triangles are equal. Furthermore, in consequence the total area of all the triangles is necessarily proportional to the time summed over all the discrete steps of size $\Delta t.$

Newton argued that this result would be true even if Δt was vanishingly small and the number of included triangles in any given time interval was proportionally increased. In such a circumstance the discrete orbit will become as close to a continuous curve to *any accuracy required*. It follows that by this approach, in the limit as $\Delta t \to 0,$ the true orbit of a particle about a point P can be obtained. It also follows that for any time intervals $[t_1, t_2]$ where $t_1 < t_2,$ the area swept out by a line connecting the orbiting particle to the centre of force S is proportional to the duration of the time interval i.e.

$$Area_{1,2} = k(t_2 - t_1), \tag{2}$$

where k is a constant. This is proof of Kepler's second law.

An important consequence of this proof is the following. Draw a tangent to the orbit at any given point P and drop a perpendicular line from the centre of force S to the tangent at $Q,$ then the velocity $\vec{v}$ of the orbiting particle at P lies along the tangent and, taking the distance SQ to be the variable $d_{perp},$ the following quantity is conserved for any point P:

$$a_v = \frac{1}{2} d_{perp} v, \tag{3}$$

where a_v is the *areal velocity* defined in equation (1).

The areal velocity can be calculated at any point in the orbit, but it is easiest to calculate when the velocity vector is naturally perpendicular to the position vector of the planet. This is the case when the planet it at perihelion – the closest point to the Sun. At this location it can be calculated from

$$a_v = \frac{1}{2} r_p v_p, \tag{4}$$

where r_p and v_p are the distance and velocity respectively of the planet at perihelion.

Part 4: The tangent to an ellipse.

The construction of a tangent to an ellipse at a point P is shown in Figure 4, but firstly some basic properties of an ellipse need to be explained.

An ellipse has its major axis aligned along its widest diameter and its minor axis aligned along its narrowest diameter. These are mutually perpendicular and cross at the centre of the ellipse. The major axis has length $2a$ and the minor axis has length $2b.$ Lengths a and b are related to the eccentricity e of the ellipse via the formula

$$b^2 = a^2(1 - e^2). \tag{5}$$

The ellipse has two focii S_1 and S_2 on the major axis, each at a distance ae from the centre and thus a distance $2ae$ apart (see Figure 4). Any point P on the ellipse is at a distance r_1 from S_1 and a distance r_2 from S_2 and for all points P

$$r_1+r_2=2a. \tag{6}$$

Relation (6) underpins the common means of constructing an ellipse using a fixed length of string and two fixed points. Note that when r_1 equals r_2, then from (6) $r_1=r_2=a,$ and since P is then on the minor axis it follows that

$$a^2=b^2+(ae)^2, \tag{7}$$

which rearranges to give the relation (5).

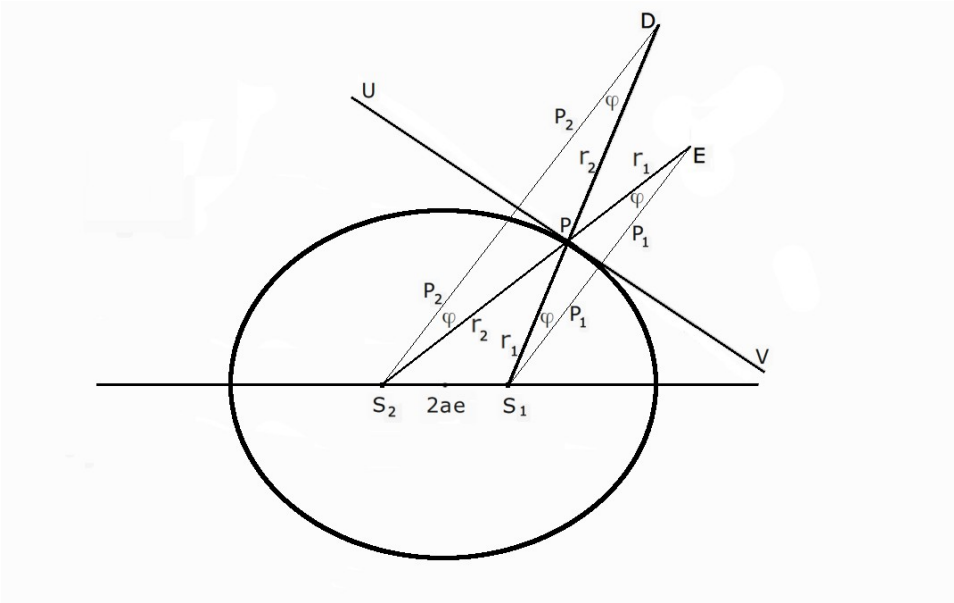

Figure 4

To construct a tangent at P in Figure 4, first extend the line S_1P a distance r_2 to the point $D.$ Now draw a line from S_2 to $D.$ Drop a perpendicular line from P to the line S_2D and extend in both directions to make the line UV in the figure. This line is the tangent at P. That this is the case is easily shown:

Line S_1D is of length r_1+r_2 by construction, and therefore by (6) equals $2a.$ For any point P' along the line $UV,$ except point $P,$ the sum of the distances $r_1'\equiv S_1P'$ and $r_2'\equiv P'D$ must be greater than $2a,$ since these two distances are not co-linear with $S_1D.$ Since the sum of the distances r_1' and r_2' violates the ellipse property (6), the point $P'\neq P$ cannot lie on the ellipse. Point P is therefore the only point on both the line UV and the ellipse and thus UV is the tangent.

Note that alternatively the tangent could have been constructed by extending the line S_2P a distance r_1 to point E and then dropping a perpendicular line from P to the line $S_1E.$ This allows us to prove a useful property of the tangent, which concerns the perpendicular distances to the tangent from the focii S_1 and $S_2.$ In Figure 4 the perpendiculars lie along the lines S_1E and S_2D respectively. Labelling the distance from S_1 to the tangent as p_1 and from S_2 to the tangent as $p_2,$ it can be shown that

$$b^2 = p_1 p_2. \tag{8}$$

Proof of relation (8) requires us first to recognise that triangles S_1PE and S_2PD are equivalent isosceles triangles. The equal angles are labelled φ in Figure 4. The equivalence of the triangles allows us to write

$$\cos\varphi = \frac{p_1}{r_1} = \frac{p_2}{r_2} = \frac{p_2}{(2a - r_1)}, \tag{9}$$

where, in the final ratio, relation (6) has been used to replace $r_2.$

Now consider the triangle $S_1ES_2.$ The cosine rule gives

$$(S_1S_2)^2 = (S_1E)^2 + (S_2E)^2 - 2(S_1E)(S_2E)\cos\varphi, \tag{10}$$

which in terms of variables defined above can be written as

$$(2ae)^2 = (2p_1)^2 + (2a)^2 - 2(2p_1)(2a)\cos\varphi. \tag{11}$$

Equation (11) reduces to

$$p_1^2 - 2ap_1\cos\varphi + a^2(1 - e^2) = 0 \tag{12}$$

From (5) the last term on the right of (12) must be $b^2,$ and from (9) replacing $\cos\varphi$ by the ratio p_1/r_1 means (12) can be rearranged to

$$b^2 = \frac{p_1^2}{r_1}(2a - r_1). \tag{13}$$

Finally, according to (9), p_1/r_1 can be replaced by $p_2/(2a - r_1)$ and so (13) reduces to (8).

Part 5: Proving the inverse square law

Kepler's first law is that a planet orbits the Sun in an ellipse, with the Sun at one focus of the ellipse. There follows Maxwell's proof that this is consistent with the inverse square law of gravity.

Figure 5 shows an elliptical orbit, where S_1, S_2 are the ellipse focii, with the Sun at $S_1.$ P is the current position of the planet. Line UV is the tangent at $P,$ r_1 and r_2 are the distances from the focii to the position P and p_1, p_2 are the perpendiculars dropped from the focii to the tangent. The line S_1P is extended by the distance r_2 to the point $D,$ making the distance S_1D equal to $2a,$ which is the length of the major axis of the ellipse. Finally a line is drawn from S_2 to $D,$ and it is known from Part 4, that this line is bisected at 90^o by the tangent $UV.$

Now, the velocity of the planet at P is $\vec{v}$ and its areal velocity a_v according to equation (1) is

$$a_v = \frac{1}{2}p_1 v, \tag{14}$$

and a_v is a constant for the orbit. Replacing p_1 using equation (8) allows (14) to be written as

$$v = \frac{2a_v}{b^2} p_2 . \qquad (15)$$

Since a_v and b are constants, it follows that the distance p_2, (which is half the line $S_2 D$ or the length $S_2 U$), is directly proportional to the velocity of the planet at P.

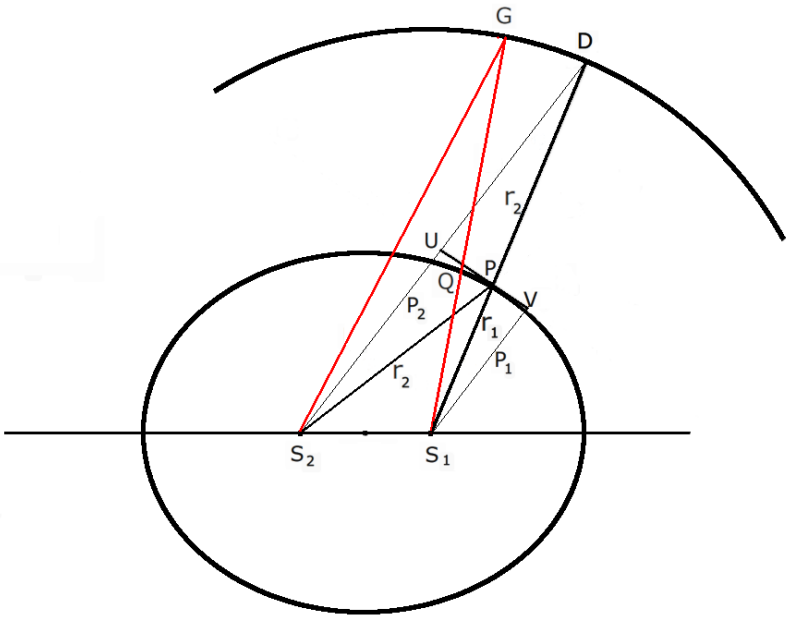

Figure 5.

This relationship means that for any point P on the ellipse, the geometric construction of the tangent also reveals the velocity. For any point P on the ellipse a circle of radius $2a$ centred on S_1 is drawn and the line $S_1 P$ extended to meet the circle at D. The line $S_2 D$ has length $2p_2$ which according to (15) gives the velocity. The direction of the velocity at P is given by the tangent, which is the perpendicular bisector of line $S_2 D$.

The circle drawn around S_1 serves for all points on the ellipse. Once it is drawn, *any* line drawn from S_2 to the circle is a measure of the velocity at some point P on the ellipse. The precise point concerned is where the perpendicular bisector of the line $S_2 P$ (i.e. the tangent) touches the ellipse. This geometric description of the velocity is called a *hodograph*. Note that while the planet's position vector revolves around the focus S_1, the velocity vector (represented by the line $S_2 D$) revolves around the focus S_2, which shows that the physically empty focus of the ellipse nevertheless has a role to play in the dynamics of the system. Note however, that the direction of line $S_2 D$ is 90^o *behind* the true velocity vector, since it is perpendicular to the tangent by construction.

Consider now another position Q on the orbit (see Figure 5) close to P. (The distance PQ represents a change in position over a very small time interval.) Extension of the line $S_1 Q$ meets the outer circle at G. While the distance PQ thus represents the change in the planet's *position*, the distance DG represents the

change in the planet's *velocity* (i.e. the *acceleration*) in the same interval of time. Being on the circumference of the outer circle, the line DG (and therefore the acceleration) is perpendicular to the circle's radius, but the hodograph is 90^o behind the true direction of the velocity, so the acceleration is actually in the direction of the Sun, as expected.

Furthermore, as the planet moves from P to $Q,$ the change in the angle of the Sun-planet vector is the same as the change in the angle between lines S_1D and $S_1G,$ and occurs in the same time interval. The angular change is therefore proportional to both the angular velocity of the planet and its acceleration, so the acceleration and angular velocities are proportional to each other.

Now, when the interval of time between positions P and Q becomes vanishingly small, the two points approach merger. In this limiting condition the areal velocity (1) can be written as

$$a_v=\frac{1}{2}r_1(\omega r_1) \quad \text{i.e.} \quad a_v=\frac{1}{2}\omega r_1^2, \tag{16}$$

where r_1 is the Sun-planet distance, ω is the angular velocity and ωr_1 is the velocity perpendicular to $r_1.$ Rearranging (16) gives

$$\omega=\frac{2a_v}{r_1^2}, \tag{17}$$

which shows that angular velocity is inversely proportional to $r_1^2.$ It has already been established that angular velocity and acceleration are proportional, so it is evident that the acceleration of the planet, and therefore the force acting upon it, is inversely proportional to the square of the Sun-planet distance. This proof of the inverse square law is surprising in that it merely requires that the orbit be an ellipse with the Sun at one focus and that the gravitational force to originate from the Sun for everything to fall into place.

Note however that this does not prove the converse: that an inverse square law implies an ellipse. Presumably Newton was concerned to establish the inverse square law of gravity as a consequence of observed facts rather than speculation, otherwise the law would appear to be a fortunate guess and perhaps not unique. Newton went on to establish that parabolic and hyperbolic orbits were also compatible with the inverse square law, which offered a means to explain the orbits of comets. Being Newton, and therefore thorough, he also proved that these were the only orbits compatible with the inverse square law.