Basic Orbital Mechanics

This essay provides an introduction to orbital mechanics, concerning the motion of planets and comets around the Sun. It is couched in the language of vectors and readers unfamliar with vector concepts are advised to read the appendix first.

Kepler's Description of Planetary Orbits

It is well known that the geometry of planetary orbits was first correctly formulated by Johannes Kepler, who succeeded in finding a geometric fit to the orbit of Mars using the data obtained by Tycho Brahe. Though his was an empirical description lacking a physical explanation, it was nevertheless a great leap forward in understanding the dynamics of orbits. His discovery is encapsulated in his three laws:

1. Each planet revolves around the Sun in an ellipse, with the Sun fixed at one focus of the ellipse.

2. The line joining the Sun's centre to the centre of the planet (the *radius vector*) sweeps out equal areas in equal times.

3. The time T to complete an orbit (also know as the period) is proportional to the planet's average distance from the Sun raised to the power of $3/2$ i.e.

   $$T = k\, a^{3/2} \tag{1}$$

   where a is the average distance and k is a constant applicable to all planets in the solar system.

Though lacking a physical foundation, these laws enabled the calculation of the positions of planets with an accuracy that was unprecedented at the time. They enabled, for instance, Jeremiah Horrocks to predict a transit of Venus in 1639 and thus become the first person ever to observe one.

The great scientific challenge that followed from Kepler's laws was to uncover their physical causes. This was finally settled by Isaac Newton and his theory of Universal Gravitation. The present article explains planetary motion in a manner consistent with Newton's theory, but cast in a modern vector form.

The Formal Solution of Orbital Motion

Newton's law of gravitation written in vector form is

$$m\ddot{\vec{r}} = -\frac{GMm}{r^3}\vec{r}, \tag{2}$$

in which G is the universal gravitational constant and M the mass of the Sun, m the mass of the planet. The vector $\vec{r}$ defines the position of the centre of the planet with respect to the centre of the Sun and r is the magnitude of $\vec{r}$ corresponding to the distance between the Sun and the planet. The vector $\ddot{\vec{r}}$ represents the second derivative of $\vec{r}$ with respect to time, or equivalently, the *acceleration* of the vector $\vec{r}$:

$$\ddot{\vec{r}} = \frac{d^2\vec{r}}{dt^2}. \tag{3}$$

It is customary to define the constant $\lambda,$ which is a parameter applicable to the Solar System:

$$\lambda = GM, \tag{4}$$

and so (2) can be written as

$$\ddot{\vec{r}} = -\frac{\lambda}{r^3}\vec{r}. \tag{5}$$

We may note that the mass of the planet m has vanished from the equation by cancellation from opposite sides.

Taking the vector cross product (see appendix) of $\vec{r}$ with both sides of equation (5) it can be seen that

$$\vec{r}\times\ddot{\vec{r}} = -\frac{\lambda}{r^3}(\vec{r}\times\vec{r}) = 0, \tag{6}$$

where we have used the fact that $\vec{a}\times\vec{a}=0$ for any vector $\vec{a}.$ It can also be shown that

$$\frac{d}{dt}(\vec{r}\times\dot{\vec{r}}) = \vec{r}\times\ddot{\vec{r}} + \dot{\vec{r}}\times\dot{\vec{r}} = \vec{r}\times\ddot{\vec{r}}. \tag{7}$$

Substituting this relation into (6) gives leads to

$$\frac{d}{dt}(\vec{r}\times\dot{\vec{r}})=0\,, \tag{8}$$

and integrating this gives

$$\vec{r}\times\dot{\vec{r}}=\vec{h}\,, \tag{9}$$

where $\vec{h}$ is a constant vector with a fixed magnitude and direction. Since $\vec{h}$ is the result of a vector cross product of $\vec{r}$ and $\dot{\vec{r}}$, it is orthogonal (i.e. perpendicular) to both vectors. It follows that

$$\vec{r}\cdot\vec{h}=0, \tag{10}$$

(see appendix) which tells us at once that the planet orbits the Sun in a flat plane perpendicular to $\vec{h}$. This is known as the *orbital plane*. Note that if equation (9) is multiplied on both sides by the planet mass m, the right hand side then represents the planet's *angular momentum* around the Sun:

$$\vec{\mu}=m\,\vec{r}\times\dot{\vec{r}}=m\vec{h}\,, \tag{11}$$

from which it follows that the angular momentum of the planet is constant.

By combining equations (5) and (9) we see that

$$\vec{h}\times\ddot{\vec{r}}=-\frac{\lambda}{r^3}\left(\vec{r}\times\dot{\vec{r}}\right)\times\vec{r}\,. \tag{12}$$

The triple vector product on the right hand side can be written as

$$\vec{h}\times\ddot{\vec{r}}=\frac{\lambda}{r^3}\left((\vec{r}\cdot\dot{\vec{r}})\vec{r}-(\vec{r}\cdot\vec{r})\dot{\vec{r}}\right), \tag{13}$$

(see appendix) which becomes

$$\vec{h}\times\ddot{\vec{r}}=-\frac{\lambda}{r^3}\left(r^2\dot{\vec{r}}-r\dot{r}\vec{r}\right), \tag{14}$$

where $\dot{r}$ is the time derivative of the magnitude r and should not be confused with the magnitude of the vector $\dot{\vec{r}}$. Equation (14) can be rearranged to

$$\vec{h}\times\ddot{\vec{r}}=-\lambda\left(\frac{\dot{\vec{r}}}{r}-\frac{\dot{r}}{r^2}\vec{r}\right). \tag{15}$$

Now we consider the derivative:

$$\frac{d}{dt}\left(\frac{\vec{r}}{r}\right)=\frac{\dot{\vec{r}}}{r}-\frac{\dot{r}}{r^2}\vec{r}\,. \tag{16}$$

Substituting (16) into (15) allows us to rewrite (15) as

$$\vec{h}\times\frac{d}{dt}\dot{\vec{r}}=-\lambda\frac{d}{dt}\left(\frac{\vec{r}}{r}\right). \tag{17}$$

Equation (17) can be integrated to give

$$\vec{h}\times\dot{\vec{r}}=-\lambda\frac{\vec{r}}{r}-\vec{c}\,, \tag{18}$$

where $\vec{c}$ is a constant vector, which like $\vec{r}$ is in the plane of the planetary orbit. At this juncture the *direction* of $\vec{c}$ in the orbit plane is arbitrary, but our preference is for $\vec{c}$ to be chosen as the *perihelion* vector, the point in the orbit where the planet makes its closest approach to the Sun. This is a unique location and easily identified. Conveniently from a mechanical point of view, the planet's velocity and position vectors are mutually orthogonal at this location.

Now we take the scalar product of the vector $\vec{r}$ with both sides of equation (18) to obtain

$$\vec{r}\cdot\vec{h}\times\dot{\vec{r}}=-\lambda\frac{\vec{r}\cdot\vec{r}}{r}-\vec{r}\cdot\vec{c}\,. \tag{19}$$

The left hand side of (19) is a triple scalar product (see appendix), which has the property that

$$\vec{r}\cdot\vec{h}\times\dot{\vec{r}}=\vec{h}\cdot\dot{\vec{r}}\times\vec{r}=-\vec{h}\cdot\vec{r}\times\dot{\vec{r}}\,. \tag{20}$$

where the reversal of the sign in the last expression is due to changing the order of vectors in the cross product. Exploiting (20) means equation (19) can be recast as

$$h^2=\lambda\,r+rc\cos\phi\,, \tag{21}$$

where we have also used the identity (9) and expanded the scalar product $\vec{r}\cdot\vec{c}$ in terms of the magnitudes r and c and the enclosed angle ϕ. Note that the angle φ is defined with respect to the fixed vector $\vec{c}$, which was defined above as the perihelion vector.

We can now write r as a function of φ by rearranging (21) into

$$r = \frac{(h^2/\lambda)}{(1+(c/\lambda)\cos\phi)}. \tag{22}$$

At this point we recognise (22) as the equation for a *conic section*[1] written in polar coordinates, which normally has the form

$$r = \frac{p}{1+e\cos\phi}, \tag{23}$$

where evidently

$$p = h^2/\lambda, \tag{24}$$

which is a constant known as the *latus rectum,* which corresponds to the variable r when $\phi = 90^o$ in (23). We also note that

$$e = c/\lambda \tag{25}$$

which is the *eccentricity* of the conic section. The eccentricity is particularly important in orbital mechanics since it defines the shape of the orbit in the following manner:

- $e=0$: the orbit is a circle,
- $e<1$: the orbit is an ellipse,
- $e=1$: the orbit is a parabola,
- $e>1$: the orbit is a hyperbola.

(26)

This solution to the orbit of a planetary body represents a proof of Kepler's first law, though it goes some way beyond Kepler by establishing that parabolas and hyperbolas as well as ellipses are acceptable orbits. This made it possible to explain the motion of other astronomical bodies, such as comets.

While equation (22) defines the planetary orbit, the time dependence of the planet's motion can be obtained from equation (9). We first define the planet's velocity $\vec{v}$ as

$$\vec{v} = \dot{\vec{r}}. \tag{27}$$

Then we can write the equation (9) in scalar form as

1 Which means the orbit *must* be either a circle, ellipse, parabola or hyperbola.

$$h = r\,v \sin\phi \tag{28}$$

where r, v and h are respectively the magnitudes of vectors $\vec{r}$, $\vec{v}$ and $\vec{h}$. Furthermore the term $v \sin\phi$ can be replaced by a variable v_p, which is the component of the vector $\vec{v}$ in a direction perpendicular to $\vec{r}$. Thus we can write

$$r\,v_p = h. \tag{29}$$

Another expression for v_p is

$$v_p = r\frac{d\phi}{dt}, \tag{30}$$

where $d\phi/dt$ is the planet's *angular velocity*. Equation (29) can therefore be written as

$$r^2\frac{d\phi}{dt} = h. \tag{31}$$

Hence by integration of (31) we have

$$t = t_0 + \frac{1}{h}\int r^2\,d\phi, \tag{32}$$

where t_0 is a constant to be regarded as the time at which observation of the orbit began. We must also mention that, of necessity, the integral in (32) is *definite* and should be evaluated over the integrating variable between points marking the start and end of the orbit section of interest. Equation (32) is in fact Kepler's second law, since the integral on the right is twice the area the radius vector sweeps out in the time $\Delta t = t - t_0$.

Substituting for r in (32) using equation (22) leads to

$$t = t_0 + h^3\int\frac{d\phi}{(\lambda + c\cos\phi)^2}. \tag{33}$$

Alternatively, we can proceed by differentiating (22) to obtain

$$dr = \frac{e}{p}r^2\sin\phi\,d\phi \tag{34}$$

and substituting for $d\phi$ in (32) to obtain

$$t = t_o + \frac{h}{c}\int \frac{dr}{\sin\phi}. \qquad (35)$$

An expression for $\sin\phi$ in terms of r follows from rearranging (22) into

$$\sin\phi = \left(1 - \frac{\lambda^2}{c^2}\left(\frac{h^2}{\lambda r} - 1\right)^2\right)^{1/2} \qquad (36)$$

Then substituting this into (35) and rearranging leads to the final result

$$t = t_0 + \int \frac{rdr}{\left(r^2(c^2 - \lambda^2)/h^2 + 2\lambda r - h^2\right)^{1/2}}. \qquad (37)$$

Both (33) and (37) represent formal solutions rather than practical ones, (though they may be amenable to numerical integration). Nevertheless in principle these afford a means of calculating the time the planet takes to move from one position to another in its orbit. This is the so-called *time of flight*.

However, what is more often required is not the time of flight, but the answer to a more important question: given a time of flight interval, how is the planet's position determined? The answer is key to constructing an astronomical *ephemeris* detailing where a planet can be found at any given time. This is best handled using Kepler's equation, which we shall deal with later.

In the following section we discuss the standard set of parameters for defining an orbit in space: the so-called *Orbital Elements*.

The Orbital Elements

The orbital elements are described with reference to Figure 1, in which the outer circle represents the celestial sphere and the blue disc represents the *plane of the ecliptic*, which is the plane of the Earth's orbit around the Sun. It is with respect to this plane that the elements of all planet orbits are defined. The vector $\vec{p}$ is perpendicular to the ecliptic plane and originates at the position of the Sun. Its magnitude is arbitrary and is usually set to unity. A second vector $\vec{\nu}$ lies in the ecliptic plane and points towards a fixed point on the celestial sphere called the *First Point of Aries*. For mathematical purposes, the vectors $\vec{\nu}$ and $\vec{p}$ represent respectively the x- and z- axes of a 3D Cartesian coordinate system, for which the y- axis (not shown) is orthogonal to both these axes. Together these axes define the *ecliptic* *coordinate system*, which has the Sun at the origin.

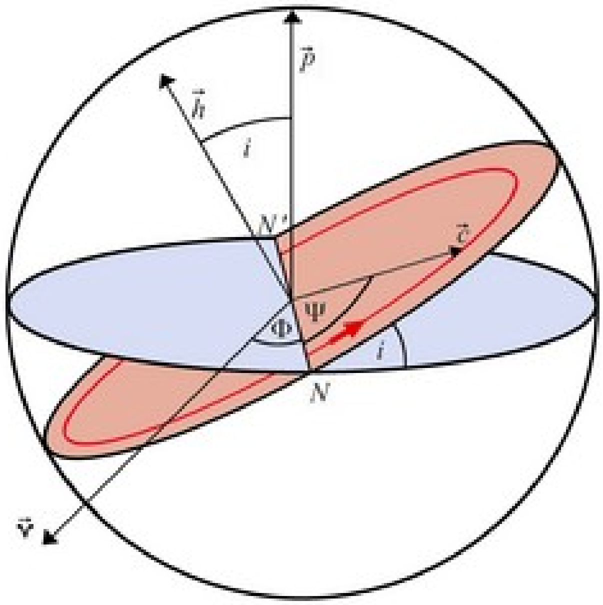

Figure 1: Orbital Elements

The amber disc in Figure 1 is the plane of the orbit of a planet other than the Earth. (The planet's orbit is drawn in red.) The vector $\vec{h}$ is perpendicular to this and also originates at the Sun's position. The intersection of the orbital plane and the ecliptic is the straight line NN', which is called the *line of nodes*, with the points N and N' as the nodes concerned. The vector $\vec{c}$ locates the perihelion - the point in the orbit where the planet is closest to the Sun. The direction of the planet's motion in its orbit is indicated in Figure 1 by a red arrow.

We can now specify the orbital elements as follows.

1. a is a constant specifying the *orbital size*. In the case of an elliptical orbit it is the length of the semi-major axis (or half the longest axis of the ellipse). For a hyperbolic orbit it is the *shortest* distance between the two branches of the hyperbola. This is perhaps a little too abstract for easy interpretation, so it is pertinent to note that the size of *any* orbit can be characterised by the distance between the planet and the Sun at perihelion. This is probably the best practical option for both hyperbolic and parabolic orbits which frequently pertain to comets.

2. e is the eccentricity, which defines the shape of the orbit. The

relation between e and the orbit shape is specified in equation (26).

3. i is the *angle of inclination* of the orbit. This is the tilt of the orbital plane with respect to the ecliptic, which is also the angle between the vectors $\vec{p}$ and $\vec{h}$ shown in Figure 1.

4. Φ is the *angle of the ascending node*, which is the angle between the line of nodes NN' and the first point of Aries $\vec{\nu}$ as shown in Figure 1. The ascending node is where the planet crosses the plane of the ecliptic travelling from a negative to a positive coordinate in the z-direction. (This is near the point N in Figure 1.)

5. Ψ is the *angle of perihelion*, which is the angle between the line of nodes NN' and the vector $\vec{c}$ locating the perihelion (see Figure 1).

6. T_0 is the *epoch of perihelion*, which is the time at which the planet is at perihelion.

Together the orbital elements constitute the minimum information required to permit a complete determination of the location of a planet in space. They represent the standard way for specifying an orbit in the astronomical literature. It is useful therefore to derive a relation between the orbital elements and the dynamical variables of planetary motion encountered in the previous section.

The element e is defined by (25) as the ratio $c/\lambda,$ where c is the perihelion distance (as determined by observation) and λ is the fixed constant for the solar system defined in equation (4).

The orbit size parameter can be obtained from the latus rectum $p,$ which is defined in equation (24). Firstly however, we must determine the vector $\vec{h},$ which is defined in terms of the basic dynamical variables $\vec{r}$ and $\dot{\vec{r}}$ through equation (9). Using $\vec{h},$ we can calculate the latus rectum using equation (24). The next step is to calculate the perihelion distance c using the orbit equation (23). Setting $\phi=0$ in (23) leads to

$$r_{min}=\frac{p}{1+e}, \tag{38}$$

where r_{min} is the required perihelion distance, which is the same as parameter $c.$ For a parabolic orbit $(e=1)$ we use (24) and (38) to obtain

$$r_{min} = \frac{p}{2} = \frac{h^2}{2\lambda} \quad \text{(parabola)}, \tag{39}$$

which is enough to specify the size of a parabolic orbit.

For a hyperbolic orbit ($e > 1$) we note that a hyperbola formally has two branches (two separate curves) and two distinct focii such that straight lines drawn from each focus to at a point on one branch obey the relation

$$d_2 - d_1 = 2a, \tag{40}$$

where d_1 and d_2 are the distances from the two focii to the point on the chosen branch and $2a$ is the closest distance between the two branches. (This property is similar to the ellipse property that requires the corresponding distances to sum to the major diagonal of the ellipse.) If the common point on the hyperbola is where the latus rectum intercepts the branch (for d_1 say, so $d_1 = p$) then the triangle formed from this point and the two focii is right angled and so by Pythagoras we can write

$$d_2^2 = p^2 + (2a + 2r_{min})^2, \tag{41}$$

where the term in brackets is the distance between the two focii and r_{min} is the perihelion distance. Thus combining equations (38), (40) and (41) leads to

$$(2a - p)^2 = p^2 + 4(a + p/(1+e))^{2.} \tag{42}$$

Expanding (42) and rearranging produces the result

$$a = \frac{p}{(e^2 - 1)} = \frac{\lambda h^2}{(c^2 - \lambda^2)}, \quad \text{(hyperbola)}, \tag{43}$$

which defines the size of the hyperbolic orbit. Note that we could also use r_{min} as the defining parameter for a hyperbola, which from (38), (24) and (25) is

$$r_{min} = \frac{h^2}{(\lambda + c)} \quad \text{(hyperbola)}. \tag{44}$$

Equation (44) could also be used with an elliptical orbit, which would be relevant to a *periodic* comet.

Lastly, for an ellipse ($e < 1$), setting $\phi = 180^0$ in (23) gives

$$r_{max}=\frac{p}{1-e}, \tag{45}$$

which represents the greatest distance between the planet and the Sun (the so-called *aphelion*). The sum of distances r_{min} and r_{max} equals $2a$, the largest diameter of the ellipse, and so from (38), (45), (24) and (25) we have

$$a=\frac{1}{2}(r_{min}+r_{max})=\frac{p}{(1-e^2)}=\frac{\lambda h^2}{(\lambda^2-c^2)}, \quad \text{(ellipse)} \tag{46}$$

which defines the size of the elliptical orbit.

In Figure 1 the vector $\vec{p}$ represents the z- axis of the ecliptic coordinate system. In which case the angle of inclination i can be obtained from

$$\cos i=\frac{\vec{p}\cdot\vec{h}}{ph}, \tag{47}$$

where p and h are the magnitudes of $\vec{p}$ and $\vec{h}$.

From Figure 1 we can define a vector $\vec{n}$ such that

$$\vec{n}=\vec{p}\times\vec{h}. \tag{48}$$

This vector points along the line of nodes NN' in the direction of the ascending node, which means we can obtain the angle of the ascending node from

$$\cos\Phi=\frac{\vec{n}\cdot\vec{\nu}}{n\nu}, \tag{49}$$

where n and ν are the magnitudes of $\vec{n}$ and $\vec{\nu}$.

The vector $\vec{n}$ lies along the line of nodes and therefore also lies in the orbital plane of the planet. It can therefore be used to obtain the angle of the perihelion Ψ using the formula

$$\cos\Psi=\frac{\vec{n}\cdot\vec{c}}{nc}. \tag{50}$$

Lastly, the epoch of perihelion T_0 is obtained by integrating equations (33) or (37) from a known position (r_0, ϕ_0) and time t_o to the perihelion position where $t=T_0$. It is much simpler however to use Kepler's equation, which is discussed in the next section.

Kepler's Equation

Kepler did not know anything about calculus, which hadn't been invented in his day, but he was a skilled geometer and mathematician. So in order to construct a planetary ephemeris, he provided his own solution, which is based on Figure 2, which we now describe.

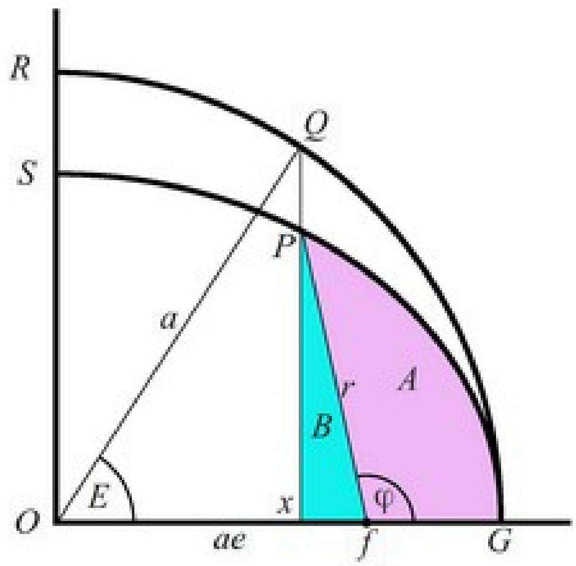

Figure 2. Kepler's Orbit Diagram

In Figure 2 the curve SPG represents part of an elliptical orbit (which is the only kind of orbit Kepler knew about!) The line OG is half the longest diameter of the ellipse and has a length a. Point O is thus at the centre of the elliptical orbit. Point f is the focus of the ellipse, where the Sun resides and the distance Of equals ae, where e is the eccentricity of the ellipse. The point P is the position of the planet on the elliptical path. The curve RQG is part of a circle, which is centred at O and has a radius a. Both curves meet at the point G on the AG axis. A perpendicular line dropped from the point P to the axis OG at x, is extended upwards to meet the circle at Q. The line drawn from O to Q has length a, since Q is on the circle. The angle E this line makes with the axis OG at O, is known as the *eccentric anomaly*. We have also drawn the line fP, which is of length r, representing the distance from the Sun to the planet at P.

The angle ϕ this line makes with axis OG at f, is called the *true anomaly*. Kepler's aim was to describe the motion of the planets using the eccentric anomaly E rather than the true anomaly ϕ. By doing this he found the time of flight problem much easier to handle.

It can be easily shown for Figure 2 that from any point on the OG axis, (such as x,) the perpendicular distance y_e to the ellipse is a constant fraction of the corresponding distance y_c to the circle, so the following ratio always holds

$$\frac{y_e}{y_c}=(1-e^2)^{1/2}. \tag{51}$$

This relation is required to derive Kepler's equation. The aim is to calculate the area marked A in Figure 2 by proceeding as follows.

Firstly the area of the slice QOG of the outer circle in Figure 2 is

$$Area \;\; QOG=\frac{a^2}{2}E. \tag{52}$$

Next the area of the triangle QOx is

$$Area \;\; QOx=\frac{a^2}{2}\sin E\cos E. \tag{53}$$

The area of the (half) sector of the circle QxG is therefore the difference between (52) and (53):

$$Area \;\; QxG=\frac{a^2}{2}(E-\sin E\cos E). \tag{54}$$

Using the relation (51) the area of the sector PxG of the ellipse is therefore

$$Area \;\; PxG=(1-e^2)^{1/2}\frac{a^2}{2}(E-\sin E\cos E). \tag{55}$$

The area of triangle marked B in Figure 2 is

$$Area \;\; B=\frac{1}{2}(ae-a\cos E)(1-e^2)^{1/2}a\sin E, \tag{56}$$

where we have again exploited the relation (51).

The area A is now the difference between (55) and (56), which is

$$Area \quad A=(1-e^2)^{1/2}\frac{a^2}{2}(E-\sin E). \tag{57}$$

According to Kepler's second law, the area A is proportional to the time the planet has taken getting from G to P. Likewise the area of the whole elliptical orbit is proportional to the period T of the whole orbit i.e.

$$\frac{t-t_0}{T}=\frac{Area \quad A}{Area \quad Ellipse}. \tag{58}$$

Using the standard ellipse area, which is given by

$$Area \quad Ellipse=\pi a^2(1-e^2)^{1/2} \tag{59}$$

and (57), the equation (58) can be rearranged into the form

$$\Delta t=t-t_0=\frac{T}{2\pi}(E-e\sin E). \tag{60}$$

This is Kepler's equation. Like equations (33) and (37) it provides a solution for the time of flight between two places in the planetary orbit. However its advantages are that it is a simple analytical form rather than an integral and it can also be used iteratively to calculate the eccentric anomaly E for a given Δt, and thus determine the position of the planet. This requires a conversion formula from the eccentric to the true anomaly. For this we consult Figure 2 and write

$$(1-e^2)^{1/2} a\sin E=r\sin\phi , \tag{61}$$

from which we obtain

$$\sin\phi=\frac{(1-e^2)^{1/2}}{r}a\sin E. \tag{62}$$

We also note that

$$r^2=(ae-e\cos E)^2+(1-e^2)a^2\sin^2 E, \tag{63}$$

which reduces to

$$r=a(1-e\cos E). \tag{64}$$

Inserting this into (62) gives the result

$$\sin\phi = \frac{\left(1-e^2\right)^{1/2}\sin E}{\left(1-e\cos E\right)}. \tag{65}$$

Thus using the eccentric anomaly E, both r and φ can be obtained from (64) and (65) respectively. Alternatively r can be obtained from (22) once φ has been obtained from (65).

We can also obtain Kepler's equation without using the area relation (58). Inserting equation (62) into (35) gives

$$\Delta t = \frac{h}{a c\left(1-e^2\right)^{1/2}}\int \frac{r}{\sin E}\, dr. \tag{66}$$

Differentiating (64) gives

$$dr = a\, e\sin E\, dE. \tag{67}$$

Substituting for r and dr in (66) using (64) and (67) gives

$$\Delta t = \frac{h e a}{c\left(1-e^2\right)^{1/2}}\int \left(1-e\cos E\right) dE. \tag{68}$$

This reduces to

$$\Delta t = \sqrt{\frac{a^3}{\lambda}}\left(E - e\sin E\right), \qquad \text{(where } e<1\text{)}, \tag{69}$$

where we have used relations (24) and (25) together with the ellipse property

$$p = a\left(1-e^2\right) \tag{70}$$

to reduce the constant terms. Equation (69) is again Kepler's equation. Comparing this with equation (60) leads to the conclusion that

$$T = 2\pi\sqrt{\frac{a^3}{\lambda}}. \tag{71}$$

Which is Newton's expression of Kepler's third law.

Kepler's equation in its original form is applicable only to elliptical orbits, but it can be adapted for hyperbolic orbits (i.e. when $e>1$). In place of equation (64) we write

$$r = a(e\cosh F - 1), \tag{72}$$

where $\cosh$ is the so-called *hyperbolic cosine* function and F is the eccentric anomaly for a hyperbola. In this case Kepler's equation (69) becomes

$$\Delta t = \sqrt{\frac{a^3}{\lambda}}(e\sinh F - F), \quad \text{(where } e > 1\text{)}, \tag{73}$$

Where $\sinh$ is the *hyperbolic sine* function. Equation (73) may once again be solved iteratively to determine F. Once F is known, the true anomaly ϕ is obtained from

$$\cos\phi = \frac{(\cosh F - e)}{(1 - e\cosh F)} \tag{74}$$

and then r is calculated using (22).

The appropriate form of Kepler's equation for parabolic orbits where $e = 1$ is due to Barker and is obtained as follows.

The equation for a parabolic orbit based on (22) is

$$r = \frac{(h^2/\lambda)}{1 + \cos\phi} = \frac{h^2}{2\lambda} sec^2 \frac{\phi}{2}. \tag{75}$$

Substituting this into (31) leads to

$$\int sec^4 \frac{\phi}{2} d\phi = \frac{4\lambda^2}{h^3} \int dt. \tag{76}$$

This equation can be expanded into

$$\int sec^2 \frac{\phi}{2}\left(1 + \tan^2 \frac{\phi}{2}\right) d\phi = \frac{4\lambda^2}{h^3} \int dt, \tag{77}$$

which integrates to

$$\tan\frac{\phi}{2} + \frac{1}{3}\tan^3\frac{\phi}{2} = \frac{2\lambda^2}{h^3}(t - t_0) = \frac{2\lambda^2}{h^3}\Delta t. \tag{78}$$

Equation (78) is Barker's equation for which tabulates solutions exist. An analytical solution is obtained using the substitution

$$\tan\frac{\varphi}{2}=2\cot 2\omega=\cot\omega-\tan\omega. \quad (79)$$

Substituting this into (78) leads to

$$\cot^3\omega-\tan^3\omega=\frac{2\lambda^2}{h^3}\Delta t. \quad (80)$$

Now let

$$\cot^3\omega=\cot\frac{s}{2}. \quad (81)$$

So (80) becomes

$$\cot\frac{s}{2}-\tan\frac{s}{2}=\frac{2\lambda^2}{h^3}\Delta t, \quad (82)$$

or

$$2\cot s=\frac{2\lambda^2}{h^3}\Delta t. \quad (83)$$

Thus solving for s in (83) provides ω from (81) and hence φ from (79). Position r is then obtained from (75).

Fitting an Orbit to Observed Planetary Positions

One of the earliest and most difficult challenges in mathematical astronomy was fitting a planterary orbit to observed planetary positions. The first person to do this correctly was Kepler, who fitted the orbit of Mars to observations made by Tycho Brahe. Previously to this, it is fair to say, the correct answer had evaded the greatest minds of antiquity. However, Kepler's breakthrough by no means made the problem simple. Thereafter much remained to be done by giants like Newton, Laplace, Gauss and others, who derived procedures of great ingenuity but intimidating complexity. The problem is a tedious one even today, even when assisted by the availability of modern computers to handle the mass of arithmetic that the work entails.

A relatively simple example of a fitting method is presented here. It is due to J.W. Gibbs, who is also famous for his work in theoretical thermodynamics. It is a mathematically elegant method but unfortunately is not immediately of use to amateur astronomers since it requires as input the three vectors $\vec{r}_1$, $\vec{r}_2$ and $\vec{r}_3$ that define, in ecliptic coordinates, three positions of the planet with respect to Sun at known times t_1, t_2 and t_3, (where $t_1 < t_2 < t_3$). Providing these vectors is nontrivial task, to say the least, but the method has uses as part of an overall strategy for fitting orbits in general.

We start by noting that we may use two positions to define a vector $\vec{g}$

$$\vec{g} = \vec{r}_1 \times \vec{r}_3 \tag{84}$$

which is aligned in the same direction as the vector $\vec{h}$ from equation (9). For this reason it may be used in place of $\vec{h}$ in the procedures given in equations (47) through to (49) to directly obtain the angle of inclination i and the longitude of the ascending node Φ. These are two of the six orbital elements we need to determine. The next stage will provide three more.

Firstly we write the orbit equation (23) in the form

$$r = \frac{p}{1 + e\cos(\phi - \phi_0)}, \tag{85}$$

where r is the Sun-planet distance, p the latus rectum of the orbit and e is the orbit eccentricity. The angle ϕ is the planet's anomaly measured with respect to the line of nodes vector (given by equation (48)) [2] and ϕ_0 is the (currently unknown) angle defining the angle of the perihelion, also with respect to the line of nodes. Equation (85) can be rearranged into

$$\left(\frac{p}{r} - 1\right) = e\cos(\phi - \phi_0). \tag{86}$$

Also, from standard trigonometry, we can write

$$\cos(\phi - \phi_0) = \cos\phi\cos\phi_0 + \sin\phi\sin\phi_0. \tag{87}$$

Let

$$u_i = \left(\frac{p}{r_i} - 1\right), \qquad \text{where} \quad i = 1,2,3\,. \tag{88}$$

and then by using (87) and (88) we may write (86) as

2 The line of nodes is already known at this stage – it was used to calculate the longitude of the ascending node.

$$
\begin{aligned}
u_1 &= e\,c_0\,c_1 + e\,s_0\,s_1\,,\\
u_2 &= e\,c_0\,c_2 + e\,s_0\,s_2\,,\\
u_3 &= e\,c_0\,c_3 + e\,s_0\,s_3\,,
\end{aligned} \tag{89}
$$

for all three planetary positions, where

$$
\begin{aligned}
c_i &= \cos\phi_i\,, \qquad & c_0 &= \cos\phi_0\,,\\
s_i &= \sin\phi_i\,, \qquad & s_0 &= \sin\phi_0\,.
\end{aligned} \tag{90}
$$

Multiplying the first equation of (89) by s_3 and the third equation by s_1 gives

$$
\begin{aligned}
u_1\,s_3 &= e\,c_0\,c_1\,s_3 + e\,s_0\,s_1\,s_3,\\
u_3\,s_1 &= e\,c_0\,s_1\,c_3 + e\,s_0\,s_1\,s_3
\end{aligned} \tag{91}
$$

and the difference between these two equations is

$$
u_1\,s_3 - u_3\,s_1 = e\,c_0(s_3\,c_1 - s_1\,c_3) = e\,c_0\sin(\phi_3 - \phi_1) = e\,c_o\,s_{13}\,, \tag{92}
$$

where evidently

$$
s_{13} = \sin(\phi_3 - \phi_1). \tag{93}
$$

Similarly we can show that

$$
u_2\,s_3 - u_3\,s_2 = e\,c_0(s_3\,c_2 - s_2\,c_3) = e\,c_0\sin(\phi_3 - \phi_2) = e\,c_o\,s_{23}\,, \tag{94}
$$

where

$$
s_{23} = \sin(\phi_3 - \phi_2). \tag{95}
$$

Combining (92) and (94) to eliminate e results in the equation

$$
s_{23}(u_1\,s_3 - u_3\,s_1) = s_{13}(u_2\,s_3 - u_3\,s_2). \tag{96}
$$

Now, expanding $u_1,$ u_2 and u_3 in (96) using (88) and rearranging the result to isolate the variable p leads after some effort to

$$
p = \frac{(s_{13}(s_2 - s_3) - s_{23}(s_1 - s_3))}{((s_2 s_{13} - s_1 s_{23})/r_3 + s_3(s_{23}/r_1 - s_{13}/r_2))}. \tag{97}
$$

By this equation the latus rectum p of the unknown orbit can be calculated. This may be substituted back into equations (88) to obtain u_1 and u_3, which we then insert into (92) to obtain

$$e\,c_0 = (u_1 s_3 - u_3 s_1)/s_{13.} \tag{98}$$

We may also obtain from (89) the following pair of equations

$$\begin{aligned} u_1 c_3 &= e\,c_0 c_1 c_3 + e\,s_0 s_1 c_{3,} \\ u_3 c_1 &= e\,c_0 c_1 c_3 + e\,s_0 c_1 s_{3.} \end{aligned} \tag{99}$$

Subtracting one from the other provides the result

$$e\,s_0 = (u_3 c_1 - u_1 c_3)/s_{13.} \tag{100}$$

Dividing (100) by (98) leads to the following

$$\tan\phi_0 = \frac{e\,s_0}{e\,c_0} = \frac{(u_3 c_1 - u_1 c_3)}{(u_1 s_3 - u_3 s_1)}. \tag{101}$$

Also, from (98) and (100) we have

$$e = ((e\,c_0)^2 + (e\,s_0)^2)^{1/2}. \tag{102}$$

Thus we now have the eccentricity $e,$ the angle of perihelion ϕ_0 and the latus rectum $p.$ As we saw in the earlier section on the orbital elements, this is sufficient information to specify the orbit size. Knowledge of this and the angle of perihelion fixes the perihelion vector and then Kepler's equation may be used to compute the epoch T_0.

What is unusual about this procedure, is that it is *overdetermined*: it should require only six coordinates fully specify an orbit, but here we are using three vectors with three components each, which is nine coordinates in all! Other methods, such as those of Laplace and Gauss require just six coordinates.

Appendix: Basic Vector Maths

Presented here is a minimalist description of vectors to assist the reader in understanding the text above. It is by no means a complete account.

A vector is a quantity that has both a magnitude and direction. Common examples of vectors include position $(\vec{r}),$ velocity $(\vec{v}),$ acceleration $(\vec{a})$ and force $(\vec{f}).$ All of these possess a direction as well as a magnitude and can be visualised as a physical 'arrow' which points in the required direction with a particular magnitude or 'strength'. The magnitude (also called the modulus), is a scalar quantity which, on its own, does not indicate a direction. Thus, for example, we have $r,$ which specifies the scalar length associated with the position vector $\vec{r}$. Another example is v which specifies the scalar speed associated with the velocity vector $\vec{v}$, and so on.

This means that a numerical representation of a vector must contain more than just one number. The numbers may be specified in many ways, but the most common (by far) is the Cartesian representation, which is a series of numbers enclosed in brackets, such as

$$\vec{r}=(x,y) \quad \text{and} \quad \vec{r}=(x,y,z). \tag{103}$$

The first of these represents a two dimensional vector that exists in a 2D plane, while the second represents a three dimensional vector in 3D space. The numbers in this case are none other than the position coordinates with respect to some chosen fixed origin and a set of mutually orthogonal axes (which is known as the reference frame). The magnitudes of these vectors are given by the Pythagoras rule:

$$r=(x^2+y^2)^{1/2} \quad \text{and} \quad r=(x^2+y^2+z^2)^{1/2}. \tag{104}$$

We can do the same with any vector e.g. the force vector:

$$\vec{f}=(f_x,f_y,f_z) \quad \text{with} \quad f=(f_x^2+f_y^2+f_z^2)^{1/2}, \tag{105}$$

where f_x, f_y, f_z are the Cartesian components of the force vector $\vec{f}$. The quantity f is thus the magnitude, or strength of the force. Most of the mathematics of vectors is concerned with calculating the properties of the vectors in terms of the components, but in this paper on orbital mechanics, this is not needed, so we will not elaborate on the subject here.

An important idea to hold on to in the context of this paper is that a vector represents a real physical entity – as much a physical entity as more familiar scalar quantities like mass (m), temperture (T) and energy (E). They can be changed by physics as much as any physical entity can. The fact that they are defined by more than one number should not confuse you. They are quite well behaved physically and do not lose their meaning because the number are changing; the vector retains the same physical identity.

Vectors can be added and subtracted in a manner similar to ordinary numbers. The addition of two vectors is written algebraically as

$$\vec{a}=\vec{b}+\vec{c}, \tag{106}$$

and this is represented geometrically by the parallelogram diagram in Figure A1.

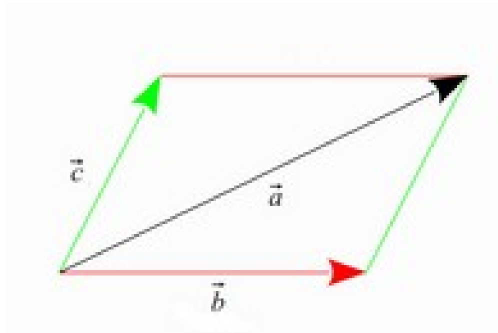

Figure A1. Vector Addition

Here the vectors $\vec{b}$ and $\vec{c}$ are joined at their origins and a parallelogram completed as show. The vector $\vec{a}$ drawn on the diagonal between $\vec{b}$ and $\vec{c}$ is the resultant vector of the sum (106).

Vector subtraction works in much the same way. So for the subtraction

$$\vec{a}=\vec{b}-\vec{c}\,, \tag{107}$$

the corresponding geometric construction is shown in Figure A2.

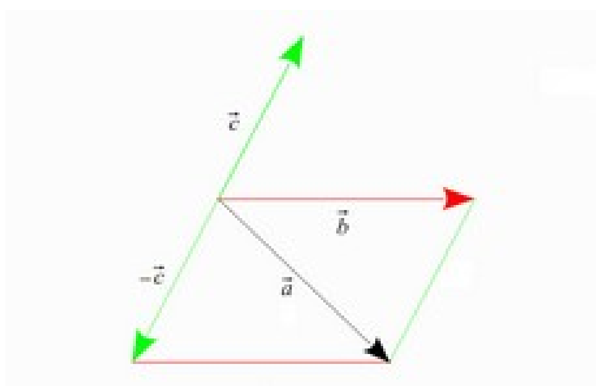

Figure A2. Vector Subtraction

Note that in this case the subtracted vector $(\vec{c})$ is first reversed in direction and then added. Another way of thinking about this is to reconsider Figure A1 and instead of drawing the diagonal between the vectors as shown, take the alternative diagonal and draw vector $\vec{a}$ from the tip of vector $\vec{c}$ to the tip of vector $\vec{b}$.

Multiplication of a vector by a scalar number is simple. The product

$$\vec{A}=c\,\vec{a}, \tag{108}$$

means that the vector $\vec{A}$ is c times bigger than vector $\vec{a}$. Division of a vector by a scalar works as you might expect.

The next thing we need to know (for this article) is how to find the derivative of a vector with respect to time. Formally this is written as

$$\frac{d\vec{r}}{dt} = lim\, \delta t \rightarrow 0 : \frac{\vec{r}(t+\delta t) - \vec{r}(t)}{\delta t}, \tag{109}$$

which is essentially the same as for a scalar variable. The main purpose for writing this out explicitly is to make it clear that the vectors $\vec{r}(t+\delta t)$ and $\vec{r}(t)$ need not point in the same direction and so the derivative necessarily encapsulates the vector's change of direction as well as its change in magnitude.

In physics it is conventional to indicate derivatives with respect to time using a 'dot' notation thus:

$$\dot{\vec{r}} = \frac{d\vec{r}}{dt} \quad \text{and} \quad \ddot{\vec{r}} = \frac{d\dot{\vec{r}}}{dt} = \frac{d^2\vec{r}}{dt^2} \quad \text{etc.} \tag{110}$$

It is rare to go beyond the second derivative with this notation.

If the vectors are written in a Cartesian representation, the derivative with respect to time is easy to specify.

$$\dot{\vec{r}} = \left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt}\right). \tag{111}$$

Thus the vector derivative is constructed from the derivatives of the components.

Next we consider multiplying a vector by a vector. There are two cases to consider: one where the result is a scalar; and one where the result is a vector. The first is commonly known as the *scalar product* and also as the *dot product* and is defined by the formula

$$a = \vec{b} \cdot \vec{c} = b\, c \cos\theta, \tag{112}$$

where a is the resultant scalar quantity, b and c are the magnitudes of their respective vectors and θ is the angle between the two vectors as indicated in Figure A3.

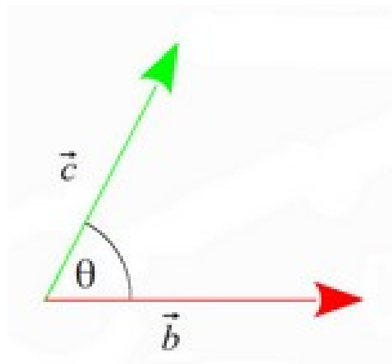

Figure A3. The Scalar or Dot Product

Note that $\theta \leq 180^o$ always and that the scalar product is zero when $\theta = 90^o$, regardless of the magnitudes of $\vec{b}$ and $\vec{c}$. When this happens vectors $\vec{b}$ and $\vec{c}$ are said to be orthogonal to each other. With reference to equation (112), we note that the quantity $b\cos\theta$ may be regarded as the component of $\vec{b}$ in the direction of $\vec{c}$ and likewise $c\cos\theta$ is the component of $\vec{c}$ the direction of $\vec{b}$. This offers a useful way of exploiting the scalar product to resolve vectors into their components.

The vector product (also called the cross product) is a more complicated operation defined as

$$\vec{a} = \vec{b} \times \vec{c} = \hat{a}\, b\, c \sin\theta, \tag{113}$$

in which $\theta \leq 180^o$ is once again the angle between the vectors $\vec{b}$ and $\vec{c}$, as for the scalar product. Scalars b and c are the magnitudes of their respective vectors and $\hat{a}$ is a so-called unit vector, which is orthogonal to both $\vec{b}$ and $\vec{c}$. Note that the order of the vectors in the product $\vec{b} \times \vec{c}$ is important. If it is written as $\vec{c} \times \vec{b}$ the result will be different . In fact we can write

$$\vec{b} \times \vec{c} = -\vec{c} \times \vec{b}, \tag{114}$$

a result which will become evident later.

The magnitude a of the resultant vector $\vec{a}$ in equation (113) is

$$a = b\, c \sin\theta. \tag{115}$$

Note that the magnitude of $\vec{a}$ is zero when $\theta = 0^o$, in which case $\vec{a}$ is called a null vector. It follows that a vector cross multiplied with itself must produce a null result since the angle between both vectors is necessarily zero.

The direction of the vector $\vec{a}$ is indicated by the unit vector $\hat{a}$. This vector has a magnitude of 1 and is related to vector $\vec{a}$ by simple scalar multiplication:

$$\vec{a}=a\,\hat{a}\,. \tag{116}$$

The direction of both $\vec{a}$ and $\hat{a}$ is as shown in Figure 4, which is to be understood in the following manner.

The vectors $\vec{b}$ and $\vec{c}$ together define a flat plane (viewed from above in Figure A4). The unit vector $\hat{a}$ and vector $\vec{a}$ are perpendicular to this plane by definition, but we also need to decide if $\hat{a}$ points above the plane or below the plane. Which one it is is determined in the following manner (which applies specifically when the cross product is written in the order $\vec{b}\times\vec{c}$).

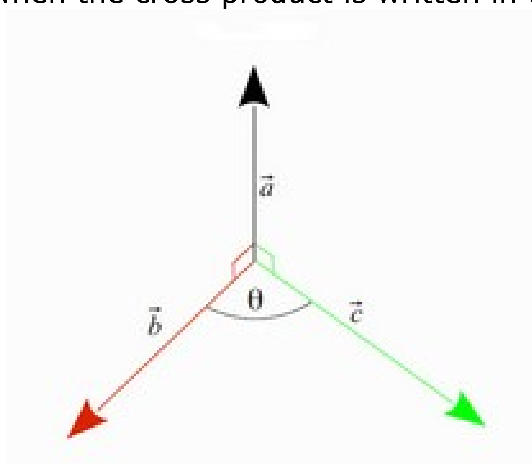

Figure A4. The Vector or Cross Product $\vec{b}\times\vec{c}$

Look down on the flat plane defined by vectors $\vec{b}$ and $\vec{c}$. Then imagine holding vector $\vec{b}$ fixed while vector $\vec{c}$ is rotated in the plane about its origin so as to decrease θ. If $\vec{c}$ rotates clockwise, then $\hat{a}$ points above the plane (as in Figure A4). If the rotation is anticlockwise, $\hat{a}$ points below the plane[3]. This is a complicated recipe that seem hard to justify. However there are innumerable physical systems in which this form of multiplication is invaluable. This is particularly so where rotational motion is involved, then it seems completely natural. So it pays to bite the bullet - learn the recipe by heart and watch the knowledge pay off in the future!

We now go beyond using the scalar or vector products in isolation. There are many occasions in physics where two or more products come together in the same mathematical expression. The most common examples are the *triple scalar product* and the *triple vector product*, which we now examine.

3 Now consider what happens when the order of vectors is $\vec{c}\times\vec{b}$ - same result?

The triple scalar product takes the following form

$$s = \vec{a} \cdot \vec{b} \times \vec{c}, \tag{117}$$

in which $\vec{a}$, $\vec{b}$ and $\vec{c}$ are the vectors and s is the scalar result. Note that the cross product $\times$ must be performed first, otherwise the expression would have no meaning. Now, according to equation (113), we can write the cross product in (117) as

$$\vec{j} = \vec{b} \times \vec{c} = \hat{j}\, bc \sin\theta, \tag{118}$$

where $\hat{j}$ is a unit vector along a direction perpendicular to both $\vec{b}$ and $\vec{c}$.

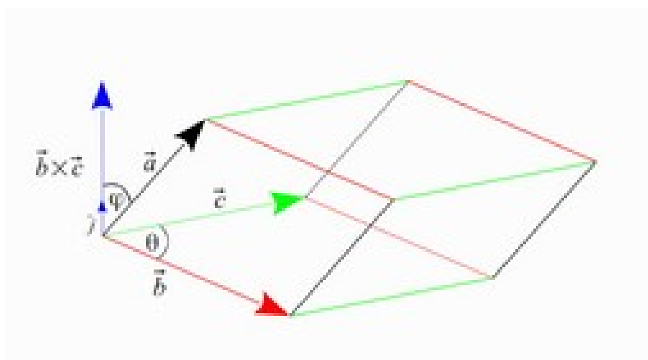

Figure A5. The Triple Scalar Product

Then the scalar product of $\vec{a}$ with the vector $\vec{j}$ is

$$s = \vec{a} \cdot \vec{j} = a \cos\phi\; bc \sin\theta, \tag{119}$$

where $a\cos\phi$ is the component of vector $\vec{a}$ in the direction of $\hat{j}$ and φ is the angle between vectors $\vec{a}$ and $\vec{j}$. Result (119) can be understood with reference to Figure 5A, where all the vectors concerned are drawn. In this figure we have constructed a parallelepiped base on the vectors $\vec{a}$, $\vec{b}$ and $\vec{c}$. We can see from this construction that the expression $bc\sin\theta$ defines the area of the face of the parallelepiped in the plane of vectors $\vec{b}$ and $\vec{c}$ while $a\cos\phi$ defines its perpendicular height. It follows that the scalar s equals the *volume* of the parallelepiped. It is easy from this identification to establish that

$$\vec{a} \cdot \vec{b} \times \vec{c} = \vec{b} \cdot \vec{c} \times \vec{a} = \vec{c} \cdot \vec{a} \times \vec{b}, \tag{120}$$

since all these triple scalar products must compute the same volume. Note however that the order of the vectors appearing in the cross products is important, since interchanging them will give a negative result for the volume (though the absolute magnitude of the result will be correct.)

The triple vector product has the form

$$\vec{u}=\vec{a}\times(\vec{b}\times\vec{c}), \tag{121}$$

where $\vec{a}$, $\vec{b}$ and $\vec{c}$ are the vectors of the triple product and $\vec{u}$ is the resultant vector. Note the use of brackets to indicate that the cross product $\vec{b}\times\vec{c}$ is obtained first and then vector $\vec{a}$ is cross multiplied with the result. (A different result is obtained if the product $\vec{a}\times\vec{b}$ is obtained first.) A Geometric interpretation of equation (121) is given in Figure A6.

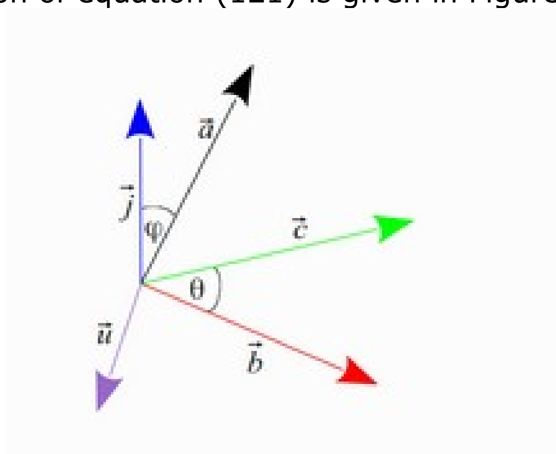

Figure A6. The Vector Triple Product

In Figure A6, The cross product $\vec{b}\times\vec{c}$ results in the vector $\vec{j}$, which is perpendicular to the plane containing both $\vec{b}$ and $\vec{c}$. Vector $\vec{a}$ then cross multiplies $\vec{j}$ to obtain the resultant vector, $\vec{u}$. Since $\vec{u}$ must be perpendicular to the plane containing both $\vec{a}$ and $\vec{j}$ it follows that it must lie in the plane containing vectors $\vec{b}$ and $\vec{c}$. Thus we see that the resultant of vector triple product is a vector that is in the same plane as $\vec{b}$ and $\vec{c}$, but is rotated with respect to both. For this reason vector $\vec{u}$ can be expressed as a vector sum of $\vec{b}$ and $\vec{c}$ i.e.

$$\vec{u}=\alpha\vec{b}+\beta\vec{c}, \tag{122}$$

where α and β are numerical constants. (This follows from a general theorem that any vector in a plane defined by two independent vectors can be expressed as a unique combination of those vectors.) Since we know that $\vec{a}$ also contributes to $\vec{u}$ through (121) we try taking the scalar product of $\vec{a}$ with $\vec{u}$ so equation (122) becomes

$$\vec{a}\cdot\vec{u}=\alpha\vec{a}\cdot\vec{b}+\beta\vec{a}\cdot\vec{c}. \tag{123}$$

Now, since we know that $\vec{a}$ and $\vec{u}$ are orthogonal, the left hand side of this equation is zero. So we can write

$$\alpha\vec{a}\cdot\vec{b}=-\beta\vec{a}\cdot\vec{c}\,. \tag{124}$$

This identity can be satisfied if we chose

$$\alpha=\vec{a}\cdot\vec{c} \quad \text{and} \quad \beta=-\vec{a}\cdot\vec{b}\,, \tag{125}$$

This is correct within an arbitrary multiplying factor, which we may assume is unity. So we can write that

$$\vec{u}=\vec{a}\times(\vec{b}\times\vec{c})=(\vec{a}\cdot\vec{c})\vec{b}-(\vec{a}\cdot\vec{b})\vec{c}\,. \tag{126}$$

This result is extremely useful in vector algebra.

It is of interest to note that since

$$(\vec{a}\times\vec{b})\times\vec{c}=-\vec{c}\times(\vec{a}\times\vec{b})\,, \tag{127}$$

which follows from equation (114), we see that, notwithstanding the change of sign, moving the brackets in the definition of the vector triple product (121) must give a resultant vector in the plane of vectors $\vec{a}$ and $\vec{b}$, whereas the original expression gives a resultant in the plane of $\vec{b}$ and $\vec{c}$, which is obviously a very different result. The position of the brackets are therefore crucial!