EVERYTHING & IN BETWEEN

This is not a help file for the E&IB web site. It is a discussion of some of the underlying physics and mathematics that the site uses. The goal is that readers will have a better understanding of what is being depicted: the real physics of the Solar System and Galaxy, the mathematics that's used to describe them, and the choices that were made in rendering the Solar System and nearby Galaxy.

Introduction.

It doesn't take a rocket scientist to understand orbits – though that community is keenly interested in them. In this introduction we'll discuss the basic concepts underlying orbits and give a guide to the jargon that suffuses and confuses the field. We won't be using any equations quite yet, but we will use a fair number of figures and diagrams to illustrate even some reasonably abstract concepts. We'll mix in a bit of astronomy, physics and mathematics. After this extended introduction, we'll go back through everything again more analytically giving the equations that connect everything together.

The ancients thought that bodies moved around one another in constant motion on circles, with the circles centered on the Earth. It was quickly apparent that just simple circles wouldn't do, no simple circular orbit could be reconciled with the observations. Circles needed to be layered on circles (epicycles) creating chains of circles though still ultimately centered on the Earth. Copernicus championed the idea that instead of the Earth being the center of all of the action, the Sun was the more natural choice. But he still believed in motion in circles, so his Sun-centered cosmology still required complex interacting circles to be able to match what was observed.

It was Kepler who added the detail that clearly showed the superiority of the Sun-centered system: his first law was the planets orbited the Sun not in circles, but in ellipses. Each planet orbit could be represented by a single ellipse, rather than a chain of interlocking circles. So let's start with ellipses.

Ellipses.

An *ellipse* is a kind of squashed circle.

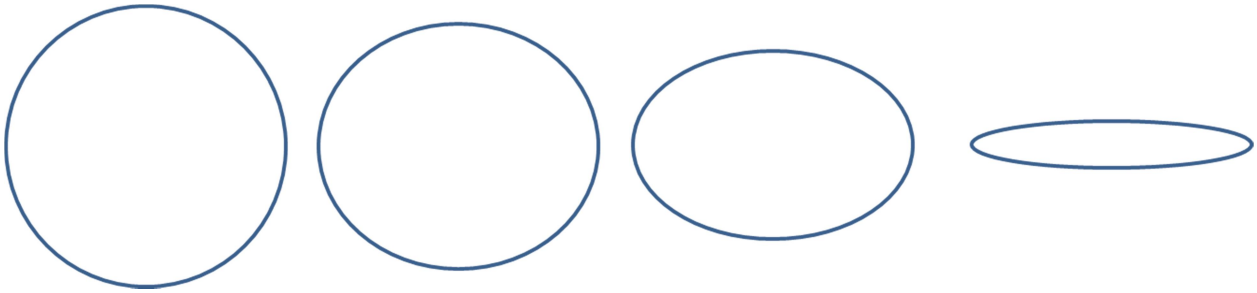


Figure 1. A circle and some ellipses

It's not egg-shaped: the ellipse is completely symmetrical with the two narrower sides looking exactly the same. A ellipse can be nearly circular, like the second figure above, or very elliptical as in the last figure. A circle is just a special case of the ellipse. Planetary orbits are generally close to circular, they tend to match something like one of the first two figures. Comet and asteroid orbits can be more stretched out as in the last two figures. It's a further tribute to Kepler and to the quality of the astronomical observations he used – made by the Danish astronomer Tycho Brahe, that he was able to elucidate the motions of the planets despite the subtleties of the orbital motions. This is all done by eye without the aid of a telescope.

Formally, an ellipse can be defined as the set of points in the plane where the sum for the distance from two specified points in the plane is constant. The two points we're measuring from are called the *foci* or *focusses* of the ellipse.

This says we take any two points in the plane, the foci. Then we can pick a third random point in the plane. We measure the sum of the distances to this third point from the two focuses. The ellipse comprises all of the points in the plane where that sum is the same. If we pick a different sum, we'll get a different ellipse for a given set of foci. And we'll get different ellipses if we change either focus.

Let's do this. Let's get a piece of paper. We secure the paper on top of a cork bulletin board, and put two pins near the middle of the paper. Now we get a loop of string that's long enough to go over both pins, but not too long or our ellipse won't fit on the paper.

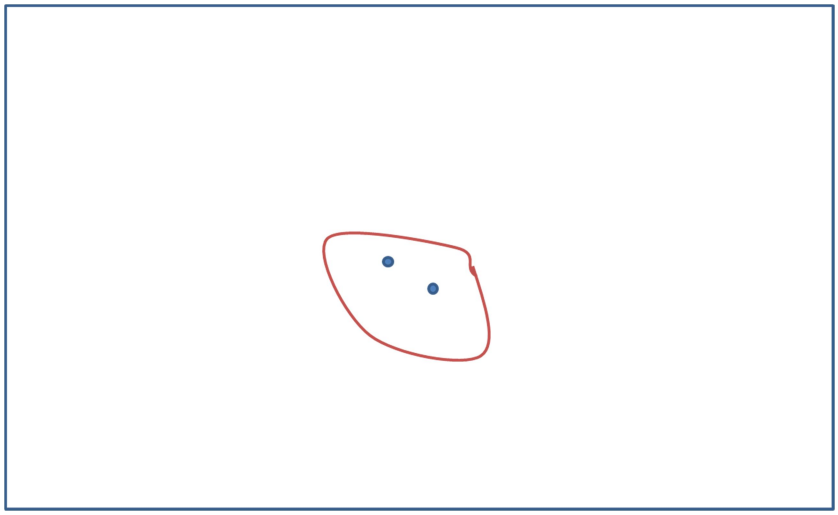


Figure 2. Two pins stuck in paper with surrounding loop.

Now we take a pencil and use it to stretch out the loop. We move drag the pencil around the pins always keeping the string taut.

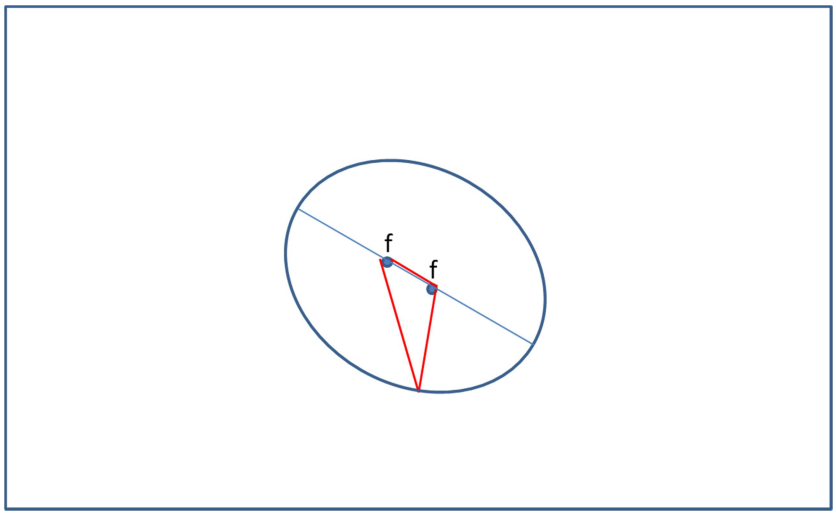


Figure 3. Ellipse drawn with pencil making the loop taut.

If you try this you might get something like the figure above. The red string loop has some total length. When we draw it taut with the pencil it forms a triangle. Two of the sides of the triangle are the distances from the foci to the current point we're drawing. The third is just the distance between the two foci. But that's a constant, so all the points on the curve have a sum of distances from the two foci that are the same. That sum is the length of the loop minus the distance between the two foci. Same total distance from two points: that was our definition of an ellipse.

If we draw an ellipse with the foci close together, we get a rounder ellipse. If the foci are coincident, we get a circle with a radius half the length of the loop. If the foci are nearly half the length of the loop apart, then we get a long narrow ellipse, since the loop is almost taut to begin with.

In the figure above we also drew the line between the foci and extended it out to the ellipse. That line is the *major axis* of the ellipse. The point midway between the two foci is the *center* of the ellipse. If we draw a line through the center perpendicular to the major axis, we get the *minor axis* of the ellipse.

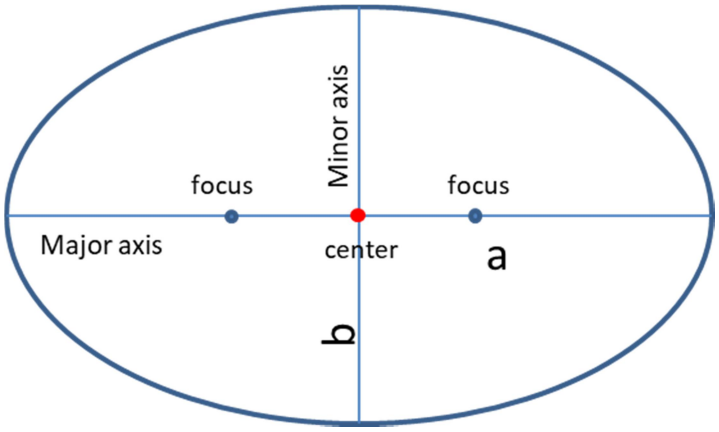


Figure 4. Labelled ellipse.

With circles, most formulae are written more conveniently using the radius not the diameter. Similarly for an ellipse we usually use the distance from the center to the edge of the ellipse along the major axis. Since this is half the major axis, it's called the *semimajor* axis. It's often designated by an **a** in formulae. The *semiminor* axis is also used sometimes: it's often designated by **b**. The length of the semimajor axis is the primary parameter we use to specify the size of the ellipse.

Using a rectangle to generate an ellipse

If we don't have a loop and pins handy we can draw an ellipse free-hand starting with any rectangle. Draw a dot at the center of each of the four sides of the rectangle and draw a smooth symmetric curve between them. It's not too hard to get a reasonable looking ellipse this way. We can think of the rectangle as generating the ellipse. The center of the rectangle is the center of the ellipse, the lengths of the sides are simply the lengths of the major and minor axes, and the orientation of the rectangle gives the orientation of the ellipse. Except for squares, whose orientation doesn't affect the associated circle, there's a one-to-one mapping between the generating rectangles and the associated ellipse.

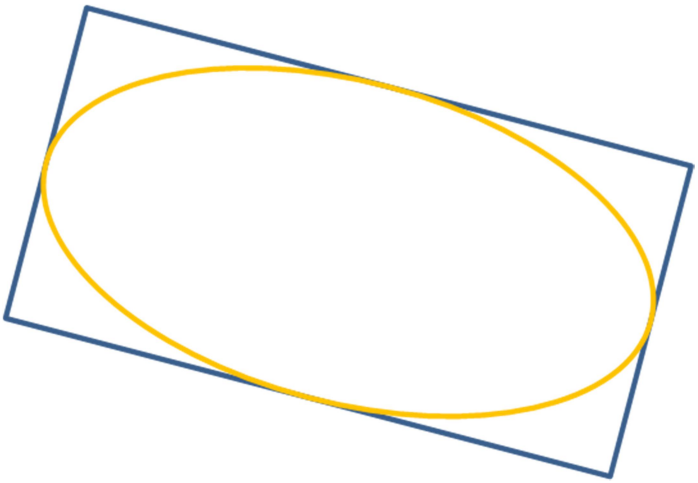


Figure 5. Ellipse and corresponding rectangle

Most drawing programs make ellipses easily, but when we deal with orbits that use hyperbolas, which aren't as commonly provided as standard figures by drawing programs, the generating rectangles can be particularly useful as a shape guide.

Eccentricity

We keep on talking about more circular or more elliptical/extended ellipses and we need some way to characterize that. We take the distance from the center red dot to one of the focus blue dots and compare it to the full semimajor axis. We'll always get a ratio that's less than 1 (since the foci are inside the ellipse). If the foci are coincident with one other, we get a circle and that ratio is 0, with the center and both foci at the same point. If the ellipse is very extended we get a ratio of just under 1. This ratio is called the *eccentricity.* It's usually designated by a little ***e*** in formulae. It's the parameter that characterizes the shape of the ellipse.

Closed orbits.

Let's get back to Kepler and talk about his first law in a bit more detail. It says that the planets orbit the Sun in closed, elliptical orbits where the Sun is at one of the foci of the orbit. If we were trying to guess what the motion of planets in the Solar System was without any preconceptions, we might not assume that the orbits would be closed. The discovery that the orbits are closed is perhaps even more important than the elliptical shape of the orbit. There is no requirement generally that orbits be closed. The orbits of satellites around the Earth aren't – they are subject to a change in some of the characteristics of the orbit that we call precession and talk about a bit below. The orbit of the Sun around the Galaxy is not closed. It's a complex three-dimensional curve. The Sun's oscillation above and below the plane of the Galaxy has a different period than its revolution around the Galactic center. Orbits can look like spirograph figures or be complex three dimensional shapes.

Life is much simpler than it might have been. Each year the Earth retraces more or less the same path around the Sun, and the other planets do the same in their own periods. The orthodoxy at the time of Kepler was that we had lots of interacting circles, predisposing the community to closed orbits, and they were pretty clear that the Sun (for the non-Copernicans) had a closed orbit since it came back to the same location every year. [Actually even the ancients knew this wasn't quite true, but the change from year to year was measured to be very, very small.] Kepler's discovery that planetary orbits were closed may not have been appreciated quite as much as it deserved.

Newton's laws of gravity and his invention of calculus, let us know that we can attribute the simplicity of the orbits in the Solar System to the overwhelming concentration of mass in a single relatively small body at the center. The Sun has 99.8% of the mass of the Solar System. While it seems peculiar to call the Sun small -- it's a million miles in diameter! – it's much smaller than the radius of the orbit of the nearest planet. So we can pretend that all of its mass is concentrated in a point at its center. And to first order we can ignore the mass of everything else.

The physical parameters of an orbit.

While we've noted that the semimajor axis and the eccentricity measure the size and shape of the orbit, it may be useful to make a short diversion to note that these are very closely related to two basic physical properties of the orbit: its energy and angular momentum.

Elliptical orbits are *bound*: the planet cannot escape from the Sun. If we wanted to pull a planet away from the Sun, we'd need to supply a lot of energy. Since we'd need to supply energy to get the components apart, the total energy of a planetary orbit is negative. Clearly there is a lot of positive kinetic energy in the motion of the planet, but gravity binding the planet to the Sun creates a negative gravitational potential energy that is even bigger. Averaged over the orbit the gravitational potential energy is twice the size of the kinetic energy.

The energy of an orbit around the Sun is entirely determined by the semimajor axis, regardless of the eccentricity. A circular orbit or a long skinny orbit with the same semimajor axis have equal energies. We would need to add the same amount of energy to a body on either of these orbits to escape from the solar system. However, it turns out that this doesn't mean that we need the same size rocket to escape from these orbits: it's easier to escape from the long skinny orbit, but that's looking ahead to our basic astronautics.

The eccentricity is a measure of the angular momentum of the orbit compared to maximum angular momentum it could have for the given energy. However before we discuss this, let's try to understand what angular momentum is. We may have heard about angular momentum and spinning ice skaters, but what does that have to do with planetary orbits.

Very loosely we can consider *angular momentum* to be the 'total torque' of a system. It's one of the most basic concepts in physics from quantum theory to cosmology. Roughly it's a product of the velocity and the size of a system. So when a skater, spinning with arms extended, pulls the arms in, the lengths in the system get smaller. Angular momentum, the product of speed and length, is conserved,

so to compensate for the decrease in lengths, the speed the skater is spinning at increases. [When the skater slows or stops, the Earth absorbs the rotation, just as the Earth absorbs the momentum of a braking car. Since the Earth is so massive, the angular momentum of the skater like the momentum of the car seems to disappear –but it does not!]

The angular momentum of an orbit is the product of the radius of the object in the orbit, and the velocity of the object perpendicular to the radius. More precisely this is the specific angular momentum, we need to multiply by the total mass of the planet to get the total angular momentum. In a circular orbit the motion of the particle is always exactly perpendicular to the radius. Circular orbits have the maximum angular momentum for a given energy. For a long skinny orbit, the motion can have a direction more parallel to the line from the Sun, so not all of the velocity will count towards the angular momentum.

If we start with a circular orbit and just decrease the velocity, we'll get somewhat more elliptical orbit, but the total energy of the orbit will also decrease (i.e., become more negative, the orbit will get more bound) – since we decreased the kinetic energy of the body. However if we simultaneously move a little further from the Sun so that the gravitational potential energy decreases to compensate for the loss of kinetic energy, we can get a sequence of orbits of constant energy but decreasing angular momentum.

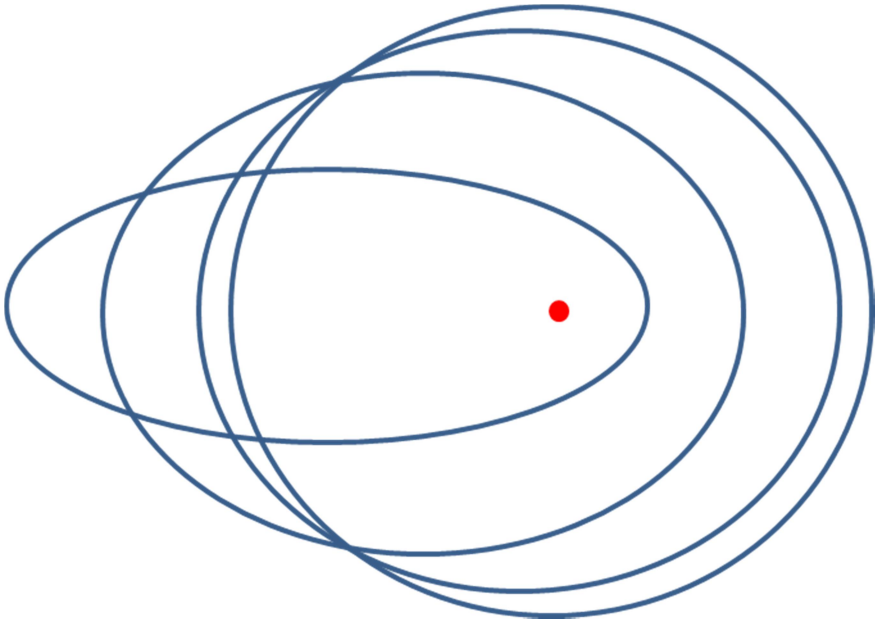


Figure 6. Orbits with the same energy but different angular momenta

In these orbits, the furthest distance from the Sun, the *aphelion* (or *apogee* if we're talking about orbits around the Earth or *apoapsis* for orbits generally) gets larger as the angular momentum gets less, but the semimajor axis and energy of the orbit remain constant. As the aphelion gets larger, the *perihelion*, the closest point to the Sun (*perigee* for Earth orbits, and *periapsis* generally) gets smaller. The sum of perihelion and aphelion distances is just the major axis and that's constant for the sequence or orbits.

An eccentricity 0 orbit has the maximum angular momentum it can have for its energy. As the eccentricity approaches 1 the angular momentum goes to 0.

Coordinate Systems

So far we've only discussed ellipses in the plane. If we only had a single orbit to worry about, then that might be enough, but the orbits of different planets are not in quite the same plane – though with the exception of Pluto they are pretty close. The orbits of comets and asteroids can be in very different orientations. So we need to be able to specify the orientation of orbits and positions in space generally. Before we can worry about orbits, we have to have some framework for specifying coordinates and orientations in space.

Let's start by reviewing the coordinates we use on the Earth.

On Earth, we can specify positions using latitude, longitude and altitude. The latitude and longitude are angles. Latitude ranges from -90 to +90, while longitude ranges from -180 to +180. The equator, the plane of the Earth's rotation is at 0 latitude. It defines a *reference plane*. The South and North poles are at -90 and +90 respectively in latitude. For any other point on the Earth we can measure the angle between the equator and the point (in the direction of the pole) to get the latitude. Latitudes are pretty well defined.

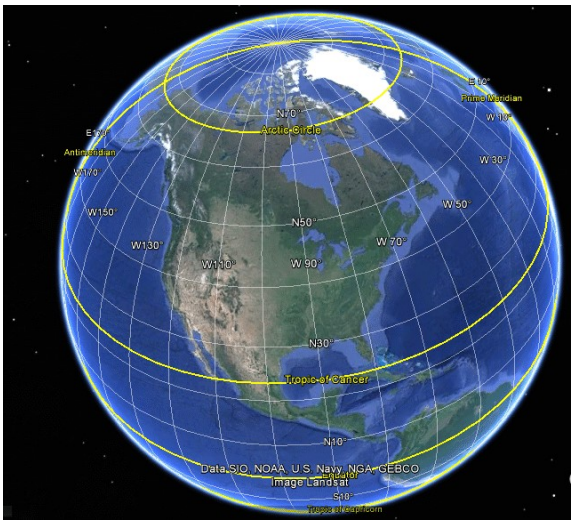


Figure 7. Coordinates on the Earth. (https://eclipse2017.nasa.gov/working-geographic-coordinates)

In principle, to measure latitude all one needs to do is measure how high in the sky the Sun is at its highest point in a day (around noon). Consult a small table that gives the latitude of the Sun on each day of the year, and a simple calculation gives our latitude.. E.g., if we measure the maximum altitude of the Sun as being at 70° above the horizon, on a day when the Sun has a known latitude of +10°, then we must be somewhere at 30° North latitude, or 10° South. The difference in the measured height of the Sun at noon from 90° is the difference between our latitude and the Sun's. Hopefully we aren't so lost that we can't tell if we are in the Northern or Southern hemispheres!

Longitude is a bit more arbitrary. Basically we just pick a point – not one of the poles -- and say that it has 0 longitude. We can draw lines due from that point due north and south to the poles to find all of the other points at 0 longitude. Then we measure how far we'd need to rotate the line we just drew to some other location to get its longitude. Since England was the premier power at the time this all got decided, the 0 of longitude was defined to be a spot in the Royal Observatory at Greenwich near London. I've trod the line at the Paris Observatory that the French were promoting as their candidate for the 0 of longitude. Maybe if Waterloo had turned out differently….

To determine longitude accurately mariners needed accurate time pieces. If we know what time it is in Greenwich when we have noon locally, then we can find the local longitude: our longitude is just the 15 times the difference in hours between our local noon time and the noon at Greenwich (ok, there's a correction associated with the equation of time and that weird 8-shaped analemma we sometimes see on globes, but we'll ignore that for this discussion. It's just a simple table lookup anyway.).

If we're working in three dimensions we need to measure the altitude of an object. Normally we measure this as some distance above the standard surface of the Earth. We could equivalently measure the distance the object is from the center of the Earth. That's what we'll do in space. Changing from altitude to radius from the center of the Earth complex because the Earth's reference surface is not a sphere, sealevel at the poles is a little closer to the center of the Earth than sealevel on the equator. Fortunately we don't run into this complication much for orbits where we measure our distances from the central point of our system.

There are two coordinate systems that we use across the entire Solar System. Usually we put the Sun at (or very close) to the center of the coordinate system. The systems are determined by the reference plane, where the latitude is 0 in the coordinate system.

Equatorial Coordinates

We can use a reference plane parallel to the Earth's plane of rotation, i.e., the plane that generates the equator. These are called Equatorial coordinates.

If we look at the sky during the night, we find the stars move in circles around the North Pole (or the South Pole if we live south of the equator). Each star moves in a circle of constant latitude. On the sky we call this *declination*. The photo below is a long exposure where the stars trail in circular arcs around the pole.

Figure 8. Star trails in the Southern Hemisphere (https://upload.wikimedia.org/wikipedia/commons/thumb/e/e5/All_In_A_Spin_Star_trail.jpg/1920px-All_In_A_Spin_Star_trail.jpg)

Astronomers built transit telescopes, where the telescope can be raised or lowered in only one direction, such that it always points due South (or North), and can move only along the line of constant longitude.

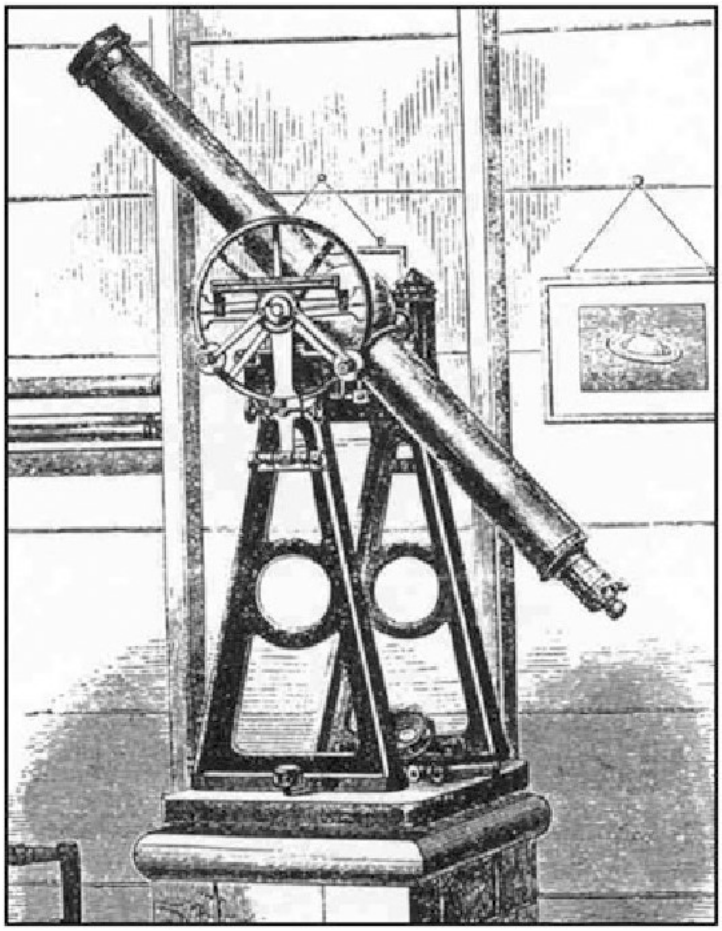


Figure 9. A transit telescope (https://www.researchgate.net/profile/Wayne-Orchiston-2/publication/310396589/figure/fig2/AS:591767783092224@1518099842381/The-transit-telescope-at-Uckfi-eld-Observatory-with-a-stand-similar-to-the-one-Tornaghi.png)

The figure above shows an example of such a transit telescope. We can use it to look at a particular star. If we leave it looking in the same location, it will see stars of the same latitude as the night passes. It's how much we raise or lower the telescope that determines the latitude we look at. That's why we call latitude 'declination', in this system. If we don't move the telescope, after just under a day we'll see the same star pass through again. This period between when we see the star pass through the telescope is called the *sidereal day* and is just a little bit shorter than the regular day. In a year there is exactly one extra sidereal day. The idea is that as the Earth moves on its orbit, the Earth needs to rotate just a little bit more than 360° to get the Sun back to the same apparent location. Since there are a fraction more than 365 days in a year, the Earth rotates by a hair under 361° every solar day with respect to the stars, with the extra degree compensating for the revolution of the Earth around the Sun. The sidereal day is the period that the Earth rotates exactly 360° with respect to the stars.

If we move the transit telescope up and down, then we see stars of the different declinations, but if we do this quickly they have the same longitude. If we have a clock that measures time in sidereal days (it runs about four minutes/day faster than a regular clock) then the sidereal time at which we see the star going through the transit telescope is a measure of its longitude. Since this is how we originally

measured the longitude of stars, longitudes in Equatorial coordinates are generally given in units of hour, minutes and seconds, with 24 hours being the same as 360° (1 hour = 15°). The longitude in Equatorial coordinates is called the *right ascension*, often abbreviated RA. Declinations are measured in 'regular' degrees however. (Sidereal hours, minutes and seconds were slightly shorter than their standard counterparts.)

Since both degrees and hours are sub-divided into differently sized minutes and seconds, there's a big opportunity for confusion. Generally we see suffixes *m* and *s* for the divisions of hours and ' and " (single and double quotes) for minutes and seconds of arc. But it pays to be careful and explicit.

Ecliptic Coordinates

The second coordinate system we use a lot in the Solar System are *Ecliptic Coordinates.* Instead of using the Equatorial Plane, the plane associated with the Earth's rotation, as the plane of 0 latitude, we use the *Ecliptic plane,* the plane of the Earth's orbit around the Sun. This is inclined by about 23.5° from the equatorial plane. We see this angle on the globe as the latitude of the Tropics of Cancer and Capricorn. The Arctic and Antarctic circles are at 90° minus this angle.

The pole of the Ecliptic Coordinates is the same 23.5° from the Equatorial pole. There is no especially bright star near the Ecliptic pole the way Polaris marks the North Equatorial pole. If we were able to see the Ecliptic pole during the day, we'd note that regardless of the time of day or year, the Sun was always exactly 90° away from it. One nice thing about ecliptic coordinates is that angles are generally measured only in degrees, and the names of the coordinates are just latitude and longitude. For once astronomers forgot to put on their obfuscation hat.

Zeroes of Longitude

So far we haven't actually said what the 0 points are for the longitudes of either of the two coordinate systems. The reference planes for the Equatorial and Ecliptic systems are not parallel, so simple geometry says that they must intersect along a line. It we extend this line to infinity, we get two points in the sky. By definition, the Earth is always at Ecliptic latitude 0, but when the Earth passes through the line where the two planes intersect, we can look back at the Sun from the Earth, and the Sun will appear from the Earth to be at exactly 0 Equatorial latitude. At one of these two points the Sun is crossing from south of the equator to north. At the other it is (apparently) moving in the opposite direction. The first happens at the time of the Vernal Equinox, the start of spring in the Northern Hemisphere. The second is at Autumnal Equinox. The point in the sky that corresponds where we see the Sun at the Vernal Equinox is called the First Point in Aries. [But it's actually now in the constellation Pisces!] This point is used as the 0 point for longitude in both the Equatorial and Ecliptic systems. It's at coordinates (0°,0°) in both equatorial and ecliptic coordinates. The only other point in the sky which is the same in both systems, is the point diametrically opposite at (180°,0°), or in equatorial coordinates (12h RA, 0° declination).

Heliocentric and Geocentric Coordinates

The equatorial or ecliptic coordinates define directions, but we're going to be working in three dimensions, so we are likely to need to define where the center of the coordinate system is. Normally we define the center as the object that we are orbiting. When we are talking about planetary orbits around the Sun, we have the Sun at the center of our coordinate system, the coordinates are *heliocentric*. When we talk about satellite orbits around the Earth we use *geocentric* coordinates. Often we don't use the exact positions of the Sun or Earth but the average positions of masses in the Solar system, or the average of the Earth and Moon weighted by their mass. These are Solar or Earth *barycentric* systems.

In planetary system animations and animations of the large moons, a reference system using the equatorial plane of the planet rather than is used.

Precession of the Earth and ICRS coordinates

While we've assumed that the Earth's rotation is stable, that's not true in detail. In particular, the Earth precesses like a spinning top with the axis of rotation wobbling very slowly. It takes 25,000 years to complete a full cycle. This changes the plane of the Earth's rotation so that the points where the rotation and revolution planes intersect also moves slowly in a circle. The pole of the Earth's rotation makes a circle in the sky with a radius of 23.5°. So our 0 of longitude changes slowly and even the equatorial latitudes of the stars change. When the Sun and Earth were astronomers' main reference points, it made sense to accommodate this. Astronomers updated their coordinate systems every 25 or 50 years or so and had to be careful comparing data taken at different times. Astronomers often use the word *epoch* for the time of something. The epoch of an observation is when the observation was taken. But traditional coordinate systems had epochs too, they were valid at some particular time. If you needed the coordinates in some other epoch there was a detailed process for converting.

Today astronomers know the position of many very distant objects very accurately to use as reference points independent of the apparent rotation of the Earth or position of the Sun. These distant objects are essentially stationary in the sky (the objects are moving, often very quickly indeed, but they are so far away that the apparent motion is infinitesimal). Using these astronomers have generally agreed to use a fixed set of coordinates. These very closely match which the characteristics of the Earth's rotation and orbit at the beginning of the year 2000. These are the *International Coordinate Reference System* or ICRS coordinates. When astronomers need to point a telescope at some object in the sky, they still need to worry about precession and all of these other details, but computers take care of it. Since objects do move in the sky, we still need to know the epoch of observations, but at least we no longer have to worry about the epoch of the coordinates as well.

Galactic Coordinates

Another coordinate system that is commonly used is Galactic coordinates. Our Galaxy is composed of a combination of a roundish bulge and a much thinner disk. The position of the Sun is close to, but not quite at the center of the disk. Galactic coordinates use a reference plane which goes through the Sun but aligned with the Galactic disk. The zero point of longitude was set to the best estimate of the center of the Galaxy. The supermassive black hole which would now likely have been set to be the center, is

just a few arcminutes away from this zero point which was set before knowledge of the black hole's exact position.

Orienting the ellipse

It's taken a while, but we now we understand how we can specify a position in space. We specify a distance and direction using Equatorial or Ecliptic coordinates with reference to some specified center.

Let's get back to understanding how we can describe an orbit. We know how to specify the size and shape of the ellipse, so we draw our ellipse on a piece of paper with the major axis along the x-axis. But how is it oriented in three-dimensional space.

Argument of Perihelion

First we have one degree of freedom to worry about in the plane.

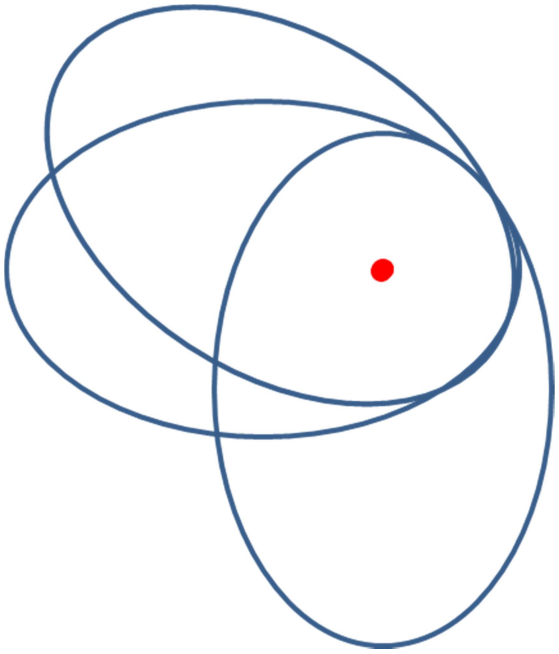


Figure 10. Orbits with the same energy and angular momentum

We might have three orbits as in the figure above, which have exactly the same semimajor axis/energy and shape/angular momentum. These figures are rotated around the Sun (or whatever they are orbiting), by different values. We measure the angle between the major axis of the ellipse and the x-axis in the orbit plane.

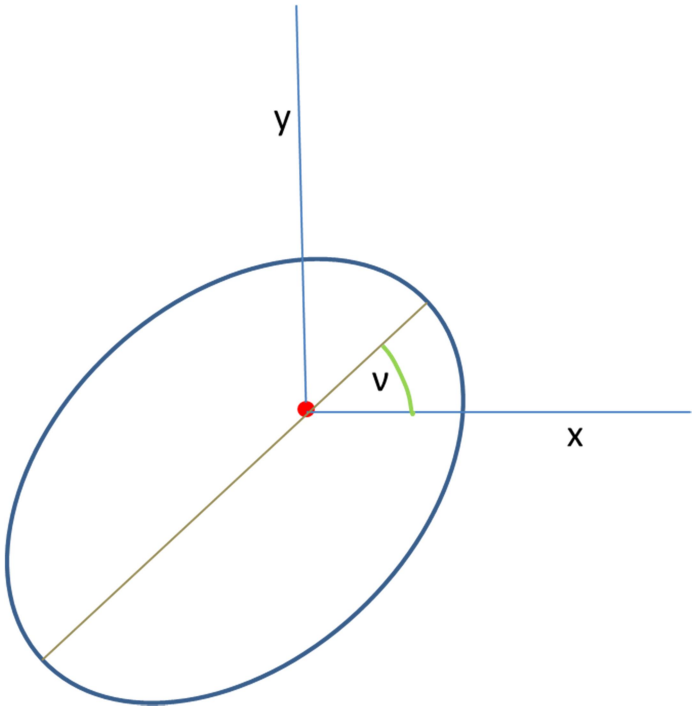


Figure 11. The argument of perihelion: all quantities are in the xy plane.

This angle is called the argument of perihelion (or perigee or periapsis), and is sometimes designated by a little Greek nu, **ν**.

We've oriented the ellipse in the plane, but how do we orient this plane in three dimensional space? Above we talked about the angular momentum of the orbit and how it determined the shape of the orbit. It also specifies the orientation of the plane. Recall that angular momentum is a kind of product of the position and velocity. Positions and velocities are vectors: they have both a magnitude and a direction. Two objects can be in the same direction but one is farther than the other, or at the same distance but in entirely different directions. Everyone running in the 100 meter sprint is running in the same direction, but some are faster than others. Cars on a highway may be all travelling at the same speed (ok that's not too likely!), but the North bound traffic is going in an entirely different direction than the South bound lanes. To specify the position/velocity we provide a distance/speed and a direction.

Vectors and Cross-products

There are different ways of multiplying vectors. The way the angular momentum is constructed the product is another vector with a magnitude that depends upon the magnitudes of the input vectors and how perpendicular these vectors are to one another. If the velocity is in the same direction as the position, then the angular momentum will be zero. We use what is called the cross-product of two vectors.

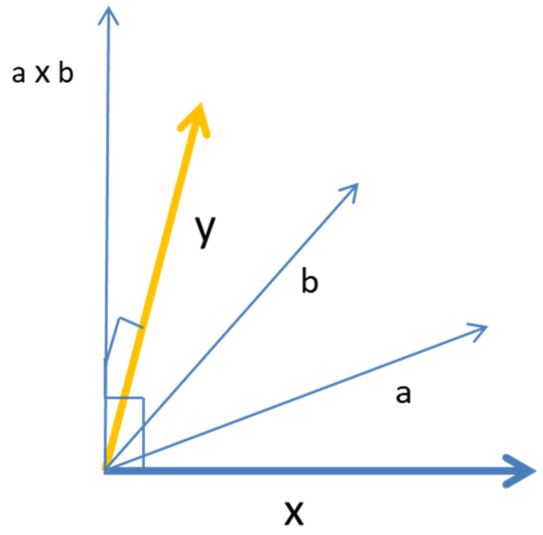


Figure 12. Cross product of two vectors in the x-y plane is a vector in the z direction.

The direction of the angular momentum is perpendicular to **both** of the input vectors, so the angular momentum of an object is a vector that's perpendicular to both the position and velocity. The position and velocity vectors are both in the plane of the orbit. If the angular momentum is perpendicular to both of them, it must be perpendicular to the orbit plane. So if we draw an orbit on a page, the angular momentum vector points perpendicular to the page. Conversely, the orbit plane is perpendicular to the angular momentum! If we know the angular momentum – and it's easily computed if we know the position and velocity – then we know the orbit plane, by looking for the plane perpendicular to it.

If we know the plane of the orbit, then we can find the line where the orbit plane intersects with our reference plane (the ecliptic or equatorial plane).

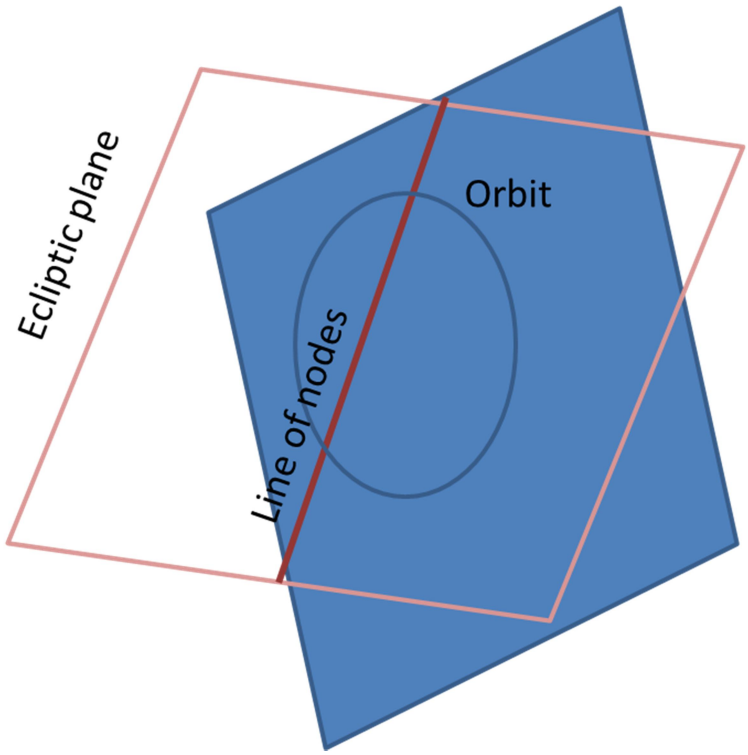


Figure 13. The intersection of the orbit with the reference plane

The intersection line where the two planes intersect is called the line of nodes. We're now ready to define the last two parameters that orient the orbit.

Inclination.

The first is just the angle between the two planes. This is called the *inclination* and is usually written as a little **i**. In the equatorial coordinate system, the Earth's orbit has an inclination of about 23.5° since that's the angle between the Earth's orbit and the reference coordinate frame. However if we measure in the ecliptic coordinates, most planets have relatively small inclinations and for the Earth the inclination is 0 by definition (since the reference plane is the plane of the Earth's orbit). Pluto's is more significant and many asteroids and comets have significant inclinations.

Prograde and Retrograde Orbits.

By convention in mathematics angles increase as we move counterclockwise in a circle. Conveniently if we view the Solar System from above the Earth's North Pole, most planets move and rotate counterclockwise. Such motion is called *prograde.* However some things move or rotate clockwise and this kind of motion is designated *retrograde*. Notably Halley's Comet and the largest moon of Neptune have retrograde orbits. To indicate that an orbit is retrograde we use the inclination.

An orbit with an inclination of 0 (e.g., the Earth in ecliptic coordinates) is moving in exactly the reference plane in a prograde direction). If we add a small inclination, the orbit is still prograde, but planet moves above and below the orbit plane during the orbit. An orbit that had an inclination of 90° would be perpendicular to the reference plane, it would be neither prograde or retrograde. If the inclination is greater than 90° then we're beginning to turn the orbit over and the motion is retrograde. An orbit with an inclination of 180° is back exactly in the reference plane, but now revolving in the retrograde direction.

The longitude of the ascending node.

The last angle that we use to specify the orientation of the orbit is called the *longitude of the ascending node.* If we look at the line where the orbit plane crosses the reference plane, the *line of nodes,* this crosses the actual orbit at two points. At one of these the planet is moving from below the reference plane to above (where North is 'above'), and vice versa at the other. These two points are called the *ascending* and *descending nodes*, respectively, and the line of nodes connects them. The vernal and autumnal equinoxes are the nodes from the Earth's orbit around the Sun – or historically the Sun's imputed orbit around the Earth.

By definition the latitude of the nodes is 0, they are on the reference plane, but they have some defined longitude. We use the longitude of the ascending node to fix the orbit in place. This quantity is sometimes designated with a capital Greek Omega, Ω. By construction the longitude of the descending node is just 180° offset from the ascending node, so there's no need to keep track of it separately.

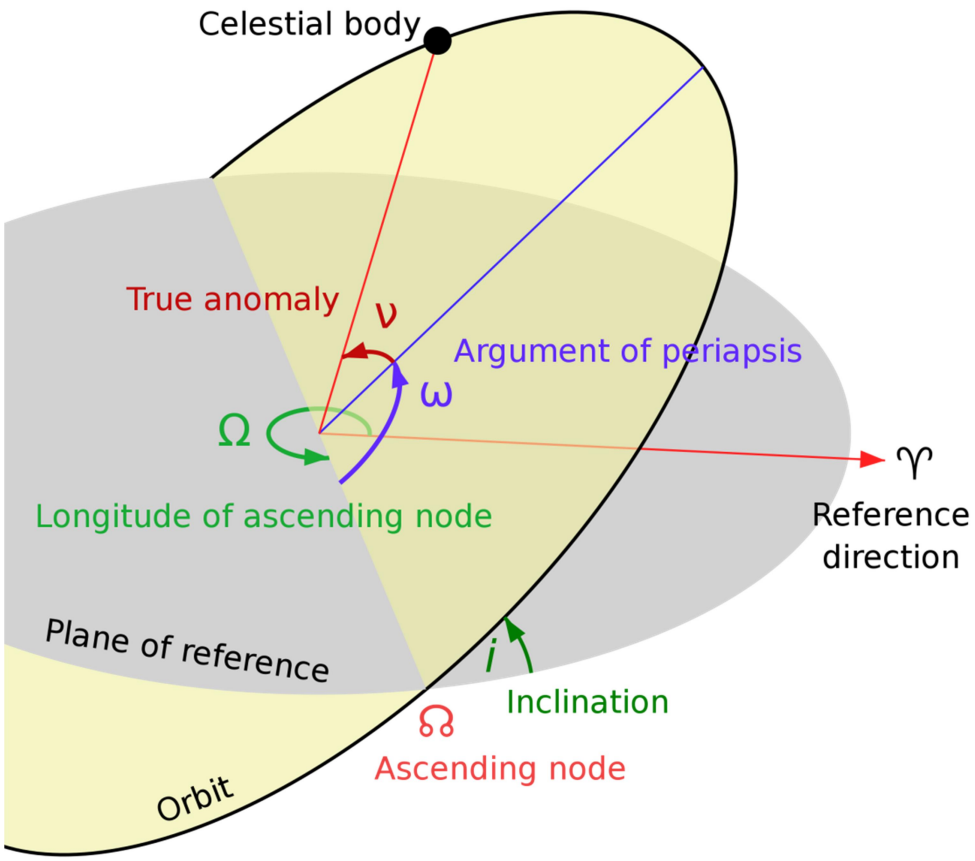


Figure 14. Angles that orient the orbit (source Wikipedia).

The lovely figure above from Wikipedia illustrates all three of these angles we've discussed above. Here the reference direction is the direction to the 0,0 coordinates in the sky. We rotate around till be get to the vector to the ascending node, then we tilt by the inclination, and do a second rotation in the orbit plane to line up with the major axis of the ellipse. This figure notes one more angle, the true anomaly that we'll get to shortly.

Phase of the orbit

At this point we have enough information to draw the orbit in space: we know the size and shape of the orbit from the semimajor axis and eccentricity, and the orientation of the ellipse from the three angles we've just discussed. But we still don't have any information about the phase of the orbit. Where on the orbit is the planet at some particular time? We need one or two more numbers to tie down the orbit.

Often we are given the last time the planet went through perihelion (or perhaps the next time it will). Or we may be given where the planet is along the orbit at some specified time, perhaps the current time, or the time some ephemeris was created. In both cases we have a time and a phase in the orbit, where the phase is implicit when we're being given the time of perihelion.

Anomalies

The phase of the orbit is usually expressed as an angle using one of three *anomalies*: true, mean or eccentric. Here astronomers use the word anomaly, to mean the offset from perihelion. We generally give either the time at which we had a perihelion passage, when all of these anomalies are 0, or we give the mean or true anomaly as at some reference time. With the phase specified, we can now say where the planet will be at any time.

True Anomaly

The *true anomaly* is the actual angle between the vector from the Sun and the planet, and the Sun to the perihelion point. So it's 0° at perihelion where it changes most rapidly. The rate of change slows till we reach aphelion at 180° where it changes slowest. Then it begins to speed up as we approach another perihelion. This is a consequence of Kepler's Second Law.

Kepler's Second Law is that an orbit sweeps out a constant area as a function of time.

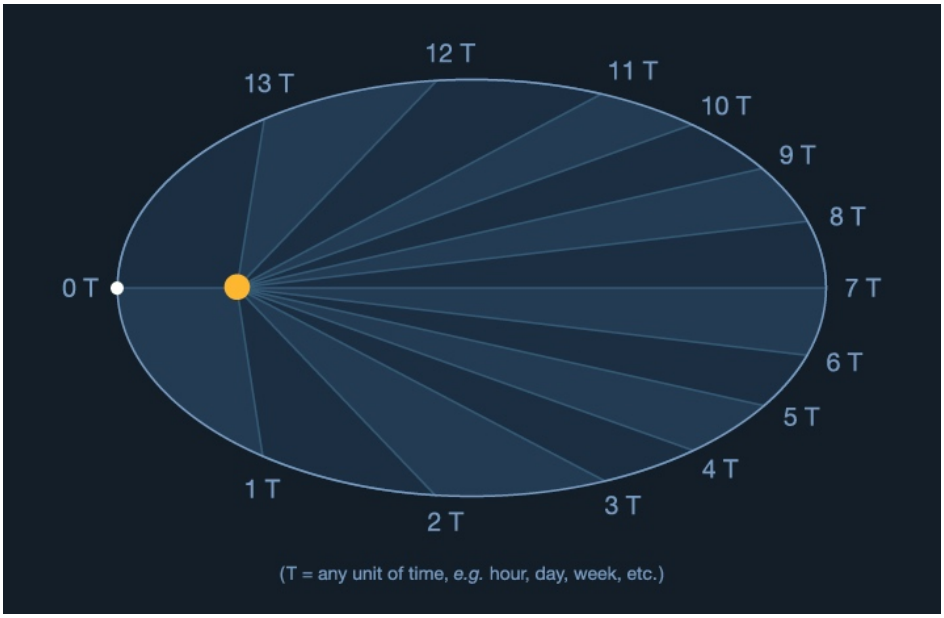


Figure 15. Kepler's Second Law. The area swept out per unit time is a constant. (From https://solarsystem.nasa.gov/basics/chapter3-3/)

If we think of a planet pulling along a taut (and if the orbit is elliptical, elastic) string connecting it with the Sun, then we can think of the string as sweeping out the orbit. The figure divides the area of the ellipse into 14 pie-like segments, which all have equal area. Kepler's Second Law says the interval spent along the edge of each of these segments must therefore be the same. So the true anomaly changes much more quickly at perihelion. The fat pie slices near the perihelion have much bigger interior angles than the long thin slices near aphelion.

There might be something familiar about all of this. The area swept as a function of time is something like the radius times the velocity – but if you think about it, only the velocity that's perpendicular to the

radius that counts. If we are moving parallel to the radius we don't sweep out any area. Radius times parallel velocity…. Kepler's Second Law says that the angular momentum of a planet in orbit is a constant.

Mean anomaly

The simple definition of the *mean anomaly* is to average the rate of change of the true anomaly to define a new abstract quantity which changes uniformly with time. So the mean anomaly will change slower than the true anomaly at perihelion, but faster at aphelion.

A more satisfying definition though, is simply to use Kepler's second law to define the mean anomaly. Then it's the fraction of the area of the ellipse that has been swept out since the last perihelion passage. So in the figure above, the mean anomaly at the point labeled 1T would be $1/14^{th}$ (or 360/14 degrees). At the point labeled 2T it would be $1/7^{th}$. At 7T, we've swept out half the circle, and both the mean and true anomalies are 180°.

Suppose we know the mean anomaly at some time: e.g., the mean anomaly is 10° on day 50, and the period of the orbit is 200 days. What's the mean anomaly on day 1000? Four full orbits after day 50 it is day 850 and we're back to mean anomaly of 10°. We need to add 150/200 or 3/4ths of a circle to the 10 degrees to get to day 1000. That's 270° plus the 10° we started at, so on day 1000, the mean anomaly is 280°. We can convert easily between time and mean anomaly.

Eccentric Anomaly

It turns out that the motion as a function of time of a planet in even a simple elliptical orbit cannot be expressed directly using time or the true or mean anomalies. What can be understood easily is an object in a circular orbit. So a third angle is used which uses a reference circle, not so much to define the phase of the orbit, but as means to convert between time (or equivalently mean anomaly) and true anomaly. This third anomaly is called the *eccentric anomaly*. We can think of generating the outer golden circle by taking the rectangle associated with the original ellipse and then expanding the shorter sides of the rectangle to make it a square, keeping the center of the rectangle/ellipse the same.

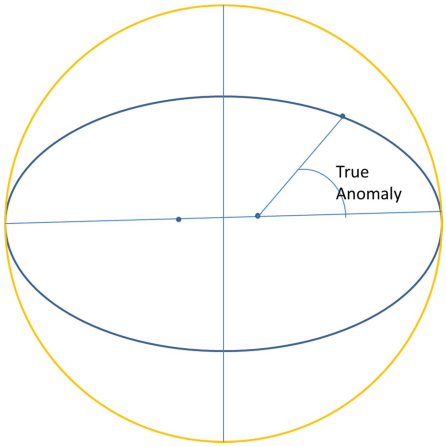


Figure 16. The true anomaly.

In the figure above we've drawn the true anomaly. But we've also embedded the ellipse in a circle that we'll use later on. The circle is tangent to the ellipse at the points along the major axis. The true anomaly is a simple angle here.

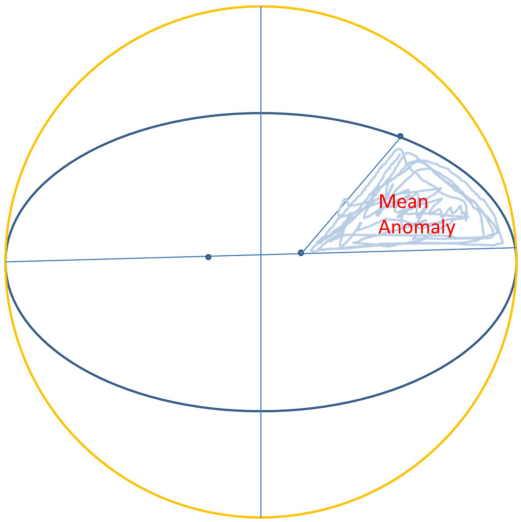


Figure 17. Areal interpretation of mean anomaly.

This figure shows the mean anomaly. It's the fractional area of the ellipse that we've swept out.

Now let's find the eccentric anomaly. We draw a line perpendicular to the major axis that goes through the point we're interested in (the one we have the mean and true anomalies for), and extend it out to a point on the enclosing circle. Then the eccentric anomaly is just the angle along the circle where we find that point.

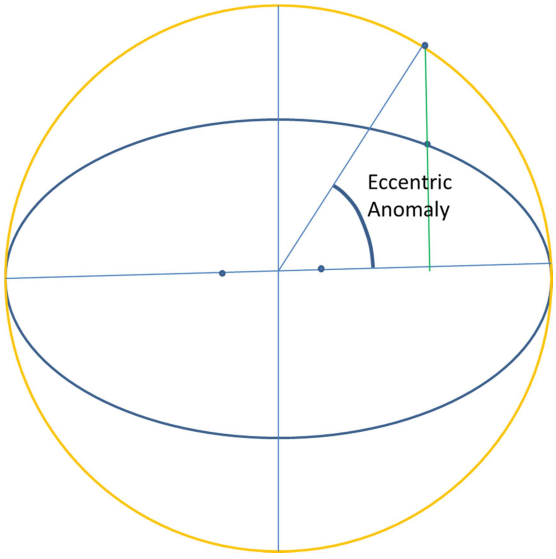


Figure 18. Angle interpretation of the eccentric anomaly.

However for a circle we the fractional area swept out is proportional to the angle. We could just as easily interpret the eccentric anomaly as the fractional area of the circle that we've swept out as we moved to the point on the circle.

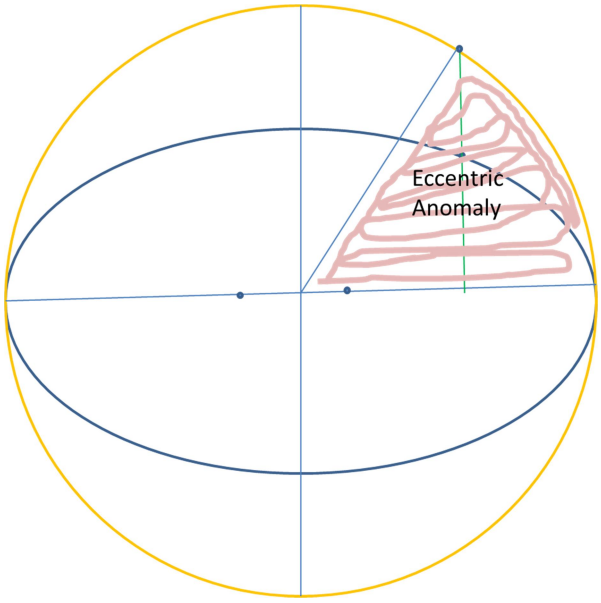


Figure 19. Areal interpretation of the eccentric anomaly.

The eccentric anomaly bridges the areal view of the mean anomaly and angular view of the true anomaly since it can be treated equally both ways.

In practice, it is easy to go either direction from the true anomaly to the eccentric anomaly, and it's easy to go from the eccentric anomaly to the mean anomaly. Going from mean anomaly to eccentric anomaly is a little more complex: typically we zero in on a value after making a few initial guesses.

So if we have a time and we want to know where on the orbit we are we

- Start with some initial reference time and the mean anomaly at that time,
- get the change in time since the reference time to calculate the mean anomaly at the time we're interested in,
- find the eccentric anomaly from the mean anomaly
- and the true anomaly from the eccentric anomaly.
- With the true anomaly and a little bit of geometry we can get the position and velocity of the planet.

Other parameters

We now have defined parameters sufficient to describe orbits and calculate the position of a body at any time, but there are a fair number of other parameters that are often used.

Period and Mean Motion

We've alluded to the period of the orbit. If an orbit is closed the *period* is just how long it takes before we return to the same point. For elliptical orbits, the period is determined only by the energy or equivalently the semimajor axis. This may be a little surprising, but maybe less when we think about the idea that the narrower the ellipse is, the less angular momentum it has and the less area. Since the angular momentum is the rate at which we're sweeping out area, it may not be too surprising that the area of the ellipse and the rate at which we're sweeping out that area both change such that the period remains the same.

Instead of the period, ephemerides sometimes give its inverse, the frequency of the orbit, or the rate at which the mean anomaly is changing. The units of the *mean motion* might be the number of orbits per day completed (common for Earth satellites), or the number of degrees the mean anomaly changes per second. Despite the very different units both of these are measures of orbital frequency and equivalent to specifying the period of the orbit.

Mean longitude and longitude of perihelion

Many planetary and natural satellite orbits have relatively small inclinations and eccentricities. If the inclination of the orbit is near 0 (or 180 for retrograde orbits), then the measured value for the longitude of the ascending node may be unstable. Similarly if the orbit is very nearly circular then the argument of perihelion may be hard to determine. This isn't saying that the orbit is ill-defined. It's like trying to measure the longitude near the North or South Poles. Very small changes in real position are dramatically magnified because the coordinates are a bit wonky at that location. But it doesn't mean that we can't stand near – or for that matter at -- the pole.

The *mean longitude* is defined as the sum of the longitude of the ascending node, the argument of perihelion and the mean anomaly. This is kind of weird. The first angle is measured along the reference plane (e.g., the ecliptic), while the next two are measured in the orbit plane. However if the inclination is small then this may still give a sense of the actual longitude of the object. Changes in one angle due to

the instability in the measurement may absorbed in one of the others. The longitude of perihelion is similar, it's just the sum of the longitude of the ascending node and the argument of perihelion.

Special cases

We've alluded to problems in defining some of the angles if the orbit is exactly circular or in the reference plane. Generally we just pick 0 for the longitude of the ascending node or argument of perihelion when the choice doesn't matter.

Radial orbits

If an orbit has no angular momentum what happens? We've taken a limiting approach which would seem to indicate that as we removed angular moment we get closer and closer to a line that drops towards to the Sun and then instantly turns around and comes back to the starting point. We should be careful when thinking about this. In such an orbit the object would be subject to infinite forces and mathematicians are a little chary of talking about infinities. What they like to do is say that if no matter how you get to the infinity, the results as you get nearer and nearer all converge, then maybe there is some real about your understanding of what's happening there.

Let's consider another way to get to our limit. Suppose we start with our object in a radial orbit, but we have the Sun be a somewhat diffuse ball which the object can pass through frictionlessly. We get to our limit by gradually concentrating the mass of the Sun – so we have the same limit as before. Now the object falls towards the Sun, but it passes through and comes out the other side. No matter how much we concentrate the Sun – so long as the Sun retains some finite extent – we get the same orbit. Which is different from the limiting orbit we considered above!

The lesson is that we probably can't say much about radial orbits where we treat the Sun as a point mass. They are at best a mathematical construct. In the real Solar System, if we have a radial orbit, then the object is going to suffer a terminal encounter with the Sun so prospects for a long term orbit are minimal anyway!

Hyperbolic orbits

So far we've only discussed bound orbits. We only rarely see natural objects on unbound orbits. There have been a few comets with very slightly unbound orbits. These had probably been very loosely bound in the Oort cloud around our Solar System and were given a slight kick by a passing star or object that pushed them into the inner Solar System with enough velocity that they would escape on the way out. In the past few years two objects that are significantly unbound and probably originated from outside the Solar System entirely have been seen passing through. The object 1I/`Oumuamua (the initial quote is intentional) raced into the Solar System and passed inside Mercury's orbit in September of 2017. It's now (September of 2021) approaching the orbit Uranus on its way out of the Solar System. The object 2I/Borisov made its closest approach to the Sun on in 2019. With improving technology and vast new surveys coming on line, we are likely to see many more in coming years.

So Kepler didn't really have much occasion to consider unbound orbits, but they have become very important for interplanetary missions. Unbound encounters with planets are used to change the orbits of satellites and get to places that would be impossible otherwise. Currently the Parker Solar probe and the BepiColombo mission to Mercury and the Lucy mission to the Trojan asteroids are using a series of encounters with the inner planets to accomplish their missions. Encounters with Jupiter were critical to fling the New Horizons spacecraft to Pluto and to put the Ulysses satellite into a high inclination orbit where we can see the poles of the Sun.

Just as the ellipse is the basic figure for a bound orbit, the hyperbola is the shape we find in unbound orbits.

Hyperbolas

What's a hyperbola? You may recall that we had a definition of an ellipse as being the set of points where the sum of the distances to two foci was the same. The definition of an hyperbola just changes one sign. Rather than looking to see where the sum of distances from two points is the same, we need to look to see where the difference in distances from two points is the same.

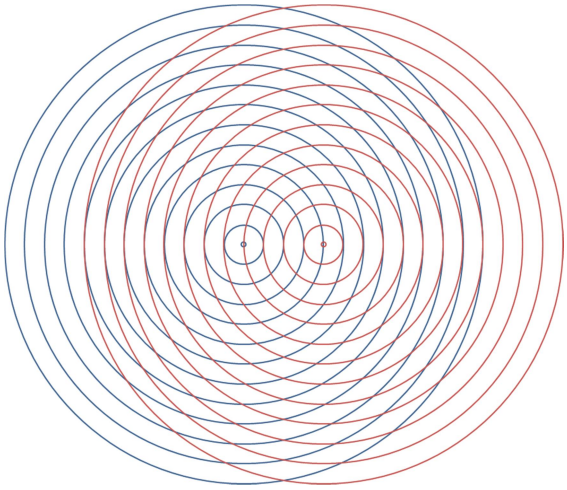


Figure 20. Rings around the two foci of a hyperbola.

In the figure above we've picked two points to be the foci of the hyperbola and drawn a set of equally spaced rings around both. The foci are about 4 units apart. Let's draw the hyperbola where the difference in distance from the two foci is 2 units. So if we go three units right from the blue focus, and 1 unit left from the red focus we find one point on the hyperbola. That's where the third blue ring is tangent to the first/innermost red ring. Let's go out to second red ring and find where it intersects the fourth blue ring. The circles are show points of constant radius from the focus, so by looking for the appropriate intersections we get the points on the hyperbola. Where does the third red ring cross the fifth blue, and so on. Below we've gone out to the sixth red and eighth blue circle. We can see how we could easily extend this to further intersections. If you squint at the figure other patterns of intersections between the circles hint at the hyperbolas where the difference in the distances is 1 or 3 or

4 units. The line bisecting the foci, is a special case hyperbola where the difference between the distances is 0.

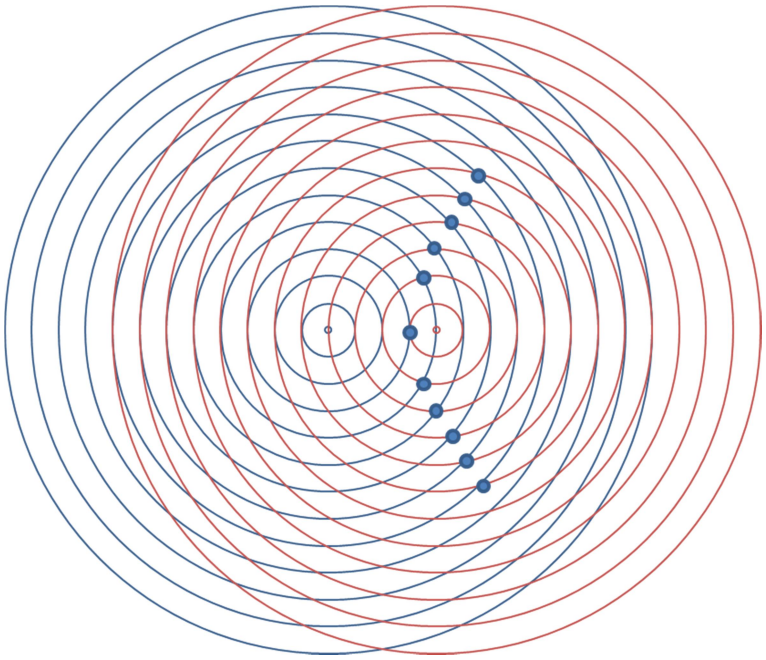


Figure 21. A few points on the hyperbola

The traditional definition of the hyperbola also includes the points where the blue center is the nearer point to the hyperbola. That adds the following points to the hyperbola.

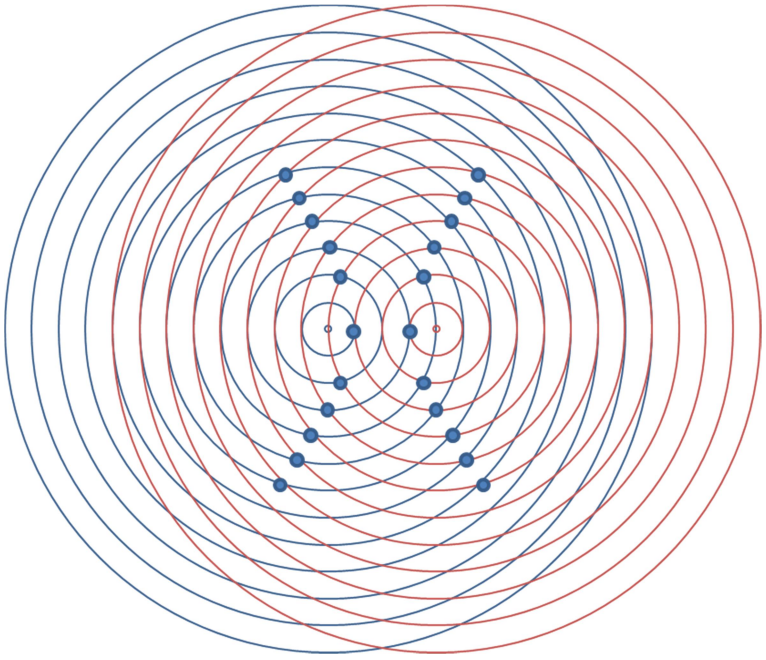


Figure 22. The second branch of the hyperbola.

The hyperbola has two branches, but since objects in orbit don't teleport, only one of the branches is used in a given orbit.

Generating rectangles are very helpful in understanding and hyperbolas.

The setup for the hyperbola is a little more complex then for ellipses. The next image gets things started…

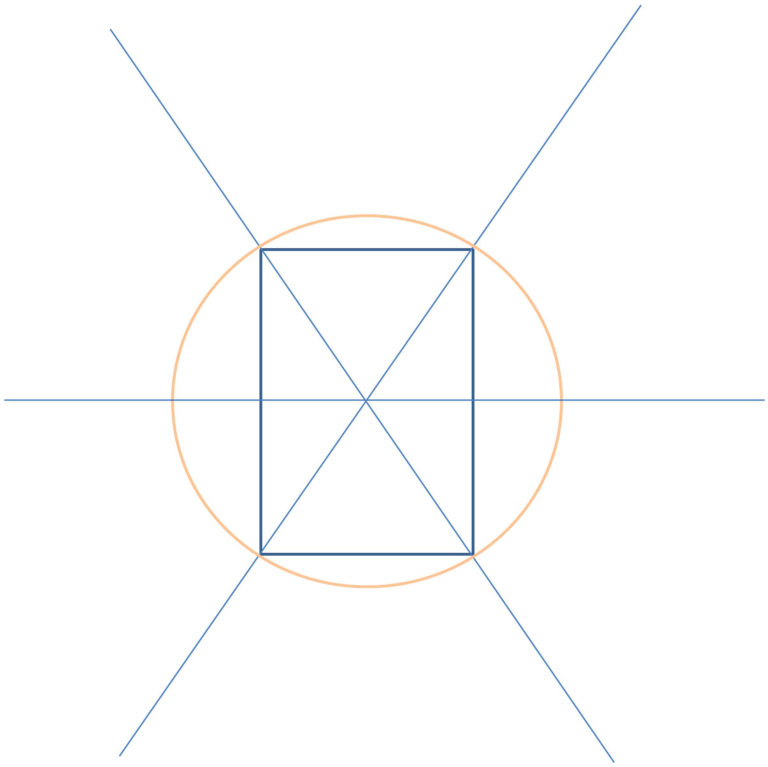


Figure 23. Setup for drawing hyperbola.

What do we have here? We start with a rectangle. This needn't be aligned with the page as it is here, but let's keep things simple. We draw the diagonals of the rectangle and extend them to infinity. We draw the circle that circumscribes the rectangle and lastly we draw a line through the center of the rectangle perpendicular to one of the sides. We could pick either one depending upon whether we want the hyperbola to open up/down or left/right. Since our line goes right/left, that's the way the hyperbola will open. We're using the hyperbola for an orbit, so we only consider one of the branches, the branch that opens to the right.

We can quickly identify a number of elements of the hyperbola.

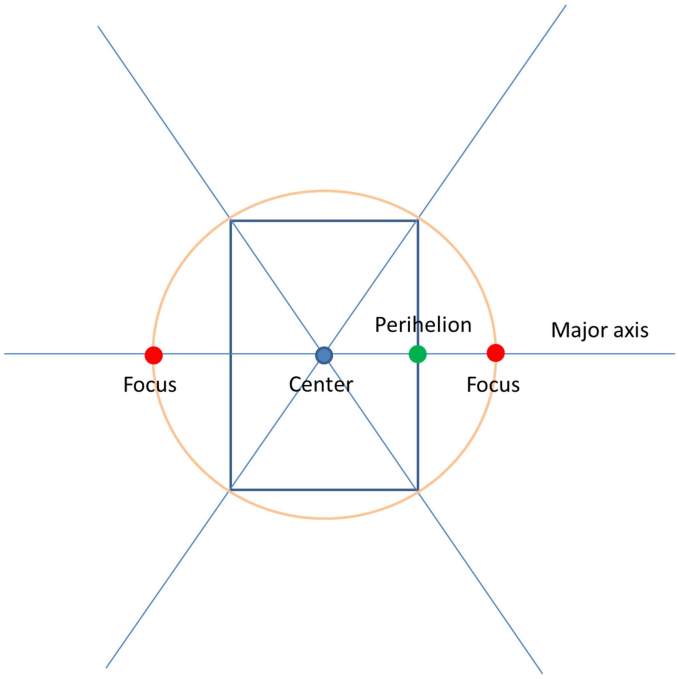


Figure 24. Elements of the hyperbola.

We see the two foci of the hyperbola where the circle that encloses the rectangle crosses the line we drew. As with ellipses, the line that joins the foci is the major axis, and the center of the hyperbola is midway between the foci. This can seem a little strange when we only draw only one branch of the hyperbola. Then the 'center' is actually to one side of the rest of the figure. We've drawn just one actual point on the hyperbola: the perihelion point will be at the point where the major axis crosses the rectangle. Here we're assuming the Sun is at the focus on the right.

Consider the green perihelion point, and its offset from the two foci. We can see that the difference in distances from the two red foci is simply the length of a side of the rectangle. Just as when we use rectangles to construct ellipses, the sides of this rectangle are the major and minor axes lengths, and so the difference in the distances from the two sides is twice the semimajor axis.

There's no simple way to draw the hyperbola, but we know that it goes through the perihelion point, and we know that at large distances the hyperbola gets closer and closer to, but never crosses the diagonals of the rectangle. Those are the *asymptotes* of the hyperbola, so far from the Sun an object on a hyperbolic orbits seems to be going essentially on a straight line.

So we can draw free hand what the hyperbola looks like…

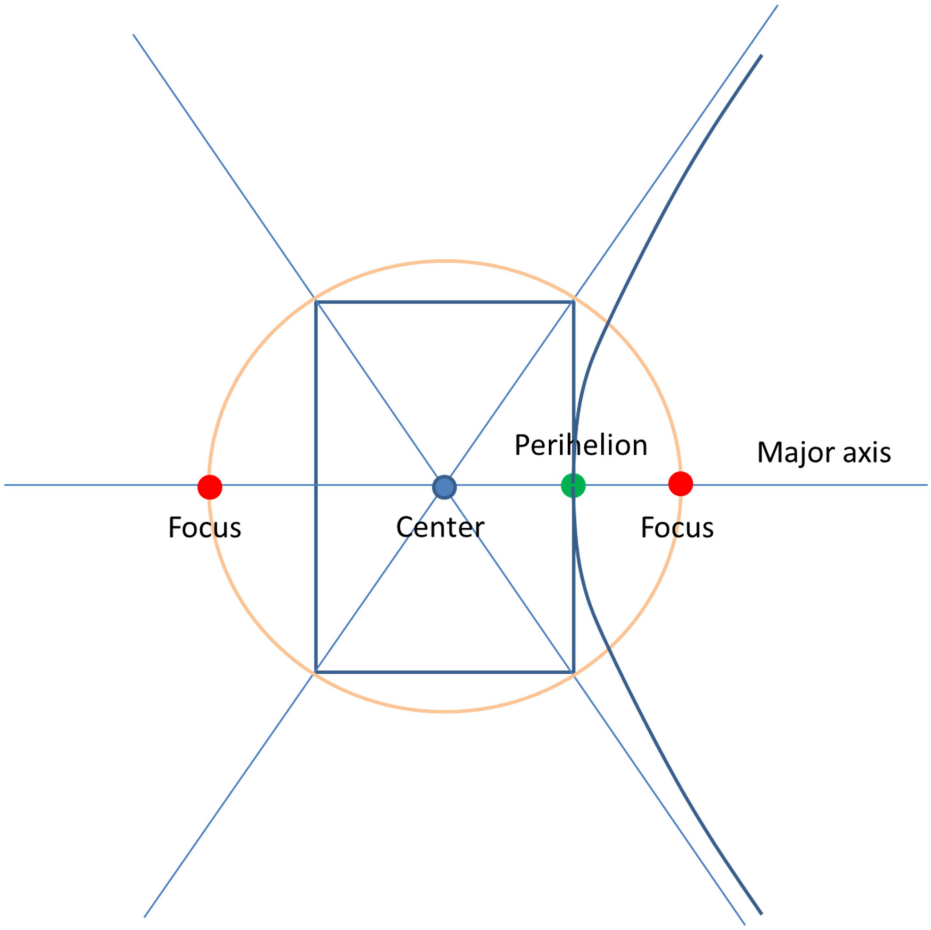


Figure 25. The hyperbolic orbit

The object comes in along one asymptote turns near the Sun and goes off along the other asymptote. We chose to draw the major axis through the shorter axis of the rectangle, so the object turns by less than 90°. If we had drawn the major axis through the longer side of the rectangle, the turn would be greater than 90°. If we had chosen a square for our rectangle, then the orbit turns by exactly 90°.

We have to rethink the values we've used for describing orbits and see if they still work for hyperbolic orbits. There is no period for the orbit, it never repeats. As noted above the lengths of the rectangle we use to generate the hyperbola give the major and minor axes lengths just as they do for ellipses, but there are a few differences. We use the length of the side of the rectangle that is parallel to the major axis as the length of the major axis. In this case we chose the shorter side, so the magnitude of the major axis will be less than the magnitude of the minor axis.

Many of the same formulae we use for orbits are unchanged for hyperbolic and elliptical orbits except that we treat the major axis length as being negative. Since the foci are now 'outside' the figure (i.e., the figure is drawn between the foci and the center) maybe that makes sense, but in any case it works.

The eccentricity for a hyperbolic orbit is simply the length of the diagonal of the generating rectangle over the length of the major axis. So it is always greater than 1. For orbits where the object doesn't

turn much, the eccentricity can be very large, for orbits where it turns a lot the eccentricity is just over 1. For the 'square' hyperbola where the orbit turns 90°, the eccentricity is √2 the length of the diagonal of a square over its side.

The orientation angles, inclination, argument of perihelion and longitude of the ascending node are mostly unaffected. We still have an orbit plane that intersects the reference plane, and we can still point to the perihelion. It's possible that there is only a descending node, the object comes in on an asymptote that's above the reference plane, and goes off on an asymptote that's below the plane only crossing the reference plane just once. In that case we just add 180° to the longitude of the descending node to get the nominal ascending node.

The true anomaly is fine – the object has a position and we can measure where it is with respect to the perihelion point. However the true anomaly never takes the values between the asymptotes of the orbit, unlike elliptical orbits where the true anomaly takes on all angles over the course of an orbit since the focus is surrounded by the ellipse.

The mean anomaly in hyperbolic orbits

If we defined the mean anomaly as the average rate of change of the true anomaly, we're in trouble with hyperbolic orbits – that's always 0 for a hyperbolic orbit since true anomaly only changes by a finite amount even over an infinite time. However our alternative definition of the mean anomaly using Kepler's Second Law is easy to adapt. We want the mean anomaly to be 0 at perihelion, so after perihelion we can define the mean anomaly as proportional to the area swept out by the orbit since perihelion, and before perihelion it is the negative of the area to be swept out till we reach perihelion. We've already seen that we have major and minor axes so we can define an area that we use to normalize the mean anomaly. That also means that although we don't have a real period for the orbit, we still can define a characteristic time associated with the orbit that can be used in most formulae that need a period.

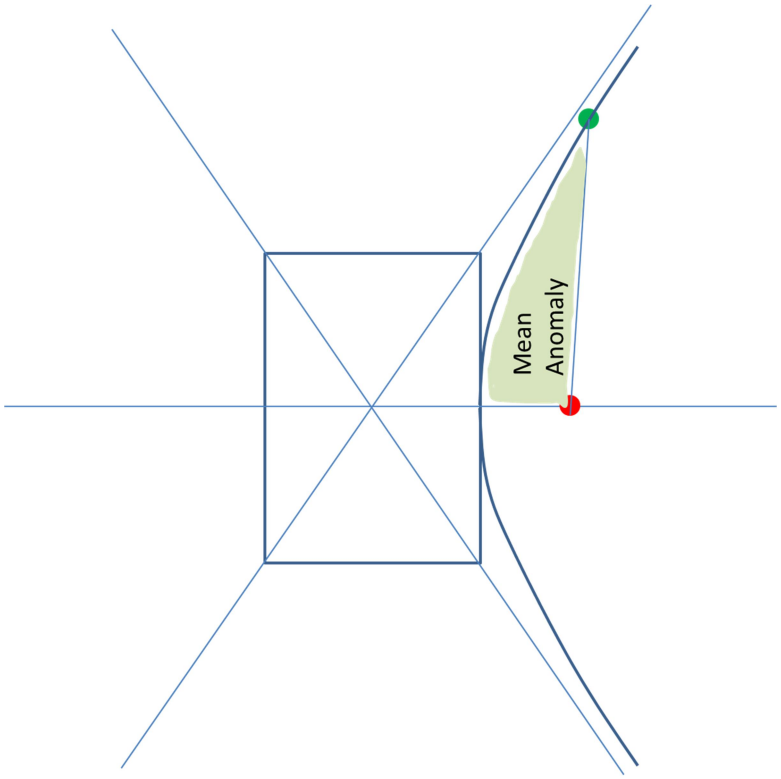


Figure 26. The hyperbolic mean anomaly.

The hyperbolic anomaly

Just as the eccentric anomaly helped us to tie together the mean and true anomalies for elliptical orbits, we use a quantity called the hyperbolic anomaly in hyperbolic orbits. The eccentric anomaly uses an ellipse with equal major and minor axes, a circle, the hyperbolic anomaly uses the equivalent hyperbola, the 'square' hyperbola we've mentioned above.

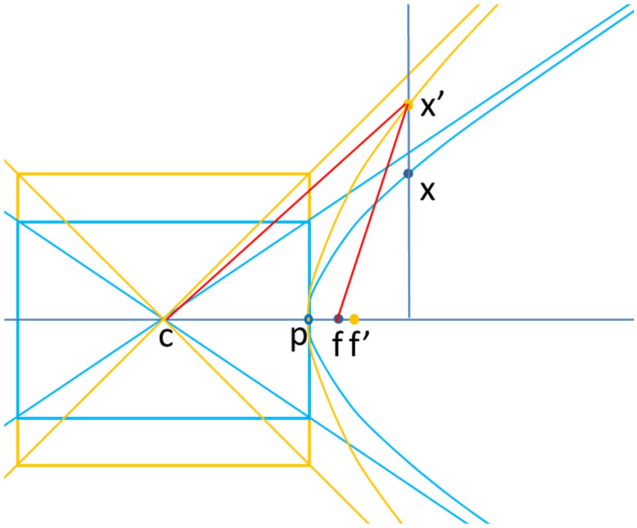


Figure 27. Setting up the hyperbolic anomaly

This figure shows two hyperbolas. The first is drawn in blue. This represents the actual orbit of the object. An object comes in along one of the arms and moves out along the second having turned significantly more than 90° in this encounter. There is a blue box used to construct the hyperbola and a blue dot on the major axis representing the focus of this hyperbola. We're interested in finding the hyperbolic anomaly associated with the blue point in the blue orbit at **x**.

There is also a golden hyperbola, which is generated by the square with the same major axis and center as the original hyperbola (just as the eccentric anomaly used the square with the same major axis and center as the original ellipse). The asymptotes are perpendicular to each other so this represents an encounter where the object turns 90°. The new hyperbola has the same perihelion and center as the original hyperbola, but the focus is different. We find a point on the golden hyperbola corresponding to **x** by drawing a line through **x** perpedicular to the major axis. This picks out the point **x'** on the golden hyperbola. We now draw a red line from the center of the golden hyperbola (where its asymptotes cross at **c**), to **x'**.

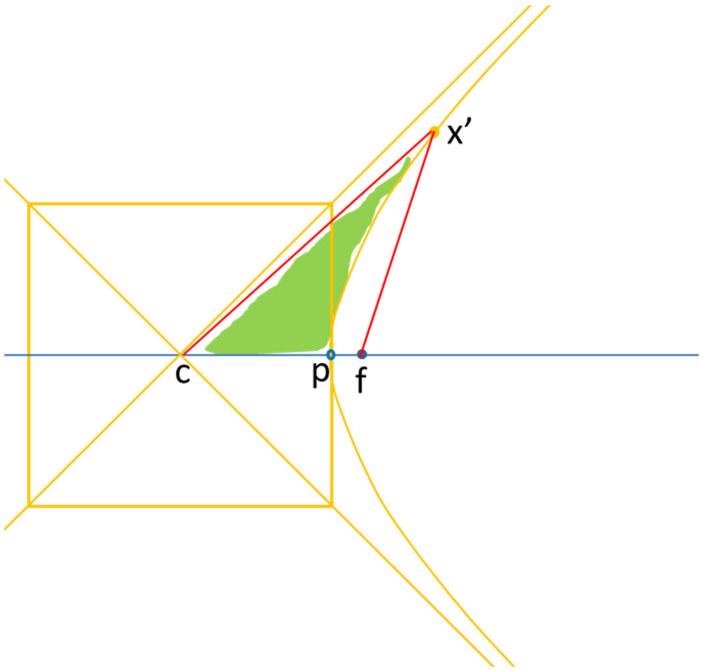


Figure 28. Hyperbolic anomaly

We've redrawn the figure getting rid of all of the information regarding the original hyperbola to simplify things.

The hyperbolic anomaly is the area of the 'triangle' bounded the red line between **c** and **x'**, the major axis between **c** and **p** and the arc of the hyperbola between **p** and **x'**. In our detailed derivation below we'll see that it's also the hyperbolic angle of the 'square', golden hyperbola between **p** and **x'** which is just twice this area. Combining the areal and angular representations, the hyperbolic anomaly bridges the mean and true anomalies of unbound orbits in the same way that the eccentric anomaly does for bound ones.

Both the mean and hyperbolic anomalies are unbounded, they go to plus and minus infinity. However while the mean anomaly changes at a constant rate, the hyperbolic anomaly changes only very slowly when we are far from the central encounter.

Parabolic Orbits

Bound particles move in elliptical orbits, unbound orbits move in hyperbolic orbits. What about orbits on the boundary between bound and unbound. If a planet had exactly 0 binding energy, it would take a parabolic path. A parabola is the transition figure between ellipses and hyperbolas. Unlike an elliptical orbit there is no maximum radius. Just as in a hyperbolic orbit the object comes in and then leaves for infinity. However unlike a hyperbolic orbit, the velocity of the object gets smaller and smaller and approaches 0 as it leaves the Sun. Nor can the behavior of the orbit at large distances be approximated by asymptotic lines as it can in a hyperbolic orbit.

In parabolic orbits the size of the semimajor axis and the period are not well defined. However unlike most elliptical and hyperbolic orbits, one can write relatively simple equations that relate the position and time, so the orbit calculation is straightforward. The eccentricity of a parabolic orbit (and all radial orbits) is exactly 1. The true anomaly takes on all values, but it takes infinite time to do so.

A few comets are calculated as having parabolic orbits. Presumably high precision orbital measurements would tip them to bound or unbound, or perturbations of the orbit by planets and non-gravitational affects may have them crossing the boundary perhaps more than once.

Transitions among orbit classes

You may recall our thought experiment where we took a circular orbit and slowed down the planet to see what would happen. What if we did the reverse?

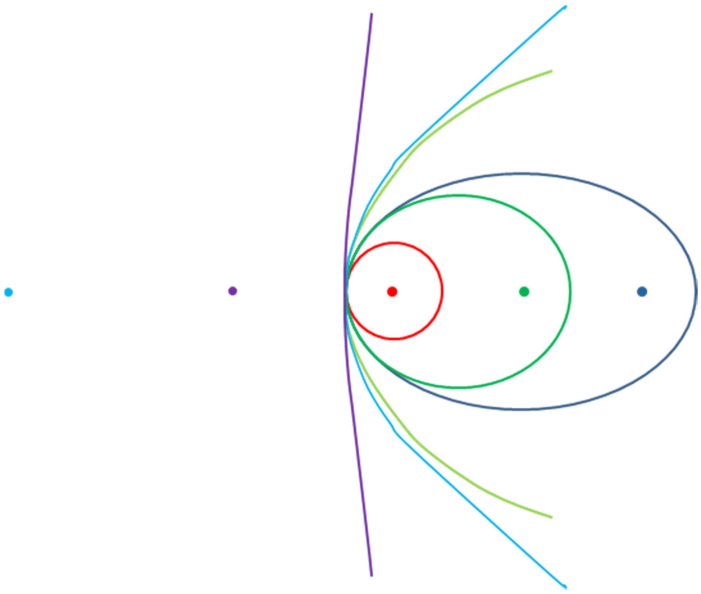


Figure 29. Increasing the energy in orbits: Orbits have one focus at the Sun/red dot and usually a second focus shown in the same color as the orbit.

We start with the red circular orbit: eccentricity 0 and the two foci are coincident. The Sun is at the red dot. If we look at the orbit of an object that was moving faster at its perihelion point than the circular orbit, we might get the dark green elliptical orbit. The second focus of the ellipse moves. It is no longer coincident with the red dot, but has moved to the right as shown. If we keep increasing the velocity we get the dark blue ellipse and the focus moves even further away, the eccentricity rises towards 1. If we increase the velocity from the original circular orbit by $\sqrt{2}$ (at the point where all the curves are tangent) then we get the light green parabolic orbit. There is no second focus to the orbit, it has moved out to an infinite distance. The eccentricity of the orbit is exactly 1.

We keep increasing the speed: now we are in a hyperbolic orbit, the other focus reappears but on the other side of the page! It starts coming back in from infinity but 180 degrees away from where it left. We see the focus for the light blue hyperbola on the left. The purple orbit is the highest energy orbit we plot; our object is moving so fast that the Sun only deflects it slightly. The second focus for the purple hyperbola moves closer to the Sun. In our little experiment, the focus always moves towards the right, except for the discontinuity at the parabolic orbit. The eccentricity continues to grow as the angle between the incoming and outgoing asymptotes of the orbit gets larger. When that angle reaches 90°, the eccentricity is $\sqrt{2}$. As the system becomes more and more unbound, the second focus returns towards the focus at the Sun. At infinite velocity, where the orbit is just a straight line unaffected by the Sun, the second focus would be placed symmetrically on the other side of the orbit line.

As the velocity of the orbit at perihelion increases we progress through all five conic figures: circle, ellipse, parabola, hyperbola and line.

Center of Mass

We know have a full framework for describing all of the cases where we have a single massless body orbiting a single point mass. This is an idealized situation which never arises in the real world, but because of the predominant mass of the Sun is a pretty good approximation for the Solar System. It turns out that we can actually solve the more realistic two-body problem using exactly the same mathematics as the ideal problem we've been looking at. When two massive bodies orbit around one another, the orbits have the same shape. One body does not revolve around the other. They both revolve around the center of mass.

The *center of mass* is along the line between the two points, and closer to the more massive object by the ratio of the mass of the two objects. E.g., if we have two objects where one is 99 times more massive then the other, then the lighter object is 99 times further from the center of mass than the heavier one. The Sun is a million times heavier than the Earth, so considering only those two bodies the center of mass is about 100 miles or 150 kilometers from the center of the Sun. Since the Sun is a more than a million kilometers in diameter, that's not a very big effect. While the Earth orbits the Sun, the Sun moves in a tiny circle around the offset point – or it would if the Earth were the main thing pulling at the Sun.

Jupiter is about one thousandth the mass of the Sun, and much further away. The center of mass of the Jupiter/Sun orbit is a bit above the surface of the Sun.

Instead of using the actual Sun as our reference point for coordinates, we often use the center of mass of the Solar system as our reference point.

Precession

We mentioned that Newton found that we get elliptical closed orbits when we are influenced by a single body with all of its mass concentrated at a point. He could relax this slightly, the mass doesn't have to be in a point though, so long as the body was exactly spherical. The orbit of satellites around the Earth and other planets are subject to substantial deviations from both of these assumptions. They are attracted to the Sun as well as the planet they are orbiting, and they are often orbiting the planet closely enough that the fact that the planet is not quite spherical – they are generally oblate spheroids like the Earth – causes the orbits to change even over quite small times scales. Fortunately most of the effects can be accommodated by adding terms where the longitude of the ascending node and the argument of perigee change linearly with time. The rates of change for the two angles are not, however the same. The inclination of the orbit, the tilt, does not change, but the change in the longitude of the ascending node means that the direction of the tilt moves in a circle around the sky, very similar to the precession of the Earth's rotation.

The change in the argument of perigee means that within the orbit plane, the perihelion point also revolves around the primary.

The simple linear fit is only an approximation but it means that we can define orbital elements for satellites with a couple of extra terms for the precession that do a much better job approximating the orbit for a reasonable time.

The rates of precession are affected by the eccentricity and periapsis of the orbit. The orbits chosen for Earth satellites are often picked to get a particular rate of precession. In particular, many satellites use what is called a Sun-synchronous orbit, where the precession of the longitude of the ascending node is 360° per year. This means that the satellite can always have the same attitude towards the Sun, so, e.g., we can avoid ever crossing the shadow of the Earth where solar panels are notoriously ineffective.

Some basic astronautics

Even with the qualitative approach we've taken so far we can deduce a few things about how we use rockets to get around the Solar System. Rockets are complex beasts, but basically all they are intended to do is change the velocity of their payloads. A given rocket can be characterized by the total change in velocity that it can impart. We can call the change we can impart ΔV. The hard part in rocketry is the take off and landing and getting through the atmosphere. We'll ignore that here! We just want to think about how can use a rocket's ΔV to get around the Solar System.

We have some initial orbit we are starting with and a final orbit we want to reach. These orbits are going to differ in energy, angular momentum and phase. Most rockets we use today are chemical and typically burn only for a few minutes at most. Since orbits have periods of hours to centuries we can consider the rockets as making instantaneous changes to the velocity. So to first order we change the velocity of the rocket while there is a negligible change in the position. So we start on an orbit with value of the velocity and we can add or subtract velocity – or move perpendicular to the current velocity. What do these do to the energy and angular momentum?

If add velocity in the direction we are already moving we move faster. Our kinetic energy increases and the binding energy of the orbit decreases, the semimajor axis and the period of the orbit increase, so even though we may think we've sped up the orbit, the average angular velocity for the orbit will be lower.

The kinetic energy goes as v^2, which means that if we need to change the energy of the orbit there may be a best time to do it. If we start with v and add ΔV our new velocity is (v+ΔV) but the kinetic energy is now $v^2 + 2v\,\Delta V + \Delta V^2$.

Suppose we have an object in a somewhat elliptical orbit. At aphelion the body is moving slowly, maybe 10 km/s. At perihelion the object is moving more quickly, say 20 km/s. If we have a rocket that give 5 km/s of ΔV, then can change the energy of the orbit by 20 x 5 +25 at aphelion, but by 40x 5 + 25 at perihelion, 125 versus 225. We get almost twice the change in energy by doing our orbit maneuver when we are moving faster. So the first general rule is that to get the most change in energy we do

burns near perihelion – or when we are moving quickly in a planetary encounter. Note that this all applies if we want to lose energy too, we just fire our rockets in the opposite direction.

Conversely the angular momentum is just the product of the radius and the velocity. Since there is not a square of the velocity, there's no benefit from doing the burn where the velocity is highest. In fact to get the largest value of $r \Delta V$ we want to make the burn when r is largest – which happens to be when the velocity is smallest. We get more leverage in the burn if we do it when we are furthest from the primary. So do burns to change angular momentum at aphelion burns to change energy at perihelion.

Another basic concept for orbits is that to the extent that the orbits are closed (e.g., ignoring precession and third bodies), the orbit is fixed when we're not doing a burn. We always come back to the last place we did an orbit maneuver.

Finally, recall our hyperbolic orbits. If our satellite is on a hyperbolic orbit with respect to a planet then the encounter will change the direction of the orbit with respect to the planet but the outgoing velocity with respect to the planet will be exactly the same as the incoming velocity. However with respect to the Sun these velocities can be very different. A hyperbolic encounter with a planet can in principle change the velocity of the spacecraft by up to twice the escape velocity of the planet at point of closest approach. This can be many km/s, so that hyperbolic encounters are key to many of our interplanetary missions.

Orbital Elements

We have two parameters related to the physics of the orbit: the semimajor axis and the eccentricity; three angles that orient the orbit in space: the inclination, argument of perihelion and longitude of the ascending node; and either a perihelion time (when the orbit phase was implicitly 0), or a pair of a time and phase. These six or seven numbers are the *orbital elements*. They almost completely define the orbit. We generally think of these as the only elements required, because we normally are assuming we know what object we are orbiting (the Sun, the Earth or one of the other planets). However in general we've got one more degree of freedom. We might have a very long period orbit around a very light body, or a very short period orbit around a super-massive star where both orbits have exactly the same shape (and where we pick the phase to have some value at a given time). So the period or the mass of the central object (or perhaps more specifically the product of the gravitational constant and the mass) needs to be specified to get a completely determined orbit. A specific source may give additional information about the orbit, or they may use some of the alternative choices for elements that we noted above. If we know the mass of the object we are orbiting, with the six or seven orbital element numbers we've completely defined the orbit.

Osculating and Principal Elements

Since the orbital elements are derived assuming that there are no bodies other than the Sun and the planet, they are not exactly correct and the elements of all of the bodies in the Solar System change with time.

If we have the position and velocity of a planet at a particular time with respect to the center of the orbit we have the *state vector* of the orbit. Position and velocity are each vectors with three components, so this state vector has seven numbers (including the time). Given a state vector we can compute the orbital elements that would exactly match the state vector. This set of orbital elements are called *osculating* elements. Osculate is a Latinate word that means to kiss. These elements produce an orbit which is exactly correct at the time we have the position and velocity for, but then gradually diverge. The osculating orbit is tangent to the real orbit, and just 'kisses' it at the reference time.

Note that the generation of the osculating elements is a straightforward bit of mathematics that relies only on the input position and velocity. That means that if we have the position and velocity of the Earth with respect to Pluto we could calculate the osculating elements for the Earth orbiting Pluto at any given time. This may seem nonsense. We need to be careful to understand that the mathematics that generate the osculating elements are a way of describing orbits that are known, they aren't a way of discovering orbits.

If we calculate the osculating elements of the Earth with respect to Pluto at one time, and then three months later, the elements are going to have changed a lot. However if we calculate the osculating elements of the Earth with respect to the center of mass of the Earth-Sun system at the those same times, we find that the orbit does indeed change but just a tiny bit; the changes to the elements are much, much smaller. To first order we can treat them and the orbit as constant.

Since the osculating elements do change with time, astronomers sometimes try to calculate elements which average out short term variations. These elements, while they may never be exactly right the way osculating elements are at their reference time, tend to be righter, longer than a randomly chosen set of osculating elements, i.e., they minimize the maximum error. Such elements are sometimes called *mean* orbital elements.

Deriving Kepler's Laws from Newton's Theory of Gravity

We've finished a broad, conceptual introduction to orbits and the vocabulary used to describe them. In this next section we're going to be a bit more formal, and describe how we get to Kepler's Laws starting with Newton. This discussion is going to become considerably more mathematical with equations and even a fair bit of calculus. However we will be infusing the mathematics with substantial discussion of what the equations mean so even if the math is a bit challenging you may be able to follow.

Newton's Law of Gravity

Newton's law of gravity can be stated as

$$\boldsymbol{F}_m = -\frac{GMm}{|\boldsymbol{r}_m - \boldsymbol{r}_M|^2}\frac{\boldsymbol{r}_m - \boldsymbol{r}_M}{|\boldsymbol{r}_m - \boldsymbol{r}_M|}$$

Here we have a body with a mass m at $\boldsymbol{r}_m$, and a second body of mass Mat $\boldsymbol{r}_M$. This equation gives the gravitational force on the mass m that arises because of the mass M. Quantities in bold are vectors: they have three components. Other quantities are scalars. Typically we're considering situations where

$m \ll M$, i.e., the m is the mass of the Earth and $M$is the mass of the Sun, but this equation is valid regardless of the masses.

The GMm numerator of the first term says that the force is proportional to the mass of both bodies. If we double either body's mass, the force is increased by a factor of two. Since $\boldsymbol{F}_m = m\boldsymbol{a}_m$, force is mass times acceleration, this also means that the acceleration of on one body due to the gravitation attraction by a second, is independent of the mass of the first body – it depends only on the mass of the body doing the attracting. This observation would ultimately be one of the basic principles of Einstein's revision of Newtonian gravity.

There is a constant of proportionality G, which gives the strength of the gravitational force. The value of G is much smaller than 1, so that we need a lot of mass to create a substantial gravitational force. The current best measured value for G is 6.67430×10^{-11} m^3 kg^{-1} s^{-2}.

If we had a test mass, m, of one kilogram, we'd need to have about 1.5×10^{10} kilograms just one meter away to exert a force about 1 tenth of the weight of the mass (i.e., the force exerted by Earth's gravity). Squeezing fifteen million tons of matter within a meter is challenging. Gravity is really weak. Direct measurements of the gravitational constant remain quite difficult and G is among the most poorly known of the basic constants in physics. In spite of that astronomers are able to make amazingly accurate predictions. Although we don't know G itself very well, the product of G and the mass of various bodies in the solar system has been measured to exquisite precision. If it turns out that G is a hair larger than the current accepted value, the masses we've determined for bodies in the Solar system will be a corresponding tiny bit smaller, or vice versa if the true value of G is a little lower.

Under the product of the masses we have a denominator which is just the square of the distance between the two points. The || operator gives the positive magnitude of a scalar or vector. So |-1| = 1, |3| = 3, and |**(1,2,3)**| is $\sqrt{14} = \sqrt{1^2 + 2^2 + 3^2}$ when **(1,2,3)** is a Cartesian vector with components 1,2,3. This term is why we call gravity an inverse-square force, it gets weaker proportionally to the square of the distance between the two bodies.

The second term, $\frac{\boldsymbol{r}_m - \boldsymbol{r}_M}{|\boldsymbol{r}_m - \boldsymbol{r}_M|}$, is simply the vector between the two bodies divided by the magnitude of that vectors. This means that this second term as a whole is a vector, but it has unit magnitude. It is a direction in space, pointing from the M body towards the m body.

The first term gives the magnitude of the force, the second term gives the direction. And we don't want to forget the lonely minus sign in front. This reverses the direction of the force, so that gravity is an attractive force: the force on mass m points to mass M.

If we consider the force on the mass M by the mass m, then we get the almost exactly the same result since $|\boldsymbol{r}_m - \boldsymbol{r}_M| = |\boldsymbol{r}_M - \boldsymbol{r}_m|$. The force has exactly the same magnitude but is pointing in the opposite direction.

$$\boldsymbol{F}_M = -\boldsymbol{F}_m$$

So the total force on the pair adds up to 0.

In general in a closed system – a system that isn't being acted upon from outside -- the sum of forces in the system is 0.

Center of Mass

If we have a system with any number of point masses labeled with i, we could define a position called the *center of mass* as

$$\boldsymbol{C}_{mass} = \frac{\sum m_i \boldsymbol{r}_i}{\sum m_i}$$

It's just the mass-weighted sum of positions.

The center of mass would be moving with a velocity

$$\boldsymbol{v}_{mass} = \frac{\sum m_i \boldsymbol{v}_i}{\sum m_i}$$

However if we calculate the acceleration of the center of mass it would be

$$\boldsymbol{a}_{mass} = \frac{\sum m_i \frac{d\boldsymbol{v}_i}{dt}}{\sum m_i} = \frac{\sum \boldsymbol{F}_i}{\sum m_i}$$

But we just noted that the total of the forces in a closed system must be 0, so the center of mass of a closed system does not accelerate.

Often when working in orbits, we will find it most convenient to work in a frame where the center of mass position vector and its velocity are 0.

If we choose this frame for our system with two bodies with masses m and Mwe have:

$m\boldsymbol{r}_m = -M\boldsymbol{r}_M$, $m\boldsymbol{v}_m = -M\boldsymbol{v}_M$ and we already had $\boldsymbol{F}_M = -\boldsymbol{F}_m$.

So the two bodies mirror one another, but the motions and velocities are scaled by the inverses of the masses. The heavier body moves slower in a proportionally smaller orbit.

Reduced mass

We can use the center of mass to reduce our two-body problem to a one-body problem.

Let $\boldsymbol{r} = \boldsymbol{r}_m - \boldsymbol{r}_M$ be the full vector between the two bodies. Then the second derivative of this vector is simply the difference of the derivatives,

$$\frac{d^2\boldsymbol{r}}{dt^2} = \frac{d^2\boldsymbol{r}_m}{dt^2} - \frac{d^2\boldsymbol{r}_M}{dt^2}$$

However we have that the forces sum to 0, so

$$m\frac{d^2\boldsymbol{r}_m}{dt^2} = -M\frac{d^2\boldsymbol{r}_M}{dt^2}$$

Or

$$\frac{d^2\boldsymbol{r}}{dt^2} = (1+\frac{m}{M})\frac{d^2\boldsymbol{r}_m}{dt^2}$$

$$\frac{d^2\boldsymbol{r}}{dt^2} = (m+M)/M\frac{d^2\boldsymbol{r}_m}{dt^2}$$

$$M/(m+M)\frac{d^2\boldsymbol{r}}{dt^2} = \frac{d^2\boldsymbol{r}_m}{dt^2}$$

If we take this last equation and multiply both sides by m we get

$$\frac{mM}{m+M}\frac{d^2\boldsymbol{r}}{dt^2} = m\frac{d^2\boldsymbol{r}_m}{dt^2} = \boldsymbol{F}_m$$

If we simple want the acceleration

$$\frac{mM}{m+M}\frac{d^2\boldsymbol{r}}{dt^2} = -\frac{GMm}{|\boldsymbol{r}|^2}\frac{\boldsymbol{r}}{|\boldsymbol{r}|}$$

$$\frac{d^2\boldsymbol{r}}{dt^2} = -\frac{G(m+M)}{|\boldsymbol{r}|^2}\frac{\boldsymbol{r}}{|\boldsymbol{r}|}$$

The total acceleration vector acts as if the total mass of the system is being used and so we can calculate its change as if we had massless body orbiting a body with the sum of the two masses.

So we can solve for the motion of both bodies as a one-body problem where the separation between the bodies is subject to the same force as one of the bodies, but reacts as if it has a little bit less mass.

The quantity $\frac{mM}{m+M}$ is called the *reduced mass* and is sometimes designated with a small Greek mu, μ. If $m \ll M$ then the reduced mass is a little bit less than m. If the two bodies are equal mass, then the reduced mass is half the original masses.

In the case where m is small, the reduced mass is approximately $m(1-\frac{m}{M})$.

So the error we make in ignoring the mass of the planets relative to the Sun, is roughly the ratio of their masses, one part in a million for the Earth, one part in a thousand for Jupiter. This isn't an acceptable error when we're trying to navigate across the Solar system: showing up a few hundred thousand miles away from where we want to be isn't going to be good for our satellite, but for understanding and

describing orbits qualitatively as we did in our introduction, it seems reasonable. For Kepler, who was working with data taken without aid of telescopes, it was fine.

Gravitational Effective Mass

In deriving the reduced mass we didn't use the form of gravitational force at all. Physicists can and do use reduced masses in dealing with particles undergoing a variety of interactions. However the form we got is not really the most convenient in terms of actually calculating the orbit. We're likely using the center of mass as the origin--the place where the coordinates are all 0-- of our coordinate system, and we'd like to know the form of the orbit using the coordinates of the planet relative to the center of mass. We can write the gravitational force on the planet as:

$$\boldsymbol{F}_m = -\frac{GMm}{|\boldsymbol{r}|^2}\frac{\boldsymbol{r}}{|\boldsymbol{r}|}$$

However $m\boldsymbol{r}_m = -M\boldsymbol{r}_M$ so we can compute $\boldsymbol{r} = (1+\frac{m}{M})\boldsymbol{r}_m$.

$$\boldsymbol{F}_m = -\frac{GMm}{\left(1+\frac{m}{M}\right)^2|\boldsymbol{r}_m|^2}$$

$$\frac{\boldsymbol{r}_m}{|\boldsymbol{r}_m|}$$

$$\boldsymbol{F}_m = -\frac{GMm}{(\frac{m+M}{M})^2|\boldsymbol{r}_m|^2}\frac{\boldsymbol{r}_m}{|\boldsymbol{r}_m|}$$

$$\boldsymbol{F}_m = -\frac{Gm\frac{M^3}{(m+M)^2}}{|\boldsymbol{r}_m|^2}\frac{\boldsymbol{r}_m}{|\boldsymbol{r}_m|}$$

This says that the orbit of the body around the center of mass, is effectively that of a massless particle around a body at the center of mass with mass $\frac{M^3}{(m+M)^2}$. This formulation allows us to directly compute in center of mass coordinates rather than having to transform back from the difference vector to a component, as we might have to do were we using the formulation we got with the reduced mass above.

Note that we simply relabeled the vectors in the second term, $\frac{\boldsymbol{r}_m}{|\boldsymbol{r}_m|}$. This ratio is a unit vector and $\boldsymbol{r}_m$ and $\boldsymbol{r}$ point in the same direction, so their associated unit vectors are the same.

Ellipse in polar coordinates

We will want to do much of our subsequent discussion in polar and spherical coordinates and it will help if we can derive the equations for ellipse and hyperbolas in radial coordinates, where the center of the coordinate system is at one a focus. This is easy if we go back to our basic definition of the ellipse and use the law of cosines.

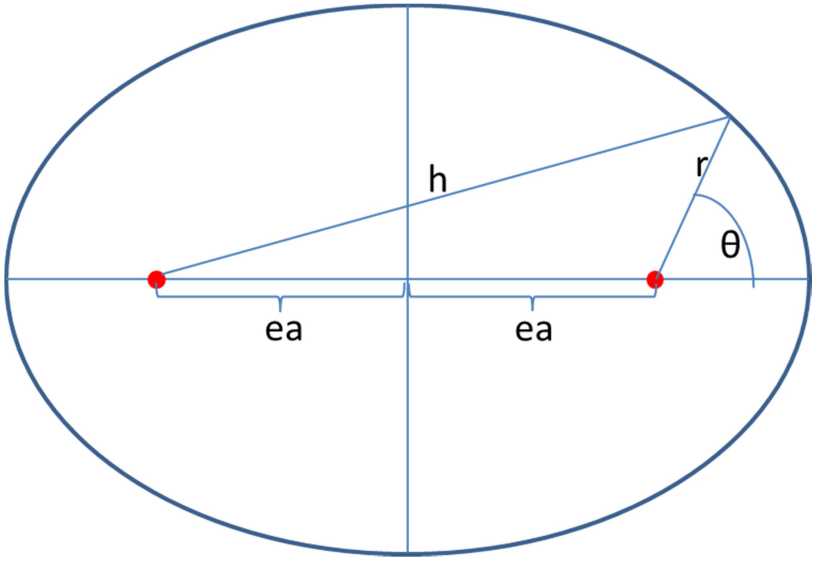


Figure 30. Ellipse in radial coordinates.

We've drawn an ellipse with origin at the right focus and shown the radial coordinates to a point on the ellipse. We've also drawn a second line of length h from the second focus to the current point on the ellipse. The ellipse has a semimajor axis a, and eccentricity e, so the distance between the center of the ellipse and the foci is simply ea. The second focus is 2ea to the left of the focus we are using as the center of coordinates. Since the sum of the distances to each point on the ellipse from the two foci is the same, then r+h must be a constant, and looking at the points on the major axis we can see that r+h=2a. We can calculate h from the law of cosines, since we know that the angle opposite the side h is just 180-θ. By the law of cosines

$$h^2 = r^2 + (2ea)^2 - 2r(2ea)\cos(180 - \theta)$$

$$h = \sqrt{r^2 + 4e^2a^2 + 4rea\cos(\theta)}$$

If we plug this into our equation for an ellipse

$$r + h = 2a$$

$$r + \sqrt{r^2 + 4e^2a^2 + 4rea\cos(\theta)} = 2a$$

$$\sqrt{r^2 + 4e^2a^2 + 4rea\cos(\theta)} = 2a - r$$

Squaring both sides of the equation

$$r^2 + 4e^2a^2 + 4rea\cos(\theta) = 4a^2 - 4ar + r^2$$

The two r^2 terms cancel and we can divide the remaining terms by $4a$ to get

$$e^2a + re\cos(\theta) = a - r$$

Rearranging to

$$r\,(1 + e\cos(\theta)) = a(1 - e^2)$$

And finally…

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

This is our equation in radial coordinates for an ellipse with a focus at the origin and the semimajor axis along the x axis. However since we're working in polar coordinates handling an ellipse at an arbitrary orientation is nothing more than adding in an appropriate offset to θ. E.g., if we offset by 90°, so that the cosine turns into a sine, then major axis will be along the y-axis. The appropriate offset will usually be the argument of perihelion. If we add 180°s to θ, then the second term in the denominator will change sign. What's important is that *e* is less than 1 so that r will always be finite.

Hyperbolas in polar coordinates

We can use the same approach to get the equation for an hyperbola in polar coordinates. Recall that for a hyperbola the difference of the distances from the foci is $2a$. So we start with h-r=2a, but proceeding the same way we ultimately get a result with just a few differences in signs that don't really affect the behavior

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

Essentially this is exactly the same equation as before. The changes in the signs simply reflect different choices for the orientation. What will make the real difference is that since *e*>1, there will be angles at which the radius goes to infinity, where $e\cos\theta = 1$ (or -1 if we switch the sign in the denominator).

Forces and acceleration in polar coordinates

The equation of motion for a body moving in a central force is a bit more complicated in polar coordinates than in Cartesian coordinates.

In Cartesian coordinates when we have some vector valued function of time and want to take derivatives, e.g., $\boldsymbol{r} = x\hat{\boldsymbol{x}} + y\hat{\boldsymbol{y}} + z\hat{\boldsymbol{z}}$,then we just have $\frac{\mathrm{d}\mathbf{r}}{\mathrm{dt}} = \frac{dx}{dt}\hat{\boldsymbol{x}} + \frac{dy}{dt}\hat{\boldsymbol{y}} + \frac{dz}{dt}\hat{\boldsymbol{z}}$. The unit vectors are constant in the Cartesian plane and through Cartesian space, so any terms involving the derivatives of the unit vectors are 0. However, this is not true for polar or spherical coordinates. The unit vectors change as we move around unless we are moving exactly radially. If we restrict ourselves to cylindrical

coordinates in a plane the two polar unit vectors are $\hat{\boldsymbol{r}}$ and $\hat{\boldsymbol{\theta}}$ with an unused $\hat{\boldsymbol{z}}$ unit vector pointing perpendicular to the plane.

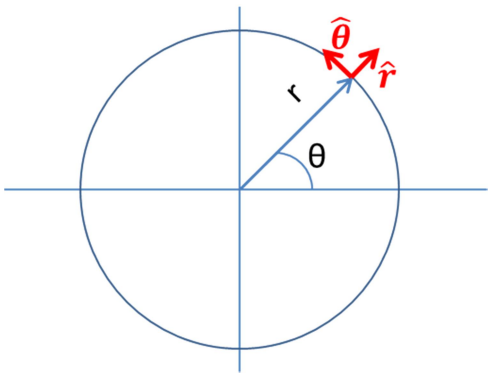


Figure 31. Unit vectors in polar coordinates

We can express the polar unit vectors in terms of the Carterian unit vectors

$$\hat{\boldsymbol{r}} = \cos\theta\ \hat{\boldsymbol{x}} + \sin\theta\,\hat{\boldsymbol{y}}$$

$$\hat{\boldsymbol{\theta}} = -\sin\theta\ \hat{\boldsymbol{x}} + \cos\theta\,\hat{\boldsymbol{y}}$$

Note that while the polar coordinates (r,θ) are components of a vector they are not themselves vectors but the unit vectors associated with them are. Note that $\boldsymbol{r}$ **(bolded)** is the full position vector while r is just the radius component of the position in polar coordinates, and $\hat{\boldsymbol{r}}$ is the unit vector showing the direction we move if we increase r slightly.

If we take the derivatives of the unit vectors we get

$$\frac{d\hat{\boldsymbol{r}}}{dt} = -\sin\theta\frac{d\theta}{dt}\,\hat{\boldsymbol{x}} + \cos\theta\frac{d\theta}{dt}\hat{\boldsymbol{y}} = \frac{d\theta}{dt}\hat{\boldsymbol{\theta}}$$

$$\frac{d\hat{\boldsymbol{\theta}}}{dt} = -\cos\theta\,\hat{\boldsymbol{x}}\frac{d\theta}{dt} - \sin\theta\frac{d\theta}{dt}\hat{\boldsymbol{y}} - \frac{d\theta}{dt}\hat{\boldsymbol{r}} = -\frac{d\theta}{dt}$$

The position is simply

$$\boldsymbol{r} = r\,\hat{\boldsymbol{r}}$$

The velocity is

$$\frac{d\boldsymbol{r}}{dt} = \frac{dr}{dt}\,\hat{\boldsymbol{r}} + r\frac{d\hat{\boldsymbol{r}}}{dt} = \frac{dr}{dt}\,\hat{\boldsymbol{r}} + r\frac{d\theta}{dt}\hat{\boldsymbol{\theta}}$$

For the acceleration we find

$$\frac{d^2\boldsymbol{r}}{dt^2} = \frac{d^2r}{dt^2}\,\hat{\boldsymbol{r}} + \frac{dr}{dt}\frac{d\theta}{dt}\hat{\boldsymbol{\theta}} + \frac{dr}{dt}\frac{d\theta}{dt}\hat{\boldsymbol{\theta}} + r\frac{d^2\theta}{dt^2}\hat{\boldsymbol{\theta}} - r\frac{d\theta}{dt}\frac{d\theta}{dt}\,\hat{\boldsymbol{r}}$$

$$\frac{d^2\boldsymbol{r}}{dt^2} = \left(\frac{d^2r}{dt^2} - r\frac{d\theta}{dt}\frac{d\theta}{dt}\right)\hat{\boldsymbol{r}} + (2\frac{dr}{dt}\frac{d\theta}{dt} + r\frac{d^2\theta}{dt^2})\hat{\boldsymbol{\theta}}$$

However we know that the gravitational acceleration is central: it has no component along the $\hat{\boldsymbol{\theta}}$ unit vector. So we can ignore those terms. The angular velocity $\frac{d\theta}{dt}$ is often written at ω. So we can simplify the acceleration from a central force to

$$\frac{d^2\boldsymbol{r}}{dt^2} = \left(\frac{d^2r}{dt^2} - r\omega^2\right)\hat{\boldsymbol{r}}$$

Setting this equal to the gravitational acceleration we have

$$\frac{d^2r}{dt^2} - r\omega^2 = -\frac{GM}{r^2}$$

With this background, we're ready to start our derivation of Kepler's laws.

Deriving Kepler's Second Law

We can now begin to derive Kepler's Laws from first principles. Kepler's Second Law, the area drawn out by the orbit is proportional to the time, is the easiest derive and is needed to get the other two.

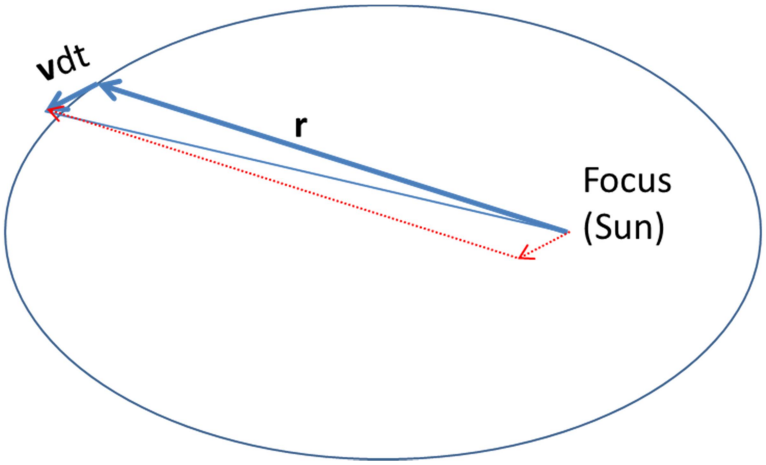


Figure 32. Deriving Kepler's Second Law

In this figure the area swept out in an interval, dt, is the blue triangle formed by the radius vector, and the distance the particle moves in a small time, $\boldsymbol{v}dt$. If we make dt small enough, the teeny bit of the triangle that's outside the ellipse becomes entirely negligible. The cross-product of the vectors $\boldsymbol{r}$ and $\boldsymbol{v}dt$ is a vector perpendicular to the ellipse (and the paper or screen on which it is drawn). The length of tha vector is the area of the parallelogram using the original vectors and the parallel red vectors. The area in the blue triangle is just half the area of the parallelogram. So we have the area being drawn out as

$$dA = |\frac{1}{2}\boldsymbol{r} \times \boldsymbol{v}dt| = |1/2\ \boldsymbol{L}/m|dt$$

Where **L** is the angular momentum of the body, defined as $\boldsymbol{L} = m\boldsymbol{r} \times \boldsymbol{v}$. Here *m* is the mass of the planet.

So the rate at which area is drawn out is just

$$\frac{dA}{dt} = |\frac{1}{2}\boldsymbol{r} \times \boldsymbol{v}|$$

Let's consider the time derivative of the vector $\boldsymbol{r} \times \boldsymbol{v}$, the angular momentum per unit mass.

$$\frac{d(\boldsymbol{r} \times \boldsymbol{v})}{dt} = \frac{d\boldsymbol{r}}{dt} \times \boldsymbol{v} + \boldsymbol{r} \times \frac{d\boldsymbol{v}}{dt}$$

But we can replace the derivatives on the right with the velocity and acceleration

$$\frac{d(\boldsymbol{r} \times \boldsymbol{v})}{dt} = \boldsymbol{v} \times \boldsymbol{v} + \boldsymbol{r} \times \boldsymbol{a}$$

The cross-product of parallel vectors is 0. So clearly the first term is 0, since any vector is parallel with itself. If the acceleration is due to a central force like gravity, then the acceleration is parallel to the radius vector, and thus the second term is also 0. So

$$\frac{d(\boldsymbol{r} \times \boldsymbol{v})}{dt} = 0$$

Angular moment is conserved and the area drawn out in an interval is proportional to the length of the interval.

Deriving Kepler's First Law[i]

We start with the force equation in radial coordinates we derived above

$$\frac{d^2r}{dt^2} - r\omega^2 = -\frac{GM}{r^2}$$

describing the acceleration of our planet. To solve this we need to make use of the constancy of angular moment and make a simple substitution.

The rate of change of the polar angle θ can be derived by our equation for angular momentum

$$\boldsymbol{L} = m\boldsymbol{r} \times \boldsymbol{v} = m\boldsymbol{r} \times r\frac{d\theta}{dt}\widehat{\boldsymbol{\theta}}$$

since only the component of the velocity perpendicular to the radius contributes to the angular moment. Since $\boldsymbol{r}$ and $\widehat{\boldsymbol{\theta}}$ are perpendicular and the second is a unit vector we can transform this into a scalar equation

$$L = mr^2\frac{d\theta}{dt}$$

where L is the magnitude of the angular momentum.

Or

$$\frac{d\theta}{dt} = L/(mr^2)$$

The chain rule in differential calculus says that if we have any function, f, we can express

$$\frac{df}{dt} = \frac{df}{d\theta}\frac{d\theta}{dt} = L/(mr^2)\frac{df}{d\theta} = \frac{Lu^2}{m}\frac{df}{d\theta}$$

so long as θ is a function of time over the appropriate intervals.

In the last step above we made the substitution $u \equiv 1/r$. Using that substitution we can express the change in r with time as

$$\frac{dr}{dt} = \frac{d(\frac{1}{u})}{dt} = -\frac{1}{u^2}\frac{du}{dt} = -\frac{1}{u^2}\frac{Lu^2}{m}\frac{du}{d\theta} = -\frac{L}{m}\frac{du}{d\theta}$$

where we used the constancy of angular momentum to change from a derivative with respect to time to one with respect to angle.

Now L and m are constants, so if we want $\frac{d\frac{dr}{dt}}{dt} = \frac{d^2r}{dt^2}$ it operates only on the $\frac{du}{d\theta}$ term in our rightmost expression and we can use the chain rule again to get

$$\frac{d^2r}{dt^2} = \frac{d(-\frac{L}{m}\frac{du}{d\theta})}{dt} = -\frac{L^2u^2}{m^2}\frac{d^2u}{d\theta^2}$$

We'll use this in place of the first term in our force equation.

Going back to the force equation, the second term is $-r\omega^2$, but this is just $r\frac{d\theta}{dt}^2$. Using the angular momentum to substitute for $\frac{d\theta}{dt}$ we can transform this to $r\frac{d\theta}{dt}^2 = r\frac{L^2}{m^2r^4} = \frac{L^2}{m^2r^3} = \frac{L^2u^3}{m^2}$

The last term is just $\frac{GM}{r^2} = GMu^2$.

Substituting back our transformed terms into the force equation we have

$$-\frac{L^2u^2}{m^2}\frac{d^2u}{d\theta^2} - \frac{L^2u^3}{m^2} = -GMu^2$$

We can divide out the $-u^2$ in each term, and the $\frac{L^2}{m^2}$ factors on the left side of the equation to get

$$\frac{d^2u}{d\theta^2} + u = \frac{GMm^2}{L^2}$$

We need some function where $\frac{d^2f}{d\theta^2} = -f$ to can solve our equation. Suppose we have such a solution f. Then if we define a second function $g = f + C$ where C is a constant. Clearly $\frac{d^2g}{d\theta^2} + g = C$. The constant term disappears when we take the derivative so it doesn't get cancelled by the second derivative term.

There are a whole set of functions that satisfy our conditions for f. Any function of the form $f = A\cos(\theta + C)$ works, where *A* and *C* are constants. Or we could use a sine function equally well, but that's really just equivalent to adding -90° to *C*. *C* effects a rotation of the system: it will be related to the argument of perihelion while *A* will be associated with the eccentricity and energy of the orbit.

We may briefly panic when we realize that we could have a more complex form for f, say $f = A\cos(\theta + C) + B\sin(\theta + D)$+… which would give us a potentially infinite number of parameters to determine. If we have two or more solutions for f then any sum of solutions is also a solution. However the laws for the addition of sines and cosines come to our aid here. If we just take the first two terms in our example we have

$$f = A\cos(\theta + C) + B\sin(\theta + D)$$

$$f = A\cos\theta\cos C - A\sin\theta\sin C + B\sin(\theta)\cos D + B\cos\theta\sin D$$

We collect all the terms for $\cos\theta$ and $\sin\theta$ to get

$$f = \cos\theta\,(A\cos C - B\sin D) + \sin\theta(-A\sin C + B\cos D)$$

We define two new constants P and Q as the coefficients rather than writing them out.

$$f = P\cos\theta + Q\sin\theta = \sqrt{P^2 + Q^2}\left(\frac{P}{\sqrt{P^2+Q^2}}\cos\theta + \frac{Q}{\sqrt{P^2+Q^2}}\sin\theta\right)$$

Now if we consider the angle δ given by $\delta = \tan^{-1}(Q/P)$, then $\frac{P}{\sqrt{P^2+Q^2}} = \cos\delta$ and $\frac{Q}{\sqrt{P^2+Q^2}} = \sin\delta$. We use the law for addition of cosines in the other direction to get $f = \sqrt{P^2 + Q^2}\cos(\theta - \delta)$.[ii]

While we only did this for two terms, no matter how many terms we initially think we have, they can all be collected into a single expression of our original form so we only have to worry about determining two constants of integration.

Let's restrict ourselves to solutions with foci along the x-axis. Then we just use the $A\cos(\theta)$ solution and we get

$$u = A\cos(\theta) + \frac{GMm^2}{L^2}$$

Or

$$r = \frac{1}{\mathrm{A}\cos(\theta) + \frac{GMm^2}{L^2}}$$

If we multiply top and bottom by $\frac{L^2}{GMm^2}$ we get

$$r = \frac{\frac{L^2}{GMm^2}}{\frac{L^2}{GMm^2}A\cos(\theta) + 1}$$

We can see that the same basic form as we had for an ellipse that we derived above. We have an equation where the orbit is an ellipse (or a hyperbola if the constants multiplying the cosine factor are greater than 1).

Deriving Kepler's Third Law

Kepler's third law states that the period of the orbit is proportional to the square root of the cube of the semimajor axis: $Period \backsim a^{\frac{3}{2}}$. This follows almost instantly for circular orbits from our force equation:

$$\frac{d^2r}{dt^2} - r\omega^2 = -\frac{GM}{r^2}$$

For a circular orbit the radius is constant and we just have

$$r\omega^2 = \frac{GM}{r^2}$$

But ω (in radians/second) is just $\omega = \frac{2\pi}{Period}$ hence $Period = 2\pi\sqrt{\frac{a^3}{GM}}$. In a circular orbit the radius is always the semimajor axis, a.

In the more general case of an elliptical orbit we need to do a bit more work. We know the rate that area is being swept out in the ellipse, and that the area of an ellipse is simply $A = \pi ab$. To get the period we need to divide the area by the rate of change. First let's express the semiminor axis in terms of the eccentricity and the semimajor axis. At the points on the ellipse on the semiminor axis we are equidistant from both foci, so the distance to each focus is *a* (since the sum of the distances must be 2*a*). We draw a right triangle with the point on the semimajor axis, the center of the ellipse and one of the foci.

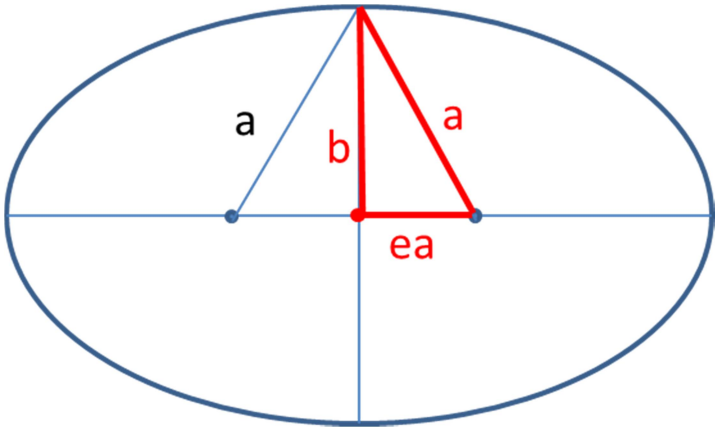


Figure 33. Finding the semiminor axis.

So we have $a^2 = (ea)^2 + b^2$. (Recall that the eccentricity was defined as the ratio of the distance from the center to a focus to the semimajor axis.) Hence $b = \sqrt{a^2 - (ea)^2} = a\sqrt{1 - e^2}$. So the area of the ellipse is $Area = \pi a^2\sqrt{1 - e^2}$.

We have the equation for an ellipse in polar coordinates:

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

and our solution for the orbit

$$r = \frac{\frac{L^2}{GMm^2}}{\frac{L^2}{GMm^2}A\cos(\theta) + 1}$$

When $\theta = 90°$, the cosines are 0 and the denominators are just 1. So combining these gives

$$a(1 - e^2) = \frac{L^2}{GMm^2}$$

We won't use this anytime soon but it's worth noting this that this gives us an expression for the magnitude of the angular momentum

$$\frac{L}{m} = \sqrt{GMa(1 - e^2)}$$

Recall the rate of area being swept out is $\frac{dA}{dt} = \frac{L}{2m}$. Using the equation above we have

$$\frac{dA}{dt} = \frac{L}{2m} = \frac{1}{2}\sqrt{GMa(1 - e^2)}$$

So

$$Period = \frac{A}{\frac{dA}{dt}} = \frac{\pi a^2\sqrt{1-e^2}}{\frac{1}{2}\sqrt{GMa(1-e^2)}} = 2\pi\sqrt{\frac{a^3}{GM}}$$

This is exactly the same as we got previously for the circular orbit. So in a given system (i.e., a set of orbits around some single massive body M), the period depends only on the major axis length. It is completely independent of the eccentricity of the orbits.

Relations of Orbital Variables and Elements

We've derived Kepler's laws of planetary motion from Newton's theory of gravity. We won't be needing to use calculus for a bit, but we want to understand how all of the myriad quantities that we've been taking about relate to one another: How do we find the position and velocity of a planet if we know its orbital elements, or how do we find the orbital elements if we know the position and velocity? What are the explicit dependencies of the semimajor axis on the energy and the eccentricity on the angular momentum?

Basically we want to compile a network of equations that allow us to express various quantities in terms of others so that we understand what we can infer from what we already know. We'll want to see how we can derive the orbital elements for physical quantities like energy and angular momentum, and from the state of the system at a particular time, i.e., the position and velocities at some reference time. We also want to be able to go the other way, and show how we can determine the positions and velocities as given the orbital elements. In discussing these transformation we'll also note where appropriate how the formulae are affected by choice of coordinate system, the notional coordinates where the mass of the second body is negligible, the center of mass coordinates where both the body and the primary move, and the total vector approach where we look at the evolution of the separation of the bodies.

Energy and Radius

The energy of the orbit is the amount of work we can get out of the system, i.e., that the bodies in the orbit could do at some location distant from the center of mass. For a bound orbit, this is negative, because we need to add energy to the system to even get the bodies far away from the center of mass. The energy comprises three elements: the negative gravitational binding energy of the pair of objects, and the positive kinetic energy of both objects involved in the orbit. When the primary is much more massive, then we might describe the potential energy as the potential energy of the orbiting body alone: we ascribe a total energy to the orbiting particle alone, but more generally we need to consider the energy as a property of the entire system.

If we assume that m is negligible compared to M, so that the massive body is stationary, then the conservation of angular momentum means that positions and velocities are perihelion and aphelion are related by

$$r_{peri}v_{peri} = r_{aph}v_{aph}$$

since the velocity is perpendicular to the velocity when we cross the major axis. But we can express both of the radii in terms of the semimajor axis and the eccentricity as

$$r_{peri} = (1-e)a$$

$$r_{ap} = (1+e)a$$

So

$$v_{aph} = \frac{1-e}{1+e}v_{peri}$$

Energy is conserved, so the energy at perihelion and aphelion is the same and we get

$$-\frac{GMm}{(1-e)a} + \frac{1}{2}m{v_{peri}}^2 = -\frac{GMm}{(1+e)a} + \frac{1}{2}m(\frac{1-e}{1+e}v_{peri})^2$$

Collecting terms and canceling out the factors of m we have

$$\frac{1}{2}{v_{peri}}^2(1-\left(\frac{1-e}{1+e}\right)^2) = \frac{GM}{a}(\frac{1}{1-e} - \frac{1}{1+e})$$

$$\frac{1}{2}{v_{peri}}^2\frac{4e}{(1+e)^2} = \frac{GM}{a}\frac{2e}{(1-e)(1+e)}$$

So we get the perihelion velocity as

$${v_{peri}}^2 = \frac{GM}{a}\frac{1+e}{1-e}$$

As we might expect if the eccentricity is 0, then the velocity of the orbit is

$$v = \sqrt{\frac{GM}{a}}$$

In this case the magnitude of the kinetic energy is exactly half the potential energy. To get the energy generally, we use the values for the perihelion radius and velocity and find the total energy for the orbit is

$$Energy = -\frac{GMm}{(1-e)a} + \frac{1}{2}m{v_{peri}}^2 = -\frac{GMm}{(1-e)a} + \frac{1}{2}\frac{GMm}{a}\frac{1+e}{1-e}$$

$$= -\frac{GMm}{a}\left(\frac{1}{1-e} - \frac{1+e}{2(1-e)}\right)$$

$$= -\frac{GMm}{2a}\left(\frac{2 - \ 1 - e}{1 - e}\right)$$

$$= -\frac{GMm}{2a}$$

The energy is independent of the eccentricity and depends only on the mass of the central body and the semimajor axis.

Let's illustrate how we might use the reduced mass or effective mass approaches.

We recall that the reduced mass approach stems from the relation

$$\frac{mM}{m+M}\frac{d^2\boldsymbol{r}}{dt^2} = m\frac{d^2\boldsymbol{r}_m}{dt^2} = -\frac{GmM}{|\boldsymbol{r}|^2}\hat{\boldsymbol{r}}$$

which we simplify to

$$\frac{d^2\boldsymbol{r}}{dt^2} = -\frac{G(m+M)}{|\boldsymbol{r}|^2}\hat{\boldsymbol{r}}$$

Recall how above we transformed the vector force equation to an equation for the polar components of the position vector. In that example we used $M$where we now have $(m+M)$. So using the same approach we find

$$\frac{d^2r}{dt^2} - r\omega^2 = -\frac{G(m+M)}{r^2}$$

Let's consider the specific example of a circular orbit where $m = M$. Then the derivative drops out and we have

$$r\omega^2 = \frac{2GM}{r^2}$$

For the circular orbit $\omega r = v$ so

$$\frac{v^2}{r} = \frac{2GM}{r^2}$$

Or

$$v = \sqrt{\frac{2GM}{r}}$$

So the total velocity increases by a factor or $\sqrt{2}$ from what we had were the mass of the secondary was negligible. However when we look at the velocity of just one of the bodies with respect to the center of

mass, then that velocity is half the total velocity. So if we use v_m to denote the velocity of one of the bodies we have $v_m = \sqrt{\frac{GM}{2r}}$. The velocity of an individual particle decreases by $\sqrt{2}$.

If we use the effective gravitational mass approach we work directly on the distance from the center of mass of one of the bodies

$$\boldsymbol{F}_m = -\frac{Gm\frac{M^3}{(m+M)^2}}{|\boldsymbol{r}_m|^2}\frac{\boldsymbol{r}_m}{|\boldsymbol{r}_m|}$$

If we just look at the acceleration of the m body then we have

$$\frac{d^2\boldsymbol{r}_m}{dt^2} = -\frac{G\frac{M^3}{(m+M)^2}}{|\boldsymbol{r}_m|^2}\hat{\boldsymbol{r}}_m$$

We proceed as above except that now the ratio $\frac{M^3}{(m+M)^2}$ plays the role of then central mass. For the circular orbit we have

$$r_m\omega^2 = \frac{{v_m}^2}{r_m} = \frac{G\frac{M^3}{(m+M)^2}}{{r_m}^2}$$

We do not need to subscript the ω term since both bodies have the same angular velocity.

If $m = M$ then $v_m = \sqrt{\frac{GM}{4r_m}}$. Since the bodies are of equal mass we know that $r_m = \frac{r}{2}$ and we get the same result as above $\sqrt{\frac{GM}{4r_m}} = \sqrt{\frac{GM}{2r}}$.

This result may seem somewhat counterintuitive. If we think of a series of circular orbits where the separation between the bodies is the same, then we can start where the mass, m, of the smaller body is negligible and then grow that mass. This result shows that the velocity gets smaller even though the forces involved have gotten greater. Although the forces involved have grown, the masses have grown correspondingly so that the effect on accelerations is not obvious. Consider the limit as m gets very large. The body whose motion we are following becomes stationary as the other body orbits around it. So the net effect of adding mass to just one body in a two body system with a circular orbit is that the increasing mass moves slower.

We can compute the total energy of the equal-mass, circular-orbit system as the potential energy plus the kinetic energies of the two bodies – which are equal. So

$$E = -\frac{GM^2}{r} + 2\left(\frac{1}{2}M(\sqrt{\frac{GM}{2r}})^2\right) = -\frac{GM^2}{2r}$$

essentially the same form as we had when the secondary mass was negligible. There is a general result in dynamics, the Virial Theorem (which we will *not* be deriving!), which states that in a gravitationally bound system the kinetic energy averaged over time will be half the average gravitational potential energy (but of the opposite sign). Our results for two body orbits are examples of the Virial Theorem in action.

In the reduced mass and effective mass computations, if we solve for the angular velocity of the circular orbit, dropping the second derivative from the force equation, we get using the reduced mass

$$r\omega^2 = -\frac{G(m+M)}{r^2}$$

$$\omega = \sqrt{-\frac{G(m+M)}{r^3}r}$$

or for the center of mass computation for one body

$$r_m\omega^2 = \frac{G\frac{M^3}{(m+M)^2}}{{r_m}^2}$$

$$\omega = \sqrt{\frac{G\frac{M^3}{(m+M)^2}}{{r_m}^3}}$$

If $m = M$, then $r_m = \frac{1}{2}r$ since the center of mass is midway between the two bodies. Then both of these are equivalent to $\omega = \sqrt{\frac{2G}{r^2}}$ or $\omega = \sqrt{\frac{GM}{2{r_m}^2}}$. Compared to the orbit where m was negligible, the angular velocity is increased by $\sqrt{2}$ or equivalently the period is decreased by that same factor.

In the discussion where m is not negligible we have been using the word radius rather than semimajor axis since it makes the relationship clearer as we switch between different coordinate system. However for circular orbits r_m would be the real semimajor axis for the orbit of one of the bodies measured in center of mass coordinates and r is the semimajor axis of the orbit of the full distance vector.

Angular Momentum and Eccentricity

While the energy determines the size and period of the orbit, the angular momentum determines its shape and orientation. The angular momentum of an orbiting body is just

$$\boldsymbol{L} = m\boldsymbol{r} \times \boldsymbol{v}$$

If we have two bodies with significant masses in the system then

$$\boldsymbol{L} = m\boldsymbol{r}_m \times \boldsymbol{v}_m + M\boldsymbol{r}_M \times \boldsymbol{v}_M$$

Unlike the situation with energy, it's reasonable to ascribe separate angular momenta to the two bodies since they each have well defined mass, position and velocity.

In the center of mass framework the angular momentum of one body is

$$\boldsymbol{L}_m = m\boldsymbol{r}_m \times \boldsymbol{v}_m$$

and the total angular momentum is

$$\boldsymbol{L} = m\boldsymbol{r}_m \times \boldsymbol{v}_m + \frac{m^2}{M}\boldsymbol{r}_m \times \boldsymbol{v}_m$$

$$\boldsymbol{L} = m\frac{m+M}{M}\boldsymbol{r}_m \times \boldsymbol{v}_m$$

If we are using the total separation and velocity vectors then we have

$$\boldsymbol{L}_m = m\frac{M}{M+m}\boldsymbol{r} \times \frac{M}{M+m}\boldsymbol{v} = \frac{mM^2}{(M+m)^2}\boldsymbol{r} \times \boldsymbol{v}$$

The total angular momentum of the system is

$$\boldsymbol{L} = \left[m\frac{M^2}{(M+m)^2} + M\frac{m^2}{(M+m)^2}\right]\boldsymbol{r} \times \boldsymbol{v}$$

$$\boldsymbol{L} = \frac{mM}{M+m}\,\boldsymbol{r} \times \boldsymbol{v}$$

Unlike the radius and velocity, the angular momenta of both bodies are in the same direction, since they are orbiting in the same direction in the same plane. The ratio of angular momenta of the two bodies is just the ratio of the masses, with the lighter body having most of the angular momentum, since the heavier body is both closer to the origin and moving slower both by the ratio of the masses.

In addition to the angular momentum of their orbits, the spin of planets and the Sun provides angular momentum in the solar system. Tidal interactions can transfer angular momentum (and energy) between planetary rotation and satellite orbits. Generally satellites which whose revolution period is less than the day of the planet they orbit lose angular momentum from their orbits, while those with a revolution period greater than the day gain angular momentum. Thus Mars' fast orbiting satellite Phobos, which orbits the planet three times a day, may ultimately crash to the Martian surface, while

the Earth's Moon, whose period of revolution is almost 30 times the Earth's rotation period, is slowly gaining angular momentum and moving to a more distant orbit.

The angular momentum determines the plane of the orbit of the satellite. We can very easily determine the inclination and the longitude of the ascending node from the angular momentum.

Orientation of the orbit plane

Consider the cross-product between the unit vector in the z direction and the angular momentum.

$$\boldsymbol{l} = \hat{\boldsymbol{z}} \times \boldsymbol{L}$$

The resulting vector, $\boldsymbol{l}$, is perpendicular to the angular momentum vector, so it is in the plane of the orbit. It is also perpendicular to the pole, so it is in the equator. If it is in both the plane and the equator, then it must point along the line of nodes, and indeed it points to the ascending node. Since this is in the equatorial plane, in Cartesian coordinates it must look like $\boldsymbol{l} = (x, y, 0)$ with no z component. The longitude of the ascending node is simply

$$\Omega = \text{atan2}(\boldsymbol{l_y}, \boldsymbol{l}_x)$$

where we use the two argument version of the arctangent as defined in many programming languages (note the order of arguments).

The magnitude of a cross-product is related to the magnitudes of the input vectors by

$$|\boldsymbol{l}| = |\boldsymbol{L}||\hat{\boldsymbol{z}}| \sin(i)$$

where i is the angle between the two input vectors. If i is 0, then the angular momentum is pointing at a pole and the orbit is in the equatorial plane. If i is 90 degrees, then the angular momentum is pointing perpendicular to the pole, and the object must be orbiting in a plane perpendicular to the equator. The dot product of the angular momentum and the $\hat{\boldsymbol{z}}$ unit vector is proportional to the product of the magnitudes of the vectors and the *cosine* of the angle between them. So

$$i = \cos^{-1}(\frac{L_z}{|\boldsymbol{L}|})$$

Using this approach we measure the inclination in the range 0-180°. Retrograde orbits have inclinations greater than 90°.

Note that one can have inclinations less than zero, but when a negative inclination is encountered, conventionally it is replaced by its absolute value and the 180° is added to the longitude of the ascending node.

The eccentricity vector

An extremely useful constant of the orbit is the eccentricity vector. This quantity allows us to quickly point to the perihelion of the orbit and to calculate its eccentricity.

Consider the vector given by

$$\boldsymbol{e} = \frac{\boldsymbol{v} \times \frac{\boldsymbol{L}}{m}}{GM} - \hat{\boldsymbol{r}}$$

This is defined as the eccentricity vector. The second term in the cross-product is the angular momentum per unit mass.

First let's consider what the value is at perihelion. The vectors in the cross-product are always perpendicular and at perihelion the velocity and the angular momentum are both perpendicular to $\hat{\boldsymbol{r}}$ so their cross-product must be parallel to that unit vector. So we can just use the scalar values to get the value at perihelion as

$$\boldsymbol{e}_{peri} = (\frac{v_{peri}\frac{L}{m}}{GM} - 1)\hat{\boldsymbol{r}}_{peri}$$

Here v_{peri} and r_{peri} are the values at perihelion. The unit vector $\hat{\boldsymbol{r}}_{peri}$ is the then current unit vector that points to perihelion.

Previously we calculated the velocity at perihelion as

$$v_{peri} = \sqrt{\frac{GM}{a}\frac{1+e}{1-e}} = \sqrt{\frac{(1+e)GM}{(1-e)a}} = \sqrt{\frac{(1+e)GM}{r_{peri}}}$$

But we also have

$$L/m = r_{peri}v_{peri}$$

So

$$\boldsymbol{e}_{peri} = (\frac{{v_{peri}}^2 r_{peri}}{GM} - 1)\hat{\boldsymbol{r}}_{peri}$$

or substituting for the perihelion velocity we have

$$\boldsymbol{e}_{peri} = (\frac{\frac{(1+e)GM}{r_{peri}}r_{peri}}{GM} - 1)\hat{\boldsymbol{r}}_{peri}$$

$$\boldsymbol{e}_{peri} = e\hat{\boldsymbol{r}}_{peri}$$

So at perihelion we know that this vector points in the direction of perihelion and has a magnitude equal to the eccentricity of the orbit. Next we prove that it is constant over the orbit.[iii]

Let's define $\boldsymbol{h} = \frac{\boldsymbol{L}}{m}$. This is the specific angular momentum which is still constant over the orbit. The time derivative of the first term in the eccentricity vector is

$$\frac{d(\boldsymbol{v}\times\boldsymbol{h})}{dt}=\frac{d\boldsymbol{v}}{dt}\times\boldsymbol{h}=\frac{d^2\boldsymbol{r}}{dt^2}\times\boldsymbol{h}$$

where the time derivative of $\boldsymbol{h}$ is 0 since it is a constant.

Substituting in the gravitational force we have

$$\frac{d(\boldsymbol{v}\times\boldsymbol{h})}{dt}=-\frac{GM}{r^2}\hat{\boldsymbol{r}}\times\boldsymbol{h}$$

$$=-\frac{GMh}{r^2}\hat{\boldsymbol{r}}\times\hat{\boldsymbol{z}}$$

where we factored out the magnitude of the angular momentum since we know that that points in the z direction in our cylindrical coordinate system. By the right-hand rule the unit vector cross-product is just

$$\hat{\boldsymbol{r}}\times\hat{\boldsymbol{z}}=\ -\widehat{\boldsymbol{\theta}}$$

in our cylindrical coordinates system.

So

$$\frac{d(\boldsymbol{v}\times\boldsymbol{h})}{dt}=\frac{GMh}{r^2}\ \widehat{\boldsymbol{\theta}}$$

Now we can use the our relation for the angular momentum $h=r^2\frac{d\theta}{dt}$ to simplify to

$$\frac{d(\boldsymbol{v}\times\boldsymbol{h})}{dt}=GM\frac{d\theta}{dt}\ \widehat{\boldsymbol{\theta}}$$

When we first started to consider derivatives in radial coordinates we noted

$$\frac{d\hat{\boldsymbol{r}}}{dt}=-\sin\theta\frac{d\theta}{dt}\ \widehat{\boldsymbol{x}}+\cos\theta\frac{d\theta}{dt}\widehat{\boldsymbol{y}}=\ \frac{d\theta}{dt}\widehat{\boldsymbol{\theta}}$$

So we just have

$$\frac{d(\boldsymbol{v}\times\boldsymbol{h})}{dt}=GM\frac{d\hat{\boldsymbol{r}}}{dt}$$

If we collect terms and divide by GM we get

$$\frac{1}{GM}\frac{d(\boldsymbol{v}\times\boldsymbol{h})}{dt}-\frac{d\hat{\boldsymbol{r}}}{dt}=0$$

And since differentiation is linear we can combine things in the derivative to get our desired result

$$\frac{d\left(\frac{1}{GM}\boldsymbol{v}\times\boldsymbol{h}-\hat{\boldsymbol{r}}\right)}{dt}=0$$

So the eccentricity vector is a constant over the orbit and must always have the value it has at perihelion.

If we wish to consider an orbit with a planet where the mass of the planet is significant, then using the reduced mass or effective gravitational mass, the GM term should be replaced by $G(M+m)$ or $\frac{GM^3}{(M+m)^2}$ respectively.

Since the eccentricity vector can be calculated directly from the position and velocity vectors it means that we can very quickly find almost all of the standard orbital elements from the position and velocity (assuming we know the mass of the primary). We use the angular momentum to calculate the line of nodes vector, and then we can use the dot and cross products of the line of nodes and the eccentricity vectors to calculate the argument of perihelion.

If $\boldsymbol{l}$ is the line of nodes vector and $\boldsymbol{e}$ is the eccentricity vector then we have

$$|\sin \nu| = \frac{|\boldsymbol{l} \times \boldsymbol{e}|}{|\boldsymbol{l}|\,|\boldsymbol{e}|}$$

and

$$\cos \nu = \frac{|\boldsymbol{l} \cdot \boldsymbol{e}|}{|\boldsymbol{l}|\,|\boldsymbol{e}|}$$

To resolve the ambiguity of the sine we look at the angle between the cross-product $\boldsymbol{l} \times \boldsymbol{e}$ and the angular momentum which defines the orientation of the orbit. The vectors $\boldsymbol{l} \times \boldsymbol{e}$ and the angular momentum, $\boldsymbol{L}$, should be parallel or antiparallel since $\boldsymbol{l}$ and $\boldsymbol{e}$ are in the plane of the orbit. When they are parallel then the sine is positive, when antiparallel, the sine is negative.

This allows us to fully define the argument of perihelion, ν.

The radius and velocity can be easily translated into an energy and thence to a semimajor axis. So we can easily get all the parameters that define the size, shape and orientation of the orbit. All that's left is the phase.

The anomalies

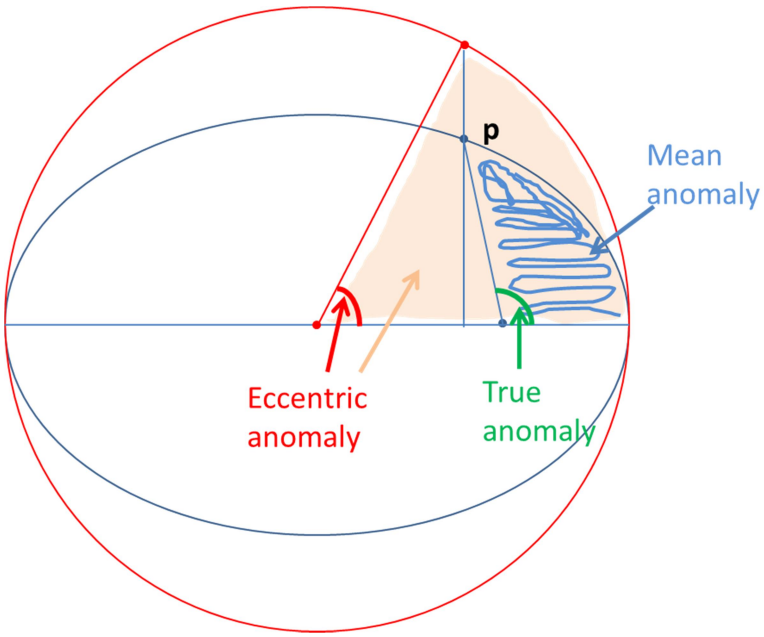


Figure 34. The three anomalies

The figure illustrates the three anomalies we use in describing the phase of the orbit. All of these measure from the last perihelion passage. For a hyperbolic orbit, the single perihelion passage is when the anomalies are 0.

We have a blue elliptical orbit with the Sun at the blue dot focus. The center of the ellipse is at the red dot. The blue ellipse is circumscribed by a red circle with radius equal to the semimajor axis, so the two figures are tangent along the major axis. The true anomaly is the green angle between the vector to perihelion and the current position, **p**, of the orbiting body where we measure the angle from the focus of the ellipse.

The mean anomaly is the area that has been swept out by the orbit since the last perihelion passage. As we have seen above, since angular momentum is conserved it is linearly related to time. Normally we will scale the mean anomaly by the total area of the ellipse so that we measure the mean anomaly from 0 to 360 degrees. The mean anomaly is shown as the blue scribbled region on the diagram.

The eccentric anomaly is defined by projecting the point at **p** onto the circumscribing circle along a line perpendicular to the major axis. The blue dot at **p** maps to the red dot on the circle. We then define the eccentric anomaly as the red angle measuring the angle between the perihelion vector and the projected point but using the center of the ellipse, not the focus as with the true anomaly. Equivalently since we are dealing with a circle, the eccentric anomaly is the fraction of the area of the circle swept out by the projected point, shown shaded tan.

Kepler's Equation: Mean anomaly/time and Eccentric anomaly

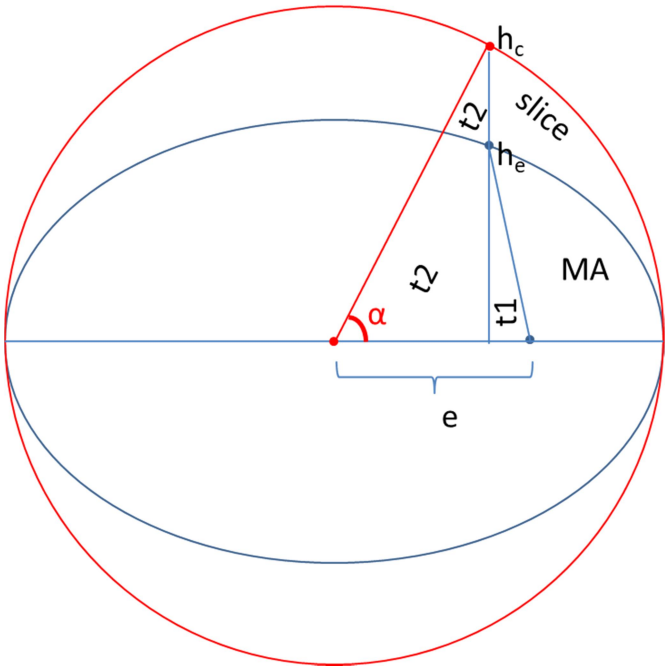


Figure 35. Calculating the eccentric anomaly

We've redrawn the previous figure highlighting the elements relating to how we get from the mean anomaly to the eccentric anomaly. At this point we wouldn't know what the true anomaly is, so we don't use that. The red angle, α, is the angular measurement of the eccentric anomaly. However if we add together four labelled areas, *MA* (the region that was the mean anomaly), *t1*, *t2* and *slice*, we'd also have the area associated with the eccentric anomaly. Note that *t2* includes the area inside and outside of the ellipse.

If the true anomaly were less than 90° (or greater than 270°), then we might have a situation like the next figure.

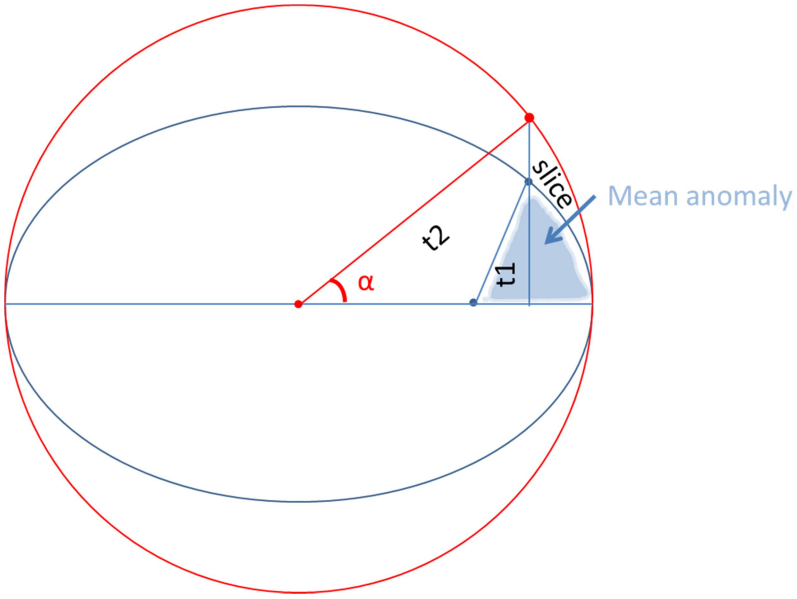


Figure 36. Eccentric anomaly near perihelion

In this situation, near perihelion, then we want to subtract the triangle t1 from the mean anomaly area, since it represents the area in the mean anomaly that's also in triangle t2.

We'll normalize the ellipse to a semimajor of 1. Then the distance between the center of the ellipse and the focus is simply *e.* In the first figure we've also marked two points, h_c and h_e. We use these to represent the distance of our points on the circle and ellipse from the major axis. We'll see that it's easy to compute the area of the two triangles, and we 'know' the mean anomaly, so the only area that is an issue is the slice area. However we don't have to compute that directly. We can transform the ellipse to the sphere, by simply multiplying the y value of each point in the ellipse by the ratio of the major axis to the minor axis. If we do this for the MA + t1 area (or MA-t1 if we are near perihelion), it will be transformed into the area MA+t1+slice. But since this is just a linear transformation in one dimension it simply increases the area by the ratio of major to the minor axis.

The circle is the unit circle, so the coordinates of the point on the circle are just $[x, y] = [\cos\alpha, \sin\alpha]$. The triangle t1 has a base of $\cos\alpha$ and an altitude of $\sin\alpha$.

$$Area_{t2} = \frac{1}{2}\cos\alpha\sin\alpha$$

The altitude for triangle t1 is h_e, but this is reduced by the ratio of the axes from the h_c. We just saw that that is $h_c = \sin\alpha$. Similarly, the base is just the difference between eccentricity and the base for t2. So base of t1 is just $e - \cos\alpha$. So

$$Area_{t1} = \frac{1}{2}(e - \cos\alpha)\frac{b}{a}\sin\alpha = \frac{b}{2a}e\sin\alpha - \frac{b}{2a}\sin\alpha\cos\alpha$$

Note than when the eccentric anomaly α is small the base of the triangle, $(\mathrm{e} - \cos\alpha)$, will be negative so when we 'add' the area for t1 in, we will actually be subtracting it, which is exactly what we want. We've left the *a*'s in the denominator, but since we've set *a* to 1, we could just drop them.

Now we need to be careful to scale the area represented by the mean anomaly properly. The mean anomaly in radians is $MA = 2\pi \frac{area\ swept\ out}{area\ of\ ellipse}$. In our case were we have *a*=1, the area of the ellipse is

$$area = \pi ab = \pi b.$$

The area in the region MA in our figure then is $\frac{MA}{2\pi} area\ of\ ellipse = area\ swept\ out.$

So the area in the region MA in the figure is related to our the measured mean anomaly of the orbit, MA_{input}, by

$$MA_{area} = MA_{input} \frac{\pi b}{2\pi} = \frac{b}{2} MA_{input}$$

So now we can scale the area inside the ellipse but outside the triangle t2, to the full circle

$$Area_{outside\ t2} = \frac{1}{b}\left(\frac{b}{2} MA_{input} + Area_{t1}\right)$$

So the total area of

$$TotalArea = \frac{1}{b}\left(\frac{b}{2} MA_{input} + \frac{b}{2a} e \sin\alpha - \frac{b}{2}\sin\alpha\cos\alpha\right) + \frac{1}{2}\cos\alpha\sin\alpha$$

$$TotalArea = \frac{1}{2} MA_{input} + \frac{1}{2} e \sin\alpha - \frac{1}{2}\sin\alpha\cos\alpha + \frac{1}{2}\cos\alpha\sin\alpha$$

$$TotalArea = \frac{1}{2} MA_{input} + \frac{1}{2} e \sin\alpha$$

Here the total area is the areal representation of the eccentric anomaly. We want the full circle to represent 2π radians, so the we have $\alpha = 2\pi \frac{TotalArea}{Area\ of\ circle} = 2\ TotalArea$. Hence we get

$$\frac{\alpha}{2} = \frac{1}{2} MA_{input} + \frac{1}{2} e \sin\alpha$$

Or more familiarly

$$MA = \alpha - e \sin\alpha$$

This is Kepler's equation giving the relationship between the mean and eccentric anomalies. As we had intuited when we looked at this qualitatively, the fact that we can treat the eccentric anomaly as both an area and angle played a significant role in our derivation.

Given the eccentric anomaly it is trivial to get the mean anomaly, but going the other way is more difficult and is usually done using some iterative process.

Kepler's Equation: Mean anomaly/time and hyperbolic anomaly

The derivation of the relationship between the mean anomaly and the hyperbolic anomaly for unbound orbits proceeds in much the same fashion as it did for the eccentric anomaly in bound orbits. Note the role of the unit circle in the derivation above for ellipses. The unit circle is defined by $x^2 + y^2 = 1$ and we can express the x and y values as $(x, y) = (\cos\theta\,, \sin\theta)$. For hyperbolic orbits we use a corresponding 'unit' hyperbola $x^2 - y^2 = 1$. Since the hyperbolas, unlike circles, have a manifest direction there is more than one 'unit' hyperbola possible. The one we've chosen is the one where the major axis of the hyperbola is along the X-axis, so that the hyperbola opens out to the right and left. We'll only consider the right branch of the hyperbola.

Consider the not coincidentally named hyperbolic sine and cosine functions.

$$\sinh\theta = \frac{e^{\theta} - e^{-\theta}}{2}$$

$$\cosh\theta = \frac{e^{\theta} + e^{-\theta}}{2}$$

If we let $(x, y) = (\cosh\theta\,, \sinh\theta)$ then

$$x^2 - y^2 = \frac{e^{2\theta} + 2e^{\theta-\theta} + e^{-2\theta} - e^{2\theta} + 2e^{\theta-\theta} - e^{-2\theta}}{4} = 1$$

The coordinates along the unit hyperbola can be represented as if they are a function of a hyperbolic angle θ, in just the same way that the coordinates of the unit circle can be parametrized by a circular angle. Note that the angle used in the hyperbolic functions is not bounded as it is for the circular trigonometric functions, the hyperbolic sine and cosine are not periodic. The hyperbolic sine ranges over all real values and the hyperbolic cosine takes on any value greater than or equal to 1.

Relating two definitions of the hyperbolic anomaly

We need to establish a key characteristic of the unit hyperbola. Consider the shaded area in the figure below. It's a trapezoidalish figure with one of the sides bound by the hyperbola. The area includes both the area we'd earlier defined as the hyperbolic anomaly and an additional triangle.

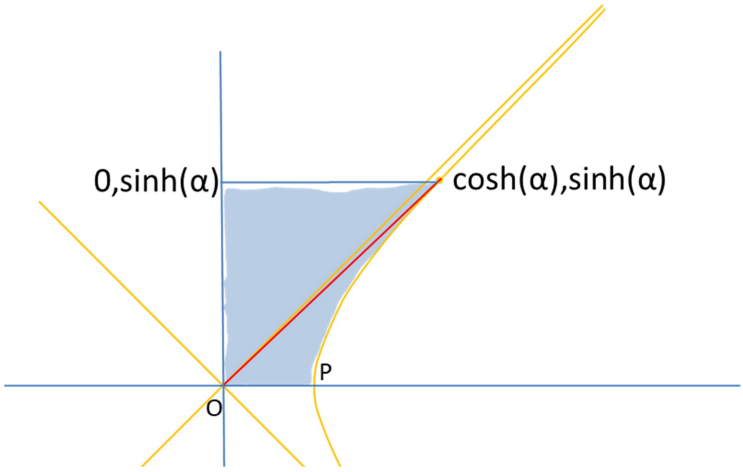


Figure 37. Area of interest

We can compute the area of this figure as

$$A = \int_0^{\sinh} \cosh\theta \, dy$$

where θ is the hyperbolic angle corresponding to a given y. We have $y = \sinh\theta$ so $dy = \cosh\theta \, d\theta$.

$$A = \int_0^{\alpha} \cosh^2\theta \, d\theta$$

Since $\cosh\theta = \frac{e^{\theta}+e^{-\theta}}{2}$, $\cosh^2\theta = \frac{e^{2\theta}+e^{-2\theta}}{4} + \frac{1}{2} = \frac{\cosh 2\theta + 1}{2}$. So we can do the integral as

$$A = \frac{\sinh 2\alpha}{4} + \frac{\alpha}{2}$$

Using the hyperbolic double angle formulas (or again just substituting the definitions for the hyperbolic functions) this is

$$A = \frac{\sinh\alpha \, \cosh\alpha}{2} + \frac{\alpha}{2}$$

This gives us the full shaded area in the figure. However consider the triangle formed by the y axis, the red line from the center to the point at $(\cosh\alpha, \sinh\alpha)$ and the horizontal top of the figure. Note that we want the red line, not the very close golden asymptote. This is just a right triangle with an area of $\frac{\sinh\alpha \, \cosh\alpha}{2}$. The remainder of the figure is the hyperbolic anomaly anomaly and so its area is simply $\frac{\alpha}{2}$. This two regions are highlighted in the figure below. So although it's a bit more subtle than for circles, there is a very direct relationship between an areal and angular definition of the hyperbolic anomaly. The factor of ½ is a bit inelegant, but makes more sense when we note that we're ignoring half of the hyperbola, if we included the other branch symmetrically, then we'd have that the area and the angles were exactly the same. However unlike mathematical abstractions, planets can only be in one place at a time, so we ignore that other half of the area and we just absorb the factor of 2 into the definition of the

hyperbolic anomaly. In the figure below we can consider the hyperbolic anomaly to be either the hyperbolic angle associated with the point along the hyperbola, or twice the green shaded region, bounded by the hyperbola, the red line and the major axis.

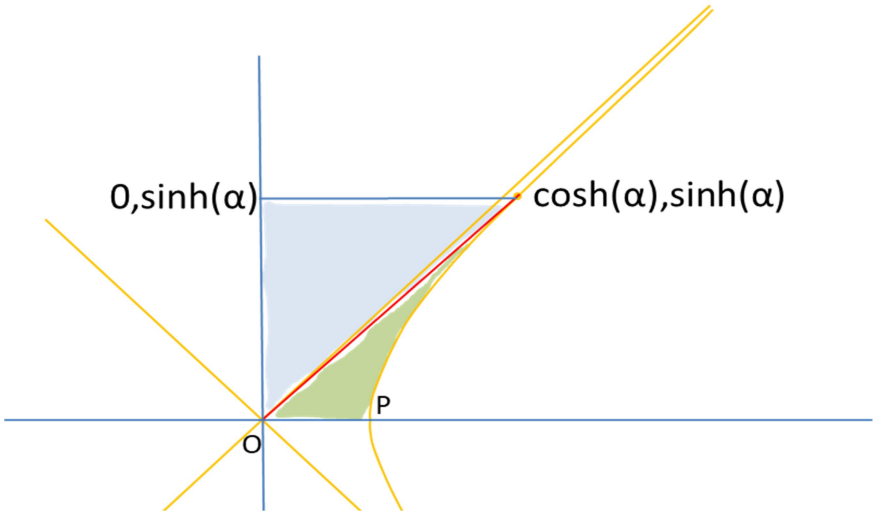


Figure 38. The hyperbolic anomaly and the triangular region

Kepler's equation for hyperbolic orbits

With this lemma in our pocket, let's start again getting a relationship for the hyperbolic anomaly.

Recall the definition of the mean anomaly in this figure.

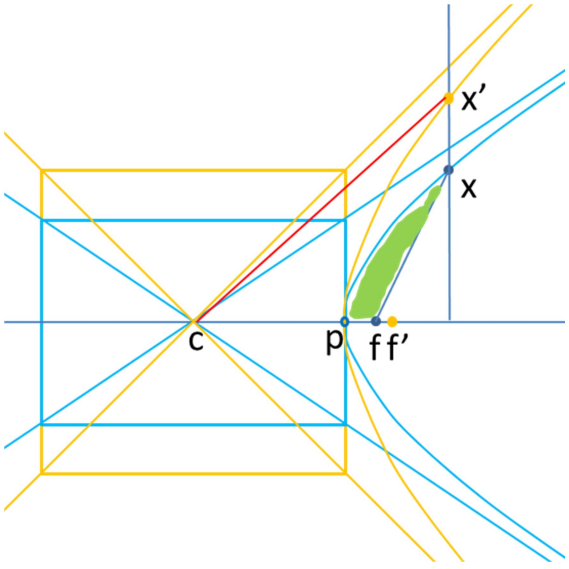


Figure 39. Recalling the Hyperbolic Mean Anomaly: $area_{unscaled}$

The mean anomaly is the green shaded region bounded by **f**, **x**, and **p.** It's the area swept out by the orbit since perihelion. We'll call this $area_{unscaled}$. Normally the mean anomaly is given as an angle, so we need to have some standard area that we will divide by to normalize. For the elliptical case we used the area of the ellipse. While there's no 'area' for a hyperbola, we can use simply use the same formula substituting in the semimajor and semiminor axes of the hyperbola. So we define the angular value of the mean anomaly as

$$MA = 2\pi \frac{area_{unscaled}}{\pi ab} = 2\frac{area_{unscaled}}{ab}$$

In addition to the original blue hyperbola, we're also showing the square/golden hyperbola with the same major axis, center, and perihelion as the original, but with a minor axis equal to the major axis. This second hyperbola is shown in gold. We use perpendiculars through the major axis to stretch the original hyperbola and transfer point **x** to **x'**.

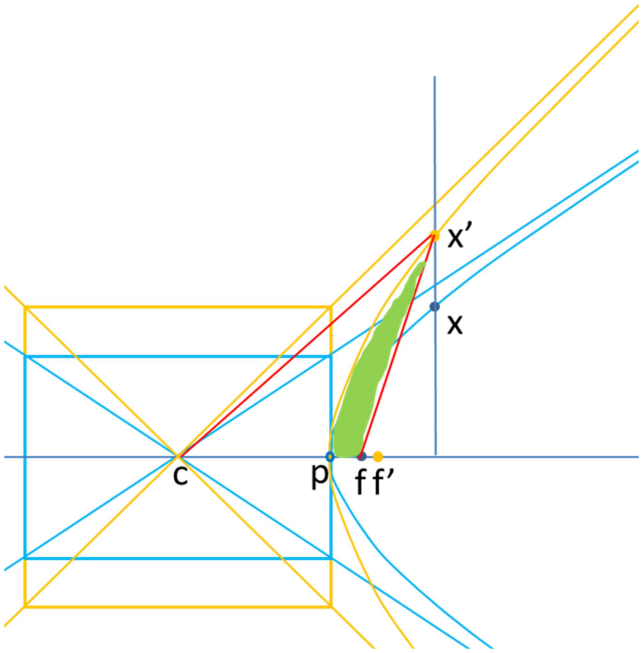


Figure 40. Scaling the hyperbolic mean anomaly: *$area_{scaled}$*

Consider the updated green region in this next figure. Note how we're now drawing a line from **f** to **x'** rather than from **f** to **x**.

Just as ellipses are squashed circles, all hyperbolas are just squashed or stretched versions of a unit hyperbola. Here we've stretched out the original blue hyperbola out into the golden hyperbola. Each point on the original hyperbola has a corresponding point on the golden hyperbola where the y-value is multiplied by a constant ratio. When we stretch the hyperbola, we stretch the mean anomaly area into this old shaded area into the new shaded area.

This stretch ratio is simply the minor axis length of the golden hyperbola over the original minor axis, i.e., since the axes are the same in the square ellipse, just *a/b*. In this case the golden hyperbola has a larger minor axis than the original hyperbola, so the new area is larger than the original area. We could have started with a tall, skinny rectangle where the major axis was smaller than the minor axis. Here the original hyperbola would have only a small turning angle, and we'd have to shrink the original hyperbola to get to the golden hyperbola. Then we'd be multiplying by a factor less than 1.

Since the green region in this second figure is just a linear y-scaling of the green area in the previous one, its area is multiplied by the same factor as the y values from the original MA area.

Note however that this scaled area, is not the mean anomaly for the orbit along the unit hyperbola. The point we are drawing from on the X axis is the focus of the original hyperbola **f**, not the focus of unit hyperbola, **f'** which is what we would need if we were computing the mean anomaly along the golden hyperbola.

For convenience let's set our major axis to 1 letting the minor axis have some value b. Then if we have an input mean anomaly as an angle, we have

$$MA = 2\frac{area_{unscaled}}{b}$$

or

$$area_{unscaled} = \frac{b\ MA}{2}$$

The scaled area in the second figure is just

$$area_{scaled} = \frac{1}{b}\frac{b\ MA}{2} = \frac{MA}{2}$$

The scaling between mean anomaly area and angle is mostly canceled out (except for a factor of 2) by the scaling of the of the original mean anomaly area to the equivalent region associated with the unit hyperbola.

Since we've set our major axis to 1, the golden hyperbola is just the unit hyperbola and if we use the center of the hyperbola as the origin the coordinates of points on it are just the hyperbolic trigonometric functions, so we have

$$\boldsymbol{x}' = (\cosh\alpha, \sinh\alpha)$$

We discovered in our lemma above that the area of the pseudo-triangle **cpx'** is just $\alpha/2$ where α is the hyperbolic anomaly.

We also have a real triangle **cfx'**. Its area is just half the base times the altitude. The altitude is $\sinh\alpha$, the y-value at **x'**. Recall that the focus of a hyperbola is on the circle we circumscribe around the generating rectangle, or half the length of the diagonal of the generating rectangle away from the center. Then the focus of the original hyperbola is a distance $\sqrt{a^2+b^2}$ away from the center or $\sqrt{1+b^2}$ after we scale the major axis. However the eccentricity, *e*, of the hyperbola is just $\frac{\sqrt{a^2+b^2}}{a}$ or also just $\sqrt{1+b^2}$. As we've scaled things the base of the triangle **cfx'** is just *e.*

This triangle **cfx'** comprises exactly the scaled mean anomaly area plus the hyperbolic anomaly area. So setting this triangle area to the sum of its two parts

$$\frac{1}{2}e\sinh\alpha = \frac{MA}{2} + \frac{\alpha}{2}$$

or

$$MA = \alpha - e\sinh\alpha$$

This is Kepler's equation for hyperbolic orbits.

As with the eccentric anomaly we get a relationship with the mean anomaly by combining angular and areal definitions of the hyperbolic anomaly.

It may still seem a little unsatisfying… The length along the arc of the hyperbola doesn't seem to be related to the hyperbolic angle in the same way as the length along the arc of the circle is related to the circular angle. But if we look deeper even that works. Usually we think of the distance along a curve as given by the integral of the infinitesimal delta's along the curve *S*

$$D_S = \int_S \sqrt{\frac{dy}{dS}^2 + \frac{dx}{dS}^2}\, dS$$

But just as we needed to switch a plus sign to a minus when we went from the definition of a circle to a hyperbola, we need to switch from our metric from the sum of distances to a difference of distances.

$$D_h = \int_h \sqrt{\frac{dy}{dS}^2 - \frac{dx}{dS}^2}\, dS$$

This kind of metric is standard in special relativity. Using this metric the 'length' of the arc along the hyperbola and the area that we use for the hyperbolic anomaly match. The radical simplifies to 1.

Getting the true anomaly: elliptic orbits

We've already done all the real work to get the true anomaly in both the elliptic and hyperbolic case. We just need to do a bit of trigonometry. The following figure draws from the figures we used in deriving Kepler's equation for an elliptic orbit, but we've labeled many of the points of interest with their coordinates assuming the origin is at the center of the ellipse. Here the ellipse has a major axis of length 1, and we've circumscribed the unit circle.

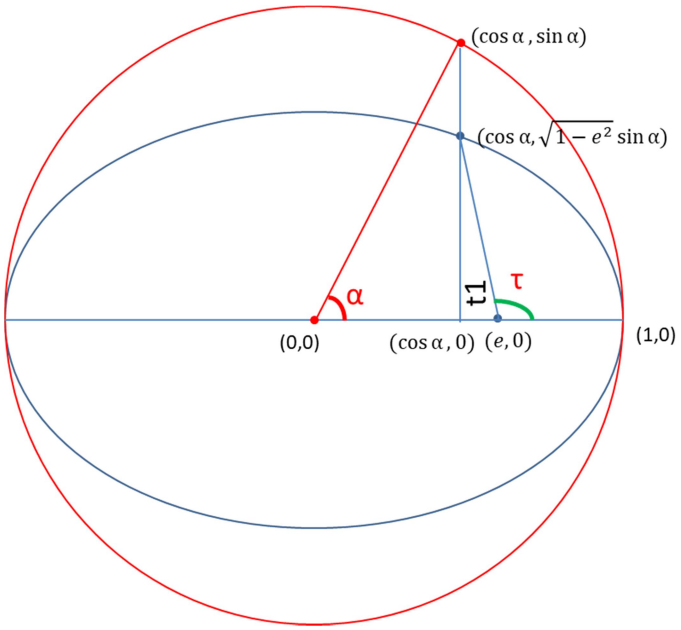


Figure 41. True and Eccentric Anomalies

The true anomaly, τ, is the green angle, while the eccentric anomaly, α, is the green angle. We won't be using the areal definition of the eccentric anomaly here. Since we want the angle at perihelion to be 0, we can immediately see that

$$\tan\tau = \frac{\sqrt{1-e^2}\sin\alpha}{\cos\alpha - e}$$

by measuring the lengths of the sides of the right triangle **t1.** In this example the base of the triangle is a negative length, as we might expect when the angle τ is between 90 and 270°.

We could just leave things here but we can derive a more symmetric form for the relationship using the tangent half angle formula

$$\tan\frac{x}{2} = \frac{1}{2}\frac{\sin x}{\cos x + 1}$$

To use the tangent formula we'll need the sine and cosine of the true anomaly. We know their ratio, and we can simply use the identity $\sin^2 x + \cos^2 x = 1$ to get the normalization. E.g., we know that $\sin\tau = C\sqrt{1-e^2}\sin\alpha$ and $\cos\tau = C(\cos\alpha - e)$ for some single value C. So if

$$\sin^2\tau + \cos^2\tau = 1 = C^2(\left(\sqrt{1-e^2}\sin\alpha\right)^2 + (\cos\alpha - e)^2)$$

So

$$\frac{1}{C^2} = (1-e^2)\sin^2\alpha + \cos^2\alpha - 2e\cos\alpha + e^2$$

$$= \sin^2\alpha - e^2\sin^2\alpha + \cos^2\alpha - 2e\cos\alpha + e^2$$

$$= 1 - e^2\sin^2\alpha - 2e\cos\alpha + e^2$$

Now substitute $\sin^2\alpha = 1 - \cos^2\alpha$

So

$$\frac{1}{C^2} = 1 - e^2 + e^2\cos^2\alpha - 2e\cos\alpha + e^2$$

$$= 1 - 2e\cos\alpha + e^2\cos^2\alpha$$

$$= (1 - e\cos\alpha)^2$$

So using this value for the normalization we take our numerator and denominator from the tangent to get

$$\sin\tau = \frac{\sqrt{1-e^2}\sin\alpha}{1-e\cos\alpha}$$

$$\cos\tau = \frac{\cos\alpha - e}{1-e\cos\alpha}$$

Using these values for the cosine and sine, we can calculate the tangent of the half angle.

$$\tan\frac{\tau}{2} = \frac{1}{2}\frac{\sin\tau}{\cos\tau + 1}$$

$$= \frac{1}{2}\frac{\frac{\sqrt{1-e^2}\sin\alpha}{1-e\cos\alpha}}{\frac{\cos\alpha - e}{1-e\cos\alpha} + 1}$$

We can simplify this to

$$= \frac{1}{2}\frac{\sqrt{1-e^2}\sin\alpha}{\cos\alpha - e + 1 - e\cos\alpha}$$

$$= \frac{1}{2} \frac{\sqrt{1-e^2}\sin\alpha}{(1-e)(1+\cos\alpha)}$$

$$= \frac{1}{2}\sqrt{\frac{1+e}{1-e}}\frac{\sin\alpha}{1+\cos\alpha}$$

but excluding the portion in the radical this is the half angle tangent formula. So

$$\tan\frac{\tau}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{\alpha}{2}$$

This gives a pleasingly symmetric relationship between the true and eccentric anomalies.

Getting the true anomaly: hyperbolic orbits

The situation with hyperbolic orbits and the hyperbolic anomaly is very similar. The figure below is adapted from our discussion of the hyperbolic Kepler equation, and notes the location of various points in the figure. We've set the major axis of our original blue hyperbola to 1. The golden hyperbola then just traces out the hyperbolic sine and cosine.

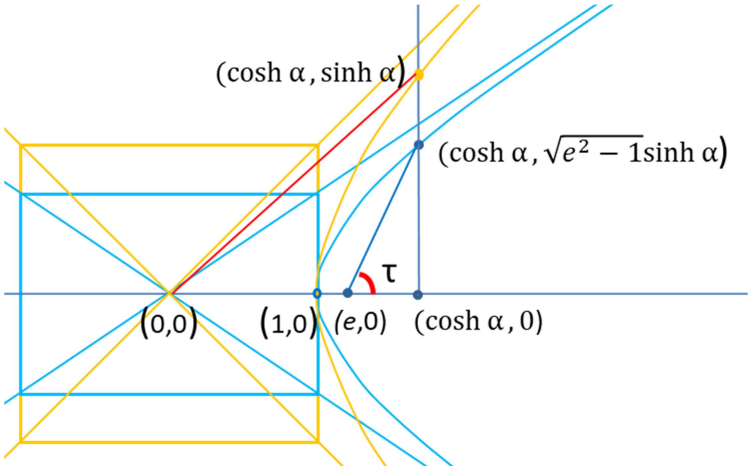


Figure 42. True anomaly in hyperbolic orbit

There's no simple visual of the hyperbolic anomaly angle, but there is a triangle corresponding to **t1** in the previous graphic which gives us a very similar formula for the true anomaly by using the opposite and adjacent sides to the true anomaly angle

$$\tan\tau = \frac{\sqrt{e^2-1}\sinh\alpha}{\cosh\alpha - e}$$

We've replaced the circular trigonometry functions of α with hyperbolic versions, and the coefficient in front of the hyperbolic sine is reversed. Since the eccentricity is now greater than 1, the elliptic orbit

version would have given us an imaginary number. The hyperbolic tangent has the same half angle rule as the circular tangent, so we can proceed exactly as before leaving us with

$$\tan\frac{\tau}{2} = \sqrt{\frac{e+1}{e-1}}\tanh\frac{\alpha}{2}$$

The formula is the same as for the bound orbits except for the use of the hyperbolic tangent, and a change in sign for the expression inside the radical.

Practical and Impractical Applications

The preceding sections have developed the basic concepts and the mathematics for describing orbits. In this section we work through examples where we use these. There are directions we can go, from state vector to orbit description or vice versa. I.e., if we know the position and velocity of an object, we can find its orbits. Conversely if we know the orbit of the object, we can find its position and velocity. As we discuss how to do things, we will occasionally note choices made in the E&IB software.

State vector to orbit

We need nine numbers to determine the osculating elements of an orbit. We need the three components of position, the three components of velocity and the time at which these are valid. That's seven.

We also need the mass of the body we are orbiting and the mass of the object in the orbit. Usually that will be treated as if it is 0, but sometimes it will be significant.

Getting center of mass values

If we are treating the mass of the object in orbit as significant, then we may need to be careful to understand what our position and velocity vectors are. Are these relative to the primary – the full separation and velocity vectors? Or are they relative to the center of mass of the system? Below we will be assuming that the values are relative to the center of mass. If our input position, $\boldsymbol{p}$, and velocity, $\boldsymbol{v}$, are the full difference vectors between two masses m and M, so $\boldsymbol{p} = \boldsymbol{p}_m - \boldsymbol{p}_M$ and similarly for $\boldsymbol{v}$, then we can get the center of mass position and velocity as $\boldsymbol{p}_m = \frac{M}{M+m}\boldsymbol{p}$, $\boldsymbol{v}_m = \frac{M}{M+m}\boldsymbol{v}$.

Units and Constants

The units in which the input position, velocities, times and masses are given may include non-SI units like astronomical units, parsecs, Earth radii, solar masses or days. The table below gives the values used in E&IB for some of the most commonly seen constants. E&IB and the calculations below use SI units.

Name	Description	Value	Units
G M_{sun}	Product of gravitational constant and solar mass	132.712 440 018 x 10^{18}	m^3s^{-2}
G	Gravitational constant	6.674 3 x 10^{-11}	$m^3kg^{-1}s^{-2}$
M_{sun}	Mass of the Sun	1.988 5 x 10^{30}	kg

M_{earth}	Mass of the Earth	5.972 19 × 10^{24}	kg
M_{sun}/ M_{earth}	Ratio of Solar and Earth masses	332 946.048 7	
AU	Astronomical unit (distance of Earth to Sun)	149.597 870 700 x 10^{9}	m
Parsec	Distance where annual *par*allax is 1 arc *sec*ond	30.856 775 812 8 x 10^{15}	m
r_{sun}	Solar radius	1.392 x 10^{9}	m
r_{earth}	Earth radius	6.378 1 x10^{6}	m
Day	Mean solar day	86 400	s
Sidereal Day	Rotation period with respect to distant stars	86 164.090 5	s
Century	Seconds in century (assuming Julian calendar)	(365*100 + 24.25)*86400 3.155 695 200 x10^{9}	s
JD0	Julian Day at 2000-01-01T00:00:00 UT	2 451 544.5	days
Epoch0	Unix epoch at JD0	946 684 800	s

Note the much higher precision for the product of the gravitational constant and the solar mass, than for each of them separately (and the similarly high precision for the ratio of solar and Earth masses).

At various places we find it convenient to measure times in either days or seconds and we use the solar day of 86400 seconds to convert between them. However note that the zero point times and calculations based on Unix times and Julian days tend to ignore leap seconds. There were 22 leap seconds between the 0 of the Unix epoch, the beginning of 1970, and the 0 point used in E&IB at the beginning of 2000. There have been only 5 leap seconds between 2000 and 2021. The E&IB code ignores leap seconds but if precisions greater than a few hundred kilometers are required, they need to be included in orbit calculations. There are a myriad other complexities to address time 'properly', but typically the kind of orbit calculations we contemplate do not require such precision.

A giant planet around a small sun

Suppose we have a planet 10,000 times the mass of the Earth moving around a star 1/3 the mass of the Sun. At time 0, we know that the full separation vector between the two objects is (2.5,1.7,0.3) astronomical units and the full velocity vector between the two objects is (4., 10.,0.1) km/s. So we have as our nine inputs the second column in the table below. Note that we will generally only be displaying the intermediate values and results to a few significant figures, but the underlying calculations are made using higher precision arithmetic.

Parameter	**Input**	**SI**	**SI_{cm}**	**SI_{offset}**
t_0	0	0 s	0 s	0 s
m	10000 M_{earth}	5.9 x 10^{28} kg	*0^{1}*	*0^{1}*
M	M_{sun}/3	6.6 x 10^{29} kg	*5.6 x 10^{29} kg^{1}*	*7.6 x 10^{29} kg^{1}*
$\mathbf{p}_x$	2.5au	370 x 10^{9} m	340 x 10^{9} m	370 x 10^{9} m
$\mathbf{p}_y$	1.7au	250 x 10^{9} m	230 x 10^{9} m	250 x 10^{9} m
$\mathbf{p}_z$	0.3au	45 x 10^{9} m	41 x 10^{9} m	45 x 10^{9} m
$\mathbf{v}_x$	4 km/s	4000 m/s	3700 m/s	4000 m/s
$\mathbf{v}_y$	10 km/s	10000 m/s	9200 m/s	10000 m/s
$\mathbf{v}_z$	0.1km/s	100 m/s	92 m/s	100 m/s

[1]Adjusted as discussed below.

In the third column we've converted our inputs into SI units. In the fourth column we've moved into the center of mass frame. The ratio of masses is about 70 to 6, so the fraction of the position and velocity vectors attributed to the lighter body is about 70/76 or 0.92. We're only keeping about two digits of precision here as we adjust the inputs. More significantly in this column, since we want to work in the center of mass frame, we've adjusted the input masses so that henceforth we can use our standard formula for a massless particle. I.e., we've set the mass of the planet to 0 and the mass of the star to $\frac{M^3}{(m+M)^2}$. This will give us the orbit of our planet only, in the center of mass frame. To get the orbit of the primary we'd simply scale the planet orbit appropriately.

The last column includes the adjustments we would make if we wanted to get results in terms of the difference vectors between the two bodies. Since that's what we were given to start, the vector lengths do not need to be adjusted, but we do need to increase the effective mass of the central body to the sum of the two masses.

Note that we have not made any assumption about what coordinate system is being used here. We could be measuring directions in Equatorial, Ecliptic or any convenient system. Our outputs will be in whatever system was used for the inputs. We will highlight the points at which we determine key orbital characteristics for our sample orbit.

Angular momentum, inclination and longitude of ascending node

The specific angular momentum of the planet is simply $\boldsymbol{L}/m = \boldsymbol{r} \times \boldsymbol{v}$. So we have

$$\boldsymbol{L}/m \approx (340, 230, 42) \times (3.7, 9.2, 0.092) \cdot 10^{12} \text{ m}^2\text{s}.$$

$$\boldsymbol{L}/m \approx (-356, 120, 2291) \cdot 10^{12} \text{ m}^2\text{s}.$$

The angular momentum vector is pointing close to the North Pole, so this will be a prograde orbit – any orbit where the angular momentum points in the northern hemisphere is prograde. We can get the inclination by taking the dot product with the $\hat{\boldsymbol{z}}$ unit vector.

The inclination is just

$$i = \cos^{-1}\left(\frac{L_z}{|\boldsymbol{L}|}\right)$$

$$i \approx \cos^{-1}\left(\frac{2291}{\sqrt{(-356)^2 + 120^2 + 2291^2}}\right)$$

We've dropped the 'x10^{12} m^{2}s' since that cancels in the numerator and denominator.

$$i \approx \cos^{-1}\left(\frac{2291}{2321}\right) \approx 9.3°$$

Next we use the cross-product with the $\hat{\boldsymbol{z}}$ unit vector to get a vector that points along the line of nodes.

$$\boldsymbol{l} = \hat{\boldsymbol{z}} \times \boldsymbol{L}$$

$$\boldsymbol{l} = (0, 0, 1) \times (-365.24, 124.12, 2277)$$

Again we drop the powers of 10 coefficient since we're only interested in the direction.

$$\boldsymbol{l} = (-119, -356, 0)$$

In the E&IB program we'd normalize the vector and return a unit vector. Then

$$\hat{\boldsymbol{l}} = (-0.318, -0.948, 0)$$

We read off the longitude of the ascending node as

$$\Omega = \text{atan2}(\boldsymbol{l}_y, \boldsymbol{l}_x)$$

$$\Omega \approx \text{atan2}(-356, -119)$$

$$\Omega \approx -108.5° = 251.5°$$

Generally we wish to keep Ω in the range 0-360°, since we normally measure longitudes in that range, and we add 360° when we initially get a negative value.

Getting the eccentricity vector

Recall the eccentricity vector

$$\boldsymbol{e} = \frac{\boldsymbol{v} \times \frac{\boldsymbol{L}}{m}}{GM} - \hat{\boldsymbol{r}}$$

This vector will give us both the eccentricity of the orbit and orientation of the major axis.

Substituting in our known values we have

$$\boldsymbol{e} \approx \frac{(3700,9200,92) \times (-356, -120, 2291) \cdot 10^{12}}{6.6743 \cdot 10^{-11} \cdot 5.6 \cdot 10^{29}} - \frac{(340,230,41)}{\sqrt{340^2 + 230^2 + 41^2}}$$

$$\approx (0.564, -0.227, 0.100) - (0.823, 0.560, 0.098)$$

$$\boldsymbol{e} \approx (-0.26, -0.79, 0.001)$$

$$e = |\boldsymbol{e}| \approx 0.83$$

We now have three of our standard orbital elements, i, Ω and e since e is the magnitude of the eccentricity vector.

The argument of periapsis

The eccentricity vector points to perihelion. So we can compare it with the vector, $\boldsymbol{l}$, to the ascending node to get ν, the argument of perihelion – or argument of periapsis might be better used here since we don't seem to be orbiting the sun. This is just the angle between these two vectors.

$$|\sin\nu| = \frac{|\boldsymbol{l}\times\boldsymbol{e}|}{|\boldsymbol{l}|\,|\boldsymbol{e}|} = \frac{|\hat{\boldsymbol{l}}\times\boldsymbol{e}|}{e}$$

$$\cos\nu = \frac{\boldsymbol{l}\cdot\boldsymbol{e}}{|\boldsymbol{l}|\,|\boldsymbol{e}|} = \frac{\hat{\boldsymbol{l}}\cdot\boldsymbol{e}}{e}$$

We resolve the ambiguity in the sine by looking at the angle between $\hat{\boldsymbol{l}}\times\boldsymbol{e}$ and $\boldsymbol{L}$. Since both $\hat{\boldsymbol{l}}$ and $\boldsymbol{e}$ are in the plane of the orbit, the cross-product is perpendicular to it, so it must be parallel or antiparallel to the angular momentum which is also perpendicular to the oribt. These vectors will be parallel when we turn less than $180°$ between the two vectors. Then the sine is positive. If the vectors are antiparallel then we know the sine is negative.

We have

$$\hat{\boldsymbol{l}}\times\boldsymbol{e} = (-0.318,-0.948,0)\times\ (-0.26,-0.79,0.001) = (-0.00079\quad 0.00026\quad 0.0051)$$

Taking the ratio element by element we have

$$\frac{\hat{\boldsymbol{l}}\times\boldsymbol{e}}{\boldsymbol{L}} = (2.21,2.21,2.21)\cdot 10^{-18}$$

So we see that this resultant vector is parallel with the angular momentum, so $\sin\nu > 0$.

We have

$$\sin\nu = \frac{|\hat{\boldsymbol{l}}\times\boldsymbol{e}|}{e} \approx 0.0062$$

$$\cos\nu = \frac{\hat{\boldsymbol{l}}\cdot\boldsymbol{e}}{e} \approx 0.999980$$

Or

$$\nu = \text{atan2}(\sin\nu,\cos\nu) \approx 0.35°$$

The very small z value in the eccentricity vector indicated that the perihelion was very close to the reference plane. Thus we find that the angle along the orbit between the ascending node, where the orbit crossed the reference plane. Remember that $\hat{\boldsymbol{l}}$ points to where the orbit is moving from below the reference plane to above it. So we'll get to the small elevation of the perihelion quickly.

The anomalies

Similarly we can get the true anomaly by looking at the vector between eccentricity vector and the position vector. Again we resolve the ambiguity in the sine by looking at the angle between $\boldsymbol{e} \times \boldsymbol{r}$ and $\boldsymbol{L}$. Here we want to start at the vector to periapsis, and then move to the current position vector.

$$|\sin\tau| = \frac{|\boldsymbol{e} \times \boldsymbol{r}|}{e|\boldsymbol{r}|}$$

$$\cos\tau = \frac{\boldsymbol{e} \cdot \boldsymbol{r}}{e|\boldsymbol{r}|}$$

We compare $\boldsymbol{e} \times \boldsymbol{r}$ with the angular momentum to get the sign of the sine. Since we're ultimately going to be taking the tangent of the sine and cosine, we don't need to normalize by the denominator if we don't want to.

We have

$$\boldsymbol{e} \times \boldsymbol{r} = (-35,\ \ 12,\ \ 227) \cdot 10^9$$

and

$$\frac{\boldsymbol{e} \times \boldsymbol{r}}{\boldsymbol{L}} = (9.96, 9.96, 9.96) \cdot 10^{-5}$$

So we expect $\sin\tau > 0$. If the ratios were negative we would use $-|\boldsymbol{e} \times \boldsymbol{r}|$ instead of $|\boldsymbol{e} \times \boldsymbol{r}|$ in the following.

So

$$\tau = \text{atan2}(\,|\boldsymbol{e} \times \boldsymbol{r}|, \boldsymbol{e} \cdot \boldsymbol{r})$$

$$\tau \approx \text{atan2}(\,2.3 \cdot 10^{11}, -2.9 \cdot 10^{11}) \approx 142°$$

We can immediately derive the eccentric anomaly, α, and mean anomaly, γ, using the relations we derived above.

$$\tan\frac{\tau}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{\alpha}{2}$$

Or

$$\tan\frac{\alpha}{2} = \sqrt{\frac{1-e}{1+e}}\tan\frac{\tau}{2}$$

$$\alpha = 2\tan^{-1}\left(\sqrt{\frac{1-e}{1+e}}\tan\frac{\tau}{2}\right)$$

So

$$\alpha \approx 2\tan^{-1}(0.307\tan 71°)$$

$$\alpha \approx 83.5°$$

The factor of two in the denominators of the tangent functions means that we do not have to worry about any ambiguity in the range of the tangent function, since this gives unique values over the full range $-180°$ to $180°$.

Kepler's equation gives us the mean anomaly instantly

$$\gamma = \alpha - e\sin\alpha$$

$$\gamma \approx 1.46\,radians - 0.83\sin 83.5°$$

$$\gamma \approx 0.635\,radians = 36.4°$$

The only issue here is recognizing that the bare form of Kepler's equation is assuming angles in radians.

In this very eccentric orbit the values of the anomalies vary dramatically. Recall that for a circular orbit all three anomalies have the same value.

We can use any of the anomalies at whatever time we measured the positions to give our reference for the phase of the orbits. Alternatively we can use our measured value of the mean anomaly to compute the previous time of periapsis passage. However to do that we need to compute the period of the orbit and we're not quite there yet.

The size and period of the orbit

With the true anomaly, we can go back to the basic equation for an ellipse in polar coordinates to get the semimajor axis, a.

$$|\boldsymbol{r}| = \frac{a(1-e^2)}{1+e\cos\tau}$$

So

$$a = \frac{|\boldsymbol{r}|(1+e\cos\tau)}{(1-e^2)}$$

$$a \approx \frac{4.2\cdot 10^{11}\cdot 0.34}{0.31} \approx 4.6\cdot 10^{11}\text{m} \approx 3.1\text{ au}$$

With the semimajor axis we have a full set of orbital elements, $a, e, \Omega, \nu, i, \tau$ or γ, and t.

Often we will want the period of the orbit which we had derived above as

$$P = 2\pi\sqrt{\frac{a^3}{GM}}$$

Plugging in we get

$$P = 3.21 \cdot 10^8 \text{s}$$

This is a hair over 10 years. With the period and mean anomaly at our reference time, we can easily find the previous (or next) periapsis passage if that is how we prefer to define the phase of the orbit. Similarly if we need some measure of the frequency of the orbit (orbits/day), then we just invert the period.

Energy

We have not explicitly computed the energy of our system. Above we had derived

$$E = -\frac{GMm}{2a}$$

Equivalently we might say

$$E/m = -\frac{GM}{2a}$$

to define a specific energy the orbiting body. In our example this would give

$$\frac{E}{m} = -\frac{GM}{2a} \approx -40 \cdot 10^6 \text{ J/kg}$$

We've highlighted this, since for most cases where the mass of the orbiting particle is negligible this is the appropriate approach. But in our example this approach is a bit naïve since unlike position or angular momentum, the potential energy of a pair of particles is not easily separable into two pieces associated with the constituents of the pair.

If we go back to the notation where $\boldsymbol{r}$ and $\boldsymbol{v}$ are the full separation vectors and $\boldsymbol{r}_m$, $\boldsymbol{r}_M$, $\boldsymbol{v}_m$, and $\boldsymbol{v}_M$ are the positions and velocities of the particles of masses m and M, then we start with the total energy of the system as

$$E = -\frac{GMm}{|\boldsymbol{r}|} + \frac{m}{2}\boldsymbol{v}_m{}^2 + \frac{M}{2}\boldsymbol{v}_M{}^2$$

If we express this in terms of the position and velocity of the mass, m, we have

$$E = -\frac{GMm}{\frac{(M+m)}{M}|\boldsymbol{r}_m|} + \frac{m}{2}\boldsymbol{v}_m{}^2 + \frac{M}{2}\left(\frac{m}{M}\boldsymbol{v}_m\right)^2$$

$$E = -\frac{GM^2m}{(M+m)|\boldsymbol{r}_m|} + \frac{m}{2}\left(1+\frac{m}{M}\right)\boldsymbol{v}_m{}^2$$

$$E = -\frac{GM^2m}{(M+m)|\boldsymbol{r}_m|} + \frac{m}{2}\left(\frac{M+m}{M}\right)\boldsymbol{v}_m{}^2$$

or looking for a comparable specific energy

$$E/m = -\frac{GM^2}{(M+m)|\boldsymbol{r}_m|} + \frac{M+m}{2M}\boldsymbol{v}_m{}^2$$

We can equivalently express this in terms of the separation vectors

$$E = -\frac{GMm}{|\boldsymbol{r}|} + \frac{m}{2}\left(\frac{M}{M+m}\boldsymbol{v}\right)^2 + \frac{M}{2}\left(\frac{m}{M+m}\boldsymbol{v}\right)^2$$

$$E = -\frac{GMm}{|\boldsymbol{r}|} + \frac{mM\boldsymbol{v}^2}{2}\left(\frac{M}{(M+m)^2} + \frac{m}{(M+m)^2}\right)$$

$$E = -\frac{GMm}{|\boldsymbol{r}|} + \frac{mM\boldsymbol{v}^2}{2(M+m)}$$

Or to get a comparable specific energy

$$E/m = -\frac{GM}{|\boldsymbol{r}|} + \frac{M\boldsymbol{v}^2}{2(M+m)}$$

Substituting in our original masses and separations, using either formula we get

$$\frac{E}{m} \approx -97\cdot 10^6 + 53\cdot 10^6 \approx -44\cdot 10^6 \text{ J/kg}$$

which is slightly different than the naïve calculation above reflecting that the smaller body has of order 10% of the mass of the system.

We see that kinetic energy is a bit more than half the potential energy at the specified instant. The virial theorem says that kinetic energy will average to half the potential energy over time, but that's not true instantaneously except at a couple of specific points in the orbit.

Orbit Elements to the State Vector

Although E&IB provides a full capability for deriving orbital elements from state vectors, most of the action goes the other way. We have elements for planets and moons and asteroids and space missions and need to find the position of these objects at a particular time. In this section we'll give an example of how we do this transformation including how to check results using the Horizons system at JPL.

To go from orbital elements to position and velocity we need the semimajor axis, eccentricity and the three orientation angles: the inclination, longitude of ascending node and argument of perihelion. We also need information on the phase of the orbit, time of last perihelion or a phase/time pair. The mass of the central body will be needed to understand how to scale the mean anomaly, and finally we will need the time at which we want the position and/or velocity. Although it would be possible to use a pair of masses, that doesn't seem to be what happens in practice. We are provided with the effective gravitational mass for our bodies orbit.

We start by using the period of the orbit and our knowledge of the phase of the orbit to get the mean anomaly at the time of interest. We then go through the eccentric anomaly to get to the true anomaly. With the true anomaly we can get the position and velocity of the orbit in the frame where the orbit is an ellipse in the XY plane with the major axis along the X-axis. Then the three orientation angles are used to generate a rotation matrix to rotate the position in the orbit plane into our standard coordinate system.

Let's start with an orbit we retrieve from JPL's extraordinary Horizons system which allows us to get osculating elements for a myriad bodies with many parameters that we can set. We will using the JPL Horizons system to extract the osculating elements for the asteroid #9460 at the time 2000-01-01. We've copied in some of the output from Horizons telnet interface. We had chosen solar barycentric, ecliptic coordinates.

Horizon example

```
Start time      : A.D. 2000-Jan-01 00:00:00.0000 TDB
Stop  time      : A.D. 2000-Jan-02 00:00:00.0000 TDB
Step-size       : 2880 minutes
*******************************************************************************
Center geodetic : 0.00000000,0.00000000,0.0000000 {E-lon(deg),Lat(deg),Alt(km)}
Center cylindric: 0.00000000,0.00000000,0.0000000 {E-lon(deg),Dxy(km),Dz(km)}
Center radii    : (undefined)
Keplerian GM    : 2.9630927493457475E-04 au^3/d^2
Small perturbers: Yes                            {source: SB441-N16}
Output units    : AU-D, deg, Julian Day Number (Tp)
Output type     : GEOMETRIC osculating elements
Output format   : 10
Reference frame : Ecliptic of J2000.0
*******************************************************************************
Initial IAU76/J2000 heliocentric ecliptic osculating elements (au, days, deg.):
  EPOCH=  2457353.5 ! 2015-Nov-27.00 (TDB)         Residual RMS= .24626
   EC= .1544932530424187   QR= 2.250897867600008   TP= 2456612.9202919132
   OM= 70.34282119796717   W=  349.7881962849056   IN= 13.6808032427285
  Equivalent ICRF heliocentric cartesian coordinates (au, au/d):
   X=-1.892594519358135E+00  Y=-2.297625358354860E+00  Z=-7.376022521104699E-01
  VX= 6.666972430273445E-03 VY=-4.504106888330518E-03 VZ=-4.134091396802909E-03
Asteroid physical parameters (km, seconds, rotational period in hours):
   GM= n.a.                RAD= 4.069              ROTPER= n.a.
   H= 12.96                G= .150                 B-V= n.a.
                           ALBEDO= .121            STYP= n.a.
*******************************************************************************
JDTDB
   EC     QR    IN
   OM     W     Tp
   N      MA    TA
```

```
   A     AD   PR
*******************************************************************************
$$SOE
2451544.500000000 = A.D. 2000-Jan-01 00:00:00.0000 TDB
 EC= 1.555906714443290E-01 QR= 2.233238856380111E+00 IN= 1.372432149577838E+01
 OM= 7.045508495808998E+01 W = 3.508486510213602E+02 Tp=  2451849.447384673171
 N = 2.293094885447135E-01 MA= 2.900726711875558E+02 TA= 2.724179649166191E+02
 A = 2.644734941760980E+00 AD= 3.056231027141850E+00 PR= 1.569930674411682E+03
$$EOE
*******************************************************************************

TIME

  Barycentric Dynamical Time ("TDB" or T_eph) output was requested. This
continuous relativistic coordinate time is equivalent to the relativistic
proper time of a clock at rest in a reference frame comoving with the
solar system barycenter but outside the system's gravity well. It is the
independent variable in the solar system relativistic equations of motion.

  TDB runs at a uniform rate of one SI second per second and is independent
of irregularities in Earth's rotation.

  Calendar dates prior to 1582-Oct-15 are in the Julian calendar system.
Later calendar dates are in the Gregorian system.

REFERENCE FRAME AND COORDINATES

  Ecliptic at the standard reference epoch

    Reference epoch: J2000.0
    X-Y plane: adopted Earth orbital plane at the reference epoch
               Note: IAU76 obliquity of 84381.448 arcseconds wrt ICRF X-Y plane
    X-axis   : ICRF
    Z-axis   : perpendicular to the X-Y plane in the directional (+ or -) sense
               of Earth's north pole at the reference epoch.

  Symbol meaning [1 au= 149597870.700 km, 1 day= 86400.0 s]:

    JDTDB    Julian Day Number, Barycentric Dynamical Time
      EC     Eccentricity, e
      QR     Periapsis distance, q (au)
      IN     Inclination w.r.t X-Y plane, i (degrees)
      OM     Longitude of Ascending Node, OMEGA, (degrees)
      W      Argument of Perifocus, w (degrees)
      Tp     Time of periapsis (Julian Day Number)
      N      Mean motion, n (degrees/day)
      MA     Mean anomaly, M (degrees)
      TA     True anomaly, nu (degrees)
      A      Semi-major axis, a (au)
      AD     Apoapsis distance (au)
      PR     Sidereal orbit period (day)
```

Sample Horizons Output

The text above includes a typical output from the Horizons system giving a single set of osculating elements for our asteroid. Horizons is perfectly happy to give any number of elements spaced at whatever time spacing a user selects, but for our purposes in illustrating how to use the elements to find positions and velocities at a given time a single instance is fine.

We are going to find the position of the asteroid at the reference epoch. Horizons will also give us state vectors for given times, so that allows us to check that our math is correct. The key elements that we will need from the output are highlighted in bold. We see the line:

```
Keplerian GM    : 2.9630927493457475E-04 au^3/d^2
```

This tells us the effective mass that we need to use for this. As we've noted above, masses and the gravitational constant are known relatively poorly, but their product is known to much higher precision. This is a small asteroid, so its own mass is entirely negligible, but even if we were looking at Jupiter we'd simply plug in the effective mass shown here.

The next critical line is

```
Output units    : AU-D, deg, Julian Day Number (Tp)
```

Angle will be in degrees, distances in AU, dates in Julian day numbers. We will need to convert appropriately.

Finally in the table of the osculating elements we have highlighted several elements

```
2451544.500000000
EC= 1.555906714443290E-01
IN= 1.372432149577838E+01
OM= 7.045508495808998E+01
W = 3.508486510213602E+02
Tp=  2451849.447384673171
MA= 2.900726711875558E+02
A = 2.644734941760980E+00
```

These are the epoch of the osculating elements, the eccentricity, inclination, longitude of ascending node, argument of perigee, time of perihelion, mean anomaly and semimajor axis. We should be able to derive the non-highlighted elements from these. The epoch/mean anomaly, and time of perihelion are redundant but illustrate the two ways that the phase in orbit can be given. Let's format all of this information in a table

Element	**Horizon Input**	**SI**	
GM	2.9630927493457475E-04 au^3/d^2	1.32890518817909e+20	
M mass of sun		1.99107799796097e+30	
Epoch of elements	2451544.500000000	0 s	
e eccentricity	0.1555906714443290	0.155590671	
i inclination	13.72432149577838	13.724321°	
Ω long.asc.node	70.45508495808998	70.455085°	
v arg.perihelion	350.8486510213602	350.848651°	
t_0 perihelion time	2451849.447384673171	26347454.03 s	
MA mean anomaly	290.0726711875558	290.072671°	
a semimajor axis	2.644734941760980	395.646716e9 m	

We have rendered times in the third column in seconds since 2000-01-01T00:00:00.

The period of the orbit

We recall that the period is given by

$$P = 2\pi\sqrt{\frac{a^3}{GM}}$$

$$P = 135642010 \text{ s} = 1569.9306 \text{ days}$$

If we look back at the Horizons output this agrees with the value specified there to the precision we've written things out. We'll be presenting results to a much high degree of precision than we did going in the other direction, since we going to be comparing with Horizons results and we want to make the agreement manifest.

Phase of the orbit

With the period we can reconcile our two mechanisms for specifying the phase of the orbit. We have the mean anomaly at the epoch as 290.072671°. We notice that the time of perihelion is positive, and we've put the epoch of the elements is the 0 of our time system. So the time of perihelion must be the time of the next perihelion after the epoch of the elements. It should be

$$t_0 = \frac{360 - 290.072671}{360} \cdot 135642010 = 26347454 \text{ s}$$

This value, in seconds, is in the table of values we got from Horizons we'd converted it from a Julian date. So we could start with either the epoch/mean anomaly pair or the time of perihelion since it's easy to convert from one to the other.

We initially want to get the position of the body at time 0 but this is the epoch of the ephemeris, so we've already got that. Say we also wanted to find the mean anomaly at the beginning of 1956-03-03. We need to find the number of seconds between the beginning of that day and the epoch. The HEASARC's date converter at https://heasarc.gsfc.nasa.gov/cgi-bin/Tools/xTime/xTime.pl provides an easy tool for converting ISO dates to Julian dates. We enter 1956-03-03 as the ISO date and it returns the corresponding Julian date: 2435535.5. We compare with the epoch date of 2451544.5 and see that our specified date is at day -16009 before the epoch, so --ignoring leap seconds-- we would have

$$t = -1383177600$$

on March 3, 1956. So using the ephemeris we have that would give a mean anomaly of

$$MA = MA_{ep} \quad + \frac{t}{P}\ 360°$$

$$MA = 290.072671° - \frac{1383177600}{135642010}\ 360° = -3380.9429° = 219.05706°$$

where we have used the modulus operator to get an angle between 0 and 360 and *t* is the time relative to the epoch at which we have a known value of the mean anomaly. Note that it will be a bad idea to rely on a position extrapolating the orbit back almost 50 years.

The Eccentric Anomaly and Kepler's equation

The next step is the only place where we cannot simply plug into an equation and get the answer. To get the eccentric anomaly we need to solve Kepler's equation, and there is no easy analytic solution going in this direction. We have

$$\gamma = \alpha - e \sin \alpha$$

where γ is the mean anomaly and α is the eccentric anomaly. Going the other way we just plugged in the value of α and γ popped out, but the inverse is quite hard despite the apparent simplicity of the equation. The usual approach is to iterate. We can start with the assumption of a circular orbit so $\gamma = \alpha$ and then use the measured error and derivatives to make a better guess until we get close enough. There are a lot of algorithms out there. For most realistic orbits convergence is generally pretty rapid. However for highly elliptic orbits things can take a while. One thing that helps a bit, is that we know that at perihelion and aphelion the anomalies are all the same. We can use Newton's method to iterate. The derivative of Kepler's equation gives us

$$\frac{d\gamma}{d\alpha} = 1 - e \cos \alpha$$

We start with $\alpha_0 = \gamma$ and compute $\gamma_0 = \alpha - e \sin \alpha_0$. Then we iterate with

$$\alpha_n = \alpha_{n-1} + (\gamma - \gamma_{n-1}) / \frac{d\gamma}{d\alpha}$$

where we evaluate the derivative at α_{n-1}. Using our input values for the mean anomaly at the epoch 2000-01-01 and the known eccentricity we have a very rapid convergence

$$\alpha_0 = 290.072662353516°$$
$$\alpha_1 = 281.227142333984°$$
$$\alpha_2 = 281.331756591797°$$
$$\alpha_3 = 281.331787109375°$$

where the last value is accurate to limit of the machine precision (i.e., $\gamma - \gamma_3 = 0$)

If we used the value for the mean anomaly at 1956-03-03 we find similar convergence

$$\alpha_0 = 219.057060241196°$$
$$\alpha_1 = 214.045464722825°$$
$$\alpha_2 = 214.063802694455°$$
$$\alpha_3 = 214.063802920931°$$

Note that in the actual calculations we treat the angles as radians and have converted them to degrees for display here. Nor do we have anything like the number of significant figures displayed in any actual measurements, but in this case we display to machine precision to show how quickly the mathematical sequence converges. Note that depending upon input values of the position and mean anomaly, convergence can be much slower.

The True Anomaly

We can just plug into the formula we derived earlier to get the true anomaly.

$$\tan\frac{\tau}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{\alpha}{2}$$

If we get a negative value for the anomaly we add 360°. We find the true anomalies are

$$\tau_{2000} = 272.417982°$$

$$\tau_{1956} = 209.348537°$$

for the two times we are looking at. If we look back at our raw Horizons output, we see the appropriate value for the true anomaly at the epoch mirrored there.

Position and velocity in the orbit plane

We now have sufficient information to determine the location and velocity of the particle in the orbit plane. We get the position from our equation for the ellipse

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

$$\boldsymbol{r} = \frac{a(1-e^2)\cos\theta}{1+e\cos\theta}\hat{\boldsymbol{x}} + \frac{a(1-e^2)\sin\theta}{1+e\cos\theta}\hat{\boldsymbol{y}}$$

Writing this in matrix notation which we will use in a bit below this would be $\boldsymbol{r} = r\begin{bmatrix}\cos\theta\\ \sin\theta\\ 0\end{bmatrix}$.

We've not discussed the velocity directly before but we have all we need to calculate it as well. We need to collect a few results from our earlier work. In addition to the equation above we recall that in polar coordinates the velocity is

$$\boldsymbol{v} = \frac{dr}{dt}\hat{\boldsymbol{r}} + r\frac{d\theta}{dt}\hat{\boldsymbol{\theta}}$$

where the polar unit vectors are related to the Cartesian unit vectors by

$$\hat{\boldsymbol{r}} = \cos\theta\,\hat{\boldsymbol{x}} + \sin\theta\,\hat{\boldsymbol{y}}$$

$$\hat{\boldsymbol{\theta}} = -\sin\theta\,\hat{\boldsymbol{x}} + \cos\theta\,\hat{\boldsymbol{y}}$$

We noted that the specific angular momentum can be written as

$$\frac{L}{m} = \sqrt{GMa(1-e^2)}$$

But we have also seen that we can get $\frac{d\theta}{dt}$ from the angular momentum,

$$r^2\frac{d\theta}{dt} = \frac{L}{m} = \sqrt{GMa(1-e^2)}$$

so

$$\frac{d\theta}{dt} = \frac{\sqrt{GMa(1-e^2)}}{r^2}$$

We have the radius as a function of angle, not time so we have $\frac{dr}{dt} = \frac{dr}{d\theta}\frac{d\theta}{dt}$ where

$$\frac{dr}{d\theta} = \frac{ae(1-e^2)\sin\theta}{(1+e\cos\theta)^2}$$

Putting all of this together we have

$$\boldsymbol{v} = \frac{d\theta}{dt}\left(\frac{dr}{d\theta}\hat{\boldsymbol{r}} + r\hat{\boldsymbol{\theta}}\right)$$

$$\boldsymbol{v} = \frac{d\theta}{dt}\left(\frac{dr}{d\theta}\cos\theta - r\ \sin\theta\right)\hat{\boldsymbol{x}} + \frac{d\theta}{dt}\left(\frac{dr}{d\theta}\sin\theta + r\ \cos\theta\right)\hat{\boldsymbol{y}}$$

This is already pretty messy. While we could substitute in the values for $\frac{d\theta}{dt}$ and $\frac{dr}{d\theta}$ to get an expression entirely in terms of the orbital constants and the true anomaly, it's easier to simply calculate these intermediates first and then plug the resulting values in.

If we try this for our two desired times for which we've derived the true anomaly we find

$$\boldsymbol{r}_{2000} = [\ 1.6181706 \cdot 10^{10}\ , -3.8320951 \cdot 10^{11}, 0]\text{m}$$

$$\boldsymbol{v}_{2000} = [18536.504, 3669.4142, 0]\text{m/s}$$

$$\boldsymbol{r}_{1956} = [\ -3.8931837 \cdot 10^{11}, -2.1890913 \cdot 10^{11}, 0]\text{m}$$

$$\boldsymbol{v}_{1956} = [9093.2263, -13285.147, 0]\text{m/s}$$

The zeros in the z values of the coordinates remind us that these are the values in the orbit plane, we still have to account for the orientation of the orbit in space.

The rotations

It's all over except for the rotations! We have found the position and velocity in the 'orbit plane' system, where the major axis is aligned with the x-axis. We now apply three rotations to transform this plane to the sky. The three rotations act as a set of Euler angles where we rotate the by a specified amount around a specified axis. Any rotation of the coordinate system can be described as a set of three Euler angle transformations and we can treat the three orientation angles are describing this set of transformations, first around the z axis, then the x axis, then around the z axis again. Note that the axes themselves move in these transformations, so that the axis of the last transformation is not the same as the first. It's easy to get the order and direction of the rotation confused, so we'll discuss this in detail as we proceed.

The position and velocity vectors we have are in the orbit plane, so the first angle we encounter is the argument of perihelion, ν, which is gives us the angle between the x axis and the major axis of the ellipse. This angle is in the orbit plane. After we apply this transformation, the x-axis should point along the line of nodes. If the argument of perihelion is some small positive value, we want to transform the vector (1,0,0) which initially points to the perihelion to something like [a little less than 1, a little more than 0, 0] . So the transformation we need to get that is

$$\boldsymbol{r}_1 = \begin{bmatrix} \cos\nu & -\sin\nu & 0 \\ \sin\nu & \cos\nu & 0 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{r}_0$$

This will transform the perihelion unit vector $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ to $\begin{bmatrix} \cos\nu \\ \sin\nu \\ 0 \end{bmatrix}$ which is what we want.

The x-axis is now pointing along the line of nodes, so it is the pivot on which we are going to do the next rotation. The line of nodes is where the reference plane and the orbit plane intersect. We are going to rotate along that hinge, by the inclination of the orbit. Since we are pointing to the ascending node, something in the orbit plane at a small positive angle from the origin should end up with a positive z value. E.g., [1,δ, 0] should end up as [x,y,>0]. So

$$\boldsymbol{r}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos i & -\sin i \\ 0 & \sin i & \cos i \end{bmatrix} \boldsymbol{r}_1$$

This transforms $\begin{bmatrix} 1 \\ \delta \\ 0 \end{bmatrix}$ to $\begin{bmatrix} 1 \\ \delta\cos i \\ \delta\sin i \end{bmatrix}$ so that a position the particle has shortly after passing the ascending node has a positive z value.

Now the x axis is still pointing along the line of nodes, but the x-y plane is now our standard reference plane. We now that if the longitude of the ascending node, Ω, is a small angle, we want the position [1,0,0] to end up in the first quadrant. So again we have

$$\boldsymbol{r}_f = \boldsymbol{r}_3 = \begin{bmatrix} \cos\Omega & -\sin\Omega & 0 \\ \sin\Omega & \cos\Omega & 0 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{r}_2$$

We can multiply the matrices together to combine the rotations

$$\boldsymbol{r}_f = \begin{bmatrix} \cos\Omega & -\sin\Omega & 0 \\ \sin\Omega & \cos\Omega & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos i & -\sin i \\ 0 & \sin i & \cos i \end{bmatrix} \begin{bmatrix} \cos\nu & -\sin\nu & 0 \\ \sin\nu & \cos\nu & 0 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{r}_0$$

$$\boldsymbol{r}_f = \begin{bmatrix} \cos\Omega\cos\nu - \sin\Omega\sin\nu\cos i & -\cos\Omega\sin\nu - \sin\Omega\cos\nu\cos i & \sin\Omega\sin i \\ \sin\Omega\cos\nu + \cos\Omega\sin\nu\cos i & -\sin\Omega\sin\nu + \cos\Omega\cos\nu\cos i & -\cos\Omega\sin i \\ \sin\nu\sin i & \cos\nu\sin i & \cos i \end{bmatrix} \boldsymbol{r}_0$$

We use this matrix to transform the input vector in the original orbit plane, into whatever standard coordinate system we are using. Since components of the matrix are constants they need to be computed only once per orbit. The matrix multiplication is quick, involving just a few additions and multiplications.

This matrix describes how we transform vectors from the system oriented with the orbit, to the standard reference system. If we want to go the other way, we have a vector in standard coordinates and we want to see what it is in the plane of the orbit, then we need to use the inverse transformation, which requires that we get the inverse of the matrix. Rotation matrices have the property that their inverse is their transpose, we just need to exchange the columns and rows. So we get that transformation pretty much for free.

If we are interested in just positions, then we can multiply by the position vector by the matrix. I.e., we had

$$\boldsymbol{r}_0 = r\begin{bmatrix} \cos\tau \\ \sin\tau \\ 0 \end{bmatrix}$$

After a bit of drudgery and use of the sine and cosine addition formulas we get

$$\boldsymbol{r}_f = r\begin{bmatrix} \cos\Omega\cos(\nu+\tau) - \sin\Omega\sin(\nu+\tau)\cos i \\ \sin\Omega\cos(\nu+\tau) + \cos\Omega\sin(\nu+t)\cos i \\ \sin(\nu+t)\sin i \end{bmatrix}$$

which simplifies things a little if we are only interested in position. We can get this result quicker if we start by combining ν and τ since they represent rotations in the same plane. Regardless, we'll need the full matrix to handle velocities.

Using this we can get the positions and velocities for our asteroid in solar barycentric ecliptic coordinates.

$$\boldsymbol{r}_{2000} = (\,2.2304052, -1.1107896, -0.60408633)\text{au}$$

Horizons can also be used to get the position directly it gives

X = 2.230405022847759E+00 Y =-1.110790089374123E+00 Z =-6.040863228231372E-01

The very slight differences are probably attributable to limited precision used in the calculation.

If we look at our date on 1956, things don't turn out so well

Horizons gives

X =1.559700791189361E+00 Y =-2.225138447971239E+00 Z =-5.341168038749972E-01

whereas we see $\boldsymbol{r}_{1956} = (0.0062533950, -2.9755849, -0.24455618)\text{au}$. This is off by a lot. If we were to look at the orbital elements, we find the mean anomaly is about 30 degrees off from our extrapolation though most of the other orbital elements have changed by only a few percent. So the basic shape and characteristics are reasonably similar to what we have, but the cumulative errors have led to a substantial offset building up.

If we use the matrix convert the velocities, we find the

$$\boldsymbol{v}_{2000} = (\,5700.0294, 18015.289, 160.05107)\text{m/s}$$

or if we convert to the preferred units for horizon

$$\boldsymbol{v}_{2000} = (0.0032920425, 0.010404700, 0.00009243722)\ \text{au/day}$$

which compares to the value Horizons gives of

VX= 3.292044365251326E-03 VY= 1.040469913882338E-02 VZ= 9.243669195736235E-05

confirming that the algorithm is consistent.

Rockets

Rockets are the epitome of Newton's third law of motion, every action has an opposite and equal reaction. There are a number of others ways to say this: in a closed system the sum of the forces is 0, or just that linear momentum is conserved.

If we throw something in one direction, we are pushed to move in the opposite direction. Suppose we were standing on a very slippery frozen lake with carrying a bunch of rocks. If we throw the rocks in one direction, we may slide a little in the other direction. We're a little 'rock'et on the lake. As often happens in our everyday experience, friction means that we will slide only a little when we throw our rocks. If we were in space and threw a rock in one direction we'd move in the opposite direction and never slow down (due to friction at least).

On the lake if we throw a bigger rock or we throw it faster, we slide a bit faster and further. The momentum of the rock we throw is the product of its mass and velocity. We'll get the same momentum in the opposite direction.

There are three elements of the rocket that we want to distinguish. The rocks are the 'reaction mass' . Every rocket throws away some reaction mass. We're the payload: the thing that we're trying to move. The third element of the rocket is the mechanism by which we accelerate the reaction mass when we throw it away. For our "rock"et, that's the muscles in our arms. In

chemical rockets the reaction mass is accelerated by burning it. The nozzle of the rocket directs the burning gas in a focused stream away from the payload, but ultimately our goal is to 'throw' away these burning gases from the rocket as quickly as possible.

Ion rockets are becoming increasingly important for satellite station keeping and may ultimately be used to dramatically decrease the length of time needed for interplanetary missions. These use electric fields to ionize and accelerate gases out of the rocket. In this case the energy source for the rocket is entirely separate from the reaction mass.

If we go back to our 'rock'et, we might have a fair number of rocks to throw. When we throw the first rock it needs to move both our body and the remaining rocks. When we throw the last rock we're only pushing our body, so we might get a little more velocity for the same momentum – assuming our arm isn't getting too tired throwing all of these.

We characterize rockets by the specific impulse of the reaction mass. I.e., how much momentum can we extract from each kilogram of our reaction mass? Since momentum is just the product of mass and velocity, the specific impulse, the momentum over the mass, is basically just the velocity at which we throw things away.

The Rocket Equation

We can write a little differential equation

$$\frac{dv}{dt} = -I_s \frac{\frac{dM_f}{dt}}{M_f + M_p}$$

This says that the instantaneous change in velocity of the rocket payload is just the specific impulse, I_s, times the amount of fuel we used over the current weight of the rocket, the current mass of the fuel, M_f, plus the mass of the payload, M_p. There's a minus sign in front, because the mass of the fuel is decreasing as we use it up (by burning it, or throwing rocks or whatever). We can ask how fast we can go if we start with some initial mass of fuel, M_0, and use it all. We just have to integrate

$$\boldsymbol{\Delta v} = \int \frac{dv}{dt} dt = \int -I_s \frac{\frac{dM_f}{dt}}{M_f + M_p} dt$$

$$\boldsymbol{\Delta v} = \int_{M_0}^{0} -I_s \frac{dM_f}{M_f + M_p}$$

The velocity change is just this integral as we burn all of our fuel to be left with none. This is an elementary integral and we're left with

$$\boldsymbol{\Delta v} = -I_s[\ln M_p - \ln(M_f + M_p)] = I_s \ln \frac{M_0 + M_p}{M_p}$$

There is good news and bad news in this equation. The good news is that we can achieve any final velocity we'd like regardless of what the specific impulse is. The bad news is that to go more than a few times faster than the specific impulse velocity we need vast amounts of reaction mass. Since chemical rockets have specific impulses of a few kilometers a second, it's virtually impossible to use them directly to accelerate to velocities of more than about 10-20 km/s. A typical specific impulse for a modern chemical rocket is something like 3 km/s. So ignoring atmospheric friction and the launch details, we would need an absolute minimum of about 8 times the payload weight in reaction mass (or fuel/oxidizer for a chemical rocket) to achieve Earth orbit.

If we consider the Falcon 9, the first stage uses about 400 tons of propellant to move about 150 tons of the first stage empty weight and a fueled second stage. So we can't expect that to get us more than about 3 km/s of velocity – completely ignoring air resistance which we can expect to substantially reduce this. The second stage has a mass of about 4 tons and carries a payload of around 10 tons with a fuel mass of about 100 tons. So it can get a $\boldsymbol{\Delta v}$ of perhaps 6 km/s and given that it works in a near vacuum we can expect it to actually achieve something close to that.

So the Falcon 9 payload can achieve the 7 km/s or 11 km/s needed for Earth orbit or escape. However if we want to get to 100 km/s, then using the current chemical rockets we're looking at ratios of fuel mass to payload mass of trillions to 1. To get to even mildly relativistic energies (say 0.001c) for a significant payload would require we use the entire mass of the universe for fuel. Other people might object.

Limits to the Specific Impulse of Chemical Rockets

The limits to the specific impulse of chemical rockets come from the underlying chemical reactions. The reactions generate a certain amount of energy per mass of fuel. The kinetic energy we gain from the rocket cannot exceed this energy.

One widely used rocket fuel is to burn oxygen and hydrogen to form water vapor. This reaction produces about 250 kJ of energy for each mole of water produced. Since a mole of water is 18 grams, we can get an absolute limit for the specific impulse of rockets using this fuel by solving

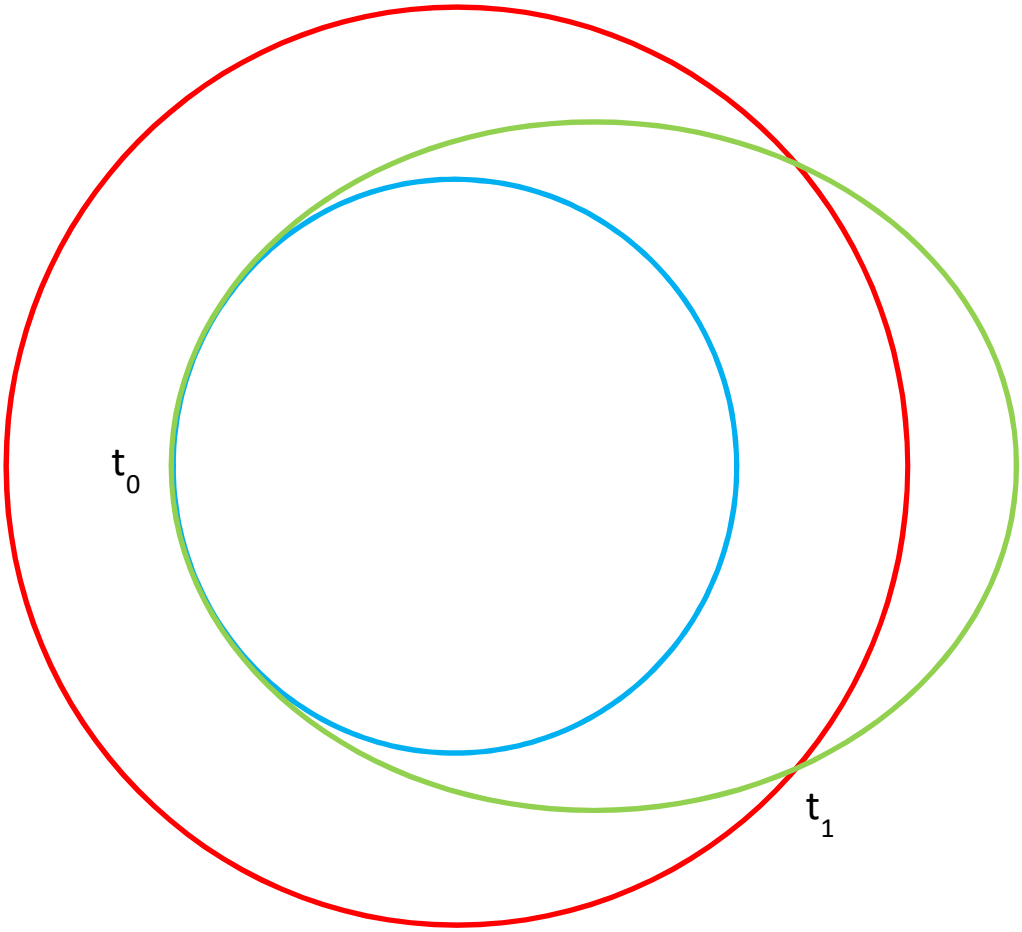
$$\frac{0.018\ v^2}{2} = 250{,}000$$

This corresponds to just over 5 km/s. Real rocket science is designing how to burn the fuel and channel the hot gas produced so that a large fraction of the total energy is converted into directed motion and not random thermal energy.

Hydrogen and oxygen is a near ideal fuel in terms of having a high energy reaction with a very small molecular weight for the products, so we aren't going to do significantly better with other chemical approaches. Fortunately we can get around the solar system with Δv's of order 10 km/s.

Transfer orbits between planets.

We often want to move from one celestial body to another. In principle one can simply add an impulse to the current orbit, which transfers us into an orbit which will intersect with the desired destination at some later time. When the intersection occurs we induce another impulse which matches our rockets orbits to the destination body. For the nonce we'll not consider the gravity of the source or destination bodies. So at time t_0 we leave the source body (say the Earth) and at t_1 we reach the destination body (say Mars). In between we are on a fixed transfer orbit with the Sun as one focus. We know the positions and times of the



impulses at t_0 and t_1. Since these two points and the Sun are all in the same plane we can treat this as a two dimensional problem. We want to know the semi-major axis, eccentricity, phase and argument of perihelion for the orbit.

Figure 43. A transfer orbit where the start is at perihelion.

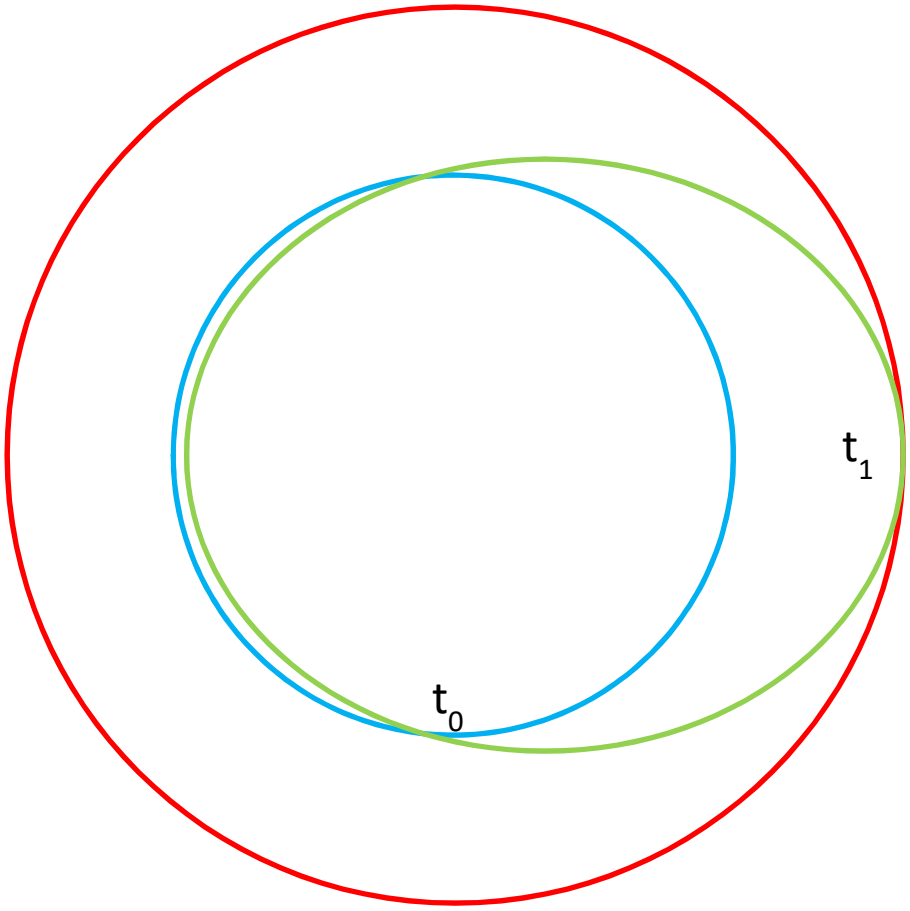


Figure 44. Transfer orbit finishing at aphelion.

Generally speaking we expect there to be such an orbit for any pair of points and times (so long as we're allowing enough time for the orbit that we don't need to worry about speed of light limitations). However if we don't give much time for the transfer, then the transfer orbit is very likely hyperbolic. E.g., if we want to get to Mars in a week, we'd need to travel more than 7 million kilometers each day, or at least 100 kilometers/second, far above the escape velocity from the Sun. So we'll want to consider both elliptical and hyperbolic trajectories.

Fortunately we have essentially the same formula for radius. r, as a function of angle, α for both elliptical and hyperbolic orbits.

$$r = \frac{(1 - e^2)a}{1 + e\cos(\alpha - \delta)}$$

Where a is the semimajor axis, e is the eccentricity, α is the time dependent angle within the ellipse (i.e., the true anomaly), and δ is a measure of the argument of perihelion, the orientation of the ellipse with the plane. For a hyperbolic orbit the term $(1 - e^2)$ is negative, but we also treat the semimajor axis as negative so the radius is still positive.

Here a, e, and δ are the constants defining the size, shape and orientation of the ellipse

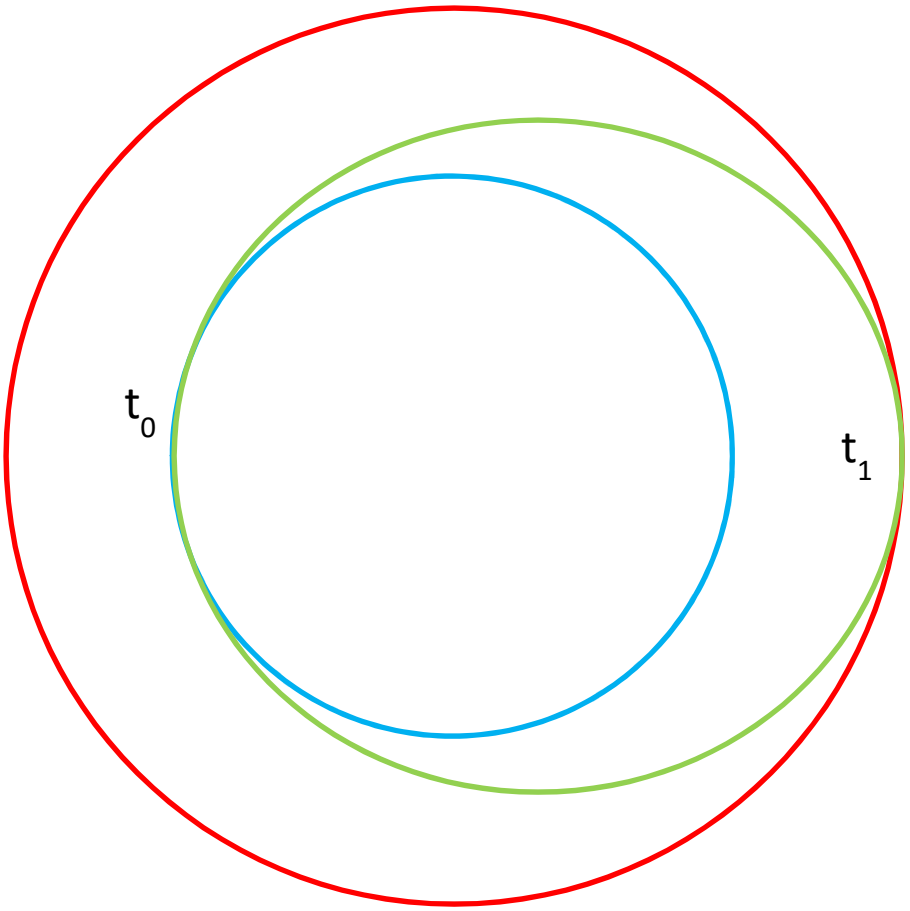


Figure 45. Transfer orbit with start at perihelion and end at aphelion, a Hohmann transfer.

respectively. At our two points we have the radii, r_0 and r_1. We also know the angle between the two points (the difference in the true anomalies) and we know the times. Solving our equation above for the angle we have

$$\alpha = \cos^{-1}\left[\frac{a(1-e^2)-r}{er}\right] + \delta$$

So if we the known difference in angles is Δα, then

$$\Delta\alpha = \cos^{-1}\frac{a(1-e^2)-r_0}{er_0} - \cos^{-1}\frac{a(1-e^2)-r_1}{er_1}$$

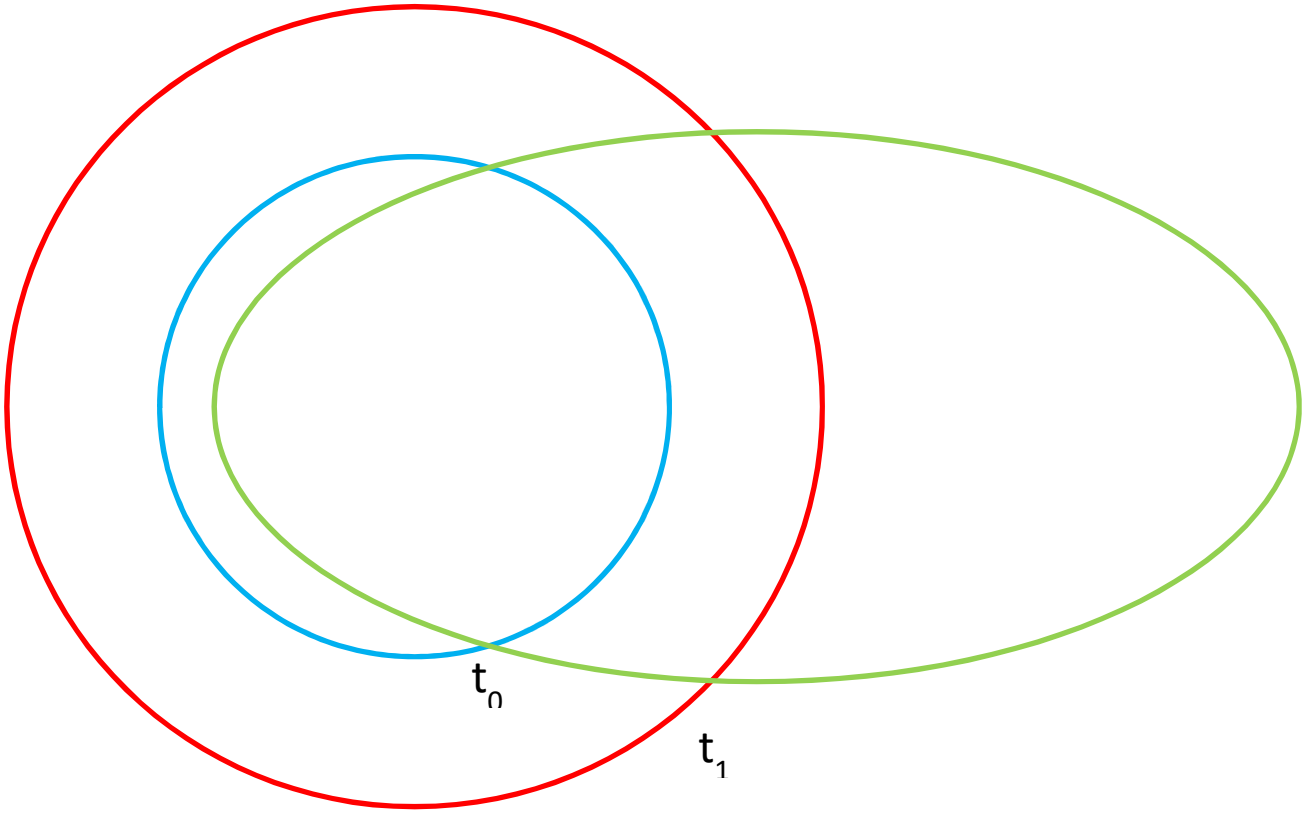


Figure 46. Rapid transfer orbit

With this we can find the appropriate value for the eccentricity as a function of the semimajor axis, or vice versa.

For each pair of eccentricity and semimajor axis we can compute the time between the intersection points on the transfer orbit. When we get a time which matches our desired interval we have found a viable transfer orbit.

Let's restate this in a more concrete example. We might be considering an orbit from the Earth to Mars. Let's assume that we want to start at some time t_0. If we want to get to Mars in a day, we need to compute an orbit between our current position and where Mars will be tomorrow. If we want to take a week, we look for the path between our current positions and Mars' in one week. For each time we consider we can evaluate the anticipated cost of the two impulses to start and finish the journey. Naturally as we allow more time, the fuel costs go down, but there is a minimum. We will find then that if we change the start time of the journey we likely can also save. We will find that for a given pair of objects there are natural start times, and journey durations which minimize the impulse required for the journey. For the Earth-Mars journey, every two years there's a minimal 8-9 month orbit.

In these cases the transfer is between perihelion to aphelion (or the reverse), and the period of the transfer is half the total period of the transfer orbit. Since the semimajor axis of the

transfer orbit is the average of the two orbits, the period can be calculated quickly. E.g., if we approximate Mars as having a semimajor axis of 1.5 au, then the semimajor axis of the transfer orbit is 1.25 au. The period of that orbit is just $1.25^{1.5}$ years or about 1.4 years. So the transfer should take about 0.7 years. Not surprisingly we tend to see new Mars missions launching every two years in just these orbits. If you run O&IB in the past, you will see Mars missions launching from the Earth a few months before each Earth/Mars opposition.

These minimal transfer orbits, called Hohmann transfers, have us boost our rocket from the near circular Earth orbit, to an elliptical orbit where the perihelion is at Earth's orbit and the aphelion is at Mars'. This first boost adds energy to the orbit – we accelerate pretty much in the direction of our orbital motion. It would be exactly in line if planet's orbits were circular and in the same plane. The boost is timed so that just as we reach Mars' orbit, the planet is there when the rocket arrives at aphelion. At this point the rocket is moving slower than Mars, so a second boost accelerates the rocket to match Mars' orbit. This acceleration is in the direction of Mars' motion. Both accelerations are in the direction of the orbit. Since the Earth and Mars orbit in slightly different plans, the boosts also need to slightly adjust the direction of the angular momentum vector of the rocket, from Earth's to the transfer orbit, and from the transfer orbit to Mars.

If we were computing a transfer orbit to Venus, our initial and final impulses would both be opposite the direction of the orbit. We'd initially slow the rocket so that the orbit would fall towards Venus. We should reach perihelion and Venus simultaneously, but the rocket will be moving faster than the planet, so a second deceleration matches the orbit.

More realistic Hohmann transfers between bodies are more complex, since the gravity of both the Earth and Mars can be used to substantially decrease the impulse required. We are typically transferring from an orbit around the Earth to an orbit around Mars. However this doesn't affect the interplanetary part of the orbit, just the details of leaving Earth orbit and entering Mars orbit.

Calculating Hohmann transfers

The characteristics of the Hohmann transfer between two reasonably circular orbits in approximately the same plane are easy to understand. The semimajor axis is the average of the semimajor axes of the source and destination. We get the ellipticity immediately since the inner planet's semimajor axis gives the perihelion and the outer planet the aphelion. So

$$\frac{1+e}{1-e} = \frac{r_1}{r_0}$$

Synodic periods

The synodic period of two planets is the period between oppositions, when the planets are closest together, aligned on the same side of the Sun. There is one opportunity for a Hohmann transfer during each synodic period. The synodic period of Earth and Mars is about 780 days or a little over two years. That's the longest synodic period between Earth and any of the other planets. The synodic period for two objects with nearly the same period is very large, but if the periods are very different, then the synodic period becomes essentially the shorter of the periods of the two bodies. Generally if the periods of two bodies are $p_0$and p_1, then the synodic period between them is

$$s_{01} = \frac{p_0 p_1}{p_1 - p_0}$$

where we have assumed p_0 is the shorter of the two periods.

This assumes the orbits are both in the same direction. If we have one prograde and one retrograde orbit, then the denominator is the sum of the two periods. If the orbits are not in substantially the same plane, or one or both orbits are highly elliptical, then while the synodic orbit may be defined, it may not be as useful a concept. The astrodynamics of a trip after one synodic period may be very different from the situation at the start.

A few satellites (e.g., Spitzer; Stereo A and B), have been launched into solar orbits very similar too, but slightly off phase from the Earth. These have very large synodic periods relative to the Earth (or each other). For two objects with very similar periods, the synodic period is essentially the mean period divided by the difference in periods. The synodic period with these satellites is many decades.

Transfer orbits with perihelion or aphelion matching a planetery orbit.

We can also consider a transfer where we the start is at the perihelion of the transfer, but the transfer orbit's aphelion is outside the destination's orbit. Since we're starting at perihelion, we know that at that time the true anomaly is 0, so we have

$$r_0 = \frac{(1 - e^2)a}{1 + e} = (1 - e)a$$

Or equivalently

$$a = \frac{r_0}{1 - e}$$

So if the angle between the initial and final impulses is Δ, then we have

$$r_1 = \frac{(1 + e)r_0}{1 + e\cos\Delta}$$

So

$$r_1 + r_1 e \cos \Delta = r_0 + r_0 e$$

And

$$r_1 - r_0 = e(r_0 - r_1 \cos \Delta)$$

Solving for the eccentricity we have:

$$e = \frac{r_1 - r_0}{r_0 - r_1 \cos \Delta}$$

Since we know r_1 is larger than r_0, the numerator is always positive, but for small values of Δ, the denominator is negative. So for $\Delta < \cos^{-1} \frac{r_0}{r_1}$ this gives a negative eccentricity. This says that if we need to change the radius a lot in a small angle, then the only orbits that match cannot have the r_0 at perihelion, counter to our starting assumption. If the angle between two points is just a little larger than this limit, then the transfer orbit will be extremely hyperbolic with large values for the eccentricity. Conversely, if the two radii do not differ by much, then the eccentricity of the orbit can be small

We'll get the same result if we try to make the outer point the aphelion of the orbit. That's only possible if the angle between the entry and exit from the transfer orbit is large enough.

More general solutions

If we return to our equation

$$\Delta\alpha = \cos^{-1} \frac{a(1-e^2) - r_0}{e r_0} - \cos^{-1} \frac{a(1-e^2) - r_1}{e r_1}$$

First consider the situation where Δα is 0. Then

$$\cos^{-1} \frac{a(1-e^2) - r_0}{e r_0} = \cos^{-1} \frac{a(1-e^2) - r_1}{e r_1}$$

We can get a solution with any eccentricity if $r_1 = r_0$, but we also satisfy the equation when $e = 1$. Then we just have

$$\cos^{-1} \frac{r_0}{r_0} = \cos^{-1} \frac{r_1}{r_1}$$

Recall that the eccentricity of all radial orbits is 1.

If we go back to our equation for the radius as a function of angle,

$$r = \frac{(1-e^2)a}{1+e\cos(\alpha-\delta)}$$

we see that we can eliminate the dependency on the semimajor axis by looking at the ratio

$$\frac{r_1}{r_0} = \frac{1+e\cos(\alpha_0-\delta)}{1+e\cos(\alpha_1-\delta)}$$

If we just define the angle $\theta \equiv \alpha_0 - \delta$ then we have $\alpha_1 - \delta = \theta + \Delta$. So

$$\frac{r_1}{r_0} = \frac{1+e\cos(\theta)}{1+e\cos(\theta+\Delta)}$$

We can now solve for the ellipticity as a function of true anomaly of the inner event. We get

$$e = \frac{r_1 - r_0}{r_1\cos(\theta+\Delta) - r_0\cos\theta}$$

The numerator is clearly positive (unless the two radii are equal), but denominator may be positive or negative. Negative values that would lead to unphysical values for the eccentricity are thus are not feasible. The maximum value for the denominator gives the minimum value for the eccentricity, and since it will go to 0 for some value of θ there will be an orbit for all larger values of eccentricity.

However the denominator is simply the sum of two out of phase cosine functions, and when we proved that orbits are elliptical we noted that sums of sines and cosines with the same frequency can be combined into a single sine or cosine function.

So

$$r_1\cos(\theta+\Delta) - r_0\cos\theta = A\cos(\theta+\Omega)$$

where

$$A = [{r_1}^2 + {r_0}^2 - 2r_1r_0\cos\Delta]^{1/2}$$

And

$$\Omega = \tan^{-1}(-r_1\sin\Delta, r_1\cos\Delta - r_0)$$

So the minimum value for the eccentricity comes when the denominator of our equation for the eccentricity is largest, when it has a value of A. Then the minimum eccentricity for a given separation angle Δ is just

$$e_{min} = \frac{r_1 - r_0}{[r_1{}^2 + r_0{}^2 - 2r_1 r_0 \cos\Delta]^{1/2}}$$

There is no maximum, since we approach a near infinite speed, near straight line trajectory between the two points as the eccentricity goes to infinity.

The denominator is negative for half of the possible angles, θ. This makes sense, for any given orbit, if we look at a point on the orbit some fixed delta of true anomaly away, then half the time that point will be closer than the current point and half the time further away. If we are measuring from what we take as the inner point, then half of these points do not meet our criteria.

The denominator is positive so long as $-\pi/2 < \theta + \Omega < \pi/2$, i.e., so long as the cosine of $\theta + \Omega$ is positive. [With proper care to make sure that we are handling the cyclic nature of the cosine appropriately.] This is defines the range of true anomalies that we should consider in computing the transfer orbit. So for a given set of transfer points, we compute the orbits as a function of the true anomalies in the allowed range. We can then select the orbit which has the appropriate change in the mean anomaly so that the orbit takes the appropriate time.

Our procedure is then:

1. Pick start time, t_0, for the transfer. The starting point, $\boldsymbol{p_0}$, is the location of the orbit of the start body at that time.
2. Pick a desired interval for the transfer.
3. Compute the position of the end body at the end of the interval at t_1. This gives the position of the endpoint of the transfer, $\boldsymbol{p_1}$.
4. Compute the radii of the start and end points, r_0 and r_1, and angle between the two vectors, Δ.
5. Calculate A and Ω for this transfer.
6. Now use the secant method to determine the orbit needed for the transfer.
7. Pick a valid true anomaly to start with. Then at each step
    a. Compute the semimajor axis.
    b. Compute the period of the orbit (or the time scale parameter for hyperbolic orbits)
    c. Find the eccentric/hyperbolic anomalies of the orbit at the start and stop times.
    d. Find the mean anomalies of the orbit at the start and start times and convert to actual elapsed time.

e. Use the actual elapsed time versus the desired elapsed time to get a better estimate of the orbit.

8. Now that the transfer orbit has been determined, compute the impulses required at the start and stop times.

We can nest this procedure in additional loops to find better durations or start times for the transfer.

Lambert's Theorem

Note that in our approach above the only inputs we use are the radii of the two points and the angle between them, or equivalently by the law of cosines, the distance between the two points. We look at possible orbits as a function of the true anomaly of the first point and iteratively get the orbit that takes some desired time. Here we are assuming orbits in a central potential whose mass is known. This result, that only the radii, separation and time of flights are needed to determine the orbit is known as Lambert's theorem, and the determination of the transfer orbit between two points in a given time is Lambert's problem. Note that there may be a couple of solutions for a given interval – typically one prograde and one retrograde.

Changes in Velocity in a Hohmann Transfer

Consider a transfer orbit from Earth to Mars. The transfer orbit has a semimajor axis that is the average of the semimajor axes of the Earth and Mars, 1 and 1.524 AU. So the transfer orbit has a semimajor axis of 1.262 AU. We know that the energy of the orbit is determined only by its semimajor axis (it's independent of eccentricity), so this orbit has a specific binding energy of 1/1.262 the binding energy of the Earth's orbit. [We can compare the circular orbits for a given semimajor axis. There the gravitational potential energy is just -GM/a, with a positive kinetic energy half of that magnitude, so the binding energy of the orbit is –GM/(2a].

If we add energy to the transfer orbit in a short period of time, the position, and thus the potential energy of the orbit does not change. So we need to add kinetic energy to the orbits such that the

$$\frac{(v_0 + \Delta v)^2}{2} - {v_0}^2 = -p\frac{{v_0}^2}{2}$$

On the right we have the desired total specific energy of the desired transfer orbit as a ratio of our starting orbit where p is the ratio of the starting specific energy of the orbit to the transfer specific energy. If we are doing transfers between two circular orbits from radius r_0 to r_1 then $p = \frac{2r_0}{r_0+r_1}$. For the Earth-Mars transfer that's about 0.79.

On the left the first term is the kinetic energy after we make a boost, while the second term is just the gravitational potential energy at the Earth (or whatever the starting orbit is), which we recall is just twice the average orbital kinetic energy. Rearranging we get

$$\Delta v^2 + 2v_0\Delta v + (p-1){v_0}^2 = 0$$

We can solve this for Δv since it's a quadratic equation we have two roots, one with a small positive change in the velocity and one with a large negative Δv.

$$v_0 = v_0\left[\pm\sqrt{2-p} - 1\right]$$

The first corresponds to the prograde transfer orbit, while the second points out that in principle we could do transfer using a retrograde orbit, but as we'll see the Δv requirements are prohibitive.

For our Earth-Mars transfer, the radical evaluates to about 1.1 and we get solutions of 0.1 v_0 and -2.1 v_0. IF we approximate the Earth's velocity as 30 km/s, then the prograde Hohmann transfer orbit requires a Δv of about 3 km/s, while the retrograde transfer orbit needs a completely unfeasible 63 km/s.

As we transfer further out in the system than Mars, p goes to 0, and the initial velocity of the prograde transfer orbit (i.e., $v_0 + \Delta v$) approaches $\sqrt{2}v_0$, the escape velocity.

If we want to make a transfer orbit to nearer the sun, then p is greater than 1, and both roots of the equation are negative. Even in the prograde transfer we still need to brake our orbital motion, by some value less than the current orbital velocity. However, even in the orbit touching the surface of the Sun, p cannot exceed 2. The semimajor axis of the transfer orbit cannot be smaller than half our starting radius, so we cannot more than double the specific total energy of the orbit. As we transfer further and further inwards we reach the limit of a radial orbit with a single solution. Here we completely dissipate all of our orbital velocity. Our probe hangs motionless for an instant before it begins its plunge towards the Sun. Since this requires a Δv of 30 km/s, it's well beyond currently feasible technologies.

We can repeat this analysis to understand the impulse required at the end of the transfer, i.e., when our probe arrives at Mars. We compute the Hohmann transfer from Mars to Earth, but since we're trying to transfer **from** the transfer orbit to the **Mars** orbit, we invert the direction of the acceleration. So instead of slowing from a Mars orbit to the transfer orbit, we accelerate the transfer orbit to match Mars'.

The Oberth Effect

We see that it takes a fairly modest 3 km/s of velocity added to the Earth's orbit to get into a transfer orbit to Mars. Since the Earth has an escape velocity of about 11 km/s, we might think we need a total Δv of 14 km/s. Certainly we could use that much… First we'd get into an escape orbit burning up our first 11 km/s of Δv. Then after we'd coasted far from the Earth moving very slowly away we'd add the 3 km/s we need to get into the transfer orbit. But this is not the most efficient use of our fuel.

What happens if, instead of starting with an initial acceleration just gives us escape velocity of 11 km/s, we immediately achieve 12 km/s. At the point (still in low Earth orbit) where we finish this initial acceleration, 11 km/s is enough to achieve escape velocity. If we're going faster than 11 km/s, then at infinity we'll retain all of the kinetic energy we have from the velocity in excess of 11 km/s. But kinetic energy goes as the square of the velocity. So if we start our orbit with 12 km/s, then at infinity we will need to retain a velocity of $\sqrt{144 - 121} = 4.8$ km/s to store all of that excess kinetic energy. That's more than we need for our transfer orbit to Mars! So our total energy budget for getting into the Hohmann transfer is just a little bit larger than the escape velocity from the Earth.

The key is that when we want to change the energy of an orbit, we want to make the changes in velocity whenever we are moving fastest. We discussed this earlier when we noted that although two orbits with the same semimajor axis and differing eccentricities have the same energy, it's easier to escape from the more eccentric orbit, since we get a bigger oomph from adding velocity at perihelion.

These gains in efficiency by making velocity changes when the rocket is moving fastest are called the Oberth effect after Hermann Oberth who first described them in the early 20th century. If we are concerned with savings of the order of a few km/s – and with current technologies we definitely are, remember that log term in the rocket equation – then the Oberth effect needs to be considered in any realistic trajectory. This comes into play in the injection and termination of transfer orbit, and also during planetary encounters: a small acceleration during a close Jovian encounter can make a major change to an orbit looking for a boost.

The Oberth effect may seem a little like voodoo: depending upon where we do our acceleration we get lots of extra energy. However there's no problem with conservation of energy here. In looking only at the payload, we've neglected that the rocket is explicitly not a closed system, we are throwing away our reaction mass and we need to think about how that affects the energy balance.

When we accelerate in low Earth orbit, the reaction mass is expelled at a point where it has a much higher (more negative) potential energy, than it would if we were to do our burn near

apogee on some extended orbit. We don't have to lug the reaction mass into the higher orbit. And when we expel the reaction mass from the rocket with some ejection velocity, v_e, the net velocity of the reaction mass is the difference between the orbital velocity and v_e, which can be much smaller than v_e, so that the ultimate kinetic energy of the reaction mass is less. The reaction mass likely ends up in a very tightly bound orbit around the Earth. Presumably it falls into and blends with the atmosphere very quickly. The reaction mass is where we find the negative energy that balances additional kinetic energy for our payload. Since the fuel/reaction mass typically takes up much or most of the weight of our rocket we cannot ignore it.

Transfer limits

In E&IB we look at the orbits between two points in the solar system: the point where we are starting (e.g., the Earth's location at the start date), and the location we wish to get to (e.g., Mars' location some specific time hence). This specifies two points in the solar system. We then consider all of the orbits that connect these two points.

There are two high-velocity limits. We simply travel very very quickly on a direct line between the two points. This is the high-eccentricity orbit, as we move faster and faster, the eccentricity goes to infinity.

We can also travel in a hyperbolic orbit plunging very close to the Sun. Since we are idealizing the Sun as a point mass, we can go very fast, but pass just infinitesimally outside the origin so that the Sun's gravity whips us around to the to the desired outward direction. The orbit effectively looks like two straight lines, one from our origin to the Sun and the second from the Sun to the destination.

We can get the asymptotes to the hyperbola the by looking at our radial equation for the hyperbola

$$r = \frac{a(1-e^2)}{1+e\cos\tau}$$

We can use this to determine the eccentricity of the Sun-grazing limit orbit, since we know the vectors to the start and stop positions. These have a known separation δ, and we want half this angle to be where the denominator goes to 0. The perihelion (at least in the limit as we approach this orbit), is offset from the starting point by $180° - \frac{\delta}{2}$. So

$$1 + e\cos\left(180° - \frac{\delta}{2}\right) = 0$$

$$e = \frac{1}{\cos\frac{\delta}{2}}$$

Thus this orbit has a well-defined, finite eccentricity. The orbit is a V with the Sun at the vertex. The true anomaly of 0, extends from the bottom of the vertex. The true anomaly of the start of the Sun-grazing transfer orbit is just $\frac{\delta}{2} - 180°$ and correspondingly the true anomaly of the end position is 180°-$\frac{\delta}{2}$. The mean and hyperbolic anomalies are not defined however.

While the straight line/direct orbit has an infinite eccentricity, the start and stop points have well defined anomalies. We can draw a perpendicular from the (straight line) orbit to the Sun, and the angles between this perpendicular and the lines drawn from the Sun to the start and end points, give the starting and ending true anomalies.

Calculating the time of a Hohman transfer.

Recall that the semimajor axis of the Hohman transfer is the average of the semimajor axes of the origin and destination planets. Its period is just

$$P_T = P_1\left[\frac{a_0 + a_1}{2a_1}\right]^{\frac{3}{2}}$$

where we are using the semimajor axes of the origin and destination planets and the period of the destination. The Hohman transfer will take exactly half of the period of the transfer orbit, so the terminus will be exactly 180 degrees around the Sun from the origin. During the transfer the destination planet will move by

$$\Delta_1 = 180°\frac{P_T}{P_1}$$

I.e., the destination planet will have moved a bit less (assuming that we're moving outward so that $P_T < P_1$) than the half a circle the spacecraft moved.

So at the start of the transfer it must have had a longitude $180° - 180°\frac{P_T}{P_1}$ degrees ahead of the origin planet's longitude. That ensures that it gets to the destination at the same time as our spacecraft. So we leave the origin planet when the longitude of the destination planet leads by

$$180°\left(1 - \left[\frac{a_0 + a_1}{2a_1}\right]^{\frac{3}{2}}\right)$$

This depends only on the ratio of the semimajor axes.

For a trip from Earth to Mars, with a major axis ratio of about 1.5, this means that we would leave the Earth when its longitude measured from the Sun is just under $45°$ behind Mars. Of course all of this is assuming perfectly circular orbits in the same plane. For the return trip, we expect the Earth to move more than the spacecraft – or Mars – but we can use the same formula (but indexing the Earth with 1) to see that the Earth should be about $70°$ behind Mars when we start.

If that's not where we are now, then we can just wait: Sometime in the next synodic period the offset will be right. The angular velocities of Earth and Mars are just reciprocals of their periods, and the difference between them gives us the longitude the Earth gains each second on Mars.

While an idealized Hohman transfer goes from perihilion to aphelion (or the reverse), this is not quite true for realistic transfers. If planet orbits were restricted exactly to the ecliptic, then we would always be able to find a perfect perihelion/aphelion transfer – where the starting location, Sun and end location would all be in an exact line. However since the planets' orbits are not co-planar, we will find that when we pick the moment where the longitudes of the planetary positions are appropriate for the transfer, e.g., the Earth is trailing Mars by the proper angle, we will find that the source and destination z-values are not exactly correct for a Hohman transfer. I.e., the z value of perihelion needs to be the negative of the z-value at aphelion modula the ratios of the perihelion and aphelion distances and this is unlikely to be exactly true for the transfer. However for planets other than Pluto orbits are close to the same plane, so that the start and end points of a realistic transfer can be within a degree or two of aphelion and perihelion.

However the three dimensional nature of real orbits means that a direct transfer orbit with no intermediate accelerations may be infeasible. Suppose the Earth is a little above the ecliptic plane (presumably just a tiny amount, since the Ecliptic plane is defined by the nominal Earth orbit) at the start of the some transfer and Mars will be above the ecliptic plane at the time we anticipate the transfer ending. It's very difficult to get an orbit that will be above the ecliptic on diametrically opposite sides of the Sun. When we try we'll get an orbit where we need to add large amounts of velocity perpendicular to the ecliptic plane, so that as we approach Mars we are plunging through the ecliptic.

This is clearly undesirable but we have little 'leverage' when we are restricted to impulses at the beginning and end of the transfer. To fix this we will calculate the transfer as a two dimensional problem and deal with the Z component of the transfer in a mid-course correction. There are undoubtedly many ways to approach this but we use the following procedure which has the advantage that it is symmetric with regard to the start and finish orbits. It should also be well defined (though not necessarily appropriate for the best transfer), except when

attempting a transfer between two bodies in the same plane where one orbit is retrograde to the other.

Three Dimensional Transfers

To accommodate the three dimensional transfer, our first step is to approximate it as a two dimensional problem. Initially we have the time we begin and end the transfer, $t_0$and t_1. Using the orbits of the planets we have a start position and velocity and an end position and velocity. We use these positions and velocities of the planets to determine the two orbital planes, and get the normal vectors to these planes. We take the average of these normals (and renormalize) to get a vector perpendicular to an average orbit plane.

We project the positions of the planets onto the average plane. However we scale the radii of the planets so that they are at the same radius as in the original orbit. We similarly project the velocities into the projection plane, but then rescale so that the speed of the planets is the same as in the original system. [It's unclear if this scaling is needed… Further research.]

In this average plane we compute the transfer orbits using our standard techniques. We find the point in the transfer orbit exactly halfway through the transfer, at t_M, and save this point,M.

We go back to our full three dimensional system, and calculate a transfer between our start point and M during the interval from t_0 to t_M and a transfer between M and our end point in the interval t_M to t_1. The two semi-transfers will have different velocities at time t_M. So we effectively add an impulse at this point: the mid-course correction. So long as the original orbit planes are close, we anticipate that this mid-course correction will be a small burn that accommodates the change in angle of the orbits. It won't put the transfer orbit in the same plane as the destination, but at this point in the orbit, there is a lot more 'leverage' in accommodating the differences in the orbit plane.

Projection to a plane.

Consider a plane which goes through the origin (as all of our orbit planes do when we put the primary at the origin). Then the plane satisfies the equation

$$ax + by + cz = 0$$

for some value of (a,b,c). We see that the origin is included in the plane, since the point (0,0,0) manifestly satisfies the equation. However if we consider the vector $\boldsymbol{A} = (a, b, c)$ and the position vector $\boldsymbol{x} = (x, y, z)$ then this just says that

$$\boldsymbol{A} \cdot \boldsymbol{x} = 0$$

So the vector $\boldsymbol{A}$ is perpendicular to the radius vector for any point in the plane, i.e., it's a normal to the plane.

Suppose we want to project some arbitrary point $\boldsymbol{P} = (p_x, p_y, p_z)$ to the plane. We want to shift along a normal to the plane to do the projection. Lets normalize $\boldsymbol{A}$ so that it has unit length. We assume that $\boldsymbol{P}$ is not on the plane so

$$\boldsymbol{A} \cdot \boldsymbol{P} = d$$

where $d \neq 0$. Suppose we now consider the point $\boldsymbol{P} - d\boldsymbol{A}$**.** This is the original point shifted some distance d along the perpendicular to the plane. If we compute

$$\boldsymbol{A} \cdot (\boldsymbol{P} - d\boldsymbol{A}) = \boldsymbol{A} \cdot \boldsymbol{P} - d\boldsymbol{A} \cdot \boldsymbol{A} = d - d = 0$$

Here we've used the used the fact that we normalized $\boldsymbol{A}$ to a unit vector. The point $\boldsymbol{P} - d\boldsymbol{A}$ is the original point $\boldsymbol{P}$ projected onto our plane along the normal to the plane. The scalar d is just the distance between the point and the plane (in whatever units we are measuring $\boldsymbol{P}$ in).

We can use the same approach to find the component of velocity (or any other vector) in the plane.

Appendices

Table 1. Table of notation and key equations

Symbol	Name	Description	Relevant equations
a	Semimajor axis	Half the length of the major axis. See also the formula for r	$P = 2\pi\sqrt{\frac{a^3}{GM}}$ $E = -\frac{GMm}{2a}$ $a = \frac{\lvert\boldsymbol{r}\rvert(1 + e\cos\tau)}{(1 - e^2)}$
b	Semiminor axis	Half the length of the minor axis	$b = a\sqrt{1 - e^2}$ (ellipse) $b = a\sqrt{e^2 - 1}$ (hyperbola)
	Center of mass		$\boldsymbol{C}_{mass} = \frac{\sum m_i \boldsymbol{r}_i}{\sum m_i}$ $\boldsymbol{v}_{mass} = \frac{\sum m_i \boldsymbol{v}_i}{\sum m_i}$
$\boldsymbol{e}$	Eccentricity vector	See equations for semiminor axis.	$\boldsymbol{e} = \frac{\boldsymbol{v} \times \boldsymbol{h}}{GM} - \hat{\boldsymbol{r}}$
e	Eccentricity	Shape parameter for ellipse	$e = \lvert\boldsymbol{e}\rvert$

α	Eccentric anomaly	(see also hyperbolic anomaly)	$MA = \alpha - e\sin\alpha$ $\tan\frac{\tau}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{\alpha}{2}$
E	Energy	Binding energy of orbit.	$E = -\frac{GMm}{2a}$ $E = -\frac{GMm}{r} + 0.5mv^2$ $\lvert L\rvert = GM\sqrt{\frac{e^2-1}{2E}}$
G	Gravitational constant		$\boldsymbol{F}_m = -\frac{GMm}{\lvert\boldsymbol{r}_m - \boldsymbol{r}_M\rvert^2}\frac{\boldsymbol{r}_m - \boldsymbol{r}_M}{\lvert\boldsymbol{r}_m - \boldsymbol{r}_M\rvert}$ $\boldsymbol{F}_m = -\frac{Gm\frac{M^3}{(m+M)^2}}{\lvert\boldsymbol{r}_m\rvert^2}\frac{\boldsymbol{r}_m}{\lvert\boldsymbol{r}_m\rvert}$ $\frac{d^2r}{dt^2} - r\omega^2 = -\frac{GM}{r^2}$
$\boldsymbol{h}$	Specific angular momentum		$\boldsymbol{h} = \frac{\boldsymbol{L}}{m}$
α	Hyperbolic anomaly	(also see eccentric anomaly)	$MA = \alpha - e\sinh\alpha$ $\tan\frac{\tau}{2} = \sqrt{\frac{e+1}{e-1}}\tanh\frac{\alpha}{2}$
i	Inclination	Tilt of the orbit relative to the reference plane	$i = \cos^{-1}(\frac{L_z}{\lvert\boldsymbol{L}\rvert})$
$\boldsymbol{L}$	Angular momentum		$\boldsymbol{L} = m\boldsymbol{r}\times\boldsymbol{v}$ $L = mr^2\frac{d\theta}{dt}$ $\frac{L}{m} = \sqrt{GMa(1-e^2)}$ $\lvert L\rvert = GM\sqrt{\frac{e^2-1}{2E}}$
$\boldsymbol{l}$	Line of Nodes	Vector pointing to ascending node	$\boldsymbol{l} = \hat{\boldsymbol{z}}\times\boldsymbol{L}$
MA	Mean Anomaly	Measure of orbital position that increases uniformly with time	$MA = \alpha - e\sin\alpha$ (ellipse) $MA = \alpha - e\sinh\alpha$ (hyperbola) $MA = 2\pi(t - t_0)/P$
Ω	Longitude of the ascending node	Longitude of point on reference plane where orbit rises above plane.	$\Omega = \text{atan2}(l_y, l_x)$

ν	Argument of periapsis	Angle between ascending node and periapsis.	$\lvert\sin\nu\rvert = \frac{\lvert\boldsymbol{l}\times\boldsymbol{e}\rvert}{\lvert\boldsymbol{l}\rvert\,\lvert\boldsymbol{e}\rvert}$ $\cos\nu = \frac{\lvert\boldsymbol{l}\cdot\boldsymbol{e}\rvert}{\lvert\boldsymbol{l}\rvert\,\lvert\boldsymbol{e}\rvert}$
P	Period	Duration of a single orbit.. A nominal period may be defined for hyperbolic orbits.	$P = 2\pi\sqrt{\frac{a^3}{GM}}$
$\boldsymbol{r}$	Position,Radius	The vector/distance between the focus of the orbit and the object	$r = \frac{a(1-e^2)}{1+e\cos\tau}$ (ellipse) $r = \frac{a(e^2-1)}{1+e\cos\tau}$ (hyperbola) $\boldsymbol{r} = r\begin{bmatrix}\cos\Omega\cos(\nu+\tau) - \sin\Omega\sin(\nu+\tau)\cos i\\ \sin\Omega\cos(\nu+\tau) + \cos\Omega\sin(\nu+t)\cos i\\ \sin(\nu+t)\sin i\end{bmatrix}$
r_{peri}, r_{aph}		Distance at perihelion and aphelion	$r_{peri} = (1-e)a$ $r_{aph} = (1+e)a$
	Rotation Matrix	Transformation from orbit plane to standard coordinates	$\begin{bmatrix}\cos\Omega\cos\nu - \sin\Omega\sin\nu\cos i & -\cos\Omega\sin\nu - \sin\Omega\cos\nu\cos i & \sin\Omega\sin i\\ \sin\Omega\cos\nu + \cos\Omega\sin\nu\cos i & -\sin\Omega\sin\nu + \cos\Omega\cos\nu\cos i & -\cos\Omega\sin i\\ \sin\nu\sin i & \cos\nu\sin i & \cos i\end{bmatrix}$
τ	True Anomaly	The angle between periapsis svector and the current location. See also equations for r.	$\tan\frac{\tau}{2} = \sqrt{\frac{1+e}{1-e}}\tan\frac{\alpha}{2}$ (ellipse) $\tan\frac{\tau}{2} = \sqrt{\frac{e+1}{e-1}}\tanh\frac{\alpha}{2}$ (hyperbola)
m, M	Masses	The mass of the orbiting and central body	
μ	Reduced mass	Effective mass of virtual particle when combining separations and velocities	$\mu = \frac{mM}{m+M}$
$\hat{\boldsymbol{r}}\hat{\boldsymbol{\theta}}\hat{\boldsymbol{x}}\hat{\boldsymbol{y}}\hat{\boldsymbol{z}}$	Unit vectors	Direction of motion when corresponding coordinate is increased infinitesimally	$\hat{\boldsymbol{r}} = \cos\theta\ \hat{\boldsymbol{x}} + \sin\theta\,\hat{\boldsymbol{y}}$ $\hat{\boldsymbol{\theta}} = -\sin\theta\ \hat{\boldsymbol{x}} + \cos\theta\,\hat{\boldsymbol{y}}$
$\boldsymbol{v}$	Velocity		${v_{peri}}^2 = \frac{GM}{a}\frac{1+e}{1-e}$ $\frac{d\theta}{dt} = \frac{\sqrt{GMa(1-e^2)}}{r^2}$ $\frac{dr}{d\theta} = \frac{ae(1-e^2)\sin\theta}{(1+e\cos\theta)^2}$

			$$\boldsymbol{v}_{orbplane} = \frac{d\theta}{dt}\left(\frac{dr}{d\theta}\cos\theta - r\ \sin\theta\right)\hat{\boldsymbol{x}} + \frac{d\theta}{dt}\left(\frac{dr}{d\theta}\sin\theta + r\ \cos\theta\right)\hat{\boldsymbol{y}}$$ $$\boldsymbol{v} = \frac{h}{r}\left[\frac{e}{a(1-e^2)}\sin\tau\,\boldsymbol{r} + \begin{bmatrix} -\cos\Omega\sin(\nu+\tau) + \sin\Omega\cos(\nu+\tau)\cos i \\ -\sin\Omega\sin(\nu+\tau) + \cos\Omega\cos(\nu+\tau)\cos i \\ \sin i\cos(\nu+\tau) \end{bmatrix}\right]$$

Deriving the double angle formulae for Sine and Cosine Type equation here.

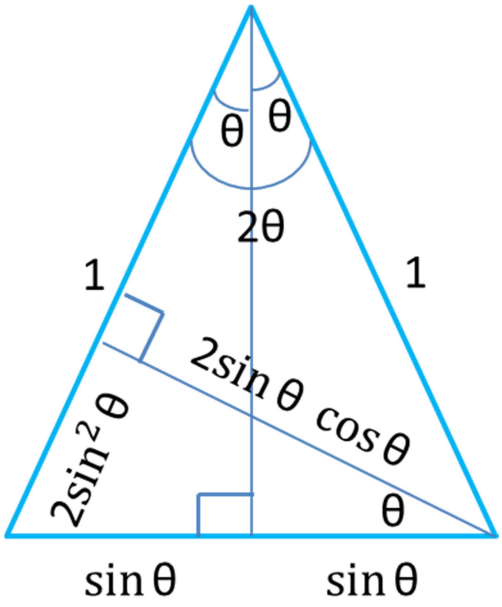


Figure 47. Double angle formula

The figure above is all we need to derive the double angle formulae for sine and cosine.

We start with an isosceles triangle where the equal sides have length 1. The third side is the base.

We can split this into two right triangles, left and right, by drawing a perpendicular to the base through the vertex joining the equal length sides. If the original vertex had an angle 2θ then each of the new triangles has an opening angle of θ. Then the base of the original isosceles triangle has a length of $2\sin\theta$.

Now we draw a line from the right base vertex perpendicular through the left side. This splits the original isosceles triangle into two new triangles, top and bottom. The hypotenuse of the new bottom triangle is just the base of the original isosceles triangle: it has a length $2\sin\theta$. However since it shares the angle in the bottom left with the left triangle, we know that the bottom triangle's angle on the right is just θ. We can immediately compute the lengths of the other two sides of this triangle as $2\sin\theta\cos\theta$ and $2\sin^2\theta$.

The hypotenuse of the new top triangle is just 1, so its sides are just the sine and cosine of 2θ. We can read these off as

$$\sin 2\theta = 2\sin\theta\cos\theta$$

$$\cos 2\theta = 1 - 2\sin^2\theta = \cos^2\theta - 1$$

Starting with these we can derive the double angle formula for the tangent and the half angle formulae for sine, cosine and tangent.

Ten minute Intro to Calculus

Occasionally in our derivations we've used a bit of calculus. This can seem pretty intimidating to those who've not studied it, but there's nothing remarkably esoteric about calculus. Say you drive half an hour and the odometer in your car increases by about 30. So you went 30 miles in half an hour. Is it surprising that your speedometer was probably averaging around 60 miles per hour? If not you basically understand differential calculus.

All it says is that if we measure how far we've gone in some small interval of time, the speed we're going at is the change in distance divided by the time interval. When we *differentiate* a function we write it as

$$\frac{df}{dt}$$

Here the d's represent a little delta, a small change. The f on the top is the thing whose rate of change we're trying to find. In our car that's the odometer reading, the distance we have travelled. The t on the bottom represents the thing we are going to change a little bit to see how f changes. So df is how much f changes if we change t by some little amount, dt. E.g., when we see that the odometer increases by .01 miles in .001 hours, then our current speed is .01/.001 miles/hour = 10 miles/hour.

There are just a few things we need to know to do most of the calculus in this presentation.

First the derivative of something with respect to itself is 1.

$$\frac{dt}{dt} = 1$$

How much does the time change each second? Time changes by 1 second every 1 second and 1/1 = 1!

Second the derivative of a constant is 0. How much does 3 change in 1 minute? 3 never changes so our numerator is always 0.

Conversely if we find that the derivative is 0, then the thing we are taking the derivative of is a constant.

Third the derivative of a sum is the sum of the derivatives.

$$\frac{d(x+y)}{dt} = \frac{dx}{dt} + \frac{dy}{dt}$$

If we're travelling in a convoy of two cars and we want to know how the total mileage changes, we just add up the changes for each car individually.

Fourth, there is the very powerful product rule

$$\frac{d(xy)}{dt} = \frac{dx}{dt}y + x\frac{dy}{dt}$$

Five, a little help with the trig functions

$$\frac{d\,\mathrm{si}}{dt} = \cos t \text{ and } \frac{d\cos}{dt} = -\sin t$$

And finally the chain rule which is a bit more esoteric, but not too bad…

Suppose we have two tables: one which records the odometer reading and the amount of fuel left in the gas tank and second that indicates the time and how much gas we had left every 10 seconds. We can treat these as two functions, $D(g)$, is the distance we travelled as a function of how much gas is left in the tank, and $g(t)$ is the amount of gas left in the tank as a function of time. Presumably this is getting smaller with time. If we want to measure our speed we can calculate it is

$$\frac{dD}{dt} = \frac{dD}{dg}\frac{dg}{dt}$$

So in the first table we find out how much distance we went as the gas tank lost a certain amount. Then we see how long it took to lose that much gas. Both of the factors on the right are negative, the factor $dg < 0$, since as we're looking at each successive interval the amount of gas gets less. We're going further as the amount of gas gets less, and the amount of gas is getting less as time is increasing. So the speed, the increase in distance with time is positive. This chain rule is true anytime we have a something which can be expressed as a function of different variables, e.g., the radius of the orbit can be expressed as either a function of time, or a function of angle from the perihelion.

These rules are really powerful when used cleverly. You now know enough, e.g., to calculate the derivative of something modestly esoteric like $\sqrt{t}$. Consider

$$1 = \frac{dt}{dt} = \frac{d(\sqrt{t}\sqrt{t})}{dt} = \frac{d\sqrt{t}}{dt}\sqrt{t} + \sqrt{t}\frac{d\sqrt{t}}{dt} = 2\sqrt{t}\frac{d\sqrt{t}}{dt}$$

We just used the product rule here… But setting the first and last expressions equal we can rearrange and get

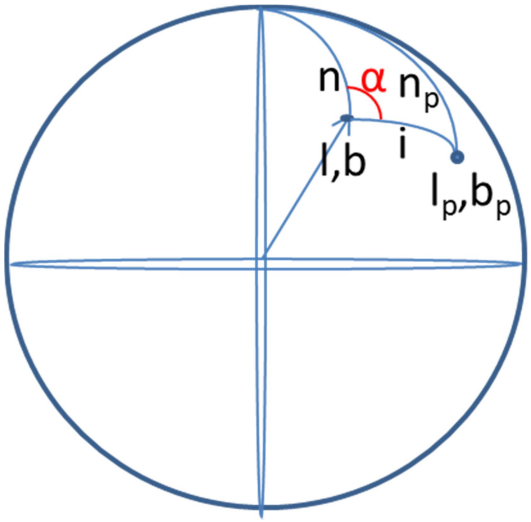
$$\frac{d\sqrt{t}}{dt} = \frac{1}{2\sqrt{t}}$$

Voila! Pretty much all of the calculus we used previously derives from these simple rules. So have fun and don't be intimidated.

References

https://phys.libretexts.org/Bookshelves/Astronomy__Cosmology/Book%3A_Celestial_Mechanics_(Tatum)/09%3A_The_Two_Body_Problem_in_Two_Dimensions/9.08%3A_Orbital_Elements_and_Velocity_Vector

Finding the pole of galaxies



We start with a galaxy at a position (l,b) and want to find where its pole is located at (l_p,b_p). We are given the position of the galaxy, the position angle,α, of the projected major axis, and the inclination of the galaxy from the line of site.

If the inclination is 0, then we are seeing the galaxy face on, so the pole of the galaxy is along the line of sight.

If we consider the colatitudes of the galaxy and the pole point, n and n_p, then

$$n = 90 - b, n_p = 90 - b_p$$

The inclination angle is given as i. The law of cosines tells us that

$$n_p = \sqrt{n^2 + i^2 - 2ni\cos\alpha}$$

This gives us the latitude of the pole. Once we have all three sides of the triangle we can get the angle between arcs to the galaxy and the pole position with either the Law or Sines or Law of Cosines. Since the position angle is constrained to be between 0 and 180 degrees using the Law of Cosines removes any ambiguities and we get

$$l_p - l = \cos^{-1}\frac{i^2 - n^2 - {n_p}^2}{2nn_p}$$

as the offset for the longitude of the pole.

KEPLER ORBIT ELEMENTS TO ECI CARTESIAN COORDINATES CONVERSION

Elliptical Case (02/02/02)

Copied from:
https://web.archive.org/web/20170810015111/http://ccar.colorado.edu/asen5070/handouts/kep2cart_2002.doc

Compute orbital position and velocity at time t given the orbital elements.

$$a, e, i, \omega, \Omega, T \Rightarrow X, Y, Z, \dot{X}, \dot{Y}, \dot{Z}$$

where Ω is the right ascension of the ascending node.

1. Compute the mean anomaly, *MA*, (0° $\rightarrow$ 360°),

$$M = n(t - T) \qquad n = \sqrt{\frac{\mu}{a^3}}$$

2. Compute the eccentric anomaly, *EA*, (0° $\rightarrow$ 360°),

$$MA = EA - e \sin EA$$

3. Compute the true anomaly, ν, (0° $\rightarrow$ 360°),

$$\tan \frac{\nu}{2} = \left[\frac{(1+e)}{(1-e)} \right]^{1/2} \tan \frac{EA}{2}$$

If *EA* < 180° then ν < 180° (There is no need to do a quadrant check. The equation above will automatically yield the correct value of ν.)

4. Compute the radius, *r*,

$$r = a(1 - e \cos EA) = \frac{p}{1 + e \cos \nu} = \frac{a(1 - e^2)}{1 + e \cos \nu}$$

5. Compute the specific angular momentum, *h,*

$$h = (\mu a(1-e^2))^{1/2}$$

6. Compute the position components, *X,Y,Z*

$$X = r(\cos\Omega\cos(\omega+\nu) - \sin\Omega\sin(\omega+\nu)\cos i)$$
$$Y = r(\sin\Omega\cos(\omega+\nu) + \cos\Omega\sin(\omega+\nu)\cos i)$$
$$Z = r(\sin i\sin(\omega+\nu))$$

7. Compute the velocity components, $\dot{X}, \dot{Y}, \dot{Z}$

$$\boldsymbol{v} = \frac{h}{r}\left[\frac{e}{a(1-e^2)}\sin\tau\,\boldsymbol{r} + \begin{bmatrix} -\cos\Omega\sin(\nu+\tau) + \sin\Omega\cos(\nu+\tau)\cos i \\ -\sin\Omega\sin(\nu+\tau) + \cos\Omega\cos(\nu+\tau)\cos i \\ \sin i\cos(\nu+\tau) \end{bmatrix}\right]$$

$$\dot{X} = \frac{Xhe}{rp}\sin\nu - \frac{h}{r}(\cos\Omega\sin(\omega+\nu) + \sin\Omega\cos(\omega+\nu)\cos i)$$

$$\dot{Y} = \frac{Yhe}{rp}\sin\nu - \frac{h}{r}(\sin\Omega\sin(\omega+\nu) - \cos\Omega\cos(\omega+\nu)\cos i)$$

$$\dot{Z} = \frac{Zhe}{rp}\sin\nu + \frac{h}{r}\sin i\cos(\omega+\nu)$$

8. To obtain ECF (Earth Centered and Fixed) coordinates

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_{ECF} = T_{ECF}^{ECI}\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_{ECI}, \begin{bmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \end{bmatrix}_{ECF} = T_{ECF}^{ECI}\begin{bmatrix} \dot{X} + \omega_{\oplus}Y \\ \dot{Y} \_ \omega_{\oplus}X \\ \dot{Z} \end{bmatrix}_{ECI}$$

where $\omega_{\oplus}$ ≡ rotation rate of the earth (7.2921158553 x 10^{-5} rad/sec)

$$T_{ECF}^{ECI} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

θ is the Greenwich hour angle of the Earth's prime meridian, i.e. the angle between the inertial X-axis and the ECF X-axis.

ECI CARTESIAN COORDINATES TO KEPLER ORBIT ELEMENTS CONVERSION
Elliptical Case (02/02/02)

[This section attempts to transcribe the following PDF back into a Word document. Note that we've generally tried to copy the original text even where that uses different conventions than our document. Since some of the text was actually pasted and other entered using Word's equation editor, the Greek fonts are inconsistent between the numbered text and the equations.

https://web.archive.org/web/20160418175843/https://ccar.colorado.edu/asen5070/handouts/cart2kep2002.pdf]

Compute orbital elements given ECI position and velocity at time t for elliptical motion

$$X, Y, Z, \dot{X}, \dot{Y}, \dot{Z} \Rightarrow a, e, i, \omega, \Omega, T$$

1. Compute the specific angular momentum and check for a degenerate orbit,

$$\vec{r} \times \vec{v} = \vec{h} = h_x \vec{u}_x + h_y \vec{u}_y + h_z \vec{u}_z$$

$$h = \left({h_x}^2 + {h_y}^2 + {h_z}^2 \right)^{1/2}$$

2. Compute the radius, *r*, and velocity, v,

$$r = (X^2 + Y^2 + Z^2)^{1/2}$$
$$v = (\dot{X}^2 + \dot{Y}^2 + \dot{Z}^2)^{1/2}$$

3. Compute the specific energy, *E*, and verify elliptical motion,

$$E = \frac{v^2}{2} - \frac{\mu}{r}$$

4. Compute semi-major axis, *a*,

$$a = -\frac{\mu}{2E}$$

5. Compute eccentricity, *e*,

$$e = \left(1 - \frac{h^2}{a\mu}\right)$$

6. Compute inclination, *i*, ($0^\circ \rightarrow 180^\circ$),

$$\cos i = \frac{h_z}{h}$$

7. Compute right ascension of the ascending node, Ω , ($0^\circ \rightarrow 360^\circ$),

$$\Omega = \text{atan2}(h_x, -h_y)$$

8. Compute argument of latitude, ω+ν , ($0^\circ \rightarrow 360^\circ$),

$$\omega + \nu = \text{atan2}\left(\frac{Z}{\sin i}, (X \cos \Omega + Y \sin \Omega)\right)$$

9. Compute true anomaly, ν, ($0^\circ \rightarrow 360^\circ$),

$$\cos\nu = \frac{a(1-e^{2)}-r}{er}$$

If $r \cdot v > 0$ then $\nu < 180°$

Or use

$$\nu = \text{atan2}\left(\sqrt{\frac{p}{\mu}}(\vec{r}.\vec{r}), p-r\right), \text{ where } p = a(1-e^2)$$

10. Compute argument of periapse, ω, ($0^\circ \rightarrow 360^\circ$),

$$\omega = (\omega + \nu) - \nu$$

11. Compute eccentric anomaly, *EA*, ($0^\circ \rightarrow 360^\circ$),

$$\tan\frac{EA}{2} = \left[\frac{1-e}{1+e}\right]^{1/2} \tan\frac{\nu}{2}$$

EA is in the same half plane as ν. This equation will yield the correct quadrant for EA.

12. Compute the time of periapse passage, *T* (note th1at EA must be in radians),

$$T = t - \frac{1}{n}(EA - e\sin EA),\ n = \sqrt{\frac{\mu}{a^3}}$$

---

[i] We follow the approach used at http://galileo.phys.virginia.edu/classes/152.mf1i.spring02/KeplersLaws.htm

[ii] The standard arctangent function is not really the proper function to use here since its range is only -90 to 90 degrees. The atan2(y,x) or arg(y,x) function that has the range -180 to 180 degrees should be used here, but may be less familiar to users.

[iii] Following https://ocw.mit.edu/courses/aeronautics-and-astronautics/16-346-astrodynamics-fall-2008/lecture-notes/lec_01.pdf