Conic Sections

Return to Contents

In spite of the obvious importance of Cartesian coordinates, we will focus most of the remainder of this and the next few parts on polar coordinates. In particular, it will be important for us to understand what "conic sections" are (ellipses, hyperbolas, and parabolas) and how they are described in polar coordinates. The reason for that is that John Kepler found it both expedient and necessary to formulate his laws concerning the motion of the planetary satellites around the Sun in terms of conic sections. He had no idea why. Newton realized, however that the reason lay in the interpretation of his Universal Law of Gravitation within the context of Polar coordinates. Of all systems of coordinates, polar coordinates seem most naturally adapted to his law. They are the easiest system of coordinates to use to deduce Kepler's laws from his single, simple principle.

Some background. In the 16th century, astronomer and monk Nicholas Copernicus quietly announced an hypothesis that would eventually shake the world. At the time (fortunately for Nicholas) no one noticed much. Until that time, conventional wisdom had it that the Sun and all the celestial bodies moved in stately (almost) circular orbits around the Earth. The Earth, *Terra Firma*, stood still. This was common sense. No one can blame the ancients for believing that.

But the ancient astronomers, going back to Ptolemy, had great difficulty hanging on to the common-sense notion, because, while common sense told them that the planets and the Sun and Moon should move along circular orbits around the Earth at uniform speeds, the actual observational data seemed to contradict this. Mars, for example, would seem to move along nicely for a while, then would slow down in the sky, and actually appear to move backwards! We will see this for ourselves later in **Harmony of the Spheres**. This could be observed for other planets also, although of course, not the Sun. As such discrepant data accumulated, Ptolemy made a valiant attempt to "save the phenomena" by postulating that the planets moved on circular paths within circular paths (wheels within wheels). Ptolemy's theory was very adept, and became the dominant hypothesis for almost 2000 years, until Copernicus, frustrated with the *ad hoc* nature of those hypotheses, put forth his bold treatise. All those "jigglings" and "perturbations" of the motions of the planets were due to the fact that *we* were moving with them. That is easy to say. Now.

Everyone believed in any case that these objects moved in circles. Circles were the only conceivable paths that the planets could take, either around the Earth or the Sun, because circles have a certain perfection in their symmetry. If the planets did not move in circles, there would have to be a **reason**. Copernicus believed that the planets moved in circles about the Sun.

John Kepler

Now John Kepler, a century or so later, set out to *prove* Copernicus' hypothesis with hard observational data. That data was collected over many years by the gifted (and erratic) observational astronomer, Tycho Brahe. To his great consternation, Kepler was not able to show that if the planets moved in circles, then we would see what we see. The curves were almost circles, but not quite. It took Kepler many years to rediscover the work of Greek Geometer Apollonius, and to learn the names of the curves that observation dictated the planets moved along. They were not circles, but **ellipses**. Kepler never found the **reason** for this departure from perfection. He left the world with a riddle, stated in his three "laws" of planetary motion. We will study that riddle in **Harmony of the Spheres**. Isaac Newton solved it. And you will, too.

In 1596 he wrote Mysterium cosmographicum, which led to discussions with Galileo and Tycho Brahe. His Astronomia nova (1609) contained the first two of what became Kepler's Laws; the third law appeared in 1619 in his Harmonice mundi. These laws were the result of calculations based on Brahe's accurate observations, which Kepler published in the Tabulae Rudolphinae (1627).

Kepler's laws of planetary motion are three mathematical statements derived from observation, and from Brahe's records. While he sought initially simply to confirm the Copernican view that planets moved in circular orbits around the Sun, he was astonished (and dismayed) to discover that the planetary paths were not circles, but were ellipses. An ellipse is a type of "conic section", as you will see below.

Kepler's Laws

Kepler's laws describe the revolutions of the planets around the sun.

1. The first law states that the shape of each planet's orbit is an ellipse with the sun at one focus.
2. The second law states that if an imaginary line is drawn from the sun to the planet, the line will sweep out equal areas in space in equal periods of time for all points in the orbit.
3. The third law states that the ratio of the cube of the semimajor axis of the ellipse (i.e., the average distance of the planet from the sun) to the square of the planet's period (the time it needs to complete one revolution around the sun) is the
same for all the planets.

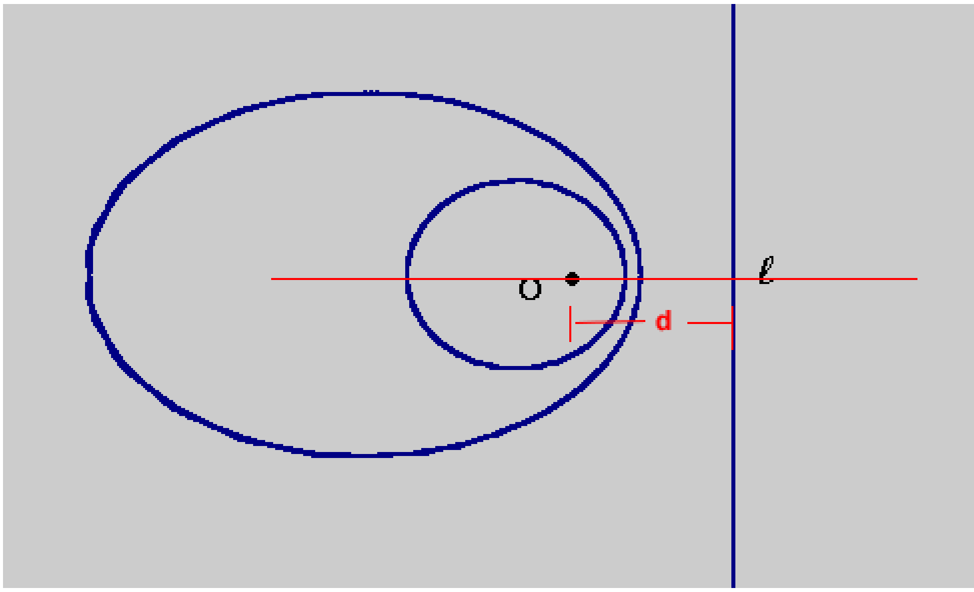
As we have said, Newton gave a physical explanation of Kepler's laws with his laws of motion and law of gravitation. In fact, Newton invented the Calculus as a mathematical system of ideas that could express his explanation.

Conic Section (description)

We want now to lay the foundation for the work to come by explaining what these curves were that Kepler saw in the sky. Apollonius had called them "conic sections" having no idea at all what their scientific significance would be 1500 years later. Imagine a cone (for example, an ice-cream cone). And imagine that you "slice" it with a plane. The intersection of these surfaces will be a curve. If the plane were perpendicular to the axis of symmetry of the cone, then this curve would be a circle (or perhaps just a point). But if it were tilted just a little, the curve would be an ***ellipse***. An ellipse is like a flattened circle. It has a remarkable pair of points called ***foci***, and this makes them useful for the design of lenses, mirrors, and satellite dishes.

Now if you tilt the plane some more, you find that it crosses the cone in two separate regions. This sort of curve, divided into two parts, is called an ***hyperbola***. That is another type of conic section. And just on the boundary between ellipses and hyperbolae, is the single-sheeted intersection called a ***parabola***. That is the third type of conic section. As it happens, all three types of conic section have a simple description using polar coordinates. We take that up now.

Imagine that a line is drawn at a distance d from the origin ***O*** in the polar coordinate plane. Call the line ***l***. Given a point ***P*** we may measure its distance from the origin, and may also measure its distance from the line *l*. The locus of points for which those two measurements are in a certain fixed ratio is a conic section. As we vary the ratio, we get different conic sections. The following picture illustrates the case for an ellipse. For the smaller ellipse, the ratio of distance to O to distance to *l* is $\frac{1}{2}$. For the larger one, that ratio is $\frac{3}{4}$.



Now if we imagine that the line perpendicular to *l* that passes through O is the ***R***–axis (angle ***A***= 0) then this description of conic sections is especially simple. Let ***d*** be the distance between O and *l*. Suppose we choose a point (A, R). Then the distance from that point to O is of course: *R*.

Question 1: Show that the distance from the point to the line *l* is $|d - R \cdot \cos(A)|$. You will not need calculus, only geometry, to do this.

End of Question

Therefore the ratio is simply

$$\frac{R}{|d - R \cdot \cos(A)|}$$

We will, from now on, measure the angle *A* in radians instead of degrees. If we consider the locus of all points (A, R)

such that

$$\left(\frac{R}{d - R \cdot \cos(A)}\right)^2 = \varepsilon^2$$

for a positive number ε, this defines a conic section of *eccentricity* ε.

When the eccentricity is less than 1, we have an ellipse, as in the pictures, when it is equal to 1 we have a parabola, and when it is greater than 1, a hyperbola.

The following simple observation is crucial for Newton's deduction, as we will see in **Harmony of the Spheres**.

Question 2: Show that the locus of points that satisfy

$$\left(\frac{R}{d - R\cdot\cos(A)}\right)^2 = \varepsilon^2$$

is the given by the polar function (representing *R* as a function of *A*) where as we said, *A* is measured in radians.

(0.1) $$R(A) = \frac{\varepsilon\cdot d}{1 + \varepsilon\cdot\cos(A)}$$

That is, solve for *R*.

End of Question

Focus-locus property of ellipses in focus-directrix form

We show a curious fact about ellipses defined by directrix *d* and eccentricity $0 < \varepsilon < 1$. They are the locus of points that satisfy

$$\left(\frac{R}{d - R\cdot\cos(A)}\right) = \varepsilon$$

where *R* is the distance from the origin and *A* is the angle (measured in radians). We saw that we can write *R* as a function of the angle:

$$R(A) = \frac{\varepsilon\cdot d}{1 + \varepsilon\cdot\cos(A)}$$

Question 3: If the origin of the conic **O** is the origin of the *x*-axis and the directrix is perpendicular to the *x*-axis so that the directrix intersects the *x*-axis at $(d, 0)$ show that in the case $0 < \varepsilon < 1$ the ellipse intersects the *x*-axis at points

$$\left(\frac{\varepsilon\cdot d}{1+\varepsilon}, 0\right) \text{ and } \left(-\frac{\varepsilon\cdot d}{1-\varepsilon}, 0\right)$$

Also show that the center of the ellipse (the average of these points) is

$$\left(-\frac{\varepsilon^2\cdot d}{1-\varepsilon^2}, 0\right)$$

Finally, show that the other focus, at equal distance from the left intersection as the origin is from the right intersection is the point

$$\left(-\frac{2\cdot\varepsilon^2\cdot d}{1-e^2},0\right)$$

End of Question

Now the foci of the ellipse are therefore

$$(0,0)\ \text{and}\ \left(-\frac{2\cdot\varepsilon^2\cdot d}{1-e^2},0\right)$$

If $P=(x,y)$ give the Cartesian coordinates of a point on the ellipse, we know that we can write it also as

$$P=\left(\frac{\varepsilon\cdot d\cdot\cos(A)}{1+\varepsilon\cdot\cos(A)},\frac{\varepsilon\cdot d\cdot\sin(A)}{1+\varepsilon\cdot\cos(A)}\right)$$

from the fact that

$$R(A)=\frac{\varepsilon\cdot d}{1+\varepsilon\cdot\cos(A)}$$

Now the curious fact about ellipses is this.

Theorem 1: For the ellipse, if we represent the foci as points C_1 and C_2 and if P is an arbitrary point on the ellipse, then

$$\|P-C_1\|+\|P-C_2\|=K$$

where *K* is constant, that is, it does not depend on *P*.

Proof: Suppose C_1 is the origin $(0,0)$ and $P=\left(\frac{\varepsilon\cdot d\cdot\cos(A)}{1+\varepsilon\cdot\cos(A)},\frac{\varepsilon\cdot d\cdot\sin(A)}{1+\varepsilon\cdot\cos(A)}\right)$. Then

$$\|P-C_1\|=\frac{\varepsilon\cdot d}{1+\varepsilon\cdot\cos(A)}$$

Let us examine $\|P-C_2\|$

This is

$$\sqrt{\left(\frac{\varepsilon\cdot d\cdot\cos(A)}{1+\varepsilon\cdot\cos(A)}+\frac{2\cdot\varepsilon^2\cdot d}{1-\varepsilon^2}\right)^2+\left(\frac{\varepsilon\cdot d\cdot\sin(A)}{1+\varepsilon\cdot\cos(A)}\right)^2}$$

Factoring the $\varepsilon\cdot d$ and simplifying, we have

$$\varepsilon \cdot d \cdot \sqrt{\left(\frac{\cos(A)}{1+\varepsilon \cdot \cos(A)} + \frac{2 \cdot \varepsilon}{1-\varepsilon^2}\right)^2 + \left(\frac{\sin(A)}{1+\varepsilon \cdot \cos(A)}\right)^2}$$

which becomes

$$\varepsilon \cdot d \cdot \sqrt{\left(\frac{1}{1+\varepsilon \cdot \cos(A)}\right)^2 + \frac{4 \cdot \varepsilon \cdot \cos(A)}{(1+\varepsilon \cdot \cos(A)) \cdot (1-\varepsilon^2)} + \frac{4 \cdot \varepsilon^2}{(1-\varepsilon^2)^2}}$$

and this is the same thing as

$$\varepsilon \cdot d \cdot \sqrt{\left(\frac{1}{1+\varepsilon \cdot \cos(A)}\right)^2 + \frac{4}{(1-\varepsilon^2)^2} - \frac{4}{(1-\varepsilon^2)}\left(\frac{1}{1+\varepsilon \cdot \cos(A)}\right)}$$

which is finally

$$\varepsilon \cdot d \cdot \sqrt{\left(\frac{1}{1+\varepsilon \cdot \cos(A)} - \frac{2}{(1-\varepsilon^2)}\right)^2}$$

Question 4: Show that under the condition $0 < \varepsilon < 1$

$$\frac{1}{1+\varepsilon \cdot \cos(A)} < \frac{2}{(1-\varepsilon^2)}$$

End of Question

We see therefore that

$$\|P - C_2\| = \frac{2 \cdot \varepsilon \cdot d}{(1-\varepsilon^2)} - \frac{\varepsilon \cdot d}{1+\varepsilon \cdot \cos(A)}$$

and so the constant sum is $\frac{2 \cdot \varepsilon \cdot d}{(1-\varepsilon^2)}$ (the distance between the left and right intersections with the axis).

End of Proof

Thus the polar equation

$$R(A) = \frac{\varepsilon \cdot d}{1+\varepsilon \cdot \cos(A)}$$

determines a condition that we will see below is satisfied for a very simple Cartesian equation when the center of the ellipse is translated to the origin. The polar equation is the one that Newton used as we shall see in **Harmony of the Spheres**.

We now return briefly to Cartesian coordinates to derive from the above curious property another equation for conic sections (here translated so that the center is at the origin). The Cartesian representation of ellipses given below will be important later in the final Chapter where we discuss integration and **Kepler's Third Law**.

Focus-locus property of ellipses in Cartesian form

Select two points in the plane separated by a distance $2 \cdot c$. Call these the foci of the conic section. For any positive number $2 \cdot a$, larger than $2 \cdot c$, the distance between these points, consider the locus of points, the sum of whose distances to these points is equal to $2 \cdot a$. This locus is an ellipse. Next consider, for any positive number $2 \cdot a$ smaller than the distance between these points, the locus of points with the property that the absolute value of the difference of distances to these two points is equal to $2 \cdot a$. The locus of these points is a hyperbola.

For our purposes here, we will consider the foci to be opposite points on the x-axis, generating a central conic. No generality is lost. Now, the single equation:

$$\frac{x^2}{a^2} + \frac{y^2}{a^2 - c^2} = 1 \tag{0.2}$$

represents both ellipses and hyperbolae with the stipulation that: $0 < c^2 < a^2$ gives an ellipse and that $0 < a^2 < c^2$ gives a hyperbola. So a single algebraic argument will give both focus interpretations, and of course, we then locate the foci at $(c,\ 0)$ and $(-c, 0)$. We step through the argument here to highlight the symmetry.

Case of an ellipse: $0 < c^2 < a^2$

In general,

$$\left(\sqrt{(x+c)^2 + y^2} + \sqrt{(x-c)^2 + y^2}\right)^2 = 2\left(x^2 + y^2 + c^2\right) + 2\sqrt{\left(x^2 + y^2 + c^2\right)^2 - 4x^2c^2}$$

also, for any (x, y) satisfying (0.2),

$$x^2 + y^2 + c^2 = a^2 + \frac{c^2}{a^2}x^2 \tag{0.3}$$

So these imply (with a little algebra) that

$$\left(\sqrt{(x+c)^2 + y^2} + \sqrt{(x-c)^2 + y^2}\right)^2 = 2\left(a^2 + \frac{c^2}{a^2}x^2\right) + 2\sqrt{\left(a^2 + \frac{c^2}{a^2}x^2\right)^2 - 4x^2c^2} = 2\left(a^2 + \frac{c^2}{a^2}x^2\right) + 2\left(\sqrt{\left(a^2 - \frac{c^2}{a^2}x^2\right)^2}\right) \tag{0.4}$$

And, in this case, $0 < c^2 < a^2$ we must have

$$a^2 - \frac{c^2}{a^2}x^2 > 0$$

so we get:

$$\left(\sqrt{(x+c)^2+y^2}+\sqrt{(x-c)^2+y^2}\right)^2 = 4a^2 \tag{0.5}$$

This says that the sum of distances from (x,y) to $(c,\ 0)$ and $(-c,0)$ is equal to $2\cdot a$.

Case of a hyperbola: $0 < a^2 < c^2$

The same algebraic argument (with the sign reversal) gives:

$$\left(\sqrt{(x+c)^2+y^2}-\sqrt{(x-c)^2+y^2}\right)^2 = 2\left(a^2+\frac{c^2}{a^2}x^2\right)-2\sqrt{\left(a^2+\frac{c^2}{a^2}x^2\right)^2-4x^2c^2} = 2\left(a^2+\frac{c^2}{a^2}x^2\right)-2\left(\sqrt{\left(a^2-\frac{c^2}{a^2}x^2\right)^2}\right) \tag{0.6}$$

which is just what is needed, because now

$$a^2 - \frac{c^2}{a^2}x^2 < 0$$

and we conclude that:

$$\left(\sqrt{(x+c)^2+y^2}-\sqrt{(x-c)^2+y^2}\right)^2 = 4a^2 \tag{0.7}$$

This says that the absolute value of the difference of distances from (x,y) to $(c,\ 0)$ and $(-c,0)$ is equal to $2\cdot a$. This also suggests an obvious duality. Starting with an ellipse:

$$E: \frac{x^2}{a^2}+\frac{y^2}{a^2-c^2} = 1 \text{ with } 0 < c^2 < a^2$$

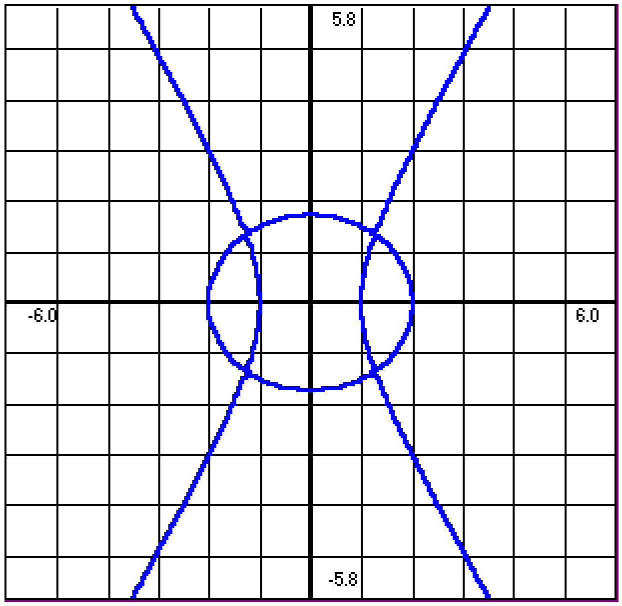
one generates a "dual" hyperbola by interchanging a and c. So we get hyperbola:

$$H: \frac{x^2}{c^2}+\frac{y^2}{c^2-a^2} = 1 \text{ with } 0 < c^2 < a^2$$

The interesting things about E and H are that the foci for E are the intersection points of H with its "major axis" and the foci for H are the intersection points of E with its "major axis" and of course, these share the same center and major axis.

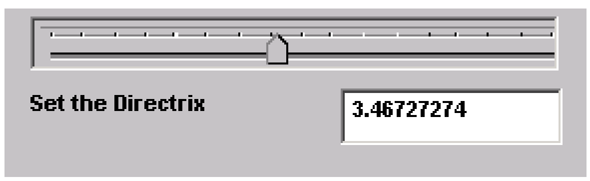
Here is a picture: for ellipse $\frac{x^2}{4}+\frac{y^2}{3}=1$ and hyperbola $x^2-\frac{y^2}{3}=1$

with foci at $(\pm 1,0)$ and (± 2,0) respectively

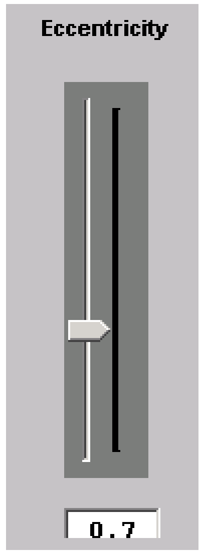


Exploration: Eccentricity and conics

Now the exercise on this page is designed to illustrate the polar representation of conics in terms of eccentricity and directrix given above. You may set the directrix by dragging the thumb in the slider below:

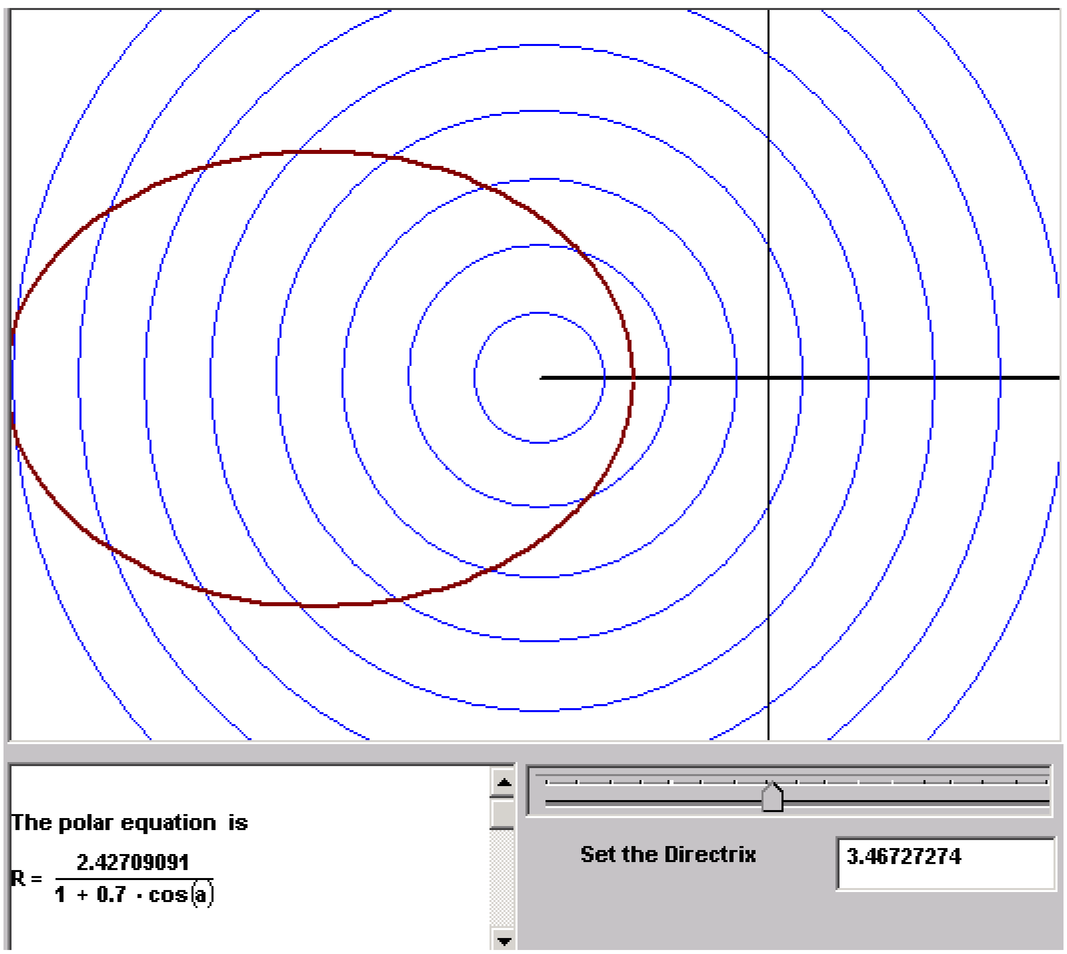


And you may set the eccentricity with the slider on the right (or by right clicking on it and accessing its Settings menu)



And when you press the Draw Conic button: 

you will see the conic and its equation:



Finally, if you click on the polar screen you will see the Cartesian and polar coordinates of the point you selected. Try to click on the conic itself. The system will report information in this form for the conic above of eccentricity **0.7** and directrix:

***x* = 3.46727274**:

```
A = 114.276 degrees, R = 3.464
x = 3.464 *cosine( 114.276 ) = -1.425 ,
y = 3.464 *sine( 114.276 ) = 3.157
Distance to d = 4.891
Ratio of R to this distance = 0.708
```

In this way, you can check for yourself whether these ratios are constant.