Mathematics 309 — Conic sections and their applicationsn

Chapter 1. Introduction to conic sections

1. The basic definitions

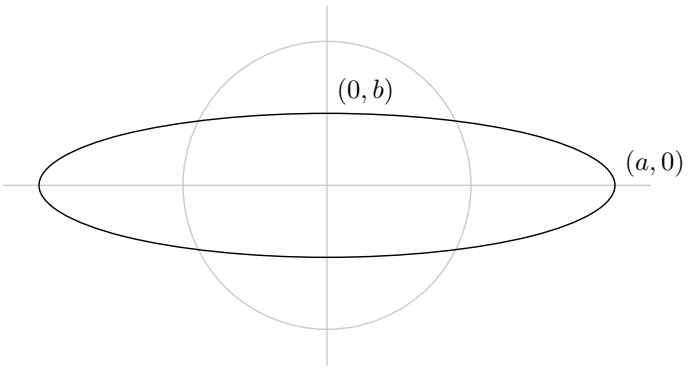
(1) An **ellipse** is obtained from a circle by scaling it in perpendicular directions, say along the coordinate axes, using possibly different scale factors along each axis. If we start with a unit circle

$$x^2 + y^2 = 1$$

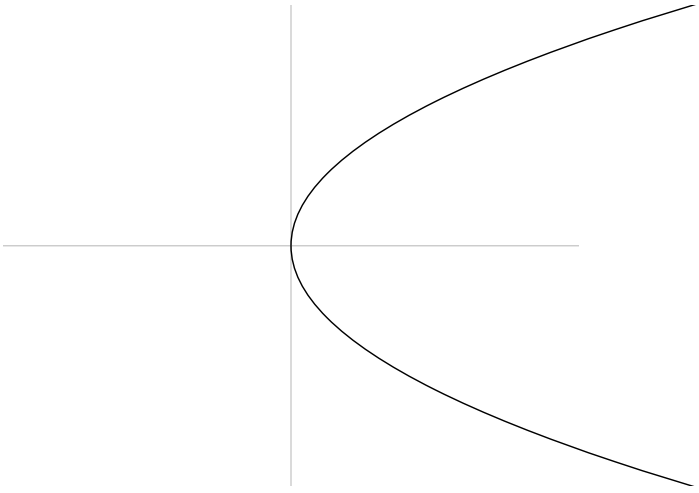
and scale x-values by a and y-values by b, we obtain the ellipse

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$$

since if (x, y) is a point on the new curve then $(x/a, y/b)$ is a point on the unit circle. If $a > b$ then the longest axis has length $2a$, and the shortest one length $2b$. The numbers a and b are called the **semi-axes** of the ellipse, a the **semi-major** axis and b the **semi-minor** axis.

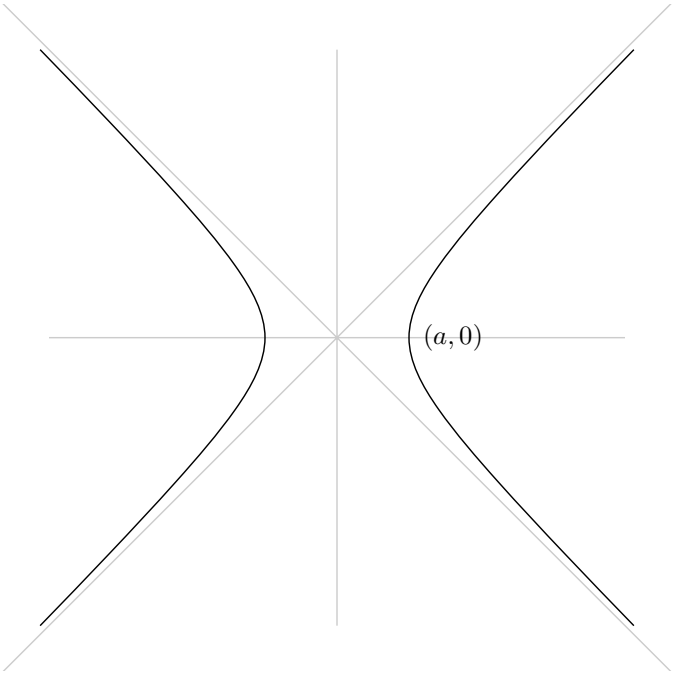


(2) We get a **parabola** from the equation $x = cy^2$.



(3) We get an **hyperbola** from the equation

$$\left(\frac{x}{a}\right)^2 - \left(\frac{y}{b}\right)^2 = 1 \, .$$



We can write the equation for an hyperbola as

$$x = \pm a\sqrt{1 + \left(\frac{y}{b}\right)^2} \, .$$

The choice of signs gives two **branches** of the curve. If $y = 0$ then $x = \pm a$, so the distance between the two branches is $2a$. If y is large then the right hand side is very close to $\pm a|y/b|$, which means that the hyperbola approaches the lines $x = \pm ay/b$ at infinity. These two lines $y = \pm bx/a$ are called the **asymptotes** of the hyperbola.

Exercise 1.1. *If y is large, what is a simple estimate for how far it is from a point (x, y) on the hyperbola to an asymptote?*

A curve is called a **conic section** if it is congruent to one of these, for suitable choices of a, b, or c.

This definitions are straightforward, and they at least allow us to sketch the curves. They do not, however, tell us what these curves have in common, why they make up all of a family of curves, or even why they are interesting. All these things will come. It turns out that neither do these definitions tell us the most important geometrical properties of the curves.

THe usual way to parametrize an ellipse is by means of

$$t \mapsto (a\cos t, b\sin t) \, .$$

In effect, the parameter t is the angle on the unit circle we derived the ellipse from. For hyperbolas, we have the associated parametrization by the hyperbolic functions

$$t \mapsto (a\cosh t, b\sinh t), \qquad \cosh t = \frac{e^t + e^{-t}}{2}, \quad \sinh t = \frac{e^t - e^{-t}}{2} \, .$$

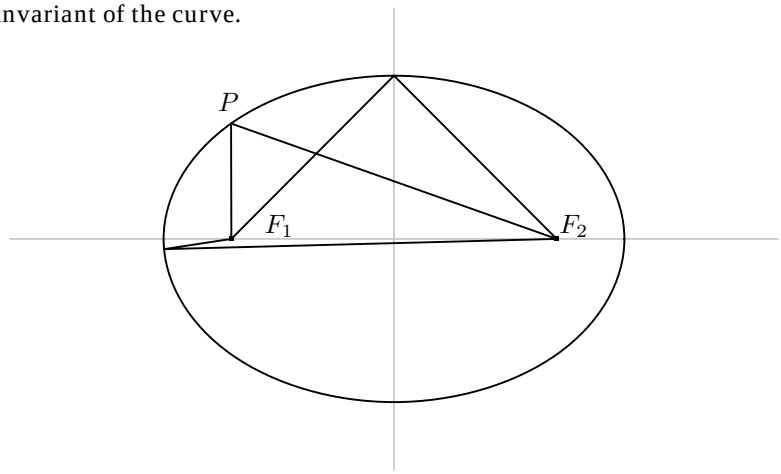
But t here has no geometrical meaning, and it is usually more convenient to parametrize the hyperbola as the union of two graphs $x = \pm a\sqrt{1 - y^2/b^2}$.

2. The focal points

To each of the conic sections we can associate one or more special points called foci (the latin plural of 'focus').

(1) An ellipse has two foci F_1 and F_2 with this property: *If we draw a line from F_1 to a point P on the curve and then another from P to F_2, the total distance $F_1P + PF_2$ doesn't depend on P.*

This sum will be an invariant of the curve.

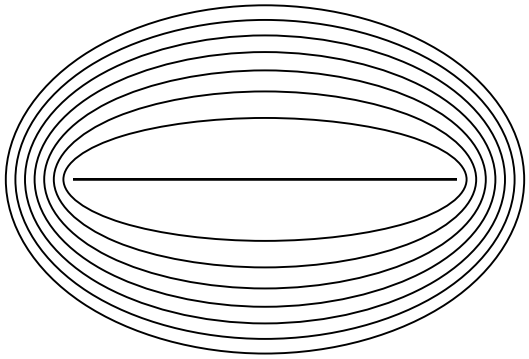


Before we begin to think about why we can find such points, we'll do some exploration. Suppose we are given to start with two points F_1 and F_2 and a distance d. Consider the set of all points P in the plane with the property that $F_1P + PF_2 = d$ (classically, the **locus** of points with this property). What can we say about it?

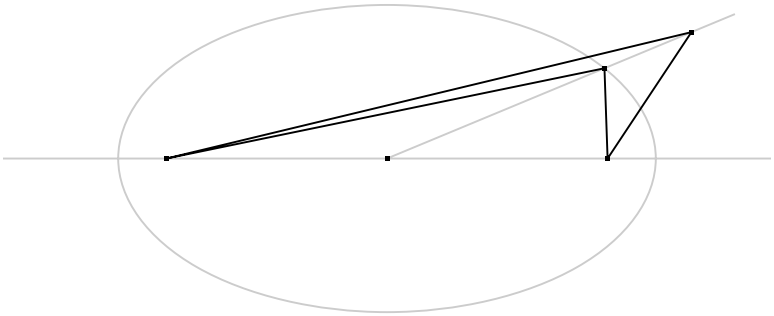
Since a straight line measures the shortest distance between two points, if the distance d is less than the distance F_1F_2, then there won't be any points P with the property we are looking for.

If $d = F_1F_2 = d$, then the points P are exactly the ones on the segment F_1F_2.

The interesting case is when $d > F_1F_2$. What we expect is that as d increases, we get larger and larger curves.



Without doing any calculating at all, I want to show that we must be looking at some kind of closed curve containing both F_1 and F_2.



Let O be the midpoint of the segment F_1F_2. Since $d > F_1F_2$, the point O certainly doesn't lie on the curve. On the other hnad, as the picture shows, if we construct a ray going out from O then as the point P moves out along the ray, $F_1P + PF_2$ increases steadily. This sum is larger than twice OP, so we canmake this sum as large as we want if we go out far enough. Therefore there exists exactly one point P where $F_1P + PF_2 = d$.

Choose a coordinate system so F_1 and F_2 lie on the x-axis and O is the origin. If P lies on the locus then its reflection in the x-axis through F_1 and F_2 will also lie on it, and so will its reflection in y-axis. The locus therefore has a four-fold symmetry with respect to the coordinate axes.

So now we can find where the foci of an ellipse must be located, if in fact it has the focal property. Use the symmetry we just noted. They must be located on the longer axis of the ellipse. Say the coordinates are $(-f, 0)$ and $(f, 0$. Now apply the characteristic property to just two points on the curve, one on each of the axes. If a and b are the semi-major axes of the ellipse, then its equation is

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1 \, .$$

We first consider the path from F_1 to the left hand end extremity $(a, 0)$ of the ellipse, then back through F_1 to F_2. Its length is $(a - f) + a + f = 2a$. Next we consider the path from F_2 to the top point $(0, b)$ and then to F_1. Its length must also be $2a$, and we conclude by Pythagoras' Theorem that we have to have

$$f = \sqrt{a^2 - b^2} \, .$$

Now we want to prove that in fact these two candidate foci are do in fact have the focal property: *If P is any point on the ellipse then $F_1P + PF_2$ is equal to $2a$.*

Let $P = (x, y)$. Then

$$F_1P = \sqrt{(x + f)^2 + y^2}, \qquad F_2P = \sqrt{(x - f)^2 + y^2}$$

and we want to calculate

$$\sqrt{(x + f)^2 + y^2} + \sqrt{(x - f)^2 + y^2} \, .$$

We can expand these out as

$$\sqrt{x^2 + 2xf + f^2 + y^2} + \sqrt{x^2 - 2xf + f^2 + y^2}$$

and then subsitute

$$y^2 = b^2 - b^2x^2/b^2, \qquad f^2 = a^2 - b^2$$

to get

$$\begin{aligned} &\sqrt{x^2 + 2xf + a^2 - b^2 + b^2 - b^2x^2/a^2} + \sqrt{x^2 - 2xf + a^2 - b^2 + b^2 - b^2x^2/a^2} \\ &\qquad = \sqrt{x^2(1 - b^2/a^2) + 2xf + a^2} + \sqrt{x^2(1 - b^2/a^2) - 2xf + a^2} \, . \end{aligned}$$

We can also write

$$f = a\sqrt{1 - b^2/a^2}$$

and then

$$x^2(1 - b^2/a^2) \pm 2xf + a^2 = \left(x\sqrt{1 - b^2/a^2} \pm a\right)^2 .$$

Therefore

$$\sqrt{x^2(1 - b^2/a^2) \pm 2xf + a^2} = \pm\left(x\sqrt{1 - b^2/a^2} \pm a\right) .$$

Since $|x| \leq a$, we choose the sign to get

$$\sqrt{x^2(1 - b^2/a^2) \pm 2xf + a^2} = \left(a \mp x\sqrt{1 - b^2/a^2}\right)$$

and therefore getting $2a$ for the sum.

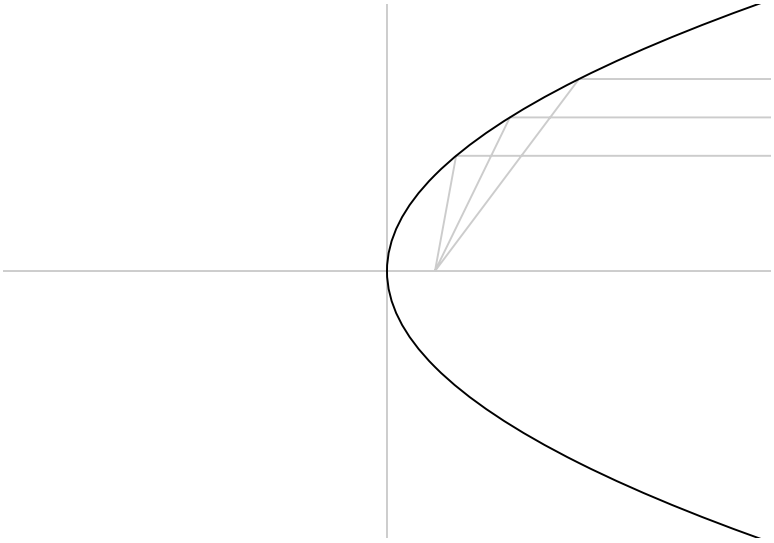
This calculation is a bit tricky and not very enlightening. We shall see later on a very elegant geometric proof of this focal property.

Exercise 2.1. *The argument shows us that*

$$F_1P = a + x\sqrt{1 - b^2/a^2} .$$

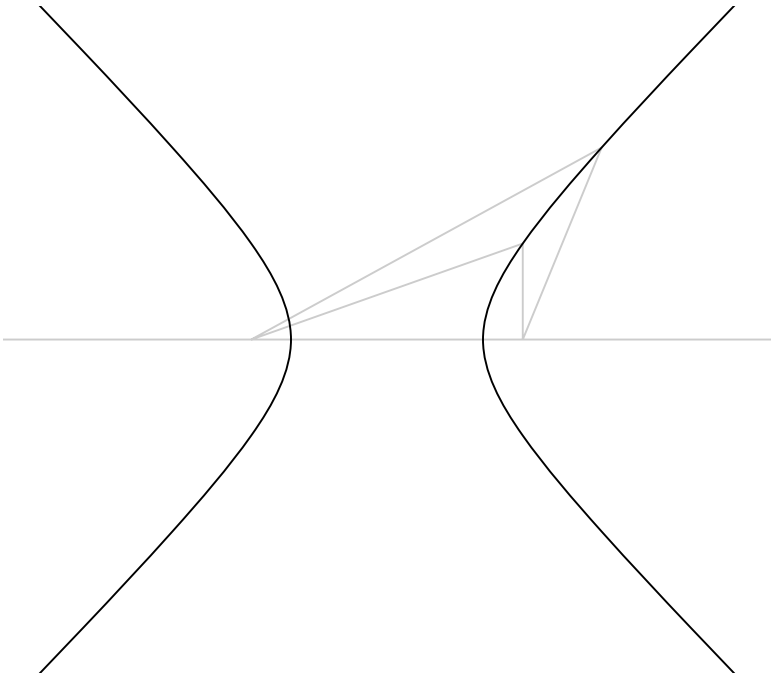
This is a surprisingly simple formula. Is there a direct geometric construction to prove it?

(2) For parabolas, we have the following: *There exists a unique point F such that all rays coming straight in from ∞ pass through F.*



Exercise 2.2. *Find where the focus of the parabola $x = cy^2$ has to be. Prove the focal property for it.*

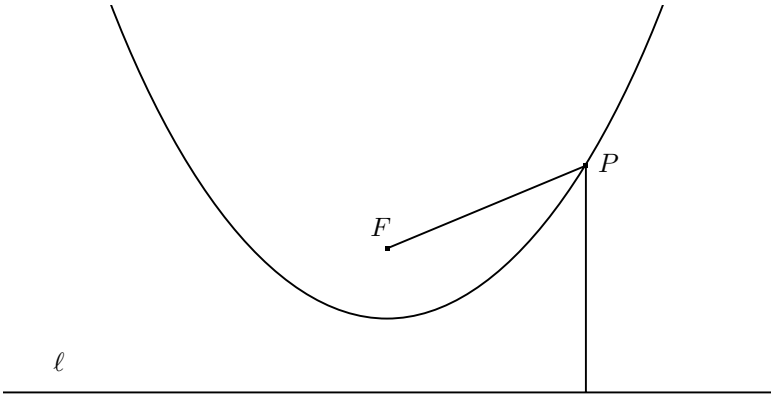
(3) For hyperbolas: *There exist two point F_1 and F_2 such that for all P on one branch of the curve, $F_1P - PF_2$ is independent of P.*



Exercise 2.3. *Find where the focus of the hyperbola $(x/a)^2 - (y/b)^2 = 1$ has to be. Prove the focal property for it. (Hint: Finding the focus is not so simple. There is no direct analogue of the point $(0, b)$ for an hyperbola. But you should be able to use the asymptotes to help you. In other words, if P is a point on an asymptote then as it moves off to infinity the distance $F_1P - PF_2$ should approach a constant. Also, you can easily figure out what this constant is in terms of a and b. At any rate, if you guess the location of the foci correctly your proof will work, and then you will know your guess was correct.)*

3. The focus and the directrix

There is another way to construct the conic sections which at least makes it clear why they all belong in a single family. Choose a number $e > 0$, and a line ℓ, say horizontal. Choose a point F above the line. Plot the curve made up of all points P such that the ratio $FP/d(P, \ell) = e$. Then the curve is a conic section and F is its focus. If $e < 1$ the curve is an ellipse, if $e = 1$ it is a parabola, and if $e > 1$ it is one half of an hyperbola.



Exercise 3.1. *Let ℓ be the x-axis, $F = (0, f)$. Find equations without square roots for the curve corresponding to the constant e. Can you figure out formulas for a and b (or c) in terms of e and f? If $e < 1$, the curve is an ellipse. What is its centre?*

Exercise 3.2. *What kind of a curve are we looking at in the figure just above?*

Exercise 3.3. *Find a formula for the lowest point on the curve in terms of e and f. Ignoring linguistic niceties, I shall call the distance from F to this point the* **perihelion** *distance.*

Exercise 3.4. *The distance from the focus F to a point P on the curve with the same value for y is called the* **semi-latus rectum** *p of the curve. Find a formula for it in terms of e and f. Find a formula for the ratio of p to the perihelion distance.*

The constant e is called the **eccentricity** of the curve. The line ℓ is called a **directrix**.

4. Agreement

How can we see that the curves described in this section are the same as the conic sections we defined in the first section?

First we explore a bit the curves defined in the last section. Suppose that ℓ is the x-axis, and that F is on the y-axis. This is something we can rrange by choosing coordinates properly. There a few points on it that we can plot easily. The condition on P is symmetric with respect to which side of F it lies on, so reflection in the y-axis takes the curve into itself. This suggests that we investigate first to see what point on the y-axis also lie on the curve. Let $F = (0, f)$ and set $P = (0, y)$. If P lies below F but above the x-axis then then $y, f - y > 0$ and P satisfies our condition if and only if

$$y/f - y = e \qquad y = ef - ey, \qquad y = \frac{f}{1+e} \ .$$

If P lies above F the we get similarly

$$y = \frac{f}{1-e} \ .$$

This requires that $e < 1$. If $e > 1$ then the point $(0, f/(1-e))$ will lie on the curve, but it will be below the x-axis. These cases cover all possibilities for P on the y-axis. These facts are consistent with the proposal that we are looking at an ellipse if $e < 1$, a parabola if $e = 1$, and an hyperbola if $e > 1$, but of course they are only weak evidence.

There are two other important points which are easy to locate. These are the ones at the same level as F. Let $P = (x, f)$. Then for $x > 0$ the condition on P is that $x/f = e$ or $x = ef$. In fact, both $(\pm ef, f)$ will be on our curve. The line from F to this point is, as we have, seen called the semi-latus rectum, and if we know where the focus of a conic section is it is easy to locate. Its length is usually called p. Notice that the ratio of p to the perihelion distance is

$$\frac{ef}{f - f/(1+e)} = 1 + e \ .$$

This means that we can tell immediately from this ratio whether we have an ellipse, a parabola, or an hyperbola.

We now want to consider in detail the question: *Suppose we are given a point $F = (0, f)$ and a number $e > 0$. What kind of a curve are we in fact looking at, and what are its parameters?* In other words, we want to justify the claim that it is a conic section.

We start with the simpler question: What is the equation satisfied by all points $P = (x, y)$ such that $FP/d(P, \ell) = e$?

The distance from (x, y) to $(0, f)$ is $\sqrt{x^2 + (y-f)^2}$. Therefore the equation satisfied by P is

$$\frac{\sqrt{x^2 + (y-f)^2}}{y} = e, \qquad x^2 + (y-f)^2 = e^2 y^2$$

and if $e \neq 1$

$$x^2 + y^2(1-e^2) - 2fy = -f^2 = x^2 + (1-e^2)\big(y^2 - 2fy/(1-e^2)\big) = -f^2$$

and after completing the square

$$x^2+(1-e^2)\big(y^2-2fy/(1-e^2)+f^2/(1-e^2)^2\big)=x^2+(1-e^2)\big(y-f/(1-e^2)\big)^2=\frac{f^2}{(1-e^2)}-f^2=\frac{f^2e^2}{(1-e^2)}\,.$$

This is the equation of a conic section with centre at $(0, f/(1-e^2))$. It is an ellipse of $1-e^2>0$ or $e<1$ and an hyperbola if $1-e^2<0$ or $e>1$. If $e<1$ then it can be written in the usual form with

$$a^2=\frac{f^2e^2}{1-e^2},\quad b^2=\frac{a^2}{1-e^2}\,.$$

Note that we now have for an ellipse the relationship

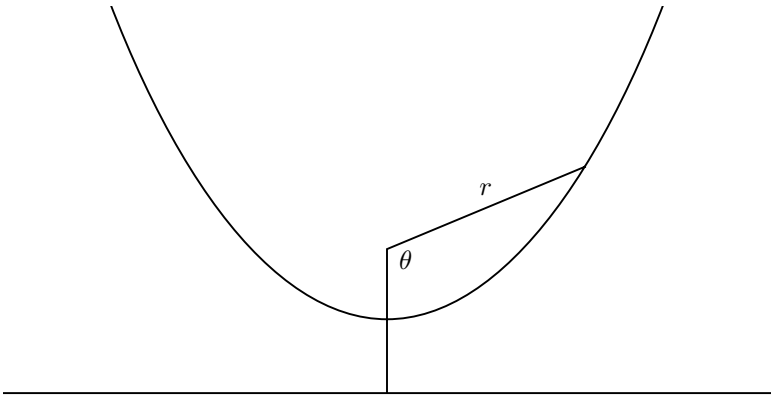
$$e=\sqrt{1-\frac{a^2}{b^2}}$$

where now we revert to earlier situation where the major axis is horizontal. Since the focus point is at $(a^2-b^2,0)$ in the earlier case, we also see that the eccentricity is the ratio of the distance of a focus from the centre to the length of semi-major axis.

Exercise 4.1. *What if $e=1$?*

5. Radial coordinates

We now shift to the following coordinate picture:



That is to say, we choose our origin at one focus and use radial coordinates, with the angle θ measured from the perihelion. This will be convenient in examining planetary motion, for example. We ask: *What is the equation of an ellipse in these coordinates?*

The condition for points on the curve can be expressed as $r/y=e$ or $r=ey$. In turn we have $y=f+r\cos\theta$, where the focus is $(0,f)$. We can also write $f=p/e$ where p is the semi-latus rectum. So we have an equation to solve for r

$$r=e(p/e+r\cos\theta),\qquad r=\frac{p}{1+e\cos\theta};\,.$$

6. Summary

There are several possible ways to define the plane curves known as conic sections. No matter how they are introduced, other descriptions wil be useful in various circumstances. In this chapter I introduced them in terms of algebraic equations more or less centred at the origin and oriented along the coordinate axes, and then gave

an alternate characterization in terms of the focus and directrix. We have not yet seen why they are called conic sections.

There are various parameters associated to any conic section. From now we'll follow these conventions:

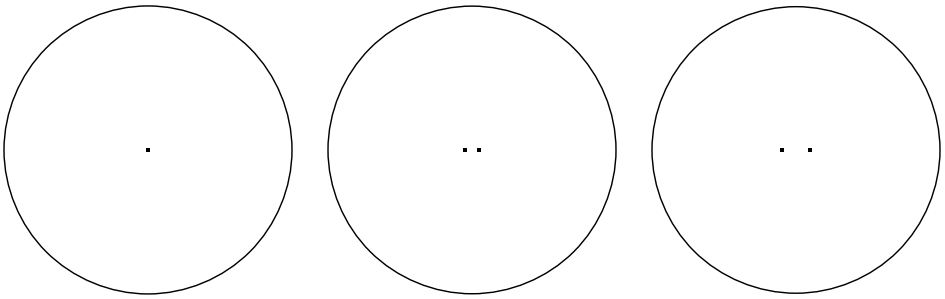
$$\begin{aligned} a &= \text{ semi-major axis} \\ b &= \text{ semi-minor axis} \\ p &= \text{ semi-latus rectum} \\ e &= \text{ eccentricity} \end{aligned}$$

Notice that while a and b don't always make sense, p and e have geometrical significance for all conic sections. Given p and e, the distance from the focus (the 'perihelion' distance) is equal to $p/(1+e)$, and the distance from the focus to the directrix is equal to p/e. For ellipses, the eccentricity is equal to $\sqrt{1-(b/a)^2}$, and the distance from the centre to the focus is $f = \sqrt{a^2 - b^2}$.

7. A remark about ellipses

As we shall see later in much more detail, the planets' orbits are very close to ellipses. Kepler's discovery of this fact was one of the great advances in astronomy. Since the conic sections were certainly well understod by the ancient Greeks, one might why it took so long for the elliptical shapes of planetary orbits to be discovered? The main point here is that although the eccentricity of a planetary orbit affects the *dynamics* of planetary motion, it has a much smaller effect on the *geometry* of that orbit. We can explain at least the second point right here, but postpone the first one until we look at Kepler's laws of planetary motion.

Here is a sequence of ellipses with eccentricities $e = 0.0, 0.05, 0.1$, with the foci plotted as well:



The separation of the foci is noticeable at this scale, but if you think the orbits themselves look pretty much like circles you're right. We have

$$e = \sqrt{1-(b/a)^2}, \qquad (b/a) = \sqrt{1-e^2}\ .$$

It is the difference between the ratio b/a and 1 which measures the difference between an orbit and a circle. This formula says that is a second order function of eccentricity, while the relative separation ae of the foci is a linear function of it.