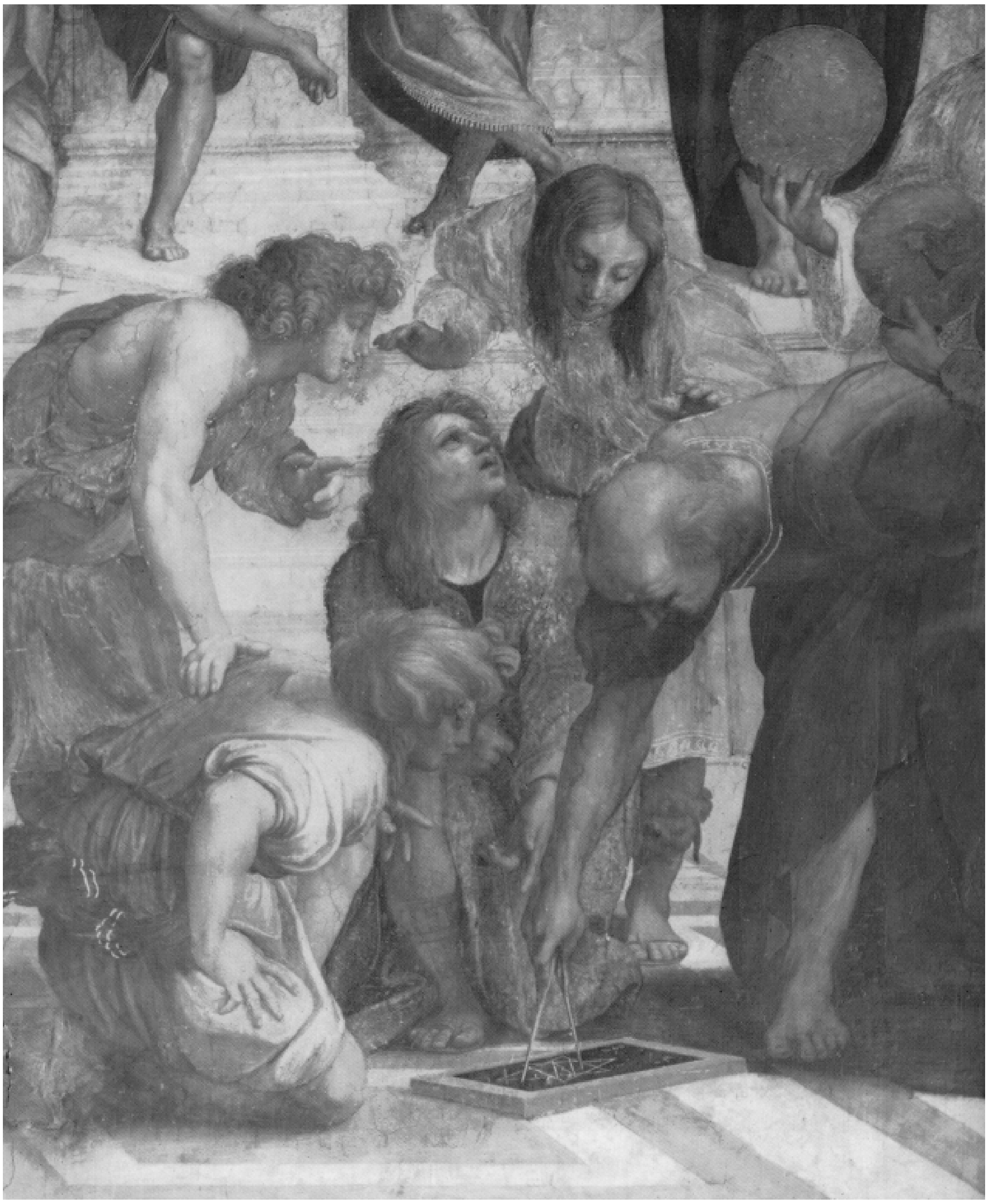
The cycloid

Scott Morrison

“The time has come”, the old man said,
“to talk of many things:
Of tangents, cusps and evolutes,
of curves and rolling rings,
and why the cycloid’s tautochrone,
and pendulums on strings.”



Everyone is well aware of the fact that pendulums are used to keep time in old clocks, and most would be aware that this is because even as the pendulum loses energy, and winds down, it still keeps time fairly well. It should be clear from the outset that a pendulum is basically an object moving back and forth tracing out a circle; hence, we can ignore the string or shaft, or whatever, that supports the bob, and only consider the circular motion of the bob, driven by gravity. It's important to notice now that the angle the tangent to the circle makes with the horizontal is the same as the angle the line from the bob to the centre makes with the vertical. The force on the bob at any moment is proportional to the sine of the angle at which the bob is currently moving. The net force is also directed perpendicular to the string, that is, in the instantaneous direction of motion. Because this force only changes the angle of the bob, and not the radius of the movement (a pendulum bob is always the same distance from its fixed point), we can write:

$$\ddot{\theta} \propto \sin\theta$$

Now, if θ is always small, which means the pendulum isn't moving much, then $\sin\theta \approx \theta$. This is very useful, as it lets us claim:

$$\ddot{\theta} \propto \theta$$

which tells us we have simple harmonic motion going on. Don't worry too much if that equation doesn't tell *you* that we have simple harmonic motion; if you worry over every equation you'll never get to the end, so just read on, peaceful and unperturbed! Now the important thing about simple harmonic motion is that given a certain proportionality constant in the above equation, the period of the motion is fixed, and completely independent of the amplitude of motion. In the context of the pendulum, this means that as it winds down (the amplitude decreases), it still keeps time (the period is unchanged). We now seem to have satisfactorily established that a pendulum will, indeed, tell you the time. Indubitably much to the relief of all the clock makers! Unfortunately, we've had to make a rather restrictive assumption in deriving our last equation; namely that the pendulum is only moving through small angles. Thus, even though we needn't keep winding up the pendulum, we can only let it swing a little, so it will need to be wound up anyway- a pendulum that's not perceptibly moving, even if it is in fact keeping perfect time, isn't much use. So, in the end, back in the real world, we have to keep driving the pendulums with a little motor or somesuch. Now, this obviously isn't a very agreeable state of affairs, if only for aesthetic reasons. I, for one, would be more impressed by a more vigorous and energetic breed of grandfather clock.

Building a better pendulum

We now have our problem; if a normal 'circular' pendulum doesn't achieve what we want, is there some other shape that does, and is it even vaguely practical? What we need of course, is some shape, such that when a particle rolls or swings along it, that object always takes the same amount of time to get from being motionless at the top of one swing to the bottom of the curve. This constant time must remain constant regardless of how high up the curve the particle starts. If we can achieve this, even as energy is drained away by friction, and the movement starts from a progressively lower position each time, the period will stay the same. And this time, we want it to be really, exactly, the same, not just when it's performing some inconsequential and measly little swing.

What this means is that if we let go of the putative balls labelled A, B and C in figure 1 at the same time, they should all reach the bottom at the same time. We can straightaway guess a few things about the curve. Firstly, it should be symmetrical about a vertical line through the middle, so the balls can roll down one side, up the other and back again. Secondly, it should be concave up. This is because for the highest ball to reach the bottom at the same time as the lower ones, it must accelerate faster right from the beginning, so it can begin to catch up.

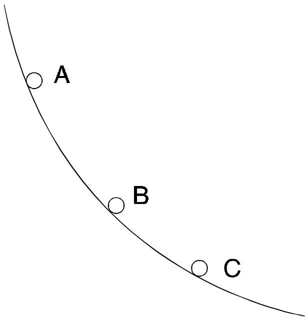


Figure 1

Things get serious

It's now time to begin our first serious incursion into mathematics land, in an attempt to come to grips with this shape. The property we want is that the time a ball takes to reach the bottom, when starting at rest, is independent of where it starts. This property is known as the 'tautochrone' property; for those who know a little Greek (ancient, that is) this should appear immediately sensible, but for others, it comes from words meaning 'the same' and 'time'. Hence, it's the curve that takes 'the same time' regardless of where the motion starts.

We'd make a good start by defining T to be the time a particle takes to reach the bottom under the influence of gravity, and writing down integral for it. (Calculus, and the calculus of integrals especially, will unfortunately have to play a considerable role in the pages to come.) We wouldn't mind an integral with respect to distance, since that's the only other dimension available, so we can now generate the inside bit of the integral by looking at units. We want time, and start with length, so we need to start with a measure of time per length, that is, the reciprocal of velocity:

$$T = \int_{start}^{finish} \frac{1}{v} ds$$

I've also made the integral definite by putting in some rather indefinite limits (it's the principle that counts). Both the velocity v and the distance s are measured in the direction of the curve. In addition to this 'curve-length' parameter, we'll need x and y coordinates, and introduce them as appropriate. We know something about the velocity, because by considering energy we can calculate the velocity from the distance the particle has fallen:

$$\tfrac{1}{2} mv^2 = mg(y_0 - y)$$

where y_0 is the initial height. For the sake of convenience we'll say the bottom of the curve is at $y=0$. We can now put some limits on the integral, but I'll always carefully write them in the form $y=a$, because the integral may not be with respect to y, or indeed obviously with respect to anything, and these limits may need to be converted into the appropriate variable. We can rest assured that we can do this at least for integrals with respect to *s* or *x*; for each *y* value there's really only one (actually two, but they're the same by symmetry) *x* or *s* value. Rearranging this last equation for *v*, we then rewrite the integral.

$$v = \sqrt{2g(y_0 - y)}$$

$$T = \int_{y=y_0}^{y=0} \frac{1}{\sqrt{2g(y_0 - y)}} ds$$

It's probably worth replacing *ds* here with something in terms of *x* and *y*.

$$ds^2 = dx^2 + dy^2$$

$$ds = \sqrt{dx^2 + dy^2}$$

(This should be clear from its obvious origin in Pythagoras.)

So we now have:

$$T = \int_{y=y_0}^{y=0} \sqrt{\frac{dx^2 + dy^2}{2g(y_0 - y)}}$$

We can turn this into an integral with respect to either *x* or *y* (more likely), but in either case I'm not at all sure where to proceed. The problem is that all we have is that integral, and the knowledge that it has to be independent of y_0. Unfortunately we can't simply differentiate with respect to y_0, because we have to be able to do the integral beforehand. We can do something productive by pretending to have done the integral, and fiddling with that. To do this, we introduce $F(y)$, which we say is the indefinite version of the above definite integral. *T* then becomes:

$$T = F(0) - F(y_0) = c$$

If *T* is independent of y_0, then $F(y)$ must be constant for all *y*, except at zero, where it must take on a different value in order to give a non-zero time. For $y_0 = 0$, of course, *T* is actually zero. Finding such a function *F* which hasn't been defined piecemeal is daunting enough, let alone relating it to our integral. I spent sufficient fruitless effort with these equations to know not to bore you with the details. Therefore, at this point we'll leave off this line of argument, and maybe return if we find an answer, and determine if it's consistent with what we've found here.

Simple(?) harmonic motion

Calculus starting out from integrals is always a bit difficult; let's have a go from the other end. We've already seen with the pendulums that we're looking for simple harmonic motion. Pulling up the first equation available:

$$\ddot{s} = -ks$$

I've written this in terms of *s* because the motion is always directed along *s*. Unfortunately this equation doesn't really help us, as although we could work out the forces, and hence $\ddot{s}$, it would probably, and in fact does, involve considerable trigonometry. Much more convenient, is:

$$P = \tfrac{1}{2}ks^2$$

where *P* is the potential energy. This equation is the integral (over *s*) of the previous one (we then have to change the sign to convert from kinetic to potential energy). Hopefully this equation will remind you of the potential energy of springs, derived from Hooke's law. Using energy in physics problems hides a wide variety of sins; here it allows us entirely to do away with forces, accelerations, and evil second derivatives. What really makes this particular equation attractive is the fact that we can do away with *P* with such ease; it is of course, simply the height above the bottom of the curve (times a few bits and pieces). Expanding *s* as a path length integral, we now get:

$$mgy = \tfrac{1}{2}k\left(\int_{y=0}^{y=y} \sqrt{dy^2 + dx^2}\right)^2$$

What we'd like to be able to do from here is generate some sort of differential equation. So that's what we'll do:

$$\frac{2mgy}{k} = \left(\int_{y=0}^{y=y} \sqrt{dy^2 + dx^2}\right)^2$$

taking the square root of both sides, and rearranging the integral,

$$\sqrt{\frac{2\mathrm{mgy}}{\mathrm{k}}} = \int_{y=0}^{y=y} \sqrt{1+\left(\frac{dx}{dy}\right)^2}\, dy$$

we can now see the benefit of factoring out the *dy* from the integral. Differentiating with respect to *y*,

$$\sqrt{\frac{2\mathrm{mg}}{\mathrm{k}}}\,\frac{1}{2\sqrt{y}} = \sqrt{1+\left(\frac{dx}{dy}\right)^2}$$

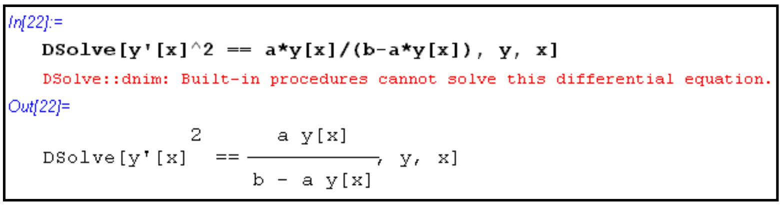
we almost have a differential equation, so squaring,

$$\frac{mg}{2ky} = 1+\left(\frac{dx}{dy}\right)^2$$

or in another form

$$\frac{dy}{dx} = \sqrt{\frac{2ky}{mg - 2ky}}$$

I've taken the positive square root in the above equation for simplicity. It is obvious that the gradient of the curve will be positive only on the right hand half of the shape. We can take the negative square root if we wish for the left side. As it turns out this won't end up mattering at all; our subsequent fiddling will give us results about both halves of the curve. At this point we can claim to have solved this problem (to a certain extent); getting a differential equation is probably a good indication the final answer is nearby. At this point we'd like to find an explicit equation relating x and y, and to do this, we'll apply one of the standard techniques of modern mathematics. Namely, asking *Mathematica*[1], a computer program, to have a look at the problem, and seeing if it can do anything with this differential equation.



```
In[22]:=
    DSolve[y'[x]^2 == a*y[x]/(b-a*y[x]), y, x]
    DSolve::dnim: Built-in procedures cannot solve this differential equation.
Out[22]=
                     2        a y[x]
    DSolve[y'[x]   == -----------, y, x]
                        b - a y[x]
```

Figure 2

I've had to rearrange the differential equation slightly, but *Mathematica* is still unable to solve our problem; instead, it flatly refuses, and spits back the question at us. This perhaps reinvigorates us, and injects new significance and meaning into our lives; after all, a computer has just beaten the world champion chess player, and the duty now devolves upon all of us to continue the struggle. This is our chance to free ourselves from the clutches of computerised mathematics. On the other hand, it may have been better if *Mathematica* had simply told us the answer, and saved us a lot of work!

Notice first that our differential equation is only meaningful for certain values of y. Because k is positive (it's the coefficient in the simple harmonic motion equation, so it has to be positive), we know that $mg > 2ky$ in order for us to be able to take the square root. Similarly, y can't be negative, which is not altogether surprising as we defined $y=0$ as the bottom of the curve. Since we have the gradient $\left(\frac{dy}{dx}\right)$, an obvious parameter to introduce would be the angle the tangent with this gradient makes with the horizontal. We'll call this angle ψ. Figure 3 shows a sketch of the curve with these boundaries marked in, and a tangent making an angle ψ with the horizontal.

Now,

$$\begin{aligned} \tan\psi &= \frac{dy}{dx} \\ &= \sqrt{\frac{2ky}{mg - 2ky}} \end{aligned}$$

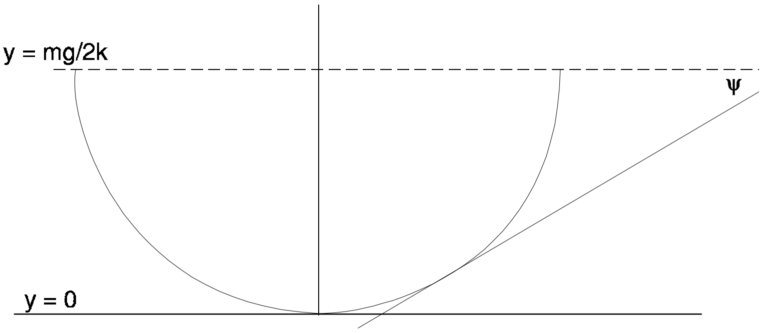


Figure 3

Using Pythagoras, as shown in figure 4, we can now work out $\sin\psi$ and $\cos\psi$

We now have $\sin\psi = \sqrt{\dfrac{2ky}{mg}}$

and

$$\cos\psi = \sqrt{\frac{mg-2ky}{mg}}$$

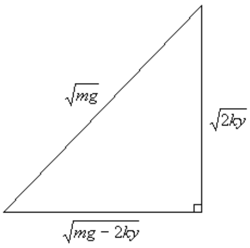


Figure 4

or, $\cos^2\psi = \dfrac{mg-2ky}{mg}$

Now, using a couple of trigonometric identities[1],

$$y = \frac{mg}{4k}(1-\cos 2\psi)$$

Differentiating this with respect to ψ ,

$$\begin{aligned}
\frac{dy}{d\psi} &= \frac{mg}{2k}\sin 2\psi \\
&= \frac{mg}{k}\sin\psi\cos\psi \\
\frac{dx}{d\psi} &= \frac{dx}{dy}\frac{dy}{d\psi} \\
&= \cot\psi\,\frac{mg}{k}\sin\psi\cos\psi \\
&= \frac{mg}{k}\cos^2\psi \\
&= \frac{mg}{2k}(\cos 2\psi + 1)
\end{aligned}$$

We can now integrate to find x

$$x = \frac{mg}{4k}(\sin 2\psi + 2\psi + c)$$

where c is simply a constant of integration. We're almost there now, but I'm going to have to ask you to trust me while I make a substitution. Although this substitution clears up the equations slightly, it isn't otherwise clear what justification there is for applying it. Once we've had a look at some of the properties of the curve I'll come back to it. At the same time, we'll introduce a new constant, r, to stand for $\frac{mg}{4k}$, so as to clean up the coefficient occurring in both equations. Why I choose the letter r will be apparent later.

Let $\theta = \pi - 2\psi$, now we get

$$y = r(1 + \cos\theta)$$
$$x = r(\sin\theta - \theta + \pi + c)$$

and by choosing $c = -\pi$, we get

$$x = r(\sin\theta - \theta)$$

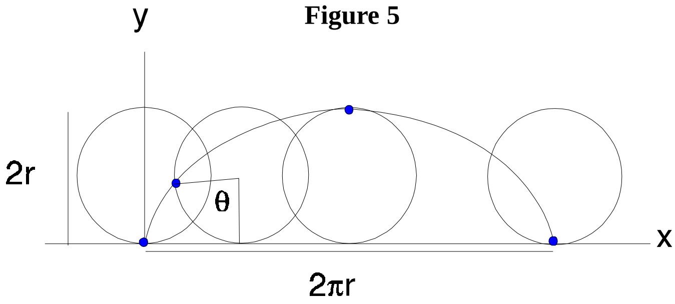
Deciphering the equations

We've done the maths, but what on earth do these things *look* like? If we ignore the θ term in the expression for x, we can see something of what's going on, because the equations are now those of a circle, radius r (hence the name) shifted up by r:

$$y = r(1 + \cos\theta)$$
$$x = r\sin\theta$$

This is the locus of a point moving around in a circle. The term in θ simply slides this point sideways as it moves along. Interestingly, by the time the point has made a complete revolution, ie θ has changed from 0 to 2π , the whole system has been shifted sideways by the diameter of the circle, because the θ has changed from 0 to 2π times r, which is the diameter. This offers us a physical explanation of the motion- it's not simply a point moving around a circle, but a point on a circle, which is rolling along. It is important that the wheel is sliding at the same rate that it is turning, because this allows us to combine these movements into the idea of 'rolling'. It could thus be imagined as the path traced out by a pebble stuck in an automobile wheel.

Figure 5



The equations we've derived don't create a curve with this same orientation, so lets look at it again and consider directions and speeds of rotation and sliding in a bit more detail. At first, just considering the circular, that is, rotational, part of the motion, the particle starts (by 'starts' I mean $\theta = 0$, with θ increasing) at the highest point of it's motion ($\cos 0 = 1$), but moving downwards. At the same time, as θ increases, so does $\sin\theta$, so the particle begins moving right (the positive *x* direction). When we reintroduce the lateral movement, we notice that the sign of the θ term is negative, telling us that the 'sliding' is to the *left*. How can this be? To reconcile this with our notion of rolling, we have to think of the wheel as rolling along the roof, towards the left! At the beginning of the motion, the $\sin\theta$ will almost exactly cancel out the $-\theta$, so we can infer that the beginning of motion is one of the steep parts of the curve. To have a look at this orientation, simply turn this page upside down (so you're still looking at this side of the page). You'll have to ignore the axes, but you'll even be able to see the way the curve is traced out as θ increases.

Some history

The cycloid has quite a long history in modern mathematics; relatively modern at least - not classical. It was first cursorily investigated by Cusa, in attempting to find the area of the circle by integration, and subsequently was studied by the whole gamut of mathematicians who started us all off on calculus. Galileo gave it its modern name, and investigated the area under each arch, but as yet its tautochrone and brachistochrone properties were unknown- Galileo actually thought the brachistochrone was in fact a circle- and very little had been discovered about its geometry. This geometry was uncovered incrementally by Fermat, Roberval, Wren, and Huygens. In 1673 Huygens discovered the tautochrone property, which has just been discussed. He actually appears to have done this without calculus, and is said to have taken half a year to construct all the requisite geometric diagrams! In 1696 the cycloid really began to take its place in mathematical history; Johann Bernoulli had discovered the brachistochrone property (more about that later) of the cycloid, and offered the problem as a challenge. Jacob (his brother), Leibniz, Newton, and L'Hôpital answered this challenge. Amongst these five mathematicians, it would be fairly safe to claim that we could find the founders of modern calculus; the cycloid has quite a distinguished history! In fact, by this stage, the cycloid had been the subject of so many challenges, competitions, arguments and rivalries, that it has been called by some the 'Helen of Geometers'[3].

A lemma

To explain why I made the substitution I did during the previous derivation, we need to consider normals to the cycloid. A normal, for those who don't know, is the line perpendicular to the curve at a given point. If we are able to construct tangents, this means it is very easy to construct the corresponding normal, because the normal is at right angles to the tangent, and we know how to construct perpendiculars easily. The interesting property of the normals is that they 'pass through the base of the generating circle'. What this means, is that if we remember the circle that we rolled along to create the cycloid, the normal to the curve will pass through the point of contact of this generating circle on the surface on which it's rolling. A diagram will be useful here:

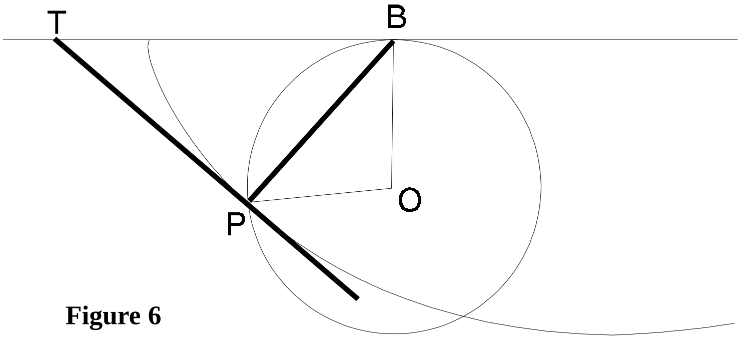


Figure 6

The best of several ways of proving that *TP* is perpendicular to *PB* is very simple; when we consider the point on the circle as the cycloid is being traced out, its instantaneous centre of motion is simply the point of contact between the circle and the ground. It follows immediately from this that the tangent is perpendicular to the line joining the moving point on the circle and the point of contact between the circle and the ground, because in rotational movement the direction of movement is always perpendicular to the direction towards the centre of motion. If the tangent is perpendicular to this direction, then the normal is itself this line. This may not appear a particularly rigorous proof, but it is entirely true; a simple piece of coordinate geometry will confirm this result. At this point, we should draw a diagram, illustrating both our old definition of ψ, as the angle the tangent makes with the horizontal, and our newfound knowledge about the normals. In this diagram, the point *P* is the variable point on the cycloid, the point *O* is the centre of the rolling circle. *B* is the point of contact between the rolling circle and the ground, and *T* is the intersection of the tangent at *P* with the ground.

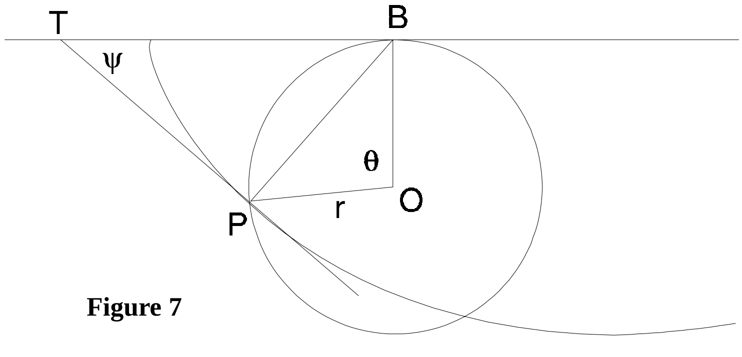


Figure 7

Why that substitution worked

Our lemma concerning the normal lets us say, after some angle chasing[4], that the angle at the centre of the circle, which we'll call θ, is related to the angle the tangent makes with the horizontal, ψ, according to the following equation:

$$\theta = \pi - 2\psi$$

The substitution simply changed the equations so that they would be in terms of the angle through which the circle had rolled, rather than the angle which the tangent makes to the horizontal. If we had to pick a parameter, this angle would probably be the one we're after, and it's worth thinking for a moment about how this substitution works. That should be enough geometry for now.

Just making sure

It's worthwhile pausing at this point to check that our result is consistent with our earlier working, namely that

$$T = \int_{y=y_0}^{y=0} \sqrt{\frac{dx^2 + dy^2}{2g(y_0 - y)}}$$

must be independent of y_0. Substituting our parametric equations in, and introducing θ_0 as the value of θ for $y = y_0$ we get

$$\begin{aligned} T &= \int_{\theta=\theta_0}^{\theta=\pi} \sqrt{\frac{r^2(\cos\theta - 1)^2 + r^2(-\sin\theta)^2}{2gr(1+\cos\theta_0 - 1 - \cos\theta)}}\,d\theta \\ &= \sqrt{\tfrac{r}{2g}} \int_{\theta=\theta_0}^{\theta=\pi} \sqrt{\frac{\cos^2\theta - 2\cos\theta + 1 + \sin^2\theta}{\cos\theta_0 - \cos\theta}}\,d\theta \\ &= \sqrt{\tfrac{r}{g}} \int_{\theta=\theta_0}^{\theta=\pi} \sqrt{\frac{1 - \cos\theta}{\cos\theta_0 - \cos\theta}}\,d\theta \end{aligned}$$

We can actually perform this integration, with the help of *Mathematica* (or a bit of a mess with pen and paper- which I'll most generously leave to the reader as an exercise), and we get

$$T = \sqrt{\tfrac{r}{g}} \left[-2\tan^{-1}\left(\frac{\sqrt{2}\cos\left(\tfrac{\theta}{2}\right)}{\sqrt{\cos\theta_0 - \cos\theta}} \right) \right]_{\theta=\theta_0}^{\theta=\pi}$$

The interesting thing here is that when we substitute $\theta = \theta_0$, the denominator of the argument becomes zero, so the argument itself becomes infinite. What's so felicitous about the inverse tan function $\left(\tan^{-1}\right)$ is that it approaches $\frac{\pi}{2}$ asymptotically as the argument goes to infinity.

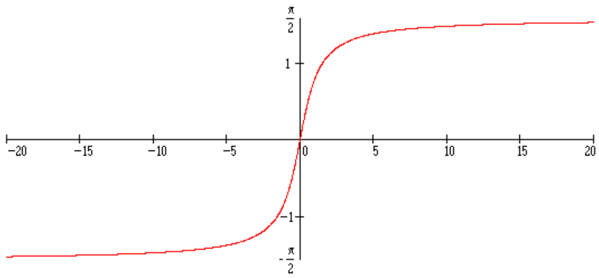


Figure 8: tan^{-1}

This result of this is that when we evaluate the expression at $\theta = \pi$, we get 0, because $\cos\left(\frac{\pi}{2}\right) = 0$, and when we evaluate it at $\theta = \theta_0$, we get something like $2\tan^{-1}\left(\frac{\sqrt{2}\cos\left(\frac{\theta_0}{2}\right)}{0}\right)$, which, if we simply grimace and continue when confronted with infinity, rather than give up, evaluates to π . The important thing here is that all the θ_0 's cancel, and the integral evaluates to $\pi\sqrt{\frac{r}{g}}$. We rejoice at this point that the time is completely independent of θ_0 , indicating that we have in fact found the tautochrone. This approach is not as aesthetically pleasing as the derivation by way of simple harmonic motion, because it relies on knowledge of the result, and only demonstrates that the cycloid is the tautochrone, but lays no claim to actually deriving the shape of the tautochrone from scratch. It is most reassuring however, that our two lines of attack have been shown to be consistent; this should give us considerable confidence in our answer.

The period of tautochrone oscillations

The other interesting information resulting from this approach concerns the period; by evaluating the integral, we can actually find the full period, which will be very useful if we later decide to actually build a tautochrone clock. The full period is clearly four times the time it takes for the particle to travel from rest to the bottom of the curve (it has to take the same amount of time rolling down one side as up the other, because it's just going in reverse, as it were). We've calculated this time as $\pi\sqrt{\frac{r}{g}}$, so the full period is $4\pi\sqrt{\frac{r}{g}}$. If we recall that the period of a normal pendulum is $2\pi\sqrt{\frac{l}{g}}$, with *l* as the length of the string, we might postulate that the quantity *4r* is analogous to *l* in some way, because we can write the period of the tautochrone as $2\pi\sqrt{\frac{4r}{g}}$.

A pause for breath

We have derived the shape of the tautochrone, by considering the nature of simple harmonic motion, and taken a cursory glance at the geometry of this curve, the cycloid. Since discovering this derivation, starting from the simple harmonic motion constraint, I've had a look around at the 'literature' as it were. What has been interesting is the two different classes of 'solutions' provided in various books. Quite a number, especially more general books on calculus, only show that the cycloid is the tautochrone. (This corresponds to the second part of our proof, where we demonstrated our answer was consistent with our earlier, aborted, approach.) On the other hand, some do provide an actual derivation (that is, showing that the tautochrone is the cycloid, rather than the other way round), and when they do, it is generally much simpler and more elegant than the solution I've found - which is quite annoying as far as I'm concerned! Personally I think it's more fulfilling to go from tautochrone to cycloid, than from cycloid to tautochrone, but that could well be a biased opinion. An advantage of the 'simple harmonic motion' method , which I used, *is that it also works perfectly well in potential fields other than normal gravity*[11]*. In such cases, the ‘guess the solution and prove it’s right’ approach doesn’t work, as the guessing becomes increasingly difficult. After we’ve looked at another amazing property of the cycloid, I ’ll say something about constructing some form of tautochrone clock.*

The next problem to which we’ll turn our attention is the curve known as the brachistochrone. This time, the word for ‘the same’, has been replaced by the word for ‘the shortest’ (the superlative of βραχύς, - εῖα, - ύ). We’re no longer interested in keeping accurate time, but simply in getting from *A* to *B* as quickly as possible, under gravity. *A* and *B* are simply two points, sharing neither a vertical nor a horizontal syzygy. The first suggestion that springs to mind is simply a straight line. After all, a straight line is the shortest distance between two points and all that, and it would certainly make life easy. If it were so easy, of course, I wouldn’t have bothered introducing the question to you. The problem with a straight line is that we could probably improve upon it by making it a little steeper initially, and flattening out later in order to arrive at *B*. This would result in a higher acceleration from the start, which would result in a higher velocity over the flatter bit, quite possibly making an improvement. Just as in the tautochrone problem, we have to work from a physical description of a property of the curve to the mathematical form. This time, it won’t be as simple as manipulating the equations for simple harmonic motion. Instead, we will have to play with an integral, (which we managed to avoid in approaching the tautochrone problem). The brachistochrone problem is often discussed in relation to the calculus of variations, which is a very powerful technique for minimising integrals, something that is very often required in physics. It can be used for problems such as proving that of all shapes of a given perimeter, the circle encloses the greatest area. Unfortunately, the calculus of variations is rather high-tech, requiring partial derivatives and other such nasties. The proof I’ve found avoids all that, and manages to do everything with nice and basic calculus. It does so by arriving at the results of the calculus of variations to the extent that they are needed in the midst of other things, and only really derives them within the context of the brachistochrone problem. Some might argue that this is a bit silly, and that instead of only considering a special case one should work out the abstract general case, and then almost as a denouement prove various examples. In defence, I offer two counterarguments; firstly, I’m not smart enough to derive the abstract case straight out, but can just manage, with just a few hints[5], how a particular problem works. Secondly, having to resort to partial derivatives is almost as bad, in my mind, as having to resort to calculus in the first instance. For this reason it’s worthwhile to consider the problem almost ‘from first principles’.

Why we need some new maths

The first thing to do is write down an expression for the time the particle takes to fall along the curve. Using exactly the same derivation as for the tautochrone problem, we get:

$$T = \int_A^B \sqrt{\frac{dx^2 + dy^2}{2g(y_A - y)}}$$

where the limits *A* and *B* represent the limits expressed in whichever variable is appropriate for the integral, and y_A is the *y* coordinate of *A*, the starting point. For simplicity we'll take *A* as the point $(0,0)$, so from here on we can leave out y_A. As a secondary result of this, *y* should always be negative (the particle has to fall *downwards*; it doesn't have wings). Our problem now is to find the curve, described by $\frac{dy}{dx}$ and *y*, which minimises *T*. Normally, to minimise a function, we differentiate the function with respect to its independent variable. When the derivative is zero, we know the function is either at a maximum or minimum, or a stationary turning point. So, why can't we simply differentiate this expression, and hence find the function? The problem here is that *T* is a function of the shape of the curve (the definite integral maps functions to numbers), so the independent variable is this shape. Therefore, to minimise *T* we need to differentiate using d(shape), or d(path), or something similarly meaningless. The calculus of variations develops a technique by which we can do something like that, but as I said before, it requires partial derivatives.

Minimisation without differentiation

To escape this problem, we have to think a little about what 'minimising' means, and about other ways of approaching it. I don't want you to think I'm just about to do a differentiation by first principles here; we'll be concerned with something perhaps even simpler, but also perhaps not at all obvious, and quite surprising! Our problem is that differentiation has become impossible; we are going to have to do calculus without the accustomed apparatus. If we think about a fairly simple curve, such as a parabola, we notice something interesting about its minimum.

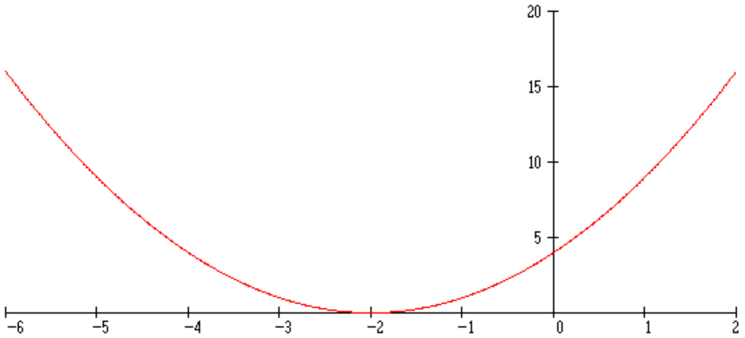


Figure 9: A parabola

As well as being the lowest point, the minimum is also the place at which the curve is flat! (This is of course what we look for when differentiating). If a curve is ‘flat’ in the way the minimum of a parabola is, if we move just slightly to either side of the minimum, the function has almost the same value as at the minimum. There is of course a very small change, but this is due to the second derivative, or concavity, of the curve, and there is no first order change due to the first derivative. At all other places on the curve, if we move slightly to either side, there are correspondingly significant changes in the value of the function. If we only move a small amount, the function only changes a small amount, but this is very different from the case at the minimum, where the function changes a negligible amount. Such a ‘negligible amount’ is referred to as a ‘second order change’, because it generally results from a small quantity raised to the second or higher power. We can use this idea to find the minimum by demanding that the point we think is the minimum show this property, namely that the value of the function changes only a very, very small amount.

If, for example, we have the curve $y(x)=(x+2)^2=x^2+4x+4$, and want to find the minimum, we say that the minimum occurs at α, and then demand that the difference between the *y* values at α and $\alpha+\Delta\alpha$ is very small, where $\Delta\alpha$ is small. The difference is given by:

$$\begin{aligned} y(\alpha)-y(\alpha+\Delta\alpha) &= \alpha^2+4\alpha+4-\left(\alpha^2+2\alpha\Delta\alpha+(\Delta\alpha)^2+4\right. \\ &= -2\alpha\Delta\alpha+(\Delta\alpha)^2+4\Delta\alpha \\ &= \Delta\alpha(4-2\alpha)+(\Delta\alpha)^2 \end{aligned}$$

Now this difference has to be negligible, which means that the squared term is allowed, because $\Delta\alpha$ is small, so $(\Delta\alpha)^2$ is really, really small, but the first term must disappear. The only way this can happen is if $\alpha=-2$, and this is indeed the minimum. The attraction of this approach is that it allows us to find maxima or minima without needing to know about differentiation, which is useful when we in fact don’t want to know about differentiation, because it has become too difficult.

It’s interesting that this method finds other interesting points as well as minima; namely maxima and stationary turning points. For our problem, maxima don’t matter- if you claim you’ve found a curve that gives a maximum for the time, I can always make a longer time by stretching out your curve. This new way of thinking about ‘minima’ allows us to revise commonplace bits of physics like Fermat’s principle in very interesting ways.[6]

Our new technique and the brachistochrone

To apply these ideas to the brachistochrone problem, we proceed as follows. First we assume that there is some curve which does minimise *T*, which is a fairly reasonable assumption. If it minimises *T*, then the graph of *T* against ‘shape’ must be almost flat around that particular shape, and hence if we modify that shape slightly, then *T* shouldn’t change much. In order to achieve this ‘modification’, we introduce a new function, which always has only small values and say that if previously the curve was defined by $x = f(y)$, then it is now given by $\bar{x} = f(y) + \eta(y)$, where $\bar{x}$ represents the modified curve. Notice we're doing something slightly strange here, in expressing *x* in terms of *y* rather than vice versa, as would be more usual. As we proceed I'll point out why this makes the mathematics much easier. Another important thing to keep in mind is the fact that $\eta(y)$ is always pretty small, so that our idea of allowing second order changes but not first order ones is still valid. For conciseness, I'll write these equations more succinctly as $\bar{x} = x + \eta$. To apply our newfound techniques, we simply find the difference in time indicated by the integral using the unshifted curve and the integral using the shifted curve, and then demand there is no first-order change in time.

The old and new integrals are:

$$T = \int_A^B \sqrt{\frac{dx^2 + dy^2}{-2gy}}$$

and

$$\bar{T} = \int_A^B \sqrt{\frac{d\bar{x}^2 + dy^2}{-2gy}}$$

Expanding out the second of these

$$\begin{aligned}\bar{T} &= \int_A^B \sqrt{\frac{d(x+\eta)^2 + dy^2}{-2gy}} \\ &= \int_A^B \sqrt{\frac{dx^2 + 2dxd\eta + d\eta^2 + dy^2}{-2gy}} \\ &= \int_A^B \sqrt{\frac{dx^2 + dy^2}{-2gy} + \frac{2dxd\eta + d\eta^2}{-2gy}}\end{aligned}$$

We'd now like to find the difference between $\bar{T}$ and *T*, which would be quite easy if it weren't for that square root! At least under the square root we've managed to construct a term that is identical to the expression under the square root for *T*. What we'll have to do is expand the square root out in some fashion so as to create several terms, some of which will, we hope, cancel out with the integrand for *T*. One or two methods to expand out such an expression may come to mind. The binomial theorem would probably do the job, but is somewhat inconvenient[7], and Taylor series offer a much clearer way forward.

Taylor series

A Taylor series provides a method for approximating the value of a function, given the value of the function and a few derivatives at another point. If we have some function $F(x)$ for which we know the value at $x=a$, we can calculate the value at $x=a+h$ according to the formula:

$$F(a+h)=F(a)+hF'(a)+\tfrac{h^2}{2!}F''(a)+\tfrac{h^3}{3!}F'''(a)+\ldots$$

$F'(a)$ here means 'the first derivative of $F(x)$ with respect to *x*, evaluated at a'. We then use as many terms as we need to achieve a desired accuracy. This series will only converge quickly, and hence be useful, if either h is quite small, or the derivates become very small. Taylor series are very interesting creatures, with many applications[8]. As an example of how they work, we can consider the function $F(x)=\sqrt{x}$. The derivatives look like $F'(x)=\dfrac{1}{2\sqrt{x}}$ and $F''\left(\dfrac{-1}{4x\sqrt{x}}\right)$, and so on and so forth, and as a result we can write:

$$\sqrt{a+h}=\sqrt{a}+\frac{h}{2\sqrt{x}}+\frac{-h^2}{8x\sqrt{x}}+\cdots$$

In our case, we want to expand the monstrous square root in the integrand of $\overline{T}$, so the function we will be using is $F(x)=\sqrt{x}$.

If we choose $a=\dfrac{dx^2+dy^2}{-2gy}$ and $h=\dfrac{2dxd\eta+d\eta^2}{-2gy}$

we can begin to expand it out. Notice that, as is required in order to use a Taylor series, h is small, being a multiple of $d\eta$, which is much smaller than dy or dx (this is simply because we have to say that η is small).

Bear with me, it will be over soon.

Expanding out $\sqrt{\dfrac{dx^2+dy^2}{-2gy}+\dfrac{2dxd\eta+d\eta^2}{-2gy}}$ using our Taylor series gives us

$$\left(\frac{dx^2+dy^2}{-2gy}\right)^{\frac{1}{2}}+\left(\frac{2dxd\eta+d\eta^2}{-2gy}\right)\left(\frac{1}{2}\right)\left(\frac{dx^2+dy^2}{-2gy}\right)^{-\frac{1}{2}}+\frac{1}{2}\left(\frac{2dx}{}\right.$$

One can say safely enough that this is a mess. Fortunately the way forward is clear; all we need do is discard anything which has as a factor any power of $d\eta$ greater than 1. This is simply because terms containing this correspond to second order changes, or even higher order changes, which we don't mind about at all. That simplifies everything enormously, and we get

$$\sqrt{\frac{dx^2+dy^2}{-2gy}+\frac{2dxd\eta+d\eta^2}{-2gy}} = \sqrt{\frac{dx^2+dy^2}{-2gy}}+\left(\frac{2dxd\eta}{-2gy}\right)\frac{\quad}{2\sqrt{\quad}}$$

Notice that as well as discarding the third and following terms from the Taylor series expansion, I've also scrubbed out part of the second term, the $d\eta^2$.

$$\sqrt{\frac{dx^2+dy^2}{-2gy}+\frac{2dxd\eta}{-2gy}} = \sqrt{\frac{dx^2+dy^2}{-2gy}}+\frac{dxd\eta}{\sqrt{-2gy\left(dx^2+dy^2\right)}}$$

What we have to do now is calculate the difference between the two times, for the shifted and unshifted curves, and then apply the requirement that this difference is zero to a first order approximation. Because we've already removed all the second order terms, all that we have to do is require that the difference is actually zero. If we say that this difference between times is δT,

$$\begin{aligned}\delta T &= 0 = \overline{T} - T\\ &= \int_A^B\left(\sqrt{\frac{dx^2+dy^2}{-2gy}}+\frac{dxd\eta}{\sqrt{-2gy\left(dx^2+dy^2\right)}}\right)-\int_A^B\sqrt{\frac{dx^2+dy^2}{-2gy}}\\ &= \int_A^B\frac{dxd\eta}{\sqrt{-2gy\left(dx^2+dy^2\right)}} = 0\end{aligned}$$

This lovely cancelling of the entire unshifted integral can only come about because of a particular trick we employed, namely shifting the *x* coordinates of the curve rather than the *y* coordinates. If we had shifted *y*, the η, as well as turning up in the numerator of all these fractions, would also turn up in the denominator. This would make it much more difficult at least, and perhaps completely impossible, to cancel terms as we did. It's very hard to convey any sense of how you can manage to pull such 'tricks' out of thin air while doing maths like this. That's simply because you just don't pull these tricks out of nowhere, nor is there always a 'moment of inspiration' when you see the beautiful and powerful way forward. Rather, you plough on ahead, get it completely wrong the first time through (I did!), and end up utterly stuck, and then go back and try to do things in a slightly different fashion in order that the next time you'll clear the next obstacle. Of course, it's very tempting not to tell anyone about those mistakes, and only write up those parts that make you look like pure genius!

In order to simplify our working, we will introduce

$$\psi = \frac{dx}{\sqrt{-2gy\left(dx^2 + dy^2\right)}}$$

The requirement concerning the difference between times now becomes

$$\int_A^B \psi d\eta = 0$$

which is beginning to look very concise, and perhaps even elegant! Even if the progress we have made from our 'huge mess' only a moment ago is progress achieved through making simplifying substitutions, that progress makes our life that much easier, and allows us to get quickly to the heart of the problem, and find our answer. It may be possible to extract some information out of this equation, and the suggestion that comes to mind is that ψ is always zero. This seems reasonable enough, and seems to be required in order to ensure that the integral is actually always zero, regardless of how we choose to define the arbitrary η function. This isn't in fact entirely true, as we know that η is zero at both *A* and *B*, which means the integral of $d\eta$ from *A* to *B* must be zero. Integrating $d\eta$ from *A* to *B* is the sort of thing this last integral equation is doing, and this restriction on η might just be enough that the integral might be zero in a wider variety of cases. For example, if ψ were to be constant, we would get

$$c\int_A^B d\eta = 0$$

which does in fact equal zero, as I just pointed out. The important thing is that we don't want to confuse restrictions on η with restrictions on ψ. Because of this slight confusion, we won't draw any immediate conclusions from this integral equation, but do something fairly clever which gets us around the problem of whether the fact that η is zero at both *A* and *B* actually lets us have a wider variety of functions for ψ.

To do this, we'll now perform an integration by parts:

$$\left|\psi\eta\right|_A^B - \int_A^B \eta d\psi = 0$$

Now, an important point which we brought up at the beginning was that η must be zero at either end of the curve. This is because this method of variations doesn't mean anything if we move the endpoints; the position of the endpoints is part of the specification of the problem, and it's not for us to try and change them. This means that the first term in the previous equation is zero, because if η is zero evaluated at either *A* or *B*, so is $\psi\eta$. This solves a large part of the problem we discussed a moment ago; we have, in a way, 'removed' (by doing the integration by parts), that part of the integral which was concerned with the restriction on η. In other words, we have now 'used' this piece of information, and it's important to know that we have 'used' all our relevant information by the time that we get to a solution. We now get

$$\int_A^B \eta d\psi = 0$$

This time, there are no hidden pitfalls; although there are restrictions on η, this equation involves some integral of η, which has no such restrictions, and may therefore be considered entirely arbitrary. Now that we have an entirely arbitrary function, we can be certain that $d\psi$ must always be zero. This conclusion basically comes from the fact that we can choose any η function at all. If $d\psi$ were not always zero, we could choose an η which was positive when $d\psi$ was positive, and negative when $d\psi$ was negative, hence ensuring a positive integral, which is not allowed (because it has to be zero!). Alternatively, we could use some sort of 'spike' function as our η to achieve much the same end. Therefore

$$d\psi = 0$$
$$\psi = c$$

We are now in a position to create a differential equation for the brachistochrone, by recalling our definition of ψ.

$$\psi = c = \frac{dx}{\sqrt{-2gy\left(dx^2 + dy^2\right)}}$$
$$-2c^2 gy\left(dx^2 + dy^2\right) = dx^2$$
$$dy^2\left(-2c^2 gy\right) = dx^2\left(1 + 2c^2 gy\right)$$
$$\frac{dy}{dx} = \sqrt{\frac{1 + 2c^2 gy}{-2c^2 gy}}$$

This is pretty much the same differential equation as we arrived at when considering the tautochrone problem. This differential equation is not exactly the same as the previous one, but can be made so by shifting the *y*-axis. Parametric equations of the same form as those derived for the tautochrone can be found.

But which bit?

The remaining problem is that we don't know which section of the cycloid we're actually interested in for the brachistochrone. This will obviously be determined in some way by the spacing and relative positions of the starting and ending points of motion. The spacing actually only changes the radius of the cycloid, and determining the specific points for the beginning and ending involves solving a tedious equation, which can only be done numerically, rather than analytically. To show how this problem is solved in principle, we first have to have another look at the parametric equations for our cycloid.

Using the same derivation as for the tautochrone problem, we get

$$y = r(1 - \cos\theta)$$
$$x = r(\theta - \sin\theta)$$

These aren't exactly the same as before, but the cycloid is quite a versatile beast, and these equations produce a translation of the tautochrone curve (we've simply chosen a different origin). We began the problem by defining A as (0, 0), and this corresponds to $\theta = 0$, which is a cusp, because we can easily find

$$\frac{dy}{dx} = \frac{\sin\theta}{1 - \cos\theta}$$

which is undefined at $\theta = 0$. We therefore know that the section of the cycloid we're interested in starts at the cusp. If we then work out the gradient from *A* to *B*, we can then find the full arc. All we do is draw a unit cycloid (*r* = 1), and draw a line of the required gradient through one cusp. This will intersect with the loop of the cycloid, and we can then scale the unit cycloid up so that the distance from *A* to *B* is as specified by the problem. An interesting consequence of this is that for very low gradient between *A* and *B*, the brachistochrone actually dips down below the level of *B*, because the line across the unit cycloid will actually intersect it on the other side, after it has gone through its minimum. Incidentally, Huygens discovered (I'm not sure about proved) that of all cycloid arcs with equal separation between the end points and equal overall gradients, the one with the steepest beginning would take the shortest time for a particle to slide along. This tells us that the cycloid arc we are interested in does in fact pass through the cusp, as we had been led to expect.

Another pause for breath

What exactly have we achieved up to this point? We have completed two main proofs, deriving the curve which has the tautochrone and brachistochrone properties. In both cases this was the cycloid. In addition, we've investigated a property of the cycloid, namely our lemma about the normals to the cycloid, and cleared up a few odds and ends, such as the period of tautochrone oscillations, and which section of a cycloid is appropriate for the brachistochrone. Perhaps more important, however, are the techniques we constructed in order to solve all these problems, especially our method of doing variational problems, without resorting to the calculus of variations. The calculus of variations is fundamentally important because of the wide variety of problems which require it, especially physical problems. In fact, Lagrangian and Hamiltonian dynamics, which are immensely powerful ways of solving very difficult problems rely very much on the calculus of variations. Finding a method, and demonstrating it, which allows us to do some of these problems without advanced maths is a worthwhile result. This method works to solve the problem of which shape with a given perimeter bounds the greatest area (the circle), and the problem of what shape is formed by a soap bubble between two metal rings (the surface of revolution of the catenary, the catenoid). It even allows us to *derive Newton's second law of motion from the principle of least action (particles are lazy, so if we minimize the integral which tells us how much action they have to take, we can discover their motion). On the other frontier, we've stumbled across the way of making a much better type of clock. This is an interesting problem, so that's where we'll go next.*

In which an answer is proposed and a clock is made

We've had a bit of a romp through calculus land in investigating the tautochrone and brachistochrone properties, so for the last major all out attack on the cycloid we'll restrict ourselves almost entirely to geometry. The whole point of what follows is to lay the foundations for building a real life tautochrone clock. We know that we an make such a clock by rolling something along a cycloid, but it would be really nice if we could somehow make a cycloid pendulum; a pendulum whose bob swung along the arc of a cycloid. Before the days of the Global Positioning System, it was immensely difficult for navigators to calculate their position on earth. A rough idea of latitude was available from the declination of the sun, but longitude was a much harder problem. Of course, if reliable clocks could be made, a shipboard clock could keep Greenwich Mean Time, and by comparison with local time, again available from the sun, the longitude could be calculated. Such an accurate clock represented an extraordinary challenge, and was the focus of attention of many mathematicians and engineers for some time. This problem is the difficult question, and the cycloid pendulum is the answer we propose. If it weren't for three or so intervening centuries, this would represent the cutting edge of mathematical research.

First of all I'll introduce the concepts of 'evolutes' and 'involutes' in general, mention one or two interesting things about them, and then use these ideas to uncover yet another amazing property of the cycloid.

Both evolutes and involutes are curves derived from another curve. Hence we can speak of the evolute of a parabola, or the involute of a circle, and so forth. The best way to get to grips with the evolute is through diagrams. The basic idea is to draw lots of normals to the curve, and see what emerges.

The evolute

This diagram shows a number of normals drawn to a cycloid. What is immediately obvious is that another curve has appeared; namely, the curve that marks the top edge of the area filled with normals. This curve is known as the 'evolute', and is officially defined as the 'envelope of the normals'. This simply means that it is tangent to every normal. Notice that each normal furnishes a small part of the evolute, and that the part it contributes lies between the two intersections between the normal we are interested in, and the normals just on either side. What this means is that if we want to find a point on the evolute that corresponds to a particular normal, we need take another normal very close by, and think about what transpires as the two normals approach each other. Because the part of the evolute associated with our normal lies just adjacent to its intersections with nearby normals, the corresponding point on the evolute must be the limiting position of the intersection between two very close normals. By

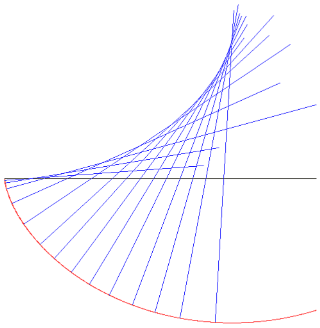


Figure 10

finding the point corresponding to each normal, we can then construct the entire evolute.[9]

The involute

The involute could be called the ‘anti-evolute’; if you take the involute of the evolute of a curve you get back the original curve. This isn’t quite exactly true, but will do as an introduction. This time, instead of taking the envelope of the normals, we find a curve which is perpendicular to all the tangents to our given curve. By considering how we might trace out such a curve, we’ll see why there are very many involutes, but only one evolute.

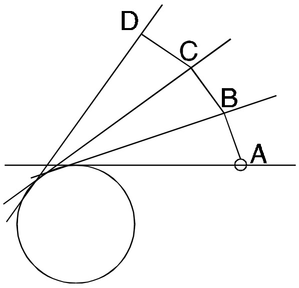


Figure 11

If we begin on a point on a particular tangent, labelled *A* in the above diagram, we need to trace out the curve so that it intersects the next tangent at right angles. We draw the perpendicular to the next tangent, meeting it at *B*, then head off in a different direction to *C*, and then on to *D*. Obviously this doesn’t not actually create the involute, but a chunky approximation of it. However, if we consider many more such tangents, much closer to each other, the curve becomes smoother, and approaches the involute. Of course, we also have to trace backwards from *A* to produce the complete involute. The reason we can have several involutes is that we are free to choose our starting point *A* anywhere along the length of its tangent. In addition, we are free to choose to trace out the involute in either a clockwise or anticlockwise direction. For those who are interested, the involute of a circle is actually an Archimedean spiral.[10] If we can find a differential equation for our curve, by simply taking the negative reciprocal (remember how we construct perpendicular lines in coordinate geometry?), we generate the differential equation of the involute.

The second lemma - arcs of cycloids

Now we have seen what evolutes and involutes are about, we have to pause and uncover another piece of the geometry of the cycloid. If we look at a section of cycloid, from one point to another, between which the generating circle has rolled through a certain angle, we can describe the element of the arc of the cycloid with two, simpler, curves. This is because the cycloid is being generated by two separate motions; the turning movement of the circle, and the concurrent ‘sliding’, which combine to create the rolling movement. Thus, a particular arc is the ‘vector addition’, as it were, of an arc of a circle, and a straight-line segment. Again, a diagram is indispensable.

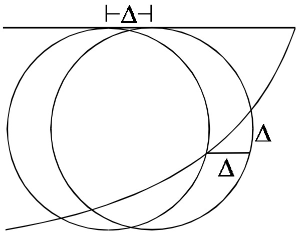


Figure 12

The horizontal interval is obviously of length Δ, and the arc of the circle is of length Δ because that arc length is the same as the arc length that has rolled across the ground. This becomes our second lemma after the one relating normals and generating circles.

It is now time to embark upon a very interesting proof. Because it is very much a geometric result, I’ll have to use the language of geometry and the structure of a formal proof, but I’ll make sure there are lots of diagrams!

The evolute of a cycloid is another cycloid

Consider half an arc of a cycloid, concave up, as *arcAPC*, inscribed in a rectangle *ABCD*, with a variable point *P*, and the corresponding generating circle, meeting the rectangle in *I* on *BC*, and *T* on *AD*.

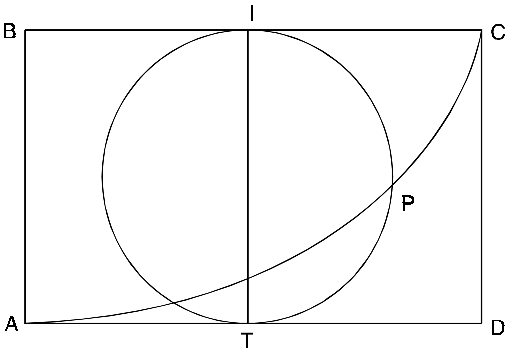


Figure 13

From our first lemma, the normal to the cycloid at *P* passes through *I*. Because $\angle IPT$ is a right angle, it is an angle in a semicircle, the tangent to the cycloid at *P* passes through *T*. We now construct a line parallel to *BC* through *P*, meeting *IT* at *R*, and also construct the tangent to the circle (not the cycloid) at *P*, which will meet *AD* in *S*.

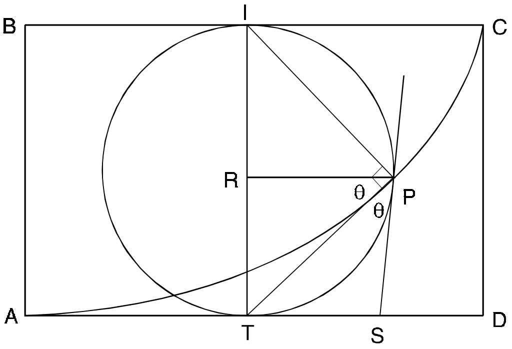


Figure 14

Since $IP \perp PT$, $\angle RPT$ and $\angle RPI$ are complementary, as are $\angle RPI$ and $\angle PIT$. As a result, $\angle RPT = \angle PIT$, and then by the alternate segment theorem, $\angle SPT = \angle RPT$.

Now consider *P* after the cycloid has rolled a further small distance, Δ, moving *I* to I', *T* to T', and *P* to P', such that $II' = TT' = \Delta$. The tangent to the cycloid at P' passes through T', and the normal through I'. Mark Q as the intersection of $I'P'$ and *RP*. If we construct a parallel to *BC* through P' meeting the circle with diameter *IT* in *V*, then $P'V = arcVP = \Delta$, by our second theorem.

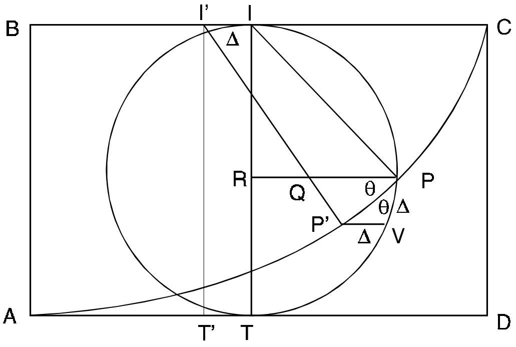


Figure 15

We now have to consider the region bounded by $P'QPV$. As Δ becomes very small, $arcVP$ becomes a straight line, and the arc of the cycloid, $arcPP'$ also becomes a straight line, and approaches the tangent to the cycloid PT. As a result we can draw $P'QPV$ as a trapezium (remembering $QP \| P'V$).

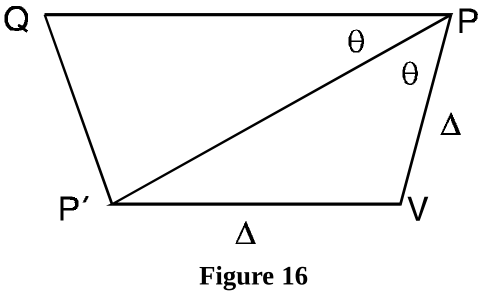


Figure 16

We know construct a line parallel to QP' through V, meeting QP in U, and PP' in W. From this we immediately have $QU = P'V = \Delta$. Now $\angle QP'P$ approaches a right angle, because as Δ decreases, $P'I'$ becomes parallel to PI, which is perpendicular to PT, and at the same time P' approaches the line PT. This means $P'I'$ becomes perpendicular to *PT*, which means $QP' \perp P'P$. As a result, $UW \perp WP$, so ΔUVP is isosceles, so $UP = VP = \Delta$, and therefore $QP = 2\Delta$.

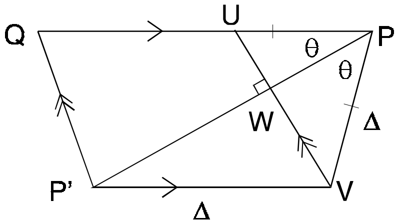


Figure 17

Now, PI and $P'I'$ are two normals to the cycloid, so as Δ goes to zero, the intersection of these two normals will give us the position of the centre of curvature, and hence a point on the evolute. We can readily find this intersection, by extending PI and $P'I'$, to meet at F. Because $QP = 2\Delta$ and $II' = \Delta$, the triangles $\Delta II'F$ and ΔQPF are similar, in a ratio of 1:2, and hence $2IF = PF$, or $IF = PI$. What this means is that the centre of curvature at any point on the cycloid lies along the normal (it couldn't be otherwise), the same distance again between the cycloid and the base of the generating circle. I haven't been able to find anyone else who's done this particular proof, but presumably any geometric proof of this fact will follow the same approach of considering two nearby normals. The centre of curvature of a cycloid can however be attacked very easily using calculus; the cycloid seems to have been designed so that the curvature equations will be tremendously simple. It's tempting at this point simply to show you a diagram, and in the tradition of great Indian mathematicians, offer a simple suggestion: "Behold!"

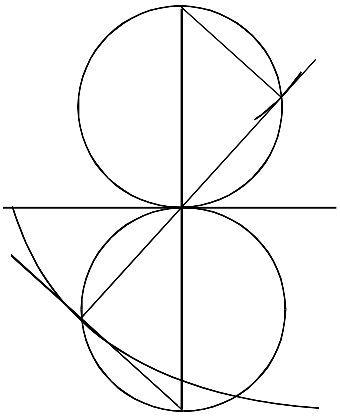


Figure 18

This new information about the centre of curvature actually allows us to prove something incredible about the cycloid; it is its own evolute. We can see this because our just completed result, as shown in the above diagram indicates an interesting symmetry. Above the generating circle we have started with, we find another one, and the interesting thing is that any tangent to the evolute (which must lie along the normal to the original) passes through the bottom of the new generating circle. This means that the normal must pass through the top of the new circle, because the normal and the tangent are perpendicular, and hence form an angle in a semicircle. If we consider the property of the cycloid as having normals passing through the base of the generating circle as a *defining* property of the cycloid, then the evolute must also be a cycloid, simply because it fulfils this property. That this property is sufficient to describe a cycloid doesn't require a huge suspension of disbelief. After all, we even know that the angle (measured from the ‘top’) at the centre of the new generating circle is supplementary to that in the original, by symmetry, so the evolute is being traced out locked in step with the original, but ‘upside down’.

A digression on self-evolute curves

The cycloid is not alone in being self-evolute. This property is also shared by classes of curves known as epicycloids and hypocycloids. These are created in a similar manner to our cycloid. Instead of a circle rolling along a plane, these involve circles rolling on circles. The epicycloids are traced out by a point on a circle rolling around the outside of a fixed circle, and for the hypocycloids on a circle rolling inside the fixed circle. The fixed and moving circles may have different radii. There are two common special cases of the epicycloid, the cardioid, for which the two circles have equal radii, and the nephroid, for which the moving circle has half the radius of the fixed circle. The hypocycloids can be considered as epicycloids but with moving circles with negative radii. An interesting case is found when the radius of the moving circle becomes insignificant compared to the radius of the fixed circle. The fixed circle now becomes like a straight line, and in fact the cycloid we have investigated is the limiting case. The cycloid could in fact be seen as the intermediate curve between the epicycloids and the hypocycloids, when the radius becomes very, very, small as it changes sign from positive to negative. The self-evolute property is in fact a property common to all these curves. One or two other curves are also self-evolute, including the equiangular spiral.

The question that might now be coming to mind is why we’ve had this detour into the geometry of evolutes. The self-evolute property seems to have little to do with the discussion of the physical properties of the cycloid. However much I would like to be able to tell you that the self-evolute property is in fact equivalent to the tautochrone and brachistochrone properties, in some fundamental way, I can’t. I’m certain however there is at least some connection, but the fact that the cycloid only has the tautochrone and brachistochrone properties under certain potential fields,[10] and the fact that some completely alien curves such as the equiangular spiral are self-evolute disinclines me to be hopeful. For now, our interest in the self-evolute property lies in another direction. The tautochrone property of the cycloid is literally begging for the cycloid to be used as a clock mechanism; this is of course why it interested early mathematicians so much. The challenge to build reliable clocks was a very important part of the longitude problem, and Huygens investigated the tautochrone property with this in mind.

Since we’ve ascertained the shape of the tautochrone, it would be most interesting to attempt to make a tautochrone clock. The obvious way to do this is by making a ramp in the shape of a cycloid. A ball bearing rolling on this ramp will then execute simple harmonic motion, hopefully keeping perfect time even as the amplitude decreases. Mr Braga made several such cycloid ramps, with 1 and 2 second periods. Undeniably these were amazing to watch. On the other hand, aesthetically there were one or two problems. In our derivation, rotational energy of the ball rolling along the ramp was ignored, and its derivation was only valid for particles sliding along wires or swinging as pendulums and so forth. However, as it turns out the rotational energy is not really a problem. The maths is slightly different, but the tautochrone is still a cycloid. More importantly these are ramps; there will presumably be considerable friction over many cycles. We wouldn't consider making a normal pendulum clock with a ball rolling in a circular arc. We can do much better now; a cycloid pendulum would be far more attractive, and the solution to the problem of how to force the bob to swing along a cycloidal path is at hand.

We need to return to involutes, and consider what the involute-evolute relationships mean 'in the real world'. Start, for example, with a circle with a length of cotton wrapped around it many times, and a pen tied to the end. Slowly unwrap the cotton, tracing the path of the pen. The resulting curve looks like figure 19.

The secret life of the involute

This is the involute of a circle. This makes sense; the involute is the curve which is always perpendicular to tangents to the given curves, just as the movement of the pen is always perpendicular to the cotton. This perpendicularity comes about because the point at which the string is unwrapping at any instant is the instantaneous centre of motion. Basically, when the string is unwrapped a small amount, the length of the string fundamentally doesn't change; although more string has unwrapped, the point at which it is unwrapping from is the same distance further away. The pen now has to move to a different tangent, but with the same length of string. Any movement that changes the angle without changing the radius of movement is essentially circular movement, and in such, the movement is always at right angles to the normal. It is easy to imagine the shapes traced out as a string unwrapped from various objects, and this will always give you the involute. The many different involutes correspond to initial different lengths of string.

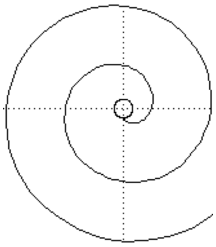


Figure 19

The cycloid pendulum

Now, to make the bob swing along a cycloid, we clearly need the string to be 'unwrapping' from a curve which is the evolute of a cycloid, so that the cycloid is the involute of this shape. And, as we have already proved, this shape is in fact another cycloid, of equal dimensions but displaced. By referring to figure 10, we can see how the two cycloids are situated relative to one another.

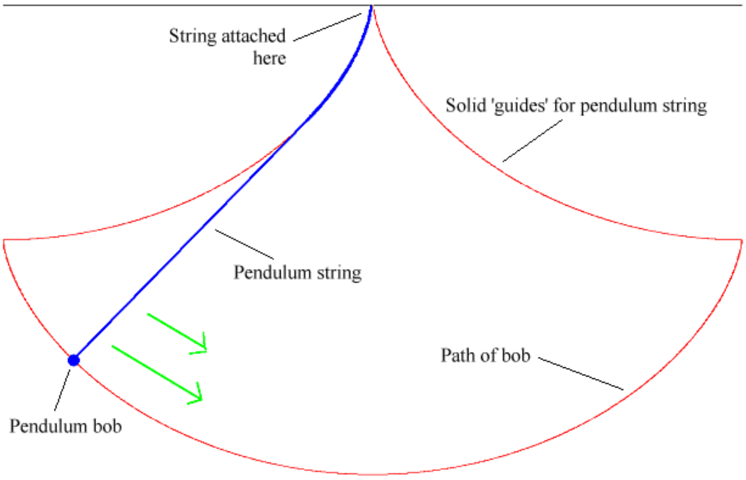


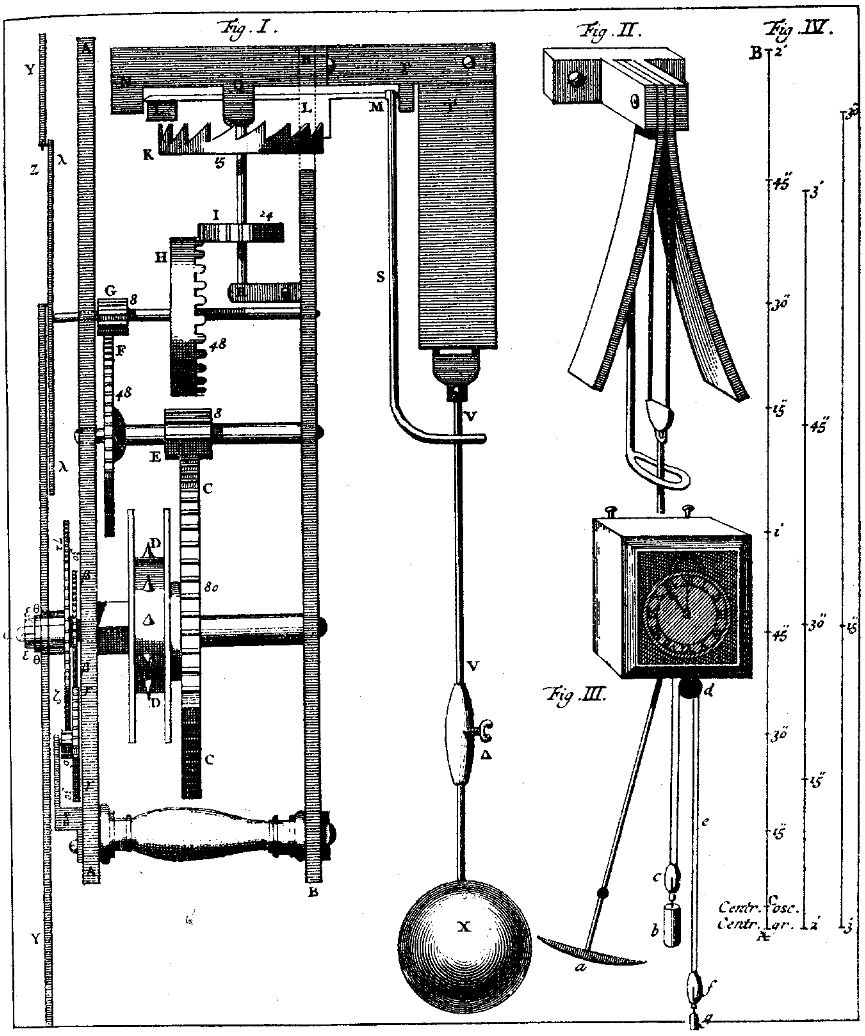
Figure 20

Just as in a circular pendulum, the string exerts a force at right angles to the velocity, and hence does no work. Because of this, we can ignore the string and know that the bob will move just as a particle sliding along a cycloid arc.

In this case, the different involutes of the upper curve correspond to the paths traced out with different length strings. However, only when the length of the string is the same as the arc length of one ‘wing’ of the guide is the path traced out by the bob a cycloid of the same size. We can calculate the period of oscillation using our early result with the tautochrone. We determined that the period would be $2\pi\sqrt{\frac{4r}{g}}$. This is interesting, because the ‘height’ of a loop of the cycloid is $2r$, so the full height of our pendulum apparatus is $4r$, which is the value that appears in the expression for the period where the length of the pendulum string would normally appear. This makes sense, because if a cycloid pendulum is only swinging through small amplitudes, we can ignore the guides, and it becomes a circular pendulum, with length $4r$. We can also tell from our diagram that the arc length of each of the guides is $4r$, because the string will wrap exactly around them. This result can be easily confirmed by calculus.

Huygens discovered this very method of utilizing the tautochrone property, although the cycloid was never employed as a practical clock due to technical difficulties. It is very interesting however, that although he eventually did this, it wasn’t without a few mistakes. While Huygens was still working out the details of the theory of involutes and evolutes, he actually tried to use a section of the involute of a circle (figure 19) as the guide for a cycloid pendulum. He knew that he needed to make the bob swing in a cycloid by this stage, but did not know about its self-evolute property. It is sort of reasonable that he might try this; after all, the involute of a circle is constructed rather like a cycloid, but instead of having a circle roll over a fixed straight line, the involute of a circle involves a straight line rolling around a fixed circle. Finally however, Huygens, who notably worked without calculus, uncovered some of the same amazing truths about the cycloid that we have.

And, finally, we reach what is for now the end of the journey. I would like to thank those of you who have read this far; I’ll make the overconfident assumption that if you’re reading the last page, you’ve read the whole article! I’ll also offer my thanks to a number of other people. Firstly, thanks to Andrew Snow and Benji Chung for assisting in building cycloid pendulums, Mr Braga for demonstrating his cycloid ramps, and Mr Ward for providing interesting mathematical literature on the subject. Dr Pender deserves thanks for allowing me to try out most of these ideas before the Maths Group- it’s invariably useful to try to explain things to other people, and occasionally results in the realisation you didn’t know anywhere near as much as you’d hoped. Finally, many thanks to Dr Bishop, for introducing me to both the tautochrone and brachistochrone problems, for being available to discuss the maths, and for the indispensable encouragement, assistance, and advice!



A drawing of Huygens' cycloidal pendulum clock

(The original appeared in *Horologium oscillatorium.* This picture was taken from E Segrè, *From Falling Bodies to Radio Waves*, **1984**, WHFreeman and Company, New York.)

Further reading:

Lagrangian Dynamics, D. A. Wells, **1967** McGraw-Hill

Rocket-fuel powered maths; the calculus of variations, and the brachistochrone.

Statics and the Dynamics of a Particle, W.D. MacMillan, **1958**, Dover

Discusses tautochrones in the most general and abstract way possible, the brachistochrone using the calculus of variations, and demonstrates annoyingly simple derivation of the cycloid as the tautochrone.

Geometry and the Imagination (Anschauliche Geometrie), D. Hilbert and S. Cohn-Vossen, translated by P. Nemenyi. Chapter IV Differential Geometry, Chapter V Kinematics

Discussion of the evolute and involute in general, and curvature, epicycloids and hypocycloids

MacTutor Famous Curves Index, http://www-groups.dcs.st-and.ac.uk/~history/Curves/Curves.html

Source of historical information about and diagrams of cycloids, epicycloids, hypocycloids, involute of the circle, and equiangular spirals

The Feynman Lectures on Physics, R P Feynman et al, Vol II, **1964** Addison-Wesley (Lecture on the principle of least action)

Best described as 'The calculus of variations for dummies'.

Footnotes

[1] *Mathematica*, student version 2.2, Wolfram Research Inc.

[2] $\sin 2\theta = 2\sin\theta\cos\theta$ and $\cos^2\theta = \frac{1}{2}(\cos 2\theta + 1)$

[3] Boyer, C. B. *A History of Mathematics*, New York: Wiley, **1968**. p. 389.

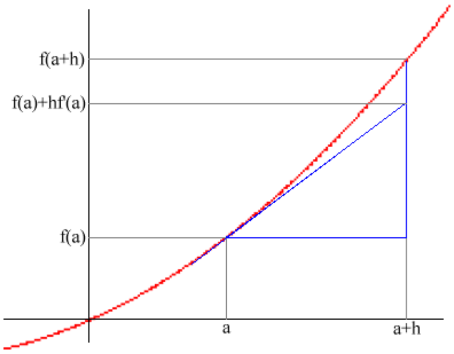
[4] Our lemma tells us that $TP \perp PB$, and of course $OB \perp BT$. Because the radii of a circle form an isosceles triangle, we also know that $\angle BPO = \angle PBO$. Simple angle chasing should now explain the substitution. If $\angle PTB = \psi$ (according to our previous definition), $\angle PBT = \frac{\pi}{2} - \angle PTB$, and $\angle OPB = \frac{\pi}{2} - \left(\frac{\pi}{2} - \angle PTB\right) = \psi$, and then by the angle sum of the isosceles triangle, $\angle POB = \pi - 2\psi$, which is exactly the substitution we made earlier.

[5] Thanks to R.P. Feynman!

[6] Fermat's principle telling us about the path of light isn't really a principle of least time, but a principle of local minima of time (this is why light reflects in a mirror!). Here the idea of minima as region where the function doesn't change much really comes into its own; light doesn't actually take paths of least time because it's in a hurry. Instead it actually goes every place at once, and as it travels along its phase changes (light will quite obligingly be a wave if that's what you expect it to be). At most regions, the bits of light that land nearby each other have different phases. This is because there *is* a first order change in time for a small change in position, so nearby beams of light take different amounts of time to reach their destination. Since the phase of the light 'oscillates' with time, this means that the light destructively interferes with itself in these areas, and as a result ends up not really being there, because it has almost entirely cancelled itself out. At local minima all the nearby rays have the same phase (now there are *no* first order changes in time over small changes in position), and these add up, constructively interfering, resulting in light's apparent behaviour as obeying a principle of least time.

[7] The binomial theorem will certainly give us a term that will cancel out with the unshifted integral. Unfortunately, when the binomial theorem is used on square roots, it gives an infinite series of terms. This isn't itself a problem, as most of these can be discarded if we are only looking for first order changes. There are still a few terms left over, however, which neither cancel nor can be immediately disregarded. In addition, it is not immediately obvious if these terms create first, or second, or whatever degree changes, because they have fractional indices. The binomial theorem was not going to provide a clear way forward. I should mention now that the binomial theorem does actually work; everything appears messier at first, but since I first got this problem out, I've had a look at using the binomial theorem, and the proof in fact proceeds in an almost exactly identical fashion.

[8] We can obtain at least a feeling for the truth of this approximation with the aid of a diagram illustrating the first two terms. You can see in the diagram how by just taking into account the first derivative we can obtain a fairly accurate approximation. Successive derivatives improve this by taking into account the concavity, and the rate of change of concavity, and so on. The form of these first two terms should hopefully be clear from the ordinary meaning of derivatives, as rates of change.



9 There are a number of other means of finding the evolute. The first of these introduces something called the circle of curvature. Just as a tangent tells us something about the gradient of a curve, the circle of curvature includes some extra information, now about the curvature of the curve. Curvature is a relative of concavity, and the two can be calculated from each other (as long as the gradient is known). You could almost say that gradient is to angle as concavity is to curvature. As it will turn out, the evolute is actually the locus of the centre of the circle of curvature. One way of regarding a tangent is as the line through two nearby points on a curve, creating a secant; as the two points approach each other, the limiting case of the secant becomes the tangent. The equivalent approach for the circle of curvature, which is often called the osculating circle, involves three points (just as two points can define a line, three will define a circle). All three of these points are on the curve, and as they come very close together, the circle through them both becomes tangent to the curve, and also takes on the same ‘curvature’ of the curve.

It is actually quite easy now to see why the definition of the evolute as the locus of the centre of the circle of curvature is consistent with our idea of it as the intersection of nearby normals. As the two normals draw closer, their lengths become the same, because they both start in the same region and end at their point of intersection. This means we have a circle, this time not defined by three points on a circle, but by two radii drawn from a common centre (the point of intersection). We can then see that this circle is actually the circle of curvature, because the fact that the normals are actually the perpendiculars to the curve means that the angle between them ‘characterises’ the curvature at that point. It should therefore be fairly safe to claim that this circle is actually an alternatively constructed circle of curvature. Again, the locus of the centre of this circle defines the evolute.

The other method uses calculus; curvature lends itself quite well to calculus, and for some curves is even more amenable to analysis than concavity. Of course, the calculus is equivalent to the geometric use of circles of curvature, but for some shapes calculus shows us the evolute very quickly.

10 Or almost; the Archimedean spiral is the pedal curve of the involute of a circle. The pedal curve is the locus of the intersections of tangents to the curve with perpendiculars from the pedal point. Taking the pedal curve of the Archimedean spiral does not change the curve much, but pulls the innermost point into the origin, and shifts the whole curve slightly clockwise.

11 The brachistochrone and tautochrone problems can also be solved for a considerable variety of arbitrary potential fields. The methods I’ve proposed work without modification for all potential fields which are only a function of height (ie, not of horizontal position).

Introducing U as the potential, and assuming that $\frac{\partial U}{\partial x} = 0$, we now derive a differential equation for the brachistochrone as

$$\frac{dy}{dx} = \sqrt{\frac{1 - kU}{kU}}$$

and for the tautochrone as

$$\frac{dy}{dx} = \sqrt{\frac{kU}{\left(\frac{dU}{dy}\right)^2 - kU}}$$

Under gravity, $U \propto y$, these differential equations become equivalent after a shift of origin. These equations make it clear that the tautochrone and brachistochrone are the same curve only under gravity-like potential. For example, under a potential field where $U \propto y^2$, the brachistochrone becomes the arc of a circle, while the tautochrone remains a cycloid.