A New Minimization Proof for the Brachistochrone

Gary Lawlor

The American Mathematical Monthly, March 1996, Volume 103, Number 3, pp. 242–249.

1. INTRODUCTION. The cycloid is the curve traced out by a point on the circumference of a rolling circle. See Figure 1. This curve has two additional names and a lot of interesting history.

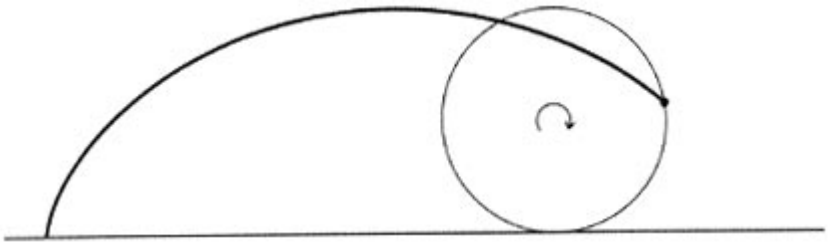

Figure 1. A cycloid, traced out by a fixed point on a rolling circle.

The other two names are the “tautochrone” and the “brachistochrone.” In order to talk about the properties of the cycloid that gave it these names, we need to turn the curve upside-down, so that it is concave upward. Thus, throughout this paper, a rolling circle will always be rolling “on the ceiling.”

The word “tautochrone” means “same time.” In the 1600s Christian Huygens discovered a pendulum whose period is independent of how high the bob swings; this is only approximately true for a regular pendulum.

Huygens found that if obstructions in the shape of half cycloids are placed so that the string of a pendulum wraps around them as it swings (see Figure 2), then the bob also traces out a cycloid. Further, the period of motion is independent of whether the bob makes the full swing or swings only part of the way back and forth. This is equivalent to saying that if we build a ramp in the shape of a cycloid and let a marble roll down it, the time is takes to reach the lowest point is independent of where on the cycloid we started the marble. See Section 8.

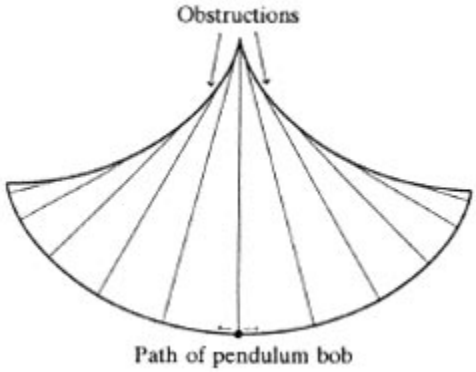

Figure 2. Christian Huygens’ pendulum, with period independent of height of swing.

Then in the late 1600s, the mathematician Johann Bernoulli posed a question and invited mathematicians of the time to solve it. He asked, "Let two points A and B be given in a vertical plane. Find the curve that a point M, moving on a path AMB must follow such that, starting from A, it reaches B in the shortest time under its own gravity."

Five prominent mathematicians of the time solved the problem, namely Johann and Jakob Bernoulli, Leibniz, L'Hopital, and Newton. They showed that the solution is also the cycloid, and gave the cycloid its name "brachistochrone," which is Greek for "quickest." The interesting history (including the above quotation from Bernoulli) and an outline of Johann Bernoulli's proof can be found in V. M. Tikhomirov's wonderful book, "Stories about Maxima and Minima" [**T**].

In this paper we will call the two endpoints P and Q. Bernoulli's "path AMB" we will call a ramp, and the point M we will call a marble.

We will give a new proof that the brachistochrone is the shortest time ramp, using the idea of slicing described in the paper [**L1**]. The philosophy of slicing is to compare two quantities by slicing both into tiny pieces that are easier to compare. We begin with an object that we hope to prove has least volume, area, length, or time among a certain class of competitor objects. We compare this "champion" object with an arbitrary competitor. The goal is to find a strategy for slicing both in such a way that each piece of the champion is smaller or shorter than each piece of the competitor.

The picture of the proof is closely related to Huygens' cycloid pendulum. We will then outline another proof that uses the same basic ideas but demonstrates some additional interesting geometry of the cycloid.

Various modifications of Bernoulli's original question can also be answered by the method of this paper; see Section 9.

We also briefly comment on why the cycloid is the tautochrone.

2. GENERAL FACTS

2.1. Proposition. *The velocity of a marble rolling without friction down a ramp is proportional to $\sqrt{|y|}$, if the marble starts at rest at a point where $y = 0$. We will choose units so that velocity is* ***equal*** *to $\sqrt{|y|}$.*

Remark. It is interesting that, in particular, velocity is independent of the shape and length of the ramp and of the x coordinate, and depends only on the y coordinate.

Proof of Proposition 2.1. The kinetic energy of the marble is proportional to mv^2, whether we consider the marble as "sliding" or rolling. The potential energy is mgy. By the law of conservation of energy, the gain in kinetic energy must equal the loss in potential energy. So

$$mg|y| = kmv^2,$$

which means that

$$v = \sqrt{g/k}\sqrt{|y|}.$$ ■

2.2. Proposition. *The tangent line to a cycloid passes through the lowest point of the rolling circle, and the normal line passes through the highest point.*

Proof: The proof is shown in Figure 3. We draw a vector A tangent to the rolling circle, with the length of A chosen so that it reaches to the horizontal line as in the figure. From its endpoint we draw the horizontal vector B that reaches to the lowest point of the circle. By a general symmetry property of circles, the vectors A and B have the same length. On the other hand, the velocity vector of the point on the rolling circle's circumference can be written as a sum of two vectors, one tangent to the circle (from the rotation) and one horizontal (from the motion of the circle's center). Since the circle rolls without slipping, these two vectors must be of the same length. This means that the velocity vector of the moving point is a multiple of $A + B$, so that the tangent line does pass through the lowest point of the circle.

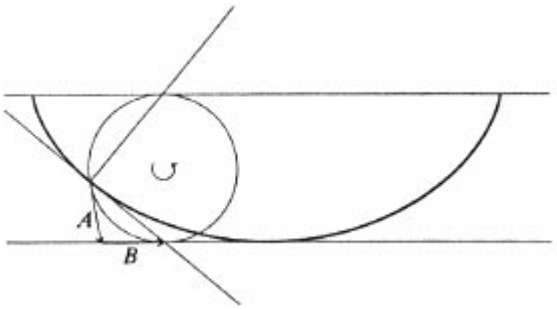

Figure 3. The tangent and normal lines to the cycloid pass through the lowest and highest points of the rolling circle.

Now by a general property of circles, the normal line automatically goes through the opposite point, namely the highest point of the circle. Alternatively, one can rescale A and B into unit vectors, and consider $A^{\perp} + B^{\perp}$. ■

2.3. Proposition. *In Figure 4, the segment KL is perpendicular to the lower cycloid and tangent to the upper cycloid. The point J is the bisector of the segment KL.*

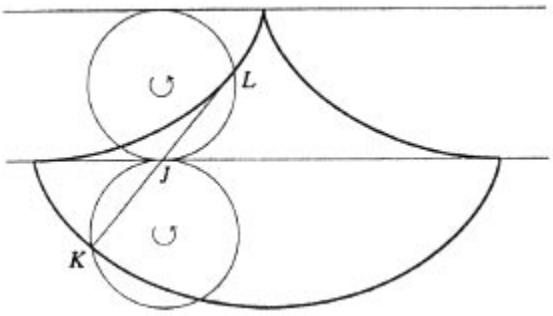

Figure 4. The tangency point J is the midpoint of segment $\overline{KL}$.

Proof: The two circles in the figure are to be thought of as rolling simultaneously to the right, both rolling "on their ceilings" (counterclockwise) and thus slipping against each other rather than rolling like gears. They trace out the upper and lower cycloids. It is interesting to note the similarity between Huygens' tautochrone pendulum and our proof of Bernoulli's brachistochrone property.

In the figure, K is defined as the fixed point on the circumference of the lower rolling circle, and L is the fixed point of the upper circle. Since the circles roll at the same speed, K and L are always opposite each other, with the tangency point J being the midpoint of the line segment between them.

By two applications of Proposition 2.2, since the line segment KL goes through J, the segment is perpendicular to the lower cycloid and tangent to the upper cycloid. ■

3. THE BRACHISTOCHRONE IS THE SHORTEST TIME RAMP. We now state and prove the main theorem.

3.1. Theorem. *Let P and Q be two given points in a vertical plane, with the y coordinate of Q no higher than that of P. Let M be the cycloid whose tangent vector points straight down at point P, and whose generating circle has the right radius so that the cycloid passes through the point Q. For any path from P to Q, consider the time it takes for a marble, starting at rest at P, to roll without friction down to Q. Among all such paths, M is the one on which the marble takes the least time.*

Proof: Our goal in this paper is to present the proof in as elementary a manner as possible. Thus, although each of the following lemmas is true exactly, in proving the second one we will ignore tiny approximation errors. In Section 7 we sketch a rigorous proof by calculus.

Figure 5 is the picture of the proof. We draw narrowly-spaced straight lines perpendicular to the cycloid, and call the lines "strings," reminiscent of Huygens' pendulum. We let P be the highest point of the cycloid on the left, and let Q be any other point of the cycloid. Now every ramp on which a marble may roll from P to Q must cross all of the strings. By Lemma 3.3, the ramp on which a marble travels most quickly from any one string to the next is the cycloid. Thus, by adding up the local results we obtain the theorem. ■

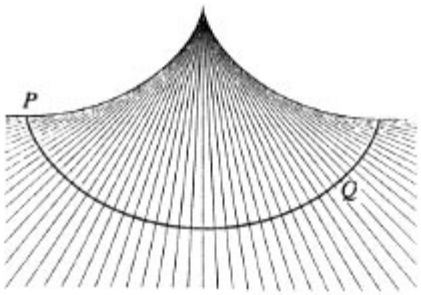

Figure 5. The main idea of the proof that the cycloid is time-minimizing.

3.2. Lemma. *Given a constant $r > 0$, the minimum value of $(r + x)/\sqrt{x}$ occurs where $x = r$.*

Proof: Of course, this is straightforward calculus. Or for a student who has not had calculus, one can write

$$\frac{r + x}{\sqrt{x}} = \frac{r}{\sqrt{x}} + \sqrt{x}.$$

In the latter form, we have a sum of two terms whose product is the constant r. Thinking of the terms as lengths of sides of a rectangle, we see that the area of the rectangle is constant and we want to minimize the (semi)-perimeter. This is accomplished by a square, so we set

$$\frac{r}{\sqrt{x}} = \sqrt{x},$$

so that $x = r$. ■

3.3. Lemma. *Given any two consecutive strings in Figure* 5*, for any ramp starting at P consider the tiny increment of time it takes for a marble to cross the gap from the one string to the next. This time increment is always minimized by the cycloid.*

Proof: Pick two consecutive strings, as in Figure 6. There are two factors working against each other in the competition for a fastest path across the gap. Up higher, the gap is narrower, but the speed will be smaller. Very near the top the speed approaches zero but the width does not, so the time will in fact be greater there. Down lower, the speed is greater but the gap is wider. Asymptotically the width grows like y and the speed grows like $\sqrt{y}$, so the time is larger there as well. Thus, somewhere in the middle the crossing time is minimized. Let U be the intersection point of the two chosen strings. Let r be the distance from U to the point V where the first string crosses $y = 0$. Let x be the distance from V to a place where some ramp crosses the gap. Now the velocity is proportional to $\sqrt{x}$, and the width of the gap is proportional to $r + x$, so the time to cross orthogonally is proportional to $(r + x)/\sqrt{x}$, which is minimized when $r = x$, which by Proposition 2.3 is where the brachistochrone crosses the gap. (Note that we have used the approximating assumptions that the speed of the marble is constant as it crosses the gap, that the quickest ramp across the gap is a line segment perpendicular to the first string, and that the two strings intersect at the point where the first one is tangent to the upper cycloid in Figure 4.)

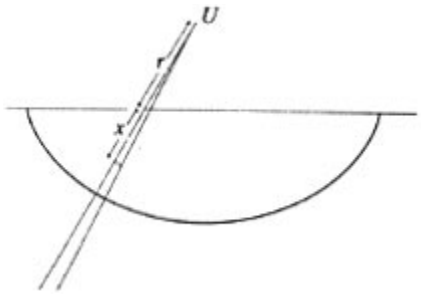

Figure 6. The quickest place to cross the gap is where $x = r$.

4. FINDING THE CYCLOID IN THE FIRST PLACE. We have shown how to prove that a cycloid is the shortest-time ramp. We have not mentioned how to discover the cycloid as a candidate in the first place. This was done by the 17th century mathematicians by setting up and solving a differential equation. Michael Kerckhove at the University of Richmond had the idea that slicing could also be used as a tool for finding such a differential equation. This approach is proving successful and is a topic of current research.

5. COMPARISON WITH OTHER PROOFS. Recall that the strategy of slicing is to compare two quantities by slicing them into tiny pieces that are easier to compare and for which the desired inequality still holds.

The solutions found in the 17th century by Leibniz and by Johann Bernoulli can both be described as slicing with horizontal lines; see [**T**], pp. 58-62. The modern technique called the calculus of variations can be viewed as slicing with vertical lines; see L. C. Young's wonderful book [**Y**]. With horizontal and with vertical slicing, it is not true that each piece of the curve will minimize time across its respective gap. Competitor curves will take a shorter time to cross some gaps and a longer time to cross others. In the calculus of variations proof, the extra bookkeeping thus required is accomplished by integration by parts. Leibniz allows only a tiny piece of the curve to vary at a time, and Bernoulli argues by analogy with the refraction of light.

Young is careful to point out that in all of these methods mentioned above a lot of work remains to complete the proof. A complete argument goes as follows: One shows that there does exist a time minimizing curve, and then shows that such a curve (if it is smooth) must satisfy a certain differential equation, and finally finds **all** curves satisfying the differential equation and compares them to see which one takes least time.

In contrast, the method of slicing does not depend on a proof of existence of a minimizing curve. And, although slicing partially localizes the problem, we still keep a sufficiently global view to keep track of all competitors throughout the process.

6. SKETCH OF ANOTHER SLICING PROOF. We now sketch a second proof, which is done by the same basic idea as the first, namely, slice up the plane with curves perpendicular to the brachistochrone ramp, such that the fastest way across the gap between consecutive curves is that taken by the brachistochrone. The geometry is interesting. The result obtained is not quite as strong as before.

For this proof, we can only take half of one cycle of a cycloid, not extending beyond the point where the cycloid becomes horizontal. This time, the slicing curves are themselves cycloids. They have the same size and shape as the brachistochrone cycloid, but are turned right side up (see Figure 7). It is a wonderful fact that no matter how far we shift the cycloids left or right, if they intersect the brachistochrone they do so orthogonally. (The two horizontal lines in Figure 7 are parallel; the apparent narrowing to the right is an illusion.)

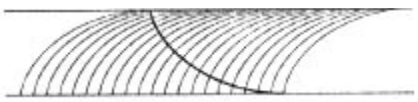

Figure 7. The picture of another slicing proof.

Now consider a fixed gap between consecutive slicing curves. It can be calculated that the shortest time path across the gap is not unique; it takes the same amount of time to cross the gap orthogonally no matter where you cross it.

If *S* is some other ramp with the same starting and ending points as the (half) brachistochrone, and if *S* does not go "below ground", i.e., below the level of the

endpoint, then S must cross all of the slicing cycloids. The best it can do is to cross all of them orthogonally, which is what the brachistochrone ramp does.

7. PROOF SKETCH OF THEOREM 3.1 USING CALCULUS. One way to include calculus is to define a function $f(x, y)$ by declaring it to equal zero at the top of the brachistochrone ramp, and at each point of the ramp let it equal the time it takes for the marble to get there. Extend the function by letting its level sets be the slicing curves (the straight strings of the first proof or the upside-down cycloids of the second proof). Now prove the inequality $\|\nabla f\| \leq 1/\sqrt{|y|}$ with equality on the brachistochrone.This can be done by implicit differentiation. Let M be the brachistochrone and S a comparison ramp with the same endpoints. Then write the inequalities

$$\text{Time}(M) = f(Q) - f(P) = \int_M \nabla f \cdot \mathbf{ds} = \int_S \nabla f \cdot \mathbf{ds}$$

$$\leq \int_S \|\nabla f\|\, ds \leq \int_S \frac{1}{\sqrt{|y|}}\, ds = \text{Time}(S).$$

This proof is an example of the method called "calibrations," whose merits were demonstrated in a 1980 landmark paper by Reese Harvey and Blaine Lawson [**HL**]. The method of calibrations (like slicing) provides a global minimization result. Further comments on the comparison between calibrations and slicing are found in the slicing paper [**L1**]. A good exposition on calibrations is the paper [**M1**].

8. THE TAUTOCHRONE. Why does it take the same amount of time for a marble to roll down to the center point C of the ramp, regardless of where we start it? The answer is that the motion is harmonic. If we set a marble at the center point C, there will be no force to move it from this equilibrium point. Now it turns out that if we set a marble at any other point of the ramp and measure its distance from C (measuring arc length along the ramp), then the force acting on the marble (the component of gravity parallel to the ramp) will be proportional to the distance away from C. So if we start marble M some distance away from C, and we simultaneously start marble N (say) twice as far out from C, then marble N will always have twice the force acting on it, so twice the acceleration, so twice the velocity, so it will go twice as far as marble M in the same amount of time, and both will reach C at the same time. The same is true with the word "twice" replaced by "α times as much," for any positive real number α.

9. OTHER QUESTIONS. The geometric proof of the time-minimizing property of the brachistochrone (the one with straight lines) opens the door to the investigation of interesting related problems. Among them are:

(1) What is the quickest path from P to Q if the initial speed at P is nonzero?
(2) What is the quickest path if the cycloid solution is ruled out because it would dip below the level of the floor?
(3) What is the quickest path from a point P to anywhere on a line L?

The first can be answered by calculating how high we should draw the horizontal line $y = 0$. That is, we pretend that a marble has gained its initial velocity by rolling down a ramp from a starting point higher than P. Having drawn the line $y = 0$, we then find the right-sized cycloid that starts vertically at $y = 0$ and then passes through the points P and Q. The proof that this is the best ramp is the same. Draw lines perpendicular to this cycloid; the cycloid is the quickest path across any gap.

The quickest ramp for the second questions consists of two pieces of cycloids with a line segment in between. Find the cycloid that is vertical at P and is tangent to the floor. Divide it in half. Move the right-hand piece to the right until it passes through Q. Join with a line segment along the floor. The proof is again by drawing lines perpendicular to the ramp everywhere. The lines perpendicular to the floor segment are parallel; the quickest path across those gaps is along the floor (since you can't go below the floor).

The third question is solved by a cycloid vertical at P and perpendicular to L, with the same proof.

The reader may enjoy devising and solving questions with more complicated obstructions (like question 2) and/or more complicated free boundary (like question 3).

10. FURTHER READING. For further study of current work in geometric measure theory in proving minimization of area, length, time, etc., I recommend the papers [**Bk**], [**K**], [**L1**], [**L2**], [**LM**], and [**M2**].

REFERENCES

[Bk] Ken Brakke, Soap films and covering spaces, Research report, The Geometry Center, University of Minnesota, July 1993.

[HL] Reese Harvey and H. Blaine Lawson, Jr., Calibrated Geometries, Acta Mathematica **148** (1982), 47–157.

[K] Michael Kerckhove, Isolated orbits of the adjoint action and area-minimizing cones, Proceedings of the AMS **121** (1994) 497–503.

[L1] Gary Lawlor, Proving area-minimization by slicing, preprint.

[L2] Gary Lawlor, A sufficient criterion for a cone to be area-minimizing, Memoirs Amer. Math. Soc. **91**, No. 446 (1991).

[LM] Gary Lawlor and Frank Morgan, Paired calibrations applied to soap films, immiscible fluids, and surfaces or networks minimizing other norms, Pacific Journal of Mathematics, **166 no.1** (1994), 55–83.

[M1] Frank Morgan, Area-minimizing surfaces, faces of Grassmannians, and calibrations, American Math. Monthly **95 no. 9** (1988), 813–822.

[M2] Frank Morgan, Soap films and mathematics, Proc. Symp. Pure Math. **54** (1993), 375–380.

[T] V. M. Tikhomirov, *Stories about Maxima and Minima. Translated from the Russian by Abe Shenitzer.,* Amer. Math. Soc. and Math Assoc. of America, 1990.

[Y] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory,* W. B. Saunders Co., Philadelphia, 1969.

Department of Mathematics
Brigham Young University
Provo, UT 84602
lawlor@math.byu.edu