THE EVOLUTION OF…

Edited by **Abe Shenitzer and John Stillwell**

The Isoperimetric Problem

Viktor Blåsjö

1. ANTIQUITY. To tell the story of the isoperimetric problem one must begin by quoting Virgil:

> At last they landed, where from far your eyes
> May view the turrets of new Carthage rise;
> There bought a space of ground, which Byrsa call'd,
> From the bull's hide they first inclos'd, and wall'd.
> (*Aeneid*, Dryden's translation)

This refers to the legend of Dido. Virgil's version has it that Dido, daughter of the king of Tyre, fled her home after her brother had killed her husband. Then she ended up on the north coast of Africa, where she bargained to buy as much land as she could enclose with an oxhide. So she cut the hide into thin strips, and then she faced, and presumably solved, the problem of enclosing the largest possible area within a given perimeter—the isoperimetric problem. But earthly factors mar the purity of the problem, for surely the clever Dido would have chosen an area by the coast so as to exploit the shore as part of the perimeter. This is essential for the mathematics as well as for the progress of the story. Virgil tells us that Aeneas, on his quest to found Rome, is shipwrecked and blown ashore at Carthage. Dido falls in love with him, but he does not return her love. He sails away and Dido kills herself. Kline concludes [**23**, p. 135]:

> And so an ungrateful and unreceptive man with a rigid mind caused the loss of a potential mathematician. This was the first blow to mathematics which the Romans dealt.

As for the mathematics of the isoperimetric problem, the Greeks pretty much solved it, by their standards, when Zenodorus proved that a circle has greater area than any polygon with the same perimeter. His work was lost. We know of it mainly through Pappus and Theon of Alexandria. Pappus's introduction to the subject (in his *Collection*, Book V) is considered a literary masterpiece. To quote Heath [**18**, p. 389]:

> It is characteristic of the great Greek mathematicians that, whenever they were free from the restraint of the technical language of mathematics, as when for instance they had occasion to write a preface, they were able to write in language of the highest literary quality, comparable with that of the philosophers, historians, and poets.

We quote Pappus's introduction from [**40**, pp. 588–593]:

> Though God has given to men, most excellent Megethion, the best and most perfect understanding of wisdom and mathematics, He has allotted a partial share to some of the unreasoning creatures as well. To men, as being endowed with reason, He granted that they should do everything in the light of reason and demonstration, but to the other unreasoning creatures He

gave only this gift, that each of them should, in accordance with a certain natural forethought, obtain so much as is needful for supporting life. This instinct may be observed to exist in many other species of creatures, but it is specially marked among bees. Their good order and their obedience to the queens who rule in their commonwealths are truly admirable, but much more admirable still is their emulation, their cleanliness in the gathering of honey, and the forethought and domestic care they give to its protection. Believing themselves, no doubt, to be entrusted with the task of bringing from the gods to the more cultured part of mankind a share of ambrosia in this form, they do not think it proper to pour it carelessly into earth or wood or any other unseemly and irregular material, but, collecting the fairest parts of the sweetest flowers growing on the earth, from them they prepare for the reception of the honey the vessels called honeycombs, [with cells] all equal, similar and adjacent, and hexagonal in form.

That they have contrived this in accordance with a certain geometrical forethought we may thus infer. They would necessarily think that the figures must all be adjacent one to another and have their sides common, in order that nothing else might fall into the interstices and so defile their work. Now there are only three rectilineal figures which would satisfy the condition, I mean regular figures which are equilateral and equiangular, inasmuch as irregular figures would be displeasing to the bees. [Pappus goes on to argue that only triangles, squares or hexagons fit around a point.] . . . the bees in their wisdom chose for their work that which has the most angles, perceiving that it would hold more honey than either of the two others.

Bees, then, know just this fact which is useful to them, that the hexagon is greater than the square and the triangle and will hold more honey for the same expenditure of material in constructing each. But we, claiming a greater share in wisdom than the bees, will investigate a somewhat wider problem, namely that, of all equilateral and equiangular plane figures having an equal perimeter, that which has the greater number of angles is always greater, and the greatest of them all is the circle having its perimeter equal to them.

Apart from fascination with bees, motivation for the isoperimetric problem came from astronomy. Theon's account of Zenodorus's proof is found in his commentary on Ptolemy's *Almagest*. (Incidentally, Arabic work on isoperimetry was also motivated by an interest in astronomy.) So let us quote Ptolemy (although the point he makes is not very clear):

The following considerations also lead us to the concept of the sphericity of the heavens. No other hypothesis but this can explain how sundial constructions produce correct results; furthermore, the motion of the heavenly bodies is the most unhampered and free of all motions, and freest motion belongs among plane figures to the circle and among solid shapes to the sphere; similarly, since of different shapes having an equal boundary those with more angles are greater [in area or volume], the circle is greater than [all other] surfaces, and the sphere greater than [all other] solids; [likewise] the heavens are greater than all other bodies.

(*Almagest*, Toomer's translation [**41**])

We must also mention that, apparently, it was a common belief in ancient times that the perimeter of a figure determines its area. Proclus says:

Such a misconception is held by geographers who infer the size of a city from the length of its walls. And the participants in a division of land have sometimes misled their partners in the distribution by misusing the longer boundary line; having acquired a lot with a shorter boundary and so, while getting more than their fellow colonists, have gained a reputation for superior honesty.

(Commentary on the *Elements*, Morrow's translation [**29**])

Gandz [**14**, p. 107] speculates that this may already have stimulated Babylonian mathematicians:

> The typical form of all the quadratic equations in Babylonian mathematics was: Given is the perimeter, $x + y = a$, and the area, $xy = b$; to find the length, x, and the breadth, y.... The great probability is, in the writer's opinion, that the origin of this archaic type of quadratic equations is to be seen as the effect of the aforementioned schemes of those who tried to cheat the plain man in the computation of the capacity of the area.

Zenodorus's polygon proof. We will now see how Zenodorus proved that a circle has greater area than any polygon with the same perimeter.

Theorem. *For regular polygons with the same perimeter, more sides imply greater area.*

Proof. Consider the apothem, the radius-like perpendicular drawn from the center to a side (see Figure 1).

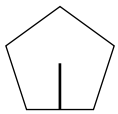


Figure 1.

Half the product of the apothem by the fixed perimeter yields the area of the polygon:

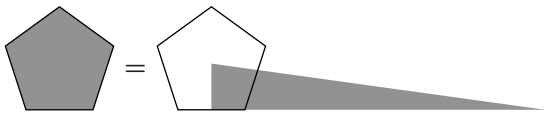


Figure 2.

The apothem is the height of the triangle in Figure 3:

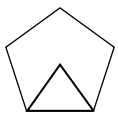


Figure 3.

If we increase the number of sides, the base of the triangle in Figure 3 is shortened and the angle is decreased. It is clear that its height increases. We would prove this by trigonometry; Zenodorus had to rely on the usual pretrig bag of tricks. It is routine for us, and it probably was for Zenodorus as well. ■

We would all be very surprised if this next theorem did not follow from the previous one, but Zenodorus's proof is so delightful that we include it anyway.

Theorem. *A circle has greater area than any regular polygon with the same perimeter.*

Proof. Archimedes proved that the cut-and-roll area formula also holds for the circle (Figure 4).

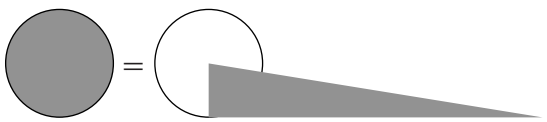


Figure 4.

So we must show that the apothem of any regular polygon is shorter than the radius of the circle with the same perimeter. Rescale the polygon so that it circumscribes the circle (Figure 5):

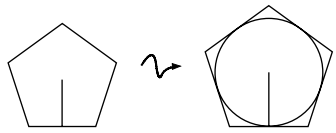


Figure 5.

The perimeter is now greater than the perimeter of the circle, and therefore greater than before the scaling. Thus the scaling was a magnification, with the apothem magnified to the size of the radius of the circle. ■

And now comes the key theorem. Following Zenodorus, we tacitly assume that among all n-gons with given perimeter there is (at least) one that has greater area than all the others. We will say more about this later.

Theorem. *A regular n-gon has greater area than all other n-gons with the same perimeter.*

Proof. Among isoperimetric triangles with the same base, the isosceles triangle covers the greatest area,

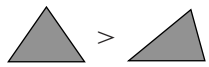


Figure 6.

so the maximal n-gon must be equilateral. Otherwise we could improve on it by making it equilateral.

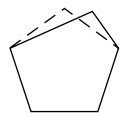


Figure 7.

We now know that the maximal n-gon must be equilateral. Suppose that it is not equiangular. Consider two dissimilar triangles like those in Figure 8:

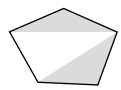


Figure 8.

Now make them similar by redistributing perimeter from the pointy to the blunt angle until the two angles are the same, as shown in Figure 9:

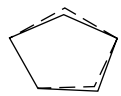


Figure 9.

This increases the area. Accordingly, the maximal n-gon must be equiangular: if not, we could improve on it. ■

For the second part of this argument we need the lemma that, if two isosceles triangles have different bases but their other sides equal, then their total area is increased when they are made similar as described in the proof. Apparently, Zenodorus phrased his lemma more generally by admitting any two dissimilar isosceles triangles (i.e., by not taking advantage of what we showed earlier, namely, that the maximal n-gon is equilateral). This generalization is false.[1] We could ignore that mistake—the generalization is pointless anyway—but to a mind less influenced by the Euclidean tradition the whole scheme for proving equiangularity seems very unnatural.

In view of Zenodorus's imperfections, the contributions of the Arabs—al-Khāzin and others after him—are quite respectable, but we will not discuss them here, since their inventiveness is also constrained by the Euclidean straitjacket. Similar approaches prevailed stubbornly until the eighteenth century (for example, we will see that there is very little difference between the first two theorems that we stated and the discussion of these very theorems in the work of Galileo [**13**]).

2. PRELIMINARIES. In this section we take care of a few things that will be common to many subsequent proofs. First, the problem itself:

The Isoperimetric Problem. *Among all figures with a given perimeter L, which one encloses the greatest area A?*

Ruining the suspense, we reply:

The Isoperimetric Theorem. *The answer is the circle of circumference L.*

To the joy of analysts everywhere, we can rephrase this theorem as an inequality:

The Isoperimetric Inequality. $L^2 - 4\pi A \geq 0$*, with equality only for the circle.*

Sometimes we will find it natural to deal instead with the *dual isoperimetric problem*: Among all figures with a given area A, determine the one that has the shortest perimeter. This is clearly equivalent to the isoperimetric problem, and the bridge between them is scaling. For suppose the circle solves only the dual problem. That would mean that there is some figure with the same perimeter as the circle but with greater area. Rescale this figure so that it has the same area as the circle. But the rescaled figure has a shorter perimeter, a contradiction. Also, obviously, the isoperimetric inequality does not discriminate between the original problem and its dual.

[1] To maximize the area of two isosceles triangles on given bases, we should not aim to make the angles equal but to achieve the following relation between the base angles α, β and the bases a, b:

$$\sin\alpha : \sin\beta = a : b.$$

This is shown geometrically by Steiner [**38**], who says it had already been shown by Lhuilier [**27**] using differential calculus. A counterexample to Zenodorus's lemma might then look something like Figure 10:

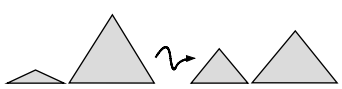


Figure 10.

Here is a lemma that will save us a lot of fuss:

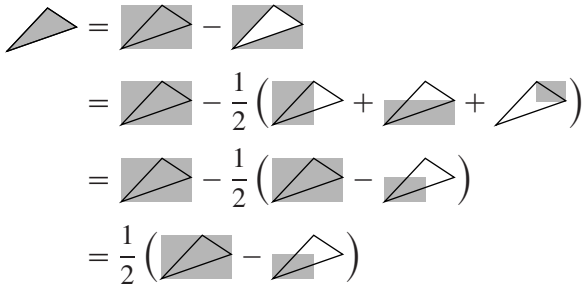
The convexity lemma. *A solution to the isoperimetric problem must be convex.*

Proof. Suppose it is not. Then taking the convex hull—that is, snapping a rubber band around it and taking that as the new figure—increases the area and decreases the perimeter. ∎

Some of our analytic approaches will use the area formula

$$A(D) = \frac{1}{2}\int_{\partial D} x\,dy - y\,dx = \int_{\partial D} x\,dy = \int_{\partial D} -y\,dx.$$

Being geometers at heart, we refuse to think of this as a corollary to Green's theorem. Instead, we derive it from its discrete analogue (which will also come in handy). We begin with triangles.



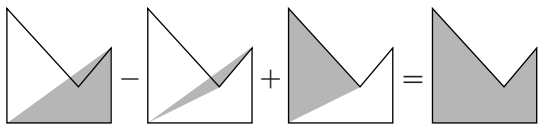
That is, the area of a triangle T, say in the first quadrant, with vertices (for simplicity) $(0, 0)$, (x_1, y_1), and (x_2, y_2) labelled counterclockwise, is given by

$$A(T) = \frac{x_1 y_2 - y_1 x_2}{2}.$$

In general, this is the signed area, as we see by labelling the vertices clockwise instead. For the area of an n-gon P with its vertices $(0, 0)$, (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n) labelled counterclockwise we obtain, by subdividing P into coherently oriented triangles and summing their signed areas,

$$A(P) = \frac{1}{2}\sum_{k=1}^{n-1} x_k y_{k+1} - y_k x_{k+1}.$$

Here is how the signed area takes care of a pathological case:



Taking the limit in the formula for $A(P)$, we get the area for any figure D with suitably smooth boundary ∂D:

$$A(D) = \frac{1}{2}\int_{\partial D} x(y + dy) - y(x + dx) = \frac{1}{2}\int_{\partial D} x\,dy - y\,dx.$$

3. STEINER. We come now to the hero of our story—Jakob Steiner. From the time of Descartes, it had been an explicit aim to reduce the solving of geometrical problems to algebraic calculation, so that the method of solution would be mechanical and general. This scheme was of course enormously successful, and the mathematics it led to is truly a powerful weapon, even in the hands of simpletons. But fiddling with formulas has romanticizing consequences, so we should expect there to have been nineteenth-century mathematicians who put up a fight for geometry. There were, and Steiner was one of them. Needless to say, the battle was lost.

The undertone of rebellion is present in the biographical note on Steiner by Geiser [**15**]:

> Unfortunately, the academic history writing also has its Achilles heel, it's not for nothing that it is called "Éloges" in France. The cool, distinguished tone requires restraint in all doubtful points and contradictions that cannot be worked around are put in the most refined wording, like the laughter of von Münchhausen in Immermann's novel, who, as we all know, eventually laughed in such a refined manner that no one could notice it anymore. In this way, these speeches of praise bring to mind modern holy pictures, where the most splendid aniline colors are used to augment picturesque drapery with magenta and azure; that the long blonde curls then appear in a most ornate arrangement need not be said.
>
> Not all heads are suitable as models for such paintings—where, for instance, would one find the comb to shape Jakob Steiner's wild hair into academic fashion?

Steiner gave five proofs of the isoperimetric theorem. Lovely as they are, he left one point open to attack: all proofs assume the existence of a solution (his strategy is always to take a figure that is not a circle and show that its area can be improved). This did not go unpunished. The analyst vultures can smell an existence assumption from miles away. To this day, many authors, revealing their sympathies, are eager to point out that existence is nontrivial. Perron [**30**] at least jokes about it:

Theorem. *Among all curves of a given length, the circle encloses the greatest area.*

Proof. For any curve that is not a circle, there is a method (given by Steiner) by which one finds a curve that encloses greater area. Therefore the circle has the greatest area. ■

Theorem. *Among all positive integers, the integer 1 is the largest.*

Proof. For any integer that is not 1, there is a method (to take the square) by which one finds a larger positive integer. Therefore 1 is the largest integer. ■

After having gone to a lot of trouble to rigorize Steiner's first proof with the help of some ugly calculations, Blaschke [**1**, p. 32] asks himself what Steiner, "who was not a friend of excessive politeness," would say about this. At best, says Blaschke, he would have quoted Faust:

> For thus your mind is trained and braced,
> In Spanish boots it will be laced,
> That on the road of thought maybe
> It henceforth creep more thoughtfully,
> . . . . . .
> Who would study and describe the living, starts
> By driving the spirit out of the parts:
> In the palm of his hand he holds all the sections,
> Lacks nothing, except the spirit's connections.
>
> (Goethe's *Faust*, Kaufmann's translation)

We could also suggest our old friend Dido's last words from the Purcell opera:

> Remember me, but ah! forget my fate.

Steiner's first article on the isoperimetric problem [**36**] clearly assumes the existence of a solution without any indication that this must be proved. I suppose that the criticism followed this publication. The following nonsense passage from his next article [**37**] should perhaps be seen as a halfhearted attempt to silence the critics.

> It is clear that there are, for a given perimeter, infinitely many figures of different form, which may also have different areas. Nevertheless, it is clear that the area, though it can be made arbitrarily small, cannot be made arbitrarily large, for one can always give a figure, proportional to the perimeter, that exceeds their area. Such a figure is for example the circle that has its midpoint on the perimeter and its radius equal to half the given perimeter. But when, for the same given perimeter, figures can have different areas, while these cannot be arbitrarily large, there must necessarily be either one figure that has greater area than all others or there must be several, differently shaped figures that have this property in common, i.e. that have equal area among themselves, but greater area than all the others.

That said, we will see that there are good reasons for considering Steiner's proofs to be essentially complete after all.

We will discuss Steiner's five proofs and try to show that they contain essentially three distinct ideas. All five proofs are presented in two articles published in 1842 [**37**], [**38**]. These articles are French translations. The German originals are published only in Steiner's collected works [**39**] (edited by Weierstrass, who complains in his preface that the French translations are not very good and that they introduce several errors that are not in the originals). The proof that is fifth in these articles had already been published in 1838 [**36**].

Steiner's four-hinge proofs.

Theorem. *Any figure with maximal area must be a circle.*

Proof. Take a figure with maximal area. Cut its perimeter in half with a line. This line will split the area in half as well, because if it did not we could take the half with the greater area together with its reflection in the line and get a figure with the same perimeter but greater area (Figure 11).

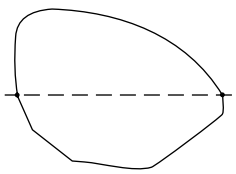


Figure 11.

Consider one of these halves. Suppose it is not a semicircle. Then there will be some point on the boundary where lines drawn from the points on the symmetry line meet at an angle that is not a right angle.

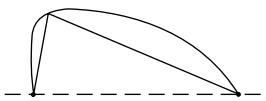


Figure 12.

Think of there being a void inside the triangle and think of the pieces on the sides as glued on. Slide the endpoints along the symmetry line to make the angle a right angle (Figure 13).

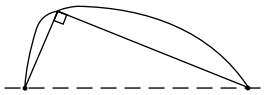


Figure 13.

This increases the area, so reflecting this gives a figure with greater area while the perimeter is still the same. That is impossible, so the halves must be semicircles and our figure must have been a circle to begin with. ■

If we draw the triangles together with their reflections then we get quadrilaterals, and we can then see the four hinges that have given the proof its name (Figure 14). Also, we could use reflection in the midpoint of the symmetry line instead, so that the quadrilateral becomes a parallelogram. Perhaps it is slightly more intuitive to argue that straightening out the parallelogram increases the area.

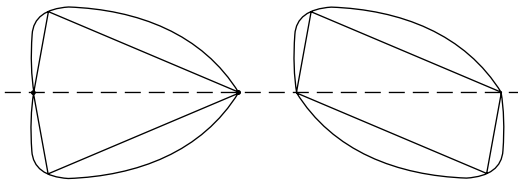


Figure 14.

Unlike Steiner, we cannot help being a bit curious as to whether this process converges. The answer is that it does, and to the circle, no less. This will be proved later.

The proof also works for showing that an optimal $2n$-gon must be regular. Suppose it is not. Then, when we cut the figure in half by drawing the line from vertex 1 to vertex $n + 1$, some vertex will not be on the semicircle from vertex 1 to vertex $n + 1$ and we can apply Steiner's method there and get a better, isoperimetric $2n$-gon.

We will not discuss the second and third proofs as they are based on essentially the same idea as the first and are not as beautiful. The second proof, for instance, takes the detour of showing that if we are given four sticks to make a quadrilateral, then, to maximize the area, we should put the vertices on a circle. Then, sure enough, the theorem follows, for if there were any four points on the optimal figure that were not on a circle, then we could improve on it (Figure 15).

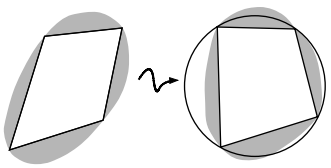


Figure 15.

Steiner's mean boundary proof.

Theorem. *Any figure with maximal area must be a circle.*

Proof. Given two curves, consider the mean curve—the curve that stays halfway between the two curves at all times (Figure 16).

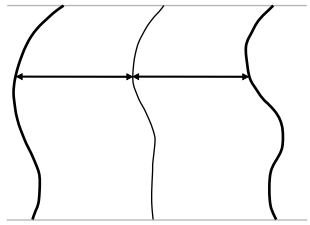


Figure 16.

The length of this curve will be less than the mean length of the two given curves (or equal if the two curves are the same), as we can see by slicing the curves into infinitesimals and confirming the claim piecewise (Figure 17).

Figure 17.

Take a figure with maximal area and cut its perimeter in half with a line.

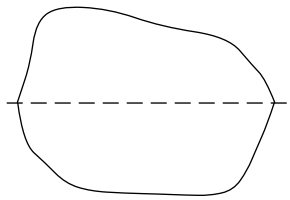


Figure 18.

As before, the line will split the area in half as well, because if it did not we could take the half with the greater area together with its reflection in the line and get a figure with the same perimeter but greater area. Suppose the two halves are not reflections of each other. Then reflect them to the same side and draw the mean curve (Figure 19).

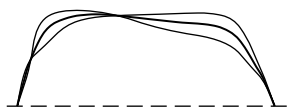


Figure 19.

This mean curve is shorter. But it encloses the same area, for it contains all the area that the two curves have in common and then half the additional area of the first curve and half the additional area of the second. However, these additional areas are just as large, ensuring that the mean curve actually encloses as much area as the the other curves.

So when we cut the perimeter of a maximal figure in half, the halves cannot be different in shape: if they were, then we could use this construction to get a figure with the same area and smaller perimeter. Thus the maximal figure cannot be anything other than a circle. ■

Steiner's snowball-packing proof. Grab a handful of snow. Put one hand on either side and compress the snow. Repeat this from all angles. You end up with a snowball. Why is it a ball? Because packing the given amount of snow tighter and tighter minimizes the surface area. We now do the same thing for plane figures, packing their area tighter and tighter, thus minimizing the perimeter.

Theorem. *Any figure with maximal area must be a circle.*

Proof. We begin with a convex figure and wish to modify it so that it becomes symmetric in a line. To do this, we think of the figure as consisting of vertical slices, and then we slide each of these slices so that half of each slice lies on either side of the line. If, for now, we consider only polygons, then the process is more down-to-earth, for we can determine the effect by examining what happens to the vertices:

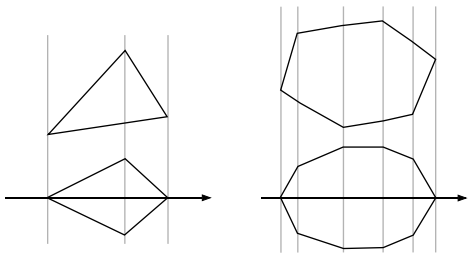


Figure 20.

The area is still the same, but as we see, triangles and trapezoids map to isosceles ones, and these, as we know, cover area more efficiently.

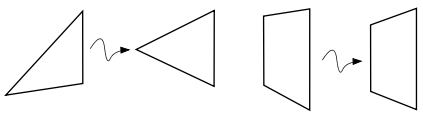


Figure 21.

For polygons then, the perimeter has decreased unless all triangles and trapezoids are already isosceles, which occurs when the original figure is symmetric in the line (up to a translation). Applying this to infinitesimals, we are persuaded that the same is true for any convex figure. Thus a solution must be symmetric in every direction, for otherwise we could improve on it. We feel that such a figure must be a circle. We can see this a bit more rigorously as follows. Take such a figure. It will be unaffected (up to translation) if we make it symmetric in the x-axis and then in the y-axis. But now it must be symmetric in every line through the origin. Surely it must be a circle. Pedantically, consider a point inside the figure and reflect it in all lines through the origin to show that all other points at the same distance from the origin must also be inside the figure. ■

For the question of convergence, and for the mathematical modelling of snowball-packing, it is interesting to start with an arbitrary figure and see if repeated symmetrizations make it into a circle (Figure 22).

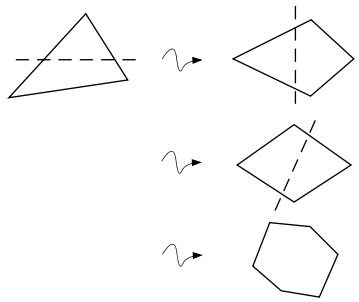


Figure 22.

Quite obviously, this process does in fact converge to a circle, and Steiner pretty much says so, even though convergence is not in his vocabulary. This is the most telling passage [**36**, p. 285]:

> [T]he difference between the smallest and the largest diameter decreases, for by the process, when the new axis is chosen (as long as it is not parallel to the prior ones), the largest diameter will be made smaller and the smallest will be made larger, as is easy to see. By appropriate choices of the new axes, the diameters can be brought closer to equality faster.

It is also implicit in what follows that Steiner considers it obvious that such cleverly chosen symmetrizations will always give improvements that are substantial, in the sense that the process will not halt prematurely before it reaches the circle. Thus Steiner knew—though he could perhaps have argued more convincingly, were he not such a principled anti-analyst—that we can begin with any figure, make it better and better and, in the limit, end up with a circle. This, of course, would prove the isoperimetric theorem without any assumption of existence.

4. THE CALCULUS OF VARIATIONS. The solution of the isoperimetric problem by means of the calculus of variations was the first proof unhampered by Euclidean sterility, and we will try to present it as such, following Euler's account from 1744 [**11**]. This stands in contrast to the modern theory of the calculus of variations, which has long been saturated with rigor and analytic trickery.

The question of the existence of a solution also belongs here. Weierstrass proved existence in his lectures on the calculus of variations in 1879, and with that the first rigorous solution of the isoperimetric problem was completed. Weierstrass never published these results. We must rely on the volume on the calculus of variations in his collected works [**42**], which is reconstructed from lecture notes of his students.

This is what Weierstrass has to say about Steiner's proofs [**42**, pp. 259–260]:

> A detailed discussion of this [the isoperimetric] problem is desirable, since Steiner was of the opinion that the methods of the calculus of variations were not sufficient to give a complete proof. [Weierstrass notes what Steiner has proved.] But the calculus of variations is in a position to prove all this, as we will show later; furthermore it can show what Steiner could not—that such a maximum really exists. In many cases this can be proved by geometric means as well, but when, for example, asked for the curve that for a given perimeter encloses the greatest area (with no further conditions), Steiner draws a line that cuts the perimeter in half and shows by symmetrizing the figure around this line that a curve of greater area is created if this line is not a symmetry axis; now it is clear that only for the circle do such lines have such properties, but this does not prove that there is an actual maximum, and not just an upper bound.

Theory. Here is the basic idea. Take a curve and wiggle it a little bit, while keeping its perimeter fixed. If the curve is the one with maximal area then we are at an optimum, so an infinitesimal wiggle will cause zero change in the area. In order to find the optimal figure, therefore, we calculate the change in area caused by an infinitesimal wiggle and set this equal to zero. This leads to a differential equation that must be satisfied by an optimal figure, and indeed that is satisfied only by circles.

Forget about isoperimetry for a while and consider the simpler problem of finding a function $y(t)$ that extremizes an integral that depends on it, say

$$\int F(y, \dot{y}, t)\, dt.$$

(The archetypal example is the brachistochrone problem—the problem of finding the curve of quickest descent from one point to another—where $y(t)$ is the curve and the integral is an expression with gravity and stuff in it that says how long it takes for a ball to roll down.) Split the t-axis into Δt-segments, and let $y(t)$ be linear on these.

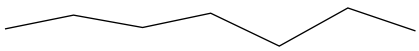


Figure 23.

The functions $y(t)$ and $\dot{y}(t)$ are then determined by the value of $y(t)$ at the break points, call them $y_1, y_2, y_3, \ldots$. Suppose that $y(t)$ is the optimal curve. Grab one of these points y_k, pull it up and down, and try to feel it click at the optimum. This occurs when the rate of change in the integral caused by the change in y_k is zero (i.e., when an infinitesimal change in y_k results in a zero change in the integral).

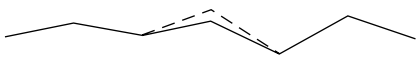


Figure 24.

Say we increase y_k by dy_k. What happens to the integral (which is now really a sum)? First, there is the direct change caused by the change in y_k, namely,

$$dy_k \left(\frac{\partial F}{\partial y_k} \right).$$

Then there are the changes caused by the changes in the derivatives on the sides of y_k, call them $\dot{y}_k$ and $\dot{y}_{k+1}$. When y_k changes by dy_k, $\dot{y}_k$ changes by $dy_k/\Delta t$, so it causes the change in the integral

$$dy_k \left(\frac{1}{\Delta t} \frac{\partial F}{\partial \dot{y}_k} \right).$$

Similarly, $\dot{y}_{k+1}$ changes by $-(dy_k/\Delta t)$ and causes the change

$$-dy_{k+1} \left(\frac{1}{\Delta t} \frac{\partial F}{\partial \dot{y}_{k+1}} \right).$$

We infer that the equation for the total change to be zero is

$$\frac{\partial F}{\partial y_k} - \frac{1}{\Delta t} \left(\frac{\partial F}{\partial \dot{y}_{k+1}} - \frac{\partial F}{\partial \dot{y}_k} \right) = 0.$$

Taking the limit and applying this everywhere (for all t) we get the Euler equation[2]

$$\frac{\partial F}{\partial y} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{y}}\right) = 0.$$

We look for solutions by solving this differential equation.

To deal with the isoperimetric problem we need a variation of this idea. We are looking for the curves $t \mapsto (x(t), y(t))$ that maximize the area

$$\int A(x, \dot{x}, y, \dot{y}, t)\, dt = \frac{1}{2}\int x\dot{y}\, dt - y\dot{x}\, dt$$

while the perimeter

$$\int L(x, \dot{x}, y, \dot{y}, t)\, dt = \int \sqrt{\dot{x}^2 + \dot{y}^2}\, dt$$

is kept fixed. As before, a solution must be such that if we wiggle it infinitesimally while keeping the perimeter fixed (that is, the change in the perimeter integral is zero), then because we are at an optimum the change in the area integral is also zero. We will consider variations in $x(t)$ and $y(t)$ separately; of course, the two are analogous. Accordingly, we should make an infinitesimal change in $y(t)$ that leaves the perimeter the same. Our old procedure of changing just one point will not do. Instead, we change two points, y_k and y_{k+1} (Figure 25).

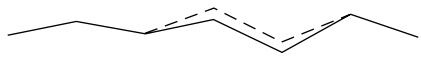


Figure 25.

The equation for the change in the perimeter to be zero is then

$$dy_k\left(\frac{\partial L}{\partial y_k} - \frac{1}{\Delta t}\left(\frac{\partial L}{\partial \dot{y}_{k+1}} - \frac{\partial L}{\partial \dot{y}_k}\right)\right) + dy_{k+1}\left(\frac{\partial L}{\partial y_{k+1}} - \frac{1}{\Delta t}\left(\frac{\partial L}{\partial \dot{y}_{k+2}} - \frac{\partial L}{\partial \dot{y}_{k+2}}\right)\right) = 0.$$

At the same time, the change in the area should also be zero

$$dy_k\left(\frac{\partial A}{\partial y_k} - \frac{1}{\Delta t}\left(\frac{\partial A}{\partial \dot{y}_{k+1}} - \frac{\partial A}{\partial \dot{y}_k}\right)\right) + dy_{k+1}\left(\frac{\partial A}{\partial y_{k+1}} - \frac{1}{\Delta t}\left(\frac{\partial A}{\partial \dot{y}_{k+2}} - \frac{\partial A}{\partial \dot{y}_{k+2}}\right)\right) = 0.$$

Let's write these last equations a bit more compactly as

$$dy_k \Delta L_k + dy_{k+1}\Delta L_{k+1} = 0, \qquad dy_k \Delta A_k + dy_{k+1}\Delta A_{k+1} = 0.$$

These equations capture the essence of a solution, but we will go a bit further, using a trick to combine them into one simple equation. Suppose that we have a solution. In this case, let λ be the number by which one has to multiply ΔL_k to get ΔA_k. Now subtract λ times the first equation from the second,

$$dy_k\,(\Delta A_k - \lambda \Delta L_k) + dy_{k+1}\,(\Delta A_{k+1} - \lambda \Delta L_{k+1}) = 0.$$

[2] Also called the Euler-Lagrange equation. Lagrange [**25**] in 1762 used a dull analytic approach that became fashionable with the nineteenth century wave of rigor and has kept its lead over Euler's approach ever since.

Since the first term is zero, the second term must also be zero. Applying this for all values of k, we see that for a solution we must have $\Delta A_i - \lambda \Delta L_i = 0$, with the same λ for all i. Taking the limit, we arrive at the differential equation that describes the solution:

$$\frac{\partial(A - \lambda L)}{\partial y} - \frac{d}{dt}\left(\frac{\partial(A - \lambda L)}{\partial \dot{y}}\right) = 0.$$

Similarly, we obtain an equation satisfied by $x(t)$:

$$\frac{\partial(A - \lambda L)}{\partial x} - \frac{d}{dt}\left(\frac{\partial(A - \lambda L)}{\partial \dot{x}}\right) = 0.$$

But these are the equations we would have arrived at if we were looking for unconstrained extrema of $\int A - \lambda L$. In other words, the problem of extremizing $\int A$ while keeping $\int L$ fixed is the same as that of extremizing $\int A - \lambda L$ without side conditions, for some λ. This is what is called "Euler's rule."

In our investigation of the isoperimetric problem, we could not help solving the more general problem of extremizing pretty much any old integral while another is kept fixed, so perhaps we should forgive the calculus of variations people for calling any such problem an "isoperimetric problem."

Calculations. We now apply the rules we have found to the isoperimetric problem.

Theorem. *Any figure with maximal area must be a circle.*

Proof. To maximize the area

$$\frac{1}{2}\int x\dot{y}\,dt - y\dot{x}\,dt$$

while the perimeter

$$\int \sqrt{\dot{x}^2 + \dot{y}^2}\,dt$$

is kept fixed, we should form the function

$$F = \frac{1}{2}(x\dot{y}\,dt - y\dot{x}\,dt) - \lambda\sqrt{\dot{x}^2 + \dot{y}^2}$$

and look for unconstrained extrema of

$$\int F\,dt.$$

We will find the extrema by solving the Euler equations

$$\frac{\partial F}{\partial x} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{x}}\right) = 0, \qquad \frac{\partial F}{\partial y} - \frac{d}{dt}\left(\frac{\partial F}{\partial \dot{y}}\right) = 0.$$

These are

$$\frac{1}{2}\dot{y} - \frac{d}{dt}\left(-\frac{1}{2}y - \lambda\frac{\dot{x}}{\sqrt{\dot{x}^2+\dot{y}^2}}\right) = 0,$$
$$-\frac{1}{2}\dot{x} - \frac{d}{dt}\left(\frac{1}{2}x - \lambda\frac{\dot{y}}{\sqrt{\dot{x}^2+\dot{y}^2}}\right) = 0.$$

If t is taken to be an arclength parameter, then these equations simplify to

$$\dot{y} + \lambda\ddot{x} = 0, \qquad -\dot{x} + \lambda\ddot{y} = 0,$$

and their solutions are

$$x = x_0 + \lambda\cos\frac{t-t_0}{\lambda}, \qquad y = y_0 + \lambda\sin\frac{t-t_0}{\lambda}.$$

These are plainly parametric equations for a circle. ∎

Following Weierstrass [**42**, pp. 70–75], we also give a proof along these lines of the isoperimetric theorem for polygons. For this, we need only the discrete versions of the foregoing theory, and then we are supposed to call it not the calculus of variations but the method of *Lagrange multipliers*. Of course, the ideas involved are the same.

Theorem. *Among all n-gons with perimeter L, the regular one has the greatest area.*

Proof. We should maximize the area

$$A = \frac{1}{2}\sum_{k=1}^{n-1} x_k y_{k+1} - y_k x_{k+1},$$

or, for convenience, twice the area, while fixing the perimeter

$$L = \sum_{k=1}^{n-1}\sqrt{(x_k - x_{k-1})^2 + (y_k - y_{k-1})^2} = \sum_{k=1}^{n} l_k.$$

The variational argument now gives,[3] for each k,

$$y_{k+1} - y_{k-1} + \lambda\left(\frac{x_k - x_{k+1}}{l_{k+1}} + \frac{x_k - x_{k-1}}{l_k}\right) = 0,$$
$$-x_{k+1} + x_{k-1} + \lambda\left(\frac{y_k - y_{k+1}}{l_{k+1}} + \frac{y_k - y_{k-1}}{l_k}\right) = 0.$$

Anticipating some messy calculations, we introduce the clever notation of writing z_k for the complex number pointing from (x_{k-1}, y_{k-1}) to (x_k, y_k),

$$z_k = x_k - x_{k-1} + i(y_k - y_{k-1}).$$

[3]Textbooks give us rules to remember these equations. We could say that we are looking for unconstrained extrema of $F = 2A - \lambda(\sum_{k=1}^{n} l_k - L)$, so the equations express the fact that the partial derivatives of F vanish. Or we could say that the gradient vectors of A and $\sum_{k=1}^{n} l_k - L$ should be parallel.

Then the equations are seen to be, respectively, the imaginary and (-1 times) the real part of

$$z_k + z_{k+1} + \lambda i \left(\frac{z_k}{l_k} - \frac{z_{k+1}}{l_{k+1}} \right) = 0.$$

Since our new notation gives the neat expression $l_k = z_k \overline{z_k}$ for the side lengths, we rewrite this as

$$z_k \left(1 + \frac{\lambda i}{l_k} \right) = -z_{k+1} \left(1 - \frac{\lambda i}{l_{k+1}} \right)$$

and multiply each side by its conjugate to get

$$l_k^2 \left(1 + \frac{\lambda^2}{l_k^2} \right) = l_{k+1}^2 \left(1 + \frac{\lambda^2}{l_{k+1}^2} \right).$$

It follows that $l_k = l_{k+1}$ (i.e., the polygon is equilateral). Now we again look at

$$z_k \left(1 + \frac{\lambda i}{l_k} \right) = -z_{k+1} \left(1 - \frac{\lambda i}{l_{k+1}} \right)$$

and see that z_{k+1}/z_k, which is just $e^{i(\arg z_{k+1} - \arg z_k)}$ now that $|z_{k+1}| = |z_k|$, is the same for all k. This means that the polygon is equiangular as well. ■

Existence for polygons. Before we plunge into the general theory dealing with existence, we note that it is easy to show existence if we restrict ourselves to n-gons. To convey some of the historical flavor, we sketch how Weierstrass proves existence for n-gons (although he does not use this approach for the general existence theorem).

Theorem. *Among all n-gons of perimeter L there is at least one with maximal area.*

Proof. All the n-gons $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ could be put inside a square—we could put the first vertex at, say, $(0, 0)$—so the areas are bounded and the coordinates of the vertices of the n-gon are bounded by the coordinates of the square. Whether the maximal area is attained or not, there is an accumulation n-gon $(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots, (x_n^*, y_n^*)$ around which the neighboring n-gons have areas arbitrarily close to the maximal area. (By focusing on the vertices of a polygon we can think of it as a point of $\mathbb{R}^{2n}$.) But this accumulation n-gon is clearly itself one of the L-perimeter n-gons and it must have the maximal area, or else the continuous area function could not take values arbitrarily close to the maximal area in its vicinity. ■

We might have phrased the proof differently—say, in terms of sequences, repeatedly extracting subsequences in order to make the vertices converge one by one. Or call it compactness, if you must. The idea is still the same.

This theorem suffices to supply what is missing in, for instance, Steiner's four-hinge proof. Take any convex figure and approximate it by a sequence of n-gons, all of the same perimeter, and let $n \to \infty$. The sequence of regular n-gons of the same perimeter exhibits greater areas term by term, and thus at least as large a limit area. We conclude

that the arbitrary figure we started with cannot have greater area than the circle with the same perimeter.

Weierstrass's sufficient condition. We can think of the calculus of variations as a generalization of ordinary calculus. From this perspective the Euler equation corresponds to the rule that the derivative should be zero at an extremum—both rules say that an infinitesimal wiggle causes a zero change if we are at an extremum. Both are just necessary conditions; we prove them by assuming that we were dealing with an optimum and then deduce that it must have this property. We could consider the analogue of taking the second derivative—we call this the second variation. But, again analogously, we should not expect this to settle the matter. We are still looking for sufficient conditions. These local investigations will not do.

Weierstrass says that the general proof of existence "was considered to be so difficult that it was almost thought that it could not be given by means of the calculus of variations." It is difficult indeed. Weierstrass's proof [**42**, pp. 257–264, 301–302] relies on his general theory, and it is too involved for us to discuss in full. Hilbert [**19**, sec. 23] simplified the theory by introducing the idea of an "invariant integral," and it is this idea that we take advantage of here. But it is still hard to grasp intuitively how this abstract theory works even when restricted to the isoperimetric problem. All we can do in this article is to try to get a feel for the basic nature of the theory.

Suppose that we wish to find the shortest path between two points. The Euler equation tells us that the solution can be nothing but a straight line, and we wish to show that this really is a solution. The first step of the general theory is to get a field of extremals, that is, we should cover the area around our supposed minimum curve with curves that satisfy the Euler equation, one through every point. In our case, these will be lines parallel to our first line.

In our example, we can proceed as follows. Take any curve joining the two points. Its length is given by summing the arclength elements along the curve: $\int ds$. Now consider the integral $\int \cos\theta \, ds$ that counts how much of the arclength goes forward. (Thus θ is the angle the curve makes with the extremal of our field of extremals that passes through the point under construction.) This integral does not depend on our choice of curve; it always gives the length of the line between the two points. There is an integral with corresponding properties in the general theory. We might as well state it: in the case of extremizing $\int F(y, \dot{y}, t)\, dt$ it is

$$\int F(y, p, t) + (\dot{y} - p)F_{\dot{y}}(y, p, t)\, dt,$$

where $p(x, y)$ is the slope of the extremal passing through (x, y). We would need a much longer discussion to understand it, but it does in fact have the two key properties of our example $\int \cos\theta \, ds$: it is independent of the path (this is quite difficult to show), and along the supposed extremum curve y^* we have of course $\dot{y}^* = p$, so the value of the integral is the supposed extremum value $\int F(y^*, \dot{y}^*, t)\, dt$.

It is now easy to conclude our example. The difference in length between an arbitrary curve and the line segment is $\int 1 - \cos\theta \, ds$, where the integral is taken along the curve. This integral is always greater than zero, unless the curve and the line are the same, so that θ is always zero. Similarly, in the general case, the difference in $\int F$ between an arbitrary curve in our field of possible extremals and the supposed extremal curve is given by, again integrating along the arbitrary curve,

$$\int F(y, \dot{y}, t) - F(y, p, t) - (\dot{y} - p)F_{\dot{y}}(y, p, t)\, dt.$$

The integrand here is the Weierstrass excess function ($\mathcal{E}$-function, E-function) that he used to formulate his sufficient condition: if the excess function is negative everywhere, then the curve is truly a maximum.

Existence by compactness. We saw earlier that one could prove existence for n-gons essentially by compactness, because a sequence of L-perimeter n-gons cannot converge to a non-L-perimeter non-n-gon. But surely a sequence of L-perimeter figures could not converge to a non-L-perimeter figure. So the L-perimeter figures are just as compact. There should be a compactness proof for them, too. Indeed there is. Spivak [**33**, pp. 441–444], for example, does this from the modern standpoint, employing the more general construction of beginning with any compact metric space, say a square in the plane, and making a new space out of all its closed subsets using the so-called Hausdorff metric, which in the plane amounts to measuring the distance between two closed sets by determining how thick a strip would have to be put around each in order to encompass the other. Then one goes on to prove that this new space is compact as well. The standard proof of this is not hard, but it does not shed any light on the isoperimetric problem. Anyway, once that is done the rest is easy. The convex sets form a closed subset, and the perimeter function, acting on this set, is quite clearly continuous with respect to this Hausdorff metric, so the inverse image of a fixed number under the perimeter function is a closed subset of a closed set in a compact space, which ensures that the continuous area function attains its maximum there.

If we don't like hiding behind an abstract theorem like that, then we could try to mimic the polygon proof. It would go something like this. Whether the maximal area is attained or not, there will be a sequence of isoperimetric figures whose areas converge to the maximal area. Take such a sequence and parametrize the boundary curves by arclength, say with the starting point, the point for $s = 0$, at the origin. Now consider the sequence of points halfway round the corresponding curves, points where $s = L/2$, and extract a subsequence from the sequence of curves for which the sequence of those points converges. In the next step, split the curves into fourths instead, and extract a subsequence for which the sequence of those points converges. Then split into eighths, and so on. In each of these steps we know that the points on the curve for $s = L/2^n, 2L/2^n, \ldots, (2^n - 1)L/2^n$ converge, so we can trap these points in small discs, say of radius one over n^{n^n}, around the limit points. This pins down the curve so that it cannot reach outside a $(L/2^n)$-strip around the vertices. In the next step, we pin down the curve even further, forcing it to remain inside a strip half as thick as, and contained inside, the previous one. Continuing in this way we trap the curve inside an arbitrarily thin strip, and the curves converge uniformly to some curve that obviously must have the same perimeter, and an area that is the limit of the areas.

5. POST-STEINER GEOMETRY. We will now see that we can handle existence by straightforward geometry just as well as by abstract theories. After this section there will be no more nagging questions about existence.

Edler's finite existence proof. Without assuming existence, Edler [**10**] proved in 1882 that a figure that is not a circle of the specified perimeter has smaller area than the circle. He reasoned as follows. Apply Steiner symmetrization (i.e., snowball-packing) to get a figure with greater area. Then inscribe a polygon that approximates the area closely enough, so that this polygon, too, has greater area than the original figure. Then replace this polygon with a regular polygon with the same perimeter but greater area. This is the tricky step, but we will soon prove that it can be done using Steiner symmetrization. But regular polygons have smaller area than the circle with the same

perimeter. We can schematize the reasoning very loosely as follows:

$$\begin{aligned} \text{anything} &< \text{something} \\ &< \text{some polygon} \\ &< \text{some regular polygon} \\ &< \text{the circle} \end{aligned}$$

Theorem. *Given any polygon, we can create a regular polygon with greater or equal area and smaller or equal perimeter.*

Proof. We begin like Steiner to make the polygon symmetric in the x- and y-axes. Now look at the triangle in Figure 26:

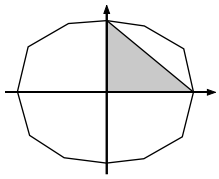


Figure 26.

Grab the hypothenuse, thinking of it as a stick with its endpoints moveable along the axes, and tilt it to make the triangle isosceles. This increases the area, unless the triangle was already isosceles. Next look at the piece attached to the hypothenuse and symmetrize it in a new axis that bisects the angle between the previous ones (Figure 27).

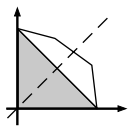


Figure 27.

Typically, this means that the top vertex ends up on the (new) axis:

Figure 28.

In some cases there is no unique top vertex:

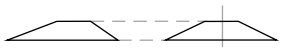


Figure 29.

These cases serve to show that if we have k vertices above the base before the symmetrization, then we will have at most $k-1$ vertices on either side of the axis after the symmetrization. This is so because each vertex can generate at most one vertex on either side. But it cannot be that all vertices do this, for either there is a top vertex or there are two vertices on the same horizontal line. In our example, the quarter-figure we started with had two vertices not on an axis, while our new one-eighth figure has only one (Figure 30).

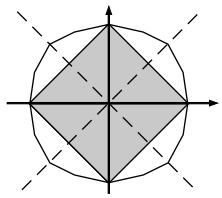


Figure 30.

We repeat this procedure, making triangles isosceles and inserting new axes between the old ones. After a finite number of steps there will be no vertices that are not on axes. If we then make the triangles isosceles we get a regular polygon, and we are done. ■

Carathéodory's convergence proof. This is a proof of Carathéodory from 1909 [**7**]. As we have discussed, showing convergence in Steiner's proofs is enough to make them perfectly rigorous.

Theorem. *The method of Steiner's first proof converges to the circle.*

Proof. As in Steiner's proof, we work with half-figures, curves of length L that start and end on a given line. For such a curve, take the convex hull and rescale it so that it has length L. Then apply Steiner's improvement procedure. Repeat these two steps and always keep the left endpoint fixed. We claim that this process converges uniformly to a semicircle of length L.

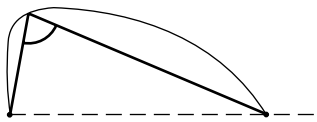


Figure 31.

When we change the angle indicated in Figure 31 to a right angle, then the area is increased. The increase is equal to

$$\frac{|\text{first side}| \cdot |\text{second side}|}{2} \left(1 - \sin(\text{the angle})\right).$$

This quantity varies continuously along the curve, so it attains its maximum, and we can agree always to take the first point where the maximum is attained as the point at which we apply our Steiner improvement.

Now consider the sequence of right endpoints. These stay bounded, so they have at least one limit point. Extract a subsequence of curves for which the right endpoints converge. Consider an epsilon-thin strip around the semicircle that begins at the left point and ends at this right limit point. Figures constructed by this method will be forced, after a finite number of steps, to stay completely inside this strip, because the improvement quantity for all possible triangles with right point in the strip and top point outside this strip has some nonzero lower bound. Were the top points of figures in our sequence to remain outside the strip, then the increases would be substantial and the areas would tend to infinity, which of course they cannot do. The semicircle that we are approximating uniformly with curves of length L must be the semicircle of length L, and the limit point of the right endpoints can be none other than the right endpoint of this semicircle. ■

Study's convergence proof. This is a proof of Study from 1909 [**7**]. Again, it uses a limiting argument to get rid of the existence assumption.

Theorem. *All areas of perimeter L are bounded by that of the circle with perimeter L.*

Proof. From any figure, we create a sequence of isoperimetric figures with increasing areas that converges to the circle. We begin with a figure, that, as usual, we make convex and symmetric with respect to the x- and y-axes, leaving the area at least as large. The idea is to introduce further symmetry axes. Consider the part of the figure in the first quadrant. The line that goes through the origin at angle of $\pi/4$ with the x-axis does not, in general, cut this perimeter in half. Translate it then, so that it does, in fact, cut the perimeter in half (Figure 32). Consider the two pieces that result:

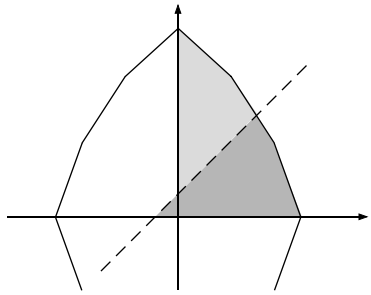


Figure 32.

Both of these pieces could be used to make a figure with the same perimeter—just put the pointy end at the origin and reflect all the way around. So we take the piece with greater area and do precisely this. Clearly, we have increased the area, or possibly left it the same, and kept the perimeter fixed. We now have four symmetry axes. Continuing in the same way poses no difficulties, giving 2^{n+1} symmetry axes at the nth step.

Now to show convergence. Consider the points where the nth figure intersects the symmetry axes. These lie on two circles (Figure 33):

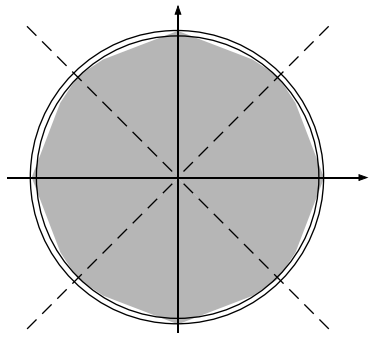


Figure 33.

These circles remain within some closed disc. From the sequence of figures extract a subsequence for which the sequences of these circles converge. They must converge to the same limit circle, for the distance between the circles at stage n is bounded by the length of the slices of the figure's perimeter, which is equal to 2^{-n-2} times the

perimeter. But when the circles converge to the same limit circle, the figures converge uniformly to this circle as well. Therefore the limit circle can be none other than the circle with the same perimeter as the figures. And so the increasing sequence of areas converges to what must be their upper bound, the area of the circle. ■

6. VECTORS.

Gromov's vector analysis proof. The paper [**21**] contains a nice two-dimensional adaptation of the n-dimensional proof of Gromov from 1986 [**16**], which we now look at. To set it up we consider a flaky physical-intuitive argument. Where the earth is perfectly spherical, we can balance a stick by placing it perpendicular to the ground. Where the earth is not spherical, say on the side of a hill, placing a stick perpendicular to the ground will cause it to tip over. Let's agree that this experience convinces us that, if we were stranded on a nonspherical planet, we could always find spots where putting a stick perpendicular to the ground would cause it to tip over. Thus it is only for the sphere that gravity always acts in the direction of the normal to the surface. Let's also agree that this still holds when the universe is flat—when planets are plane figures.

To capture the mathematics of gravity, we should think of this in terms of vector fields, and to make it easier for us we consider the negative of the gravitational field—just take the ordinary gravitational field and multiply the vectors by -1, pretending that we are in a dual universe where gravity pushes rather than pulls. Now take all figures with a given area and fill them with cement. Then they all produce the same amount of negative gravity. This negative gravity has flow lines out of the figure, but only for the circle do they always flow out along the normal. For any other figure, the lines flow out askew, which we feel is an inefficient use of perimeter. So, perhaps, this will force the perimeter to be greater than that of the circle of the same area.

How do we exploit these ideas to make a proof? Well, the amount of negative gravity being produced can be calculated by summing over the boundary ∂D of the figure D the outward flow $\partial F/\partial \vec{\nu}$ along the normal:

$$\int_{\partial D} \frac{\partial F}{\partial \vec{\nu}}\, ds.$$

We wish to show that a noncircular figure spreads the outflow over a greater perimeter. This would follow at once if we could show that the flow along the normal from a point on a circle is greater than the flow along the normal from any point on the boundary of any figure of the same area. That would mean that not only does a noncircular figure dilute the outflow at some places but also that it cannot concentrate the flow somewhere else. Indeed, this is the case, as we are about to show by calculation.

Fix a point on the boundary and its normal (Figure 34):

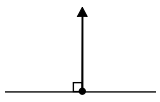


Figure 34.

We now have a given area to distribute under this point to make the negative gravitational flow in the direction of the normal as large as possible. What will an arbitrary infinitesimal square (Figure 35) contribute to the flow if we choose to include it?

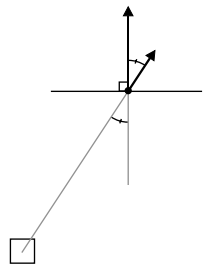


Figure 35.

The force the square exerts on the point is proportional to the area of the square divided by its distance from the point.[4] Then the angle θ the force vector makes with the normal determines the part of the force that acts in the direction of the normal. That part is the magnitude of the force times $\cos\theta$.

Suppose that we have found the optimal shape of the area. Walk straight back to the last infinitesimal square included in that direction, and do the same thing for an angle θ (Figure 36).

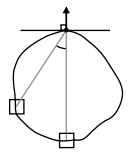


Figure 36.

These two squares contribute equally to the negative gravitational flow in the direction of the normal. This is so because if one of them contributed more, then the figure wouldn't be optimal (for then we would have included more area in that direction). Say that the straight-back square is at a distance r_0, while the other square is at distance r_θ. For their contributions to be equal, it must be the case that

$$\frac{\cos\theta}{r_\theta} = \frac{1}{r_0},$$

so

$$r_\theta = r_0 \cos\theta.$$

This means that the figure is a circle, because it says that if we break the r_0-line in half and tip the bottom half towards the r_θ-line, then they meet tip-to-tip (Figure 37). Solve these triangles if you must:

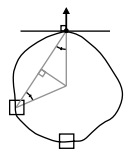


Figure 37.

[4]The pseudo-explanation is as follows. In three-dimensional space, gravity is inversely proportional to the square of the distance because, at a given distance from the source, the gravitational force distributes itself across the surface of a sphere. In two-dimensional space it is distributed only across a circle, so it is inversely proportional to the distance.

That's it, we're done. Since we got so carried away that we forgot to put the proof in a proper proof environment, we summarize it here.

Theorem. *Of all figures with a given area, the circle has the shortest perimeter.*

Proof. Consider the collection of all figures D with a given area. Fill each with negative cement. Then they all generate the same amount of negative gravitational flow out through the boundary ∂D of D (i.e., the integral

$$\int_{\partial D} \frac{\partial F}{\partial \vec{v}}\, ds$$

is the same for all of them). But the integrand is uniquely maximized by the circle, so all other figures must have greater perimeter. ∎

Schmidt's projection proof. This is a two-dimensional version of the n-dimensional proof of Schmidt from 1939 [**32**]. The following is a physical analogy vaguely connected with this proof. Tape pins on a balloon so that they point perpendicularly outward from the surface. When the balloon is not inflated it is all wobbly, and the pins point in whatever direction they please. But when we inflate it, then the pins all point away from the center of the balloon. Similarly, it is because the circle is so packed with area that its normals are forced to point away from its midpoint.

Theorem. $L^2 - 4\pi A \geq 0$*, with equality only for the circle.*

Proof. Begin with an arbitrary figure and parameterize the boundary by arclength $t \mapsto (x(t), y(t))$ (i.e., use a unit speed parameterization, one with $\dot{x}^2 + \dot{y}^2 = 1$). Construct a circle that is as wide as the figure and project the boundary vertically onto it, as in Figure 38:

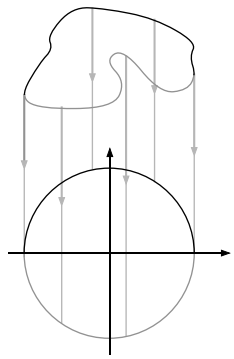


Figure 38.

Under this projection the x-coordinate is kept fixed and the y-coordinate is sent to, say, $\hat{y}$, in a way that puts $(x, \hat{y})$ on the circle. As (x, y) traces out our figure, the vector from O to $(x, \hat{y})$ points to the projection of (x, y) on the circle. The key insight is now that if the figure is a circle, then its normal is always parallel to $(x, \hat{y})$, but if it is not a circle, then its normal is not always parallel to $(x, \hat{y})$. In other words, the circle maximizes the projection of $(x, \hat{y})$ onto the normal $(\dot{y}, -\dot{x})$. Aha, a maximization property of the

circle! Looks like we're on to something. To formalize this, we take the inner product of $(x, \hat{y})$ and the normal $(\dot{y}, -\dot{x})$. This is less than the product of the lengths

$$x\dot{y} - \hat{y}\dot{x} \leq r,$$

and our geometric intuition tells us that equality occurs pointwise only for the circle. Let's see what we can do with that. We have

$$A = \int_0^L x\dot{y}\, dt \leq \int_0^L r + \hat{y}\dot{x}\, dt.$$

The integral in the first term is of course just Lr, and the second integral is the negative area of the circle. Now we are just one ad hoc factorization away from the isoperimetric inequality:

$$A \leq Lr - \pi r^2 = \frac{L^2}{4\pi} - \frac{\pi}{4}\left(\frac{L}{\pi} - 2r\right)^2 \leq \frac{L^2}{4\pi}. \qquad \blacksquare$$

7. DISSECTION.

Lawlor's dissection proof. If you are getting tired of physical analogies, then skip ahead, because this will be our most far-fetched one yet. If not, imagine a pizza. The isoperimetric theorem says that the circular shape of pizzas maximizes the topping-to-crust ratio. We bake a noncircular pizza with the same amount of crust as a round one, and now we wish to show that less topping fits on this one. A naive attempt would be to slice the round pizza into slices in the customary manner, although perhaps thinner, and then to arrange these slices so that their crusts cover the crust of the noncircular pizza. The isoperimetric theorem would follow if the interior of the noncircular pizza were always completely covered in the process. This will obviously not be so, but the idea is not beyond salvation. We will sense its spirit in the following proof of Lawlor from 1998 [**26**].

We restrict ourselves to polygons, by approximation, and also to quarter figures. This is alright because, given any figure, we can easily create a figure with the same perimeter and at least as large an area that is symmetric with respect to the x- and y-axes. Specifically, draw a horizontal line that cuts the perimeter in half. Take as the new figure the half with the greater area together with its reflection. Next, draw a vertical line that cuts the new perimeter in half. Again, take as the new figure the half with the greater area together with its reflection.

Theorem. *Among all fixed-length, equilateral n-gon arcs in the first quadrant, the regular one encloses the greatest area.*

Proof. We cover our figure with triangles. We triangulate the regular n-gon arc by drawing rays from the vertices to the origin:

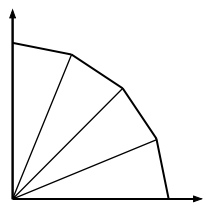


Figure 39.

For the general case, we draw the same rays, only translated to start at the vertices.

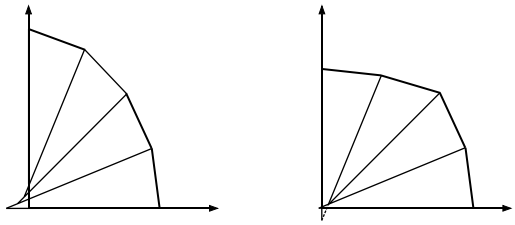


Figure 40.

This gives us triangles that cover our figure, perhaps with excess and overlap, as shown in Figure 40. The picture contains the idea. Formally, convexity ensures that no ray can sneak under the preceding ray. As for covering, a point in the interior will be to the left of the arc and between the first and last rays (those that are parallel to the axes), and thus it will be trapped between two consecutive rays and covered by the associated triangle.

By construction, all these triangles have the same base and the same opposite angle, and among such triangles the isosceles triangle has the greatest area. So the regular n-gon is greater piece by piece, and thus greater as a whole. ■

8. SERIES. The series approaches that follow are not very enlightening, but they do suggest a systematic scheme for attacking inequalities:

Generic inequality. *This + that* ≥ 0*, with equality when so-and-so.*

Generic proof. Express the terms analytically and expand in series. Prove the inequality by comparing coefficients—the point is that this is now essentially a matter of arithmetic. Force equality and see what that means for the series. ■

Hurwitz [**22**], in the opening paragraph of the article containing his Fourier analysis proof, also tries to explain to us why a series approach is natural:

> Fourier series and analogous expansions intervene quite naturally in the general theory of curves and surfaces. Indeed, this theory, conceived from the point of view of analysis, obviously deals with the study of arbitrary functions. I was thus led to apply Fourier series to some geometric problems, and I obtained in this way some results which will be presented in this work. It will be noticed that my considerations hardly form but a beginning in a certain direction of research, which undoubtedly will give many new results.

Hurwitz's Fourier series proof. This is the proof of Hurwitz from 1902 [**22**]. First we address a couple of notational issues. We use old-school notation for Fourier series:

$$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos nt + b_n \sin nt.$$

With this notation Parseval's theorem reads

$$\frac{1}{\pi}\int_0^{2\pi} f(x)g(x)\,dx = \frac{a_0 b_0}{2} + \sum_{n=1}^{\infty} a_n b_n + c_n d_n.$$

Theorem. *$L^2 - 4\pi A \geq 0$, with equality only for the circle.*

Proof. Choose a parameterization $t \mapsto (x(t), y(t))$ that traces out a given curve of length L in time 2π with constant speed. This translates to

$$\dot{x}^2 + \dot{y}^2 = \left(\frac{L}{2\pi}\right)^2.$$

Let the Fourier series of the coordinate functions be

$$x(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos nt + b_n \sin nt,$$

$$y(t) = \frac{1}{2}c_0 + \sum_{n=1}^{\infty} c_n \cos nt + d_n \sin nt.$$

Then the Fourier series for the derivatives are

$$\dot{x}(t) = \sum_{n=1}^{\infty} n(b_n \cos nt - a_n \sin nt),$$

$$\dot{y}(t) = \sum_{n=1}^{\infty} n(d_n \cos nt - c_n \sin nt).$$

Next we use Parseval's theorem to express L^2 and A in terms of these coefficients.

$$L^2 = 4\pi^2 \left(\frac{L}{2\pi}\right)^2 = 2\pi^2 \left(\frac{1}{\pi}\int_0^{2\pi} \left(\frac{L}{2\pi}\right)^2 dt\right) = 2\pi^2 \left(\frac{1}{\pi}\int_0^{2\pi} \dot{x}^2 + \dot{y}^2\, dt\right)$$

$$= 2\pi^2 \left(\sum_{n=1}^{\infty} n^2(a_n^2 + b_n^2 + c_n^2 + d_n^2)\right),$$

$$A = \int_0^{2\pi} x\dot{y}\, dt = \pi \sum_{n=1}^{\infty} n(a_n d_n - b_n c_n).$$

We hope that the isoperimetric inequality will follow when we combine these expressions. It does, as follows:

$$L^2 - 4\pi A = 2\pi^2 \left(\sum_{n=1}^{\infty} n^2(a_n^2 + b_n^2 + c_n^2 + d_n^2) - 2n(a_n d_n - b_n c_n)\right)$$

$$= 2\pi^2 \left(\sum_{n=1}^{\infty} (na_n - d_n)^2 + (nb_n + c_n)^2 + (n^2 - 1)(c_n^2 + d_n^2)\right)$$

$$\geq 0.$$

When is there equality? Looking at the term for $n = 1$ we get

$$a_1 = d_1, \quad b_1 = -c_1.$$

For all larger n the coefficients vanish

$$a_n = b_n = c_n = d_n = 0,$$

that is,

$$x(t) = \frac{1}{2}a_0 + a_1 \cos t + b_1 \sin t,$$
$$y(t) = \frac{1}{2}c_0 - b_1 \cos t + a_1 \sin t.$$

The curve described by these equations is a circle. ∎

Fourier analysis is not indispensable for this proof. Hardy, Littlewood, and Pólya [**17**, pp. 186–187] manage without it, but as they say themselves:

> The proof is in principle that of Hurwitz, but differs (a) in that we do not use the theory of Fourier series and (b) in our unsymmetrical treatment of x and y.

We could suggest adding: (c) in that it lacks the aesthetic appeal of Fourier analysis.

Carleman's power series proof. Of all conformal mappings of the (complex) plane, the linear fractional transformations are those that send circles to circles. That is, they preserve the perfection of the circle while other conformal mappings smudge things up. We now map the unit circle conformally onto some (simply connected) figure. If the figure is not a circle, we hope to use one of those smudging mappings to reveal the imperfection $L^2 - 4\pi A > 0$. But if the figure is a circle, a linear fractional transformation will do, resulting in the preservation of "perfection" (i.e., $L^2 - 4\pi A = 0$). To prove this, we resort to using power series. This is the proof of Carleman from 1921 [**8**].[5]

Theorem. *$L^2 - 4\pi A \geq 0$, with equality only for the circle.*

Proof. Given a simple closed curve of finite length L, we map the unit disk onto its interior by a conformal mapping φ and express the length and area in terms of φ'. By standard formulas from complex analysis,

$$L = \int_0^{2\pi} |\varphi'(z)|\, d\theta, \qquad A = \int_0^{2\pi}\int_0^1 |\varphi'(z)|^2 r\, dr\, d\theta.$$

To be able to compare the power series of the two terms we consider a function ψ that is analytic in the unit disk and satisfies $\psi^2 = \varphi'$. Then ψ has a power series expansion about the origin

$$\psi(z) = \sum_{n=0}^{\infty} a_n z^n.$$

[5] Kraus [**24**] in 1932 gives exactly the same proof and says: "If the proof of the isoperimetric inequality has already been carried out in this way, then I have not been able to find it in the literature."

Let $b_n = a_n a_0 + a_{n-1}a_1 + \cdots + a_0 a_n$, so that

$$\left(\sum_{n=0}^{\infty} a_n z^n\right)\left(\sum_{n=0}^{\infty} a_n z^n\right) = \sum_{n=0}^{\infty} b_n z^n.$$

Then

$$L = \int_0^{2\pi} |\psi^2(z)|\, d\theta = 2\pi \sum_{n=0}^{\infty} |a_n|^2,$$

$$A = \int_0^{2\pi}\int_0^1 |\psi^2(z)|^2 r\, dr\, d\theta = \int_0^{2\pi}\int_0^1 |b_n|^2 r^{2n+1}\, dr\, d\theta$$

$$= \pi \sum_{n=0}^{\infty} \frac{|b_n|^2}{n+1}.$$

The isoperimetric inequality becomes

$$\sum_{n=0}^{\infty} \frac{|b_n|^2}{n+1} \le \left(\sum_{n=0}^{\infty} |a_n|^2\right)\left(\sum_{n=0}^{\infty} |a_n|^2\right),$$

and we can prove it by making termwise comparisons and invoking Cauchy's inequality:

$$\begin{aligned} |b_n|^2 &= |a_n a_0 + a_{n-1}a_1 + \cdots + a_0 a_n|^2 \\ &\le \left(|a_n|^2|a_0|^2 + |a_{n-1}|^2|a_1|^2 + \cdots + |a_0|^2|a_n|^2\right)(n+1). \end{aligned}$$

Equality occurs when

$$\begin{aligned} &|a_n a_0 + a_{n-1}a_1 + \cdots + a_0 a_n|^2 \\ &\quad = \left(|a_n|^2|a_0|^2 + |a_{n-1}|^2|a_1|^2 + \cdots + |a_0|^2|a_n|^2\right)(n+1), \end{aligned}$$

that is, when all the terms in the sum $a_n a_0 + a_{n-1}a_1 + \cdots + a_0 a_n$ are equal for all n. This means that $\{a_n\}$ is a geometric sequence, say with ratio q. The consequence for φ' is that

$$\varphi'(z) = \psi(z)^2 = \frac{a_0^2}{(1-qz)^2},$$

which we recognize as the derivative of a linear fractional transformation. So an area maximizing curve is the image of the unit circle under a linear fractional transformation—a circle. ∎

9. CONVEXITY. Here's what we're going to do. We take an arbitrary noncircular figure. We manipulate it with clever reflections and things to get a nonconvex figure of the same area and perimeter. A nonconvex figure cannot have maximal area, so we will then have shown that no noncircular figure can have the maximal area.

Convexity and symmetry. I saw this proof in [**21**]. Its origin is apparently unclear.

Theorem. *If there is a smooth solution to the isoperimetric problem, then it is a circle.*

Proof. Take a smooth solution. Place it so that the x-axis cuts the perimeter in half. Make a new figure from each half by combining it with its reflection in the x-axis. Each of these two new figures must have the maximal area as well, for together they have twice the maximal area and neither can have more than maximal area. Since they have maximal area, they must be convex. Therefore, the x-axis must have cut the original figure at right angles, for otherwise one of our new figures would have a dent there (Figure 41).

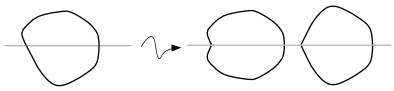


Figure 41.

Take the new figures and repeat the process, only this time split the perimeter with the y-axis instead. This gives four, still smooth, figures with maximal area that are symmetric in both the x- and y-axes. Because of this symmetry, any line through the origin cuts the perimeters of these figures in half, as demonstrated in Figure 42.

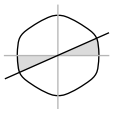


Figure 42.

Therefore, any line through the origin must meet each of the four figures (when placed with their "centers" at the the origin) at right angles. Otherwise, by reflecting the halves we would again obtain a dented figure with maximal area. We conclude that the figures are all circles, so we must have had a circle to begin with. ■

Convexity and tangents. Here is a more general but less clean convexity proof of Demar from 1975 [**9**].

Theorem. *If there is a solution to the isoperimetric problem, then it is a circle.*

Proof. Take a convex figure and extend two of its tangents until they meet (Figure 43).

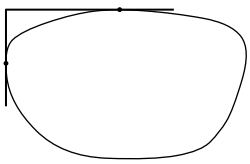


Figure 43.

Now draw the line segment between the two points of tangency and its perpendicular bisector (Figure 44):

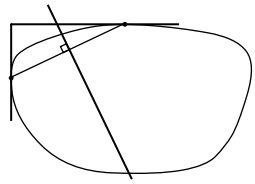


Figure 44.

For the figure to have maximal area, the tangents must meet on this perpendicular line, for otherwise we could make a nonconvex figure with maximal area by reflecting the cut-off piece in it (Figure 45):

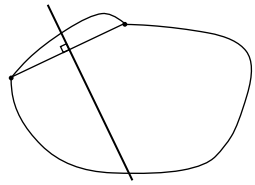


Figure 45.

This construction rules out many figures as candidates for maximal area. In particular, it rules out all figures with any straight segments, for if one of our points of tangency were on such a segment, we could move it a little bit, causing a change in the perpendicular bisector, while the tangent would still be the same, so the tangents could not always meet on the perpendicular.

The foregoing construction also demonstrates that if we have the figure with the maximal area, then the tangents meet on this line and are of equal length. Moreover, this also holds for the normals at the points of tangency (Figure 46):

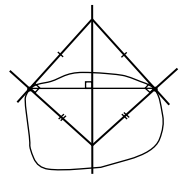


Figure 46.

We now wish to pick a third, arbitrary point on the boundary and show that it and the two points of tangency are on a circle. For the construction to work, we cannot pick this new point completely arbitrarily, however. The solution curve must have a tangent at this point. This is not a problem, since there is a dense set of points on the curve at which it has tangents (only nonconvexity could mess that up). Also, the tangent cannot be parallel to the two tangents we have already. This is not a problem either, since this can happen only at isolated points now that we have ruled out figures with straight segments.

With our new, third point, the situation looks something like Figure 47:

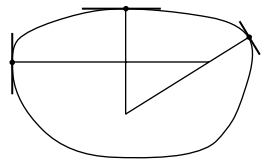


Figure 47.

The important thing is that the normals form a triangle that we now prove is not a triangle at all, but a point. We have just shown that the normals are pairwise equal:

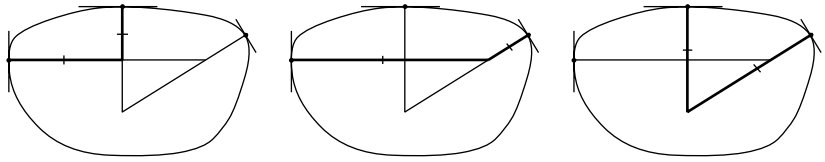
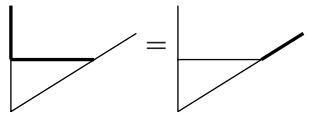
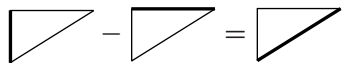


Figure 48.

We combine the first two of these relations to get



Subtracting this from the third relation gives



The conclusion: one side is the sum of the other two, so the triangle is not a proper triangle, and it is certainly not a line, so it must in fact be a point. It follows that all three points are equidistant from this point, and therefore they lie on a circle. ■

10. PARALLEL CURVES. Take a convex figure and roll a circle along its boundary. The curve that is traced out by the center of the circle is what we call a *parallel curve* (Figure 49).

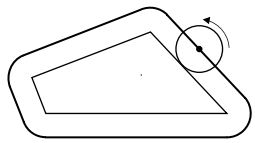


Figure 49.

Or, if you prefer, dip the circle in paint and have it bounce around with its midpoint trapped inside the figure. It will then paint the inside of the parallel figure.

The beautiful thing about parallel curves is that they preserve the quantity $L^2 - 4\pi A$. The isoperimetric inequality $L^2 - 4\pi A \geq 0$ suggests that we think of $L^2 - 4\pi A$ as a quantity that shows how far the figure is from being a circle. Then a parallel figure is as far from being a circle as the original figure. Intuitively, we feel that this is only natural, since taking the parallel curve means, in a sense, "adding a circle" (and for circles we have $L^2 - 4\pi A = 0$).

Let's establish the preservation of $L^2 - 4\pi A$. For the rest of this section we will, for the sake of simplicity, deal only with parallel curves of convex polygons. So take a convex polygon and construct the parallel curve using a circle of radius r (Figure 50).

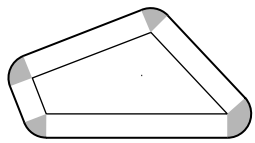


Figure 50.

The new area is the old area plus the area of the strips plus the area of the shaded pieces. This is all easy to calculate once we realize that the shaded pieces fit together to form a disk. This is so because the sum of the angles of the shaded pieces is the amount by which we turn when we walk around the figure once (Figure 51).

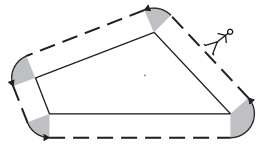


Figure 51.

The preservation of $L^2 - 4\pi A$ is now easy to verify, for we have the relations between the old area a and perimeter l and their new counterparts A and L:

$$A = a + lr + r^2\pi, \qquad L = l + 2\pi r,$$

giving

$$L^2 - 4\pi A = l^2 + 4l\pi r + 4\pi^2 r^2 - 4\pi a - 4\pi lr - 4\pi^2 r^2 = l^2 - 4\pi a.$$

This preservation of $L^2 - 4\pi A$ seems like a good thing when our aim is to prove the isoperimetric inequality, but outer parallel curves are not of much use. What we really need is inner parallel curves, which essentially allow us to move inwards until the area is exhausted, for then surely $L^2 - 4\pi A \geq 0$.

We consider two ways of setting this up. First, we come up with a contrived definition of inner parallel curves, ensuring that they, too, magically preserve $L^2 - 4\pi A$. We also give a more straightforward definition of inner parallel curves for which $L^2 - 4\pi A$ is not preserved. Nevertheless, this new definition gets the job done.

Parallel curves with a twist. One way to think about outer parallel curves is this: the parallel curve is traced out by a moving normal, and when we reach a corner we let the normal turn continuously to its new direction (Figure 52).

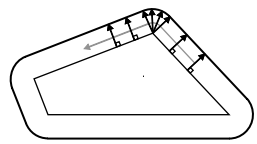


Figure 52.

We can do the same thing with an inward-pointing normal (Figure 53).

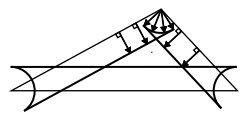


Figure 53.

Now see how clever we have been. We decide that the perimeter along the circular arcs should be counted negatively, and so too the areas they bound. (This makes some sense in terms of orientation.) Then the new area a and perimeter l will be

$$a = A - Lr + r^2\pi, \qquad l = L - 2\pi r.$$

We have thus managed to preserve $L^2 - 4\pi A$:

$$l^2 - 4\pi a = L^2 - 4L\pi r + 4\pi^2 r^2 - 4\pi A + 4\pi Lr - 4\pi^2 r^2 = L^2 - 4\pi A.$$

We can now prove the isoperimetric inequality by taking any polygon and showing that $L^2 - 4\pi A \geq 0$ for one of its parallel curves. We imagine taking inner parallel curves further and further in until the area has shrunken to nothing. For a square, for instance, we should use a normal of half the side length to trace out the parallel curve (Figure 54).

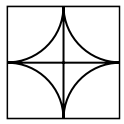


Figure 54.

There is then no positive area left, so indeed $L^2 - 4\pi A \geq 0$. But the picture says more. The negative area we are left with is composed precisely of the pieces we would have to take away from the square for it to become a circle. This works for any polygon that circumscribes a circle (Figure 55).

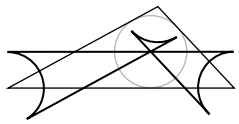


Figure 55.

But this pretty property is lost for polygons not circumscribing circles (Figure 56).

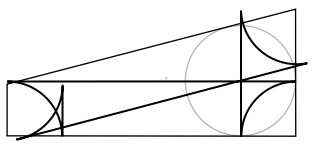


Figure 56.

This figure is all messed up on the left: there is a little bit of twice negative area for instance (three sides have swept past it). Nevertheless, it is still intuitively obvious that when we take the inner parallel curve with an offset that is the radius of the greatest circle we can inscribe, then we run out of positive area and get $L^2 - 4\pi A \geq 0$, as has been proved, among others, by Bonnesen in 1921 [**3**], [**4**].

Parallel curves the no-nonsense way. For this approach we follow Bol's article from 1943 [**2**]. Bol is surprised that this "truly simple proof" has not been given earlier.

Theorem. $L^2 - 4\pi A \geq 0$ *for convex polygons.*

Proof. Given a convex polygon, we create parallel inner polygons by sliding all sides perpendicularly inward. This process gives polygons of fewer and fewer sides, eventually terminating in a point or a line (Figure 57).

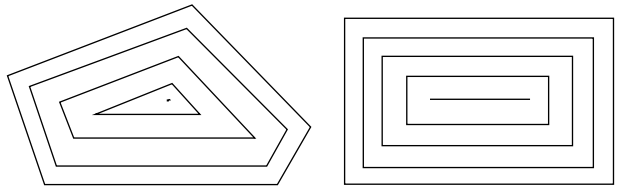


Figure 57.

Points and lines have no area, hence for them the isoperimetric inequality is certainly true. The idea is now to show that the quantity $L^2 - 4\pi A$ decreases as we move inward. Then $L^2 - 4\pi A$ must have been nonnegative to begin with, because we make it smaller and smaller and still end up with something nonnegative.

We illustrate the procedure in a case where the number of sides is reduced in order to stress that the reasoning is always the same. Let's say that the perimeters are L for the big polygon and l for the small one, and that we have moved the sides inward by r (Figure 58).

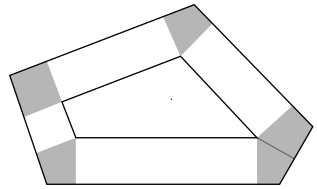


Figure 58.

The big area A is the small area a plus the area of the strips lr plus the area of the shaded pieces. As before, we see that the shaded pieces fit together, but this time to form something like the shape in Figure 59:

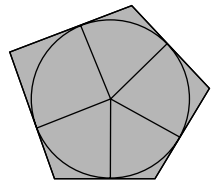


Figure 59.

The spokes are equally long—they have length r—so we are dealing with an inscribed circle. Inspired by that, we write $r^2\hat{\pi}$ for the area, where $\hat{\pi}$ must be slightly larger than π. This means that the perimeter is $2r\hat{\pi}$, because the area is r times the perimeter divided by two (Figure 60),

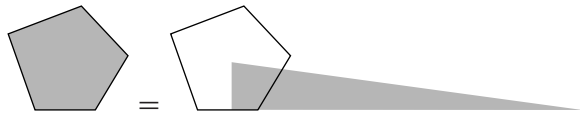


Figure 60.

as we see by summing the areas of the triangles in Figure 61:

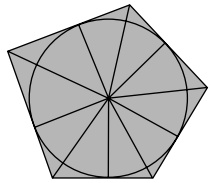


Figure 61.

Now we have all we need to be able to tell what happens to $L^2 - 4\pi A$, for we have

$$A = a + lr + r^2\hat{\pi}, \qquad L = l + 2r\hat{\pi},$$

which gives

$$\begin{aligned} L^2 - 4\pi A &= l^2 + 4lr\hat{\pi} + 4r^2\hat{\pi}^2 - 4\pi a - 4\pi lr - 4\pi r^2\hat{\pi} \\ &= l^2 - 4\pi a + 4rl(\hat{\pi} - \pi) + 4r^2\hat{\pi}(\hat{\pi} - \pi) \\ &> l^2 - 4\pi a, \end{aligned}$$

confirming that this quantity decreases, as claimed. ■

11. INTEGRAL GEOMETRY. In prerevolutionary France, Georges Louis Leclerc, Comte de Buffon, spent his bourgeois leisure time tossing needles on the floor. In 1777 he wrote ([**5**], quoted from [**6**, p. 473]):

> I suppose that in a room in which the parquet floor is simply divided by parallel joints, one throws a stick in the air and one of the players bets that the stick will not cross any of the parallels of the parquet floor, and that the other bets on the contrary that the stick will cross some of these parallels; the chances of these two players are asked for.

Santaló's integral geometry proof. This is a proof of Santaló [**31**, pp. 37–38].

Theorem. $L^2 - 4\pi A \geq 0$.

Proof. Consider a convex figure D of perimeter L. Now, instead of throwing needles at it, throw circles. And use circles of the same perimeter as D (i.e., with radius $r = L/2\pi$; see Figure 62, where D is the figure with the boldface boundary).

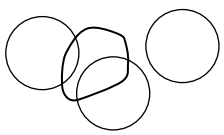


Figure 62.

We hope that this will give us information about the area of D—the more circles intersect it the greater its area—and we capture this intuition by considering the weighted area of the plane,

$$\iint_{\mathbb{R}^2} I(x, y)\, dx\, dy,$$

where we determine the weight $I(x, y)$ of an infinitesimal square by putting the center of one of our circles there and counting how many times it intersects the boundary ∂D of D. (One pesky detail is that the intersection-counting function $I(x, y)$ might be infinite at some points, but of course it will be so almost nowhere, so this cannot hurt the integral.)

An approximation to this weighted area immediately occurs to those who paid careful attention when reading the last section. Only the region at most r away from D is given nonzero weight, that is, the area of the region D_r inside the outer parallel curve at distance r from D. This is, as we recall, $A + Lr + r^2\pi$. Also, if we drop one of our circles with its center anywhere inside this parallel figure, then it will intersect the boundary of D at least twice (there will be no zero-area in the middle, for the circle cannot fit inside the figure). In terms of formulas,

$$\iint_{\mathbb{R}^2} I(x, y)\, dx\, dy \geq \iint_{D_r} 2\, dx\, dy = 2\left(A + Lr + r^2\pi\right).$$

Now we calculate the weighted area by brute force. Consider a point on ∂D. Which circles intersect this point? They are, of course, the circles with centers at distance r from it (Figure 63).

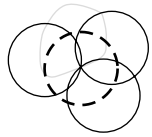


Figure 63.

Consider an infinitesimal ds-segment of ∂D. Which circles intersect this segment? They must be circles with centers that lie in two strips like those in Figure 64:

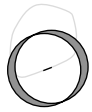


Figure 64.

The strips have height $2r$ (in the direction perpendicular to the boundary segment ds) and constant width ds, so the ds-segment contributes a total amount of weighted area $4r\,ds$. Summing these contributions we obtain

$$\iint_{\mathbb{R}^2} I(x, y)\, dx\, dy = \int_{\partial D} 4r\, ds = 4rL.$$

We now have two expressions for the weighted area that we can combine,

$$4rL \geq 2\left(A + Lr + r^2\pi\right),$$

and, remembering that we chose $r = L/2\pi$, this reduces to the isoperimetric inequality. ■

There is more to be extracted from these ideas. We write A_n for the area with weight n, giving a total weighted area of

$$A_1 + 2A_2 + 3A_3 + \cdots,$$

although all the odd terms will be zero of course (almost any circle intersects ∂D an even number of times). We have just calculated the total weighted area to be $4rL$, from which we conclude that

$$4rL = 2A_2 + 4A_4 + 6A_6 + \cdots.$$

And the area inside the parallel curve consists of all the areas with nonzero weight,

$$A + rL + r^2\pi = A_2 + A_4 + A_6 + \cdots.$$

Taking half the first of these equations and subtracting the second gives

$$rL - A - r^2\pi = A_4 + 2A_6 + 3A_8 + \cdots.$$

Since $r = L/2\pi$, this gives us an expression for $L^2 - 4\pi A$ in terms of weighted areas:

$$L^2 - 4\pi A \geq 4\pi\,(A_4 + 2A_6 + 3A_8 + \cdots)\,.$$

Now we see that there can be equality only for the circle, for that's the only case where there is no four-area A_4. We can see this as follows. Place a circle on D as shown in Figure 65:

Figure 65.

Slide the circle along ∂D, always keeping just a little bit of it outside. For there to be no four-area, the other side of the circle must never hit ∂D. This works when D and the circle are the same, or when D contains the circle, which is of course impossible.

We have done our best to stay clear of generalizations of the isoperimetric theorem, but we cannot resist the temptation to mention one pretty generalization connected with this proof, a generalization that is also important for understanding much of the isoperimetric literature. Recall one of our definitions of parallel curves: we took a circle, dipped it in paint, and moved it about with its midpoint inside a figure in order to paint the parallel figure. Obviously, we could generalize this procedure, taking something other than the circle as our brush. We could then take any point in its interior to be the point that is not allowed to leave the figure. Nor are we allowed to twist the brush. The result is what is known as *Minkowski addition*. The behavior of this procedure is captured by the Minkowski inequality:

$$\sqrt{\text{painted area}} \geq \sqrt{\text{area of the figure}} + \sqrt{\text{area of the brush}}.$$

If we choose the brush to be a circle, then this inequality asserts that

$$\sqrt{\text{area inside the parallel curve}} \geq \sqrt{\text{area of the figure}} + \sqrt{\text{area of the circle}},$$

and it readily reduces to the isoperimetric inequality. The most beautiful proof of the Minkowski inequality that I know of is a very natural generalization of the proof we have given here—we just drop this brush figure instead of the circle and everything works out. A very readable exposition of this proof is found in [**12**].

1. W. Blaschke, *Kreis und Kugel*, 1916; reprinted by Chelsea, New York, 1949.
2. G. Bol, Einfache Isoperimetriebeweise für Kreis und Kugel, *Abh. Math. Sem. Univ. Hamburg* **15** (1943) 27–36.
3. T. Bonnesen, Über eine Verschärfung der isoperimetrischen Ungleichheit des Kreises in der Ebene und auf der Kugeloberfläche nebst einer Anwendung auf eine Minkowskische Ungleichheit für konvexe Körper, *Math. Annalen* **84** (1921) 216–227.
4. ———, *Les problèmes des isopérimètres et des isépiphanes*, Gauthier-Villars & Cie, Paris, 1929.
5. G. Buffon, *Essai d'arithmétique morale*, 1777.
6. ———, *Œuvres philosophiques de Buffon*, Presses universitaires de France, Paris, 1954.
7. C. Carathéodory and E. Study, Zwei Beweise des Satzes, daß der Kreis unter allen Figuren gleichen Umfanges den größten Inhalt hat, *Math. Annalen* **68** (1909) 133–140.
8. T. Carleman, Zur Theorie der Minimalflächen, *Math. Z.* **9** (1921) 154–160.
9. R. F. Demar, A simple approach to isoperimetric problems in the plane, *Math. Mag.* **48** (1975) 1–12.
10. F. Edler, Vervollständigung der Steinerschen elementargeometrischen Beweise für den Satz, daß der Kreis grösseren Flächeninhalt besitzt, als jede andere ebene Figur gleich grossen Umfanges, *Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen* (1882) 73–80.
11. L. Euler, *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes sive solutio problematis isoperimetrici latissimo sensu accepti*, 1744, German translation in [**34**].
12. H. Flanders, A proof of Minkowski's inequality for convex curves, this MONTHLY **75** (1968) 581–593.
13. G. Galilei, *Discorsi e dimostrazioni matematiche intorno a due nuove scienze*, Elsevier Press, Leiden, 1638.
14. S. Gandz, Isoperimetric problems and the origin of quadratic equations, *Isis* **32** (1940) 103–115.
15. C. F. Geiser, Zur Erinnerung an Jakob Steiner, *Verhandlungen der Schweizerischen Naturforschenden Gesellschaft* **56** (1874) 215–251.
16. M. Gromov, Isoperimetric inequalities in Riemannian manifolds, 1986, appendix in [**28**].
17. G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, 1934.
18. T. L. Heath, *A History of Greek Mathematics*, vol. 2, Clarendon Press, Oxford, 1921.
19. D. Hilbert, Mathematische Probleme. *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen* (1900) 253–297; English translation in [**20**].
20. ———, Mathematical problems, *Bull. Amer. Math. Soc.* **8** (1902) 437–479.
21. H. Howards, M. Hutchings, and F. Morgan, The isoperimetric problem on surfaces, this MONTHLY **106** (1999) 430–439.
22. A. Hurwitz, Sur quelques applications géométriques des séries de Fourier, *Annales Scientifiques de l'École Normale Supérieure* **19** (1902) 357–408.
23. Morris Kline, *Mathematics for the Nonmathematician*, Dover, New York, 1985.
24. W. Kraus, Über den Zusammenhang einiger Charakteristiken eines einfach zusammenhängenden Bereiches mit der Kreisabbildung, *Mitt. Math. Sem. Giessen* **21** (1932) 1–28.
25. J. L. Lagrange, Essai d'une nouvelle méthode pour déterminer les maxima et les minima des formules intégrales indéfinies., *Miscellanea Philosophica-Mathematica Societatis Privatae Taurinensis* **2** (1762) 173–195; German translation in [**35**].
26. G. Lawlor, A new area maximization proof for the circle, *Math. Intelligencer* **20** (1998) 29–31.
27. S. Lhuilier, *De relatione mutua capacitatis et terminorum figurarum etc*, Warsaw, 1782.
28. V. D. Milman and G. Schechtman, *Asymptotic Theory of Finite Dimensional Normed Spaces*, Lecture Notes in Mathematics, no. 1200, Springer-Verlag, Berlin, 1986.
29. G. R. Morrow, *Proclus: A Commentary on the First Book of Euclid's Elements*, Princeton University Press, Princeton, 1970.
30. O. Perron, Zur Existenzfrage eines Maximums oder Minimums, *Jahresber. Deutsch. Math.-Verein.* **22** (1913) 140–144.
31. L. Santaló, *Introduction to Integral Geometry*, Hermann & Cie, Paris, 1953.
32. E. Schmidt, Über das isoperimetrische Problem im Raum von n Dimensionen, *Math. Z.* **44** (1939) 689–788.
33. M. Spivak, *A Comprehensive Introduction to Differential Geometry*, vol. 4, 2nd ed., Publish or Perish, Berkeley, CA, 1979.
34. P. Stäckel, *Abhandlungen über Variations-Rechnung. Erster Theil.*, vol. 46 of *Ostwald's Klassiker der exakten Wissenschaften*, Engelman, Leipzig, 1894.
35. ———, *Abhandlungen über Variations-Rechnung. Zweiter Theil.*, vol. 47 of *Ostwald's Klassiker der exakten Wissenschaften*, Engelman, Leipzig, 1894.

36. J. Steiner, Einfache Beweise der isoperimetrischen Hauptsätze, *J. Reine Angew. Math.* **18** (1838) 281–296.
37. ———, Sur le maximum et le minimum des figures dans le plan, sur la sphère et dans l'espace en général. premier mémoire, *J. Reine Angew. Math.* **24** (1842) 93–162.
38. ———, Sur le maximum et le minimum des figures dans le plan, sur la sphère et dans l'espace en général. second mémoire, *J. Reine Angew. Math.* **24** (1842) 189–250.
39. ———, *Gesammelte Werke*, vol. 2, Berlin, 1882.
40. I. Thomas, *Selections Illustrating the History of Greek Mathematics*, vol. 2, Harvard University Press, Cambridge, 1941.
41. G. J. Toomer, *Ptolemy's Almagest*, Princeton University Press, Princeton, 1998.
42. K. Weierstrass, *Mathematische Werke*, vol. 7, Mayer & Müller, Berlin, 1927.

VIKTOR BLÅSJÖ recently received his undergraduate degree from Stockholm University. Seeing his work warmly received by the MONTHLY, he feels that perhaps there is a place in this world after all for a humble student who just wants to understand things.
viktorblasjo@mac.com

To August Ferdinand

If a circle you send
From the deserts of sand
Of a land Oriental and Gobi-ous,
With a circle you'll end,
As we now understand,
Thanks to nonorientable Möbius
That umlaut is hard, and woe be us—
We always misprOnounce your name.
Ah, Möbius, Möbius, Möbius—
You one-sided stripper of fame!

——Submitted by Seth Braver, University of Montana (Missoula)