The Isoperimetric Problem on Surfaces

Hugh Howards, Michael Hutchings, and Frank Morgan

1. Introduction. The isoperimetric problem on a surface is to enclose a given area with the shortest possible curve. The classical isoperimetric theorem asserts that in the plane the unique solution is a circle. On curved surfaces the isoperimetric problem is harder and much remains open. Even on the simplest paraboloid the "obvious" solution was proved only in 1996 by Benjamini and Cao ([2, Thms. 5, 8]; see also [24, Prop. 7], [22, Thm. 3.1], [30, Thm. 1], [29], [26]):

Theorem 1.1 (Benjamini and Cao). The unique least-perimeter way to enclose given area in the paraboloid of revolution

\[
P = \{z = x^2 + y^2\} \subset \mathbb{R}^3
\]

is a horizontal circle \(z = c\).

This article gives our three favorite proofs of the classical isoperimetric theorem in the plane and then presents some recent results on other surfaces, including a new proof for the paraboloid. Section 2 uses an amazingly simple symmetry argument to show that a nice minimizer must be a circle. Unfortunately this approach needs to assume that a nice minimizer exists. Section 3 gives a very simple, complete proof without assuming a nice minimizer exists, following the undergraduate thesis of Howards [15]. Section 4 provides another complete proof, a slight twist on a magical proof of Gromov [10].

In general surfaces the existence of a nice, one-component perimeter-minimizing curve has been astonishingly problematic. Fortunately a relatively easy approach is now available from [12], as explained in Section 5. One has to allow the curve to bump up against itself.

Sections 6—8 solve the isoperimetric problem for cylinders, cones, flat tori, and Klein bottles. Section 9 treats the paraboloid and
certain other surfaces of revolution. Section 10 discusses hyperbolic surfaces.

This work was partly inspired by a more difficult question we heard from J.C.C. Nitsche about the soap film between a large wire boundary and a small, moveable loop of thread. The thread wants to position itself to minimize the area of the soap film outside it. If the thread were constrained to lie in a fixed surface bounded by the wire (which unfortunately is not the case), then the thread would want to be an isoperimetric curve in that surface.

Osserman [23] provides a marvelous survey on the isoperimetric inequality.

2. The circle in the plane, assuming smooth existence. We assume that there is a compact minimizer $C$ among smooth curves of finitely many components and enclosed area $\pi$, and use symmetry to prove it must be a single round unit circle; existence is a nontrivial assumption, a fact overlooked by some early workers. The proof uses a symmetry argument we heard from Brian White and Luen-fai Tam, who thought it originated with Blaschke (see [9, Thm. 3.4], [17, Thm. 5.3], and [16, §2]); we have been unable to trace its origin and would be grateful to anyone who could help.

Suppose $C$ is not a round circle. Take a horizontal line splitting the enclosed area in half. Each half must have the same length, or the shorter half, together with its reflection, would be shorter than $C$. Replacing $C$ by half plus its reflection if necessary, we may assume that $C$ is symmetric across the horizontal line. Similarly we may assume that $C$ is symmetric across a vertical line. We may assume the lines meet at the origin. Now $C$ is symmetric under the composition of the two reflections, i.e., under 180-degree rotation around the origin. Hence every line through the origin splits the area in half. $C$ must meet every line through the origin orthogonally; otherwise, one half of $C$, together with its reflection, would not be convex, and its convex hull would have less perimeter and more area. It follows that $C$ consists of circles about the origin. A single circle is best. We conclude that the original $C$ is a round circle.

This argument can be generalized to prove that a round hypersphere is perimeter-minimizing for given volume in $\mathbb{R}^n$, in the round sphere $\mathbb{S}^n$, and in hyperbolic space $\mathbb{H}^n$. More generally, it
shows that a minimizing cluster of k bubbles enclosing k < n
prescribed volumes in \( \mathbb{R}^n \) has \( O(n-k+1) \) symmetry, assuming known
but difficult existence and regularity [16, Thm. 2.6]. It played an
essential role in the recent proof by Hass, Hutchings, and Schlaflpy of
the equal volumes case of the still open Double Bubble Conjecture,
which says that the familiar standard double soap bubble is the
least-area way to enclose and separate two given volumes of air
([11], [16], [14], [18], [13]).

3. The circle in the plane, without assuming existence. To
prove that the circle is perimeter-minimizing (but not necessarily
uniqueness), by approximation it suffices to show that the shortest
n-gon enclosing given area is the regular n-gon. In his undergraduate
thesis, Howards [15] gave the following geometric proof free of
variational calculus, including ideas that we have since traced back to
Zenodorus about 200 BC, Steiner in 1838 ([27], [28, p. 105 and Fig.
6]), and Courant and Hilbert [6, p. 166]; see the interesting "A history
of the classical isoperimetric problem" by Porter [25] and Bonnesen
and Fenchel [5, §57].

By compactness, there is a shortest n-gon in the 2n-
dimensional space of vertices. It is convex. Consider two adjacent
sides, which determine a triangle, and the line L through the common
vertex and parallel to the third side of the triangle. These two sides
must constitute the shortest path to L and back, since all such
constructions yield triangles of the same area. The first side, together
with the reflection of the second across L, must form a straight line.
Hence the two sides have the same length. Therefore the n-gon is
equilaterial.

To prove that the equilateral n-gon is regular, we begin with n
even. For opposite vertices P, Q, the line PQ must have the same area
above as below, or a reflection of the larger half would enclose more
area (or, scaled down, the same area with less length). For an
intermediate vertex M, the angle PMQ must be 90°, or replacing it
with a 90° angle and reflecting as in Figure 3.1 would increase the
area enclosed. Therefore the n-gon is inscribed in a circle and must
be regular.
Figure 3.1
The angle PMQ must be 90°, or replacing it with a 90° angle and reflecting would increase the area enclosed.
Finally suppose \( n \) is odd. A regular \( 2n \)-gon comes from putting little triangles on the sides of the regular \( n \)-gon. If a perimeter-minimizing \( n \)-gon, known to be equilateral, were not regular, putting those little triangles on its sides would yield a non-regular, perimeter-minimizing \( 2n \)-gon, the final contradiction.

This completes the proof that the circle is perimeter minimizing. In fact, now that we know that a minimizer exists, we can use the above arguments to prove uniqueness. Consider any minimizer. It must be convex. As above, a line bisecting its perimeter must bisect the area, and any inscribed angle must be 90°. Therefore, the minimizer must be a circle.

4. The circle in the plane, another proof without assuming existence. Gromov ([10]; see [3, 12.11.4] or [20, 10.5]) gave a proof of the isoperimetric theorem in \( R^n \) by direct comparison. The strategy in \( R^2 \), for example, is to find a vectorfield \( v \) on any competing region \( R \) of area \( \pi \) with smooth boundary \( C \) and outward unit normal \( n \) such that

\[
\text{(4.1) } \text{div}(v) \geq 2,
\]

\[
\text{(4.2) } v \cdot n \leq 1,
\]

with equality everywhere only if \( R \) is a disc. If such a \( v \) can be found, then the isoperimetric inequality follows immediately from Stokes' theorem:

\[
\text{length}(C) \geq \int_C v \cdot n = \int_R \text{div}(v) \geq 2 \text{area}(R) = 2\pi,
\]

with equality only if \( R \) is a disc.

The Gromov proof finds such a \( v \) by a very clever construction, but the resulting \( v \) is not canonical. We now show that there is a canonical such \( v \) when \( n = 2 \).

The canonical \( v \) is the negative of the gravitational field induced by a substance of constant density filling the region \( R \). More precisely,
\[ v(x) = \frac{1}{\pi} \int_{y \in \mathbb{R}} \frac{x-y}{|x-y|^2} \, dy. \]

By the two-dimensional analog of Gauss's law, \( \text{div}(v) = 2 \) in \( \mathbb{R} \), so it now suffices to prove (4.2).

Fix a point \( x \) in the boundary, and choose polar coordinates \((r, \theta)\) around \( x \) so that \( n \) points in the direction \( \theta = \pi \). Then

\[ v(x) \cdot n = \frac{1}{\pi} \int_{y \in \mathbb{R}} \cos \theta \frac{r}{r} \, dy. \]

Since the area of \( \mathbb{R} \) is fixed, this integral is maximized if we put the points of \( \mathbb{R} \) where \( (\cos \theta)/r \) is largest. The level sets of \( (\cos \theta)/r \) are circles tangent to \( C \) at \( x \), with smaller circles giving larger values of \( (\cos \theta)/r \). So clearly a disc of the given area uniquely maximizes the integral, completing the proof.

5. **Proof of existence of nice least-perimeter enclosures.** In the Euclidean plane and in other special cases where all candidates can be convexified, the existence of a (convex) region of least perimeter and prescribed area follows from Blaschke's selection theorem ([4, p. 38], [8, Chapt. 4]). A general smooth Riemannian surface \( S \) requires a more general argument. There must be some hypothesis to prevent the solution from disappearing to infinity as in Figure 5.1. Suppose for now that \( S \) is a compact surface, perhaps with convex boundary.

For the moment we restrict to images of the unit circle parametrized by arclength. Later we consider curves of several components. Then compactness properties of Lipschitz functions (Ascoli-Arzelà Theorem) immediately produce a minimizer. The only problem is that in theory the limit might bump up against itself too wildly to permit the standard variational argument that it has constant geodesic curvature. The solution may actually bump up against itself, as in the cylinder of Figure 5.2. This technical difficulty delayed for 75 years the completion of Poincaré's proof that every smooth convex sphere contains a simple closed geodesic. In 1982 C. Croke [7] gave a complete proof by minimizing a combination of length and energy in a class of piecewise geodesic curves.
Figure 5.1
In the surface of revolution generated by $y = 1/x$, for any given area, there is a sequence of annuli disappearing to infinity with perimeters going to 0.

Figure 5.2
Some least perimeter enclosures on the cylinder bump up against themselves.
More recently, Hass and Morgan ([12]; see also [22, Lemma 2.2]) have provided a very simple approach to more general existence and regularity using local convexification. Away from the boundary of $S$, a minimizing enclosure is an embedded curve of constant geodesic curvature $\kappa_0$, except possibly for finitely many geodesic arcs or isolated points where it bumps up against itself but remains $C^1$. Even at the boundary of $S$ the curve remains $C^1$ and the geodesic curvature satisfies $\kappa \leq \kappa_0$ (weakly). If for bounded area the curve is allowed a large number of components enclosing disjoint regions, no curve bumps itself on the inside and $\kappa \leq \kappa_0$ everywhere. If the curve is allowed several nested components enclosing multiply connected regions, it never bumps itself. A word on the proof: if local convexification causes two pieces of curve to cross, the longer one is rerouted along the shorter. This process reduces length unless the curves were convex to begin with. Given convexity, standard variational arguments prove the rest.

In many noncompact manifolds, such as the Euclidean or hyperbolic plane, one can work inside a large convex set. Hyperbolic surfaces and other surfaces can have thin cusps to infinity with nonconvex truncations, but as long as the area of the cusp is finite, any sequence of curves going off to infinity has area going to 0 and may be discarded.

Existence and some regularity hold as well for clusters in $\mathbb{R}^2$ (enclosing and separating several regions of prescribed areas [21]) and in general dimensions by the techniques of geometric measure theory [19, Chapt. 13]. In higher dimensions you cannot hope to prescribe the topological type; for example, regions connected by thin tubes can disconnect in the limit. Even for curves in the plane, such general techniques do not have the topological control we need.

6. Circular cylinders. On the cylinder $\{x^2 + y^2 = a^2\} \subset \mathbb{R}^3$, the least-perimeter enclosure of area $A$ is a small (round) circle for $A \leq 4\pi a^2$ and two horizontal circles for $A \geq 4\pi a^2$.

Proof. We know that any solution consists of closed curves of constant curvature. If one curve is homotopically trivial and hence is a small round circle, it is the only one, or it could be translated to touch another and contradict regularity. If all the curves are
homotopically nontrivial, there must be at least two of them to enclose area, and two horizontal circles are best. The transition occurs when the circumference of the small circle $2\sqrt{\pi A}$ equals the length of two horizontal circles $4\pi a$, i.e., $A = 4\pi a^2$.

7. Flat tori and Klein bottles (Howards [15, Thm. 3.1]). Let $S$ be a flat torus or Klein bottle with shortest closed geodesic of length $a$. Given $0 < A < \text{area } S$, the least-perimeter region of area $A$ is

(1) a circular disc if $0 < A \leq a^2/\pi$;

(2) a band (possibly Möbius) with geodesic boundary if $a^2/\pi \leq A \leq \text{area } S - a^2/\pi$;

(3) the complement of a circular disc if area $S - a^2/\pi \leq A \leq \text{area } S$.

Proof. Any solution consists of closed curves of constant curvature. As in the argument in Section 6, if one is homotopically trivial and therefore a small round circle, it is the only one or it could be translated to touch another and contradict regularity. If all the components are nontrivial, for any given area the perimeter is uniquely minimized by a single geodesic band with perimeter $2a$. The transitions between types occur when the circle has circumference $2a$.

Remark. The round sphere and round projective plane may be treated by similar arguments [15, Thms. 4.1, 5.1] or by the methods of Section 9 [22, Thms. 3.1, 3.3].

8. Circular cones. On the circular cone \( \{ z = a\sqrt{x^2 + y^2} \} \subset \mathbb{R}^3 \), the least-perimeter enclosure of area $A$ is a horizontal circle.

Proof. If any component does not encircle the vertex, it must be the only component (or it could be translated to touch another component, contradicting regularity), and hence it must be a circle of length $2\sqrt{\pi A}$. Consider a constant-curvature curve that encircles the vertex. It must be symmetric about the line through the vertex and a point most distant from the vertex, as in Figure 8.1, so it must be a horizontal circle. Clearly a single horizontal circle would have less
perimeter than several. Since one horizontal circle has length less than $2\sqrt{\pi A}$, it must be the minimizer.

\[\text{Figure 8.1}\]

A constant-curvature curve about the vertex of the cone must be a circle of smaller circumference than a planar circle of the same area.

\textbf{Remark.} Actually for a simply-connected domain $D$ on a surface with Gauss curvature $K$, the perimeter $L$ satisfies

$$L^2 \geq 4\pi A - 2A \int_D \max\{K, 0\},$$

with equality for the singular limit case of the cone [23, Thm. 4.3].
9. The paraboloid and other surfaces of revolution. Sections 9 and 10 provide some new examples. The following theorem and corollary include the paraboloid. The proof integrates the Gauss-Bonnet theorem.

**Theorem 9.1** ([24, Prop. 7], [22, Thm. 2.1], [29]). Consider the plane with smooth, rotationally symmetric, complete metric such that the Gauss curvature is a strictly decreasing function of the distance from the origin. Then the unique length-minimizing simple closed curve enclosing a given area is a circle centered at the origin.

By Section 5, inside a surface of finite area or inside a large convex ball $B$, for bounded area, there is a minimizer among $C^1$ curves of $m \leq m_0$ components, enclosing $m$ disjoint regions. Away from $\partial B$, it is an embedded curve of constant geodesic curvature $\kappa_0$, except possibly for finitely many geodesic arcs or isolated points where it bumps up against itself. If $m_0$ is large, the curvature $\kappa \leq \kappa_0$ everywhere.

A short proof of the following standard technical lemma is in [20, §9.7, p. 112] (cf. [22, Lemma 2.3]). The idea is that the rate of change of the perimeter is essentially the geodesic curvature, which is controlled by the Gauss-Bonnet theorem.

**Lemma 9.2.** Let $L(A)$ denote the least perimeter of a one-component curve enclosing area $A$. Then $L(A)$ is differentiable almost everywhere and

$$ L(A)^2 \geq 2 \int_0^A L \ L' . $$

**Proof of Theorem 9.1** [22]. In the surface or inside a large convex ball, for $m_0$ large, let $L(A)$ denote the length of a shortest curve of $m \leq m_0$ components enclosing $m$ disjoint regions of total area $A$. First we claim that if $L$ is differentiable at $A$, then $L'(A)$ is the geodesic curvature $\kappa_0$. One geometric interpretation of geodesic curvature is the rate of change of length with respect to area under perturbations of the given minimizer [20, Chapt. 2]. Hence for $\Delta A > 0$, the new minimizers must do at least as well as perturbations of the old one, and $L'(A) \leq \kappa_0$. Similarly for $\Delta A < 0$, $-L'(A) \leq -\kappa_0$. Therefore $L'(A) = \kappa_0$. 
Now Gauss-Bonnet tells us that the total Gauss curvature of the enclosed region equals:

$$2\pi - \int \kappa \geq 2\pi - L(A) \kappa_0 = 2\pi - L(A) L'(A).$$

Let $G(A)$ denote the total Gauss curvature of a disc of area $A$ centered at the origin. Since the Gauss curvature is a decreasing function of radius, any other region with the same area must have less total Gauss curvature. So we have

$$2\pi - L(A) L'(A) \leq G(A),$$

and

$$L(A) L'(A) \geq 2\pi - G(A).$$

By Lemma 9.2, integration from $A = 0$ to $A_1$ yields

$$L(A)^2 \geq 2L(A)L'(A) \geq 4\pi A_1 - 2 \int G(A).$$

This inequality is sharp for a circle centered at the origin, as we can see by integrating the Gauss-Bonnet formula for circles centered at the origin with area $A$ from 0 to $A_1$. Hence equality holds in (9.2), $L$ is differentiable everywhere, and equality holds in (9.1). Therefore a minimizer encloses Gauss curvature $G(A)$ and must be a circle about the origin.

The following general extension to several, perhaps multiply connected regions is deduced in [22, Thm. 3.1]. Here we give a proof for the easy case of positive Gauss curvature, which includes the paraboloid.

**Corollary 9.3.** Among unions of disjoint, perhaps multiply connected regions, a perimeter minimizer exists and is a (round) disc, disc complement, or annulus about the origin. If the Gauss curvature is positive or the total Gauss curvature of every compact region is less than $2\pi$, then the minimizer is a disc.

**Proof for the case of positive Gauss curvature.** By the Gauss-Bonnet theorem, the perimeter $P(r)$ and geodesic curvature $\kappa(r)$ of a circle about the origin of radius $r$ bounding a disc $D$ satisfy
\[(9.3) \quad P' = \kappa P = 2\pi - \int_D K.\]

The total Gauss curvature is at most \(2\pi\), since otherwise eventually \(P' \leq -a < 0\) and \(P\) hits 0. By \(9.3\), \(\kappa\) is positive and decreasing. Consider any collection of simple closed curves enclosing area \(A_1\). By discarding any curves inside others, enclose area \(A_2 \geq A_1\). By Theorem 9.1, each curve alone would be shortest if a circle about the origin. Since \(\kappa = dP/dA\) is decreasing, one single circle about the origin is best. Since \(A_1 \leq A_2\), the circle of area \(A_1\) is best of all.

10. **Hyperbolic manifolds.** We consider geometrically finite complete hyperbolic surfaces (curvature \(-1\)). Such surfaces may be compact or have finitely many ends: cusps (with exponentially shrinking thickness and finite area) or flared ends (asymptotic to the hyperbolic plane).

**Theorem 10.1** [1, Thm. 2.2]. Let \(S\) be a hyperbolic surface. For given area \(0 < A < \text{area}(S)\), a perimeter-minimizing system of embedded rectifiable curves bounding a region of area \(A\) consists of curves of the same constant curvature of one of four types:

(I) a circle,

(II) horocycles around cusps,

(III) two "neighboring curves" at constant distance from a geodesic, bounding an annulus or complement,

(IV) geodesics or single neighboring curves.

The total perimeter \(L\) satisfies

\[(10.1) \quad L \leq \sqrt{A^2 + 4\pi A},\]

with equality for a circle of area \(A\). If \(S\) has at least one cusp, then cases (I) and (III) do not occur and \(L \leq A\); if moreover \(A < \pi\), then a minimizer consists of horocycles bounding neighborhoods of an arbitrary collection of cusps and has perimeter \(L = A\).
Proof sketch. The constant-curvature curves on a hyperbolic surface are circles bounding discs ($\kappa > 1$) or the complement ($\kappa < -1$), horocycles around cusps ($\kappa = 1$) or the complement ($\kappa = -1$), and constant-curvature curves around necks ($|\kappa| < 1$, including the geodesics around the middle of necks with $\kappa = 0$).

A minimizer cannot have more than one circle, since sliding one until it hits another (or itself) would contradict regularity. Since for other types, $dL/\delta A = \kappa$ is less than it is for a circle, (10.1) always holds, and there is an $A_0 \geq 0$ such that if $A < A_0$ the minimizer is a circle, while if $A > A_0$ it is not a circle and (for $\Delta A > 0$)

(10.2) $\Delta L/\Delta A < 1$.

Now a computation shows that an annulus (or complement) as in Case (III) must occur alone, or an operation such as discarding it would contradict (10.2). Therefore the minimizer falls into one of the four asserted cases.

Henceforth assume $S$ has a cusp. Case (I) cannot occur, because sliding the circle out the cusp until it hits itself would contradict regularity. Hence the minimizer always has $|\kappa| \leq 1$, and always $L(A) \leq A$. A computation shows that Case (III) cannot occur.

Finally assume $A < \pi$. We claim there is no minimizer with $-1 \leq \kappa < 1$ and length $L \leq A$, so $-A + \kappa L < 0$. Otherwise, applying Gauss-Bonnet to the enclosed region yields

$$2\pi \chi = -A + \kappa L < 0,$$

$\chi \leq -1$, $-A + \kappa L \leq -2\pi$, $\kappa L \leq -\pi$, $\kappa < 0$, $L \geq \pi > A$, a contradiction.

The remaining possibilities, systems of curves with $\kappa = 1$, consist of horocycles bounding cusp neighborhoods. Since $\kappa = 1$, as you slide a horocycle out a cusp $dL/\delta A = 1$, and therefore its length equals the area of the cusp neighborhood. By the claim, such systems remain minimizing as long as they exist, either for all $A < \pi$ or until they bump up against themselves at some $A_1 < \pi$. If one bumps, by regularity the minimizer has perimeter less than $A_1$, contradicting the claim and proving the theorem.
References


28. J. Steiner, Sur le maximum et minimum des figures dans le plan, sur le sphere, et l'espace en general, *Crelle's J.* 24 (1842) 93-152, 189-250 (figures separate at end of fascicle).


Wake Forest University, Winston-Salem, NC 27109
howards@wfu.edu

Stanford University, Stanford, CA 94305
hutching@math.harvard.edu

Williams College, Williamstown, MA 01267
Frank.Morgan@williams.edu
Brief descriptive summary

The Isoperimetric Problem on Surfaces
by Hugh Howards, Michael Hutchings, and Frank Morgan

The authors begin with their three favorite proofs that in the plane the circle minimizes perimeter for given area, including a slight twist on a magical proof of Gromov. In a paraboloid of revolution, the "obvious" solution turns out to be a recent result, though proved here using little more than calculus and the Gauss-Bonnet formula. There are also recent examples in hyperbolic surfaces.

howards@wfu.edu
hutching@math.harvard.edu
Frank.Morgan@williams.edu
Biographical sketch

The three authors all spent time at Williams College, where Howards and Hutchings participated in Morgan's NSF undergraduate research Geometry Group and Howards wrote his undergraduate thesis.

Howards, who went to Williams and UC San Diego, is assistant professor of Mathematics at Wake Forest University. Hutchings, who went to Harvard, is Szego Assistant Professor of Mathematics at Stanford. Morgan, who went to MIT and Princeton, is Meenan Third Century Professor of Mathematics at Williams College. In January, 1993, he received one of the first MAA national awards for distinguished teaching.
