



A planimeter is a drafting instrument that measures the area enclosed in a region. Its design is based on Green's theorem, one of several global results about curves presented in this chapter.

Additional Topics in Curves

This chapter presents several excursions that delve more deeply into the geometry of curves, including some of the famous theorems in the field. The theory of curves is an old and extremely well developed mathematical topic. Our aim is simply to describe a few fundamental and interesting highlights.

1. Theorems of Hopf and Jordan

This section is devoted to proving the following two historically significant global theorems:

THEOREM 2.1.

Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a simple closed plane curve. Let $C = \gamma([a, b])$ denote its trace.

- (1) (**Hopf's Umlaufsatz**) The rotation index of γ is either 1 or -1 .
- (2) (**The Jordan Curve Theorem**) $\mathbb{R}^2 - C = \{\mathbf{p} \in \mathbb{R}^2 \mid \mathbf{p} \notin C\}$ has exactly two path-connected components. Their common boundary is C . One component (which we call the **interior**) is bounded, while the other (which we call the **exterior**) is unbounded.

Each theorem provides a method to meaningfully distinguish between the two possible orientations of γ , and the methods they provide are equivalent:

DEFINITION 2.2.

A simple closed plane curve $\gamma : [a, b] \rightarrow \mathbb{R}^2$ is called **positively oriented** if it satisfies the following equivalent conditions:

- (1) The rotation index of γ equals 1.
- (2) The interior is on one's left as one traverses γ ; more precisely, for each $t \in [a, b]$, $R_{90}(\gamma'(t))$ points toward the interior in the sense that there exists $\delta > 0$ such that $\gamma(t) + sR_{90}(\gamma'(t))$ lies in the interior for all $s \in (0, \delta)$.

Otherwise, γ is **negatively oriented**, in which case its rotation index equals -1 , and $R_{90}(\gamma'(t))$ points toward the exterior for all $t \in [a, b]$.

The equivalence of these two conditions will follow from ideas in the proofs of Hopf's Umlaufsatz and the Jordan curve theorem (Exercise 2.1).

The curve in Fig. 2.1 hints that Theorem 2.1 is not as obvious as it might at first appear. Although this curve performs many full clockwise and many full counterclockwise turns, most of them cancel each other, leaving a net counterclockwise rotation of one turn. This curve is therefore positively oriented, which is more easily verified by observing that its interior is on its left.

The remainder of this section is devoted to (1) sketching the proofs of these two fundamental theorems, which could be skipped on a first read, and (2) generalizing Hopf's Umlaufsatz to piecewise-regular curves, which is an important prerequisite for Chap. 6.

Recall that the velocity function of a unit-speed closed plane curve $\gamma : [a, b] \rightarrow \mathbb{R}^2$ can be regarded as a function $\mathbf{v} : [a, b] \rightarrow S^1$ with $\mathbf{v}(a) = \mathbf{v}(b)$, where $S^1 = \{(\cos \theta, \sin \theta) \mid \theta \in \mathbb{R}\}$. This viewpoint allowed us in the previous chapter to construct a global angle function $\theta : [a, b] \rightarrow \mathbb{R}$, contrived so that $\mathbf{v}(t) = (\cos \theta(t), \sin \theta(t))$ for all $t \in [a, b]$. From this, we defined the rotation index of the plane curve as $\frac{1}{2\pi}(\theta(b) - \theta(a))$.

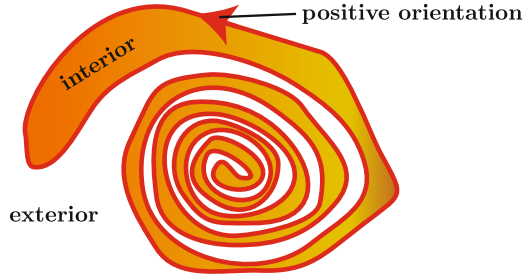


FIGURE 2.1. Perhaps it's not obvious how to prove Theorem 2.1

Now forget about the plane curve, and imagine instead that you began with an arbitrary continuous function from $[a, b]$ to S^1 whose values at a and b agree. Don't assume that it is the velocity function of anything. We claim that the above steps still apply:

PROPOSITION AND DEFINITION 2.3.

If $\mathbf{f} : [a, b] \rightarrow S^1$ is a continuous function with $\mathbf{f}(a) = \mathbf{f}(b)$, then there exists a continuous **angle function** $\theta : [a, b] \rightarrow \mathbb{R}$ such that for all $t \in [a, b]$, we have

$$\mathbf{f}(t) = (\cos \theta(t), \sin \theta(t)).$$

This function is unique up to adding an integer multiple of 2π . The **degree** of \mathbf{f} is defined as the integer $\frac{1}{2\pi} (\theta(b) - \theta(a))$.

If \mathbf{f} is smooth, then the claim follows from Proposition 1.39 (on page 35), since integrating \mathbf{f} yields a unit-speed plane curve whose velocity function is \mathbf{f} . The proof for continuous functions is outlined in Exercise 2.2. In summary, the *degree* of a continuous function $\mathbf{f} : [a, b] \rightarrow S^1$ with $\mathbf{f}(a) = \mathbf{f}(b)$ is an integer that represents roughly the number of times the domain is wrapped counterclockwise around the circle. Notice that the rotation index of a closed plane curve γ (as defined in Exercise 1.55 on page 40) equals the degree of its unit tangent function $t \mapsto \mathbf{t}(t)$.

We will repeatedly use the idea that two functions from $[a, b]$ into S^1 with sufficiently close outputs must have the same degree. In fact, if their outputs never point in opposite directions, then they must have the same degree:

LEMMA 2.4.

Let $\mathbf{f}_1, \mathbf{f}_2 : [a, b] \rightarrow S^1$ be continuous functions with $\mathbf{f}_1(a) = \mathbf{f}_1(b)$ and $\mathbf{f}_2(a) = \mathbf{f}_2(b)$. If \mathbf{f}_1 and \mathbf{f}_2 have different degrees, then $\mathbf{f}_1(t_0) = -\mathbf{f}_2(t_0)$ for some $t_0 \in [a, b]$.

PROOF. Let $\theta_1, \theta_2 : [a, b] \rightarrow \mathbb{R}$ be angle functions for \mathbf{f}_1 and \mathbf{f}_2 . Consider the difference $\delta(t) = \theta_2(t) - \theta_1(t)$. Since the degrees are different,

$$|\delta(b) - \delta(a)| = \underbrace{|\theta_2(b) - \theta_2(a)|}_{2\pi(\text{degree } \mathbf{f}_2)} - \underbrace{|\theta_1(b) - \theta_1(a)|}_{2\pi(\text{degree } \mathbf{f}_1)} \geq 2\pi.$$

Since δ has a net change of at least 2π , there must be an *odd* integer multiple of π between $\delta(a)$ and $\delta(b)$. The intermediate value theorem implies that δ achieves this value for some $t_0 \in [a, b]$, so $\mathbf{f}_1(t_0) = -\mathbf{f}_2(t_0)$. \square

PROOF OF HOPF'S UMLAUFSAZ. Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a simple closed plane curve. Let C denote its trace. Let $\mathbf{p} \in C$ be a point such that C is entirely on one side of the tangent line, L , to C at \mathbf{p} . One can find such a point by considering a circle (centered anywhere in \mathbb{R}^2) with radius large enough to contain C , and then shrinking the radius until the circle first touches C . The point at which it first touches C will have the desired property.

We can assume without loss of generality that γ is parametrized by arc length with $\gamma(a) = \mathbf{p}$. Consider the triangle

$$T = \{(t_1, t_2) \mid a \leq t_1 \leq t_2 \leq b\}.$$

Define the function $\psi : T \rightarrow S^1$ as follows:

$$\psi(t_1, t_2) = \begin{cases} \gamma'(t_1) & \text{if } t_1 = t_2 \\ \frac{\gamma(t_2) - \gamma(t_1)}{|\gamma(t_2) - \gamma(t_1)|} & \text{if } t_1 \neq t_2 \text{ and } \{t_1, t_2\} \neq \{a, b\}, \\ -\gamma'(a) & \text{if } \{t_1, t_2\} = \{a, b\}. \end{cases}$$

For most inputs, $\psi(t_1, t_2)$ is the unit vector pointing in the direction from $\gamma(t_1)$ to $\gamma(t_2)$. The rest of the definition just ensures that ψ is continuous. For example, according to Proposition 1.7 (on page 4), the correct way to extend ψ continuously to a point (t, t) on the hypotenuse of T is $\psi(t, t) = \gamma'(t)$.

Let $\alpha_0 : [0, 1] \rightarrow T$ be a parametrization of the line segment from (a, a) to (b, b) . Let $\alpha_1 : [0, 1] \rightarrow T$ be a parametrization of the line segment from (a, a) to (a, b) followed by the line segment from (a, b) to (b, b) . It is possible to interpolate continuously between α_0 and α_1 by a family of paths, $\alpha_s : [0, 1] \rightarrow T$, $s \in [0, 1]$, each of which goes from (a, a) to (b, b) ; see Fig. 2.2 (left). Here “continuously” means that $(s, t) \mapsto \alpha_s(t)$ is a continuous function from $[0, 1] \times [0, 1]$ to T .

For each $s \in [0, 1]$, let $D(s)$ denote the degree of $\psi \circ \alpha_s : [0, 1] \rightarrow S^1$. Lemma 2.4 can be used to show that $s \mapsto D(s)$ is locally constant and therefore continuous on $[0, 1]$. Since D is integer-valued and continuous, it follows from Proposition A.19 of the appendix (on page 353) that D must be constant on $[0, 1]$, so $D(1) = D(0)$.

By definition, $D(0)$ equals the degree of the unit tangent function of γ , which equals the rotation index of γ . It remains to explain why $D(1)$ equals 1 or -1 . In Fig. 2.2 (right), as α_1 first goes from (a, a) to (a, b) , the path $\psi \circ \alpha_1$ follows the blue vectors, tracing the top half of S^1 counterclockwise. Then as α_1 goes from (a, b) to (b, b) , the path $\psi \circ \alpha_1$ follows the negatives

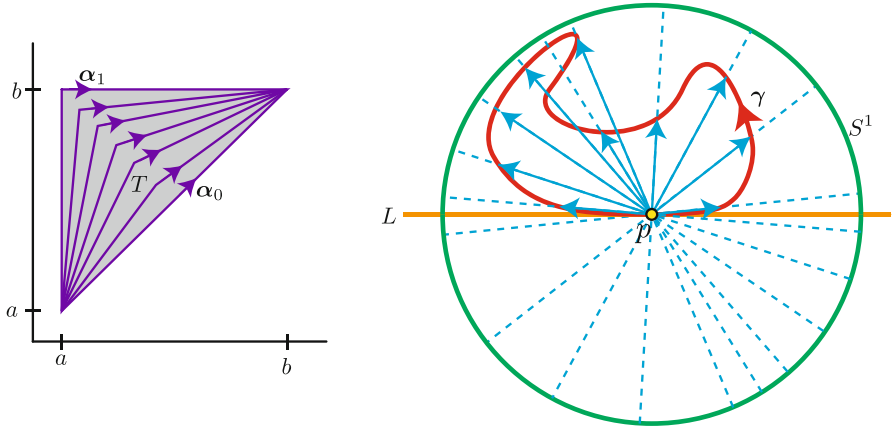


FIGURE 2.2. The proof of Hopf's Umlaufsatz

of the blue vectors, tracing the bottom half of S^1 counterclockwise. Thus, $D(1) = 1$. If γ had the other orientation, then $D(1) = -1$. This completes the proof.¹ \square

For the next proof, we require the idea of a *tubular neighborhood*. Suppose that $\gamma : [a, b] \rightarrow \mathbb{R}^2$ is a simple closed plane curve. For small $\epsilon > 0$, consider the function $\varphi : (-\epsilon, \epsilon) \times [a, b] \rightarrow \mathbb{R}^2$ defined as

$$\varphi(s, t) = \gamma(t) + s \cdot R_{90}(\mathbf{v}(t)).$$

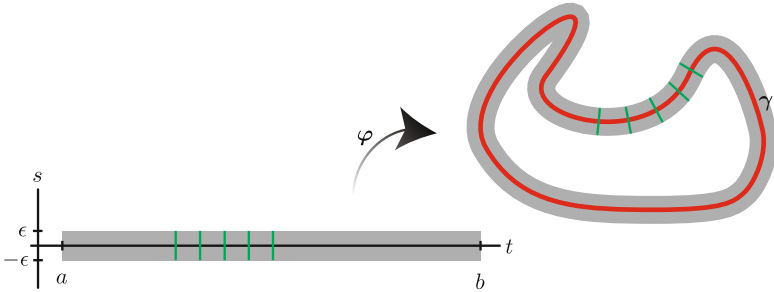


FIGURE 2.3. A tubular neighborhood

For fixed t_0 , the function $s \mapsto \varphi(t_0, s)$ parametrizes a small line segment that crosses the trace of γ orthogonally; in Fig. 2.3, it is shown in green for five choices of t_0 . The important fact is that these green lines do not intersect each other. In other words, we have the following:

PROPOSITION 2.5.

For sufficiently small $\epsilon > 0$, φ is injective.

¹To help visualize the function $\psi \circ \alpha_s$ and the degree of this function, a nice animation is available at <http://www.mathematik.com/Hopf/index.html>

We will postpone the proof of this claim until Exercise 3.11, after we discuss the inverse function theorem. The phenomenon is both local and global. The value ϵ must be chosen small enough to ensure that the green lines remain disjoint locally as the curve bends sharply, and also globally as the curve loops back close to itself. The image of φ is called a **tubular neighborhood** of γ .

If the trace of γ is removed from the tubular neighborhood, then what is left has two path-connected components, namely $\varphi((-\epsilon, 0) \times [a, b])$ and $\varphi((0, \epsilon) \times [a, b])$. Each is path-connected, because φ identifies it with a rectangle.

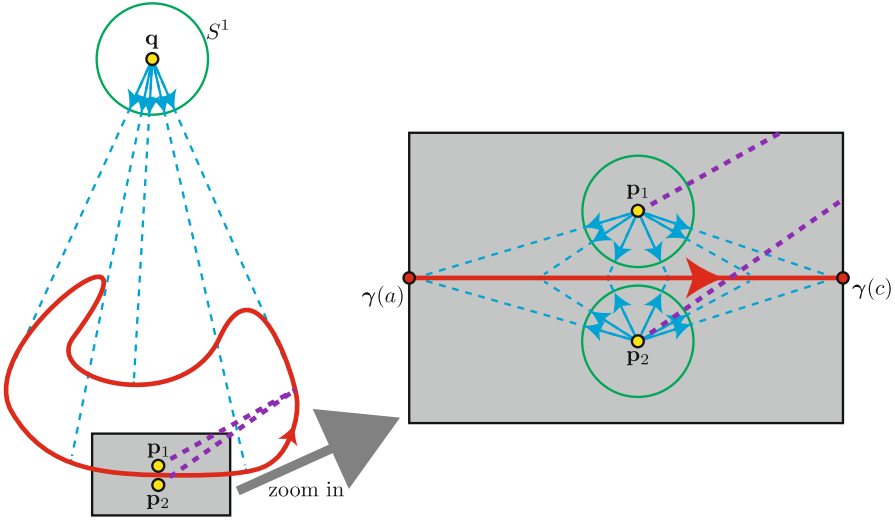
PROOF OF THE JORDAN CURVE THEOREM. Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a simple closed plane curve. Let C denote its trace. For any $\mathbf{p} \in \mathbb{R}^2 - C$, consider the function $\mathbf{f}_{\mathbf{p}} : [a, b] \rightarrow S^1$ defined as

$$\mathbf{f}_{\mathbf{p}}(t) = \frac{\gamma(t) - \mathbf{p}}{|\gamma(t) - \mathbf{p}|}.$$

Let $W(\mathbf{p})$ denote the degree of $\mathbf{f}_{\mathbf{p}}$. Intuitively, if you stand at \mathbf{p} while keeping your finger pointing at a friend who traverses C , then $W(\mathbf{p})$ is the net number of counterclockwise rotations that this activity forces you to perform. Lemma 2.4 can be used to verify that W is locally constant and therefore continuous on its domain $\mathbb{R}^2 - C$. It is constant on every path-connected component of this domain, because along every path in the domain, W changes continuously but is also integer-valued, so Proposition A.19 from the appendix (on page 353) implies that it must be constant. Our aim is to show that $\mathbb{R}^2 - C$ has exactly two path-connected components, one on which $W = 0$ and the other on which $W = 1$ or $W = -1$ (depending on the orientation of γ).

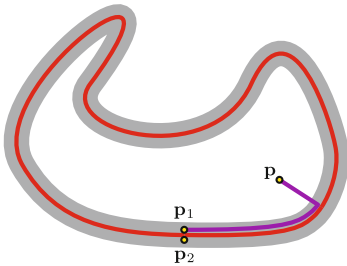
If \mathbf{q} is sufficiently far from C , like the point labeled “ \mathbf{q} ” in Fig. 2.4, then $\mathbf{f}_{\mathbf{q}}$ is not surjective, because its image is constrained to an arc of S^1 , so Lemma 2.4 implies that $W(\mathbf{q})$ equals the degree of a constant function, which is 0. We will demonstrate next that W also attains at least one nonzero value, so $\mathbb{R}^2 - C$ has at least two connected components.

Imagine zooming in with sufficient magnification at a point of C so that the zooming window lies within a tubular neighborhood, and C is well approximated by its tangent line within this window. Choose a pair of points $\mathbf{p}_1, \mathbf{p}_2$ in this window that are close to each other but are on opposite sides of C . We claim that the window and the points can be chosen such that $|W(\mathbf{p}_1) - W(\mathbf{p}_2)| = 1$. To prove this, let ϵ denote any quantity that approaches zero as the zooming window shrinks and as $|\mathbf{p}_1 - \mathbf{p}_2|$ becomes small relative to the size of the window. Just to make the discussion more specific, we assume that the curve is positioned and oriented as in Fig. 2.4 (with C horizontal traversed left to right, \mathbf{p}_1 above, and \mathbf{p}_2 below) and is parametrized so that $\gamma(t)$ lies in this window for time parameters $t \in [a, c]$ (where $a < c < b$). Restricted to $[a, c]$, both $\mathbf{f}_{\mathbf{p}_1}$ and $\mathbf{f}_{\mathbf{p}_2}$ begin within distance ϵ of $(-1, 0)$ and end within distance ϵ of $(1, 0)$; $\mathbf{f}_{\mathbf{p}_1}$ approximately covers the bottom half of

FIGURE 2.4. $W(\mathbf{q}) = 0$, while $|W(\mathbf{p}_1) - W(\mathbf{p}_2)| = 1$

S^1 counterclockwise, while $\mathbf{f}_{\mathbf{p}_2}$ approximately covers the top half of S^1 clockwise. Furthermore, for $t \in [c, b]$ we have $|\mathbf{f}_{\mathbf{p}_1}(t) - \mathbf{f}_{\mathbf{p}_2}(t)| < \epsilon$, as indicated by the dashed purple lines in Fig. 2.4.

We will perform some small perturbations to $\mathbf{f}_{\mathbf{p}_1}$ and $\mathbf{f}_{\mathbf{p}_2}$ that (by Lemma 2.4) do not alter the degree of either function. First, we can modify both functions on $[a, c]$ so that they begin exactly at $\mathbf{f}_{\mathbf{p}_1}(a) = \mathbf{f}_{\mathbf{p}_2}(a) = (-1, 0)$ and end exactly at $\mathbf{f}_{\mathbf{p}_1}(c) = \mathbf{f}_{\mathbf{p}_2}(c) = (1, 0)$. Next, we can redefine $\mathbf{f}_{\mathbf{p}_2}$ to equal $\mathbf{f}_{\mathbf{p}_1}$ on $[c, b]$. After these modifications, their degrees differ by one. To understand why, imagine traversing $\mathbf{f}_{\mathbf{p}_1}$ followed by the reverse orientation of $\mathbf{f}_{\mathbf{p}_2}$. This path, denoted by $\mathbf{f}_{\mathbf{p}_1} - \mathbf{f}_{\mathbf{p}_2}$, traverses the bottom half of S^1 counterclockwise, then does something else, then does that same something else in reverse, then traverses the top half of S^1 counterclockwise. The net result is one counterclockwise rotation. Since the degree of $\mathbf{f}_{\mathbf{p}_1} - \mathbf{f}_{\mathbf{p}_2}$ equals 1, it follows that $\text{degree}(\mathbf{f}_{\mathbf{p}_1}) - \text{degree}(\mathbf{f}_{\mathbf{p}_2}) = 1$.

FIGURE 2.5. Every $\mathbf{p} \in \mathbb{R}^2 - C$ can be joined to either \mathbf{p}_1 or \mathbf{p}_2 with a path that avoids C

We now know that \mathbf{p}_1 and \mathbf{p}_2 are in different path-connected components of $\mathbb{R}^2 - C$. We claim that these are the only components. In other words, every other point $\mathbf{p} \in \mathbb{R}^2 - C$ can be connected to either \mathbf{p}_1 or \mathbf{p}_2 by a continuous path in $\mathbb{R}^2 - C$. To see this, choose a shortest path from \mathbf{p} to C . Before reaching C , this path will reach a fixed tubular neighborhood of C , inside of which it can be connected to \mathbf{p}_1 or \mathbf{p}_2 ; see Fig. 2.5.

Thus, $\mathbb{R}^2 - C$ has exactly two path-connected components. Let $B \subset \mathbb{R}^2$ denote any ball large enough to contain C . Clearly, one component of $\mathbb{R}^2 - C$ contains the complement of B , and is therefore unbounded, so the other component is contained in B and is therefore bounded. \square

The remainder of this section is devoted to generalizing its main theorems. The Jordan curve theorem remains true if $\gamma : [a, b] \rightarrow \mathbb{R}^2$ is only a *continuous* function with $\gamma(a) = \gamma(b)$ that is one-to-one on $[a, b)$, but the proof is more difficult in this setting.

Hopf's Umlaufsatz does not make sense when γ is only continuous, but it can at least be generalized to *piecewise-regular* curves:

DEFINITION 2.6.

A **piecewise-regular curve** in \mathbb{R}^n is a continuous function $\gamma : [a, b] \rightarrow \mathbb{R}^n$ with a partition, $a = t_0 < t_1 < \dots < t_n = b$, such that the restriction, γ_i , of γ to each subinterval $[t_i, t_{i+1}]$ is a regular curve. It is called **closed** if additionally $\gamma(a) = \gamma(b)$, and **simple** if γ is one-to-one on the domain $[a, b)$. It is said to be **of unit speed** if each γ_i is of unit speed.

In other words, there might be finitely many times at which γ is only continuous but not smooth. The definition of “closed” does not require the derivatives of γ to agree at a and b ; this allows the possibility that $t = a$ might correspond to one of the nonsmooth points.

A piecewise-regular simple closed *plane* curve γ is called **positively oriented** if $R_{90}(\gamma'(t))$ points toward the interior for all values of t that correspond to smooth points (all values except the partition endpoints), as in Fig. 2.6. Otherwise, it is called **negatively oriented**.

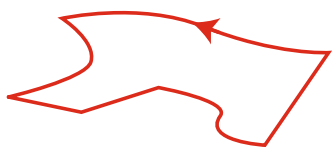


FIGURE 2.6. A positively oriented piecewise-regular simple closed plane curve

Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a piecewise-regular *plane* curve with partition denoted by $a = t_0 < t_1 < \dots < t_n = b$. Each nonsmooth point $\gamma(t_i)$ is called a **corner** of γ . At this corner, there are *two* velocity vectors, coming from the left- and right-hand limits:

$$\mathbf{v}^-(t_i) = \lim_{h \rightarrow 0^-} \frac{\gamma(t_i + h) - \gamma(t_i)}{h} = \lim_{t \rightarrow t_i^-} \gamma'(t),$$

$$\mathbf{v}^+(t_i) = \lim_{h \rightarrow 0^+} \frac{\gamma(t_i + h) - \gamma(t_i)}{h} = \lim_{t \rightarrow t_i^+} \gamma'(t).$$

By the regularity hypothesis, neither is zero. The **signed angle** at $\gamma(t_i)$, denoted by $\alpha_i \in [-\pi, \pi]$, is defined such that its absolute value equals the smallest determination of the angle between $\mathbf{v}^-(t_i)$ and $\mathbf{v}^+(t_i)$. The sign of

α_i is defined to be positive if $\mathbf{v}^+(t_i)$ is a counterclockwise rotation of $\mathbf{v}^-(t_i)$ through this angle (and to be negative if it is clockwise); see Fig. 2.7. Notice that reversing the orientation of γ would change the sign of the signed angle at each corner, but would not affect the absolute value.

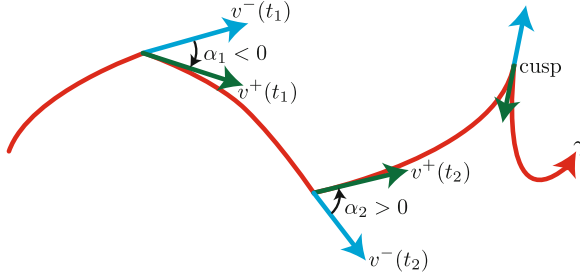


FIGURE 2.7. A piecewise-regular plane curve with three corners

The corner $\gamma(t_i)$ is called a **cusp** if $\mathbf{v}^+(t_i)$ is a negative scalar multiple of $\mathbf{v}^-(t_i)$. The specification of whether the signed angle at a cusp equals π or $-\pi$ is easiest to describe when γ is simple, closed, and positively oriented. Under these added hypotheses, $\alpha_i = \pi$ if $\mathbf{v}^-(t_i)$ points toward the exterior, or $\alpha_i = -\pi$ if $\mathbf{v}^-(t_i)$ points toward the interior; see Fig. 2.8. The sign convention is the opposite if γ is negatively oriented.

If γ is closed and $\gamma'(a) \neq \gamma'(b)$, then $\gamma(a)$ counts as a corner, and the corresponding signed angle is defined exactly as above, but with $\mathbf{v}^-(a)$ replaced by $\mathbf{v}^-(b)$.

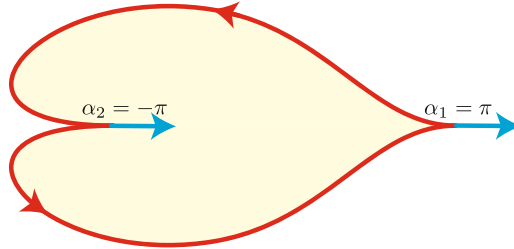


FIGURE 2.8. The sign convention at a cusp

THEOREM 2.7 (Generalized Hopf's Umlaufsatz).

Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a unit-speed positively oriented piecewise-regular simple closed plane curve. Let κ_s denote its signed curvature function, and let $\{\alpha_i\}$ be the list of signed angles at its corners. Then

$$\int_a^b \kappa_s(t) dt + \sum_i \alpha_i = 2\pi.$$

Here “ $\int_a^b \kappa_s(t) dt$ ” is shorthand for $\sum_i \left(\int_{t_i}^{t_{i+1}} \kappa_s(t) dt \right)$, which means the sum of the integral of κ_s over the smooth segments of γ .

If γ is regular (no corners) and θ denotes a global angle function of γ , recall from Sect. 6 of Chap. 1 that $\kappa_s = \theta'$, so

$$\int_a^b \kappa_s(t) dt = \int_a^b \theta'(t) dt = \theta(b) - \theta(a) = 2\pi \cdot (\text{rotation index}).$$

So in this case, Theorem 2.7 says that the rotation index equals 1, which we knew from Theorem 2.1.

When γ has corners, it is still possible to define an “angle function” θ that has a jump discontinuity at each corner by an amount equal to the corresponding signed angle, and that elsewhere satisfies $\kappa_s = \theta'$. The expression $\int_a^b \kappa_s(t) dt + \sum_i \alpha_i$ equals the net change in this (discontinuous) angle function. The proof of Theorem 2.7 involves smoothing the corners so that this angle function becomes continuous:

PROOF IDEA. The visual idea of the proof is to smooth γ in neighborhoods of the corners, as illustrated in Fig. 2.9. If $\tilde{\gamma}$ denotes a smoothed version of γ (in which neighborhoods of the corners have been replaced with the dashed lines shown in the figure), and $\tilde{\kappa}_s$ is the signed curvature function of $\tilde{\gamma}$, then our original version of Hopf’s Umlaufsatz says that $\int_a^b \tilde{\kappa}_s(t) dt = 2\pi$. Although we will not discuss the analytic details, it is visually believable that the smoothing can be constructed such that

$$\int_a^b \kappa_s(t) dt + \sum_i \alpha_i = \int_a^b \tilde{\kappa}_s(t) dt = 2\pi.$$

□

The above proof helps explain our previous definition of the signed angle at a corner. The definition was essentially contrived to match the net change in the angle function after smoothing; in other words, α_i is the net counterclockwise rotation of the smoothed corner (interpreted as clockwise if α_i is negative), in the limit as the smoothing occurs in a smaller and smaller neighborhood of the corner. This description applies equally well at a cusp.

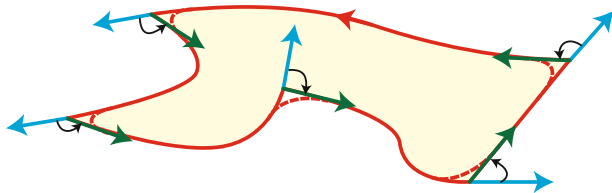


FIGURE 2.9. Hopf’s Umlaufsatz is generalized by smoothing the curve at the corners

It is sometimes convenient to rephrase the previous theorem in terms of *interior angles*. Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a piecewise-smooth simple closed plane curve with signed angles denoted by $\{\alpha_i\}$. The i th **interior angle** of γ , denoted by $\beta_i \in [0, 2\pi]$, is defined as in Fig. 2.10. Notice that changing the orientation of γ would change the sign of each signed angle, but would not affect the interior angles. Interior angles are related to signed angles as follows:

$$\beta_i = \begin{cases} \pi - \alpha_i & \text{if } \gamma \text{ is positively oriented,} \\ \pi + \alpha_i & \text{if } \gamma \text{ is negatively oriented.} \end{cases}$$

In Fig. 2.8, for example, $\beta_1 = 0$ and $\beta_2 = 2\pi$.

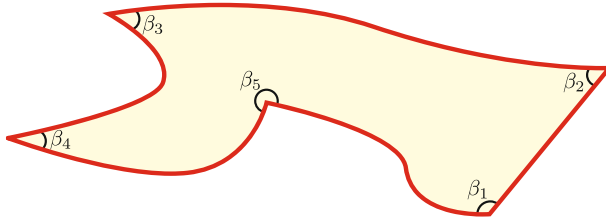


FIGURE 2.10. Interior angles

In Theorem 2.7, γ is assumed to be positively oriented, so the theorem becomes

$$\int_a^b \kappa_s(t) dt = \sum_i \beta_i - (n-2)\pi,$$

where n is the number of corners. If the smooth segments of γ are straight line segments, then this becomes

$$\sum_i \beta_i = (n-2)\pi,$$

which is a well-known formula for the sum of the interior angles of a polygon.

EXERCISES

EXERCISE 2.1. Prove that the two conditions in Definition 2.2 are equivalent, as claimed. *HINT:* Using the existence of a tubular neighborhood, prove that $R_{90}(\gamma'(t))$ either points to the interior for all $t \in [a, b]$ or points to the exterior for all $t \in [a, b]$, so it suffices to consider a single value of t . Choose the value corresponding to the point labeled “p” in Fig. 2.2.

EXERCISE 2.2. Prove Proposition 2.3. *HINT:* Use a compactness argument to divide $[a, b]$ into finitely many subintervals, on each of which the image of \mathbf{f} is completely contained in one of the following four half-circles: top, bottom, right, left. Define a local angle function on each subinterval.

Working from left to right, add the correct integer multiple of 2π to each local angle function so they match to form a global angle function.

EXERCISE 2.3. Let $\mathbf{f} : [a, b] \rightarrow S^1$ be a continuous function with $\mathbf{f}(a) = \mathbf{f}(b)$, and let $\mathbf{p} \in S^1$. If the degree of \mathbf{f} equals n , what is the minimal possible size of the set $\{t \in [a, b] \mid \mathbf{f}(t) = \mathbf{p}\}$?

□

2. Convexity and the Four Vertex Theorem (Optional)

In this section, we describe one of the earliest global results in differential geometry, which provides a restriction on the number of vertices of a simple closed plane curve.

DEFINITION 2.8.

Let $\gamma : I \rightarrow \mathbb{R}^2$ be a regular plane curve. A point, $\gamma(t)$, on its trace is called a **vertex** if the signed curvature function has a local maximum or local minimum at t .

This definition is independent of parametrization (Exercise 2.5). As calculated in Exercise 2.4, an ellipse has exactly four vertices, two at which κ_s is maximal and two at which κ_s is minimal. The polar coordinate graph of $r = 1 - 2\sin(\theta)$ has exactly two vertices; see Fig. 2.11. Two is the smallest number of vertices that a closed curve could have, because the signed curvature function must achieve its global maximum and global minimum on its compact domain (according to Corollary A.25 on page 356 of the appendix). Notice that every point of a circle qualifies as a vertex that is both a local maximum and a local minimum, so a circle has infinitely many vertices. For the same reason, so does a straight line.

The main theorem of this section says that a *simple* closed plane curve must have at least as many vertices as an ellipse has.

THEOREM 2.9 (The Four Vertex Theorem).

Every simple closed plane curve has at least four vertices.

This section is devoted to the proof of this theorem, but only in the special in which the curve is *convex*:

DEFINITION 2.10.

A simple closed plane curve is called **convex** if its trace lies entirely on one closed side of each of its tangent lines.

The term “closed side” means that the tangent line itself is considered part of either “side” into which it divides the plane, since of course the tangent line at \mathbf{p} intersects the trace at \mathbf{p} (and possibly also at nearby points if the trace is a straight line segment in a neighborhood of \mathbf{p}); see Fig. 2.12.

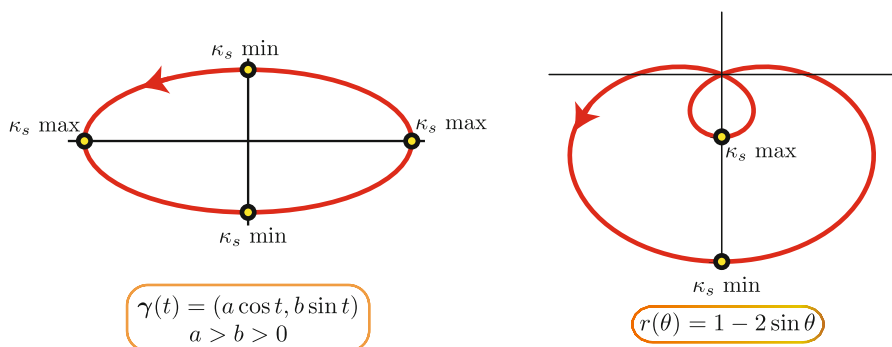


FIGURE 2.11. An ellipse has four vertices (*left*), while a nonsimple curve can have two (*right*)

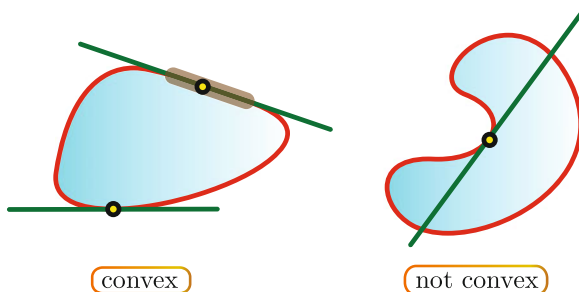


FIGURE 2.12. A convex curve lies on one closed side of each of its tangent lines

We'll require the following consequence of convexity:

LEMMA 2.11.

Let C be the trace of a simple closed convex plane curve, and let L be a line.

- (1) If L is tangent to C at two distinct points, then C contains the entire segment of L between these two points.
- (2) If $C \cap L$ contains more than two points, then it contains the entire segment of L between any pair of these points.

PROOF. Part (1) is left to the reader in Exercise 2.7 (with hints). For part (2), suppose that $L \cap C$ contains at least three points, and order them $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$ so that \mathbf{p}_2 is between \mathbf{p}_1 and \mathbf{p}_3 along L . Notice that the tangent line to C at \mathbf{p}_2 must equal L , for otherwise, the other two points would lie on opposite sides of this tangent line. Since L is now a tangent line, convexity implies that C cannot cross L at either of the other two points, so L must be tangent to C at all three points. The result now follows from part (1). \square

The final ingredient that we'll require for the proof is a formulation of the familiar definition of signed curvature in terms of the separate x - and y -components of the curve. For a regular plane curve $\gamma(t) = (x(t), y(t))$, recall that

$$\mathbf{v}(t) = (x'(t), y'(t)), \quad \mathbf{a}(t) = (x''(t), y''(t)), \quad R_{90}(\mathbf{v}(t)) = (-y'(t), x'(t)).$$

When γ is of unit speed, its signed curvature is defined by $\mathbf{a}(t) = \kappa_s(t)R_{90}(\mathbf{v}(t))$ (Eq. 1.9 on page 33). The separate x - and y -components of this equation are

$$(2.1) \quad x''(t) = -\kappa_s(t)y'(t), \quad y''(t) = \kappa_s(t)x'(t).$$

PROOF OF THE FOUR VERTEX THEOREM FOR CONVEX CURVES. Let γ be a simple closed convex plane curve and let C denote its trace. Assume without loss of generality that γ is positively oriented. We can assume that κ_s is not constant on any interval, since every element of such an interval would correspond to a vertex, so C would have infinitely many vertices. In particular, we can assume that no segment of C is a straight line segment.

If γ had exactly three vertices, then two consecutive ones along C would be of the same type (both local maxima or both local minima). But this is impossible, because a non-locally-constant smooth real-valued function cannot have two consecutive local extrema of the same type.

Now suppose that γ has fewer than three vertices; in other words, its only vertices are the point $\mathbf{p} \in C$ at which κ_s attains its global maximum and the point $\mathbf{q} \in C$ at which κ_s attains its global minimum. Notice that $\mathbf{p} \neq \mathbf{q}$, for otherwise, κ_s would be constant. Let L denote the line through \mathbf{p} and \mathbf{q} . Lemma 2.11(2) implies that L intersects C only at \mathbf{p} and \mathbf{q} .

Choose an orientation-preserving unit-speed parametrization, $\gamma : [0, l] \rightarrow \mathbb{R}^2$, such that $\gamma(0) = \gamma(l) = \mathbf{p}$ and $\gamma(a) = \mathbf{q}$ for some $a \in (0, l)$. Notice that $\kappa'_s \leq 0$ on $(0, a)$ (since κ_s decreases from its maximum to its minimum) and $\kappa'_s \geq 0$ on (a, l) (since κ_s increases from its minimum to its maximum). We can assume without loss of generality (by applying a proper rigid motion) that \mathbf{p} is the origin and L is the x -axis.

Write γ in terms of its component functions: $\gamma(t) = (x(t), y(t))$. Notice that $y(t)$ changes sign only at $t = a$, since γ lies above L on $(0, a)$ and below L on (a, l) , or possibly vice versa, depending on whether \mathbf{p} lies to the right or left of \mathbf{q} . In either case, notice that $y(t)\kappa'_s(t)$ never changes sign; this expression is either ≤ 0 on all of $[0, l]$ or it is ≥ 0 on all of $[0, l]$; see Fig. 2.13.

Furthermore, the expression $y(t)\kappa'_s(t)$ equals zero only when $\kappa'_s = 0$, which does not occur on any interval of nonzero length. Thus, this expression has a nonzero average value:

$$\int_0^l y(t)\kappa'_s(t) dt \neq 0.$$

However, integrating by parts and using Eq. 2.1 together with the fact that the functions involved are periodic (they have the same values at 0 and l), we get

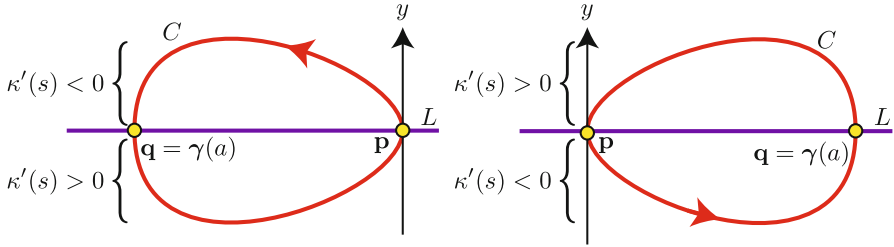


FIGURE 2.13. Either $y(t)\kappa'_s(t) \leq 0$ (left), or $y(t)\kappa'_s(t) \geq 0$ (right)

$$\int_0^l y(t)\kappa'_s(t) dt = y(t)\kappa_s(t) \Big|_{t=0}^{t=l} - \int_0^l \kappa_s(t)y'(t) dt = 0 + \int_0^l x''(t) dt = x'(t) \Big|_{t=0}^{t=l} = 0.$$

This contradiction shows that γ must have at least four vertices. \square

Since the concept of convexity is of independent importance, we end this section with some equivalent formulations of its definition:

PROPOSITION 2.12.

Let γ be a simple closed plane curve. Let C denote its trace. Let \mathcal{I} denote its interior. The following are equivalent:

- (1) γ is convex; that is, C lies on one closed side of each of its tangent lines.
- (2) The line segment joining any two points of \mathcal{I} lies entirely in \mathcal{I} ; see Fig. 2.14
- (3) κ_s does not change sign; that is, either $\kappa_s \geq 0$ on the whole domain or $\kappa_s \leq 0$ on the whole domain.

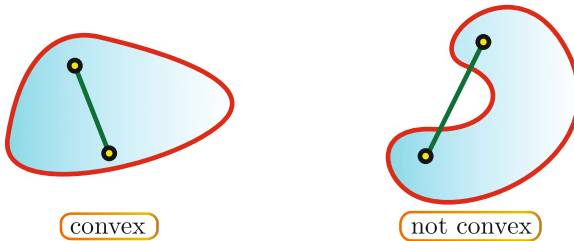


FIGURE 2.14. The line segment joining any two points of the interior of a convex curve must lie entirely in the interior

PROOF. $(1) \implies (2)$ (by contradiction): Suppose that γ is convex yet there is a pair $\mathbf{p}, \mathbf{q} \in \mathcal{I}$ such that the line segment joining them does *not* lie in \mathcal{I} . The (infinite) line, L , containing \mathbf{p} and \mathbf{q} must intersect C in at least three points (colored purple in Fig. 2.15). Lemma 2.11(2) implies that C contains the corresponding segment of L , so $\mathbf{p}, \mathbf{q} \in C$, contradicting the fact that they are interior points.

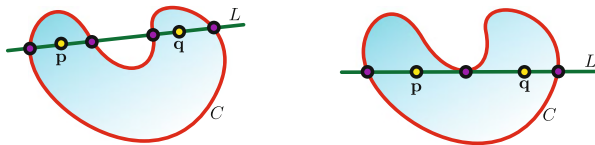


FIGURE 2.15. $L \cap C$ could contain four points (left) or three points (right), but no fewer

(2) \implies (1) Exercise 2.9.

(3) \implies (1) (by contradiction): Assume that κ_s does not change sign yet C lies on both sides of its tangent line at some point $\mathbf{p} = \gamma(t_0)$. The vector $\mathbf{n} = -R_{90}(\mathbf{v}(t_0))$ is orthogonal to C at \mathbf{p} . Consider the function

$$h(t) = \langle \gamma(t) - \gamma(t_0), \mathbf{n} \rangle.$$

Intuitively, $h(t)$ is the “height” of $\gamma(t)$ above the tangent line to C at $\gamma(t_0)$, with \mathbf{n} considered the “up” direction. Notice that $h(t_0) = 0$. Since C lies on both sides of the tangent line, h attains positive and negative values, so its global minimum and maximum occur at time values, called t_1 and t_2 respectively, that are distinct from each other and from t_0 . It is straightforward to show that the velocity vectors $\{\mathbf{v}(t_0), \mathbf{v}(t_1), \mathbf{v}(t_2)\}$ are mutually parallel; these velocity vectors are purple in Fig. 2.16.

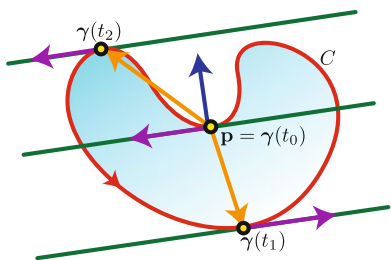


FIGURE 2.16. If C lies on both sides of its tangent line at \mathbf{p} , then the three purple velocities vectors are parallel

constant (and hence C is a straight line segment) on the segment of C between some pair of $\{\gamma(t_0), \gamma(t_1), \gamma(t_2)\}$. But this contradicts the fact that h has different values at all three points.

(1) \implies (3) (by contradiction): Assume that γ is convex yet $\kappa_s = \theta'$ changes sign. Then it is possible to choose nearby times $t_1 \neq t_2$, between which θ' changes sign, such that $\theta(t_1) = \theta(t_2)$, which means that $\mathbf{v}(t_1)$ and $\mathbf{v}(t_2)$ point in the same direction. By Hopf's Umlaufsatz, there exists a time t_3 such that $\mathbf{v}(t_3)$ points in the opposite of this direction. The tangent lines to C at the three points $\{t_1, t_2, t_3\}$ are parallel. If these three tangent

Thus, some pair of these three velocity vectors must point in the same direction (rather than opposite directions). It doesn't matter which pair, so let's say that $\mathbf{v}(t_0)$ and $\mathbf{v}(t_2)$ point in the same direction, as in the figure. This means that a global angle function, θ , changes by an integer multiple of 2π between times t_0 and t_2 . By Hopf's Umlaufsatz, θ changes by exactly $\pm 2\pi$ on the entire domain. Since $\kappa_s = \theta'$ does not change sign, θ does all of this changing monotonically. This is possible only if θ is constant

lines were all distinct, then the middle one would contradict convexity, so some two of them must coincide (here “middle” means with respect to their positions as subsets of \mathbb{R}^2). Lemma 2.11 implies that the trace of γ is a straight line segment between these times. But γ can't be straight between t_1 and t_2 , because θ' changes sign between them. Nor can γ be straight between t_3 and either other time, since θ changes by π between them. This is a contradiction. \square

EXERCISES

EXERCISE 2.4. Suppose $p > q > 0$ and consider the ellipse $\gamma(t) = (p \cos(t), q \sin(t))$. The *foci* of this ellipse are the two points on the x -axis with x -coordinates $\pm\sqrt{p^2 - q^2}$, colored purple in Fig. 2.17.

- (1) Prove that the sum of the distances from $\gamma(t)$ to these two foci is independent of t .
- (2) Prove that the signed curvature function of the ellipse is

$$\kappa_s(t) = \frac{pq}{(p^2 \sin^2(t) + q^2 \cos^2(t))^{\frac{3}{2}}}.$$

- (3) Prove that the critical points of the signed curvature function occur at $t \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$. The corresponding points on the ellipse are its intersections with the x - and y -axes.

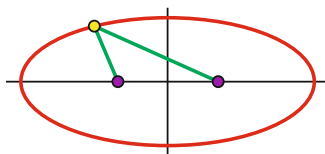


FIGURE 2.17. An ellipse is the set of points with constant summed distance to its two foci

EXERCISE 2.5. Prove that the definition of *vertex* is independent of parametrization.

EXERCISE 2.6. Let $f : (a, b) \rightarrow \mathbb{R}$ be a smooth function, and let $\gamma(t) = (t, f(t))$ be the natural parametrization of its graph. *Prove or disprove:*

- (1) A critical point of f is a vertex of γ .
- (2) A vertex of γ is a critical point of f .

EXERCISE 2.7. Prove Lemma 2.11(1). *HINT: Let $\mathbf{p}_1, \mathbf{p}_2$ denote the two points at which L is tangent to C . Let \mathbf{s} denote the last point of L past \mathbf{p}_1 that is contained in C (which could be \mathbf{p}_1 itself). If \mathbf{s} comes before \mathbf{p}_2 , show that just past \mathbf{s} , there would be a point, \mathbf{q} , such that \mathbf{p}_1 and \mathbf{p}_2 lie on different sides of its tangent line, contradicting convexity; see Fig. 2.18.*

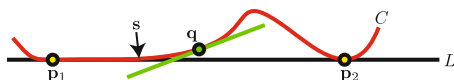


FIGURE 2.18. Moving away from p_1 toward p_2 along L , as soon as C were to separate from L , there would be a point $q \in C$ with p_1 and p_2 on different sides of its tangent line. Thus, C cannot separate from L

EXERCISE 2.8. In the proof of the four vertex theorem, we chose an orientation-preserving unit-speed reparametrization and a proper rigid motion. Why did the reparametrization need to be orientation-preserving? Why did the rigid motion need to be proper?

EXERCISE 2.9. Prove (2) \implies (1) in Proposition 2.12.

EXERCISE 2.10. Let γ be a closed plane curve whose signed curvature does not change sign (either $\kappa_s \geq 0$ on its whole domain or $\kappa_s \leq 0$ on its whole domain). If the rotation index of γ equals ± 1 , prove that γ is simple.

EXERCISE 2.11 (Convexity for a Piecewise-Regular Curve). Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a positively oriented piecewise-regular simple closed plane curve, with interior denoted by \mathcal{I} . Prove that the following are equivalent characterizations of what it means for γ to be **convex**:

- (1) The line segment joining any two points of \mathcal{I} lies entirely in \mathcal{I} .
- (2) $\kappa_s \geq 0$ and all signed angles are positive.

EXERCISE 2.12. Describe the history of the four vertex theorem and its converse. Discuss the ideas behind the proof in the nonconvex case. An excellent reference is [4].

□

3. Fenchel's Theorem (Optional)

It is natural to consider the total amount that a curve curves, measured as follows:

DEFINITION 2.13.

The **total curvature** of a regular curve $\gamma : [a, b] \rightarrow \mathbb{R}^n$ is defined as

$$\text{total curvature} = \int_a^b \kappa(t) |\gamma'(t)| dt.$$

The total curvature of γ is unchanged by reparametrization, so we will usually assume that γ is of unit speed, in which case its total curvature is $\int_a^b \kappa(t) dt$.

In order for a curve to be closed, it must return to where it started. How much total curvature does this require? The answer for a simple closed plane curves is a quick consequence of some previous theorems:

LEMMA 2.14.

The total curvature of a simple closed plane curve is $\geq 2\pi$, with equality if and only if it is convex.

PROOF. Let γ be a unit-speed simple closed plane curve. The integral of the *signed* curvature comes from Hopf's Umlaufsatz:

$$\int_a^b \kappa_s(t) dt = \int_a^b \theta'(t) dt = \theta(b) - \theta(a) = 2\pi \cdot (\text{rotation index}) = \pm 2\pi.$$

This is related to the integral of the *unsigned* curvature as follows:

$$\int_a^b \kappa(t) dt = \int_a^b |\kappa_s(t)| dt \geq \left| \int_a^b \kappa_s(t) dt \right| = 2\pi,$$

with equality if and only if κ_s does not change sign, which by Proposition 2.12 occurs if and only if γ is convex. \square

The goal of this section is to prove the following generalization to curves in \mathbb{R}^n :

THEOREM 2.15 (Fenchel's Theorem).

The total curvature of a closed curve in \mathbb{R}^n is $\geq 2\pi$, with equality if and only if it is a simple closed convex curve contained in a plane in \mathbb{R}^n (which means a translate of a two-dimensional subspace of \mathbb{R}^n).

The term “convex” was previously defined only for curves in \mathbb{R}^2 , but it also makes perfect sense for a curve contained in an arbitrary plane. For simplicity, we will prove Fenchel's theorem in the case $n = 3$. The general case follows from essentially the same argument.

We will require some vocabulary and facts related to the geometry of the sphere:

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$

A “curve in S^2 ” means a space curve whose trace is a subset of S^2 . A “great circle” means the intersection of S^2 with a two-dimensional subspace of \mathbb{R}^3 . For example, the equator $E = \{(x, y, 0) \mid x^2 + y^2 = 1\}$ is a great circle. If $\mathbf{p}, \mathbf{q} \in S^2$, then $\overline{\mathbf{p}\mathbf{q}}$ will denote their intrinsic distance in S^2 , which means the smallest possible arc length of a regular curve in S^2 between \mathbf{p} and \mathbf{q} . We will require the following fact:

LEMMA 2.16.

If $\mathbf{p}, \mathbf{q} \in S^2$ is a pair of distinct points, then $\overline{\mathbf{p}\mathbf{q}} \leq \pi$, with equality if and only if $\mathbf{p} = -\mathbf{q}$. There exists a segment of a great circle from \mathbf{p} to \mathbf{q} with arc length $\overline{\mathbf{p}\mathbf{q}}$, and this segment is unique if $\mathbf{p} \neq -\mathbf{q}$. The trace of every curve in S^2 from \mathbf{p} to \mathbf{q} with arc length $\overline{\mathbf{p}\mathbf{q}}$ is a segment of a great circle.

This lemma will be easily proven when we discuss shortest paths in general curved surfaces in Chap. 5. For now, we'll assume that it is likely familiar to most readers. It essentially just says that segments of great circles are the unique shortest paths in S^2 , as all transatlantic pilots know.

Most of the work of proving Fenchel's theorem lies in proving the following fact:

LEMMA 2.17.

Let β be a regular closed curve in S^2 . If the trace of β intersects every great circle of S^2 , then the arc length of β is $\geq 2\pi$, with equality if and only if β is a simple parametrization of a great circle.

PROOF OF LEMMA 2.17. Assume without loss of generality that $\beta : [0, l] \rightarrow S^2$ is parametrized by arc length. Define $\mathbf{p} = \beta(0) = \beta(l)$ and $\mathbf{q} = \beta(l/2)$. Let β_1, β_2 denote the restrictions of β to the domains $[0, l/2]$ and $[l/2, l]$ respectively.

If $\mathbf{p} = -\mathbf{q}$, then $\overline{\mathbf{p}\mathbf{q}} = \pi$, so β_1 and β_2 each have arc length $\geq \pi$, with simultaneous equality if and only if the trace of each is half of a great circle. Thus the arc length of β is $\geq 2\pi$, with equality if and only if the trace of β equals two halves of great circles, which must be halves of the same great circle because β is smooth.

If $\mathbf{p} \neq -\mathbf{q}$, then there is a unique great circle C containing \mathbf{p} and \mathbf{q} . Let G denote the great circle that is orthogonal to C and equidistant to \mathbf{p} and \mathbf{q} ; see Fig. 2.19. The trace of β intersects G at some point \mathbf{r} (because it intersects every great circle). Notice that $\overline{\mathbf{r}\mathbf{p}} = \overline{-\mathbf{r}\mathbf{q}}$, because the 180-degree rotation about the illustrated axis is a rigid motion mapping $\mathbf{p} \mapsto \mathbf{q}$ and $\mathbf{r} \mapsto -\mathbf{r}$. It follows that

$$\overline{\mathbf{r}\mathbf{p}} + \overline{\mathbf{r}\mathbf{q}} = \overline{-\mathbf{r}\mathbf{q}} + \overline{\mathbf{r}\mathbf{q}} = \pi.$$

Either β_1 or β_2 travels between \mathbf{p} and \mathbf{q} via \mathbf{r} , so its arc length must be at least $\overline{\mathbf{r}\mathbf{p}} + \overline{\mathbf{r}\mathbf{q}} = \pi$. In fact, its arc length must be $> \pi$ (because equality would force the angle labeled θ to equal 180° in order for β to be smooth, contradicting the assumption that $\mathbf{p} \neq -\mathbf{q}$). Since this is half of β , the full arc length of β must be $> 2\pi$, as desired. \square

PROOF OF FENCHEL'S THEOREM FOR SPACE CURVES. Let $\gamma : [0, l] \rightarrow \mathbb{R}^3$ be a unit-speed closed curve. Since γ is of unit speed, its velocity function, \mathbf{v} , is a path in S^2 (visualized as in Fig. 1.17 from Sect. refch1:sec5 of Chap. 1).

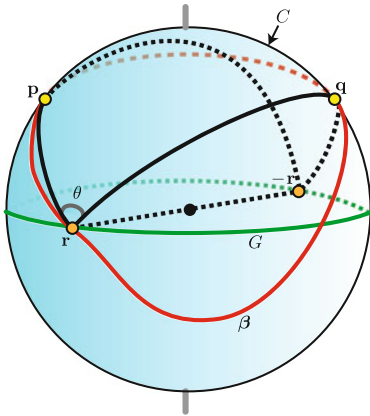


FIGURE 2.19. The proof of Lemma 2.17

We will use the fact that γ is closed to prove that the trace of \mathbf{v} intersects every great circle. For this, let $\mathcal{P} \subset \mathbb{R}^3$ be an arbitrary two-dimensional subspace, so that $G = \mathcal{P} \cap S^2$ is an arbitrary great circle. Let \mathbf{n} be a normal vector to \mathcal{P} . Notice that a point of S^2 lies on G if and only if it is orthogonal to \mathbf{n} . Since $\frac{d}{dt} \langle \gamma(t), \mathbf{n} \rangle = \langle \mathbf{v}(t), \mathbf{n} \rangle$, the fundamental theorem of calculus gives

$$\int_0^l \langle \mathbf{v}(t), \mathbf{n} \rangle dt = \langle \gamma(l), \mathbf{n} \rangle - \langle \gamma(0), \mathbf{n} \rangle = 0 \quad (\text{because } \gamma \text{ is periodic}).$$

Since the average value of $\langle \mathbf{v}(t), \mathbf{n} \rangle$ equals zero, we must have $\langle \mathbf{v}(t_0), \mathbf{n} \rangle = 0$ for some $t_0 \in [0, l]$. Thus, the trace of \mathbf{v} intersects every great circle. By Lemma 2.17, the arc length of \mathbf{v} is $\geq 2\pi$.

Since γ is of unit speed, $\kappa(t) = |\mathbf{v}'(t)|$, so we have

$$(\text{total curvature of } \gamma) = \int_0^l \kappa(t) dt = \int_0^l |\mathbf{v}'(t)| dt = (\text{arc length of } \mathbf{v}) \geq 2\pi.$$

If equality holds, then \mathbf{v} must be a simple parametrization of a great circle, say the great circle $G = S^2 \cap \mathcal{P}$. Since $\gamma(t) = \int_0^t \mathbf{v}(u) du + \gamma(0)$, the trace of γ must lie in the plane $\{\gamma(0) + z \mid z \in \mathcal{P}\}$. After applying a rigid motion, we can assume that this plane is the xy -plane, so we can consider γ to be a plane curve. Its velocity function \mathbf{v} is a *simple* parametrization of the unit circle S^1 . This implies that γ has rotation index ± 1 and has a monotonic global angle function (or equivalently, its signed curvature does not change sign). Exercise 2.10 (on page 78) implies that γ is simple, and then Lemma 2.14 implies that γ is convex. \square

EXERCISES

EXERCISE 2.13. Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a (not necessarily closed) regular plane curve with $\gamma(a) = \gamma(b)$. What is the minimal possible total curvature of γ ?

EXERCISE 2.14. Prove the $n = 2$ (plane curve) case of Fenchel's theorem. *Hint: The $n = 3$ proof involved showing that \mathbf{v} intersects every great circle of S^2 (that is, it meets the intersection of S^2 with every two-dimensional subspace of \mathbb{R}^3). The $n = 2$ analogue is that \mathbf{v} intersects every pair of antipodal points of S^1 (that is, it meets the intersection of S^1 with every one-dimensional subspace of \mathbb{R}^2).*

EXERCISE 2.15. Prove that every great circle of S^2 is the image of the equator under a rigid motion.

EXERCISE 2.16. Prove that the length, L , of a regular closed curve in \mathbb{R}^n with nowhere vanishing curvature satisfies

$$L \geq \frac{2\pi}{\kappa_{\max}},$$

where $\kappa_{\max} > 0$ is the global maximum of its curvature function.

\square

4. Green's Theorem (Calculus Background)

This section is devoted to Green's theorem, a powerful global theorem about vector fields on \mathbb{R}^2 . As a consequence, we will prove in the next section that a circle is the least-perimeter way to enclose a given area in the plane.

For consistency with other sources, in the remainder of this chapter we will sometimes use the term “curve” to mean the trace of a regular parametrized curve, and the term “oriented curve” to mean such a trace together with a choice of one of the two possible directions in which the trace could be traversed. A more formal and precise way to formulate this notion was discussed in Exercise 1.35 (on page 24), but the informal version is sufficient for our purposes.

We begin with some basic facts about vector fields on \mathbb{R}^n , which is also good preparation for our later study of vector fields on curved surfaces.

DEFINITION 2.18.

A **vector field** on an open set $U \subset \mathbb{R}^n$ is a smooth function $\mathbf{F} : U \rightarrow \mathbb{R}^n$.

Thus, \mathbf{F} associates to each point $\mathbf{p} \in U$ a vector $\mathbf{F}(\mathbf{p})$, which should be visualized with its tail drawn at \mathbf{p} . For example, a vector field on $U \subset \mathbb{R}^2$ will have the form $\mathbf{F}(x, y) = (P(x, y), Q(x, y))$, where $P, Q : U \rightarrow \mathbb{R}$ are called the **component functions** of \mathbf{F} (we’ll write this as $\mathbf{F} = (P, Q)$); see Fig. 2.20. Smoothness of \mathbf{F} means that the component functions are smooth in the sense that all partial derivatives of all orders exist. Similarly, a vector field on $U \subset \mathbb{R}^3$ has the form $\mathbf{F}(x, y, z) = (P(x, y, z), Q(x, y, z), R(x, y, z))$, shorthand as $\mathbf{F} = (P, Q, R)$.

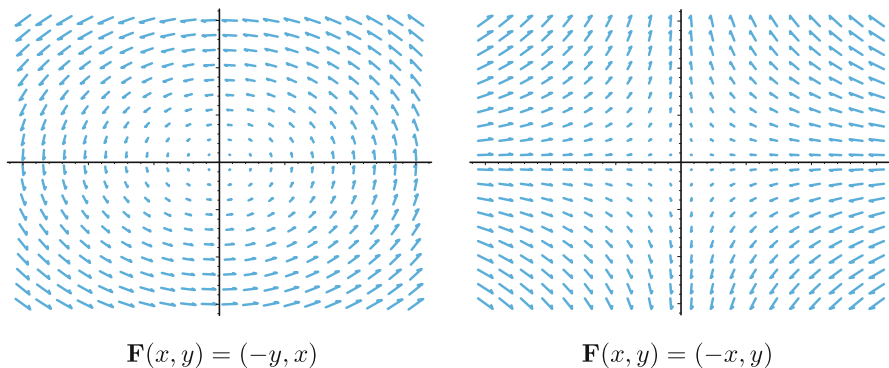


FIGURE 2.20. Two vector fields on \mathbb{R}^2 , with all vectors drawn about one-tenth their correct lengths for legibility

In physics applications, a vector field often represents a **force field**; that is, $\mathbf{F}(\mathbf{p})$ represents the force that would act on an object placed at position \mathbf{p} . The object might be a satellite acted on by gravitational forces of nearby planets, a paperclip acted on by the forces of nearby magnets, or a piece of tumbleweed acted on by wind currents. But the tumbleweed example is a stretch, because wind currents tend to change with time, while vector fields model forces that change only with position.

When a constant (vector) force, \mathbf{F} , moves an object along the displacement vector \mathbf{D} (which points from the starting to the ending position), the **work** done is defined as

$$W = |\mathbf{F}||\mathbf{D}| \cos(\theta) = \langle \mathbf{F}, \mathbf{D} \rangle.$$

To understand why this definition is reasonable, first imagine lifting a five-pound statue three feet off the ground. This requires 15 foot-pounds of work against gravity (\mathbf{F} and \mathbf{D} point in the same direction here—straight up—so their inner product is their regular product). If you instead move the five-pound statue along a diagonal line so that it ends up three feet up and seven feet to the right, as in Fig. 2.21, then $\mathbf{F} = (0, 5)$ and $\mathbf{D} = (7, 3)$, so $W = \langle (0, 5), (7, 3) \rangle = 15$. It's not surprising that the answer stayed the same—only the component of the displacement in the direction of the force is relevant (the horizontal component of the displacement requires no work against gravity). This is the same as saying that only the component of the force in the direction of the displacement is relevant. In any case, this example should help explain why our definition of work is reasonable.

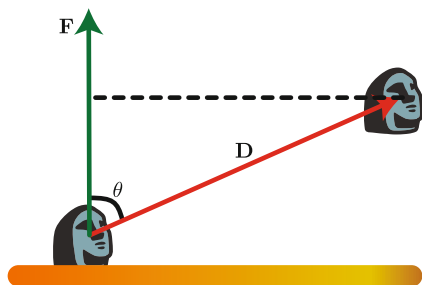


FIGURE 2.21. $W = |\mathbf{F}||\mathbf{D}| \cos(\theta) = \langle \mathbf{F}, \mathbf{D} \rangle$

How much work is required to send the statue to the Moon? It would need to travel a curved path, γ , along which the gravitational force vector, \mathbf{F} , would change as it moves farther from Earth's tug and closer to the Moon's tug. In situations like this, the work is calculated with a line integral.

DEFINITION 2.19.

If C is an oriented plane curve parametrized as $\gamma : [a, b] \rightarrow \mathbb{R}^2$ and \mathbf{F} is a vector field whose domain contains C , then the **line integral** of \mathbf{F} along C is defined as

$$\int_C \mathbf{F} \cdot d\gamma = \int_a^b \langle \mathbf{F}(\gamma(t)), \gamma'(t) \rangle dt.$$

When C is a simple closed curve, the line integral is also denoted by $\oint_C \mathbf{F} \cdot d\gamma$, and is called the **circulation** of \mathbf{F} around C .

EXAMPLE 2.20. Consider the vector field on \mathbb{R}^2 defined as $\mathbf{F}(x, y) = (-y, x)$, previously illustrated in Fig. 2.20. Let C denote the counterclockwise circle of radius 3 about the origin of \mathbb{R}^2 . To compute $\oint_C \mathbf{F} \cdot d\gamma$, we first parametrize C as $\gamma(t) = (\underbrace{3 \cos(t)}_{x(t)}, \underbrace{3 \sin(t)}_{y(t)})$, $t \in [0, 2\pi]$, and write

$$\begin{aligned} \oint_C \mathbf{F} \cdot d\gamma &= \int_0^{2\pi} \langle \mathbf{F}(x(t), y(t)), (x'(t), y'(t)) \rangle \\ &= \int_0^{2\pi} \langle (-3 \sin t, 3 \cos t), (-3 \sin t, 3 \cos t) \rangle = \int_0^{2\pi} 9 = 18\pi. \end{aligned}$$

The line integral represents *the work done by the force field \mathbf{F} in moving the object along the curve C* . This interpretation is reasonable, because a Riemann sum for the line integral has the form

$$\sum_i \langle \mathbf{F}(\gamma(t_i)), \gamma'(t_i) \Delta t_i \rangle = \sum_i \left\langle \mathbf{F}(\gamma(t_i)), \underbrace{\gamma'(t_i) \Delta t_i}_{\approx \text{displacement}} \right\rangle,$$

where t_i is a sample point from the i th subinterval into which $[a, b]$ is partitioned, and Δt_i is the length of this subinterval. The restriction of γ to each such subinterval is a subarc of C . When Δt_i is small, the forces along the i th subarc are approximately constant at the sample value $\mathbf{F}(\gamma(t_i))$, and this subarc is itself an approximately straight displacement by $\gamma'(t_i) \Delta t_i$ (because this vector points in the right direction and has the right length). Thus, the i th term of this Riemann sum approximates the work done in moving the object along the i th subarc, so the entire Riemann sum approximates the work done in moving the object along all of C ; see Fig. 2.22.

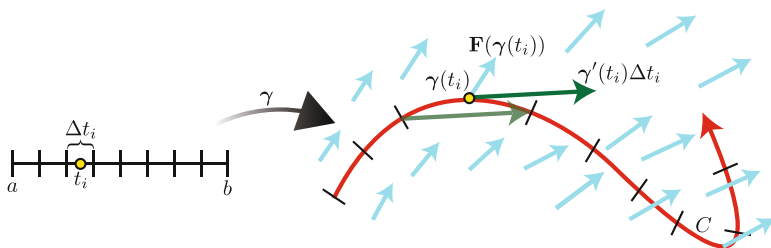


FIGURE 2.22. A Riemann sum for $\int_C \mathbf{F} \cdot d\gamma$ approximates the work done by the force field \mathbf{F} in moving the object along C

If C is only a *piecewise*-regular curve, then $\int_C \mathbf{F} \cdot d\gamma$ is defined as the sum of the line integrals of \mathbf{F} along the smooth segments of C . Line integrals are unchanged by orientation-preserving reparametrizations (Exercise 2.17). An orientation-reversing reparametrization would change the sign of the line integral; that is, $\int_{-C} \mathbf{F} \cdot d\gamma = -\int_C \mathbf{F} \cdot d\gamma$, where “ $-C$ ” represents C with opposite orientation (traversed in the opposite direction).

If the vectors of \mathbf{F} (encountered along the way as C is traversed) mostly point in the direction of motion, then the line integral will be positive, and is interpreted as the work done by \mathbf{F} in moving the object along the curve. This was the case in Example 2.20. If \mathbf{F} mostly points against the direction of motion, then the line integral will be negative, and its absolute value is interpreted as the work required for some independent force to move the object against \mathbf{F} along the curve.

If C is a small counterclockwise circle, you could think of \mathbf{F} as modeling the flow of a water current, and imagine C as the rim of a paddle wheel set in the current. The circulation is roughly the force with which the current spins the paddle wheel counterclockwise; see Fig. 2.23.

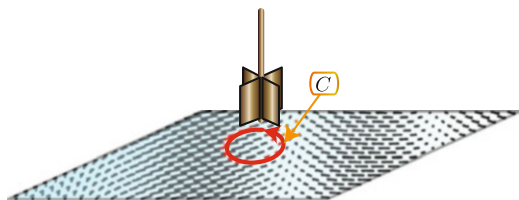


FIGURE 2.23. The circulation $\oint_C \mathbf{F} \cdot d\gamma$ is roughly the force with which the current spins the paddle wheel counterclockwise

Notice that $\int_C \mathbf{F} \cdot d\gamma$ is defined for arbitrary paths and arbitrary vector fields, which need not be related to each other. But the story becomes more natural when there are no forces other than \mathbf{F} . In this case, the path γ is completely determined by \mathbf{F} and by the initial conditions $\{\gamma(a), \gamma'(a)\}$; see Fig. 2.24. The next example provides a physical interpretation of the line integral in this situation.

EXAMPLE 2.21 (Line Integrals with No Other Forces). If \mathbf{F} is the only force, then Newton's law says that $\mathbf{F}(\gamma(t)) = m\mathbf{a}(t)$ for all $t \in [a, b]$, where m is the object's mass. One could use this to solve for γ , given initial conditions and a formula for \mathbf{F} , but instead we will use it here to derive a general meaning for the line integral:

$$\begin{aligned} \int_C \mathbf{F} \cdot d\gamma &= \int_a^b \langle \mathbf{F}(\gamma(t)), \gamma'(t) \rangle dt = \int_a^b \langle m\mathbf{a}(t), \mathbf{v}(t) \rangle dt = \frac{m}{2} \int_a^b \frac{d}{dt} \langle \mathbf{v}(t), \mathbf{v}(t) \rangle dt \\ &= \frac{m}{2} \int_a^b \frac{d}{dt} |\mathbf{v}(t)|^2 dt = \frac{m}{2} |\mathbf{v}(b)|^2 - \frac{m}{2} |\mathbf{v}(a)|^2. \end{aligned}$$

Since an object's **kinetic energy** is defined as $\frac{m}{2} |\mathbf{v}|^2$, we learn that $\int_C \mathbf{F} \cdot d\gamma$ equals the object's net change in kinetic energy between times a and b .

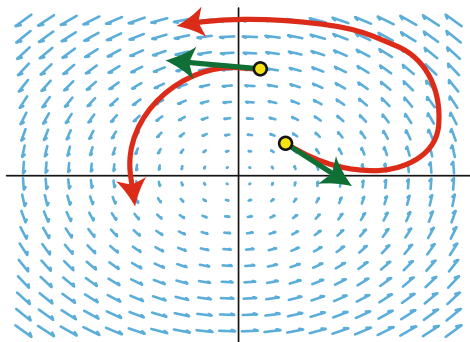


FIGURE 2.24. If there are no forces other than \mathbf{F} , then the object's path is determined by its initial position and initial velocity; it flows with the current

This physical interpretation agrees with our discussion earlier. If the force vectors are mostly in the direction of motion, then they push the object along, increasing its kinetic energy. If they are mostly against the direction of motion, then they slow the object down, decreasing its kinetic energy. Since work and energy are measured with the same units, this also agrees with the intuition that line integrals should represent work.

Another situation in which the line integral has a natural interpretation occurs when the vector field is *conservative*.

DEFINITION 2.22.

Let $U \subset \mathbb{R}^n$ be an open set, and let $f : U \rightarrow \mathbb{R}$ be a smooth function. The **gradient** of f , denoted by ∇f , is the vector field on U whose i th component function is the partial derivative of f with respect to the i th input variable. A vector field, \mathbf{F} , on U is called **conservative** if it is the gradient of some smooth function, f , on U . In this case, f is called a **potential function** of \mathbf{F} .

For example, the gradient of $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ is $\nabla f = (f_x, f_y)$, where $f_x = \frac{\partial f}{\partial x}$ and $f_y = \frac{\partial f}{\partial y}$. Similarly, the gradient of $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ is $\nabla f = (f_x, f_y, f_z)$. The following should be familiar from multivariable calculus:

LEMMA 2.23.

If $U \subset \mathbb{R}^n$ is an open set, $f : U \rightarrow \mathbb{R}$ is a smooth function, and $\gamma : I \rightarrow U$ is a regular curve, then for all $t \in I$,

$$(2.2) \quad \frac{d}{dt} f(\gamma(t)) = \langle \nabla f(\gamma(t)), \gamma'(t) \rangle.$$

At a particular time $t_0 \in I$, if we let $\mathbf{p}_0 = \gamma(t_0)$ and $\mathbf{v}_0 = \gamma'(t_0)$, this derivative is denoted by $df_{\mathbf{p}_0}(\mathbf{v}_0)$:

$$(2.3) \quad df_{\mathbf{p}_0}(\mathbf{v}_0) = \left. \frac{d}{dt} f(\gamma(t)) \right|_{t=t_0} = \langle \nabla f(\mathbf{p}_0), \mathbf{v}_0 \rangle.$$

We call $df_{\mathbf{p}_0}(\mathbf{v}_0)$ the **directional derivative** of f at \mathbf{p}_0 in the direction of \mathbf{v}_0 (although some books reserve this term for the case that \mathbf{v}_0 is of unit length). It represents the initial rate at which f changes along *any* regular

curve passing through \mathbf{p}_0 with initial velocity vector \mathbf{v}_0 . Notice that the value $\frac{d}{dt}\big|_{t=t_0} f(\gamma(t))$ depends only on the vectors $\nabla f(\mathbf{p}_0)$ and $\mathbf{v}_0 = \gamma'(t_0)$. It does not depend on $\gamma''(t_0)$ or on any other higher-order derivative information.

Figure 2.25 shows a contour diagram (a collection of level curves) for the function $f(x, y) = -\frac{1}{2}x^2 + \frac{1}{2}y^2$ together with its gradient $\nabla f(x, y) = (-x, y)$. This figure illustrates the following general geometric relationship between gradients and contour diagrams:

LEMMA 2.24.

If $U \subset \mathbb{R}^n$ is open, $f : U \rightarrow \mathbb{R}$ is smooth, and $\mathbf{p} \in U$ is such that $\nabla f(\mathbf{p}) \neq \mathbf{0}$, then:

- (1) $\nabla f(\mathbf{p})$ is orthogonal to the level set $S_{\mathbf{p}} = \{\mathbf{q} \in U \mid f(\mathbf{q}) = f(\mathbf{p})\}$ in this sense: if γ is any regular curve with $\gamma(0) = \mathbf{p}$ whose trace lies in $S_{\mathbf{p}}$, then $\langle \gamma'(0), \nabla f(\mathbf{p}) \rangle = 0$.
- (2) $\nabla f(\mathbf{p})$ points in the direction of greatest increase of f . More precisely, the directional derivative $df_{\mathbf{p}}(\mathbf{u})$ is maximized among all unit vectors \mathbf{u} by the choice $\mathbf{u} = \frac{\nabla f(\mathbf{p})}{|\nabla f(\mathbf{p})|}$.
- (3) The norm of the gradient equals the rate of increase of f in this maximizing direction; that is, $|\nabla f(\mathbf{p})| = df_{\mathbf{p}}(\mathbf{u})$, where $\mathbf{u} = \frac{\nabla f(\mathbf{p})}{|\nabla f(\mathbf{p})|}$.

PROOF. For (1), since $f(\gamma(t))$ is constant, $0 = \frac{d}{dt}\big|_{t=0} f(\gamma(t)) = \langle \nabla f(\mathbf{p}), \gamma'(0) \rangle$. For (2) and (3), notice that $df_{\mathbf{p}}(\mathbf{u}) = \langle \nabla f(\mathbf{p}), \mathbf{u} \rangle = |\nabla f(\mathbf{p})| \cos(\theta)$ has maximal value $|\nabla f(\mathbf{p})|$ occurring when the angle θ between \mathbf{u} and $\nabla f(\mathbf{p})$ equals zero. \square

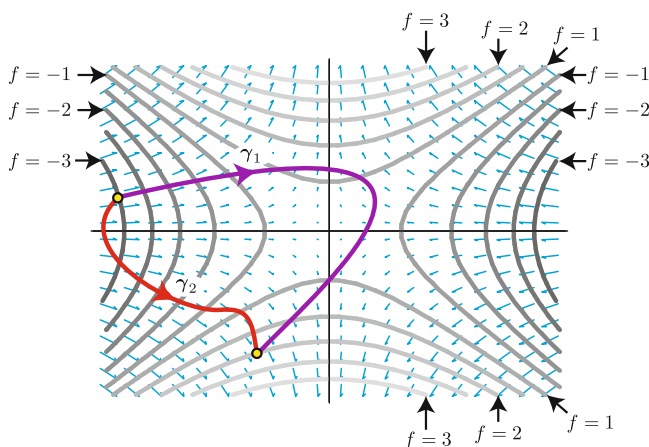


FIGURE 2.25. A contour diagram of $f(x, y) = -\frac{1}{2}x^2 + \frac{1}{2}y^2$ together with its gradient $\nabla f(x, y) = (-x, y)$. The line integral of ∇f along either γ_1 or γ_2 equals $2 - (-3) = 5$ (the net change in f)

The following proposition says that the line integral of a conservative vector field equals the net change of its potential function, as illustrated in Fig. 2.25:

PROPOSITION 2.25.

If $U \subset \mathbb{R}^n$ is open, $f : U \rightarrow \mathbb{R}$ is smooth, and $\gamma : [a, b] \rightarrow U$ is a parametrization of the piecewise-regular oriented curve C , then

$$\int_C \nabla f \cdot d\gamma = f(\gamma(b)) - f(\gamma(a)).$$

In particular, this line integral is **path-independent**—it would have the same value if C were replaced by any other piecewise-regular oriented curve in U with the same starting point $\gamma(a)$ and the same ending point $\gamma(b)$.

PROOF. We will prove this proposition in the case that C is smooth (the piecewise-regular case follows easily from this case). For this, we combine Eq. 2.2 with the fundamental theorem of calculus:

$$\int_C \nabla f \cdot d\gamma = \int_a^b \langle \nabla f(\gamma(t)), \gamma'(t) \rangle dt = \int_a^b \frac{d}{dt} f(\gamma(t)) dt = f(\gamma(b)) - f(\gamma(a)).$$

□

If γ is a closed path, then $\gamma(a) = \gamma(b)$, so $\oint_C \nabla f \cdot d\gamma = 0$. In fact, we have this:

PROPOSITION 2.26.

The following are equivalent properties for a vector field \mathbf{F} on an open path-connected set $U \subset \mathbb{R}^n$:

- (1) \mathbf{F} is conservative.
- (2) $\oint_C \mathbf{F} \cdot d\gamma = 0$ for every piecewise-regular closed curve C in U .
- (3) Line integrals of \mathbf{F} are path-independent; i.e., if C_1, C_2 are piecewise-regular curves in U with the same start and end points, then $\int_{C_1} \mathbf{F} \cdot d\gamma = \int_{C_2} \mathbf{F} \cdot d\gamma$.

PROOF. Exercise 2.18.

□

We can now justify the term “conservative”—these are the vector fields for which there is a conservation of energy law.

EXAMPLE 2.27 (Conservation of Energy). Suppose that $\mathbf{F} = \nabla f$ is a conservative vector field. Physicists call $\rho = -f$ the “potential energy function.” If there are no forces other than \mathbf{F} , then the curve γ that an object will follow

is determined by the object's initial position $\gamma(0)$ and its initial velocity $\gamma'(0)$. In this case, Example 2.21 combines with Proposition 2.25 to yield

$$\int_C \mathbf{F} \cdot d\gamma = \frac{m}{2} |\mathbf{v}(b)|^2 - \frac{m}{2} |\mathbf{v}(a)|^2 = -\rho(b) + \rho(a).$$

Thus, the total energy (kinetic plus potential) is the same at times a and b :

$$\frac{m}{2} |\mathbf{v}(a)|^2 + \rho(a) = \frac{m}{2} |\mathbf{v}(b)|^2 + \rho(b).$$

In summary, for a conservative vector field, it is possible to define a position-dependent “potential energy” that obeys a conservation law: potential energy gets traded off against kinetic energy as an object moves under the influence of only the field.

EXAMPLE 2.28 (Gravity). According to Newton's law, the magnitude of the gravitational force between two objects (with masses denoted by m and M) equals $\frac{C}{r^2}$, where r is the distance between their centers of mass, $C = mM G$, and G is the universal gravitational constant.

Consider a large object (such as the Earth) with mass M centered at $(0, 0, 0)$. Let $\mathbf{F}(\mathbf{p})$ denote the force it exerts on a small object (such as a grapefruit) with mass m centered at $\mathbf{p} = (x, y, z)$. Since $\mathbf{F}(\mathbf{p})$ has magnitude $\frac{C}{|\mathbf{p}|^2}$ and points in the direction of the center-pointing unit vector $-\frac{\mathbf{p}}{|\mathbf{p}|}$, we have

$$\mathbf{F}(\mathbf{p}) = -\frac{C}{|\mathbf{p}|^3} \mathbf{p} = -C (x^2 + y^2 + z^2)^{-3/2} (x, y, z).$$

To demonstrate that \mathbf{F} is a conservative vector field on its domain, $\mathbb{R}^3 - \{(0, 0, 0)\}$, we must construct a smooth real-valued potential function f on this domain such that $\mathbf{F} = \nabla f$. Trial and error suffices to come up with

$$f(\mathbf{p}) = \frac{C}{|\mathbf{p}|} - K = C (x^2 + y^2 + z^2)^{-1/2} - K,$$

where $K \in \mathbb{R}$ is an arbitrary constant. Therefore, $\rho(\mathbf{p}) = -f(\mathbf{p}) = K - \frac{C}{|\mathbf{p}|}$ is the potential energy function. It is often convenient to choose K such that the potential energy equals zero on the surface of the Earth. The conservation law for this situation says that the total energy

$$\underbrace{\frac{1}{2} m |\mathbf{v}|^2}_{\text{kinetic}} + \underbrace{K - \frac{C}{|\mathbf{p}|}}_{\text{potential}}$$

remains constant for an object under the influence of only gravity (no air resistance, no smashing into the Earth's surface, etc.).

Now look back at the graph of the vector field $\mathbf{F}(x, y) = (-y, x)$ in Fig. 2.20. How could you verify that this vector field is *not* conservative? Visually, you could observe that the line integral is not zero around a circle centered at the origin (as confirmed in Example 2.20). Or algebraically, you could use the following:

LEMMA 2.29.

If $\mathbf{F} = (P, Q)$ is a conservative vector field on an open set $U \subset \mathbb{R}^2$, then $Q_x = P_y$ at every point of U .

PROOF. Since $\mathbf{F} = \nabla f$, we have $P = f_x$ and $Q = f_y$. Since mixed partial derivatives commute,

$$Q_x = (f_y)_x = (f_x)_y = P_y.$$

□

Thus, two things are true for a conservative vector field on \mathbb{R}^2 . First, the quantity $Q_x - P_y$ vanishes, and second, the circulation around closed curves vanish. The following definition hints at the relationship between these two things:

DEFINITION 2.30.

The **infinitesimal circulation** of the vector field $\mathbf{F} = (P, Q)$ is the real-valued function defined as $Q_x - P_y$ on the domain of \mathbf{F} .

This (nonstandard) term is appropriate because the infinitesimal circulation equals the limit circulation around smaller and smaller circles:

COROLLARY 2.31.

If \mathbf{F} is a vector field defined in a neighborhood of $\mathbf{p} \in \mathbb{R}^2$, and C_r denotes the counterclockwise circle of radius r centered at \mathbf{p} , then

$$(Q_x - P_y)(\mathbf{p}) = \lim_{r \rightarrow 0} \frac{1}{\pi r^2} \oint_{C_r} \mathbf{F} \cdot d\gamma.$$

This result is labeled a corollary because Green's theorem will be required to prove it. Nevertheless, the geometric meaning of Green's theorem will be easier to comprehend after understanding the content of this corollary. Returning to the paddle-wheel metaphor, Corollary 2.31 roughly says that the infinitesimal circulation at \mathbf{p} measures the counterclockwise force (per unit of paddle-wheel area) of the current on a small paddle wheel placed at \mathbf{p} . Why might you expect the expression $Q_x - P_y$ to measure such a thing? This expression is positive when Q_x is positive and P_y is negative. Both cause counterclockwise spin. $Q_x > 0$ means that the y -component of \mathbf{F} increases as x increases, so the current has more upward push against the right side of the paddle wheel than the left, causing counterclockwise spin. And $P_y < 0$ means that the x -component of \mathbf{F} decreases as y increases, so the current has more rightward push against the bottom side of the paddle wheel than the top, again causing counterclockwise spin. These two phenomena (the spin caused by vertical variations in the horizontal component of force, and the spin caused by horizontal variations in the vertical component of force) are separated out in the top two vector fields displayed in Fig. 2.26, while these phenomena combine additively in the bottom vector field. Is it visually

believable that a paddle wheel placed at any position in any of these three vector fields will experience the same force spinning it counterclockwise?

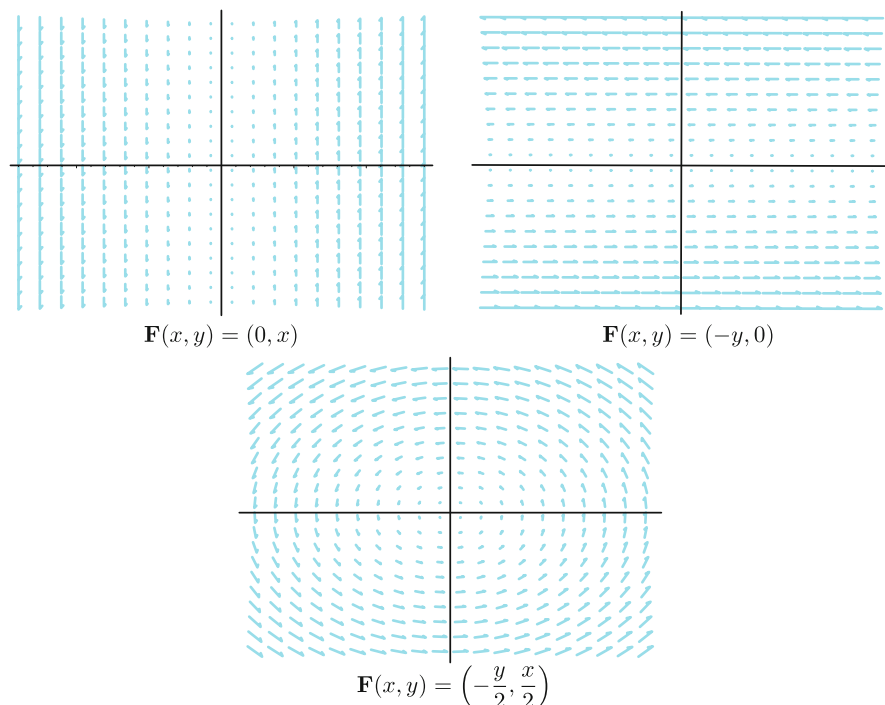


FIGURE 2.26. Three vector fields with constant infinitesimal circulation equal to 1

THEOREM 2.32 (Green's Theorem).

Let C be a positively oriented piecewise-regular simple closed curve in the plane. Let D denote the interior of C . Let $\mathbf{F} = (P, Q)$ be a vector field defined on an open set containing $D \cup C$. Then,

$$\oint_C \mathbf{F} \cdot d\gamma = \iint_D (Q_x - P_y) dA.$$

Green's theorem says that the circulation of \mathbf{F} around C equals the integral over D of the infinitesimal circulation. So if the average value over D of the infinitesimal circulation is positive (paddle wheels mostly spin counterclockwise), then the circulation around C is positive (the vectors encountered while traversing the curve mostly align with the direction of motion).

EXAMPLE 2.33. *Green's theorem provides an alternative way to compute the line integral of Example 2.20. Since the infinitesimal circulation of the vector field in this example is constant at 2, we have*

$$\oint_C \mathbf{F} \cdot d\boldsymbol{\gamma} = \iint_D 2 \, dA = 2 \cdot \text{Area}(D) = 18\pi,$$

where D is the interior of C .

PROOF OF COROLLARY 2.31 USING GREEN'S THEOREM. Let D_r denote the interior of C_r . If r is sufficiently small, then $(Q_x - P_y)$ is approximately constant over D_r at the sample value $(Q_x - P_y)(\mathbf{p})$. Green's theorem gives

$$\oint_{C_r} \mathbf{F} \cdot d\boldsymbol{\gamma} = \iint_{D_r} (Q_x - P_y) \, dA \approx \text{area}(D_r)((Q_x - P_y)(\mathbf{p})),$$

from which the result follows. \square

We will next prove Green's theorem in the special case that C is a rectangle, and then give a nonrigorous indication of how the general case follows.

PROOF OF GREEN'S THEOREM WHEN C IS A RECTANGLE. Suppose that D is the region $\{(x, y) \in \mathbb{R}^2 \mid a \leq x \leq b, c \leq y \leq d\}$, and C is its boundary. Denote the four segments of C as in Fig. 2.27, so $C = C_B + C_R - C_T - C_L$.

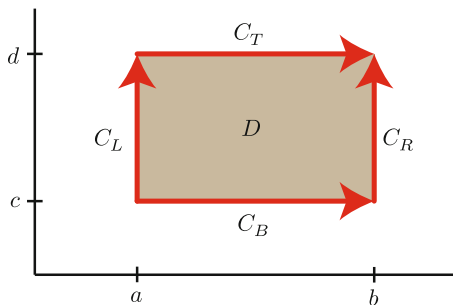


FIGURE 2.27. $C = C_B + C_R - C_T - C_L$

We parametrize each segment of C in the most natural manner. For example, we parametrize C_B as $\boldsymbol{\gamma}(t) = (t, c)$, $t \in [a, b]$, so that

$$\int_{C_B} \mathbf{F} \cdot d\boldsymbol{\gamma} = \int_a^b \langle \mathbf{F}(\boldsymbol{\gamma}(t)), \boldsymbol{\gamma}'(t) \rangle \, dt = \int_a^b \langle (P(t, c), Q(t, c)), (1, 0) \rangle \, dt = \int_a^b P(t, c) \, dt.$$

After similarly expressing the line integrals along the other three segments, we get

$$(2.4) \quad \int_C \mathbf{F} \cdot d\boldsymbol{\gamma} = \underbrace{\int_c^d (Q(b, t) - Q(a, t)) \, dt}_{\int_{C_R} \mathbf{F} \cdot d\boldsymbol{\gamma} - \int_{C_L} \mathbf{F} \cdot d\boldsymbol{\gamma}} + \underbrace{\int_a^b (P(t, c) - P(t, d)) \, dt}_{\int_{C_B} \mathbf{F} \cdot d\boldsymbol{\gamma} - \int_{C_T} \mathbf{F} \cdot d\boldsymbol{\gamma}}.$$

On the other hand,

$$\begin{aligned}\iint_D (Q_x - P_y) dA &= \iint_D Q_x dA - \iint_D P_y dA \\ &= \int_{y=c}^{y=d} \left(\int_{x=a}^{x=b} Q_x dx \right) dy - \int_{x=a}^{x=b} \left(\int_{y=c}^{y=d} P_y dy \right) dx.\end{aligned}$$

Applying the fundamental theorem of calculus to both inner integrals turns this last expression into the expression in Eq. 2.4. \square

We now provide a (nonrigorous) indication of how the general case of Green's theorem follows from the rectangle case. The idea is to find a collection of rectangles, $\{R_1, R_2, \dots, R_k\}$, whose union closely approximates D (as in Fig. 2.28), so that the following is a good approximation:

$$\iint_D (Q_x - P_y) dA \approx \iint_{R_1 \cup R_2 \cup \dots \cup R_k} (Q_x - P_y) dA = \sum_i \iint_{R_i} (Q_x - P_y) dA.$$

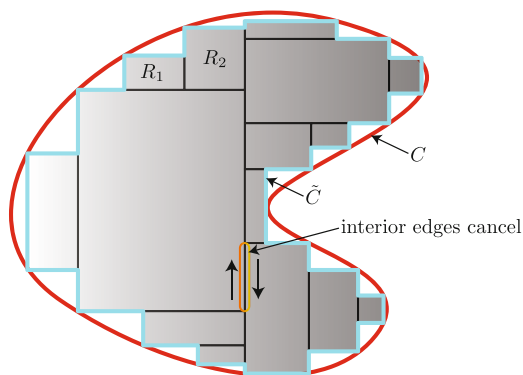


FIGURE 2.28. An “Etch A Sketch” approximation of C

Denote the boundaries of these rectangles by $\{C_1, C_2, \dots, C_k\}$, all oriented counterclockwise. Let \tilde{C} denote the outer edge (which is colored light blue in the figure and looks like something drawn on an Etch A Sketch). It is possible to ensure that the following is a good approximation: $\oint_C \mathbf{F} \cdot d\gamma \approx \oint_{\tilde{C}} \mathbf{F} \cdot d\gamma$. We won't prove this, but think about why you believe it; the arc lengths of \tilde{C} and C will *not* become close to each other as the picture is refined using more and more smaller rectangles, so why should the line integrals become close to each other?

Notice that $\sum_i \oint_{C_i} \mathbf{F} \cdot d\gamma = \oint_{\tilde{C}} \mathbf{F} \cdot d\gamma$, because all interior edges cancel. In other words, each interior edge receives opposite orientations from the two rectangles that share it. Since Green's theorem holds on each rectangle, we have

$$\oint_C \mathbf{F} \cdot d\boldsymbol{\gamma} \approx \oint_{\tilde{C}} \mathbf{F} \cdot d\boldsymbol{\gamma} = \sum_i \oint_{C_i} \mathbf{F} \cdot d\boldsymbol{\gamma} = \sum_i \iint_{R_i} (Q_x - P_y) dA \approx \iint_D (Q_x - P_y) dA.$$

This completes our proof-sketch of Green's theorem.

We end this section by discussing the alternative flux version of Green's theorem. We begin with the concept of *flux*. Let C be a simple closed plane curve parametrized as $\boldsymbol{\gamma} : [a, b] \rightarrow \mathbb{R}^2$. Let D be the interior of C . Let \mathbf{F} be a vector field defined on an open set containing $D \cup C$. For each $t \in [a, b]$, let $\mathbf{n}(t)$ denote the outward-pointing unit-length vector orthogonal to $\mathbf{v}(t)$, which is purple in Fig. 2.29.

DEFINITION 2.34.

The **flux** of \mathbf{F} across C is defined as

$$\text{flux} = \int_a^b \langle \mathbf{F}(\boldsymbol{\gamma}(t)), \mathbf{n}(t) \rangle |\mathbf{v}(t)| dt.$$

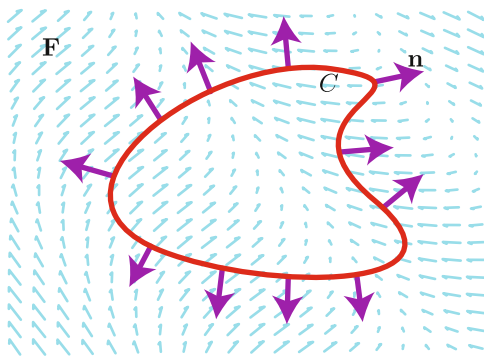


FIGURE 2.29. The flux measures the net outflow of \mathbf{F} across C

The flux of \mathbf{F} across C is independent of the parametrization of C . For a unit-speed parametrization, it is simply the integral of $\langle \mathbf{F}, \mathbf{n} \rangle$ along $\boldsymbol{\gamma}$; see Fig. 2.29. So the flux is positive if the vectors of \mathbf{F} encountered while traversing C mostly point outward. The flux is negative if they mostly point inward. For this reason, the flux is sometimes called the “net outflow” of \mathbf{F} across C .

Here is an imperfect but useful metaphor: if C is a net, and \mathbf{F} represents the velocity vectors of a school of small fish, then the flux is roughly the net rate at which fish are escaping from the net. The flux would be negative if fish were mostly swimming into the net.

DEFINITION 2.35.

The **divergence** of the vector field $\mathbf{F} = (P, Q)$ is the real-valued function defined as $P_x + Q_y$ on the domain of \mathbf{F} .

The analogy with Definition 2.30 would be tighter if “divergence” were instead called “infinitesimal flux,” because of its geometric meaning:

COROLLARY 2.36.

If \mathbf{F} is a vector field defined in a neighborhood of $\mathbf{p} \in \mathbb{R}^2$, and C_r denotes the circle of radius r centered at \mathbf{p} , then

$$(P_x + Q_y)(\mathbf{p}) = \lim_{r \rightarrow 0} \frac{1}{\pi r^2} (\text{flux of } \mathbf{F} \text{ across } C_r).$$

The proof of Corollary 2.36 using the flux version of Green's theorem (which we are about to state) is essentially identical to our previous proof of Corollary 2.31 using the original version of Green's theorem.

THEOREM 2.37 (Flux Version of Green's Theorem).

Let C be a simple closed curve in the plane. Let D denote the interior of C . Let $\mathbf{F} = (P, Q)$ be a vector field defined on an open set containing $D \cup C$. Then,

$$(\text{flux of } \mathbf{F} \text{ across } C) = \iint_D (P_x + Q_y) dA.$$

Thus, the flux of \mathbf{F} across C equals the integral of the divergence over D . Returning to the fish metaphor, the divergence at a point roughly measures the extent to which fish are swimming away from that point (diverging from that point), so the theorem roughly says that net outflow of fish across a net equals the integral of the rate at which fish are diverging from all the points inside the net.

The following proof of Theorem 2.37 can be summarized as “rotate all of the vectors 90° and then apply Green's theorem.”

PROOF. Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a unit-speed positively oriented parametrization of C . Denote the components of γ by $\gamma(t) = (x(t), y(t))$. Notice that

$$\mathbf{n}(t) = -R_{90}(\mathbf{v}(t)) = (y'(t), -x'(t)).$$

Let $\tilde{\mathbf{F}}$ denote the vector field obtained from \mathbf{F} by rotating all individual vectors 90° counterclockwise; that is, $\tilde{\mathbf{F}} = R_{90}(\mathbf{F}) = (-Q, P)$. We have

$$\begin{aligned} \text{flux} &= \int_a^b \langle \mathbf{F}(\gamma(t)), \mathbf{n}(t) \rangle dt = \int_a^b \langle (P(\gamma(t)), Q(\gamma(t))), (y'(t), -x'(t)) \rangle dt \\ &= \int_a^b \langle (-Q(\gamma(t)), P(\gamma(t))), (x'(t), y'(t)) \rangle dt \\ &= \int_a^b \langle \tilde{\mathbf{F}}(\gamma(t)), \mathbf{v}(t) \rangle dt \\ &= \underbrace{\oint_C \tilde{\mathbf{F}} \cdot d\gamma}_{\text{apply Green's theorem to } \tilde{\mathbf{F}}} = \iint_D (P_x + Q_y) dA. \end{aligned}$$

This completes the proof, but notice that the proof's heart can be concisely rewritten by expressing the integrand as

$$\langle \mathbf{F}, \mathbf{n} \rangle = \langle \mathbf{F}, -R_{90}\mathbf{v} \rangle = \langle R_{90}\mathbf{F}, -R_{90}^2\mathbf{v} \rangle = \langle R_{90}\mathbf{F}, \mathbf{v} \rangle = \langle \tilde{\mathbf{F}}, \mathbf{v} \rangle.$$

Here we're using the fact that the rotation map $R_{90} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is a rigid motion with the property that $R_{90} \circ R_{90} = -(\text{the identity map})$. \square

EXERCISES

EXERCISE 2.17. Prove that line integrals are unchanged by orientation-preserving reparametrizations.

EXERCISE 2.18. Prove Proposition 2.26.

EXERCISE 2.19. What is the **escape velocity** of a rocket on the Earth's surface (the upward velocity needed to escape the planet's gravity, so that it will never be pulled back to the planet, neglecting air resistance and other forces)?

EXERCISE 2.20. For the vector field $\mathbf{F}(x, y) = (x, y + 2)$:

- (1) Calculate directly the line integral along the top half of the unit circle from $(1, 0)$ to $(-1, 0)$.
- (2) Calculate directly the line integral along the straight line from $(1, 0)$ to $(-1, 0)$.
- (3) Recalculate the above line integrals by finding a potential function for \mathbf{F} and applying Proposition 2.25.

EXERCISE 2.21. For the vector field $\mathbf{F}(x, y) = (2y + 3, x)$:

- (1) Calculate the line integral along the top half of the unit circle from $(1, 0)$ to $(-1, 0)$.
- (2) Calculate the line integral along the straight line from $(1, 0)$ to $(-1, 0)$.
- (3) Calculate the line integral around the loop that first traverses the top half of the unit circle from $(1, 0)$ to $(-1, 0)$ and then traverses the straight line from $(-1, 0)$ to $(1, 0)$. Solve this by subtracting the previous two answers, and also solve this using Green's theorem.

EXERCISE 2.22. Let $p_0 \in \mathbb{R}^n$ and let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ denote the “distance to p_0 ” function; that is, $f(p) = \text{dist}(p, p_0)$.

- (1) If $p \in \mathbb{R}^n$ with $p \neq p_0$, verify that $\nabla f(p) = \frac{p - p_0}{|p - p_0|}$.
- (2) Suppose that C is a simple closed curve in \mathbb{R}^n with $p_0 \notin C$. If $p \in C$ is the point of C that is closest to p_0 , prove that $p - p_0$ is orthogonal to the tangent line to C at p .

EXERCISE 2.23. With and without Green's theorem, show that a constant vector field has zero line integral around every circle in \mathbb{R}^2 .

EXERCISE 2.24. With and without Green's theorem, calculate the line integral of $\mathbf{F}(x, y) = (xy, x^2)$ around the triangle with vertices $(1, 1)$, $(1, 5)$, and $(3, 4)$.

EXERCISE 2.25. Prove the following partial converse to Lemma 2.29: If \mathbf{F} is a vector field whose domain is all of \mathbb{R}^2 with the property that $Q_x(p) = P_y(p)$ for all $p \in \mathbb{R}^2$, then \mathbf{F} is conservative.

EXERCISE 2.26. Consider the vector field \mathbf{F} with domain $\mathbb{R}^2 - \{(0, 0)\}$ defined as $\mathbf{F}(x, y) = \left(\frac{-y}{x^2+y^2}, \frac{x}{x^2+y^2} \right)$.

- (1) Verify that $Q_x - P_y = 0$ at every point of this domain.
- (2) Verify that the line integral of \mathbf{F} is *not* zero around a circle centered at the origin.
- (3) By Green's theorem, around what types of loops must the line integral of \mathbf{F} equal zero?

EXERCISE 2.27. Prove Green's theorem (Theorem 2.32) as a corollary of the flux version of Green's theorem (Theorem 2.37).

□

5. The Isoperimetric Inequality (Optional)

In this section, we prove the classic isoperimetric inequality in the plane as an application of Green's theorem.

We begin with a corollary of Green's theorem that is useful for computing area:

COROLLARY 2.38.

Let C be a positively oriented simple closed plane curve parametrized as $\gamma(t) = (x(t), y(t))$, $t \in [a, b]$. The area of the interior, D , of C equals

$$\text{Area}(D) = \int_a^b x(t)y'(t) dt = - \int_a^b y(t)x'(t) dt.$$

PROOF. To obtain these two formulas for area, apply Green's theorem separately to the two vector fields $\mathbf{F}_1(x, y) = (0, x)$ and $\mathbf{F}_2(x, y) = (-y, 0)$. Each of these vector fields has constant infinitesimal circulation of 1 and is illustrated in Fig. 2.26 on page 91. □

The vector field $\mathbf{F}_1(x, y) = (0, x)$, which was used in the above proof, is also illustrated in Fig. 2.30. As you look at this image, think about how you could design a mechanical device that measures the area inside C when it is pushed around C . To measure the arc length of C would be much easier—a wheel on a stick with an odometer would suffice, like the one shown in Fig. 2.30 (left). But it should also be mechanically possible to measure the area by measuring $\oint_C \mathbf{F}_1 \cdot d\gamma = \int_a^b x(t)y'(t) dt$. The wheel and odometer would need to be modified to be sensitive only to the vertical component of motion, with sensitivity proportional to the distance to the y -axis. Exercise 2.33 discusses a **planimeter**, which is a mechanical device that uses Green's theorem to measure area (although not exactly in the manner suggested above).

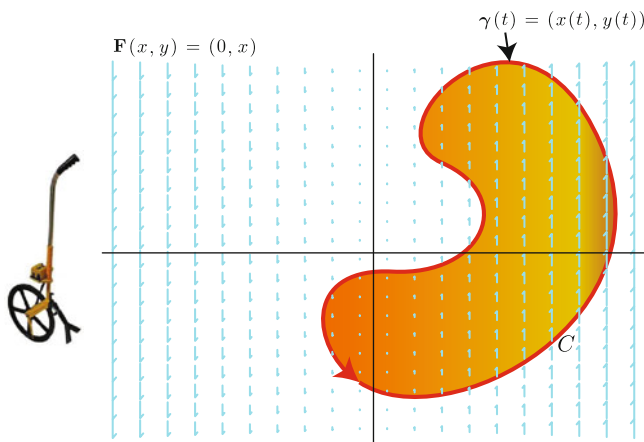


FIGURE 2.30. $\text{Area} = \oint_C \mathbf{F} \cdot d\boldsymbol{\gamma} = \int_a^b x(t)y'(t) dt$

You can check that the length l and area A of a circle of any radius are related by the equation $l^2 = 4\pi A$ (here we're using "length" as an abbreviation for "arc length," which is also often called "perimeter"). The remainder of this section is devoted to proving that every other simple closed curve has larger length than the circle that encloses the same area. More precisely, we have the following theorem.

THEOREM 2.39 (The Isoperimetric Inequality).

Let C be a simple closed plane curve. If l denotes the length of C and A denotes the area of the interior of C , then

$$l^2 \geq 4\pi A,$$

with equality if and only if C is a circle.

Thus, if C is not a circle, then its length is greater than the length of a circle with the same area as C . You could also read the theorem this way: if C is not a circle, then its area is less than the area of a circle with the same length as C . This second viewpoint justifies the term "isoperimetric" which means "same perimeter"—among all curves with the same perimeter, the circle has the largest area.

PROOF. Take two vertical lines that do not intersect C and move them together until they first touch C , so that C becomes tangent to both lines and lies between them. Let S^1 be a circle that is also tangent to both of these lines. Its radius, r , equals half the distance between the lines. Assume without loss of generality that the origin is the center of S^1 . Figure 2.31 shows C and S^1 as nonintersecting, but the vertical position of S^1 will be irrelevant for the proof.

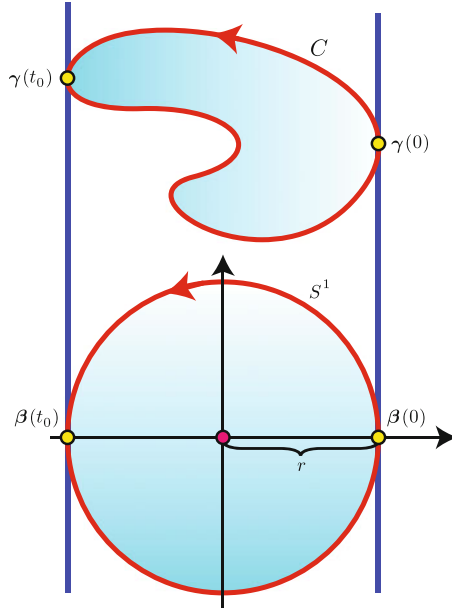


FIGURE 2.31. The proof of the isoperimetric inequality

Let $\gamma(t) = (x(t), y(t))$, $t \in [0, l]$, denote a parametrization of C by arc length such that $\gamma(0)$ is the intersection of C with one of the vertical lines, and $\gamma(t_0)$ is the intersection with the other for some $t_0 \in (0, l)$. We will choose a parametrization of S^1 , denoted by $\beta : [0, l] \rightarrow \mathbb{R}^2$, contrived so that for all $t \in [0, l]$, the x -coordinates of γ and β agree; that is, we will traverse S^1 so as to remain vertically aligned with someone who is traversing C at unit speed. This is achieved by defining $\beta(t) = (x(t), \tilde{y}(t))$, where

$$\tilde{y}(t) = \pm \sqrt{r^2 - x(t)^2},$$

with the sign depending on whether $t \in [0, t_0]$ or $t \in [t_0, l]$. Notice that β is not necessarily a regular parametrization, but this will not affect the calculations that follow.

By Corollary 2.38, the area of C is $A = \int_0^l x(t)y'(t) dt$, while the area of S^1 is $\pi r^2 = -\int_0^l \tilde{y}(t)x'(t) dt$ (check that this is valid even though the parametrization of C is not necessarily simple or regular). Adding these areas together (and suppressing the input variable) yields

$$\begin{aligned}
A + \pi r^2 &= \int_0^l (xy' - \tilde{y}x') \, dt \leq \int_0^l |xy' - \tilde{y}x'| \, dt = \int_0^l \sqrt{(xy' - \tilde{y}x')^2} \, dt \\
&= \int_0^l \sqrt{x^2 y'^2 - 2xy' \tilde{y}x' + \tilde{y}^2 x'^2} \, dt \\
&= \int_0^l \sqrt{(x^2 + \tilde{y}^2)(x'^2 + y'^2) - (xx' + \tilde{y}y')^2} \, dt \\
&\leq \int_0^l \sqrt{(x^2 + \tilde{y}^2)(x'^2 + y'^2)} \, dt \\
&= \int_0^l \sqrt{x^2 + \tilde{y}^2} \, dt && \text{because } \gamma \text{ is unit-speed} \\
&= \int_0^l |\beta(t)| \, dt = \int_0^l r \, dt = lr && \text{because } \beta(t) = (x(t), \pm \sqrt{r^2 - x(t)^2}).
\end{aligned}$$

In summary, $A + \pi r^2 \leq lr$. We now use the fact that the geometric mean of two positive numbers is bounded below by the arithmetic mean:

$$\sqrt{A}\sqrt{\pi r^2} \leq \frac{1}{2}(A + \pi r^2) \leq \frac{1}{2}lr.$$

This gives that $l^2 \geq 4\pi A$, as desired.

It remains to discuss the equality case in which $l^2 = 4\pi A$. In this case, all inequalities above become equalities. In particular, $(xx' + \tilde{y}y')^2 = 0$. Therefore,

$$0 = xx' + \tilde{y}y' = \langle (x, \tilde{y}), (x', y') \rangle = \langle \beta, \gamma' \rangle.$$

Since $|\beta|$ is constant, we also know that $\langle \beta, \beta' \rangle = 0$. Since both β' and γ' are orthogonal to β , they must be parallel to each other. But since they have identical first components (namely x'), this implies that they are equal to each other, at least when their common first component x' is nonzero.

In summary, $\beta'(t) = \gamma'(t)$ for all $t \in [0, l]$ at which $\gamma'(t)$ is not vertical. By continuity, the same must be true at isolated times when $\gamma'(t)$ is vertical. In fact, all such times must be isolated, for otherwise γ would be vertical on an interval, so β would have zero speed on that interval, yet would have unit speed everywhere it did not have zero speed, contradicting the continuity of its speed.

Thus, $\beta'(t) = \gamma'(t)$ for all $t \in [0, l]$. Since antiderivatives are unique up to an additive constant, $\gamma = \beta + \mathbf{w}$ for some constant vector $\mathbf{w} \in \mathbb{R}^2$. In other words, C is a translation of the circle. \square

EXERCISES

EXERCISE 2.28. If $l, A \in \mathbb{R}$ are positive numbers for which $l^2 \geq 4\pi A$, prove that there exists a simple closed plane curve with length l and area A .

EXERCISE 2.29. For $p > q > 0$, consider the ellipse $\gamma(t) = (p \cos t, q \sin t)$, exactly as in Exercise 2.4 on page 77. Let A denote its area and l its length.

- (1) Use Green's theorem to calculate A in terms of p and q .
- (2) Set $q = 1$, and use a computer algebra system to plot the graph of $\frac{4\pi A}{l^2}$ as a function of p . *COMMENT: a computer is necessary because the arc-length integral does not evaluate to an elementary closed-form expression for general p .*

EXERCISE 2.30. Use Green's theorem to find area of the region bounded by the x -axis and the trace of the curve

$$\gamma(t) = (t - \sin(t), 1 - \cos(t)), \quad t \in [0, 2\pi].$$

COMMENT: This curve is called a "cycloid" and is illustrated in the next section.

EXERCISE 2.31.

- (1) Show that the line integral of $\mathbf{F}(x, y) = (-y, x)$ along the line segment from (x_1, y_1) to (x_2, y_2) equals $x_1 y_2 - x_2 y_1$.
- (2) If C is a polygon with vertices denoted by $(x_1, y_1), \dots, (x_n, y_n)$ (ordered counterclockwise), prove that the area A of the polygon is given by

$$2A = (x_1 y_2 - x_2 y_1) + (x_2 y_3 - x_3 y_2) + \cdots + (x_{n-1} y_n - x_n y_{n-1}) + (x_n y_1 - x_1 y_n).$$

EXERCISE 2.32. Describe the history of isoperimetric problems including alternative proofs. An excellent reference is [3].

EXERCISE 2.33. A **planimeter** is a mechanical drafting instrument used to determine the area enclosed in a region; see the figure on page 61. Describe how these devices work and how their function is related to Green's theorem.



6. Huygens's Tautochrone Clock (Optional)

For the major countries of Europe, the seventeenth and eighteenth centuries represented an era of naval exploration. The main impediment to navigation at sea was the **longitude problem**—while it was relatively easy to establish one's latitude at sea, there was no effective known method to determine one's longitude. This limitation resulted in countless maritime disasters. As the lost life and lost gold piled up, several countries offered prizes and established observatories and scientific centers devoted to solving the problem. For example, in 1714, the British government established a *Board of Longitude* and offered a *longitude prize* of 20,000 pounds, writing:

The discovery of the longitude is of such consequence to Great Britain for the safety of the navy and merchant ships as well as for the improvement of trade that for want thereof many ships have been retarded in their voyages, and many lost...

Potential solution methods based on astronomical observations were championed by Galileo, Newton, Halley, and others. Work in this direction led to several key scientific discoveries including the speed of light. The main alternative approach involved attempting to build a better clock. If a clock could be constructed that maintained accurate time during long voyages at sea, then longitude could be determined each day from the exact time of sunrise or high noon.

Christiaan Huygens is credited with inventing pendulum clocks, which could keep accurate time on land but not at sea. The motion of waves rocking a ship would render a pendulum erratic—on some swings, the pendulum bob would traverse a wider circular arc than others. A wider swing takes slightly more time than a narrow one, eventually leading to inaccurate readings.

In fact, Huygens was the first to prove that a circular arc is not quite isochronous (wide swings take slightly more time than narrow swings). The best way to model this phenomenon is to forget about the string. Pretend that the circular arc traced by the pendulum bob is an actual physical track (colored blue in Fig. 2.32) and that the bob is a frictionless object sliding down the track. Huygens proved that the time to reach the bottom increases when the bob begins higher up the track. This might sound obvious, but it's not true of every track shape. He discovered a track shape with the property that the time to reach the bottom is independent of the object's starting position on the track. He called this shape a **tautochrone curve** (Greek for “same time”).

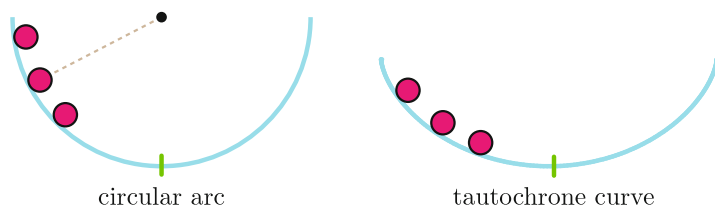


FIGURE 2.32. On the circular arc, the higher object takes the most time to slide to the bottom. On the tautochrone curve, all objects take the same amount of time

Huygens's next idea was to build bumpers against which the pendulum string would wrap, causing the pendulum bob to traverse a tautochrone curve rather than a circular arc; see Fig. 2.33. He constructed a **tautochrone clock** based on this design, with hopes of solving the longitude problem. He believed that his clock would keep accurate time at sea because all pendulum swings would take the same amount of time, even though the waves would cause some swings to be wider than others. Unfortunately, his clock did not function as accurately as he hoped on its trial voyage. Perhaps the added friction of the string against the bumpers offset the theoretical advantages, or perhaps the storms were severe enough to cause the bob to jerk about.

The longitude prize was eventually awarded to John Harrison for designing a seaworthy clock based on springs and balances.²

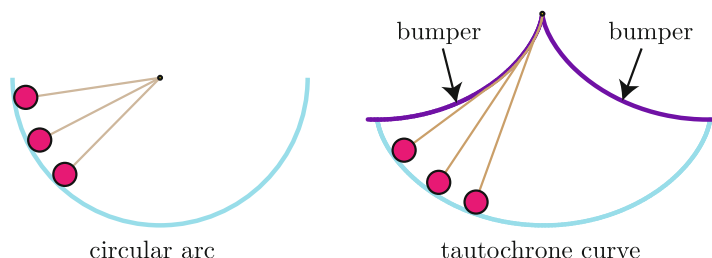


FIGURE 2.33. The pendulum string wraps against the bumpers, causing the bob to traverse a tautochrone curve

Although Huygens failed to win the longitude prize, the mathematics he invented to design his tautochrone clock survived and later found other interesting applications. The remainder of this section is devoted to discussing this mathematics and some of these applications.

The mathematical story begins with the *cycloid*. Imagine a wheel of radius 1 initially centered at $(0, 1)$, so it is tangent to the x -axis at the origin. Imagine marking this point of tangency on the wheel with a chalk mark and then letting the wheel roll without slipping along the x -axis. The curve traced by the chalk mark is called the **cycloid**; see Fig. 2.34. Parametrizing the cycloid is a simple matter of adding together the two purple vectors:

$$(2.5) \quad \gamma(t) = (t, 1) + (-\sin(t), -\cos(t)) = (t - \sin(t), 1 - \cos(t)).$$

The cycloid is regular on the domain $(0, 2\pi)$, which corresponds to the portion shown in Fig. 2.34.

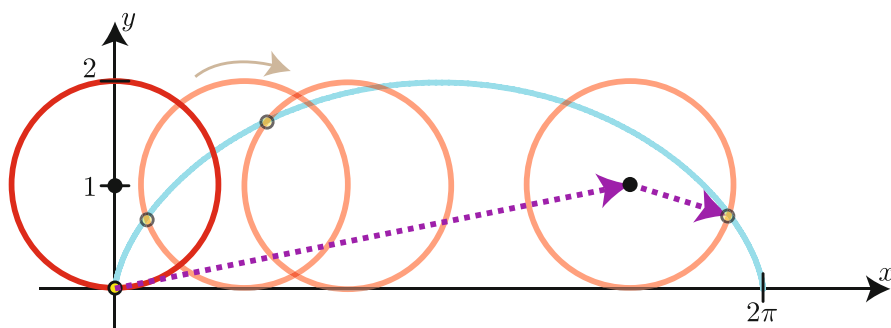


FIGURE 2.34. A cycloid is the path traced by a chalk mark on the edge of a wheel that rolls along a straight line

²We recommend [7] for the history of the longitude problem and John Harrison's story.

The cycloid is turned upside down by the rigid motion $(x, y) \mapsto (x, -y)$. The resultant **inverted cycloid** is parametrized as $\hat{\gamma}(t) = (t - \sin(t), \cos(t) - 1)$. Huygens proved that this is a solution to the tautochrone problem:

THEOREM 2.40.

The inverted cycloid, $\hat{\gamma}(t) = (\underbrace{t - \sin(t)}_x, \underbrace{\cos(t) - 1}_y)$, $t \in (0, 2\pi)$, is a tautochrone curve.

PROOF. Consider an object beginning at $\hat{\gamma}(t_0) = (x_0, y_0)$ and sliding under the influence only of gravity (without friction) to the bottom of the curve $\hat{\gamma}(\pi)$. We must show that the time required is independent of t_0 . Since the right half of the curve is the mirror image of the left, it suffices to assume that $t_0 \in (0, \pi)$.

Our first job is to relate the following variables that change as the curve is traversed: t, x, y, s, T . Here, s is the arc-length parameter defined as $s(t) = \int_{t_0}^t |\hat{\gamma}'(t)| dt$, while T is the time parameter, so that $T(t)$ denotes the time required to slide from position $\hat{\gamma}(t_0)$ to position $\hat{\gamma}(t)$. Notice that

$$\left(\frac{ds}{dt}\right)^2 = \left|\frac{d\hat{\gamma}}{dt}\right|^2 = \left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 = (1 - \cos(t))^2 + (-\sin(t))^2 = 2 - 2\cos(t).$$

Kinetic energy is traded for potential energy such that the following is always true:

$$(2.6) \quad \frac{1}{2}mv^2 = mg(y_0 - y),$$

where g is the Earth's gravitational constant (Exercise 2.35). Here v is the object's speed, measured as

$$\frac{ds}{dT} = v = \sqrt{2g(y_0 - y)}.$$

The chain rules says that $\frac{ds}{dt} = \frac{ds}{dT} \cdot \frac{dT}{dt}$. Solving for $\frac{dT}{dt}$ gives

$$\frac{dT}{dt} = \frac{\left(\frac{ds}{dt}\right)}{\left(\frac{ds}{dT}\right)} = \frac{\sqrt{2 - 2\cos(t)}}{\sqrt{2g(y_0 - y)}} = \frac{\sqrt{2 - 2\cos(t)}}{\sqrt{2g(\cos(t_0) - \cos(t))}}.$$

So the time required to reach the bottom is

$$T(\pi) = \int_{t_0}^{\pi} \left(\frac{dT}{dt}\right) dt = \int_{t_0}^{\pi} \frac{\sqrt{2 - 2\cos(t)} dt}{\sqrt{2g(\cos(t_0) - \cos(t))}} = \frac{1}{\sqrt{g}} \int_{t_0}^{\pi} \sqrt{\frac{1 - \cos(t)}{\cos(t_0) - \cos(t)}} dt.$$

Using integration tricks or computer assistance, this final integral can be shown to be independent of t_0 , and it evaluates to π , so we have $T(\pi) = \frac{\pi}{\sqrt{g}}$. \square

Bernoulli and Euler later proved the converse: every tautochrone curve is a segment of the inverted cycloid (possibly resized or translated).

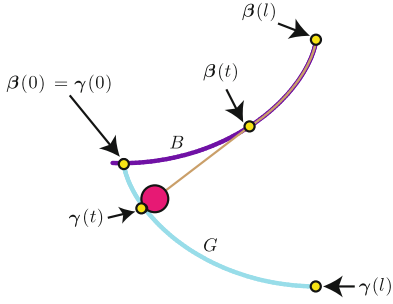


FIGURE 2.35. G is an involute of B with $(\lambda_0 = 0)$

Huygens's second mathematical problem was to determine the shape of the bumpers. For this, let $\beta : [0, l] \rightarrow \mathbb{R}^2$ be a parametrized plane curve (the bumper curve) and let $\gamma : [0, l] \rightarrow \mathbb{R}^2$ be the induced parametrization of the path that the pendulum bob will follow. It is simplest to imagine that initially at $t = 0$, the string is completely wrapped along the bumper curve, ending at $\beta(0)$, and is about to begin unwrapping. So the initial position of the pendulum bob is $\gamma(0) = \beta(0)$. Since the trace of γ will depend only on the trace of β , we can assume that β is parametrized by

arc length. At every $t \in [0, l]$, the length of unwrapped string equals t , so the bob's position has distance t from $\beta(t)$ in the direction of $-\beta'(t)$; see Fig. 2.35. That is, $\gamma(t) = \beta(t) - t\beta'(t)$.

The following definition slightly generalizes the above discussion by allowing an initial tail of unwrapped string with length λ_0 .

DEFINITION 2.41.

Let B be an oriented plane curve with nonzero curvature. An *involute* of B is a plane curve G that has a parametrization $\gamma : (0, l) \rightarrow \mathbb{R}^2$ of the form

$$(2.7) \quad \gamma(t) = \beta(t) - (t + \lambda_0)\beta'(t),$$

where $\beta : (0, l) \rightarrow \mathbb{R}^2$ is a unit-speed parametrization of B , and $\lambda_0 \in \mathbb{R}$ is a constant with $\lambda_0 \notin (-l, 0)$.

Notice that $\gamma'(t) = -(t + \lambda_0)\beta''(t)$, which is nonzero because $\lambda_0 \notin (-l, 0)$ and because B has nonzero curvature, so γ is regular. For example, the nonzero curvature hypothesis does not allow B to be a straight line, because then G would be a single point. The case $\lambda_0 \geq 0$ corresponds to an “unwrap string,” as was previously illustrated in Fig. 2.35. The case $\lambda_0 \leq -l$ corresponds to a string that is wrapped along B .

If $\beta : (0, l) \rightarrow \mathbb{R}^2$ is a regular (not necessarily unit-speed) parametrization of B , one can avoid reparametrizing it by arc length by replacing Eq. 2.7 with

$$(2.8) \quad \gamma(t) = \beta(t) - (s(t) + \lambda_0) \frac{\beta'(t)}{|\beta'(t)|},$$

where $s(t) = \int_0^t |\beta'(u)| du$ is the arc-length function.

Huygens really needed to solve the inverse problem: find the bumper curve B whose involute is the tautochrone curve G . He solved this by understanding the inverse relationship between involutes and evolutes. Here is a more precise formulation of Eq. 1.8 on page 30:

DEFINITION 2.42.

Let G be an oriented plane curve parametrized as $\gamma : (0, l) \rightarrow \mathbb{R}^2$. Assume for all $t \in (0, l)$ that $\kappa(t) \neq 0$ and $\kappa'(t) \neq 0$. Let \mathbf{n} denote the unit normal to γ . The **evolute**, B , of G is the plane curve with parametrization $\beta : (0, l) \rightarrow \mathbb{R}^2$ given by

$$\beta(t) = \gamma(t) + \frac{1}{\kappa(t)} \mathbf{n}(t) = \text{the center of the osculating circle of } G \text{ at } \gamma(t).$$

You can check that $\beta'(t) = -\frac{\kappa'(t)}{\kappa(t)^2} \mathbf{n}(t)$, so the hypothesis that κ' is nonvanishing ensures that β is regular. For example, this hypothesis does not allow G to be a circle, because then B would be a single point—its center. The variable choices have already hinted at the following inverse relationship between involutes and evolutes:

PROPOSITION 2.43.

- (1) With the assumptions of Definition 2.41, B is the evolute of G .
- (2) With the assumptions of Definition 2.42, G is an involute of B .

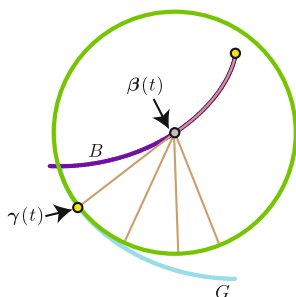


FIGURE 2.36. Pinning the string to $\beta(t)$ at time t makes the pendulum bob begin traversing the osculating circle of G at $\gamma(t)$

Before beginning the proof, we mention a visual reason to believe claim (1) that B is the evolute of G in Fig. 2.35. If the string were pinned to $\beta(t)$ at time t , then the pendulum bob would begin traversing the circle of radius $t + \lambda_0$ centered at $\beta(t)$, colored green in Fig. 2.36. This circle approximates G well at $\gamma(t)$ (in fact, it is the osculating circle) because second derivatives don't detect the difference between pinning and not pinning. This suggests that $\beta(t)$ is the center of the osculating circle of G at $\gamma(t)$.

PROOF OF PROPOSITION 2.43(2). To prove part (2), let G be an oriented plane curve parametrized as $\gamma : (0, l) \rightarrow \mathbb{R}^2$, and assume that its curvature function satisfies $\kappa(t) \neq 0$ and $\kappa'(t) \neq 0$ for all $t \in (0, l)$. We will consider the case that κ' is strictly negative (the other case, that κ' is strictly

positive, is handled similarly). The evolute of G is the curve B parametrized as

$$(2.9) \quad \beta(t) = \gamma(t) + \frac{1}{\kappa(t)} \mathbf{n}(t).$$

As previously mentioned, $\beta'(t) = -\frac{\kappa'(t)}{\kappa(t)^2} \mathbf{n}(t)$. In particular, $\frac{\beta'(t)}{|\beta'(t)|} = \mathbf{n}(t)$; see Fig. 2.37. Thus, Eq. 2.9 can be rewritten as

$$\gamma(t) = \beta(t) - \frac{1}{\kappa(t)} \frac{\beta'(t)}{|\beta'(t)|}.$$

Comparing to Eq. 2.8, to prove that G is an involute of B , it will suffice to verify that

$$\frac{1}{\kappa(t)} = s(t) + \lambda_0$$

for some $\lambda_0 \notin (-l, 0)$, where $s(t) = \int_0^t |\beta'(u)| du$ is the arc-length function of β . For this, observe that

$$\frac{d}{dt} \left(\frac{1}{\kappa(t)} - s(t) \right) = -\frac{\kappa'(t)}{\kappa(t)^2} - |\beta'(t)| = -\frac{\kappa'(t)}{\kappa(t)^2} - \left| \frac{\kappa'(t)}{\kappa(t)^2} \right| = 0.$$

Thus, $\frac{1}{\kappa(t)} - s(t)$ is a constant, and by considering the time $t = 0$, one confirms that this constant is ≥ 0 .

The proof of part (1) is left to the reader in Exercise 2.44. \square

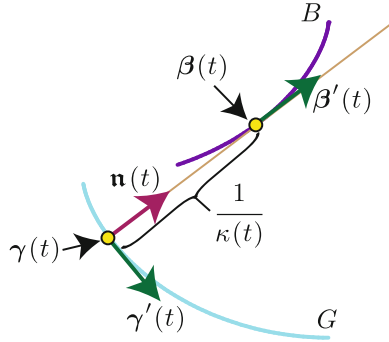


FIGURE 2.37. B is the evolute of G ; therefore, G is an involute of B

In light of Proposition 2.43, Huygens could choose his bumper curve to be the evolute of the tautochrone curve (the inverted cycloid). Figure 2.38 zooms out to show more of the curves G and B from Fig. 2.33, leading one to guess the following:

PROPOSITION 2.44.

The evolute of the inverted cycloid is a translation of the inverted cycloid.

PROOF. Exercise 2.46. □

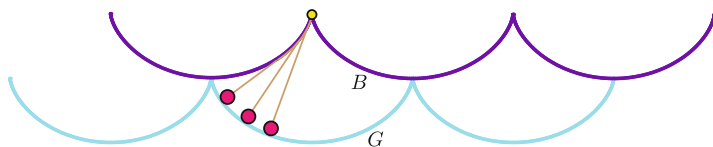


FIGURE 2.38. The evolute, B , of G is a translation of G

The intended interpretation of Proposition 2.44 should be clear from Fig. 2.38. But to be pedantic about satisfying the hypotheses of Definition 2.42, we should restrict the inverted cycloid to the domain $(0, \pi)$; in other words, we should consider only the left half of one period of the inverted cycloid, in which case its evolute is a translation of the right half.

Huygens's tautochrone clock did not win the longitude prize, but the underlying mathematics has found other interesting applications through the centuries. For example, evolutes are important in the field of optics. To understand the relevance, consider an alternative method of illustrating the evolute of the inverted cycloid G . Figure 2.39 (left) shows G together with its *normal line* (the line in the direction of \mathbf{n}) at 40 points along G . These 40 purple lines intersect in a pattern that tricks your eye into seeing the curved shape of the evolute, B , of G . If the purple lines are light rays, then B is a curve of focused brightness called a *caustic* (Latin for “burn,” because focused sun rays can burn). Intuitively, the normal lines of G should focus along B because G is well approximated at each point by its osculating circle, whose normal lines all focus on its center.

To describe this focusing more precisely, let \mathcal{F} denote the family of all normal lines (not just at 40 points, but at all points of G). Then B is called the **envelope** of \mathcal{F} , which means the unique curve whose tangent lines are all members of \mathcal{F} . In computer graphics, modern ray-tracing software is capable of rendering envelopes as bright caustics. In Fig. 2.39 (right), the bright caustic is the envelope of a certain family, \mathcal{F} , of lines emanating from the curved mirror. In contrast to our cycloid example, these lines are not normal to the mirror, but rather point in the directions that appropriately model how light rays from a single source would bounce off the mirror.

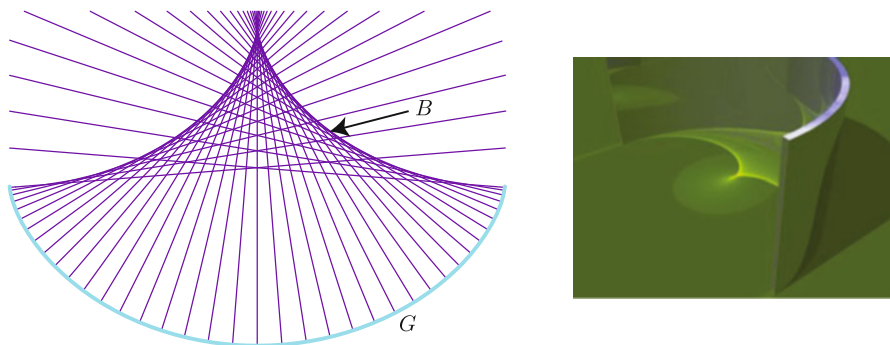


FIGURE 2.39. *Left:* the evolute, B , of G is the envelope of its normal lines. *Right:* the bright caustic is the envelope of the light rays reflected off the curved mirror

Another application involves the shape of gears. Old-fashioned gears with triangular or rectangular teeth clanged against each other with excessive vibration and noise. To solve this problem, Euler invented “involute gears” whose edges have the shape of a segment of an involute of a circle. This shape ensures that a pair of interlocking teeth meet each other at a single contact point that varies along a straight line as the gears turn, greatly reducing vibration and wear. Most gears in use today are involute gears.

We hope that the illustrations in this section were helpful, but don’t settle for still images. It is easy to find online animated illustrations of these concepts. We particularly recommend the animations found on the Wikipedia pages for “cycloid,” “tautochrone curve,” “involute,” “evolute,” and “involute gear.” Also check out the animated tautochrone clock on the Wolfram MathWorld “Tautochrone Problem” page.

EXERCISES

EXERCISE 2.34. The cycloid in Eq. 2.5 can be generalized as

$$\gamma(t) = (at - b \sin(t), a - b \cos(t))$$

for arbitrary constants $a, b > 0$.

- (1) Interpret in terms of the path of a chalk mark on (not necessarily the edge of) a rolling wheel.
- (2) Use a computer graphing application to plot the graph for several choices with $a < b$ and several choices with $a > b$.

EXERCISE 2.35. How can Eq. 2.6 be derived from the conservation of energy law in Example 2.28 on page 89?

EXERCISE 2.36. Verify the claim after Definition 2.42 that $\beta'(t) = -\frac{\kappa'(t)}{\kappa(t)^2} \mathbf{n}(t)$.

EXERCISE 2.37. Definition 2.42 defines only the evolute of a *plane* curve. The evolute of a *space* curve $\gamma : (0, l) \rightarrow \mathbb{R}^3$ can be defined by the same formula: $\beta(t) = \gamma(t) + \frac{1}{\kappa(t)}\mathbf{n}(t)$. Modify the formula for $\beta'(t)$ from the previous exercise to make it valid for space curves. Describe the general condition under which β is regular for space curves. Determine the evolute of the helix from Example 1.3 on page 2.

EXERCISE 2.38. Let $\beta : (0, l) \rightarrow \mathbb{R}^2$ be a unit-speed plane curve with nonzero curvature. Let $\gamma : (0, l) \rightarrow \mathbb{R}^2$ be a regular plane curve such that for all $t \in (0, l)$, $\gamma(t)$ meets the tangent line to $\beta(t)$ at a right angle (that is, the tangent line to the trace of β at $\beta(t)$ contains $\gamma(t)$ and is orthogonal to $\gamma'(t)$). Prove that γ is an involute of β .

EXERCISE 2.39. Prove that the evolute of the parabola $y = x^2$ is the graph of $y = \frac{1}{2} + 3 \left| \frac{x}{4} \right|^{2/3}$, illustrated in Fig. 2.40.

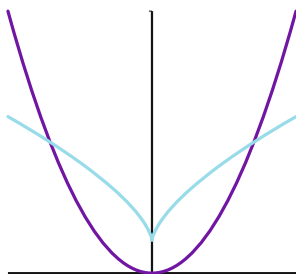


FIGURE 2.40. The evolute of the parabola

EXERCISE 2.40. Let $p > q > 0$ and consider the ellipse $\gamma(t) = (p \cos(t), q \sin(t))$, $t \in [0, 2\pi]$. prove that its evolute is an astroid (defined in Exercise 1.12 on page 8.)

EXERCISE 2.41. For the parabola and ellipse from the previous two exercises, construct a graph similar to Fig. 2.39 showing a family of normal lines intersecting to form the shape of the evolute.

EXERCISE 2.42. Describe an infinite family of plane curves that all have the same evolute.

EXERCISE 2.43. Use a computer graphing application to plot several involutes of a circle of radius 1.

EXERCISE 2.44. Prove Proposition 2.43(1).

EXERCISE 2.45. If the string of Huygens's tautochrone clock is lengthened, will the bob still traverse a tautochrone curve?

EXERCISE 2.46. Prove Proposition 2.44.

EXERCISE 2.47. An **epicycloid** (respectively **hypocycloid**) is the path followed by a chalk mark on the rim of a circle of radius b that rolls without slipping outside (respectively inside) a circle of radius a , as illustrated

in Fig. 2.41. Find formulas for these curves and use a computer graphing application to plot them for several choices of the constants a, b with $b < a$.

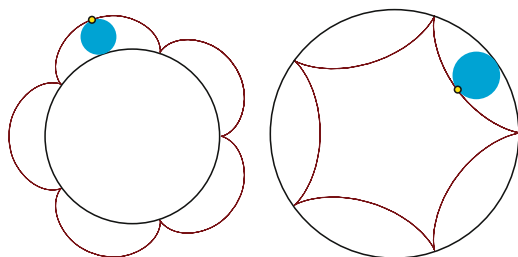


FIGURE 2.41. An epicycloid (*left*) and a hypocycloid (*right*)



Differential Geometry of Curves and Surfaces

Tapp, K.

2016, VIII, 366 p. 186 illus. in color., Hardcover

ISBN: 978-3-319-39798-6