

CHAPTER IV
INTEGRATION

§4.1 The Cantor Set

The subject of this chapter is the measurement of the areas of subsets of the plane \mathbf{R}^2 . The areas of elementary geometric figures, such as squares, rectangles, and triangles, are already known to us. By *known to us* we mean that, e.g., by defining the area of a rectangle to be the product of the lengths of its sides, we obtain quantities that agree with our intuition. Since every right-angle triangle is half a rectangle, the areas of right-angle triangles are also known to us. Similarly, we can obtain the area of a general triangle.

How does one approach the problem of measuring the area of an unfamiliar figure or subset of \mathbf{R}^2 , say a subset that cannot be broken up into triangles? For example, how does one measure the area of the unit disk

$$D = \{(x, y) : x^2 + y^2 < 1\}?$$

One solution is to arbitrarily define the area of D to equal whatever one feels is right. The Egyptian book of Ahmes (~ 1900 B.C.) states that the area of D is $(16/9)^2$. In the Indian Śulbastras (written down ~ 500 B.C.), the area of D is taken to equal $(26/15)^2$. Albrecht Dürer (1471–1528) of Nuremberg solved a related problem which amounted to taking the area of D to equal $25/8$.

Which of these answers should we accept as the area of D ? If we treat these answers as *estimates* of the area of D , then, in our minds, we must have the presumption that such a quantity — the area of D — has a meaningful existence. In that case, we have no way of judging the merit of an estimate except by the quality of the reasoning leading to it.

Realizing this, by reasoning that remains perfectly valid today, Archimedes (~ 250 B.C.) carefully established,

$$\frac{223}{71} < \text{area}(D) < \frac{22}{7}.$$

In §4.4, we show that $\text{area}(D) = \pi$, where π , “Archimedes’ constant”, is the real number defined in §3.5.

At the basis of the Greek mathematicians’ computations of area was the *method of exhaustion*. This asserted that the area of a set $A \subset \mathbf{R}^2$ could be computed as the limit of areas of a sequence of inscribed sets (A_n) that filled out more and more of A as $n \nearrow \infty$ (Figure 4.1). Nevertheless the Greeks were apparently uncomfortable with the concept of infinity and never used this method as stated. Instead, for example, in dealing with D , Archimedes used inscribed and circumscribed polygons

fig4.1.eps

FIGURE 4.1. The method of exhaustion.

with 96 sides to obtain the above result. He never explicitly passed to the limit. It turns out, however, that the method of exhaustion is so important to integration that in §4.5 we give a careful derivation of it.

Now, the unit disk is not a totally unfamiliar set to the reader. But, if we are presented with some genuinely unfamiliar subset C , the situation changes, and we may no longer have any clear conception of the area of C . If we are unable to come up with a procedure leading us to the area, then, we may be forced to reexamine our intuitive notion of area. In particular, we may be led to the conclusion that the “true area” of C may have no meaning. Let us describe such a subset.

Recall (§2.1) that an *open rectangle* is a subset of the form $(a, b) \times (c, d)$ where a or c may equal $-\infty$ and b or d may equal $+\infty$. A *compact rectangle* is a subset of the form $[a, b] \times [c, d]$.

Let C_0 denote the compact unit square $[0, 1] \times [0, 1]$. Divide C_0 into nine equal subsquares and take out from C_0 all but the four compact corner subsquares. Let C_1 be the remainder, i.e., the union of the four remaining compact subsquares. Repeat this process with each of the four subsquares. Divide each subsquare into nine equal compact sub-subsquares and take out, in each subsquare, all but the four compact corner sub-subsquares. Call the union of the remaining sixteen sub-subsquares C_2 . Continuing in this manner yields a sequence $C_0 \supset C_1 \supset C_2 \supset \dots$. The *Cantor set* is the common part (Figure 4.2) of all these sets, i.e., their intersection

$$C = \bigcap_{n=1}^{\infty} C_n.$$

At first glance it is not clear that C is not empty. But $(0, 0) \in C$! Moreover the sixteen corners of the set C_1 are in C . Similarly, any corner of any subsquare, at any level, lies in C . But the set of such points is countable (§1.7), and it turns out that there is much more: There are as many points in C as there are in the unit square C_0 . In particular C is uncountable.

fig4.2.eps

FIGURE 4.2. The Cantor set.

To see this, recall the concept of ternary expansions (§1.6). Let $a \in [0, 1]$. We say that

$$a = .a_1a_2\dots$$

is the ternary expansion of a if the naturals a_n are ternary digits 0, 1, 2, and

$$a = \sum_{n=1}^{\infty} a_n 3^{-n}.$$

Now, let $(a, b) \in C_0$, and let

$$a = .a_1a_2\dots$$

and

$$b = .b_1b_2\dots$$

be ternary expansions of a and b . If $a_1 \neq 1$ and $b_1 \neq 1$, then, (a, b) is in C_1 . Similarly, in addition, if $a_2 \neq 1$ and $b_2 \neq 1$, $(a, b) \in C_2$. Continuing in this manner, we see that, if $a_n \neq 1$ and $b_n \neq 1$ for all $n \geq 1$, $(a, b) \in C$. Conversely, $(a, b) \in C$ implies that there are ternary expansions of a and b as stated. Thus, $(a, b) \in C$ iff a and b have ternary expansions in which the digits are equal to 0 or 2.

Now, although some reals may have more than one ternary expansion, a real a cannot have more than one ternary expansion $.a_1a_2\dots$ where $a_n \neq 1$ for all $n \geq 1$ because any two ternary expansions yielding the same real must have their n th digits differing by 1 for some $n \geq 1$ (Exercise 2 of §1.6). Thus, the mapping

$$(a, b) = (.a_1a_2a_3\dots, .b_1b_2b_3\dots) \mapsto \left(\sum_{n=1}^{\infty} a'_n 2^{-n}, \sum_{n=1}^{\infty} b'_n 2^{-n} \right),$$

where $a'_n = a_n/2$, $b'_n = b_n/2$, $n \geq 1$, is well defined. Since any real in $[0, 1]$ has a binary expansion, this mapping is a surjection of the Cantor set C onto the unit square C_0 . Since C_0 is uncountable (Exercise 4 of §1.7), we conclude that C is uncountable (§1.7).

The difficulty of measuring the size of the Cantor set underscores the difficulty in arriving at a consistent notion of area. Above, we saw that the Cantor set is uncountable. In this sense, the Cantor set is “big”. On the other hand, note that the areas of the subsquares removed from C_0 to obtain C_1 sum to $5/3^2$. Similarly, the areas of the sub-subsquares removed from C_1 to obtain C_2 sum to $20/9^2$. Similarly, at the next stage, we remove squares with areas summing to $80/27^2$. Thus, the sum of the areas of all the removed squares is

$$\frac{5}{9} + \frac{20}{9^2} + \frac{80}{27^2} + \dots = \frac{5}{9} \left(1 + \frac{4}{9} + \left(\frac{4}{9} \right)^2 + \dots \right) = 1.$$

Since C is the complement of all these squares in C_0 and C_0 has area 1, the area of C is $1 - 1 = 0$. Thus, in the sense of area, the Cantor set is “small”.

This argument is perfectly reasonable, except for one aspect. We are assuming that areas can be added and subtracted in the usual manner, even when there are *infinitely* many sets involved. In §4.2, we show that, with an appropriate definition of area, this argument can be modified to become correct, and the area of C is in fact zero.

Another indication of the smallness of C is the fact that C has no interior. To explain this, given any set $A \subset \mathbf{R}^2$, let us say that A has interior if we can fit some open rectangle Q within A , i.e., $Q \subset A$. If we cannot fit an open rectangle, no matter how small, within A , then, we say that A has no interior. For example, the unit disk has interior but a line segment has no interior. The Cantor set C has no interior, because there is a point in every open rectangle whose coordinate ternary expansions contain at least one digit 1. Alternatively, if C contained a rectangle Q , then, the area of C would be at least as much as the area of Q , which is positive. But we saw above that the area of C equals zero.

Since this reasoning applies to any set, we see that if $A \subset \mathbf{R}^2$ has interior, then, the area of A is positive. The surprising fact is that the converse of this statement

is false. There are sets $A \subset \mathbf{R}^2$ that have positive area but have no interior. Such a set is described in Exercise 2.

These issues are discussed to point out the existence of unavoidable phenomena involving area where things do not behave as simply as triangles. In the first three decades of this century, these issues were finally settled. The solution to *the problem of area*, analyzed extensively by Archimedes more than two thousand years ago, can now be explained in a few pages. Why did it take so long for the solution to be discovered? It should not be too surprising that one missing ingredient was the completeness property of the set of real numbers, the importance of which was not fully realized until the nineteenth century.

Exercises 4.1.

1. Let $C_0 = [0, 1] \times [0, 1]$ denote the unit square, and let C'_1 be obtained by throwing out from C_0 the middle subrectangle $(1/3, 2/3) \times [0, 1]$ of width $1/3$ and height 1. Then, C'_1 consists of two compact subrectangles. Let C'_2 be obtained from C'_1 by throwing out, in each of the subrectangles, the middle sub-subrectangles $(1/9, 2/9) \times [0, 1]$ and $(7/9, 8/9) \times [0, 1]$, each of width $1/3^2$ and height 1. Then, C'_2 consists of four compact sub-subrectangles. Similarly C'_3 consists of eight compact sub-sub-subrectangles, obtained by throwing out from C'_2 the middle sub-sub-subrectangles of width $1/3^3$ and height 1. Continuing in this manner, we have $C'_1 \supset C'_2 \supset C'_3 \supset \dots$. Let $C' = \bigcap_{n=1}^{\infty} C'_n$. Show that $\text{area}(C') = 0$ and C' has no interior.

2. Fix a real $0 < \alpha < 1$ (e.g., $\alpha = .7$) and let $C_0 = [0, 1] \times [0, 1]$ be the unit square. Let C_1^α be obtained from C_0 by throwing out the middle subrectangle of width $\alpha/3$ and height 1. Then, C_1^α consists of two subrectangles. Let C_2^α be obtained from C_1^α by throwing out, in each of the subrectangles, the *middle* sub-subrectangles of width $\alpha/3^2$ and height 1. Then, C_2^α consists of four sub-subrectangles. Similarly C_3^α consists of eight sub-sub-subrectangles, obtained by throwing out from C_2^α the middle sub-sub-subrectangles of width $\alpha/3^3$ and height 1. Continuing in this manner, we have $C_1^\alpha \supset C_2^\alpha \supset C_3^\alpha \supset \dots$. Let $C^\alpha = \bigcap_{n=1}^{\infty} C_n^\alpha$. Show that $\text{area}(C^\alpha) > 0$, but C^α has no interior.

3. For $A \subset \mathbf{R}^2$ let

$$A + A = \{(x + x', y + y') : (x, y) \in A, (x', y') \in A\}$$

be the *set of sums*. Show that $C + C = [0, 2] \times [0, 2]$ (Exercise 1.6.6).

§4.2 Area

Recall that $(a, b) = \{x : a < x < b\}$ and $[a, b] = \{x : a \leq x \leq b\}$. An open rectangle is a subset (Figure 4.3) of the form $(a, b) \times (c, d)$, where a or c may equal $-\infty$ and b or d may equal ∞ . In particular, \mathbf{R}^2 is an open rectangle. We say that an open rectangle is *bounded* if a, b, c , and d are reals. A compact rectangle is a subset of the form $[a, b] \times [c, d]$ where a, b, c , and d are reals. In particular, the vertical line segment $\{a\} \times [c, d]$ is a compact rectangle. Similarly, so is a horizontal line segment. A single point is a compact rectangle. If $Q = (a, b) \times (c, d)$ is a bounded, nonempty, open rectangle, the compact rectangle $\bar{Q} = [a, b] \times [c, d]$ is its *compactification*. Note that an open rectangle may be empty, for example $(a, a) \times (c, d)$ is empty.

fig4.3.eps

FIGURE 4.3. A and B are rectangles, but C is not.

Let A be a subset of \mathbf{R}^2 . A *cover* of A is a sequence of sets (A_n) , such that A is contained in their union,

$$A \subset \bigcup_{n=1}^{\infty} A_n.$$

In a given cover, the sets (A_n) may overlap, i.e., intersect. If for some N , $A_n = \emptyset$ for $n > N$, we say that (A_1, \dots, A_N) is a *finite cover* (Figure 4.4).

fig4.4.eps

FIGURE 4.4. A finite cover.

A *paving* of A is a cover (Q_n) , where the sets Q_n , $n \geq 1$, are *open* rectangles. A *finite paving* is a finite cover that is also a paving (Figure 4.5). Every subset $A \subset \mathbf{R}^2$ has at least one (not very interesting) paving $Q_1 = \mathbf{R}^2$, $Q_2 = \emptyset$, $Q_3 = \emptyset$, \dots

fig4.5.eps

FIGURE 4.5. A finite paving.

For any nonempty, open rectangle $Q = (a, b) \times (c, d)$ or compact rectangle $Q = [a, b] \times [c, d]$, let

$$\|Q\| = (b - a) \cdot (d - c).$$

Then, $\|Q\|$ is a positive real or 0 or ∞ . We also take $\|\emptyset\| = 0$.

Let A be a subset of \mathbf{R}^2 . The *area* of A is defined by

$$\text{area}(A) = \inf \left\{ \sum_{n=1}^{\infty} \|Q_n\| : \text{all pavings } (Q_n) \text{ of } A \right\}.$$

This definition of area is at the basis of all that follows. It is necessarily complicated because it applies to all subsets A of \mathbf{R}^2 . As an immediate consequence of the definition, $\text{area}(\emptyset) = 0$. Similarly, the area of a finite vertical line segment A is zero since A can be covered by a thin open rectangle of arbitrarily small area.

In words, the definition says that to find the area of a set A , we cover A by a sequence Q_1, Q_2, \dots of open rectangles, measure the sum of their areas, and take this sum as an estimate for the area of A . Of course, we expect that this sum will be an overestimate of the area of A for two reasons. The paving may cover a superset of A , and we are not taking into account any overlaps when computing the sum. Then, we define the area of A to be the inf of these sums.

Of course, carrying out this procedure explicitly, even for simple sets A , is completely impractical. Because of this, we almost never use the definition directly to compute areas. Instead, as is typical in mathematics, we derive the elementary properties of area from the definition, and we use them to compute areas.

We emphasize that according to the above definitions, a rotated rectangle A is not a rectangle. Hence, $\|A\|$ is not defined unless the sides of the rectangle A are parallel to the axes. Nevertheless, we will see, below, that area is rotation-invariant. Hence, $\text{area}(A)$ turns out as expected.

Whether or not we can compute the area of a given set, the above definition applies consistently to every subset A . In particular, this is so whether A is a rectangle, a triangle, a smooth graph, or the Cantor set C . Now, let us derive the properties of area that follow immediately from the definition.

Since every open rectangle Q is a paving of itself, $\text{area}(Q) \leq \|Q\|$. Below, we obtain $\text{area}(Q) = \|Q\|$ for a rectangle Q . Until we establish this, we refer to $\|Q\|$ as the *naive area* of Q . Note that, although area is defined for every subset, naive area is defined only for (nonrotated) rectangles.

If (a, b) is a point in \mathbf{R}^2 and $A \subset \mathbf{R}^2$, the set

$$A + (a, b) = \{(x + a, y + b) : (x, y) \in A\}$$

is the *translate* of A by (a, b) . Then, $[A + (a, b)] + (c, d) = A + (a + c, b + d)$ and, for any rectangle Q , $\|Q + (a, b)\| = \|Q\|$. From this follows the *translation invariance of area*,

$$\text{area}[A + (a, b)] = \text{area}(A), \quad A \subset \mathbf{R}^2.$$

To see this, let (Q_n) be a paving of A . Then, $(Q_n + (a, b))$ is a paving of $A + (a, b)$, so

$$\text{area}[A + (a, b)] \leq \sum_{n=1}^{\infty} \|Q_n + (a, b)\| = \sum_{n=1}^{\infty} \|Q_n\|.$$

Since $\text{area}(A)$ is the inf of the sums on the right, $\text{area}[A + (a, b)] \leq \text{area}(A)$. Now, in this last inequality, replace, in order, (a, b) by $(-a, -b)$ and A by $A + (a, b)$. We obtain $\text{area}(A) \leq \text{area}[A + (a, b)]$. Hence, $\text{area}(A) = \text{area}[A + (a, b)]$, establishing translation invariance (Figure 4.6).

fig4.6.eps

FIGURE 4.6. $\text{area}(A) = \text{area}[A + (a, b)]$.

If $k > 0$ is real and $A \subset \mathbf{R}^2$, the set

$$kA = \{(kx, ky) : (x, y) \in A\}$$

is the *dilate* of A by k . Then, $k(cA) = (kc)A$ for k and c positive, and $\|kQ\| = k^2\|Q\|$ for every rectangle Q . From this follows the *dilation invariance of area*,

$$\text{area}(kA) = k^2 \cdot \text{area}(A), \quad A \subset \mathbf{R}^2.$$

To see this, let (Q_n) be a paving of A . Then, (kQ_n) is a paving of kA so

$$\text{area}(kA) \leq \sum_{n=1}^{\infty} \|kQ_n\| = k^2 \left(\sum_{n=1}^{\infty} \|Q_n\| \right).$$

Since $\text{area}(A)$ is the inf of the sums on the right, we obtain $\text{area}(kA) \leq k^2 \cdot \text{area}(A)$. Now, in this last inequality, replace, in order, k by $1/k$ and A by kA . We obtain $k^2 \cdot \text{area}(A) \leq \text{area}(kA)$. Hence, $\text{area}(kA) = k^2 \cdot \text{area}(A)$, establishing dilation invariance.

fig4.7.eps

FIGURE 4.7. $\text{area}(kA) = k^2 \cdot \text{area}(A)$.

Instead of dilation from the origin, we can dilate from any point in \mathbf{R}^2 . In particular, by elementary geometry, certain subsets such as rectangles, triangles, and parallelograms have well defined centers. Given a subset A with a center (a, b) , its *centered dilation* kA is the set (Figure 4.8) obtained by translating (a, b) to $(0, 0)$, dilating as above by the factor $k > 0$, then, translating $(0, 0)$ back to (a, b) . Then, $\text{area}(kA) = k^2 \cdot \text{area}(A)$ for centered dilations as well. For example, if $k < 1$, kA is A shrunk towards its center.

fig4.8.eps

FIGURE 4.8. Centered dilation.

The diagonal line segment $A = \{(x, y) : 0 \leq x \leq 1, y = x\}$ has zero area (Figure 4.9). To see this, choose $n \geq 1$ and let $Q_k = \{(x, y) : (k-1)/n < x < k/n, (k-1)/n < y < k/n\}$, $k = 1, \dots, n$. Then, (Q_1, \dots, Q_n) is almost a paving of A , except for the corner points of Q_k in A , $k = 1, \dots, n$, which are not covered. To remedy this, for $c > 1$, let cQ_k denote the centered dilations of Q_k , $k = 1, \dots, n$. Then, (cQ_1, \dots, cQ_n) is a paving of A . Hence, $\text{area}(A) \leq \|cQ_1\| + \dots + \|cQ_n\| = n \cdot c^2 \cdot \frac{1}{n^2} = c^2/n$. Since $n \geq 1$ may be arbitrarily large, we conclude that $\text{area}(A) = 0$. Similarly the area of any finite line segment is zero.

fig4.9.eps

FIGURE 4.9. Area of a diagonal line segment.

As with dilation, setting $-A = \{(-x, -y) : (x, y) \in A\}$, we have *reflection invariance of area*,

$$\text{area}(-A) = \text{area}(A), \quad A \subset \mathbf{R}^2,$$

and *monotonicity*,

$$\text{area}(A) \leq \text{area}(B), \quad A \subset B \subset \mathbf{R}^2.$$

Another property is *subadditivity*. For any cover (A_n) of a given set A ,

$$(2.1) \quad \text{area}(A) \leq \sum_{n=1}^{\infty} \text{area}(A_n).$$

Here, the sets A_n , $n \geq 1$, need not be rectangles. For future reference, we call the sum on the right side of (2.1) the *area of the cover* (A_n) .

In particular, since (A, B) is a cover of $A \cup B$,

$$\text{area}(A \cup B) \leq \text{area}(A) + \text{area}(B).$$

Similarly, (A_1, A_2, \dots, A_n) is a cover of $A_1 \cup A_2 \cup \dots \cup A_n$, so,

$$\text{area}(A_1 \cup A_2 \cup \dots \cup A_n) \leq \text{area}(A_1) + \text{area}(A_2) + \dots + \text{area}(A_n).$$

To obtain subadditivity, note that, if the right side of (2.1) is ∞ , there is nothing to show, since in that case (2.1) is true. Hence, we may safely assume $\text{area}(A_n) < \infty$ for all $n \geq 1$. Let $\epsilon > 0$, and, for each $k \geq 1$, choose a paving $(Q_{k,n})$ of A_k satisfying

$$(2.2) \quad \sum_{n=1}^{\infty} \|Q_{k,n}\| < \text{area}(A_k) + \epsilon 2^{-k}.$$

This is possible since $\text{area}(A_k)$ is the inf of sums of the form $\sum_{n=1}^{\infty} \|Q_n\|$. Then, the double sequence $(Q_{k,n})$ is a cover by open rectangles, hence, a paving of A . Summing (2.2) over $k \geq 1$, we obtain

$$\begin{aligned} \text{area}(A) &\leq \sum_{k=1}^{\infty} \left(\sum_{n=1}^{\infty} \|Q_{k,n}\| \right) \\ &\leq \sum_{k=1}^{\infty} (\text{area}(A_k) + \epsilon 2^{-k}) = \left(\sum_{k=1}^{\infty} \text{area}(A_k) \right) + \epsilon. \end{aligned}$$

Since $\epsilon > 0$ is arbitrary, subadditivity follows.

From subadditivity and monotonicity, we see that the area of any (infinite) line is zero and $\text{area}(Q) = \text{area}(\overline{Q})$ for any bounded open rectangle Q . Moreover, $\text{area}(Q) \leq \|Q\|$ for any compact rectangle Q .

THEOREM. *The area of a (compact or open) rectangle equals the product of the lengths of its sides.*

We see, below, that this holds for rotated rectangles as well. Before we derive this Theorem, we make some elementary observations concerning the naive area (Figure 4.10).

Let A be a compact rectangle. If $A = Q_1 \cup \dots \cup Q_N$ with $A = [a, b] \times [c, d]$, $Q_n = [a_{n-1}, a_n] \times [c, d]$, $n = 1, \dots, N$, and $a = a_0 < a_1 < \dots < a_N = b$, then, by definition of $\|\cdot\|$,

$$(2.3) \quad \|A\| = \|Q_1\| + \dots + \|Q_N\|.$$

Let us call such a cover (Q_1, \dots, Q_N) of A horizontal type I. Similarly, we can define vertical type I covers. (Note that, since the rectangles are compact, type I covers are not pavings.) If $A = Q_1 \cup \dots \cup Q_N$ is a vertical type I cover and each Q_i can be broken into a horizontal type I cover, $Q_i = Q_{i1} \cup \dots \cup Q_{iN_i}$, $i = 1, \dots, N$, we call $(Q_{ij} : 1 \leq i \leq N, 1 \leq j \leq N_i)$ a type II cover of A . Clearly, for a type II cover, $\|A\|$ equals the sum of $\|Q_{ij}\|$ over all $1 \leq j \leq N_i$, $1 \leq i \leq N$. If we rewrite the rectangles in a type II cover as (Q_1, \dots, Q_N) , we see that (2.3) holds also for any type II cover of A . If $A = Q_1 \cup \dots \cup Q_N$, where Q_n , $n = 1, \dots, N$, are possibly overlapping compact rectangles, then, we say that (Q_1, \dots, Q_N) is a type III cover. Given a type III cover of A , by extending all the horizontal edges of all the rectangles in the cover till they intersect the edges of A , we obtain a type II cover. Moreover, because the overlaps are counted more than once and for type II covers we have (2.3), we conclude that

$$(2.4) \quad \|A\| \leq \|Q_1\| + \dots + \|Q_N\|$$

for type III covers, where $A = Q_1 \cup \dots \cup Q_N$.

fig4.10.eps

FIGURE 4.10. Type I, type II, type III covers.

Now, we can derive the theorem. Let A be a compact rectangle. We already know (Exercise 2) that $\text{area}(A) \leq \|A\|$, so, we need only derive $\|A\| \leq \text{area}(A)$. To establish this we need to show that

$$(2.5) \quad \|A\| \leq \sum_{n=1}^{\infty} \|Q_n\|$$

for every paving (Q_n) of A . Let (Q_n) be a paving of A . We want to apply (2.4) to (Q_n) . To this end, we need to make four reductions that turn the paving (Q_n) into a type III cover.

If any one of the rectangles Q_n is unbounded, then, the right side of (2.5) is infinite, so, (2.5) is true. Therefore, we may assume that all the rectangles Q_n , $n \geq 1$, are bounded. This is the first reduction. Now, since the rectangles Q_n are open, there is (§2.1) a natural N , such that (Q_1, \dots, Q_N) is a finite paving of A . This is the second reduction. Then, $(\overline{Q}_1, \dots, \overline{Q}_N)$ is a finite cover of A by compact rectangles. This is the third reduction. Now, cutting off the excess leads us to the type III cover $(\overline{Q}_1 \cap A, \dots, \overline{Q}_N \cap A)$. This is the fourth reduction. Hence, by (2.4), we conclude that

$$\|A\| \leq \sum_{n=1}^N \|\overline{Q}_n \cap A\| \leq \sum_{n=1}^N \|\overline{Q}_n\| = \sum_{n=1}^N \|Q_n\| \leq \sum_{n=1}^{\infty} \|Q_n\|,$$

establishing (2.5). Taking the inf over all pavings (Q_n) of A in (2.5), we obtain $\|A\| \leq \text{area}(A)$, hence, $\text{area}(A) = \|A\|$ for A compact. Since $\text{area}(A) = \text{area}(\overline{A})$ and $\|A\| = \|\overline{A}\|$ for A bounded, open, we also obtain the result for bounded, open rectangles A . When A is an unbounded open rectangle, we know that $\|A\| = \infty$. On the other hand, since there are bounded subrectangles Q in A for which $\text{area}(Q) = \|Q\|$ is arbitrarily large, we obtain $\text{area}(A) = \infty$ by monotonicity. Hence, $\text{area}(A) = \|A\|$ for this case also. \square

Now, we can compute the area of a triangle A with a horizontal base of length b and height h by constructing a cover of A consisting of thin horizontal strips (Figure 4.11). Let S^- and S^+ denote the nonhorizontal sides, where we take S^- to the left of S^+ and let $\|A\|$ denote the naive area of A , i.e., $\|A\| = hb/2$. By reflection invariance, we may assume A lies above its base.

For each $n \geq 1$ on S^- , mark $n + 1$ equally spaced points (x_i^-, y_i) , $i = 0, 1, \dots, n$. Similarly, on S^+ , mark (x_i^+, y_i) , $i = 0, 1, \dots, n$. Here, we take $y_0 = 0$, $y_n = h$, $y_0 < y_1 < \dots < y_n$ and $y_i - y_{i-1} = h/n$, $i = 1, \dots, n$. Then, (x_0^-, y_0) and (x_0^+, y_0) are the endpoints of the base, and $x_0^+ - x_0^-$ is the length of the base. Two cases can occur. Either S^- has positive slope, in which case $x_0^- < x_1^- < \dots < x_n^-$, or S^- has negative slope, in which case $x_0^- > x_1^- > \dots > x_n^-$. To take care of both cases simultaneously, let $x_i^{-*} = \min(x_{i-1}^-, x_i^-)$ for each $i = 1, \dots, n$. Similarly, let $x_i^{+*} = \max(x_{i-1}^+, x_i^+)$, $i = 1, \dots, n$.

fig4.11.eps

FIGURE 4.11. Cover of a triangle.

Now, let Q_i denote the compact rectangle $[x_i^{-*}, x_i^{+*}] \times [y_{i-1}, y_i]$, $i = 1, \dots, n$. Then, we obtain a cover (Q_1, \dots, Q_n) of A (the Q_i 's cover by definition of $x_i^{\pm*}$).

Since the height of each Q_i is h/n and the width is $x_i^{+*} - x_i^{-*}$, by subadditivity, we obtain

$$\text{area}(A) \leq h \cdot \frac{1}{n} \sum_{i=1}^n (x_i^{+*} - x_i^{-*})$$

which says $\text{area}(A)$ is less or equal to h times the average of the widths of the rectangles Q_i , $i = 1, \dots, n$. But the average of the widths of the rectangles Q_i , $i = 1, \dots, n$, differs from $b/2$, at most, by b/n . Now, let $n \nearrow \infty$. Then, the sequence of averages converges to $b/2$, and we obtain $\text{area}(A) \leq \|A\|$.

To obtain the reverse inequality, draw two other triangles B, C with horizontal bases, such that $A \cup B \cup C$ is a rectangle and A, B , and C intersect only along their edges. Then, by simple arithmetic, the sum of the naive areas of A, B , and C equals the naive area of $A \cup B \cup C$, so, by subadditivity of area,

$$\begin{aligned} \|A\| + \|B\| + \|C\| &= \|A \cup B \cup C\| \\ &= \text{area}(A \cup B \cup C) \\ &\leq \text{area}(A) + \text{area}(B) + \text{area}(C) \\ &\leq \text{area}(A) + \|B\| + \|C\|. \end{aligned}$$

Cancelling $\|B\|, \|C\|$, we obtain the reverse inequality $\|A\| \leq \text{area}(A)$.

THEOREM. *The area of a triangle equals half the product of the length of its base and its height.* \square

We have derived this theorem assuming that the base of the triangle is horizontal. The general case follows from rotation invariance, which we do below. Now, let P be a parallelogram with horizontal base and let $\|P\|$ denote its naive area, i.e., the product of the length of its base and its height. Then, we leave it as an exercise to show that $\text{area}(P) = \|P\|$.

THEOREM. *The area of a parallelogram equals the product of the length of its base and height.* \square

By the next theorem, the areas of rectangles, triangles and parallelograms are given by the last three theorems, even when rotated. Recall that rotations were defined in §3.5.

THEOREM (ROTATION INVARIANCE). *Let $A \subset \mathbf{R}^2$. If A is rotated into A' , then, $\text{area}(A) = \text{area}(A')$.*

To see this, first, assume that Q is a bounded rectangle. Hence, Q' is a rotated rectangle. Now, decompose Q' into the union of a parallelogram P and two triangles S, T , all with horizontal bases and intersecting along their edges (Figure 4.12). Then, by simple arithmetic, the sum of the naive areas of S, T , and P equals the naive area of Q , so by subadditivity

$$\begin{aligned} \text{area}(Q') &\leq \text{area}(S) + \text{area}(P) + \text{area}(T) \\ &= \|S\| + \|P\| + \|T\| \\ &= \|Q\| = \text{area}(Q). \end{aligned}$$

We have just shown that $\text{area}(Q') \leq \text{area}(Q)$ for any bounded rectangle Q . If Q is an unbounded rectangle, this last inequality is clearly true. Hence, $\text{area}(Q') \leq \text{area}(Q)$ for every rectangle Q .

Now, let A be any subset, and let (Q_n) be a paving of A . Suppose that the rotation sending A to A' sends Q_n to Q'_n , $n \geq 1$. Then, $A' \subset \bigcup_{n=1}^{\infty} Q'_n$. So, by subadditivity

$$\text{area}(A') \leq \sum_{n=1}^{\infty} \text{area}(Q'_n) \leq \sum_{n=1}^{\infty} \text{area}(Q_n).$$

Here, we used the inequality derived in the previous paragraph. Taking the inf over all pavings of A , we obtain $\text{area}(A') \leq \text{area}(A)$. If we apply this last inequality to A' instead of A and to the inverse rotation, we obtain the reverse inequality $\text{area}(A) \leq \text{area}(A')$. We conclude that $\text{area}(A') = \text{area}(A)$. \square

fig4.12.eps

FIGURE 4.12. Rotation invariance of area.

Area also has good invariance properties under other types of dilation. For example, let $k > 0$ and set $H(x, y) = (kx, y)$, $V(x, y) = (x, ky)$. Then, the mappings $H : \mathbf{R}^2 \rightarrow \mathbf{R}^2$, $V : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ dilate area in the sense $\text{area}[H(A)] = k \cdot \text{area}(A) = \text{area}[V(A)]$. To see this, first, check that this is so on rectangles. Then, use the definition of area to arrive at the general case.

Our last item is additivity. In general, we do not expect that $\text{area}(A \cup B) = \text{area}(A) + \text{area}(B)$ because A and B may overlap, i.e., intersect. If A and B are disjoint, one expects to have additivity. Here, we establish additivity only for the case when A and B are well separated. Exercises 12 and 13 of §4.5 discuss a broader case.

If $A \subset \mathbf{R}^2$ and $B \subset \mathbf{R}^2$, set

$$d(A, B) = \inf \sqrt{(a - c)^2 + (b - d)^2},$$

where the inf is over all points $(a, b) \in A$ and points $(c, d) \in B$. We say A and B are *well separated* if $d(A, B)$ is positive (Figure 4.13). For example, although $\{(2, 0)\}$ and the unit disk are well separated, $\mathbf{Q} \times \mathbf{Q}$ and $\{(\sqrt{2}, 0)\}$ are disjoint but not well separated. Note that, since $\inf \emptyset = \infty$, A empty implies $d(A, B) = \infty$. Hence, the empty set is well separated from any subset of \mathbf{R}^2 .

fig4.13.eps

FIGURE 4.13. Well Separated sets.

If the lengths of the sides of a rectangle Q are a and b , by the *diameter* of Q , we mean the length of the diagonal $\sqrt{a^2 + b^2}$.

THEOREM (ADDITIVITY). *If A and B are well separated, then,*

$$\text{area}(A \cup B) = \text{area}(A) + \text{area}(B).$$

By subadditivity, $\text{area}(A \cup B) \leq \text{area}(A) + \text{area}(B)$, so, we need show only that

$$(2.6) \quad \text{area}(A \cup B) \geq \text{area}(A) + \text{area}(B).$$

If $\text{area}(A \cup B) = \infty$ (2.6) is true, so, assume $\text{area}(A \cup B) < \infty$. In this case, to compute the area of $A \cup B$, we need consider only pavings involving bounded open rectangles, since the sum of the areas of open rectangles with at least one unbounded, open rectangle, is ∞ . Let $\epsilon = d(A, B) > 0$. If (Q_n) is a paving of $A \cup B$ with bounded, open rectangles Q_n , $n \geq 1$, decompose the compactification \overline{Q}_n of each bounded open rectangle Q_n into a type II cover (see above) where the compact subrectangles are all of diameter less than ϵ . Since $\|Q_n\|$, then, equals the sum of the areas of its subrectangles, by replacing each Q_n by its subrectangles, we obtain a cover (\overline{Q}'_n) of $A \cup B$ by compact rectangles, where each \overline{Q}'_n has diameter less than ϵ and

$$\sum_{n=1}^{\infty} \|\overline{Q}'_n\| = \sum_{n=1}^{\infty} \|Q_n\|.$$

Thus, for each n , \overline{Q}'_n intersects A or B or neither but not both. For each n , if \overline{Q}'_n intersects A , denote it \overline{Q}_n^A . If it intersects B , denote it \overline{Q}_n^B . Because no \overline{Q}'_n intersects both A and B , (\overline{Q}_n^A) must be a cover of A and (\overline{Q}_n^B) must be a cover of B . Hence, by subadditivity,

$$\begin{aligned} \text{area}(A) + \text{area}(B) &\leq \sum_{n=1}^{\infty} \text{area}(\overline{Q}_n^A) + \sum_{n=1}^{\infty} \text{area}(\overline{Q}_n^B) \\ &\leq \sum_{n=1}^{\infty} \text{area}(\overline{Q}'_n) \\ &= \sum_{n=1}^{\infty} \|Q_n\|. \end{aligned}$$

Taking the inf over all pavings (Q_n) of $A \cup B$, we obtain the result. \square

As an application, let A denote the unit square $[0, 1] \times [0, 1]$ and B the triangle obtained by joining the three points $(1, 0)$, $(1, 1)$ and $(2, 1)$. We already know that $\text{area}(A) = 1$ and $\text{area}(B) = 1/2$, and we want to conclude that $\text{area}(A \cup B) = 1 + 1/2 = 3/2$ (Figure 4.14). But A and B are not well separated, so, we do not have additivity directly. Instead, we dilate A by a factor $0 < \alpha < 1$ towards its center. Then, the shrunken set αA and B are well separated. Moreover, $\alpha A \subset A$, so,

$$\begin{aligned} \text{area}(A \cup B) &\geq \text{area}((\alpha A) \cup B) \\ &= \text{area}(\alpha A) + \text{area}(B) \\ &= \alpha^2 \cdot \text{area}(A) + \text{area}(B) \\ &= \alpha^2 + 1/2. \end{aligned}$$

Since α can be arbitrarily close to 1, we conclude that $\text{area}(A \cup B) \geq 3/2$. Since subadditivity yields $\text{area}(A \cup B) \leq \text{area}(A) + \text{area}(B) = 1 + 1/2 = 3/2$, we obtain the result we seek, $\text{area}(A \cup B) = 3/2$.

fig4.14.eps

FIGURE 4.14. Area of $A \cup B$.

Additivity holds by induction for several sets. If A_1, A_2, \dots, A_n are pairwise well separated subsets of \mathbf{R}^2 , then,

$$(2.7) \quad \text{area}(A_1 \cup \dots \cup A_n) = \text{area}(A_1) + \dots + \text{area}(A_n).$$

To see this, (2.7) is trivially true for $n = 1$, so, assume (2.7) is true for a particular $n \geq 1$, and let A_1, \dots, A_{n+1} be pairwise well separated. Let $\epsilon_j = d(A_j, A_{n+1}) > 0$, $j = 1, \dots, n$. Since

$$d(A_1 \cup \dots \cup A_n, A_{n+1}) = \min(\epsilon_1, \dots, \epsilon_n) > 0,$$

A_{n+1} and $A_1 \cup \dots \cup A_n$ are well separated. Hence, by the inductive hypothesis,

$$\begin{aligned} \text{area}(A_1 \cup \dots \cup A_{n+1}) &= \text{area}(A_1 \cup \dots \cup A_n) + \text{area}(A_{n+1}) \\ &= \text{area}(A_1) + \dots + \text{area}(A_n) + \text{area}(A_{n+1}). \end{aligned}$$

By induction, this establishes (2.7) for all $n \geq 1$.

More generally, if (A_n) is a sequence of pairwise well separated sets, then,

$$(2.8) \quad \text{area}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \text{area}(A_n).$$

To see this, subadditivity yields

$$\text{area}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \text{area}(A_n).$$

For the reverse inequality, apply (2.7) and monotonicity to the first N sets, yielding

$$\text{area}\left(\bigcup_{n=1}^{\infty} A_n\right) \geq \text{area}\left(\bigcup_{n=1}^N A_n\right) = \sum_{n=1}^N \text{area}(A_n).$$

Now let $N \nearrow \infty$, obtaining

$$\text{area}\left(\bigcup_{n=1}^{\infty} A_n\right) \geq \sum_{n=1}^{\infty} \text{area}(A_n).$$

This establishes (2.8).

As an application of (2.8), we can now compute the area of the Cantor set C . The Cantor set is constructed by removing, at successive stages, smaller and smaller open subsquares of $C_0 = [0, 1] \times [0, 1]$. Denote these subsquares Q_1, Q_2, \dots (at what stage they are removed is not important). Then, for each n , C and Q_n are disjoint, so, for $0 < \alpha < 1$, C and the centered dilations αQ_n are well separated. Moreover for each m, n the centered dilations αQ_n and αQ_m are well separated. But the union of C with all the squares αQ_n , $n \geq 1$, lies in the unit square C_0 . Hence, by (2.8),

$$\text{area}(C) + \sum_{n=1}^{\infty} \text{area}(\alpha Q_n) = \text{area}\left(C \cup \left(\bigcup_{n=1}^{\infty} \alpha Q_n\right)\right) \leq \text{area}(C_0) = 1.$$

In the previous section we obtained, $\sum_{n=1}^{\infty} \text{area}(Q_n) = 1$. By dilation invariance, this implies $\text{area}(C) + \alpha^2 \leq 1$. Letting $\alpha \nearrow 1$, we obtain $\text{area}(C) + 1 \leq 1$ or $\text{area}(C) = 0$.

THEOREM. *The area of the Cantor set is zero.* \square

Exercises 4.2.

1. Establish reflection invariance and monotonicity of area.
2. Show that the area of a bounded line segment is zero, the area of any line is zero, and $\text{area}(Q) = \text{area}(\overline{Q})$ for any bounded open rectangle Q . Show also that $\text{area}(Q) \leq \|Q\|$ for any compact rectangle Q .
3. Let P be a parallelogram with a horizontal base, and let $\|P\|$ denote the product of the length of its base and its height. Then, $\text{area}(P) = \|P\|$.
4. Compute the area of a trapezoid.
5. If A and B are rectangles, then, $\text{area}(A \cup B) = \text{area}(A) + \text{area}(B) - \text{area}(A \cap B)$.
6. For $k \in \mathbf{R}$, define $H : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ and $V : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ by $H(x, y) = (kx, y)$ and $V(x, y) = (x, ky)$. Then, $\text{area}[V(A)] = |k| \cdot \text{area}(A) = \text{area}[H(A)]$ for every $A \subset \mathbf{R}^2$.
7. \star A mapping $L : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ is *linear* if it is of the form $L(x, y) = (ax+by, cx+dy)$ with $a, b, c, d \in \mathbf{R}$. Show that a linear mapping sends lines to (possibly collapsed) lines and parallelograms to (possibly collapsed) parallelograms. Show that a linear mapping L is invertible (i.e., a bijection §1.1) iff the real $\det(L) = ad - bc$ is not zero. In this case, show that the inverse K of L is linear and compute $\det(K)$.
8. \star Let $L : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ be an invertible linear mapping. Show that

$$(2.9) \quad \text{area}[L(A)] = |\det(L)| \cdot \text{area}(A), \quad A \subset \mathbf{R}^2.$$

Thus, L is area-preserving iff $\det(L) = \pm 1$. Such an L is called *affine*, and this result is *affine-invariance* of area. (Do this for rectangles first.)

fig4.15.eps

FIGURE 4.15. Affine invariance of area.

9. \star Let $L : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ be a noninvertible linear mapping. Show that $L(\mathbf{R}^2)$ is contained in a line and (2.9) holds.
10. Show that $\{(\sqrt{2}, 0)\}$ and the unit disk are well separated, but $\{(\sqrt{2}, 0)\}$ and $\mathbf{Q} \times \mathbf{Q}$ are not.
11. Let D be the unit disk, and let $D^+ = \{(x, y) : x^2 + y^2 < 1 \text{ and } y > 0\}$. Show that $\text{area}(D) = 2 \cdot \text{area}(D^+)$.
12. Compute the area of the sets C' and C^α described in Exercises 1 and 2 of §4.1, using the properties of area.
13. Let $P_k = (\cos(2\pi k/n), \sin(2\pi k/n))$, $k = 0, 1, \dots, n$. Then, P_0, P_1, \dots, P_n are evenly spaced points on the unit circle with $P_n = P_0$. Let D_n denote the n -sided polygon obtained by joining the points P_k . Compute $\text{area}(D_n)$.
14. Let $A \subset \mathbf{R}^2$. A *triangular paving* of A is a cover (T_n) of A where each T_n , $n \geq 1$, is a triangle (oriented arbitrarily). With $\text{area}(A)$ as defined previously, show that

$$\text{area}(A) = \inf \left\{ \sum_{n=1}^{\infty} \|T_n\| : \text{all triangular pavings } (T_n) \text{ of } A \right\}.$$

Here, $\|T\|$ denotes the naive area of the triangle T , i.e., half the product of the length of the base times the height.

15. ★ Let $A \subset \mathbf{R}^2$. If $\text{area}(A) > 0$ and $0 < \alpha < 1$, there is *some* open rectangle Q , such that $\text{area}(Q \cap A) > \alpha \cdot \text{area}(Q)$. (Argue by contradiction, and use the definition of area.)

§4.3 The Integral

Let $f : (a, b) \rightarrow \mathbf{R}$ be a function defined on an open interval (a, b) , where, as usual, a may equal $-\infty$ or b may equal ∞ . We say f is *bounded* if $|f(x)| \leq M$, $a < x < b$, for some real M . If f is *nonnegative*, i.e., if $f(x) \geq 0$, $a < x < b$, the *subgraph of f over (a, b)* is the set (Figure 4.16)

$$G = \{(x, y) : a < x < b, 0 < y < f(x)\} \subset \mathbf{R}^2.$$

Note that the inequalities in this definition are strict.

fig4.16.eps

FIGURE 4.16. Subgraphs of nonnegative functions.

For nonnegative f , we define the *integral* $\int_a^b f(x) dx$ of f from a to b to be the area of its subgraph G ,

$$\int_a^b f(x) dx = \text{area}(G).$$

Then, the integral is either 0, a positive real, or ∞ . The reason for the unusual notation is explained below.

Thus, according to our definition, *every nonnegative function has an integral* and integrals of nonnegative functions are areas — nothing more, nothing less — of certain subsets of \mathbf{R}^2 .

Since the empty set has zero area, we always have $\int_a^a f(x) dx = 0$. For each $k \geq 0$, the subgraph of $f(x) = k$, $a < x < b$, over (a, b) is an open rectangle, so,

$$\int_a^b k dx = k(b - a).$$

Since the area is monotone, so is the integral: If $0 \leq f \leq g$ on (a, b) ,

$$\int_a^b f(x) dx \leq \int_a^b g(x) dx.$$

In particular, $0 \leq f \leq M$ on (a, b) implies $0 \leq \int_a^b f(x) dx \leq M(b - a)$.

A nonnegative function f is *integrable* over (a, b) if $\int_a^b f(x) dx < \infty$. For example, we have just seen that every bounded nonnegative f is integrable over a bounded interval (a, b) . Now, we discuss the integral of a *signed* function, i.e., a function that takes on positive and negative values.

Given a function $f : (a, b) \rightarrow \mathbf{R}$, we set

$$f^+(x) = \max[f(x), 0],$$

fig4.17.eps

FIGURE 4.17. Positive and negative parts of $\sin x$.

and

$$f^-(x) = \max[-f(x), 0].$$

These are (Figure 4.17) the *positive part* and the *negative part* of f , respectively. Note that $f^+ - f^- = f$ and $f^+ + f^- = |f|$.

We say a signed function f is *integrable over* (a, b) if

$$\int_a^b |f(x)| dx < \infty.$$

In this case, $\int_a^b f^\pm(x) dx \leq \int_a^b |f(x)| dx$ are both finite. For integrable f , we define the *integral* $\int_a^b f(x) dx$ of f from a to b by

$$\int_a^b f(x) dx = \int_a^b f^+(x) dx - \int_a^b f^-(x) dx.$$

From this follows

$$\int_a^b [-f(x)] dx = - \int_a^b f(x) dx$$

for every integrable function f , since $g = -f$ implies $g^+ = f^-$ and $g^- = f^+$.

We warn the reader that, although $\int_a^b f(x) dx = \int_a^b f^+(x) dx - \int_a^b f^-(x) dx$ is a definition, the identity $\int_a^b |f(x)| dx = \int_a^b f^+(x) dx + \int_a^b f^-(x) dx$ is not true for general integrable f . However, it is true when f is continuous. This property (linearity) is discussed in the next section.

Also, from the above discussion, we see that *every bounded (signed) function is integrable over a bounded interval*. For example, $\sin x$ and $\sin x/x$ are integrable over $(0, \pi)$. In fact, both functions are integrable over $(0, b)$ for any finite b and, hence, $\int_0^b \sin x dx$ and $\int_0^b (\sin x/x) dx$ are defined.

It is reasonable to expect that $\sin x$ is not integrable over $(0, \infty)$. Indeed the subgraph of $|\sin x|$ consists of a union of sets G_n , $n \geq 1$, where each G_n denotes the subgraph over $((n-1)\pi, n\pi)$. By translation invariance, the sets G_n , $n \geq 1$, have the same positive area and the sets G_1, G_3, G_5, \dots , are well separated. Hence, we obtain

$$\int_0^\infty (\sin x)^+ dx = \text{area} \left(\bigcup_{n=1}^\infty G_{2n-1} \right) = \sum_{n=1}^\infty \text{area}(G_{2n-1}) = \infty.$$

By considering, instead, G_2, G_4, G_6, \dots , we obtain $\int_0^\infty (\sin x)^- dx = \infty$. Thus,

$$\int_0^\infty (\sin x)^+ dx - \int_0^\infty (\sin x)^- dx = \infty - \infty.$$

Hence, $\int_0^\infty \sin x dx$ cannot be defined as a difference of two areas.

It turns out that $\sin x/x$ is also not integrable over $(0, \infty)$. To see this, let G_n denote the subgraph of $|\sin x/x|$ over $((n-1)\pi, n\pi)$, $n \geq 1$ (Figure 4.18). In each G_n , we can insert a rectangle Q_n of area $\sqrt{2}/(4n-1)$ so that the rectangles are well separated. By additivity, then, we obtain

$$\int_0^\infty \frac{|\sin x|}{x} dx \geq \text{area} \left(\bigcup_{n=1}^\infty Q_n \right) = \sum_{n=1}^\infty \text{area}(Q_n) = \sum_{n=1}^\infty \frac{\sqrt{2}}{4n-1} = \infty,$$

by comparison with the harmonic series. Thus, $\sin x/x$ is not integrable over $(0, \infty)$. More explicitly, this reasoning also shows that

$$\int_0^\infty \left(\frac{\sin x}{x} \right)^+ dx \geq \text{area}(Q_1) + \text{area}(Q_3) + \text{area}(Q_5) + \dots = \infty$$

and

$$\int_0^\infty \left(\frac{\sin x}{x} \right)^- dx \geq \text{area}(Q_2) + \text{area}(Q_4) + \text{area}(Q_6) + \dots = \infty.$$

Thus,

$$\int_0^\infty \left(\frac{\sin x}{x} \right)^+ dx - \int_0^\infty \left(\frac{\sin x}{x} \right)^- dx = \infty - \infty,$$

hence, $\int_0^\infty \sin x/x dx$ also cannot be defined as a difference of two areas.

fig4.18.eps

FIGURE 4.18. The graphs of $\sin x/x$ and $|\sin x|/x$.

To summarize, *the integral of an integrable function is the area of the subgraph of its positive part minus the area of the subgraph of its negative part*. Every property of $\int_a^b f(x) dx$ ultimately depends on a corresponding property of area.

Frequently, one checks integrability of a given f by first applying one or more of the properties below to the nonnegative function $|f|$. For example, consider the function $g(x) = 1/x^2$ for $x > 1$, and, for each $n \geq 1$, let G_n denote the compact rectangle $[n, n+1] \times [0, 1/n^2]$. Then, (G_n) is a cover of the subgraph of g over $(1, \infty)$ (Figure 4.19). Hence,

$$\int_1^\infty \frac{1}{x^2} dx \leq \sum_{n=1}^\infty \frac{1}{n^2},$$

which is finite (§1.6). Thus, g is integrable over $(1, \infty)$. Since the signed function $f(x) = \cos x/x^2$ satisfies $|f(x)| \leq g(x)$ for $x > 1$, by monotonicity, we conclude that

$$(3.1) \quad \int_1^\infty \left| \frac{\cos x}{x^2} \right| dx < \infty.$$

Hence, $\cos x/x^2$ is integrable over $(1, \infty)$.

fig4.19.eps

FIGURE 4.19. A cover of the subgraph of $1/x^2$ over $(1, \infty)$.

Of course, functions may be unbounded and integrable. For example, the function $f(x) = 1/\sqrt{x}$ is integrable over $(0, 1)$. To see this, let G_n denote the compact rectangle $[1/(n+1)^2, 1/n^2] \times [0, n+1]$. Then, (G_n) is a cover of the subgraph of f over $(0, 1)$ (Figure 4.20). Hence,

$$\int_0^1 \frac{1}{\sqrt{x}} dx \leq \sum_{n=1}^{\infty} (n+1) \left(\frac{1}{n^2} - \frac{1}{(n+1)^2} \right) = \sum_{n=1}^{\infty} \frac{2n+1}{n^2(n+1)} \leq 2 \sum_{n=1}^{\infty} \frac{1}{n^2},$$

which is finite. Thus, f is integrable over $(0, 1)$.

fig4.20.eps

FIGURE 4.20. A cover of the subgraph of $1/\sqrt{x}$ over $(0, 1)$.

THEOREM (MONOTONICITY). *Suppose that f and g are both nonnegative or both integrable on (a, b) . If $f \leq g$ on (a, b) , then,*

$$\int_a^b f(x) dx \leq \int_a^b g(x) dx.$$

If $0 \leq f \leq g$, we already know this. For the integrable case, note that $f \leq g$ implies $f^+ = \max(f, 0) \leq \max(g, 0) = g^+$ and $g^- = \max(-g, 0) \leq \max(-f, 0) = f^-$ on (a, b) . Hence,

$$\int_a^b f^+(x) dx \leq \int_a^b g^+(x) dx,$$

and

$$\int_a^b f^-(x) dx \geq \int_a^b g^-(x) dx.$$

Subtracting the second inequality from the first, the result follows. \square

Since $\pm f \leq |f|$, the theorem implies

$$\pm \int_a^b f(x) dx = \int_a^b \pm f(x) dx \leq \int_a^b |f(x)| dx$$

which yields

$$\left| \int_a^b f(x) dx \right| \leq \int_a^b |f(x)| dx$$

for every integrable f .

THEOREM (TRANSLATION AND DILATION INVARIANCE). *Let f be nonnegative or integrable on (a, b) . Choose $c \in \mathbf{R}$ and $k > 0$. Then,*

$$\begin{aligned} \int_a^b f(x+c) dx &= \int_{a+c}^{b+c} f(x) dx, \\ \int_a^b kf(x) dx &= k \int_a^b f(x) dx, \end{aligned}$$

and

$$\int_a^b f(kx) dx = \frac{1}{k} \int_{ka}^{kb} f(x) dx.$$

If f is nonnegative, let G denote the subgraph of $f(x + c)$ over (a, b) (Figure 4.21). Then, the translate $G + (c, 0)$ equals

$$\{(x, y) : a + c < x < b + c, 0 < y < f(x)\},$$

which is the subgraph of $f(x)$ over the interval $(a+c, b+c)$. By translation invariance of area, we obtain translation invariance of the integral in the nonnegative case. If f is integrable, by the nonnegative case,

$$\int_a^b f^+(x + c) dx = \int_{a+c}^{b+c} f^+(x) dx,$$

and

$$\int_a^b f^-(x + c) dx = \int_{a+c}^{b+c} f^-(x) dx.$$

Now, if $g(x) = f(x + c)$, then, $g^+(x) = f^+(x + c)$ and $g^-(x) = f^-(x + c)$. So, subtracting the last equation from the previous one, we obtain translation invariance in the integrable case.

For the second equation and f nonnegative, recall that from the previous section the dilation mapping $V(x, y) = (x, ky)$, and let G denote the subgraph of f over (a, b) . Then, $V(G) = \{(x, y) : a < x < b, 0 < y < kf(x)\}$. Hence, $\text{area}(V(G)) = \int_a^b kf(x) dx$. Now, dilation invariance of the area (Exercise 6 of §4.2) yields $\int_a^b kf(x) dx = k \int_a^b f(x) dx$ for f nonnegative. For integrable f , the result follows by applying, as above, the nonnegative case to f^+ and f^- .

For the third equation, let $H(x, y) = (kx, y)$, and let G denote the subgraph of $f(kx)$ over (a, b) . Then, $H(G) = \{(x, y) : ka < x < kb, 0 < y < f(x)\}$. The third equation now follows, as before, by dilation invariance. For integrable f , the result follows by applying the nonnegative case to f^\pm . \square

fig4.21.eps

FIGURE 4.21. Translation and dilation invariance of integrals.

By similar reasoning, one can also derive (Figure 4.22)

$$\int_{-b}^{-a} f(-x) dx = \int_a^b f(x) dx,$$

valid for f nonnegative or integrable over (a, b) .

fig4.22.eps

FIGURE 4.22. Reflection invariance of integrals.

The next property is additivity.

THEOREM (ADDITIVITY). *Suppose that f is nonnegative or integrable over (a, b) , and choose $a < c < b$. Then,*

$$\int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx.$$

To see this, first, assume that f is nonnegative. Since the vertical line $x = c$ has zero area, subadditivity yields

$$\int_a^b f(x) dx \leq \int_a^c f(x) dx + \int_c^b f(x) dx.$$

So, we need only show that

$$(3.2) \quad \int_a^b f(x) dx \geq \int_a^c f(x) dx + \int_c^b f(x) dx.$$

If f is not integrable, (3.2) is immediate since the left side is infinite, so, assume f is nonnegative and integrable. Now, choose any strictly increasing sequence $a < c_1 < c_2 < \dots$ converging to c . Then, for $n \geq 1$, the subgraph of f over (a, c_n) and the subgraph of f over (c, b) are well separated (Figure 4.23). So, by monotonicity and well separated additivity,

$$(3.3) \quad \int_a^b f(x) dx \geq \int_a^{c_n} f(x) dx + \int_c^b f(x) dx.$$

We wish to send $n \nearrow \infty$ in (3.3). To this end, for each $n \geq 1$, let G_n denote the subgraph of f over (c_n, c_{n+1}) , and let L_n denote the vertical line segment $\{(x, y) : x = c_n, 0 < y < f(x)\}$. Since G_2, G_4, G_6, \dots , are pairwise well separated,

$$\text{area}(G_2) + \text{area}(G_4) + \text{area}(G_6) + \dots \leq \int_a^c f(x) dx < \infty.$$

Since G_1, G_3, G_5, \dots , are pairwise well separated,

$$\text{area}(G_1) + \text{area}(G_3) + \text{area}(G_5) + \dots \leq \int_a^c f(x) dx < \infty.$$

Adding the last two inequalities yields the convergence of $\sum_{n=1}^{\infty} \text{area}(G_n)$. Hence, the tail (§1.6) goes to zero:

$$\lim_{n \nearrow \infty} \sum_{k=n}^{\infty} \text{area}(G_k) = 0.$$

Since the subgraph of f over (c_n, c) equals $G_n \cup L_{n+1} \cup G_{n+1} \cup L_{n+2} \cup G_{n+2} \cup L_{n+3} \cup \dots$, subadditivity implies

$$0 \leq \int_{c_n}^c f(x) dx \leq \sum_{k=n}^{\infty} \text{area}(G_k), \quad n \geq 1.$$

Hence, we obtain

$$(3.4) \quad \lim_{n \nearrow \infty} \int_{c_n}^c f(x) dx = 0.$$

Since by monotonicity and subadditivity, again,

$$\int_a^{c_n} f(x) dx \leq \int_a^c f(x) dx \leq \int_a^{c_n} f(x) dx + \int_{c_n}^c f(x) dx, \quad n \geq 1,$$

we conclude that

$$\lim_{n \nearrow \infty} \int_a^{c_n} f(x) dx = \int_a^c f(x) dx.$$

Now, sending $n \nearrow \infty$ in (3.3) yields (3.2). Hence, the result for f nonnegative. If f is integrable, apply the nonnegative case to f^+ and f^- . Then,

$$\int_a^b f^+(x) dx = \int_a^c f^+(x) dx + \int_c^b f^+(x) dx,$$

and

$$\int_a^b f^-(x) dx = \int_a^c f^-(x) dx + \int_c^b f^-(x) dx.$$

Subtracting the second equation from the first, we obtain the result in the integrable case. \square

fig4.23.eps

FIGURE 4.23. Additivity of integrals.

The bulk of the derivation above involves establishing (3.4). If f is bounded, say by M , then, the integral in (3.4) is no more than $M(c - c_n)$, hence, trivially, goes to zero. The delicacy is necessary to handle unbounded situations.

The main point in the derivation is that, although the subgraph G of f over (a, c) and the subgraph G' of f over (c, b) are not well separated, we still have additivity, because we know something — the existence of the vertical edges — about the geometry of G and G' .

In the previous section, when we wanted to apply additivity to several sets (for example, when we computed the area of the Cantor set) that were not well separated, we dilated them by a factor $0 < \alpha < 1$ and, then, applied additivity to the shrunken sets.

Why don't we use the same trick here for G or G' ? The reason is that if the graph of f is sufficiently “jagged” (Figure 4.24), we do not have $\alpha G \subset G$, a necessary step in applying the shrinking trick of the previous section.

fig4.24.eps

FIGURE 4.24. A “jagged” function.

By induction, additivity holds for a partition (§2.2) of (a, b) : If $a = x_0 < x_1 < \dots < x_{n+1} = b$ and f is nonnegative or integrable over (a, b) , then,

$$\int_a^b f(x) dx = \sum_{k=1}^{n+1} \int_{x_{k-1}}^{x_k} f(x) dx.$$

Since the right side does not involve the values of f at the points defining the partition, we conclude that *the integrals of two functions $f : (a, b) \rightarrow \mathbf{R}$ and $g : (a, b) \rightarrow \mathbf{R}$ are equal, whenever they differ only on finitely many points $a < x_1 < \dots < x_n < b$.*

Another application of additivity is to piecewise constant functions. A function $f : (a, b) \rightarrow \mathbf{R}$ is *piecewise constant* if there is a partition $a = x_0 < x_1 < \dots < x_{n+1} = b$, such that f , restricted to each open subinterval (x_{i-1}, x_i) , $i = 1, \dots, n+1$, is constant. (Note that the values of a piecewise constant function at the partition points x_i , $1 \leq i \leq n$, are not restricted in any way.) In this case, additivity implies

$$\int_a^b f(x) dx = \sum_{i=1}^{n+1} c_i \Delta x_i,$$

where $\Delta x_i = x_i - x_{i-1}$, $i = 1, \dots, n+1$. Since a continuous function can be closely approximated by a piecewise constant function (§2.3), the integral should be thought of as a sort of sum with Δx_i “infinitely small,” hence, the notation dx replacing Δx_i and \int replacing \sum .

This view is supported by Exercise 3. Indeed, by defining integrals as areas of subgraphs, we capture the intuition that integrals are approximately sums of areas of rectangles in *any* paving, and not just finite vertical pavings as given by the “Riemann sums” of Exercise 3.

Also, since the integral is, by definition, a combination of certain areas and the notation $\int_a^b f(x) dx$ is just a mnemonic device, the variable inside the integral sign is a “dummy” variable, i.e., $\int_a^b f(x) dx = \int_a^b f(t) dt$. Nevertheless, the interpretation of the integral as a “continuous sum” is basic, useful, and important.

Let us go back to the integrals of $\sin x$ and $\sin x/x$ over $(0, \infty)$. Above, we saw that these functions were not integrable over $(0, \infty)$, and, so, $\int_0^\infty \sin x dx$ and $\int_0^\infty \sin x/x dx$ could not be defined as the difference of the areas of the positive and the negative parts. An alternate approach is to consider $F(b) = \int_0^b \sin x dx$ and to take the limit $F(\infty) = \lim_{b \rightarrow \infty} F(b)$. However, since the areas of the sets G_n , $n \geq 1$, are equal, by additivity, $F(n\pi) = \text{area}(G_1) - \text{area}(G_2) + \dots \pm \text{area}(G_n)$ equals $\text{area}(G_1)$ or zero according to whether n is odd or even. Thus, the limit $F(\infty)$ does not exist and this approach fails for $\sin x$.

For $\sin x/x$, however, it is a different story. Let $F(b) = \int_0^b \sin x/x dx$, and let G_n denote the subgraph of $|\sin x|/x$ over $((n-1)\pi, n\pi)$ for each $n \geq 1$. Then, by additivity $F(n\pi) = \text{area}(G_1) - \text{area}(G_2) + \dots \pm \text{area}(G_n)$. Hence,

$$\lim_{n \nearrow \infty} \int_0^{n\pi} \frac{\sin x}{x} dx = \text{area}(G_1) - \text{area}(G_2) + \text{area}(G_3) - \dots$$

But this last series has a finite sum since it is alternating with decreasing terms! Thus,

$$(3.5) \quad \int_0^\infty \frac{\sin x}{x} dx \neq \lim_{n \nearrow \infty} \int_0^{n\pi} \frac{\sin x}{x} dx$$

since the left side is not defined and the right side is a well defined, finite real. The limit $\lim_{n \nearrow \infty} F(n\pi)$ is computed in Exercise 12 of §5.4.

On the other hand, when f is nonnegative or integrable, its integral over an interval (a, b) can be obtained as a limit of integrals over subintervals (a_n, b_n) (Figure 4.25), and the behavior (3.5) does not occur.

THEOREM (CONTINUITY AT THE ENDPOINTS). *If f is nonnegative or integrable on (a, b) and $a_n \rightarrow a+$, $b_n \rightarrow b-$, then,*

$$(3.6) \quad \int_a^b f(x) dx = \lim_{n \nearrow \infty} \int_{a_n}^{b_n} f(x) dx.$$

If f is integrable on (a, b) and $a_n \rightarrow a+$, $b_n \rightarrow b-$, then, in addition,

$$(3.7) \quad \lim_{n \nearrow \infty} \int_a^{a_n} f(x) dx = 0,$$

and

$$(3.8) \quad \lim_{n \nearrow \infty} \int_{b_n}^b f(x) dx = 0.$$

fig4.25.eps

FIGURE 4.25. Continuity at the endpoints.

To see this, first assume that f is nonnegative and $b_n \nearrow b$, and fix $a < c < b$. Since area is monotone, the sequence $\int_c^{b_n} f(x) dx$, $n \geq 1$, is increasing and

$$\lim_{n \nearrow \infty} \int_c^{b_n} f(x) dx \leq \int_c^b f(x) dx.$$

For the reverse inequality, let G_n denote the subgraph of f over (b_n, b_{n+1}) , $n \geq 1$. By additivity,

$$\int_c^{b_n} f(x) dx = \int_c^{b_1} f(x) dx + \sum_{k=1}^{n-1} \text{area}(G_k),$$

so, taking the limit and using subadditivity,

$$\begin{aligned} \lim_{n \nearrow \infty} \int_c^{b_n} f(x) dx &= \int_c^{b_1} f(x) dx + \sum_{k=1}^{\infty} \text{area}(G_k) \\ &\geq \int_c^{b_1} f(x) dx + \int_{b_1}^b f(x) dx \geq \int_c^b f(x) dx. \end{aligned}$$

Hence,

$$(3.9) \quad \lim_{n \nearrow \infty} \int_c^{b_n} f(x) dx = \int_c^b f(x) dx.$$

In general, if $b_n \rightarrow b-$, then, $b_{n^*} \nearrow b$ (§1.5), and $b_{n^*} \leq b_n < b$. Hence,

$$\int_c^{b_{n^*}} f(x) dx \leq \int_c^{b_n} f(x) dx \leq \int_c^b f(x) dx, \quad n \geq 1,$$

which implies (3.9), for general $b_n \rightarrow b-$. Since

$$\int_{a_n}^c f(x) dx = \int_{-c}^{-a_n} f(-x) dx,$$

applying what we just learned to $f(-x)$ yields

$$\lim_{n \nearrow \infty} \int_{a_n}^c f(x) dx = \int_a^c f(x) dx.$$

Hence,

$$\begin{aligned} \int_a^b f(x) dx &= \int_a^c f(x) dx + \int_c^b f(x) dx \\ &= \lim_{n \nearrow \infty} \int_{a_n}^c f(x) dx + \lim_{n \nearrow \infty} \int_c^{b_n} f(x) dx \\ &= \lim_{n \nearrow \infty} \int_{a_n}^{b_n} f(x) dx. \end{aligned}$$

For the integrable case, apply (3.6) to f^\pm to get (3.6) for f . Since

$$\int_a^{a_n} f(x) dx = \int_a^b f(x) dx - \int_{a_n}^b f(x) dx,$$

we get (3.7). Similarly, we get (3.8). \square

For example,

$$\int_0^1 x^r dx = \lim_{a \rightarrow 0^+} \int_a^1 x^r dx,$$

and

$$\int_1^\infty x^r dx = \lim_{b \rightarrow \infty} \int_1^b x^r dx,$$

both for r real.

When f is integrable, the last theorem can be improved: We have continuity of the integral at every point in (a, b) .

THEOREM (CONTINUITY). *Suppose that f is integrable over (a, b) , and set*

$$F(t) = \int_a^t f(x) dx, \quad a < t < b.$$

Then, F is continuous on (a, b) .

To see this, fix $a < c < b$, and let $c_n \rightarrow c-$. Applying the previous theorem on (a, c) we obtain $F(c_n) \rightarrow F(c)$. Hence, we obtain continuity of F from the left at every real in (a, b) .

Now, let $g(x) = f(-x)$, $-b < x < -a$, and

$$G(t) = \int_t^{-a} g(x) dx, \quad -b < t < -a.$$

Since, by additivity,

$$G(t) = \int_{-b}^{-a} g(x) dx - \int_{-b}^t g(x) dx, \quad -b < t < -a,$$

by the previous paragraph applied to g , the function G is continuous from the left at every point in $(-b, -a)$. Thus, the function

$$G(-t) = \int_{-t}^{-a} f(-x) dx = \int_a^t f(x) dx = F(t), \quad a < t < b,$$

is continuous from the right at every point in (a, b) . This establishes continuity of F on (a, b) . \square

Our last item is the integral test for positive series.

INTEGRAL TEST. *Let $f : (0, \infty) \rightarrow (0, \infty)$ be decreasing. Then,*

$$(3.10) \quad \gamma = \lim_{n \nearrow \infty} \left[\sum_{k=1}^n f(k) - \int_1^{n+1} f(x) dx \right]$$

exists and $0 \leq \gamma \leq f(1)$. In particular, the integral $\int_1^\infty f(x) dx$ is finite iff the sum $\sum_{n=1}^\infty f(n)$ converges.

fig4.26.eps

FIGURE 4.26. Integral test.

For each $n \geq 1$, let $B_n = (n, n+1) \times (0, f(n))$, $B'_n = (n, n+1) \times [f(n+1), f(n)]$, and let G_n denote the subgraph of f over $(n, n+1)$ (Figure 4.26). Since f is decreasing, $G_n \subset B_n \subset G_n \cup B'_n$, for all $n \geq 1$. Then, the quantity whose limit is the right side of (3.10), equals

$$\sum_{k=1}^n [\text{area}(B_k) - \text{area}(G_k)],$$

which is clearly increasing with n (here, we used additivity). Hence, the limit $\gamma \geq 0$ exists. On the other hand, by subadditivity, we get

$$\text{area}(B_k) - \text{area}(G_k) \leq \text{area}(B'_k) = f(k) - f(k+1).$$

So,

$$\gamma = \sum_{n=1}^\infty [\text{area}(B_k) - \text{area}(G_k)] \leq \sum_{n=1}^\infty [f(k) - f(k+1)] = f(1).$$

Thus, $\gamma \leq f(1)$. If either $\int_1^\infty f(x) dx$ or $\sum_{n=1}^\infty f(n)$ is finite, (3.10) simplifies to

$$\gamma = \sum_{n=1}^\infty f(n) - \int_1^\infty f(x) dx,$$

which shows that the sum is finite iff the integral is finite. \square

Exercises 4.3.

1. Show that $\int_0^\infty f(kx)x^{-1} dx = \int_0^\infty f(x)x^{-1} dx$ for $k > 0$ and $f(x)/x$ nonnegative or integrable over $(0, \infty)$.
2. Show that $\int_{-b}^{-a} f(-x) dx = \int_a^b f(x) dx$ for f nonnegative or integrable over (a, b) .
3. ★ Let $f : [a, b] \rightarrow \mathbf{R}$ be continuous. If $a = x_0 < x_1 < \cdots < x_{n+1} = b$ is a partition of $[a, b]$, a *Riemann sum* corresponding to this partition is the real (Figure 4.27)

$$\sum_{i=1}^{n+1} f(x_i^\#)(x_i - x_{i-1}),$$

where $x_i^\#$ is arbitrarily chosen in (x_{i-1}, x_i) , $i = 1, \dots, n+1$. Let $I = \int_a^b f(x) dx$. Show that, for every $\epsilon > 0$, there is a $\delta > 0$, such that

$$(3.11) \quad \left| I - \sum_{i=1}^{n+1} f(x_i^\#)(x_i - x_{i-1}) \right| \leq \epsilon$$

for any partition $a = x_0 < x_1 < \cdots < x_{n+1} = b$ of mesh less than δ and choice of points $x_1^\#, \dots, x_{n+1}^\#$. (Approximate f by a piecewise constant f_ϵ as in §2.3.)

fig4.27.eps

FIGURE 4.27. Riemann sums.

4. Let $f : (0, 1) \rightarrow \mathbf{R}$ be given by

$$f(x) = \begin{cases} x & \text{if } x \text{ irrational,} \\ 0 & \text{if } x \text{ rational.} \end{cases}$$

Compute $\int_0^1 f(x) dx$.

5. ★ Let $f : (a, b) \rightarrow \mathbf{R}$ be nonnegative, and suppose that $g : (a, b) \rightarrow \mathbf{R}$ is nonnegative and piecewise constant. Use additivity to show that

$$\int_a^b [f(x) + g(x)] dx = \int_a^b f(x) dx + \int_a^b g(x) dx.$$

(First, do this for g constant.)

6. Let $f : (0, \infty) \rightarrow \mathbf{R}$ be nonnegative and equal to a constant c_n on each subinterval $(n-1, n)$ for $n = 1, 2, \dots$. Then,

$$\int_0^\infty f(x) dx = \sum_{n=1}^\infty c_n.$$

Instead, if f is integrable, then, $\sum_{n=1}^\infty c_n$ is absolutely convergent and the equality holds.

7. ★ A function $f : (a, b) \rightarrow \mathbf{R}$ is *Riemann integrable over (a, b)* if there is a real I satisfying the following property: For all $\epsilon > 0$, there is a $\delta > 0$, such that

(3.11) holds for any partition $a = x_0 < x_1 < \cdots < x_{n+1} = b$ of mesh less than δ and choice of intermediate points $x_1^\#, \dots, x_{n+1}^\#$. Thus, Exercise 3 says every function continuous on a compact interval $[a, b]$ is Riemann integrable over (a, b) . Let $f(x) = 0$ for $x \in \mathbf{Q}$ and $f(x) = 1$ for $x \notin \mathbf{Q}$. Show that this f is *not* Riemann integrable over $(0, 1)$.

8. Let $g : (0, \infty) \rightarrow (0, \infty)$ be decreasing and bounded. Show that

$$\lim_{\delta \rightarrow 0^+} \delta \sum_{n=1}^{\infty} g(n\delta) = \int_0^{\infty} g(x) dx.$$

(Apply the integral test to $f(x) = g(x\delta)$.)

9. Let $f : (-b, b) \rightarrow \mathbf{R}$ be nonnegative or integrable. If f is even, then, $\int_{-b}^b f(x) dx = 2 \int_0^b f(x) dx$. Now, let f be integrable. If f is odd, then, $\int_{-b}^b f(x) dx = 0$.

10. Show that $\int_{-\infty}^{\infty} e^{-a|x|} dx < \infty$ for $a > 0$.

11. ★ If $f : \mathbf{R} \rightarrow \mathbf{R}$ is superlinear (Exercise 20 of §2.3) and continuous, the *Laplace transform*

$$L(s) = \int_{-\infty}^{\infty} e^{sx} e^{-f(x)} dx$$

is finite for all $s \in \mathbf{R}$. (Write $\int_{-\infty}^{\infty} = \int_{-\infty}^a + \int_a^b + \int_b^{\infty}$ for appropriately chosen a and b .)

12. ★ A function $\delta : \mathbf{R} \rightarrow \mathbf{R}$ is a *Dirac delta function* if it is nonnegative and satisfies

$$(3.12) \quad \int_{-\infty}^{\infty} \delta(x) f(x) dx = f(0)$$

for *all continuous* nonnegative $f : \mathbf{R} \rightarrow \mathbf{R}$. Show that there is no such function. (Construct continuous f 's which take on the two values 0 or 1 on most or all of \mathbf{R} , and insert them into (3.12).)

§4.4 The Fundamental Theorem of Calculus

By constructing appropriate covers, Archimedes was able to compute areas and integrals in certain situations. For example, he knew that $\int_0^1 x^2 dx = 1/3$. On the other hand, Archimedes was also able to compute tangent lines to certain curves and surfaces. However, he apparently had no idea that these two processes were intimately related, through the fundamental theorem of calculus. It was the discovery of the fundamental theorem, in the seventeenth century, that turned the computation of areas from a mystery to a simple and straightforward reality.

In this section, all functions will be continuous. Since we will use f^+ and f^- repeatedly, it is important to note that (§2.3) *a function is continuous iff both its positive and negative parts are continuous*.

Let f be continuous on (a, b) , and let $[c, d]$ be a compact subinterval. Since (§2.3) continuous functions map compact intervals to compact intervals, f is bounded on $[c, d]$, hence, integrable over (c, d) .

Let f be continuous on (a, b) , fix $a < c < b$, and set

$$F_c(x) = \begin{cases} \int_c^x f(t) dt, & c \leq x < b, \\ -\int_x^c f(t) dt, & a < x \leq c. \end{cases}$$

By the previous paragraph, $F_c(x)$ is finite for all $a < x < b$. From the previous section, we know that F_c is continuous. Here, we show that F_c is differentiable and $F'_c(x) = f(x)$ on (a, b) (Figure 4.28). We will need the modulus of continuity μ_x (§2.3) of f at x . To begin, by additivity, $F_c(y) - F_c(x) = F_x(y) - F_x(x)$ for any two points x, y in (a, b) , whether they are to the right or the left of c .

Then, for $a < x < t < y < b$, $|f(t) - f(x)| \leq \mu_x(y - x)$. Thus, $f(t) \leq f(x) + \mu_x(y - x)$. Hence,

$$\begin{aligned} \frac{F_c(y) - F_c(x)}{y - x} &= \frac{F_x(y) - F_x(x)}{y - x} \\ &= \frac{1}{y - x} \int_x^y f(t) dt \\ &\leq \frac{1}{y - x} \int_x^y [f(x) + \mu_x(y - x)] dt = f(x) + \mu_x(y - x). \end{aligned}$$

Similarly, since $a < x < t < y < b$ implies $f(x) - \mu_x(y - x) \leq f(t)$,

$$\frac{F_c(y) - F_c(x)}{y - x} \geq f(x) - \mu_x(y - x).$$

Combining the last two inequalities, we obtain

$$\left| \frac{F_c(y) - F_c(x)}{y - x} - f(x) \right| \leq \mu_x(y - x)$$

for $a < x < y < b$. If $a < y < x < b$, repeating the same steps yields

$$\left| \frac{F_c(y) - F_c(x)}{y - x} - f(x) \right| \leq \mu_x(x - y).$$

Hence, if $a < x \neq y < b$,

$$\left| \frac{F_c(y) - F_c(x)}{y - x} - f(x) \right| \leq \mu_x(|y - x|),$$

which implies, by continuity of f at x ,

$$\lim_{y \rightarrow x} \frac{F_c(y) - F_c(x)}{y - x} = f(x).$$

Hence, $F'_c(x) = f(x)$. We have established the following result, first mentioned in §3.6.

fig4.28.eps

FIGURE 4.28. The derivative at x of the integral of f is $f(x)$.

THEOREM. Every continuous $f : (a, b) \rightarrow \mathbf{R}$ has a primitive on (a, b) . \square

When f is continuous and integrable on (a, b) , we can do better.

THEOREM. Let $f : (a, b) \rightarrow \mathbf{R}$ be continuous and integrable. Then,

$$F(x) = \int_a^x f(t) dt, \quad a < x < b,$$

implies

$$F'(x) = f(x), \quad a < x < b,$$

and

$$F(x) = \int_x^b f(t) dt, \quad a < x < b,$$

implies

$$F'(x) = -f(x), \quad a < x < b.$$

To see this, for the first implication, write $\int_a^x f(t) dt = \int_a^c f(t) dt + F_c(x)$, and use $F'_c(x) = f(x)$. Since, by additivity, $\int_a^x f(t) dt + \int_x^b f(t) dt$ equals the constant $\int_a^b f(x) dx$, the second implication follows. \square

For example, if

$$F(x) = \int_0^{\tan x} e^{-t^2} dt, \quad 0 < x < \frac{\pi}{2},$$

then, $F'(x) = e^{-\tan^2 x} \sec^2 x$ by the above theorem combined with the chain rule. We will need this in §5.4.

The last two results show that integrals yield primitives. This is one version of the *fundamental theorem of calculus*. The other version of the fundamental theorem states that primitives yield integrals. When one is seeking areas or integrals, it is this version that is all-important.

FUNDAMENTAL THEOREM OF CALCULUS. Let f be nonnegative or integrable over (a, b) . Suppose that f is continuous on (a, b) , and let F be any primitive of f on (a, b) . Then, $F(b-)$ and $F(a+)$ exist, and

$$\int_a^b f(x) dx = F(b-) - F(a+).$$

To see this, first, assume that f is nonnegative. Then, F is increasing ($F' = f \geq 0$). Hence, $F(b-)$ and $F(a+)$ exist for any primitive F . In particular, with F_c as above, $F_c(b-)$ and $F_c(a+)$ exist. Since $F_c - F = k$ is a constant, by continuity at the endpoints,

$$\begin{aligned} \int_a^b f(x) dx &= \int_a^c f(x) dx + \int_c^b f(x) dx \\ &= \lim_{n \nearrow \infty} \int_{a+1/n}^c f(x) dx + \lim_{n \nearrow \infty} \int_c^{b-1/n} f(x) dx \end{aligned}$$

$$\begin{aligned}
&= -\lim_{n \nearrow \infty} F_c(a + 1/n) + \lim_{n \nearrow \infty} F_c(b - 1/n) \\
&= F_c(b-) - F_c(a+) \\
&= (F(b-) + k) - (F(a+) + k) = F(b-) - F(a+).
\end{aligned}$$

For the integrable case, let F^\pm denote primitives of f^\pm (here, F^\pm are *not* the positive and negative parts of F). Then, $F^+ - F^-$ differs from any primitive F of f by a constant k . Since $F^\pm(b-)$ and $F^\pm(a+)$ exist, so do $F(b-)$ and $F(a+)$. Hence,

$$\begin{aligned}
\int_a^b f(x) dx &= \int_a^b f^+(x) dx - \int_a^b f^-(x) dx \\
&= F^+(b-) - F^+(a+) - F^-(b-) + F^-(a+) \\
&= F(b-) + k - F(a+) - k = F(b-) - F(a+). \quad \square
\end{aligned}$$

Note that, in the fundamental theorem, as stated above, a or b or $F(a+)$ or $F(b-)$ may be infinite.

When a , b , $F(b-)$ and $F(a+)$ are all finite, the fundamental theorem simplifies slightly. Indeed, in this case, by defining $F(b) = F(b-)$, $F(a) = F(a+)$, the primitive F extends to a continuous function on the compact interval $[a, b]$ and the fundamental theorem becomes

$$\int_a^b f(x) dx = F(b) - F(a).$$

In particular, this simpler form of the fundamental theorem applies when f and (a, b) are both bounded. All primitives displayed below were obtained in §3.6.

For example, $\sin x$ is bounded and has the primitive $-\cos x$ on $(0, \pi)$. So,

$$\int_0^\pi \sin x dx = (-\cos \pi) - (-\cos 0) = 2.$$

Similarly, x^n , $n \geq 0$, is bounded and has the primitive $x^{n+1}/(n+1)$ over any bounded interval (a, b) , so,

$$(4.1) \quad \int_a^b x^n dx = \frac{b^{n+1}}{n+1} - \frac{a^{n+1}}{n+1}, \quad n \geq 0.$$

Below, it is convenient to denote $F(b) - F(a) = F(x)|_a^b$. Since, in §3.6, a primitive of f was written $\int f(x) dx$, the fundamental theorem becomes

$$\int_a^b f(x) dx = \int f(x) dx \Big|_a^b.$$

This explains the notation $\int f(x) dx$ for primitives. (The notation $\int_a^b f(x) dx$ for integrals was explained in §4.3.)

Also $f(x) = 1/\sqrt{x(1-x)} > 0$ has the primitive $F(x) = 2 \arcsin \sqrt{x}$ continuous over $[0, 1]$, so,

$$\int_0^1 \frac{dx}{\sqrt{x(1-x)}} = 2 \arcsin \sqrt{x} \Big|_0^1 = 2 \arcsin 1 = \pi.$$

Similarly, since $f(x) = 1/(1+x^2)$ is nonnegative and has the primitive $F(x) = \arctan x$ over \mathbf{R} ,

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \arctan x \Big|_{-\infty}^{\infty} = \frac{\pi}{2} - \left(-\frac{\pi}{2}\right) = \pi.$$

The unit disk

$$D = \{(x, y) : x^2 + y^2 < 1\}$$

is the disjoint union of a horizontal line segment and the two half-disks

$$D^{\pm} = \{(x, y) : x^2 + y^2 < 1, \pm y > 0\}.$$

Then, $\text{area}(D) = 2 \cdot \text{area}(D^+)$ (Exercise 11 of §4.2). But D^+ is the subgraph of $f(x) = \sqrt{1-x^2}$ over $(-1, 1)$, which has a primitive continuous on $[-1, 1]$. Hence,

$$\int_{-1}^1 \sqrt{1-x^2} dx = \frac{1}{2} \left(\arcsin x + x\sqrt{1-x^2} \right) \Big|_{-1}^1 = \frac{\pi}{2}.$$

This yields the following.

THEOREM. *The area of the unit disk is π . \square*

Of course, by translation and dilation invariance, the area of any disk of radius $r > 0$ is πr^2 . Another integral is

$$\int_0^1 (-\log x) dx = (x - x \log x) \Big|_0^1 = 1 + \lim_{x \rightarrow 0^+} x \log x = 1 + 0 = 1.$$

Our next item is the linearity of the integral.

THEOREM (LINEARITY). *Suppose that f, g are continuous on (a, b) . If f and g are both nonnegative or both integrable over (a, b) , then,*

$$\int_a^b [f(x) + g(x)] dx = \int_a^b f(x) dx + \int_a^b g(x) dx.$$

To see this, let F and G be primitives corresponding to f and g . Then, $f + g = F' + G' = (F + G)'$. So, $F + G$ is a primitive of $f + g$. By the fundamental theorem,

$$\begin{aligned} \int_a^b [f(x) + g(x)] dx &= F(b-) + G(b-) - F(a+) - G(a+) \\ &= \int_a^b f(x) dx + \int_a^b g(x) dx. \quad \square \end{aligned}$$

We say $f : (a, b) \rightarrow \mathbf{R}$ is *piecewise continuous* if there is a partition $a = x_0 < x_1 < \cdots < x_{n+1} = b$, such that f is continuous on each subinterval (x_{i-1}, x_i) , $i = 1, \dots, n+1$. Now, by additivity, the integral \int_a^b can be broken up into $\int_{x_{i-1}}^{x_i}$, $i = 1, \dots, n+1$. We conclude that *linearity also holds for piecewise continuous functions*.

By induction, linearity holds for finitely many (piecewise) continuous functions. If f_1, \dots, f_n are (piecewise) continuous and all nonnegative or all integrable over (a, b) , then,

$$\int_a^b \sum_{k=1}^n f_k(x) dx = \sum_{k=1}^n \int_a^b f_k(x) dx.$$

Since primitives are connected to integrals by the fundamental theorem, there is an integration by parts (§3.6) result for integrals.

THEOREM (INTEGRATION BY PARTS). *Let f and g be differentiable on (a, b) with $f'g$ and fg' continuous. If $f'g$ and fg' are both nonnegative or both integrable, then,*

$$\int_a^b f(x)g'(x) dx = f(x)g(x) \Big|_{a+}^{b-} - \int_a^b f'(x)g(x) dx.$$

This follows by applying the fundamental theorem to $f'g + fg' = (fg)'$ and using linearity. \square

Since primitives are connected to integrals by the fundamental theorem, there is a substitution (§3.6) result for integrals. Recall (§2.3) that continuous strictly monotone functions map open intervals to open intervals.

THEOREM (SUBSTITUTION). *Let g be differentiable and strictly monotone on an interval (a, b) with g' continuous, and let $(m, M) = g[(a, b)]$. Let $f : (m, M) \rightarrow \mathbf{R}$ be continuous. If f is nonnegative or integrable over (m, M) , then, $f[g(t)]|g'(t)|$ is nonnegative or integrable over (a, b) , and*

$$(4.2) \quad \int_m^M f(x) dx = \int_a^b f[g(t)]|g'(t)| dt.$$

To see this, first, assume that g is strictly increasing and f is nonnegative, let F be a primitive of f , let $H(t) = F[g(t)]$, and let $h(t) = f[g(t)]g'(t)$. Then, $(m, M) = (g(a+), g(b-))$ and $H'(t) = F'[g(t)]g'(t) = f[g(t)]g'(t) = h(t)$ by the chain rule. Hence, H is a primitive for h . Moreover, h is continuous and nonnegative, $F(M-) = H(b-)$, and $F(m+) = H(a+)$. By the fundamental theorem,

$$\begin{aligned} \int_m^M f(x) dx &= F(M-) - F(m+) \\ &= H(b-) - H(a+) \\ &= \int_a^b h(t) dt \\ &= \int_a^b f[g(t)]g'(t) dt. \end{aligned}$$

Since $|g'(t)| = g'(t)$, this establishes the case with g strictly increasing and f nonnegative. If f is integrable, apply the nonnegative case to f^\pm . Since the positive and negative parts of $f[g(t)]g'(t)$ are $f^\pm[g(t)]g'(t)$, the integrable case follows.

If g is strictly decreasing, then, $(m, M) = (g(b-), g(a+))$. Now, $h(t) = g(-t)$ is strictly increasing, $h((-b, -a)) = (m, M)$, and $h'(-t) = -g'(-t) = |g'(t)|$ is nonnegative on (a, b) . Applying what we just learned to f and h over $(-b, -a)$ yields

$$\begin{aligned} \int_m^M f(x) dx &= \int_{-b}^{-a} f[h(t)]h'(t) dt \\ &= \int_a^b f[h(-t)]h'(-t) dt \\ &= \int_a^b f[g(t)]|g'(t)| dt. \quad \square \end{aligned}$$

If g is not monotone, then, (4.2) has to be reformulated (Exercise 21). To see what happens, let us consider a simple example with $f(x) = 1$. Let $g : (a, b) \rightarrow (m, M)$ be *piecewise linear* with line segments inclined at $\pm\pi/4$. By this, we mean g is continuous on (a, b) and the graph of g is a line segment with slope ± 1 on each subinterval (t_{i-1}, t_i) , $i = 1, \dots, n+1$, for some partition $a = t_0 < t_1 < \dots < t_{n+1} = b$ of (a, b) (Figure 4.29). Then, $|g'(t)| = 1$ for all but finitely many t , so, $\int_a^b |g'(t)| dt = b - a$. On the other hand, substituting $f(x) = 1$ in (4.2) gives $\int_a^b |g'(t)| dt = M - m$. Thus, in such a situation, (4.2) cannot be correct unless the domain and the range have the same length, i.e., $M - m = b - a$.

To fix this, we have to take into account the extent to which g is not a bijection. To this end, for each x in (m, M) , let $\#(x)$ denote the number of points in the inverse image $g^{-1}(\{x\})$. Since (m, M) is the range of g , $\#(x) \geq 1$ for all x . The correct replacement for (4.2) with $f(x) = 1$ is

$$(4.3) \quad \int_m^M \#(x) dx = \int_a^b |g'(t)| dt.$$

This holds as long as g is continuous on (a, b) and there is a partition $a = t_0 < t_1 < \dots < t_{n+1} = b$ of (a, b) with g differentiable, g' continuous, and g strictly monotone on each subinterval (t_{i-1}, t_i) , for each $i = 1, \dots, n+1$ (Exercise 21).

For example, supposing $g : (a, b) \rightarrow (m, M)$ piecewise linear with slopes ± 1 , reduces (4.3) to $\int_m^M \#(x) dx = b - a$. Dividing by $M - m$ yields

$$(4.4) \quad \frac{1}{M - m} \int_m^M \#(x) dx = \frac{b - a}{M - m}.$$

Now, the left side of (4.4) may be thought of as the average value of $\#(x)$ over (m, M) . We conclude that, for a piecewise linear g with slopes ± 1 , the average value of the number of inverse images equals the ratio of the lengths of the domain over the range.

fig4.29.eps

FIGURE 4.29. Piecewise linear: $\#(x) = 4$, $\#(x') = 3$.

Now, we derive the integral version of

TAYLOR'S THEOREM. *Let $n \geq 0$ and suppose that f is $(n+1)$ times differentiable on (a, b) , with $f^{(n+1)}$ continuous on (a, b) . Suppose that $f^{(n+1)}$ is nonnegative or integrable over (a, b) , and fix $a < c < b$. Then,*

$$\begin{aligned} f(x) = & f(c) + f'(c)(x - c) + \frac{f''(c)}{2!}(x - c)^2 + \dots \\ & \dots + \frac{f^{(n)}(c)}{n!}(x - c)^n + \frac{h_{n+1}(x)}{(n+1)!}(x - c)^{n+1}, \end{aligned}$$

where

$$h_{n+1}(x) = (n+1) \int_0^1 (1-s)^n f^{(n+1)}[c + s(x-c)] ds.$$

To see this, recall, in §3.4, that we obtained $R_{n+1}(x, x) = 0$ and (here, ' denotes derivative with respect to t)

$$R'_{n+1}(x, t) = -\frac{f^{(n+1)}(t)}{n!}(x-t)^n.$$

Now, apply the fundamental theorem to $-R'_{n+1}(x, t)$ and substitute $t = c + s(x-c)$, $dt = (x-c)ds$, obtaining

$$\begin{aligned} R_{n+1}(x, c) &= \frac{1}{n!} \int_c^x f^{(n+1)}(t)(x-t)^n dt \\ &= \frac{(x-c)^{n+1}}{n!} \int_0^1 f^{(n+1)}(c+s(x-c))(1-s)^n ds \\ &= \frac{(x-c)^{n+1}}{(n+1)!} h_{n+1}(x). \quad \square \end{aligned}$$

In contrast with the Lagrange and Cauchy forms (§3.4) of the remainder, here, we need continuity and nonnegativity or integrability of $f^{(n+1)}$.

Our last item is the integration of power series. Since we already know (§3.6) how to find primitives of power series, the fundamental theorem and (4.1) yield the following.

THEOREM. *Suppose that $R > 0$ is the radius of convergence of*

$$f(x) = \sum_{n=0}^{\infty} a_n x^n.$$

If $[a, b] \subset (-R, R)$, then,

$$(4.5) \quad \int_a^b f(x) dx = \sum_{n=0}^{\infty} \int_a^b a_n x^n dx. \quad \square$$

For example, substituting $-x^2$ for x in the exponential series,

$$e^{-x^2} = 1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \frac{x^8}{4!} - \dots$$

Integrating this over $(0, 1)$, we obtain

$$\int_0^1 e^{-x^2} dx = 1 - \frac{1}{1!3} + \frac{1}{2!5} - \frac{1}{3!7} + \frac{1}{4!9} - \dots$$

This last result is, in general, false if $a = -R$ or $b = R$. For example, with $f(x) = e^{-x} = \sum_{n=0}^{\infty} (-1)^n x^n / n!$ and $(a, b) = (0, \infty)$, (4.5) reads $1 = \infty - \infty + \infty - \infty + \dots$. Under additional assumptions, however, (4.5) is true, even in these cases (see §5.2).

Exercises 4.4.

1. Compute $\int_0^\infty e^{-sx} dx$ for $s > 0$.
2. Compute $\int_0^1 x^{r-1} dx$ and $\int_1^\infty x^{r-1} dx$ and $\int_0^\infty x^{r-1} dx$ for r real. (There are three cases, $r < 0$, $r = 0$, and $r > 0$.)
3. Suppose that f is continuous over (a, b) , and let F be any primitive. If f and (a, b) are both bounded, then, f is integrable, and $F(a+)$ and $F(b-)$ are finite.
4. Let $f(x) = \sin x/x$, $x > 0$, and let $F(b) = \int_0^b f(x) dx$, $b > 0$. Show that $F(\infty) = \lim_{b \rightarrow \infty} F(b)$ exists and is finite. (Write $F(b) = \int_0^1 f(x) dx + \int_1^b f(x) dx$, integrate the second integral by parts, and use (3.1). This limit is computed in Exercise 12 of §5.4.)
5. For f continuous and nonnegative or integrable over $(0, 1)$,

$$\int_0^1 f(x) dx = \int_0^\infty e^{-t} f(e^{-t}) dt.$$

6. Compute $\int_0^\infty e^{-sx} x^n dx$ for $s > 0$ and $n \geq 0$. (Integration by parts.)
7. Compute $\int_0^\infty e^{-nx} \sin(sx) dx$ and $\int_0^\infty e^{-nx} \cos(sx) dx$ for $n \geq 1$. (Integration by parts.)
8. Show that $\int_0^\infty e^{-t^2/2} t^x dt = (x-1) \int_0^\infty e^{-t^2/2} t^{x-2} dt$ for $x > 1$. Use this to derive

$$\int_0^\infty e^{-t^2/2} t^{2n+1} dt = 2^n n!, \quad n \geq 0.$$

(Integration by parts.)

9. Compute $\int_0^1 (1-t)^n t^{x-1} dt$ for $x > 0$ and $n \geq 1$. (Integration by parts.)
10. Compute $\int_0^1 (-\log x)^n dx$.
11. ★ Show that

$$\int_{-1}^1 (x^2 - 1)^n dx = (-1)^n \frac{2n \cdot (2n-2) \cdot \dots \cdot 2}{(2n+1) \cdot (2n-1) \cdot \dots \cdot 3} \cdot 2.$$

(Integrate by parts.)

12. ★ For $n \geq 0$, the *Legendre polynomial* P_n (of degree n) is given by $P_n(x) = f^{(n)}(x)/2^n n!$, where $f(x) = (x^2 - 1)^n$. Show that

$$\int_{-1}^1 P_n(x)^2 dx = \frac{2}{2n+1}.$$

13. Use the integral test (§4.3) to show that

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}, \quad s > 1,$$

converges.

14. Use the integral test (§4.3) to show that

$$\gamma = \lim_{n \nearrow \infty} \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \log n \right)$$

exists and $0 < \gamma < 1$. This particular real γ is *Euler's constant*.

15. Compute $\int_{-\pi}^{\pi} x \cos(nx) dx$ and $\int_{-\pi}^{\pi} x \sin(nx) dx$ for $n \geq 0$. (Integration by parts.)

16. ★ Compute $\int_{-\pi}^{\pi} f(nx)g(mx) dx$, $n, m \geq 0$, with $f(x)$ and $g(x)$ equal to $\sin x$ or $\cos x$ (three possibilities — use (5.3) in §3.5).

17. ★ If $f, g : (a, b) \rightarrow \mathbf{R}$ are nonnegative and continuous, derive the *Cauchy-Schwarz inequality*

$$\left[\int_a^b f(x)g(x) dx \right]^2 \leq \left[\int_a^b f(x)^2 dx \right] \cdot \left[\int_a^b g(x)^2 dx \right].$$

(Use the fact that $q(t) = \int_a^b [f(x) + tg(x)]^2 dx$ is a nonnegative quadratic polynomial and Exercise 5 of §1.4.)

18. For $n \geq 1$, show that

$$\int_0^n \frac{1 - (1 - t/n)^n}{t} dt = 1 + \frac{1}{2} + \cdots + \frac{1}{n}.$$

19. We say $f : (a, b) \rightarrow \mathbf{R}$ is *piecewise differentiable* if there is a partition $a = x_0 < x_1 < \cdots < x_{n+1} = b$, such that f restricted to (x_{i-1}, x_i) is differentiable for $i = 1, \dots, n+1$. Let $f : (a, b) \rightarrow \mathbf{R}$ be piecewise continuous and integrable. Show that $F(x) = \int_a^x f(t) dt$, $a < x < b$, is continuous on (a, b) and piecewise differentiable on (a, b) .

20. ★ If $f : (a, b) \rightarrow \mathbf{R}$ is nonnegative and $g : [a, b] \rightarrow \mathbf{R}$ is nonnegative and continuous, then,

$$\int_a^b [f(x) + g(x)] dx = \int_a^b f(x) dx + \int_a^b g(x) dx.$$

(Use Exercise 5 of §4.3 and approximate g by a piecewise constant g_ϵ as in §2.3. Since f is arbitrary, linearity may not be used directly.)

21. ★ Suppose that $g : (a, b) \rightarrow (m, M)$ is continuous, and suppose that there is a partition $a = t_0 < t_1 < \cdots < t_{n+1} = b$ of (a, b) , such that g is differentiable, g' is continuous, and g is strictly monotone on each subinterval (t_{i-1}, t_i) , for each $i = 1, \dots, n+1$. For each x in (m, M) , let $\#(x)$ denote the number of points in the inverse image $g^{-1}(\{x\})$. Also let $f : (m, M) \rightarrow \mathbf{R}$ be continuous and nonnegative. Then,

$$\int_m^M f(x)\#(x) dx = \int_a^b f[g(t)]|g'(t)| dt.$$

(Use additivity on the integral \int_a^b .)

22. ★ Let f be differentiable with f' continuous on an interval containing $[a, b]$. Show that the variation of f corresponding to any partition in (a, b) (Exercise 4 of §2.2) is bounded by $\int_a^b |f'(x)| dx$. Use Exercise 3 of §4.3 to show that the total variation of f over (a, b) equals $\int_a^b |f'(x)| dx$. (Rewrite the variation of f over a given partition as a Riemann sum for $|f'|$.)

§4.5 The Method of Exhaustion

In this section, we compute the area of the unit disk D via the Method of Exhaustion.

For $n \geq 3$, let $P_k = (\cos(2\pi k/n), \sin(2\pi k/n))$, $0 \leq k \leq n$. Then, the points P_k are evenly spaced about the unit circle $\{(x, y) : x^2 + y^2 = 1\}$, and $P_n = P_0$. Let $D_n \subset D$ be the interior of the inscribed regular n -sided polygon obtained by joining the points P_0, P_1, \dots, P_n (we do not include the edges of D_n in the definition of D_n). Then (Exercise 13 of §4.2),

$$\text{area}(D_n) = \frac{n}{2} \sin(2\pi/n) = \pi \cdot \frac{\sin(2\pi/n)}{2\pi/n}.$$

Since $\lim_{x \rightarrow 0} \frac{\sin x}{x} = \sin' 0 = \cos 0 = 1$, we obtain

$$(5.1) \quad \lim_{n \nearrow \infty} \text{area}(D_n) = \pi.$$

Since

$$(5.2) \quad D_4 \subset D_8 \subset D_{16} \subset \dots, \quad \text{and} \quad D = \bigcup_{n=2}^{\infty} D_{2^n},$$

it is reasonable to make the guess that

$$(5.3) \quad \text{area}(D) = \lim_{n \nearrow \infty} \text{area}(D_{2^n}),$$

and, hence, conclude that $\text{area}(D) = \pi$. The reasoning that leads from (5.2) to (5.3) is generally correct. The result is called the *Method of Exhaustion*.

Although $\text{area}(D)$ was computed in the previous section using the fundamental theorem, in Chapter 5 we will need the Method to compute other areas.

We say that a sequence of sets (A_n) is *increasing* (Figure 4.30) if $A_1 \subset A_2 \subset A_3 \subset \dots$.

fig4.30.eps

FIGURE 4.30. An increasing sequence of sets.

THEOREM (METHOD OF EXHAUSTION). *If $A_1 \subset A_2 \subset \dots$ is an increasing sequence of subsets of \mathbf{R}^2 , then,*

$$\text{area} \left(\bigcup_{n=1}^{\infty} A_n \right) = \lim_{n \nearrow \infty} \text{area}(A_n).$$

We warn the reader that the result is false, in general, for decreasing sequences. For example, take $A_n = (n, \infty) \times (-\infty, \infty)$, $n \geq 1$. Then, $\text{area}(A_n) = \infty$ for all $n \geq 1$, but $\bigcap_{n=1}^{\infty} A_n = \emptyset$ so $\text{area}(\bigcap_{n=1}^{\infty} A_n) = 0$. This lack of symmetry between increasing and decreasing sequences is a reflection of the lack of symmetry in the definition of area: $\text{area}(A)$ is defined as an inf of over-estimates, not as a sup of underestimates.

The Method is established in three stages: first, when (definitions below) the sets A_n , $n \geq 1$, are open; then, when the sets A_n , $n \geq 1$, are interopen; finally, for arbitrary sets A_n , $n \geq 1$. *Open* and *interopen* are structural properties of sets that we describe below.

We call a set $G \subset \mathbf{R}^2$ *open* if every point $(a, b) \in G$ can be surrounded by a nonempty, open rectangle wholly contained in G . For example an open rectangle is an open set, but a compact rectangle Q is not, since no point on the edges of Q can be surrounded by a rectangle wholly contained in Q . The n -sided polygon D_n , considered above, is an open set as is the unit disk D . Since there are no points in \emptyset for which the open criterion fails, \emptyset is open.

For our purposes, the most important example of an open set is given by the following.

THEOREM. *If $f \geq 0$ is continuous on (a, b) , its subgraph is an open subset of \mathbf{R}^2 .*

To see this, pick x and y with $a < x < b$ and $0 < y < f(x)$. We have to find a rectangle Q containing (x, y) and contained in the subgraph. Pick $y < y_1 < f(x)$. We claim there is a $c > 0$, such that $|t - x| < c$ implies $a < t < b$ and $f(t) > y_1$. If not, then, for all $n \geq 1$, we can find a real t_n in the interval $(x - 1/n, x + 1/n)$ contradicting the stated property, i.e., satisfying $f(t_n) \leq y_1$. Then, $t_n \rightarrow x$, so, by continuity $f(t_n) \rightarrow f(x)$. Hence, $f(x) \leq y_1$, contradicting our initial choice of y_1 . Thus, there is a $c > 0$, such that the rectangle $Q = (x - c, x + c) \times (0, y_1)$ contains (x, y) and lies in the subgraph. \square

Thus, *the integral of a continuous nonnegative function is the area of an open set.*

An alternative description of open sets is in terms of distance. If (a, b) is a point and A is a set, then, the distance $d((a, b), A)$ between the point (a, b) and A , by definition, is the distance between the set $\{(a, b)\}$ and the set A (§4.2). For example, if Q is an open rectangle and $(a, b) \in Q$, then, $d((a, b), Q^c)$ is positive. Here and below, $A^c = \mathbf{R}^2 \setminus A$.

THEOREM. *A set G is open iff $d((a, b), G^c) > 0$ for all points $(a, b) \in G$.*

Indeed, if $(a, b) \in G$ and $Q \subset G$ contains (a, b) , then, $Q^c \supset G^c$, so, $d((a, b), G^c) \geq d((a, b), Q^c) > 0$. Conversely, if $d = d((a, b), G^c) > 0$, then, the disk R of radius $d/2$ and center (a, b) lies wholly in G . Now, choose any rectangle Q in R containing (a, b) . \square

If G, G' are open subsets, so are $G \cup G'$ and $G \cap G'$. In fact, if (G_n) is a sequence of open sets, then, $G = \bigcup_{n=1}^{\infty} G_n$ is open. To see this, if $(a, b) \in G$, then, $(a, b) \in G_n$ for some specific n . Since the specific G_n is open, there is a rectangle Q with $(a, b) \in Q \subset G_n \subset G$. Hence, G is open. Thus, an infinite union of open sets is open. If G_1, \dots, G_n are finitely many open sets, then, $G = G_1 \cap G_2 \cap \dots \cap G_n$ is open. To see this, if $(a, b) \in G$, then, $(a, b) \in G_k$ for all $1 \leq k \leq n$, so, there are open rectangles Q_k with $(a, b) \in Q_k \subset G_k, 1 \leq k \leq n$. Hence, $Q = \bigcap_{k=1}^n Q_k$ is an open rectangle containing (a, b) and contained in G (a finite intersection of open rectangles is an open rectangle). Thus, a finite intersection of open sets is open. However, an infinite intersection of open sets need not be open.

If $A \subset \mathbf{R}^2$ is any set and $\epsilon > 0$, by definition of area, we can find an open set G containing A and satisfying $\text{area}(G) \leq \text{area}(A) + \epsilon$ (Exercise 6). If we had additivity and $\text{area}(A) < \infty$, writing $\text{area}(G) = \text{area}(A) + \text{area}(G \setminus A)$, we would conclude that $\text{area}(G \setminus A) \leq \epsilon$. Conversely, if we are seeking properties of sets that guarantee additivity, we may, instead, focus on those sets M in \mathbf{R}^2 satisfying the above approximability condition: *For all $\epsilon > 0$, there is an open superset G of M , such that $\text{area}(G \setminus M) \leq \epsilon$.* Instead of doing this, however, it will be quicker for us to start with an alternate equivalent (Exercise 16) formulation.

We say a set $M \subset \mathbf{R}^2$ is *measurable* if

$$(5.4) \quad \text{area}(A) = \text{area}(A \cap M) + \text{area}(A \cap M^c), \quad \text{for all } A \subset \mathbf{R}^2.$$

For example, the empty set is measurable and M is measurable iff M^c is measurable. Below, we show that every open set is measurable. Measurability may be looked upon as a strengthened form of additivity, since the equality in (5.4) is required to hold for *every* $A \subset \mathbf{R}^2$. Note that the trick, below, of summing alternate areas C_1, C_3, C_5, \dots was already used in derivating additivity in §4.3. Compare the next derivation with that derivation!

In §4.2, we established additivity when the sets were well separated. Now, we establish a similar result involving open sets.

THEOREM. *If G is open, then, G is measurable.*

To see this, we need show only that

$$(5.5) \quad \text{area}(A) \geq \text{area}(A \cap G) + \text{area}(A \cap G^c)$$

for every $A \subset \mathbf{R}^2$, since the reverse inequality follows by subadditivity. Let $A \subset \mathbf{R}^2$ be arbitrary. If $\text{area}(A) = \infty$, (5.5) is immediate, so, let us assume that $\text{area}(A) < \infty$. Let G_n be the set of points in G whose distance from G^c is at least $1/n$. Since $A \cap G_n$ and $A \cap G^c$ are well separated (Figure 4.31),

$$\text{area}(A) \geq \text{area}(A \cap G_n) + \text{area}(A \cap G^c).$$

By subadditivity,

$$\text{area}(A \cap G) \leq \text{area}(A \cap G_n) + \text{area}(A \cap G \cap G_n^c).$$

Combining the last two inequalities, we obtain

$$(5.6) \quad \text{area}(A) \geq \text{area}(A \cap G) + \text{area}(A \cap G^c) - \text{area}(A \cap G \cap G_n^c).$$

Thus, if we show that

$$(5.7) \quad \lim_{n \nearrow \infty} \text{area}(A \cap G \cap G_n^c) = 0,$$

letting $n \nearrow \infty$ in (5.6), we obtain (5.5), hence, the result.

To obtain (5.7), let C_n be the set of points (a, b) in G satisfying $1/(n+1) \leq d((a, b), G^c) < 1/n$. Since G is open, $d((a, b), G^c) > 0$ for every point in G . Thus,

$$G \cap G_n^c = C_n \cup C_{n+1} \cup C_{n+2} \cup \dots$$

But the sets $C_n, C_{n+2}, C_{n+4}, \dots$, are well separated. Hence,

$$\text{area}(A \cap C_n) + \text{area}(A \cap C_{n+2}) + \text{area}(A \cap C_{n+4}) + \dots \leq \text{area}(A \cap G).$$

Since $C_{n+1}, C_{n+3}, C_{n+5}, \dots$, are well separated,

$$\text{area}(A \cap C_{n+1}) + \text{area}(A \cap C_{n+3}) + \text{area}(A \cap C_{n+5}) + \dots \leq \text{area}(A \cap G).$$

Adding the last two inequalities, by subadditivity, we obtain

$$(5.8) \quad \text{area}(A \cap G \cap G_n^c) \leq \sum_{k=n}^{\infty} \text{area}(A \cap C_k) \leq 2 \text{area}(A \cap G) < \infty.$$

Now, (5.8) with $n = 1$ shows that the series $\sum_{k=1}^{\infty} \text{area}(A \cap C_k)$ converges. Thus, the tail series, starting from $k = n$ in (5.8), approaches zero, as $n \nearrow \infty$. This establishes (5.7). \square

fig4.31.eps

FIGURE 4.31. An open set is measurable.

Now we establish the Method for measurable, hence, for open sets. In fact, we need to establish a strengthened form of the Method for measurable sets.

THEOREM (MEASURABLE METHOD OF EXHAUSTION). *If $M_1 \subset M_2 \subset \dots$ is an increasing sequence of measurable subsets of \mathbf{R}^2 and $A \subset \mathbf{R}^2$ is arbitrary, then,*

$$\text{area} \left[A \cap \left(\bigcup_{n=1}^{\infty} M_n \right) \right] = \lim_{n \nearrow \infty} \text{area}(A \cap M_n).$$

As stated previously, the Method is obtained by choosing $A = \mathbf{R}^2$. To derive this, let $M_\infty = \bigcup_{n=1}^{\infty} M_n$. Since $A \cap M_n \subset A \cap M_\infty$, by monotonicity, the sequence $(\text{area}(A \cap M_n))$ is increasing and bounded above by $\text{area}(A \cap M_\infty)$. Thus,

$$\lim_{n \nearrow \infty} \text{area}(A \cap M_n) \leq \text{area}(A \cap M_\infty).$$

To obtain the reverse inequality, apply (5.4) with M and A , there, replaced by M_1 and $A \cap M_2$ respectively, obtaining

$$\text{area}(A \cap M_2) = \text{area}(A \cap M_2 \cap M_1) + \text{area}(A \cap M_2 \cap M_1^c).$$

Since $A \cap M_2 \cap M_1 = A \cap M_1$, this implies that

$$\text{area}(A \cap M_2) = \text{area}(A \cap M_1) + \text{area}(A \cap M_2 \cap M_1^c).$$

Now, apply (5.4) with M and A , there, replaced by M_2 and $A \cap M_3$ respectively, obtaining

$$\begin{aligned} \text{area}(A \cap M_3) &= \text{area}(A \cap M_2) + \text{area}(A \cap M_3 \cap M_2^c) \\ &= \text{area}(A \cap M_1) + \text{area}(A \cap M_2 \cap M_1^c) \\ &\quad + \text{area}(A \cap M_3 \cap M_2^c). \end{aligned}$$

Proceeding in this manner, we obtain

$$\text{area}(A \cap M_n) = \text{area}(A \cap M_1) + \sum_{k=2}^n \text{area}(A \cap M_k \cap M_{k-1}^c).$$

Sending $n \nearrow \infty$, we obtain

$$\lim_{n \nearrow \infty} \text{area}(A \cap M_n) = \text{area}(A \cap M_1) + \sum_{k=2}^{\infty} \text{area}(A \cap M_k \cap M_{k-1}^c).$$

Since

$$M_1 \cup (M_2 \cap M_1^c) \cup (M_3 \cap M_2^c) \cup \cdots = M_{\infty},$$

subadditivity implies that

$$\text{area}(A \cap M_{\infty}) \leq \text{area}(A \cap M_1) + \sum_{k=2}^{\infty} \text{area}(A \cap M_k \cap M_{k-1}^c).$$

Hence, we obtain the reverse inequality

$$\lim_{n \nearrow \infty} \text{area}(A \cap M_n) \geq \text{area}(A \cap M_{\infty}). \quad \square$$

By choosing $A = \mathbf{R}^2$, we conclude that the Method is valid for measurable, hence, open sets. *This completes stage one of the derivation of the Method.*

Next we establish the Method for interopen sets. A set $I \subset \mathbf{R}^2$ is *interopen* if I is the infinite intersection of a sequence of open sets (G_n) , $I = \bigcap_{n=1}^{\infty} G_n$. Of course, every open set is interopen. Also, every compact rectangle is interopen (Exercise 5). The key feature of interopen sets is that any set A can be covered by some interopen set I , $A \subset I$, having the same area, $\text{area}(A) = \text{area}(I)$ (Exercise 7).

THEOREM. *If (M_n) is a sequence of measurable sets, then, $\bigcap_{n=1}^{\infty} M_n$ is measurable.*

To derive this theorem, we start with two measurable sets M, N , and we show that $M \cap N$ is measurable. First, note that

$$(5.9) \quad (M \cap N)^c = (M \cap N^c) \cup (M^c \cap N) \cup (M^c \cap N^c).$$

Let $A \subset \mathbf{R}^2$ be arbitrary. Since N is measurable, write (5.4) with $A \cap M$ and N replacing A and M , respectively, obtaining

$$\text{area}(A \cap M) = \text{area}(A \cap M \cap N) + \text{area}(A \cap M \cap N^c).$$

Now, write (5.4) with $A \cap M^c$ and N replacing A and M respectively, obtaining

$$\text{area}(A \cap M^c) = \text{area}(A \cap M^c \cap N) + \text{area}(A \cap M^c \cap N^c).$$

Now, insert the last two equalities in (5.4). By (5.9) and subadditivity, we obtain

$$\begin{aligned} \text{area}(A) &= \text{area}(A \cap M) + \text{area}(A \cap M^c) \\ &= \text{area}(A \cap (M \cap N)) + \text{area}(A \cap (M \cap N^c)) \\ &\quad + \text{area}(A \cap (M^c \cap N)) + \text{area}(A \cap (M^c \cap N^c)) \\ &\geq \text{area}(A \cap (M \cap N)) + \text{area}(A \cap (M \cap N)^c). \end{aligned}$$

Hence,

$$\text{area}(A) \geq \text{area}(A \cap (M \cap N)) + \text{area}(A \cap (M \cap N)^c).$$

Since the reverse inequality is an immediate consequence of subadditivity, we conclude that $M \cap N$ is measurable.

Now, let (M_n) be a sequence of measurable sets and set $N_n = \bigcap_{k=1}^n M_k$, $n \geq 1$. Then, N_n , $n \geq 1$, are measurable. Indeed $N_1 = M_1$ is measurable. For the inductive step, suppose that N_n is measurable. Since $N_{n+1} = N_n \cap M_{n+1}$, we conclude that N_{n+1} is measurable. Hence, by induction, N_n is measurable for all $n \geq 1$. Now, $M_\infty = \bigcap_{n=1}^\infty M_n = \bigcap_{n=1}^\infty N_n$ and $N_1 \supset N_2 \supset \dots$. Hence, $N_1^c \subset N_2^c \subset \dots$, so, by the measurable Method, we obtain

$$\begin{aligned} \text{area}(A \cap M_\infty^c) &= \text{area} \left[A \cap \left(\bigcap_{n=1}^\infty N_n \right)^c \right] \\ &= \text{area} \left(A \cap \bigcup_{n=1}^\infty N_n^c \right) \\ (5.10) \qquad &= \lim_{n \nearrow \infty} \text{area}(A \cap N_n^c). \end{aligned}$$

Here, we used De Morgan's law (§1.1). Now, for each $n \geq 1$,

$$\begin{aligned} \text{area}(A) &= \text{area}(A \cap N_n) + \text{area}(A \cap N_n^c) \\ (5.11) \qquad &\geq \text{area}(A \cap M_\infty) + \text{area}(A \cap N_n^c). \end{aligned}$$

Sending $n \nearrow \infty$ in (5.11) and using (5.10) yields

$$\text{area}(A) \geq \text{area}(A \cap M_\infty) + \text{area}(A \cap M_\infty^c).$$

Since the reverse inequality follows from subadditivity, we conclude that $M_\infty = \bigcap_{n=1}^\infty M_n$ is measurable. \square

By choosing (M_n) in the theorem to consist of open sets, we see that every interopen set is measurable. Hence, we conclude that the Method is valid for interopen sets. *This completes stage two of the derivation of the Method.*

The third and final stage of the derivation of the Method is to establish it for an increasing sequence of arbitrary sets. To this end, let $A_1 \subset A_2 \subset \dots$ be an arbitrary increasing sequence of sets. For each $n \geq 1$, by Exercise 7, choose an interopen set I_n containing A_n and having the same area: $I_n \supset A_n$ and $\text{area}(I_n) = \text{area}(A_n)$. For each $n \geq 1$, let

$$J_n = I_n \cap I_{n+1} \cap I_{n+2} \cap \dots$$

Then, J_n is interopen, $A_n \subset J_n \subset I_n$, and $\text{area}(J_n) = \text{area}(A_n)$, for all $n \geq 1$. Moreover $J_n = I_n \cap J_{n+1}$. Hence (and this is the reason for introducing the sequence (J_n)), the sequence (J_n) is increasing. Thus, by applying the Method for interopen sets,

$$\begin{aligned} \lim_{n \nearrow \infty} \text{area}(A_n) &= \lim_{n \nearrow \infty} \text{area}(J_n) \\ (5.12) \qquad &= \text{area} \left(\bigcup_{n=1}^{\infty} J_n \right) \geq \text{area} \left(\bigcup_{n=1}^{\infty} A_n \right). \end{aligned}$$

On the other hand, by monotonicity, the sequence $(\text{area}(A_n))$ is increasing and bounded above by $\text{area}(\bigcup_{n=1}^{\infty} A_n)$. Hence,

$$\lim_{n \nearrow \infty} \text{area}(A_n) \leq \text{area} \left(\bigcup_{n=1}^{\infty} A_n \right).$$

Combining this with (5.12), we conclude that

$$\lim_{n \nearrow \infty} \text{area}(A_n) = \text{area} \left(\bigcup_{n=1}^{\infty} A_n \right).$$

This completes stage three, hence, the derivation of the Method. \square

We end by describing the connection between the areas of the inscribed and circumscribed polygons of the unit disk D , as the number of sides doubles. Let

$$P_k = \left(\frac{\cos(2\pi k/n)}{\cos(\pi/n)}, \frac{\sin(2\pi k/n)}{\cos(\pi/n)} \right), \quad 0 \leq k \leq n.$$

Then, the points P_k are evenly spaced about the circle $\{(x, y) : x^2 + y^2 = \sec^2(\pi/n)\}$, and $P_n = P_0$. Let D'_n denote the interior of the regular n -sided polygon obtained by joining the points P_0, \dots, P_n by line segments. Then, $D'_n \supset D$ and $D'_n = c \cdot D_n$ with $c = \sec(\pi/n)$. Hence, by dilation invariance, we obtain

$$\text{area}(D'_n) = c^2 \cdot \text{area}(D_n) = n \tan(\pi/n)$$

which also goes to π as $n \nearrow \infty$.

Let a_n, a'_n denote the areas of the inscribed and circumscribed n -sided polygons D_n, D'_n , respectively. Then, using trigonometry, one obtains (Exercise 11)

$$(5.13g) \qquad a_{2n} = \sqrt{a_n a'_n}$$

and

$$(5.13h) \qquad \frac{1}{a'_{2n}} = \frac{1}{2} \left(\frac{1}{a_{2n}} + \frac{1}{a'_n} \right).$$

Since $a_4 = 2$ and $a'_4 = 4$, we obtain $a_8 = 2\sqrt{2}$ and $a'_8 = 8(\sqrt{2} - 1)$. Thus,

$$2\sqrt{2} < \pi < 8(\sqrt{2} - 1).$$

Continuing in this manner, one obtains approximations to π . These identities are very similar to those leading to Gauss' *arithmetic-geometric mean*, which we discuss in §5.3.

Exercises 4.5.

1. If Q is an open rectangle and $(x, y) \in Q$, then, $d((x, y), Q^c) > 0$.
2. Find a sequence (A_n) of open sets, such that $\bigcap_{n=1}^{\infty} A_n$ is not open.
3. ★ A set A is *closed* if A^c is open. Show that a compact rectangle is closed, an infinite intersection of closed sets is closed, and a finite union of closed sets is closed. Find a sequence (A_n) of closed sets, such that $\bigcup_{n=1}^{\infty} A_n$ is not closed. (You will need De Morgan's law (§1.1).)
4. ★ Given a real a , let L_a denote the vertical infinite line through a , $L_a = \{(x, y) : x = a, y \in \mathbf{R}\}$. Also set $L_{-\infty} = L_{+\infty} = \emptyset$. Let f be nonnegative and continuous on (a, b) . Show that

$$C = \{(x, y) : a < x < b, 0 \leq y \leq f(x)\} \cup L_a \cup L_b$$

is a closed set and

$$\int_a^b f(x) dx = \text{area}(C).$$

This shows that the integral of a continuous nonnegative function is also the area of a closed set. (Compare C with the subgraph of $f(x) + \epsilon/(1+x^2)$ for $\epsilon > 0$ small.)

5. ★ Show that C is closed iff

$$d((x, y), C) = 0 \quad \iff \quad (x, y) \in C.$$

If C is closed and $G_n = \{(x, y) : d((x, y), C) < 1/n\}$, then, G_n is open and $C = \bigcap_{n=1}^{\infty} G_n$. Thus, every closed set is interopen.

6. Let $A \subset \mathbf{R}^2$ be arbitrary. Use the definition of $\text{area}(A)$ to show: For all $\epsilon > 0$, there is an open superset G of A satisfying $\text{area}(G) \leq \text{area}(A) + \epsilon$. Conclude that

$$\text{area}(A) = \inf\{\text{area}(G) : A \subset G, G \text{ open}\}.$$

7. Let $A \subset \mathbf{R}^2$ be arbitrary. Show that there is an interopen set I containing A and having the same area as A (use Exercise 6).
8. If (M_n) is a sequence of measurable sets, then, $\bigcup_{n=1}^{\infty} M_n$ is measurable.
9. ★ The Cantor set is closed.
10. Show that $D'_n \supset D$.
11. Derive (5.13g) and (5.13h).
12. If A and B are disjoint and A is measurable, then, $\text{area}(A \cup B) = \text{area}(A) + \text{area}(B)$.
13. ★ If (A_n) is a sequence of disjoint measurable sets, then,

$$\text{area}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \text{area}(A_n).$$

14. If A and B are measurable, then, $\text{area}(A \cup B) = \text{area}(A) + \text{area}(B) - \text{area}(A \cap B)$. ■
15. ★ Let A, B, C, D be measurable subsets of \mathbf{R}^2 . Obtain expressions for $\text{area}(A \cup B \cup C)$ and $\text{area}(A \cup B \cup C \cup D)$ akin to the result in the previous Exercise.
16. ★ Show that M is measurable iff, for all $\epsilon > 0$, there is an open superset G of M , such that $\text{area}(G \setminus M) \leq \epsilon$.
17. ★ Let $A \subset \mathbf{R}^2$ be measurable. If $\text{area}(A) > 0$, there is an $\epsilon > 0$, such that $\text{area}(A \cap A') > 0$ for all translates $A' = A + (a, b)$ of A with $|a| < \epsilon$ and $|b| < \epsilon$. (Start with A a rectangle, and use Exercise 15 of §4.2.)
18. ★ If $A \subset \mathbf{R}^2$ is measurable and $\text{area}(A) > 0$, let

$$A - A = \{(x - x', y - y') : (x, y) \text{ and } (x', y') \in A\}$$

be the *set of differences*. Note that $A - A$ contains the origin. Then, for some $\epsilon > 0$, $A - A$ must contain the open rectangle $Q_\epsilon = (-\epsilon, \epsilon) \times (-\epsilon, \epsilon)$. (Use Exercise 17.)