

FIGURE 12

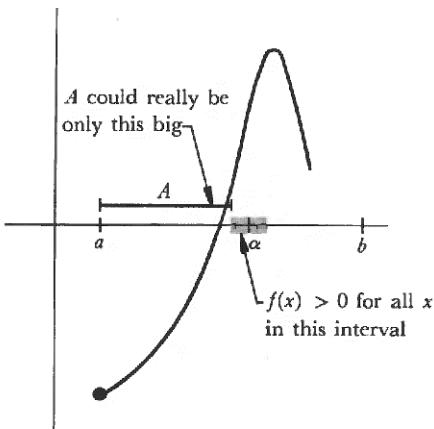


FIGURE 13

Now suppose α is the smallest number which is greater than all members of A ; clearly $a < \alpha < b$. We claim that $f(\alpha) = 0$, and to prove this we only have to eliminate the possibilities $f(\alpha) < 0$ and $f(\alpha) > 0$.

Suppose first that $f(\alpha) < 0$. Then, by Theorem 6-3, $f(x)$ would be less than 0 for all x in a small interval containing α , in particular for some numbers bigger than α (Figure 12); but this contradicts the fact that α is bigger than every member of A , since the larger numbers would also be in A . Consequently, $f(\alpha) < 0$ is false.

On the other hand, suppose $f(\alpha) > 0$. Again applying Theorem 6-3, we see that $f(x)$ would be positive for all x in a small interval containing α , in particular for some numbers smaller than α (Figure 13). This means that these smaller numbers are all *not* in A . Consequently, one could have chosen an even smaller α which would be greater than all members of A . Once again we have a contradiction; $f(\alpha) > 0$ is also false. Hence $f(\alpha) = 0$ and, we are tempted to say, Q.E.D.

We know, however, that something must be wrong, since no new properties of \mathbf{R} were ever used, and it does not require much scrutiny to find the dubious point. It is clear that we can choose a number α which is greater than all members of A (for example, we can choose $\alpha = b$), but it is not so clear that we can choose a *smallest* one. In fact, suppose A consists of all numbers $x \geq 0$ such that $x^2 < 2$. If the number $\sqrt{2}$ did not exist, there would not be a least number greater than all the members of A ; for any $y > \sqrt{2}$ we chose, we could always choose a still smaller one.

Now that we have discovered the fallacy, it is almost obvious what additional property of the real numbers we need. All we must do is say it properly and use it. That is the business of the next chapter.

PROBLEMS

1. For each of the following functions, decide which are bounded above or below on the indicated interval, and which take on their maximum or minimum value. (Notice that f *might* have these properties even if f is not continuous, and even if the interval is not a closed interval.)

(i) $f(x) = x^2$ on $(-1, 1)$.

(ii) $f(x) = x^3$ on $(-1, 1)$.

(iii) $f(x) = x^2$ on \mathbf{R} .

(iv) $f(x) = x^2$ on $[0, \infty)$.

(v) $f(x) = \begin{cases} x^2, & x \leq a \\ a+2, & x > a \end{cases}$ on $(-a-1, a+1)$. (It will be necessary to consider several possibilities for a .)

(vi) $f(x) = \begin{cases} x^2, & x < a \\ a+2, & x \geq a \end{cases}$ on $[-a-1, a+1]$.

(vii) $f(x) = \begin{cases} 0, & x \text{ irrational} \\ 1/q & x = p/q \text{ in lowest terms} \end{cases}$ on $[0, 1]$.

$$(viii) \quad f(x) = \begin{cases} 1, & x \text{ irrational} \\ 1/q & x = p/q \text{ in lowest terms} \end{cases} \text{ on } [0, 1].$$

$$(ix) \quad f(x) = \begin{cases} 1, & x \text{ irrational} \\ -1/q & x = p/q \text{ in lowest terms} \end{cases} \text{ on } [0, 1].$$

$$(x) \quad f(x) = \begin{cases} x, & x \text{ rational} \\ 0 & x \text{ irrational} \end{cases} \text{ on } [0, a].$$

$$(xi) \quad f(x) = \sin^2(\cos x + \sqrt{a + a^2}) \text{ on } [0, a^3].$$

$$(xii) \quad f(x) = [x] \text{ on } [0, a].$$

2. For each of the following polynomial functions f , find an integer n such that $f(x) = 0$ for some x between n and $n + 1$.
- (i) $f(x) = x^3 - x + 3$.
 - (ii) $f(x) = x^5 + 5x^4 + 2x + 1$.
 - (iii) $f(x) = x^5 + x + 1$.
 - (iv) $f(x) = 4x^2 - 4x + 1$.
3. Prove that there is some number x such that
- (i) $x^{179} + \frac{163}{1 + x^2 + \sin^2 x} = 119$.
 - (ii) $\sin x = x - 1$.
4. This problem is a continuation of Problem 3-7.
- (a) If $n - k$ is even, and ≥ 0 , find a polynomial function of degree n with exactly k roots.
 - (b) A root a of the polynomial function f is said to have **multiplicity m** if $f(x) = (x - a)^m g(x)$, where g is a polynomial function that does *not* have a as a root. Let f be a polynomial function of degree n . Suppose that f has k roots, counting multiplicities, i.e., suppose that k is the sum of the multiplicities of all the roots. Show that $n - k$ is even.
5. Suppose that f is continuous on $[a, b]$ and that $f(x)$ is always rational. What can be said about f ?
6. Suppose that f is a *continuous* function on $[-1, 1]$ such that $x^2 + (f(x))^2 = 1$ for all x . (This means that $(x, f(x))$ always lies on the unit circle.) Show that either $f(x) = \sqrt{1 - x^2}$ for all x , or else $f(x) = -\sqrt{1 - x^2}$ for all x .
7. How many continuous functions f are there which satisfy $(f(x))^2 = x^2$ for all x ?
8. Suppose that f and g are continuous, that $f^2 = g^2$, and that $f(x) \neq 0$ for all x . Prove that either $f(x) = g(x)$ for all x , or else $f(x) = -g(x)$ for all x .
9. (a) Suppose that f is continuous, that $f(x) = 0$ only for $x = a$, and that $f(x) > 0$ for some $x > a$ as well as for some $x < a$. What can be said about $f(x)$ for all $x \neq a$?

- (b) Again assume that f is continuous and that $f(x) = 0$ only for $x = a$, but suppose, instead, that $f(x) > 0$ for some $x > a$ and $f(x) < 0$ for some $x < a$. Now what can be said about $f(x)$ for $x \neq a$?
 *(c) Discuss the sign of $x^3 + x^2y + xy^2 + y^3$ when x and y are not both 0.
10. Suppose f and g are continuous on $[a, b]$ and that $f(a) < g(a)$, but $f(b) > g(b)$. Prove that $f(x) = g(x)$ for some x in $[a, b]$. (If your proof isn't very short, it's not the right one.)
11. Suppose that f is a continuous function on $[0, 1]$ and that $f(x)$ is in $[0, 1]$ for each x (draw a picture). Prove that $f(x) = x$ for some number x .
12. (a) Problem 11 shows that f intersects the diagonal of the square in Figure 14 (solid line). Show that f must also intersect the other (dashed) diagonal.
 (b) Prove the following more general fact: If g is continuous on $[0, 1]$ and $g(0) = 0$, $g(1) = 1$ or $g(0) = 1$, $g(1) = 0$, then $f(x) = g(x)$ for some x .
13. (a) Let $f(x) = \sin 1/x$ for $x \neq 0$ and let $f(0) = 0$. Is f continuous on $[-1, 1]$? Show that f satisfies the conclusion of the Intermediate Value Theorem on $[-1, 1]$; in other words, if f takes on two values somewhere on $[-1, 1]$, it also takes on every value in between.
 *(b) Suppose that f satisfies the conclusion of the Intermediate Value Theorem, and that f takes on each value *only once*. Prove that f is continuous.
 *(c) Generalize to the case where f takes on each value only finitely many times.
14. If f is a continuous function on $[0, 1]$, let $\|f\|$ be the maximum value of $|f|$ on $[0, 1]$.
 (a) Prove that for any number c we have $\|cf\| = |c| \cdot \|f\|$.
 *(b) Prove that $\|f + g\| \leq \|f\| + \|g\|$. Give an example where $\|f + g\| \neq \|f\| + \|g\|$.
 (c) Prove that $\|h - f\| \leq \|h - g\| + \|g - f\|$.
- *15. Suppose that ϕ is continuous and $\lim_{x \rightarrow \infty} \phi(x)/x^n = 0 = \lim_{x \rightarrow -\infty} \phi(x)/x^n$.
 (a) Prove that if n is odd, then there is a number x such that $x^n + \phi(x) = 0$.
 (b) Prove that if n is even, then there is a number y such that $y^n + \phi(y) \leq x^n + \phi(x)$ for all x .
- Hint: Of which proofs does this problem test your understanding?
- *16. Let f be any polynomial function. Prove that there is some number y such that $|f(y)| \leq |f(x)|$ for all x .
- *17. Suppose that f is a continuous function with $f(x) > 0$ for all x , and $\lim_{x \rightarrow \infty} f(x) = 0 = \lim_{x \rightarrow -\infty} f(x)$. (Draw a picture.) Prove that there is some number y such that $f(y) \geq f(x)$ for all x .

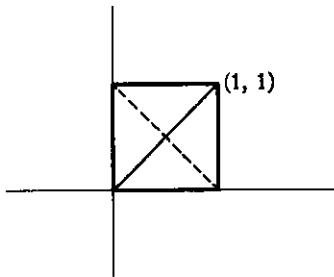


FIGURE 14

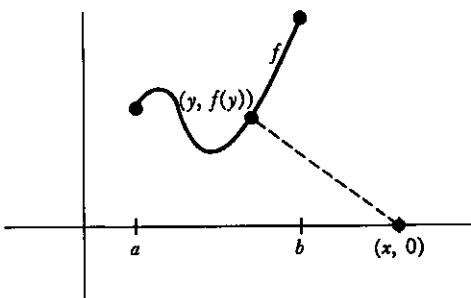


FIGURE 15

- *18. (a) Suppose that f is continuous on $[a, b]$, and let x by any number. Prove that there is a point on the graph of f which is closest to $(x, 0)$; in other words there is some y in $[a, b]$ such that the distance from $(x, 0)$ to $(y, f(y))$ is \leq distance from $(x, 0)$ to $(z, f(z))$ for all z in $[a, b]$. (See Figure 15.)
- (b) Show that this same assertion is not necessarily true if $[a, b]$ is replaced by (a, b) throughout.
- (c) Show that the assertion is true if $[a, b]$ is replaced by \mathbf{R} throughout.
- (d) In cases (a) and (c), let $g(x)$ be the minimum distance from $(x, 0)$ to a point on the graph of f . Prove that $g(y) \leq g(x) + |x - y|$, and conclude that g is continuous.
- (e) Prove that there are numbers x_0 and x_1 in $[a, b]$ such that the distance from $(x_0, 0)$ to $(x_1, f(x_1))$ is \leq the distance from $(x_0', 0)$ to $(x_1', f(x_1'))$ for any x_0', x_1' in $[a, b]$.

- **19. (a) Suppose that f is continuous on $[0, 1]$ and $f(0) = f(1)$. Let n be any natural number. Prove that there is some number x such that $f(x) = f(x + 1/n)$, as shown in Figure 16 for $n = 4$. Hint: Consider the function $g(x) = f(x) - f(x + 1/n)$; what would be true if $g(x) \neq 0$ for all x ?
- (b) Suppose $0 < a < 1$, but that a is not equal to $1/n$ for any natural number n . Find a function f which is continuous on $[0, 1]$ and which satisfies $f(0) = f(1)$, but which does not satisfy $f(x) = f(x + a)$ for any x .

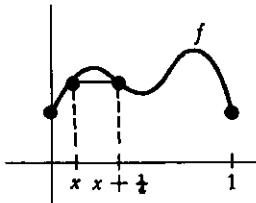


FIGURE 16

- **20. (a) Prove that there does not exist a continuous function f defined on \mathbf{R} which takes on every value exactly twice. Hint: If $f(a) = f(b)$ for $a < b$, then either $f(x) > f(a)$ for all x in (a, b) or $f(x) < f(a)$ for all x in (a, b) . Why? In the first case all values close to $f(a)$, but slightly larger than $f(a)$, are taken on somewhere in (a, b) ; this implies that $f(x) < f(a)$ for $x < a$ and $x > b$.
- (b) Refine part (a) by proving that there is no continuous function f which takes on each value either 0 times or 2 times, i.e., which takes on exactly twice each value that it does take on. Hint: The previous hint implies that f has either a maximum or a minimum value (which must be taken on twice). What can be said about values close to the maximum value?
- (c) Find a continuous function f which takes on every value exactly 3 times. More generally, find one which takes on every value exactly n times, if n is odd.
- (d) Prove that if n is even, then there is no continuous f which takes on every value exactly n times. Hint: To treat the case $n = 4$, for example, let $f(x_1) = f(x_2) = f(x_3) = f(x_4)$. Then either $f(x) > 0$ for all x in two of the three intervals (x_1, x_2) , (x_2, x_3) , (x_3, x_4) , or else $f(x) < 0$ for all x in two of these three intervals.

CHAPTER 8 LEAST UPPER BOUNDS

This chapter reveals the most important property of the real numbers. Nevertheless, it is merely a sequel to Chapter 7; the path which must be followed has already been indicated, and further discussion would be useless delay.

DEFINITION

A set A of real numbers is **bounded above** if there is a number x such that

$$x \geq a \quad \text{for every } a \text{ in } A.$$

Such a number x is called an **upper bound** for A .

Obviously A is bounded above if and only if there is a number x which is an upper bound for A (and in this case there will be lots of upper bounds for A); we often say, as a concession to idiomatic English, that “ A has an upper bound” when we mean that there is a number which is an upper bound for A .

Notice that the term “bounded above” has now been used in two ways—first, in Chapter 7, in reference to functions, and now in reference to sets. This dual usage should cause no confusion, since it will always be clear whether we are talking about a set of numbers or a function. Moreover, the two definitions are closely connected: if A is the set $\{f(x) : a \leq x \leq b\}$, then the function f is bounded above on $[a, b]$ if and only if the set A is bounded above.

The entire collection \mathbf{R} of real numbers, and the natural numbers \mathbf{N} , are both examples of sets which are *not* bounded above. An example of a set which *is* bounded above is

$$A = \{x : 0 \leq x < 1\}.$$

To show that A is bounded above we need only name some upper bound for A , which is easy enough; for example, 138 is an upper bound for A , and so are 2, $1\frac{1}{2}$, $1\frac{1}{4}$, and 1. Clearly, 1 is the least upper bound of A ; although the phrase just introduced is self-explanatory, in order to avoid any possible confusion (in particular, to ensure that we all know what the superlative of “less” means), we define this explicitly.

DEFINITION

A number x is a **least upper bound** of A if

- (1) x is an upper bound of A ,
- and (2) if y is an upper bound of A , then $x \leq y$.

The use of the indefinite article “a” in this definition was merely a concession to temporary ignorance. Now that we have made a precise definition, it is easily seen that if x and y are both least upper bounds of A , then $x = y$. Indeed, in this case

$$\begin{aligned} x \leq y, & \quad \text{since } y \text{ is an upper bound, and } x \text{ is a least upper bound,} \\ \text{and } y \leq x, & \quad \text{since } x \text{ is an upper bound, and } y \text{ is a least upper bound;} \end{aligned}$$

it follows that $x = y$. For this reason we speak of *the* least upper bound of A . The term **supremum** of A is synonymous and has one advantage. It abbreviates quite nicely to

$$\sup A \quad (\text{pronounced "soup } A\text{"})$$

and saves us from the abbreviation

$$\text{lub } A$$

(which is nevertheless used by some authors).

There is a series of important definitions, analogous to those just given, which can now be treated more briefly. A set A of real numbers is **bounded below** if there is a number x such that

$$x \leq a \quad \text{for every } a \text{ in } A.$$

Such a number x is called a **lower bound** for A . A number x is the **greatest lower bound** of A if

- (1) x is a lower bound of A ,
- and (2) if y is a lower bound of A , then $x \geq y$.

The greatest lower bound of A is also called the **infimum** of A , abbreviated

$$\inf A;$$

some authors use the abbreviation

$$\text{glb } A.$$

One detail has been omitted from our discussion so far—the question of which sets have at least one, and hence exactly one, least upper bound or greatest lower bound. We will consider only least upper bounds, since the question for greatest lower bounds can then be answered easily (Problem 2).

If A is not bounded above, then A has no upper bound at all, so A certainly cannot be expected to have a least upper bound. It is tempting to say that A does have a least upper bound if it has *some* upper bound, but, like the principle of mathematical induction, this assertion can fail to be true in a rather special way. If $A = \emptyset$, then A is bounded above. Indeed, any number x is an upper bound for \emptyset :

$$x \geq y \quad \text{for every } y \text{ in } \emptyset$$

simply because there is no y in \emptyset . Since *every* number is an upper bound for \emptyset , there is surely no least upper bound for \emptyset . With this trivial exception however,

our assertion is true—and very important, definitely important enough to warrant consideration of details. We are finally ready to state the last property of the real numbers which we need.

(P13) (The least upper bound property) If A is a set of real numbers, $A \neq \emptyset$, and A is bounded above, then A has a least upper bound.

Property P13 may strike you as anticlimactic, but that is actually one of its virtues. To complete our list of basic properties for the real numbers we require no particularly abstruse proposition, but only a property so simple that we might feel foolish for having overlooked it. Of course, the least upper bound property is not really so innocent as all that; after all, it does *not* hold for the rational numbers \mathbf{Q} . For example, if A is the set of all rational numbers x satisfying $x^2 < 2$, then there is no *rational* number y which is an upper bound for A and which is less than or equal to every other *rational* number which is an upper bound for A . It will become clear only gradually how significant P13 is, but we are already in a position to demonstrate its power, by supplying the proofs which were omitted in Chapter 7.

THEOREM 7-1

If f is continuous on $[a, b]$ and $f(a) < 0 < f(b)$, then there is some number x in $[a, b]$ such that $f(x) = 0$.

PROOF

Our proof is merely a rigorous version of the outline developed at the end of Chapter 7—we will locate the smallest number x in $[a, b]$ with $f(x) = 0$.

Define the set A , shown in Figure 1, as follows:

$$A = \{x : a \leq x \leq b, \text{ and } f \text{ is negative on the interval } [a, x]\}.$$

Clearly $A \neq \emptyset$, since a is in A ; in fact, there is some $\delta > 0$ such that A contains all points x satisfying $a \leq x < a + \delta$; this follows from Problem 6-15, since f is continuous on $[a, b]$ and $f(a) < 0$. Similarly, b is an upper bound for A and, in fact, there is a $\delta > 0$ such that all points x satisfying $b - \delta < x \leq b$ are upper bounds for A ; this also follows from Problem 6-15, since $f(b) > 0$.

From these remarks it follows that A has a least upper bound α and that $a < \alpha < b$. We now wish to show that $f(\alpha) = 0$, by eliminating the possibilities $f(\alpha) < 0$ and $f(\alpha) > 0$.

Suppose first that $f(\alpha) < 0$. By Theorem 6-3, there is a $\delta > 0$ such that $f(x) < 0$ for $\alpha - \delta < x < \alpha + \delta$ (Figure 2). Now there is some number x_0 in A which satisfies $\alpha - \delta < x_0 < \alpha$ (because otherwise α would not be the *least* upper bound of A). This means that f is negative on the whole interval $[a, x_0]$. But if x_1 is a number between α and $\alpha + \delta$, then f is also negative on the whole interval $[x_0, x_1]$. Therefore f is negative on the interval $[a, x_1]$, so x_1 is in A . But this contradicts the fact that α is an upper bound for A ; our original assumption that $f(\alpha) < 0$ must be false.

Suppose, on the other hand, that $f(\alpha) > 0$. Then there is a number $\delta > 0$ such that $f(x) > 0$ for $\alpha - \delta < x < \alpha + \delta$ (Figure 3). Once again we know that there is an x_0 in A satisfying $\alpha - \delta < x_0 < \alpha$; but this means that f is negative on $[a, x_0]$,

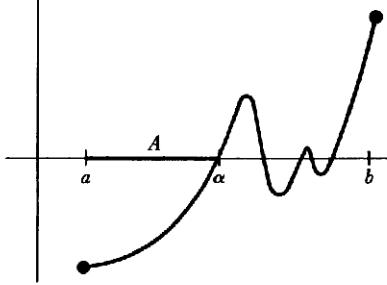


FIGURE 1

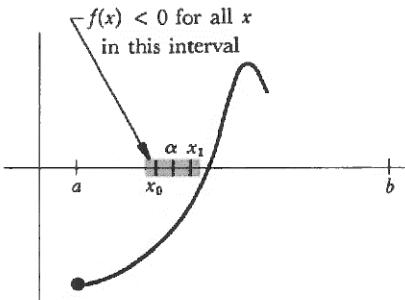


FIGURE 2

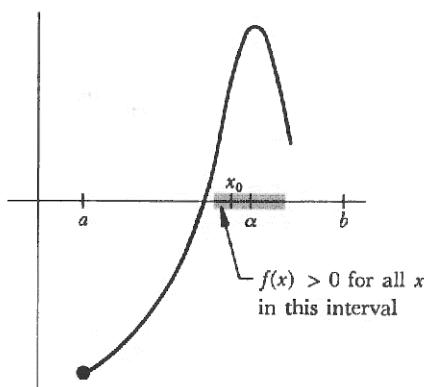


FIGURE 3

which is impossible, since $f(x_0) > 0$. Thus the assumption $f(\alpha) > 0$ also leads to a contradiction, leaving $f(\alpha) = 0$ as the only possible alternative. ■

The proofs of Theorems 2 and 3 of Chapter 7 require a simple preliminary result, which will play much the same role as Theorem 6-3 played in the previous proof.

THEOREM 1 If f is continuous at a , then there is a number $\delta > 0$ such that f is bounded above on the interval $(a - \delta, a + \delta)$ (see Figure 4).

PROOF Since $\lim_{x \rightarrow a} f(x) = f(a)$, there is, for every $\varepsilon > 0$, a $\delta > 0$ such that, for all x ,

$$\text{if } |x - a| < \delta, \text{ then } |f(x) - f(a)| < \varepsilon.$$

It is only necessary to apply this statement to some particular ε (any one will do), for example, $\varepsilon = 1$. We conclude that there is a $\delta > 0$ such that, for all x ,

$$\text{if } |x - a| < \delta, \text{ then } |f(x) - f(a)| < 1.$$

It follows, in particular, that if $|x - a| < \delta$, then $f(x) - f(a) < 1$. This completes the proof: on the interval $(a - \delta, a + \delta)$ the function f is bounded above by $f(a) + 1$. ■

It should hardly be necessary to add that we can now also prove that f is bounded below on some interval $(a - \delta, a + \delta)$, and, finally, that f is bounded on some open interval containing a .

A more significant point is the observation that if $\lim_{x \rightarrow a^+} f(x) = f(a)$, then there

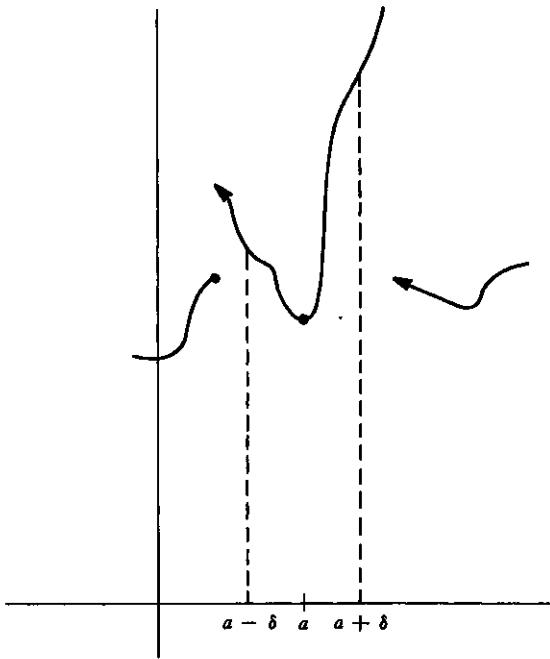


FIGURE 4

is a $\delta > 0$ such that f is bounded on the set $\{x : a \leq x < a + \delta\}$, and a similar observation holds if $\lim_{x \rightarrow b^-} f(x) = f(b)$. Having made these observations (and assuming that you will supply the proofs), we tackle our second major theorem.

THEOREM 7-2 If f is continuous on $[a, b]$, then f is bounded above on $[a, b]$.

PROOF Let

$$A = \{x : a \leq x \leq b \text{ and } f \text{ is bounded above on } [a, x]\}.$$

Clearly $A \neq \emptyset$ (since a is in A), and A is bounded above (by b), so A has a least upper bound α . Notice that we are here applying the term “bounded above” both to the set A , which can be visualized as lying on the horizontal axis, and to f , i.e., to the sets $\{f(y) : a \leq y \leq x\}$, which can be visualized as lying on the vertical axis (Figure 5).

Our first step is to prove that we actually have $\alpha = b$. Suppose, instead, that $\alpha < b$. By Theorem 1 there is $\delta > 0$ such that f is bounded on $(\alpha - \delta, \alpha + \delta)$. Since α is the least upper bound of A there is some x_0 in A satisfying $\alpha - \delta < x_0 < \alpha$. This means that f is bounded on $[a, x_0]$. But if x_1 is any number with $\alpha < x_1 < \alpha + \delta$, then f is also bounded on $[x_0, x_1]$. Therefore f is bounded on $[a, x_1]$, so x_1 is in A , contradicting the fact that α is an upper bound for A . This contradiction shows that $\alpha = b$. One detail should be mentioned: this demonstration implicitly assumed that $a < \alpha$ [so that f would be defined on some interval $(\alpha - \delta, \alpha + \delta)$]; the possibility $a = \alpha$ can be ruled out similarly, using the existence of a $\delta > 0$ such that f is bounded on $\{x : a \leq x < a + \delta\}$.

The proof is not quite complete—we only know that f is bounded on $[a, x]$ for every $x < b$, not necessarily that f is bounded on $[a, b]$. However, only one small argument needs to be added.

There is a $\delta > 0$ such that f is bounded on $\{x : b - \delta < x \leq b\}$. There is x_0 in A such that $b - \delta < x_0 < b$. Thus f is bounded on $[a, x_0]$ and also on $[x_0, b]$, so f is bounded on $[a, b]$. ■

To prove the third important theorem we resort to a trick.

THEOREM 7-3 If f is continuous on $[a, b]$, then there is a number y in $[a, b]$ such that $f(y) \geq f(x)$ for all x in $[a, b]$.

PROOF We already know that f is bounded on $[a, b]$, which means that the set

$$\{f(x) : x \text{ in } [a, b]\}$$

is bounded. This set is obviously not \emptyset , so it has a least upper bound α . Since $\alpha \geq f(x)$ for x in $[a, b]$ it suffices to show that $\alpha = f(y)$ for some y in $[a, b]$.

Suppose instead that $\alpha \neq f(y)$ for all y in $[a, b]$. Then the function g defined by

$$g(x) = \frac{1}{\alpha - f(x)}, \quad x \text{ in } [a, b]$$

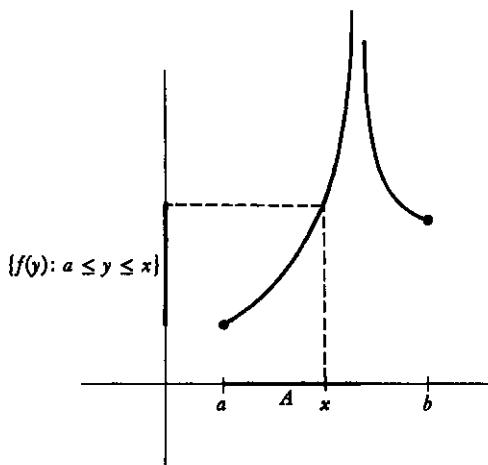


FIGURE 5

is continuous on $[a, b]$, since the denominator of the right side is never 0. On the other hand, α is the least upper bound of $\{f(x) : x \text{ in } [a, b]\}$; this means that

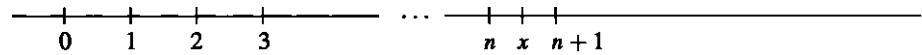
for every $\varepsilon > 0$ there is x in $[a, b]$ with $\alpha - f(x) < \varepsilon$.

This, in turn, means that

for every $\varepsilon > 0$ there is x in $[a, b]$ with $g(x) > 1/\varepsilon$.

But *this* means that g is not bounded on $[a, b]$, contradicting the previous theorem. ■

At the beginning of this chapter the set of natural numbers \mathbf{N} was given as an example of an unbounded set. We are now going to *prove* that \mathbf{N} is unbounded. After the difficult theorems proved in this chapter you may be startled to find such an “obvious” theorem winding up our proceedings. If so, you are, perhaps, allowing the geometrical picture of \mathbf{R} to influence you too strongly. “Look,” you may say, “the real numbers look like



so every number x is between two integers $n, n + 1$ (unless x is itself an integer).” Basing the argument on a geometric picture is not a proof, however, and even the geometric picture contains an assumption: that if you place unit segments end-to-end you will eventually get a segment larger than any given segment. This axiom, often omitted from a first introduction to geometry, is usually attributed (not quite justly) to Archimedes, and the corresponding property for numbers, that \mathbf{N} is not bounded, is called the *Archimedean property* of the real numbers. This property is *not* a consequence of P1–P12 (see reference [17] of the Suggested Reading), although it does hold for \mathbf{Q} , of course. Once we have P13 however, there are no longer any problems.

THEOREM 2 \mathbf{N} is not bounded above.

PROOF Suppose \mathbf{N} were bounded above. Since $\mathbf{N} \neq \emptyset$, there would be a least upper bound α for \mathbf{N} . Then

$$\alpha \geq n \quad \text{for all } n \text{ in } \mathbf{N}.$$

Consequently,

$$\alpha \geq n + 1 \quad \text{for all } n \text{ in } \mathbf{N},$$

since $n + 1$ is in \mathbf{N} if n is in \mathbf{N} . But this means that

$$\alpha - 1 \geq n \quad \text{for all } n \text{ in } \mathbf{N},$$

and *this* means that $\alpha - 1$ is also an upper bound for \mathbf{N} , contradicting the fact that α is the least upper bound. ■

There is a consequence of Theorem 2 (actually an equivalent formulation) which we have very often assumed implicitly.

THEOREM 3 For any $\varepsilon > 0$ there is a natural number n with $1/n < \varepsilon$.

PROOF Suppose not; then $1/n \geq \varepsilon$ for all n in \mathbf{N} . Thus $n \leq 1/\varepsilon$ for all n in \mathbf{N} . But this means that $1/\varepsilon$ is an upper bound for \mathbf{N} , contradicting Theorem 2. ■

A brief glance through Chapter 6 will show you that the result of Theorem 3 was used in the discussion of many examples. Of course, Theorem 3 was not available at the time, but the examples were so important that in order to give them some cheating was tolerated. As partial justification for this dishonesty we can claim that this result was never used in the proof of a *theorem*, but if your faith has been shaken, a review of all the proofs given so far is in order. Fortunately, such deception will not be necessary again. We have now stated every property of the real numbers that we will ever need. Henceforth, no more lies.

PROBLEMS

- Find the least upper bound and the greatest lower bound (if they exist) of the following sets. Also decide which sets have greatest and least elements (i.e., decide when the least upper bound and greatest lower bound happens to belong to the set).

- (i) $\left\{ \frac{1}{n} : n \text{ in } \mathbf{N} \right\}$.
- (ii) $\left\{ \frac{1}{n} : n \text{ in } \mathbf{Z} \text{ and } n \neq 0 \right\}$.
- (iii) $\{x : x = 0 \text{ or } x = 1/n \text{ for some } n \text{ in } \mathbf{N}\}$.
- (iv) $\{x : 0 \leq x \leq \sqrt{2} \text{ and } x \text{ is rational}\}$.
- (v) $\{x : x^2 + x + 1 \geq 0\}$.
- (vi) $\{x : x^2 + x - 1 < 0\}$.
- (vii) $\{x : x < 0 \text{ and } x^2 + x - 1 < 0\}$.
- (viii) $\left\{ \frac{1}{n} + (-1)^n : n \text{ in } \mathbf{N} \right\}$.

- Suppose $A \neq \emptyset$ is bounded below. Let $-A$ denote the set of all $-x$ for x in A . Prove that $-A \neq \emptyset$, that $-A$ is bounded above, and that $-\sup(-A)$ is the greatest lower bound of A .
 - If $A \neq \emptyset$ is bounded below, let B be the set of all lower bounds of A . Show that $B \neq \emptyset$, that B is bounded above, and that $\sup B$ is the greatest lower bound of A .
- Let f be a continuous function on $[a, b]$ with $f(a) < 0 < f(b)$.
 - The proof of Theorem 1 showed that there is a smallest x in $[a, b]$ with $f(x) = 0$. Is there necessarily a second smallest x in $[a, b]$ with

$f(x) = 0$? Show that there is a largest x in $[a, b]$ with $f(x) = 0$. (Try to give an easy proof by considering a new function closely related to f .)

- (b) The proof of Theorem 1 depended upon consideration of $A = \{x : a \leq x \leq b \text{ and } f \text{ is negative on } [a, x]\}$. Give another proof of Theorem 1, which depends upon consideration of $B = \{x : a \leq x \leq b \text{ and } f(x) < 0\}$. Which point x in $[a, b]$ with $f(x) = 0$ will this proof locate? Give an example where the sets A and B are not the same.

- *4. (a) Suppose that f is continuous on $[a, b]$ and that $f(a) = f(b) = 0$. Suppose also that $f(x_0) > 0$ for some x_0 in $[a, b]$. Prove that there are numbers c and d with $a \leq c < x_0 < d \leq b$ such that $f(c) = f(d) = 0$, but $f(x) > 0$ for all x in (c, d) . Hint: The previous problem can be used to good advantage.

- (b) Suppose that f is continuous on $[a, b]$ and that $f(a) < f(b)$. Prove that there are numbers c and d with $a \leq c < d \leq b$ such that $f(c) = f(d) = f(a)$ and $f(d) = f(b)$ and $f(a) < f(x) < f(d)$ for all x in (c, d) .

5. (a) Suppose that $y - x > 1$. Prove that there is an integer k such that $x < k < y$. Hint: Let l be the largest integer satisfying $l \leq x$, and consider $l + 1$.

- (b) Suppose $x < y$. Prove that there is a rational number r such that $x < r < y$. Hint: If $1/n < y - x$, then $ny - nx > 1$. (Query: Why have parts (a) and (b) been postponed until this problem set?)

- (c) Suppose that $r < s$ are rational numbers. Prove that there is an irrational number between r and s . Hint: As a start, you know that there is an irrational number between 0 and 1.

- (d) Suppose that $x < y$. Prove that there is an irrational number between x and y . Hint: It is unnecessary to do any more work; this follows from (b) and (c).

- *6. A set A of real numbers is said to be **dense** if every open interval contains a point of A . For example, Problem 5 shows that the set of rational numbers and the set of irrational numbers are each dense.

- (a) Prove that if f is continuous and $f(x) = 0$ for all numbers x in a dense set A , then $f(x) = 0$ for all x .

- (b) Prove that if f and g are continuous and $f(x) = g(x)$ for all x in a dense set A , then $f(x) = g(x)$ for all x .

- (c) If we assume instead that $f(x) \geq g(x)$ for all x in A , show that $f(x) \geq g(x)$ for all x . Can \geq be replaced by $>$ throughout?

7. Prove that if f is continuous and $f(x + y) = f(x) + f(y)$ for all x and y , then there is a number c such that $f(x) = cx$ for all x . (This conclusion can be demonstrated simply by combining the results of two previous problems.) Point of information: There *do* exist *noncontinuous* functions f satisfying $f(x + y) = f(x) + f(y)$ for all x and y , but we cannot prove this now; in fact, this simple question involves ideas that are usually never mentioned in any undergraduate course. The Suggested Reading contains references.

- *8. Suppose that f is a function such that $f(a) \leq f(b)$ whenever $a < b$ (Figure 6).

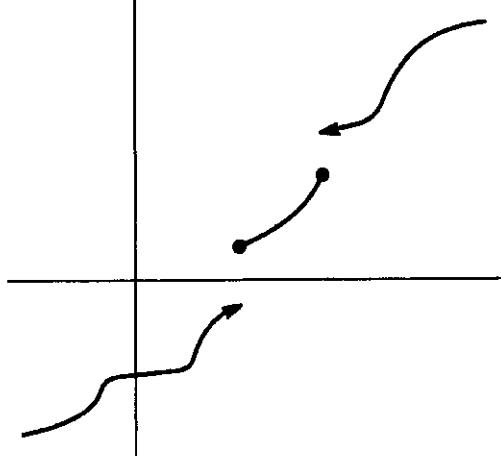


FIGURE 6

- (a) Prove that $\lim_{x \rightarrow a^-} f(x)$ and $\lim_{x \rightarrow a^+} f(x)$ both exist. Hint: Why is this problem in this chapter?

- (b) Prove that f never has a removable discontinuity (this terminology comes from Problem 6-16).

- (c) Prove that if f satisfies the conclusions of the Intermediate Value Theorem, then f is continuous.

- *9. If f is a bounded function on $[0, 1]$, let $\|f\| = \sup\{|f(x)| : x \text{ in } [0, 1]\}$. Prove analogues of the properties of $\|\cdot\|$ in Problem 7-14.

10. Suppose $\alpha > 0$. Prove that every number x can be written uniquely in the form $x = k\alpha + x'$, where k is an integer, and $0 \leq x' < \alpha$.

11. (a) Suppose that a_1, a_2, a_3, \dots is a sequence of positive numbers with $a_{n+1} \leq a_n/2$. Prove that for any $\varepsilon > 0$ there is some n with $a_n < \varepsilon$.

- (b) Suppose P is a regular polygon inscribed inside a circle. If P' is the inscribed regular polygon with twice as many sides, show that the difference between the area of the circle and the area of P' is less than half the difference between the area of the circle and the area of P (use Figure 7).

- (c) Prove that there is a regular polygon P inscribed in a circle with area as close as desired to the area of the circle. In order to do part (c) you will need part (a). This was clear to the Greeks, who used part (a) as the basis for their entire treatment of proportion and area. By calculating the areas of polygons, this method ("the method of exhaustion") allows computations of π to any desired accuracy; Archimedes used it to show that $\frac{223}{71} < \pi < \frac{22}{7}$. But it has far greater theoretical importance:

- (d) Using the fact that the areas of two regular polygons with the same number of sides have the same ratio as the square of their sides, prove that the areas of two circles have the same ratios as the square of their radii. Hint: Deduce a contradiction from the assumption that the ratio of the areas is greater, or less, than the ratio of the square of the radii by inscribing appropriate polygons.

12. Suppose that A and B are two nonempty sets of numbers such that $x \leq y$ for all x in A and all y in B .

- (a) Prove that $\sup A \leq y$ for all y in B .

- (b) Prove that $\sup A \leq \inf B$.

13. Let A and B be two nonempty sets of numbers which are bounded above, and let $A+B$ denote the set of all numbers $x+y$ with x in A and y in B . Prove that $\sup(A+B) = \sup A + \sup B$. Hint: The inequality $\sup(A+B) \leq \sup A + \sup B$ is easy. Why? To prove that $\sup A + \sup B \leq \sup(A+B)$ it suffices to prove that $\sup A + \sup B \leq \sup(A+B) + \varepsilon$ for all $\varepsilon > 0$; begin by choosing x in A

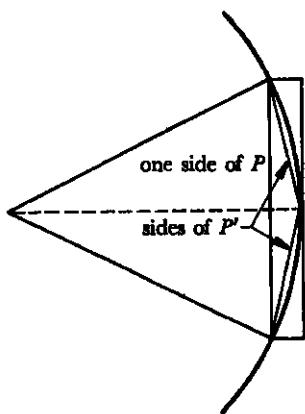
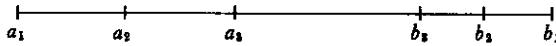


FIGURE 7

and y in B with $\sup A - x < \varepsilon/2$ and $\sup B - y < \varepsilon/2$.

FIGURE 8 

14. (a) Consider a sequence of closed intervals $I_1 = [a_1, b_1], I_2 = [a_2, b_2], \dots$. Suppose that $a_n \leq a_{n+1}$ and $b_{n+1} \leq b_n$ for all n (Figure 8). Prove that there is a point x which is in every I_n .
 (b) Show that this conclusion is false if we consider open intervals instead of closed intervals.

The simple result of Problem 14(a) is called the “Nested Interval Theorem.” It may be used to give alternative proofs of Theorems 1 and 2. The appropriate reasoning, outlined in the next two problems, illustrates a general method, called a “bisection argument.”

- *15. Suppose f is continuous on $[a, b]$ and $f(a) < 0 < f(b)$. Then either $f((a+b)/2) = 0$, or f has different signs at the end points of the interval $[a, (a+b)/2]$, or f has different signs at the end points of $[(a+b)/2, b]$. Why? If $f((a+b)/2) \neq 0$, let I_1 be one of the two intervals on which f changes sign. Now bisect I_1 . Either f is 0 at the midpoint, or f changes sign on one of the two intervals. Let I_2 be such an interval. Continue in this way, to define I_n for each n (unless f is 0 at some midpoint). Use the Nested Interval Theorem to find a point x where $f(x) = 0$.
- *16. Suppose f were continuous on $[a, b]$, but not bounded on $[a, b]$. Then f would be unbounded on either $[a, (a+b)/2]$ or $[(a+b)/2, b]$. Why? Let I_1 be one of these intervals on which f is unbounded. Proceed as in Problem 15 to obtain a contradiction.
17. (a) Let $A = \{x : x < \alpha\}$. Prove the following (they are all easy):
 (i) If x is in A and $y < x$, then y is in A .
 (ii) $A \neq \emptyset$.
 (iii) $A \neq \mathbf{R}$.
 (iv) If x is in A , then there is some number x' in A such that $x < x'$.
 (b) Suppose, conversely, that A satisfies (i)–(iv). Prove that $A = \{x : x < \sup A\}$.
- *18. A number x is called an **almost upper bound** for A if there are only finitely many numbers y in A with $y \geq x$. An **almost lower bound** is defined similarly.
 (a) Find all almost upper bounds and almost lower bounds of the sets in Problem 1.
 (b) Suppose that A is a bounded infinite set. Prove that the set B of all almost upper bounds of A is nonempty, and bounded below.

- (c) It follows from part (b) that $\inf B$ exists; this number is called the **limit superior** of A , and denoted by $\overline{\lim} A$ or $\limsup A$. Find $\overline{\lim} A$ for each set A in Problem 1.
- (d) Define $\underline{\lim} A$, and find it for all A in Problem 1.
- *19. If A is a bounded infinite set prove
- $\underline{\lim} A \leq \overline{\lim} A$.
 - $\overline{\lim} A \leq \sup A$.
 - If $\overline{\lim} A < \sup A$, then A contains a largest element.
 - The analogues of parts (b) and (c) for $\underline{\lim}$.

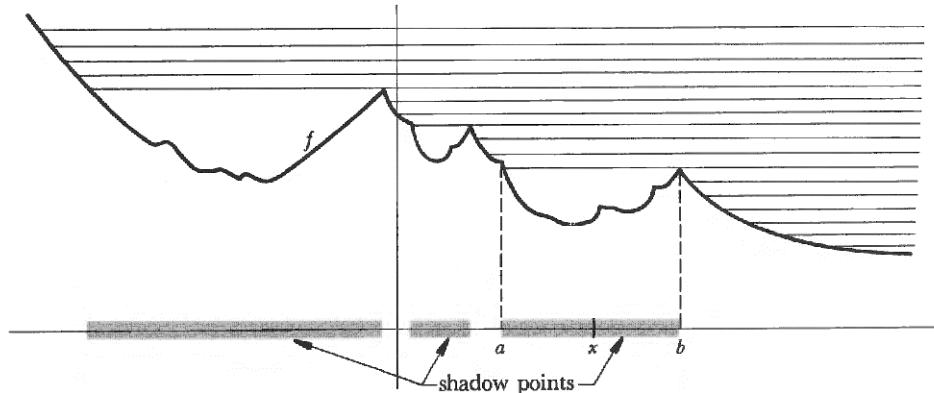


FIGURE 9

- *20. Let f be a continuous function on \mathbf{R} . A point x is called a **shadow point** of f if there is a number $y > x$ with $f(y) > f(x)$. The rationale for this terminology is indicated in Figure 9; the parallel lines are the rays of the sun rising in the east (you are facing north). Suppose that all points of (a, b) are shadow points, but that a and b are not shadow points.
- For x in (a, b) , prove that $f(x) \leq f(b)$. Hint: Let $A = \{y : x \leq y \leq b \text{ and } f(x) \leq f(y)\}$. If $\sup A$ were less than b , then $\sup A$ would be a shadow point. Use this fact to obtain a contradiction to the fact that b is not a shadow point.
 - Now prove that $f(a) \leq f(b)$. (This is a simple consequence of continuity.)
 - Finally, using the fact that a is not a shadow point, prove that $f(a) = f(b)$.

This result is known as the **Rising Sun Lemma**. Aside from serving as a good illustration of the use of least upper bounds, it is instrumental in proving several beautiful theorems that do not appear in this book; see page 443.

APPENDIX. UNIFORM CONTINUITY

Now that we've come to the end of the "foundations," it might be appropriate to slip in one further fundamental concept. This notion is not used crucially in the rest of the book, but it can help clarify many points later on.

We know that the function $f(x) = x^2$ is continuous at a for all a . In other words,

if a is any number, then for every $\varepsilon > 0$ there is some $\delta > 0$ such that, for all x , if $|x - a| < \delta$, then $|x^2 - a^2| < \varepsilon$.

Of course, δ depends on ε . But δ also depends on a —the δ that works at a might not work at b (Figure 1). Indeed, it's clear that given $\varepsilon > 0$ there is no one $\delta > 0$ that works for all a , or even for all positive a . In fact, the number $a + \delta/2$ will certainly satisfy $|x - a| < \delta$, but if $a > 0$, then

$$\left| \left(a + \frac{\delta}{2} \right)^2 - a^2 \right| = \left| a\delta + \frac{\delta^2}{4} \right| \geq a\delta,$$

and this won't be $< \varepsilon$ once $a > \varepsilon/\delta$. (This is just an admittedly confusing computational way of saying that f is growing faster and faster!)

On the other hand, for any $\varepsilon > 0$ there *will* be one $\delta > 0$ that works for all a in any interval $[-N, N]$. In fact, the δ which works at N or $-N$ will also work everywhere else in the interval.

As a final example, consider the function $f(x) = \sin 1/x$, or the function whose graph appears in Figure 18 on page 62. It is easy to see that, so long as $\varepsilon < 1$, there will not be one $\delta > 0$ that works for these functions at all points a in the open interval $(0, 1)$.

These examples illustrate important distinctions between the behavior of various continuous functions on certain intervals, and there is a special term to signal this distinction.

DEFINITION

The function f is **uniformly continuous on an interval A** if for every $\varepsilon > 0$ there is some $\delta > 0$ such that, for all x and y in A ,

$$\text{if } |x - y| < \delta, \text{ then } |f(x) - f(y)| < \varepsilon.$$

We've seen that a function can be continuous on the whole line, or on an open interval, without being uniformly continuous there. On the other hand, the function $f(x) = x^2$ did turn out to be uniformly continuous on any closed interval. This shouldn't be too surprising—it's the same sort of thing that occurs when we ask whether a function is bounded on an interval—and we would be led to suspect that any continuous function on a closed interval is also uniformly continuous on that interval. In order to prove this, we'll need to deal first with one subtle point.

Suppose that we have two intervals $[a, b]$ and $[b, c]$ with the common endpoint b , and a function f that is continuous on $[a, c]$. Let $\varepsilon > 0$ and suppose that

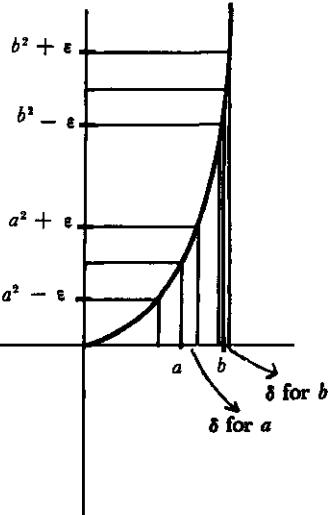


FIGURE 1

the following two statements hold:

- (i) if x and y are in $[a, b]$ and $|x - y| < \delta_1$, then $|f(x) - f(y)| < \varepsilon$,
- (ii) if x and y are in $[b, c]$ and $|x - y| < \delta_2$, then $|f(x) - f(y)| < \varepsilon$.

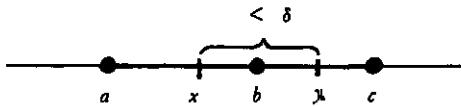


FIGURE 2

We'd like to know if there is some $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon$ whenever x and y are points in $[a, c]$ with $|x - y| < \delta$. Our first inclination might be to choose δ as the minimum of δ_1 and δ_2 . But it is easy to see what goes wrong (Figure 2): we might have x in $[a, b]$ and y in $[b, c]$, and then neither (i) nor (ii) tells us anything about $|f(x) - f(y)|$. So we have to be a little more cagey, and also use continuity of f at b .

LEMMA Let $a < b < c$ and let f be continuous on the interval $[a, c]$. Let $\varepsilon > 0$, and suppose that statements (i) and (ii) hold. Then there is a $\delta > 0$ such that,

$$\text{if } x \text{ and } y \text{ are in } [a, c] \text{ and } |x - y| < \delta, \text{ then } |f(x) - f(y)| < \varepsilon.$$

PROOF Since f is continuous at b , there is a $\delta_3 > 0$ such that,

$$\text{if } |x - b| < \delta_3, \text{ then } |f(x) - f(b)| < \frac{\varepsilon}{2}.$$

It follows that

$$(iii) \quad \text{if } |x - b| < \delta_3 \text{ and } |y - b| < \delta_3, \text{ then } |f(x) - f(y)| < \varepsilon.$$

Choose δ to be the minimum of δ_1 , δ_2 , and δ_3 . We claim that this δ works. In fact, suppose that x and y are any two points in $[a, c]$ with $|x - y| < \delta$. If x and y are both in $[a, b]$, then $|f(x) - f(y)| < \varepsilon$ by (i); and if x and y are both in $[b, c]$, then $|f(x) - f(y)| < \varepsilon$ by (ii). The only other possibility is that

$$x < b < y \quad \text{or} \quad y < b < x.$$

In either case, since $|x - y| < \delta$, we also have $|x - b| < \delta$ and $|y - b| < \delta$. So $|f(x) - f(y)| < \varepsilon$ by (iii). ■

THEOREM 1 If f is continuous on $[a, b]$, then f is uniformly continuous on $[a, b]$.

PROOF It's the usual trick, but we've got to be a little bit careful about the mechanism of the proof. For $\varepsilon > 0$ let's say that f is ε -good on $[a, b]$ if there is some $\delta > 0$ such that, for all y and z in $[a, b]$,

$$\text{if } |y - z| < \delta, \text{ then } |f(y) - f(z)| < \varepsilon.$$

Then we're trying to prove that f is ε -good on $[a, b]$ for all $\varepsilon > 0$.

Consider any particular $\varepsilon > 0$. Let

$$A = \{x : a \leq x \leq b \text{ and } f \text{ is } \varepsilon\text{-good on } [a, x]\}.$$

Then $A \neq \emptyset$ (since a is in A), and A is bounded above (by b), so A has a least upper bound α . We really should write α_ε , since A and α might depend on ε . But we won't since we intend to prove that $\alpha = b$, no matter what ε is.

Suppose that we had $\alpha < b$. Since f is continuous at α , there is some $\delta_0 > 0$ such that, if $|y - \alpha| < \delta_0$, then $|f(y) - f(\alpha)| < \varepsilon/2$. Consequently, if $|y - \alpha| < \delta_0$ and $|z - \alpha| < \delta_0$, then $|f(y) - f(z)| < \varepsilon$. So f is surely ε -good on the interval $[\alpha - \delta_0, \alpha + \delta_0]$. On the other hand, since α is the least upper bound of A , it is also clear that f is ε -good on $[a, \alpha - \delta_0]$. Then the Lemma implies that f is ε -good on $[a, a + \delta_0]$, so $a + \delta_0$ is in A , contradicting the fact that α is an upper bound.

To complete the proof we just have to show that $\alpha = b$ is actually in A . The argument for this is practically the same: Since f is continuous at b , there is some $\delta_0 > 0$ such that, if $|b - y| < \delta_0$, then $|f(y) - f(b)| < \varepsilon/2$. So f is ε -good on $[b - \delta_0, b]$. But f is also ε -good on $[a, b - \delta_0]$, so the Lemma implies that f is ε -good on $[a, b]$. ■

PROBLEMS

1. (a) For which of the following values of α is the function $f(x) = x^\alpha$ uniformly continuous on $[0, \infty)$: $\alpha = 1/3, 1/2, 2, 3$?
 (b) Find a function f that is continuous and bounded on $(0, 1]$, but not uniformly continuous on $(0, 1]$.
 (c) Find a function f that is continuous and bounded on $[0, \infty)$ but which is not uniformly continuous on $[0, \infty)$.
2. (a) Prove that if f and g are uniformly continuous on A , then so is $f + g$.
 (b) Prove that if f and g are uniformly continuous and bounded on A , then fg is uniformly continuous on A .
 (c) Show that this conclusion does not hold if one of them isn't bounded.
 (d) Suppose that f is uniformly continuous on A , that g is uniformly continuous on B , and that $f(x)$ is in B for all x in A . Prove that $g \circ f$ is uniformly continuous on A .
3. Use a “bisection argument” (page 140) to give another proof of Theorem 1.
4. Derive Theorem 7-2 as a consequence of Theorem 1.

PART 3

DERIVATIVES AND INTEGRALS

*In 1604, at the height of
his scientific career, Galileo argued
that for a rectilinear motion
in which speed increases proportionally
to distance covered,
the law of motion should be
just that ($x = ct^2$)
which he had discovered
in the investigation of falling bodies.
Between 1695 and 1700
not a single one of the monthly issues
of Leipzig's *Acta Eruditorum* was published
without articles of Leibniz,
the Bernoulli brothers
or the Marquis de l'Hôpital treating,
with notation only slightly different from
that which we use today,
the most varied problems of
differential calculus, integral calculus
and the calculus of variations.
Thus in the space of almost precisely
one century
infinitesimal calculus or,
as we now call it in English,
The Calculus,
the calculating tool *par excellence*,
had been forged;
and nearly three centuries of
constant use have not completely dulled
this incomparable instrument.*

NICHOLAS BOURBAKI

CHAPTER 9 DERIVATIVES

9

DERIVATIVES

The derivative of a function is the first of the two major concepts of this section. Together with the integral, it constitutes the source from which calculus derives its particular flavor. While it is true that the concept of a function is fundamental, that you cannot do anything without limits or continuity, and that least upper bounds are essential, everything we have done until now has been preparation—if adequate, this section will be easier than the preceding ones—for the really exciting ideas to come, the powerful concepts that are truly characteristic of calculus.

Perhaps (some would say “certainly”) the interest of the ideas to be introduced in this section stems from the intimate connection between the mathematical concepts and certain physical ideas. Many definitions, and even some theorems, may be described in terms of physical problems, often in a revealing way. In fact, the demands of physics were the original inspiration for these fundamental ideas of calculus, and we shall frequently mention the physical interpretations. But we shall always first define the ideas in precise mathematical form, and discuss their significance in terms of mathematical problems.

The collection of all functions exhibits such diversity that there is almost no hope of discovering any interesting general properties pertaining to all. Because continuous functions form such a restricted class, we might expect to find some nontrivial theorems pertaining to them, and the sudden abundance of theorems after Chapter 6 shows that this expectation is justified. But the most interesting and most powerful results about functions will be obtained only when we restrict our attention even further, to functions which have even greater claim to be called “reasonable,” which are even better behaved than most continuous functions.

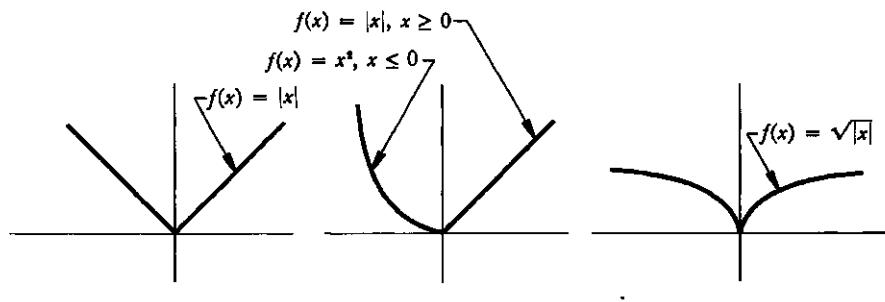


FIGURE 1 (a) (b) (c)

Figure 1 illustrates certain types of misbehavior which continuous functions can display. The graphs of these functions are “bent” at $(0, 0)$, unlike the graph of Figure 2, where it is possible to draw a “tangent line” at each point. The quotation marks have been used to avoid the suggestion that we have defined “bent” or

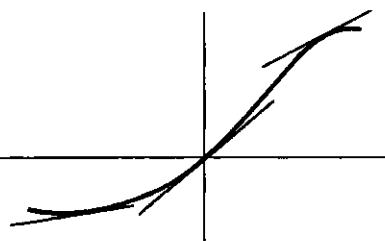


FIGURE 2

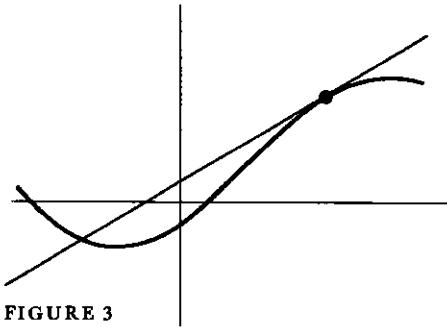


FIGURE 3

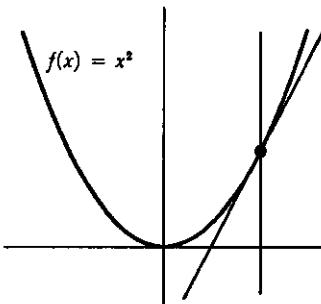


FIGURE 4

"tangent line," although we are suggesting that the graph might be "bent" at a point where a "tangent line" cannot be drawn. You have probably already noticed that a tangent line cannot be defined as a line which intersects the graph only once—such a definition would be both too restrictive and too permissive. With such a definition, the straight line shown in Figure 3 would not be a tangent line to the graph in that picture, while the parabola would have two tangent lines at each point (Figure 4), and the three functions in Figure 5 would have more than one tangent line at the points where they are "bent."

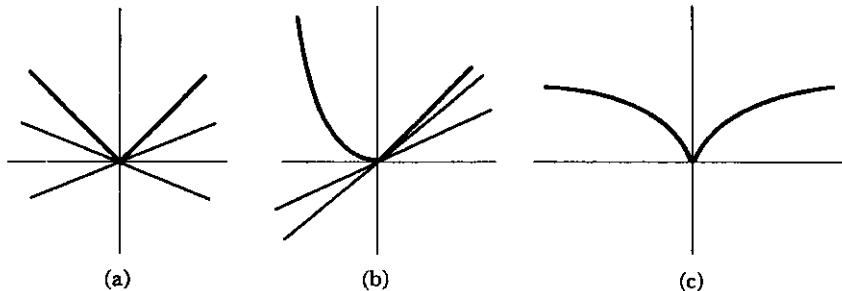


FIGURE 5

A more promising approach to the definition of a tangent line might start with "secant lines," and use the notion of limits. If $h \neq 0$, then the two distinct points $(a, f(a))$ and $(a + h, f(a + h))$ determine, as in Figure 6, a straight line whose slope is

$$\frac{f(a + h) - f(a)}{h}.$$

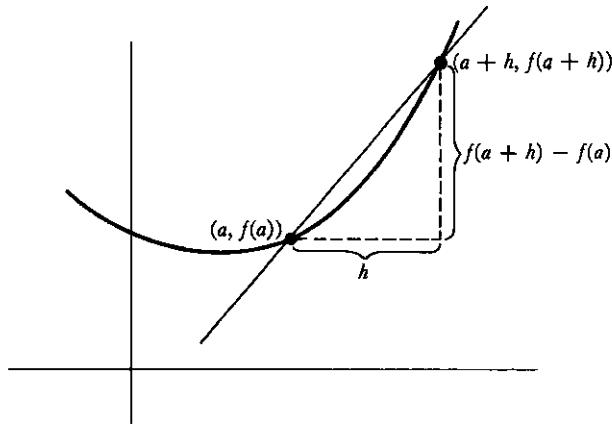


FIGURE 6

As Figure 7 illustrates, the "tangent line" at $(a, f(a))$ seems to be the limit, in some sense, of these "secant lines," as h approaches 0. We have never before talked about a "limit" of lines, but we *can* talk about the limit of their slopes: the

slope of the tangent line through $(a, f(a))$ should be

$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}.$$

We are ready for a definition, and some comments.

DEFINITION

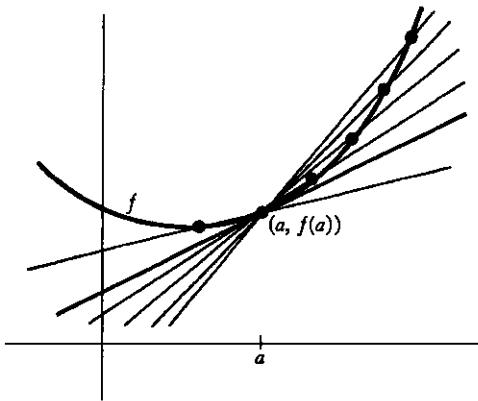


FIGURE 7

The function f is **differentiable at a** if

$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \text{ exists.}$$

In this case the limit is denoted by $f'(a)$ and is called the **derivative of f at a** . (We also say that f is **differentiable** if f is differentiable at a for every a in the domain of f .)

The first comment on our definition is really an addendum; we define the **tangent line** to the graph of f at $(a, f(a))$ to be the line through $(a, f(a))$ with slope $f'(a)$. This means that the tangent line at $(a, f(a))$ is defined only if f is differentiable at a .

The second comment refers to notation. The symbol $f'(a)$ is certainly reminiscent of functional notation. In fact, for any function f , we denote by f' the function whose domain is the set of all numbers a such that f is differentiable at a , and whose value at such a number a is

$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}.$$

(To be very precise: f' is the collection of all pairs

$$\left(a, \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \right)$$

for which $\lim_{h \rightarrow 0} [f(a+h) - f(a)]/h$ exists.) The function f' is called the **derivative** of f .

Our third comment, somewhat longer than the previous two, refers to the physical interpretation of the derivative. Consider a particle which is moving along a straight line (Figure 8(a)) on which we have chosen an “origin” point O , and a direction in which distances from O shall be written as positive numbers, the distance from O of points in the other direction being written as negative numbers. Let $s(t)$ denote the distance of the particle from O , at time t . The suggestive notation $s(t)$ has been chosen purposely; since a distance $s(t)$ is determined for each

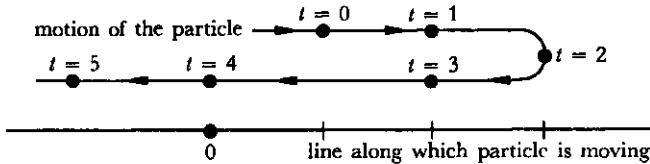


FIGURE 8(a)

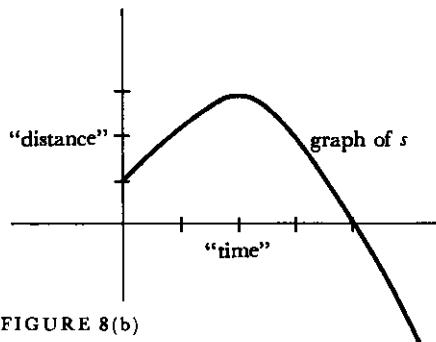


FIGURE 8(b)

number t , the physical situation automatically supplies us with a certain function s . The graph of s indicates the distance of the particle from O , on the vertical axis, in terms of the time, indicated on the horizontal axis (Figure 8(b)).

The quotient

$$\frac{s(a+h) - s(a)}{h}$$

has a natural physical interpretation. It is the “average velocity” of the particle during the time interval from a to $a + h$. For any particular a , this average speed depends on h , of course. On the other hand, the limit

$$\lim_{h \rightarrow 0} \frac{s(a+h) - s(a)}{h}$$

depends only on a (as well as the particular function s) and there are important physical reasons for considering this limit. We would like to speak of the “velocity of the particle at time a ,” but the usual definition of velocity is really a definition of average velocity; the only reasonable definition of “velocity at time a ” (so-called “instantaneous velocity”) is the limit

$$\lim_{h \rightarrow 0} \frac{s(a+h) - s(a)}{h}$$

Thus we *define* the (**instantaneous**) **velocity** of the particle at a to be $s'(a)$. Notice that $s'(a)$ could easily be negative; the absolute value $|s'(a)|$ is sometimes called the (**instantaneous**) **speed**.

It is important to realize that instantaneous velocity is a theoretical concept, an abstraction which does not correspond precisely to any observable quantity. While it would not be fair to say that instantaneous velocity has nothing to do with average velocity, remember that $s'(t)$ is not

$$\frac{s(t+h) - s(t)}{h}$$

for any particular h , but merely the limit of these average velocities as h approaches 0. Thus, when velocities are measured in physics, what a physicist really measures is an average velocity over some (very small) time interval; such a procedure cannot be expected to give an exact answer, but this is really no defect, because physical measurements can never be exact anyway.

The velocity of a particle is often called the “rate of change of its position.” This notion of the derivative, as a rate of change, applies to any other physical situation in which some quantity varies with time. For example, the “rate of change of mass” of a growing object means the derivative of the function m , where $m(t)$ is the mass at time t .

In order to become familiar with the basic definitions of this chapter, we will spend quite some time examining the derivatives of particular functions. Before proving the important theoretical results of Chapter 11, we want to have a good idea of what the derivative of a function looks like. The next chapter is devoted exclusively to one aspect of this problem—calculating the derivative of complicated functions. In this chapter we will emphasize the concepts, rather than the

calculations, by considering a few simple examples. Simplest of all is a constant function, $f(x) = c$. In this case

$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} = \lim_{h \rightarrow 0} \frac{c - c}{h} = 0.$$

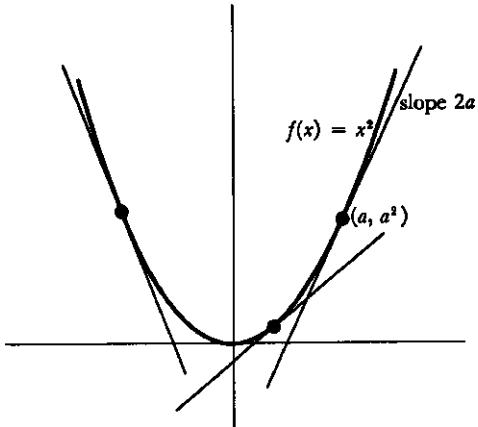
Thus f is differentiable at a for every number a , and $f'(a) = 0$. This means that the tangent line to the graph of f always has slope 0, so the tangent line always coincides with the graph.

Constant functions are not the only ones whose graphs coincide with their tangent lines—this happens for any linear function $f(x) = cx + d$. Indeed

$$\begin{aligned} f'(a) &= \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \\ &= \lim_{h \rightarrow 0} \frac{c(a+h) + d - [ca + d]}{h} \\ &= \lim_{h \rightarrow 0} \frac{ch}{h} = c; \end{aligned}$$

the slope of the tangent line is c , the same as the slope of the graph of f .

A refreshing difference occurs for $f(x) = x^2$. Here



$$\begin{aligned} f'(a) &= \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \\ &= \lim_{h \rightarrow 0} \frac{(a+h)^2 - a^2}{h} \\ &= \lim_{h \rightarrow 0} \frac{a^2 + 2ah + h^2 - a^2}{h} \\ &= \lim_{h \rightarrow 0} \frac{2ah + h^2}{h} \\ &= 2a. \end{aligned}$$

Some of the tangent lines to the graph of f are shown in Figure 9. In this picture each tangent line appears to intersect the graph only once, and this fact can be checked fairly easily: Since the tangent line through (a, a^2) has slope $2a$, it is the graph of the function

$$\begin{aligned} g(x) &= 2a(x - a) + a^2 \\ &= 2ax - a^2. \end{aligned}$$

Now, if the graphs of f and g intersect at a point $(x, f(x)) = (x, g(x))$, then

$$\begin{aligned} x^2 &= 2ax - a^2 \\ \text{or } x^2 - 2ax + a^2 &= 0; \\ \text{so } (x - a)^2 &= 0 \\ \text{or } x &= a. \end{aligned}$$

In other words, (a, a^2) is the only point of intersection.

FIGURE 9

The function $f(x) = x^2$ happens to be quite special in this regard; usually a tangent line will intersect the graph more than once. Consider, for example, the function $f(x) = x^3$. In this case

$$\begin{aligned} f'(a) &= \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \\ &= \lim_{h \rightarrow 0} \frac{(a+h)^3 - a^3}{h} \\ &= \lim_{h \rightarrow 0} \frac{a^3 + 3a^2h + 3ah^2 + h^3 - a^3}{h} \\ &= \lim_{h \rightarrow 0} \frac{3a^2h + 3ah^2 + h^3}{h} \\ &= \lim_{h \rightarrow 0} 3a^2 + 3ah + h^2 \\ &= 3a^2. \end{aligned}$$

Thus the tangent line to the graph of f at (a, a^3) has slope $3a^2$. This means that the tangent line is the graph of

$$\begin{aligned} g(x) &= 3a^2(x - a) + a^3 \\ &= 3a^2x - 2a^3. \end{aligned}$$

The graphs of f and g intersect at the point $(x, f(x)) = (x, g(x))$ when

$$\begin{aligned} x^3 &= 3a^2x - 2a^3 \\ \text{or } x^3 - 3a^2x + 2a^3 &= 0. \end{aligned}$$

This equation is easily solved if we remember that one solution of the equation has got to be $x = a$, so that $(x - a)$ is a factor of the left side; the other factor can then be found by dividing. We obtain

$$(x - a)(x^2 + ax - 2a^2) = 0.$$

It so happens that $x^2 + ax - 2a^2$ also has $x - a$ as a factor; we obtain finally

$$(x - a)(x - a)(x + 2a) = 0.$$

Thus, as illustrated in Figure 10, the tangent line through (a, a^3) also intersects the graph at the point $(-2a, -8a^3)$. These two points are always distinct, except when $a = 0$.

We have already found the derivative of sufficiently many functions to illustrate the classical, and still very popular, notation for derivatives. For a given function f , the derivative f' is often denoted by

$$\frac{df(x)}{dx}.$$

For example, the symbol

$$\frac{dx^2}{dx}$$

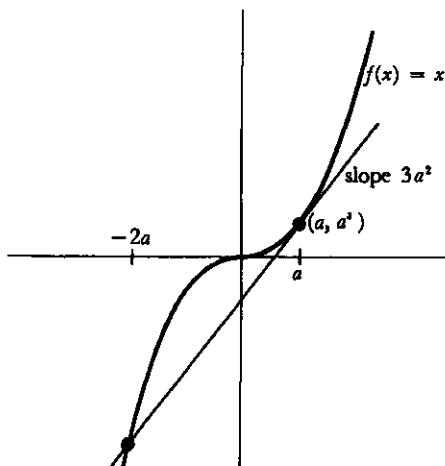


FIGURE 10

denotes the derivative of the function $f(x) = x^2$. Needless to say, the separate parts of the expression

$$\frac{df(x)}{dx}$$

are not supposed to have any sort of independent existence—the d 's are *not* numbers, they *cannot* be canceled, and the entire expression is *not* the quotient of two other numbers “ $df(x)$ ” and “ dx .” This notation is due to Leibniz (generally considered an independent co-discoverer of calculus, along with Newton), and is affectionately referred to as Leibnizian notation.* Although the notation $df(x)/dx$ seems very complicated, in concrete cases it may be shorter; after all, the symbol dx^2/dx is actually more concise than the phrase “the derivative of the function $f(x) = x^2$.”

The following formulas state in standard Leibnizian notation all the information that we have found so far:

$$\begin{aligned}\frac{dc}{dx} &= 0, \\ \frac{d(ax+b)}{dx} &= a, \\ \frac{dx^2}{dx} &= 2x, \\ \frac{dx^3}{dx} &= 3x^2.\end{aligned}$$

Although the meaning of these formulas is clear enough, attempts at literal interpretation are hindered by the reasonable stricture that an equation should not contain a function on one side and a number on the other. For example, if the third equation is to be true, then either $df(x)/dx$ must denote $f'(x)$, rather than f' , or else $2x$ must denote, not a number, but the function whose value at x is $2x$. It is really impossible to assert that one or the other of these alternatives is intended; in practice $df(x)/dx$ sometimes means f' and sometimes means $f'(x)$, while $2x$ may denote either a number or a function. Because of this ambiguity, most authors are reluctant to denote $f'(a)$ by

$$\frac{df(x)}{dx}(a);$$

instead $f'(a)$ is usually denoted by the barbaric, but unambiguous, symbol

$$\left. \frac{df(x)}{dx} \right|_{x=a}$$

*Leibniz was led to this symbol by his intuitive notion of the derivative, which he considered to be, not the limit of quotients $[f(x+h) - f(x)]/h$, but the “value” of this quotient when h is an “infinitely small” number. This “infinitely small” quantity was denoted by dx and the corresponding “infinitely small” difference $f(x+dx) - f(x)$ by $df(x)$. Although this point of view is impossible to reconcile with properties (P1)–(P13) of the real numbers, some people find this notion of the derivative congenial.

In addition to these difficulties, Leibnizian notation is associated with one more ambiguity. Although the notation dx^2/dx is absolutely standard, the notation $df(x)/dx$ is often replaced by df/dx . This, of course, is in conformity with the practice of confusing a function with its value at x . So strong is this tendency that functions are often indicated by a phrase like the following: “consider the function $y = x^2$.” We will sometimes follow classical practice to the extent of using y as the name of a function, but we will nevertheless carefully distinguish between the function and its values—thus we will always say something like “consider the function (defined by) $y(x) = x^2$.”

Despite the many ambiguities of Leibnizian notation, it is used almost exclusively in older mathematical writing, and is still used very frequently today. The staunchest opponents of Leibnizian notation admit that it will be around for quite some time, while its most ardent admirers would say that it will be around forever, and a good thing too! In any case, Leibnizian notation cannot be ignored completely.

The policy adopted in this book is to disallow Leibnizian notation within the text, but to include it in the Problems; several chapters contain a few (immediately recognizable) problems which are expressly designed to illustrate the vagaries of Leibnizian notation. Trusting that these problems will provide ample practice in this notation, we return to our basic task of examining some simple examples of derivatives.

The few functions examined so far have all been differentiable. To fully appreciate the significance of the derivative it is equally important to know some examples of functions which are *not* differentiable. The obvious candidates are the three functions first discussed in this chapter, and illustrated in Figure 1; if they turn out to be differentiable at 0 something has clearly gone wrong.

Consider first $f(x) = |x|$. In this case

$$\frac{f(0+h) - f(0)}{h} = \frac{|h|}{h}.$$

Now $|h|/h = 1$ for $h > 0$, and $|h|/h = -1$ for $h < 0$. This shows that

$$\lim_{h \rightarrow 0} \frac{f(h) - f(0)}{h} \text{ does not exist.}$$

In fact,

$$\begin{aligned} \lim_{h \rightarrow 0^+} \frac{f(h) - f(0)}{h} &= 1 \\ \text{and } \lim_{h \rightarrow 0^-} \frac{f(h) - f(0)}{h} &= -1. \end{aligned}$$

(These two limits are sometimes called the **right-hand derivative** and the **left-hand derivative**, respectively, of f at 0.)

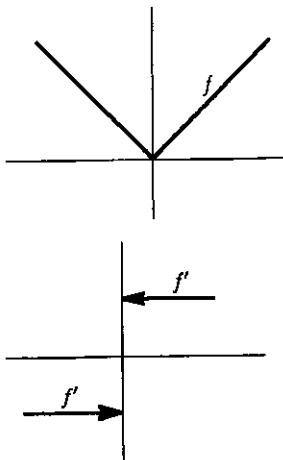


FIGURE 11

If $a \neq 0$, then $f'(a)$ does exist. In fact,

$$\begin{aligned} f'(x) &= 1 && \text{if } x > 0, \\ f'(x) &= -1 && \text{if } x < 0. \end{aligned}$$

The proof of this fact is left to you (it is easy if you remember the derivative of a linear function). The graphs of f and of f' are shown in Figure 11.

For the function

$$f(x) = \begin{cases} x^2, & x \leq 0 \\ x, & x \geq 0, \end{cases}$$

a similar difficulty arises in connection with $f'(0)$. We have

$$\frac{f(h) - f(0)}{h} = \begin{cases} \frac{h^2}{h} = h, & h < 0 \\ \frac{h}{h} = 1, & h > 0. \end{cases}$$

Therefore,

$$\lim_{h \rightarrow 0^-} \frac{f(h) - f(0)}{h} = 0, \\ \text{but } \lim_{h \rightarrow 0^+} \frac{f(h) - f(0)}{h} = 1.$$

Thus $f'(0)$ does not exist; f is not differentiable at 0. Once again, however, $f'(x)$ exists for $x \neq 0$ —it is easy to see that

$$f'(x) = \begin{cases} 2x, & x < 0 \\ 1, & x > 0. \end{cases}$$

The graphs of f and f' are shown in Figure 12.

Even worse things happen for $f(x) = \sqrt{|x|}$. For this function

$$\frac{f(h) - f(0)}{h} = \begin{cases} \frac{\sqrt{h}}{h} = \frac{1}{\sqrt{h}}, & h > 0 \\ \frac{\sqrt{-h}}{h} = -\frac{1}{\sqrt{-h}}, & h < 0. \end{cases}$$

In this case the right-hand limit

$$\lim_{h \rightarrow 0^+} \frac{f(h) - f(0)}{h} = \lim_{h \rightarrow 0^+} \frac{1}{\sqrt{h}}$$

does not exist; instead $1/\sqrt{h}$ becomes arbitrarily large as h approaches 0. And, what's more, $-1/\sqrt{-h}$ becomes arbitrarily large in absolute value, but negative (Figure 13).

FIGURE 12

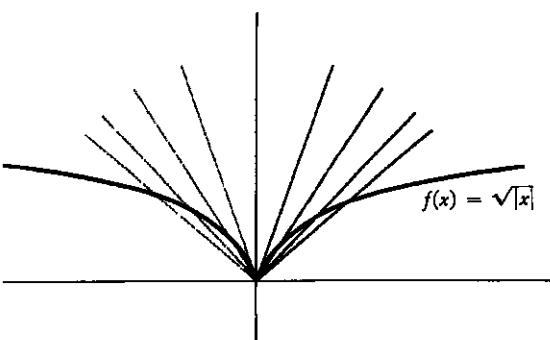


FIGURE 13

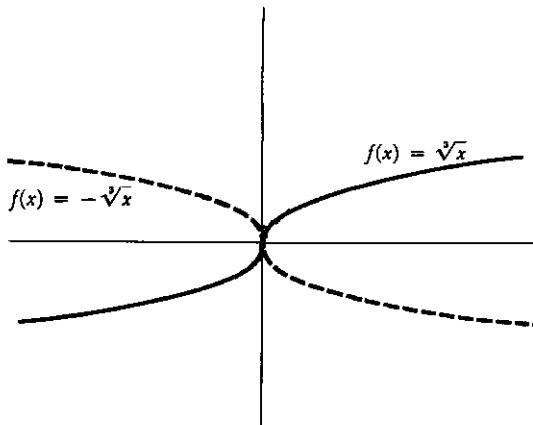


FIGURE 14

The function $f(x) = \sqrt[3]{x}$, although not differentiable at 0, is at least a little better behaved than this. The quotient

$$\frac{f(h) - f(0)}{h} = \frac{\sqrt[3]{h}}{h} = \frac{h^{1/3}}{h} = \frac{1}{h^{2/3}} = \frac{1}{(\sqrt[3]{h})^2}$$

simply becomes arbitrarily large as h goes to 0. Sometimes one says that f has an “infinite” derivative at 0. Geometrically this means that the graph of f has a “tangent line” which is parallel to the vertical axis (Figure 14). Of course, $f(x) = -\sqrt[3]{x}$ has the same geometric property, but one would say that f has a derivative of “negative infinity” at 0.

Remember that differentiability is supposed to be an improvement over mere continuity. This idea is supported by the many examples of functions which are continuous, but not differentiable; however, one important point remains to be noted:

THEOREM 1 If f is differentiable at a , then f is continuous at a .

PROOF

$$\begin{aligned}\lim_{h \rightarrow 0} f(a+h) - f(a) &= \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \cdot h \\ &= \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \cdot \lim_{h \rightarrow 0} h \\ &= f'(a) \cdot 0 \\ &= 0.\end{aligned}$$

As we pointed out in Chapter 5, the equation $\lim_{h \rightarrow 0} f(a+h) - f(a) = 0$ is equivalent to $\lim_{x \rightarrow a} f(x) = f(a)$; thus f is continuous at a . ■

It is very important to remember Theorem 1, and just as important to remember that the converse is not true. A differentiable function is continuous, but a continuous function need not be differentiable (keep in mind the function $f(x) = |x|$, and you will never forget which statement is true and which false).

The continuous functions examined so far have been differentiable at all points with at most one exception, but it is easy to give examples of continuous functions which are not differentiable at several points, even an infinite number (Figure 15). Actually, one can do much worse than this. There is a function which is *continuous*

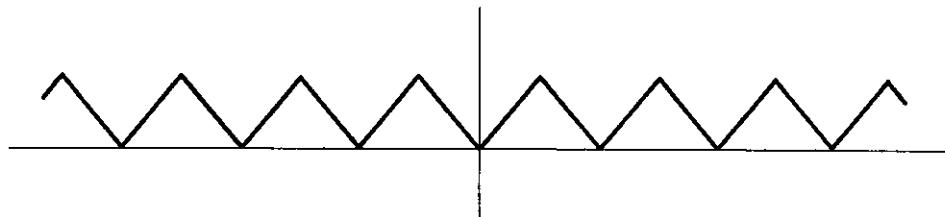


FIGURE 15

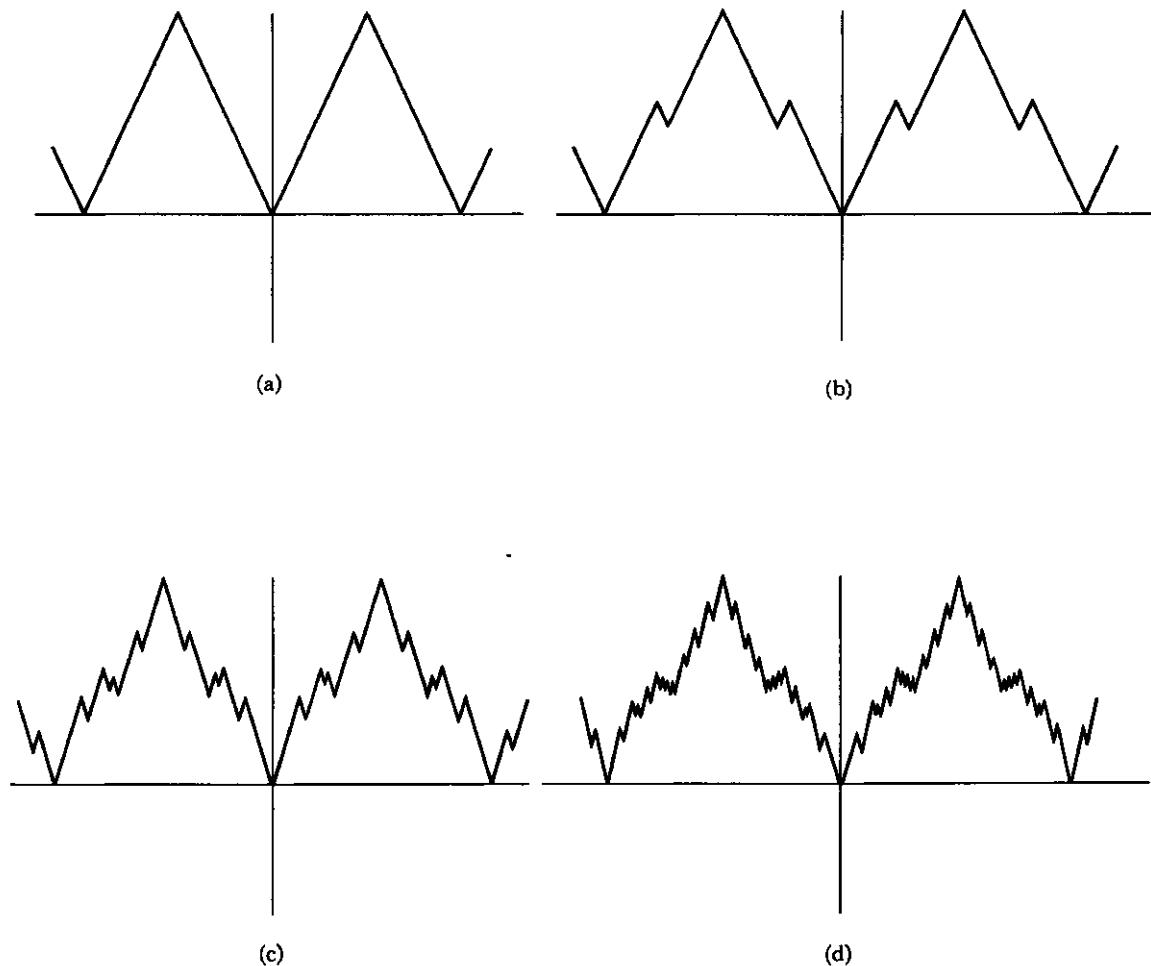


FIGURE 16

everywhere and differentiable nowhere! Unfortunately, the definition of this function will be inaccessible to us until Chapter 24, and I have been unable to persuade the artist to draw it (consider carefully what the graph should look like and you will sympathize with her point of view). It is possible to draw some rough approximations to the graph, however; several successively better approximations are shown in Figure 16.

Although such spectacular examples of nondifferentiability must be postponed, we can, with a little ingenuity, find a continuous function which is not differentiable at infinitely many points, *all of which are in* $[0, 1]$. One such function is illustrated in Figure 17. The reader is given the problem of defining it precisely; it is a straight line version of the function

$$f(x) = \begin{cases} x \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

This particular function f is itself quite sensitive to the question of differentiability. Indeed, for $h \neq 0$ we have

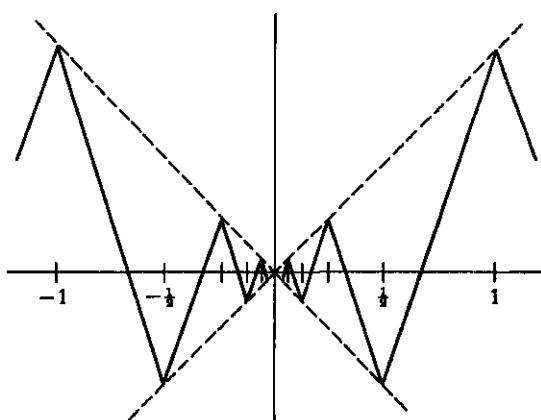


FIGURE 17

$$\frac{f(h) - f(0)}{h} = \frac{h \sin \frac{1}{h} - 0}{h} = \sin \frac{1}{h}.$$

Long ago we proved that $\lim_{h \rightarrow 0} \sin 1/h$ does not exist, so f is not differentiable at 0.

Geometrically, one can see that a tangent line cannot exist, by noting that the secant line through $(0, 0)$ and $(h, f(h))$ in Figure 18 can have any slope between -1 and 1 , no matter how small we require h to be.

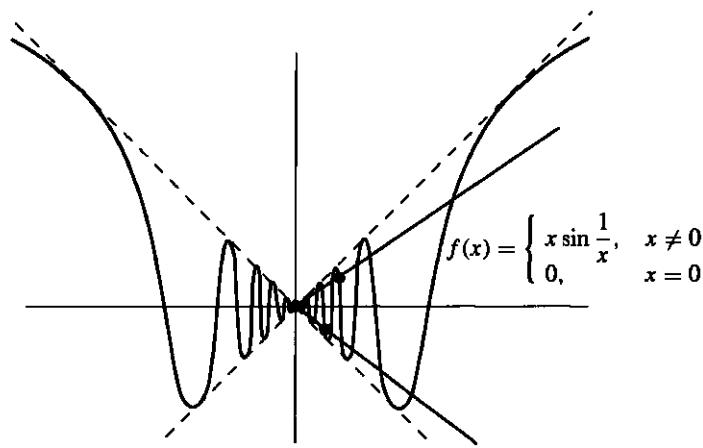


FIGURE 18

This finding represents something of a triumph; although continuous, the function f seems somehow quite unreasonable, and we can now enunciate one mathematically undesirable feature of this function—it is not differentiable at 0. Nevertheless, one should not become too enthusiastic about the criterion of differentiability. For example, the function

$$g(x) = \begin{cases} x^2 \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0 \end{cases}$$

is differentiable at 0; in fact $g'(0) = 0$:

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{g(h) - g(0)}{h} &= \lim_{h \rightarrow 0} \frac{h^2 \sin \frac{1}{h}}{h} \\ &= \lim_{h \rightarrow 0} h \sin \frac{1}{h} \\ &= 0. \end{aligned}$$

The tangent line to the graph of g at $(0, 0)$ is therefore the horizontal axis (Figure 19).

This example suggests that we should seek even more restrictive conditions on a function than mere differentiability. We can actually use the derivative to formulate such conditions if we introduce another set of definitions, the last of this chapter.

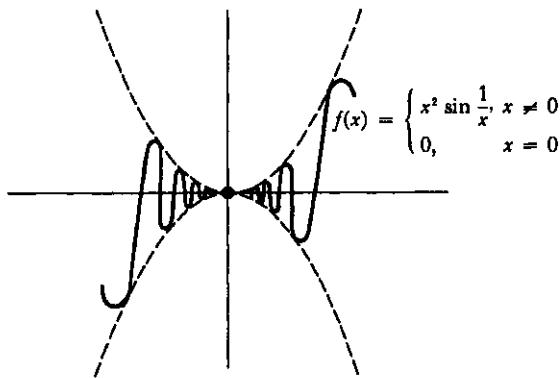


FIGURE 19

For any function f , we obtain, by taking the derivative, a new function f' (whose domain may be considerably smaller than that of f). The notion of differentiability can be applied to the function f' , of course, yielding another function $(f')'$, whose domain consists of all points a such that f' is differentiable at a . The function $(f')'$ is usually written simply f'' and is called the **second derivative** of f . If $f''(a)$ exists, then f is said to be 2-times differentiable at a , and the number $f''(a)$ is called the **second derivative of f at a** .

In physics the second derivative is particularly important. If $s(t)$ is the position at time t of a particle moving along a straight line, then $s''(t)$ is called the **acceleration** at time t . Acceleration plays a special role in physics, because, as stated in Newton's laws of motion, the force on a particle is the product of its mass and its acceleration. Consequently you can feel the second derivative when you sit in an accelerating car.

There is no reason to stop at the second derivative—we can define $f''' = (f'')'$, $f'''' = (f''')'$, etc. This notation rapidly becomes unwieldy, so the following abbreviation is usually adopted (it is really a recursive definition):

$$\begin{aligned} f^{(1)} &= f', \\ f^{(k+1)} &= (f^{(k)})'. \end{aligned}$$

Thus

$$\begin{aligned} f^{(1)} &= f' \\ f^{(2)} &= f'' = (f')' \\ f^{(3)} &= f''' = (f'')' \\ f^{(4)} &= f'''' = (f''')' \\ &\text{etc.} \end{aligned}$$

The various functions $f^{(k)}$, for $k \geq 2$, are sometimes called **higher-order derivatives** of f .

Usually, we resort to the notation $f^{(k)}$ only for $k \geq 4$, but it is convenient to have $f^{(k)}$ defined for smaller k also. In fact, a reasonable definition can be made for $f^{(0)}$, namely,

$$f^{(0)} = f.$$

Leibnizian notation for higher-order derivatives should also be mentioned. The natural Leibnizian symbol for $f''(x)$, namely,

$$\frac{d \left(\frac{df(x)}{dx} \right)}{dx},$$

is abbreviated to

$$\frac{d^2 f(x)}{(dx)^2}, \quad \text{or more frequently to} \quad \frac{d^2 f(x)}{dx^2}$$

Similar notation is used for $f^{(k)}(x)$.

The following example illustrates the notation $f^{(k)}$, and also shows, in one very simple case, how various higher-order derivatives are related to the original function. Let $f(x) = x^2$. Then, as we have already checked,

$$\begin{aligned} f'(x) &= 2x, \\ f''(x) &= 2, \\ f'''(x) &= 0, \\ f^{(k)}(x) &= 0, \quad \text{if } k \geq 3. \end{aligned}$$

Figure 20 shows the function f , together with its various derivatives.

A rather more illuminating example is presented by the following function, whose graph is shown in Figure 21(a):

$$f(x) = \begin{cases} x^2, & x \geq 0 \\ -x^2, & x \leq 0. \end{cases}$$

It is easy to see that

$$\begin{aligned} f'(a) &= 2a & \text{if } a > 0, \\ f'(a) &= -2a & \text{if } a < 0. \end{aligned}$$

Moreover,

$$\begin{aligned} f'(0) &= \lim_{h \rightarrow 0} \frac{f(h) - f(0)}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(h)}{h}. \end{aligned}$$

Now

$$\begin{aligned} \lim_{h \rightarrow 0^+} \frac{f(h)}{h} &= \lim_{h \rightarrow 0^+} \frac{h^2}{h} = 0 \\ \text{and } \lim_{h \rightarrow 0^-} \frac{f(h)}{h} &= \lim_{h \rightarrow 0^-} \frac{-h^2}{h} = 0, \end{aligned}$$

so

$$f'(0) = \lim_{h \rightarrow 0} \frac{f(h)}{h} = 0.$$

This information can all be summarized as follows:

$$f'(x) = 2|x|.$$

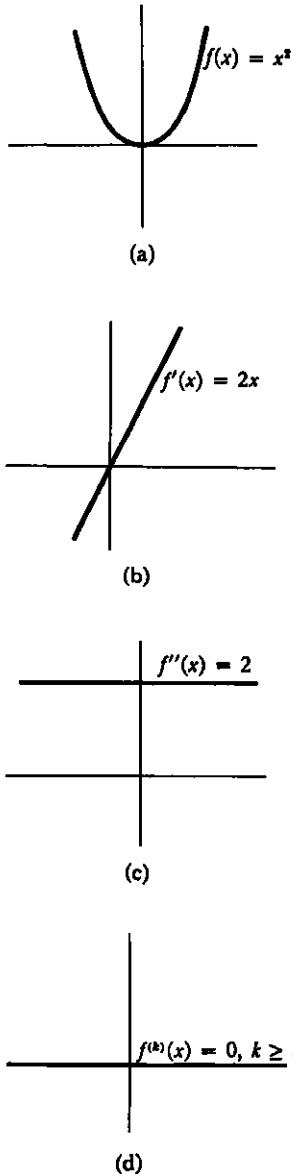


FIGURE 20

It follows that $f''(0)$ does not exist! Existence of the second derivative is thus a rather strong criterion for a function to satisfy. Even a "smooth looking" function like f reveals some irregularity when examined with the second derivative. This suggests that the irregular behavior of the function

$$g(x) = \begin{cases} x^2 \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0 \end{cases}$$

might also be revealed by the second derivative. At the moment we know that $g'(0) = 0$, but we do not know $g'(a)$ for any $a \neq 0$, so it is hopeless to begin computing $g''(0)$. We will return to this question at the end of the next chapter, after we have perfected the technique of finding derivatives.

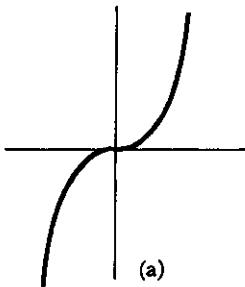
PROBLEMS

1. (a) Prove, working directly from the definition, that if $f(x) = 1/x$, then $f'(a) = -1/a^2$, for $a \neq 0$.
 (b) Prove that the tangent line to the graph of f at $(a, 1/a)$ does not intersect the graph of f , except at $(a, 1/a)$.
2. (a) Prove that if $f(x) = 1/x^2$, then $f'(a) = -2/a^3$ for $a \neq 0$.
 (b) Prove that the tangent line to f at $(a, 1/a^2)$ intersects f at one other point, which lies on the opposite side of the vertical axis.
3. Prove that if $f(x) = \sqrt{x}$, then $f'(a) = 1/(2\sqrt{a})$, for $a > 0$. (The expression you obtain for $[f(a+h) - f(a)]/h$ will require some algebraic face lifting, but the answer should suggest the right trick.)
4. For each natural number n , let $S_n(x) = x^n$. Remembering that $S_1'(x) = 1$, $S_2'(x) = 2x$, and $S_3'(x) = 3x^2$, conjecture a formula for $S_n'(x)$. Prove your conjecture. (The expression $(x+h)^n$ may be expanded by the binomial theorem.)
5. Find f' if $f(x) = [x]$.
6. Prove, starting from the definition (and drawing a picture to illustrate):
 - (a) if $g(x) = f(x) + c$, then $g'(x) = f'(x)$;
 - (b) if $g(x) = cf(x)$, then $g'(x) = cf'(x)$.
7. Suppose that $f(x) = x^3$.
 - (a) What is $f'(9)$, $f'(25)$, $f'(36)$?
 - (b) What is $f'(3^2)$, $f'(5^2)$, $f'(6^2)$?
 - (c) What is $f'(a^2)$, $f'(x^2)$?

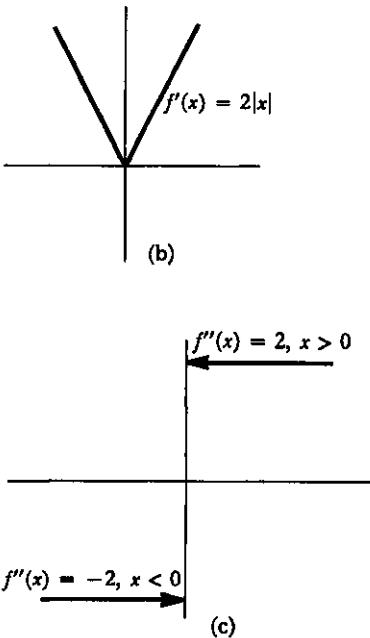
If you do not find this problem silly, you are missing a very important point: $f'(x^2)$ means the derivative of f at the number which we happen to be calling x^2 ; it is *not* the derivative at x of the function $g(x) = f(x^2)$. Just to drive the point home:

- (d) For $f(x) = x^3$, compare $f'(x^2)$ and $g'(x)$ where $g(x) = f(x^2)$.

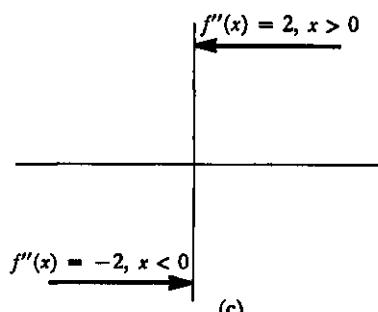
$$f(x) = \begin{cases} x^3, & x \geq 0 \\ -x^3, & x \leq 0 \end{cases}$$



(a)



(b)



(c)

FIGURE 21

8. (a) Suppose $g(x) = f(x+c)$. Prove (starting from the definition) that $g'(x) = f'(x+c)$. Draw a picture to illustrate this. To do this problem you must write out the definitions of $g'(x)$ and $f'(x+c)$ correctly. The purpose of Problem 7 was to convince you that although this problem is easy, it is not an utter triviality, and there is something to prove: you cannot simply put prime marks into the equation $g(x) = f(x+c)$. To emphasize this point:
- (b) Prove that if $g(x) = f(cx)$, then $g'(x) = c \cdot f'(cx)$. Try to see pictorially why this should be true, also.
- (c) Suppose that f is differentiable and periodic, with period a (i.e., $f(x+a) = f(x)$ for all x). Prove that f' is also periodic.
9. Find $f'(x)$ and also $f'(x+3)$ in the following cases. Be very methodical, or you will surely slip up somewhere. Consult the answers (after you do the problem, naturally).
- (i) $f(x) = (x+3)^5$.
- (ii) $f(x+3) = x^5$.
- (iii) $f(x+3) = (x+5)^7$.
10. Find $f'(x)$ if $f(x) = g(t+x)$, and if $f(t) = g(t+x)$. The answers will *not* be the same.
11. (a) Prove that Galileo was wrong: if a body falls a distance $s(t)$ in t seconds, and s' is proportional to s , then s cannot be a function of the form $s(t) = ct^2$.
- (b) Prove that the following facts are true about s if $s(t) = (a/2)t^2$ (the first fact will show why we switched from c to $a/2$):
- (i) $s''(t) = a$ (the acceleration is constant).
- (ii) $[s'(t)]^2 = 2as(t)$.
- (c) If s is measured in feet, the value of a is 32. How many seconds do you have to get out of the way of a chandelier which falls from a 400-foot ceiling? If you don't make it, how fast will the chandelier be going when it hits you? Where was the chandelier when it was moving with half that speed?
12. Imagine a road on which the speed limit is specified at every single point. In other words, there is a certain function L such that the speed limit x miles from the beginning of the road is $L(x)$. Two cars, A and B , are driving along this road; car A 's position at time t is $a(t)$, and car B 's is $b(t)$.
- (a) What equation expresses the fact that car A always travels at the speed limit? (The answer is *not* $a'(t) = L(t)$.)
- (b) Suppose that A always goes at the speed limit, and that B 's position at time t is A 's position at time $t-1$. Show that B is also going at the speed limit at all times.
- (c) Suppose B always stays a constant distance behind A . Under what conditions will B still always travel at the speed limit?

13. Suppose that $f(a) = g(a)$ and that the left-hand derivative of f at a equals the right-hand derivative of g at a . Define $h(x) = f(x)$ for $x \leq a$, and $h(x) = g(x)$ for $x \geq a$. Prove that h is differentiable at a .
14. Let $f(x) = x^2$ if x is rational, and $f(x) = 0$ if x is irrational. Prove that f is differentiable at 0. (Don't be scared by this function. Just write out the definition of $f'(0)$.)
- *15. (a) Let f be a function such that $|f(x)| \leq x^2$ for all x . Prove that f is differentiable at 0. (If you have done Problem 14 you should be able to do this.)
(b) This result can be generalized if x^2 is replaced by $|g(x)|$, where g has what property?
16. Let $\alpha > 1$. If f satisfies $|f(x)| \leq |x|^\alpha$, prove that f is differentiable at 0.
17. Let $0 < \beta < 1$. Prove that if f satisfies $|f(x)| \geq |x|^\beta$ and $f(0) = 0$, then f is not differentiable at 0.
- *18. Let $f(x) = 0$ for irrational x , and $1/q$ for $x = p/q$ in lowest terms. Prove that f is not differentiable at a for any a . Hint: It obviously suffices to prove this for irrational a . Why? If $a = m.a_1a_2a_3\dots$ is the decimal expansion of a , consider $[f(a+h) - f(a)]/h$ for h rational, and also for

$$h = -0.00\dots 0a_{n+1}a_{n+2}\dots$$

19. (a) Suppose that $f(a) = g(a) = h(a)$, that $f(x) \leq g(x) \leq h(x)$ for all x , and that $f'(a) = h'(a)$. Prove that g is differentiable at a , and that $f'(a) = g'(a) = h'(a)$. (Begin with the definition of $g'(a)$.)
(b) Show that the conclusion does not follow if we omit the hypothesis $f(a) = g(a) = h(a)$.
20. Let f be any polynomial function; we will see in the next chapter that f is differentiable. The tangent line to f at $(a, f(a))$ is the graph of $g(x) = f'(a)(x - a) + f(a)$. Thus $f(x) - g(x)$ is the polynomial function $d(x) = f(x) - f'(a)(x - a) - f(a)$. We have already seen that if $f(x) = x^2$, then $d(x) = (x - a)^2$, and if $f(x) = x^3$, then $d(x) = (x - a)^2(x + 2a)$.
(a) Find $d(x)$ when $f(x) = x^4$, and show that it is divisible by $(x - a)^2$.
(b) There certainly seems to be some evidence that $d(x)$ is always divisible by $(x - a)^2$. Figure 22 provides an intuitive argument: usually, lines parallel to the tangent line will intersect the graph at two points; the tangent line intersects the graph only once near the point, so the intersection should be a "double intersection." To give a rigorous proof, first note that

$$\frac{d(x)}{x - a} = \frac{f(x) - f(a)}{x - a} - f'(a).$$

Now answer the following questions. Why is $f(x) - f(a)$ divisible by $(x - a)$? Why is there a polynomial function h such that $h(x) =$

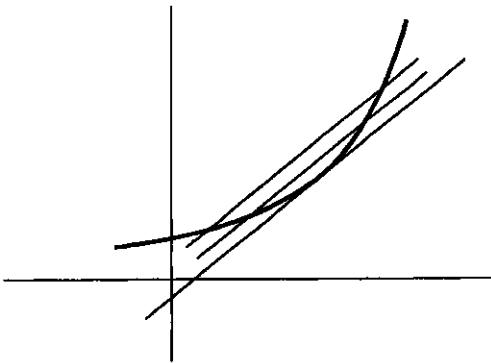


FIGURE 22

$d(x)/(x - a)$ for $x \neq a$? Why is $\lim_{x \rightarrow a} h(x) = 0$? Why is $h(a) = 0$? Why does this solve the problem?

21. (a) Show that $f'(a) = \lim_{x \rightarrow a} [f(x) - f(a)]/(x - a)$. (Nothing deep here.)
 (b) Show that derivatives are a “local property”: if $f(x) = g(x)$ for all x in some open interval containing a , then $f'(a) = g'(a)$. (This means that in computing $f'(a)$, you can ignore $f(x)$ for any particular $x \neq a$. Of course you can’t ignore $f(x)$ for all such x at once!)

- *22. (a) Suppose that f is differentiable at x . Prove that

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x - h)}{2h}.$$

Hint: Remember an old algebraic trick—a number is not changed if the same quantity is added to and then subtracted from it.

- **(b) Prove, more generally, that

$$f'(x) = \lim_{h,k \rightarrow 0^+} \frac{f(x + h) - f(x - k)}{h + k}.$$

- *23. Prove that if f is even, then $f'(x) = -f'(-x)$. (In order to minimize confusion, let $g(x) = f(-x)$; find $g'(x)$ and then remember what other thing g is.) Draw a picture!

- *24. Prove that if f is odd, then $f'(x) = f'(-x)$. Once again, draw a picture.

25. Problems 23 and 24 say that f' is even if f is odd, and odd if f is even. What can therefore be said about $f^{(k)}$?

26. Find $f''(x)$ if

- (i) $f(x) = x^3$.
- (ii) $f(x) = x^5$.
- (iii) $f'(x) = x^4$.
- (iv) $f(x + 3) = x^5$.

27. If $S_n(x) = x^n$, and $0 \leq k \leq n$, prove that

$$\begin{aligned} S_n^{(k)}(x) &= \frac{n!}{(n-k)!} x^{n-k} \\ &= k! \binom{n}{k} x^{n-k}. \end{aligned}$$

- *28. (a) Find $f'(x)$ if $f(x) = |x|^3$. Find $f''(x)$. Does $f'''(x)$ exist for all x ?
 (b) Analyze f similarly if $f(x) = x^4$ for $x \geq 0$ and $f(x) = -x^4$ for $x \leq 0$.
 *29. Let $f(x) = x^n$ for $x \geq 0$ and let $f(x) = 0$ for $x \leq 0$. Prove that $f^{(n-1)}$ exists (and find a formula for it), but that $f^{(n)}(0)$ does not exist.

30. Interpret the following specimens of Leibnizian notation; each is a restatement of some fact occurring in a previous problem.

$$(i) \quad \frac{dx^n}{dx} = nx^{n-1}$$

$$(ii) \quad \frac{dz}{dy} = -\frac{1}{y^2} \text{ if } z = \frac{1}{y}.$$

$$(iii) \quad \frac{d[f(x) + c]}{dx} = \frac{df(x)}{dx}.$$

$$(iv) \quad \frac{d[cf(x)]}{dx} = c \frac{df(x)}{dx}.$$

$$(v) \quad \frac{dz}{dx} = \frac{dy}{dx} \text{ if } z = y + c.$$

$$(vi) \quad \left. \frac{dx^3}{dx} \right|_{x=a^2} = 3a^4.$$

$$(vii) \quad \left. \frac{df(x+a)}{dx} \right|_{x=b} = \left. \frac{df(x)}{dx} \right|_{x=b+a}.$$

$$(viii) \quad \left. \frac{df(cx)}{dx} \right|_{x=b} = c \cdot \left. \frac{df(x)}{dx} \right|_{x=cb}.$$

$$(ix) \quad \left. \frac{df(cx)}{dx} \right|_{x=b} = c \cdot \left. \frac{df(y)}{dy} \right|_{y=cx}.$$

$$(x) \quad \frac{d^k x^n}{dx^k} = k! \binom{n}{k} x^{n-k}.$$

CHAPTER 10 DIFFERENTIATION

The process of finding the derivative of a function is called *differentiation*. From the previous chapter you may have the impression that this process is usually laborious, requires recourse to the definition of the derivative, and depends upon successfully recognizing some limit. It is true that such a procedure is often the only possible approach—if you forget the definition of the derivative you are likely to be lost. Nevertheless, in this chapter we will learn to differentiate a large number of functions, without the necessity of even recalling the definition. A few theorems will provide a mechanical process for differentiating a large class of functions, which are formed from a few simple functions by the process of addition, multiplication, division, and composition. This description should suggest what theorems will be proved. We will first find the derivative of a few simple functions, and then prove theorems about the sum, products, quotients, and compositions of differentiable functions. The first theorem is merely a formal recognition of a computation carried out in the previous chapter.

THEOREM 1 If f is a constant function, $f(x) = c$, then

$$f'(a) = 0 \quad \text{for all numbers } a.$$

PROOF
$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} = \lim_{h \rightarrow 0} \frac{c - c}{h} = 0. \blacksquare$$

The second theorem is also a special case of a computation in the last chapter.

THEOREM 2 If f is the identity function, $f(x) = x$, then

$$f'(a) = 1 \quad \text{for all numbers } a.$$

PROOF
$$\begin{aligned} f'(a) &= \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \\ &= \lim_{h \rightarrow 0} \frac{a+h-a}{h} \\ &= \lim_{h \rightarrow 0} \frac{h}{h} = 1. \blacksquare \end{aligned}$$

The derivative of the sum of two functions is just what one would hope—the sum of the derivatives.

THEOREM 3 If f and g are differentiable at a , then $f + g$ is also differentiable at a , and

$$(f + g)'(a) = f'(a) + g'(a).$$

PROOF

$$\begin{aligned} (f + g)'(a) &= \lim_{h \rightarrow 0} \frac{(f + g)(a + h) - (f + g)(a)}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(a + h) + g(a + h) - [f(a) + g(a)]}{h} \\ &= \lim_{h \rightarrow 0} \left[\frac{f(a + h) - f(a)}{h} + \frac{g(a + h) - g(a)}{h} \right] \\ &= \lim_{h \rightarrow 0} \frac{f(a + h) - f(a)}{h} + \lim_{h \rightarrow 0} \frac{g(a + h) - g(a)}{h} \\ &= f'(a) + g'(a). \blacksquare \end{aligned}$$

The formula for the derivative of a product is not as simple as one might wish, but it is nevertheless pleasantly symmetric, and the proof requires only a simple algebraic trick, which we have found useful before—a number is not changed if the same quantity is added to and subtracted from it.

THEOREM 4 If f and g are differentiable at a , then

$$(f \cdot g)'(a) = f'(a) \cdot g(a) + f(a) \cdot g'(a).$$

PROOF

$$\begin{aligned} (f \cdot g)'(a) &= \lim_{h \rightarrow 0} \frac{(f \cdot g)(a + h) - (f \cdot g)(a)}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(a + h)g(a + h) - f(a)g(a)}{h} \\ &= \lim_{h \rightarrow 0} \left[\frac{f(a + h)[g(a + h) - g(a)]}{h} + \frac{[f(a + h) - f(a)]g(a)}{h} \right] \\ &= \lim_{h \rightarrow 0} f(a + h) \cdot \lim_{h \rightarrow 0} \frac{g(a + h) - g(a)}{h} + \lim_{h \rightarrow 0} \frac{f(a + h) - f(a)}{h} \cdot \lim_{h \rightarrow 0} g(a) \\ &= f(a) \cdot g'(a) + f'(a) \cdot g(a). \end{aligned}$$

(Notice that we have used Theorem 9-1 to conclude that $\lim_{h \rightarrow 0} f(a + h) = f(a)$.) \blacksquare

In one special case Theorem 4 simplifies considerably:

THEOREM 5 If $g(x) = cf(x)$ and f is differentiable at a , then g is differentiable at a , and

$$g'(a) = c \cdot f'(a).$$

PROOF If $h(x) = c$, so that $g = h \cdot f$, then by Theorem 4,

$$\begin{aligned} g'(a) &= (h \cdot f)'(a) \\ &= h(a) \cdot f'(a) + h'(a) \cdot f(a) \\ &= c \cdot f'(a) + 0 \cdot f(a) \\ &= c \cdot f'(a). \blacksquare \end{aligned}$$

Notice, in particular, that $(-f)'(a) = -f'(a)$, and consequently $(f - g)'(a) = (f + [-g])'(a) = f'(a) - g'(a)$.

To demonstrate what we have already achieved, we will compute the derivative of some more special functions.

THEOREM 6 If $f(x) = x^n$ for some natural number n , then

$$f'(a) = na^{n-1} \quad \text{for all } a.$$

PROOF The proof will be by induction on n . For $n = 1$ this is simply Theorem 2. Now assume that the theorem is true for n , so that if $f(x) = x^n$, then

$$f'(a) = na^{n-1} \quad \text{for all } a.$$

Let $g(x) = x^{n+1}$. If $I(x) = x$, the equation $x^{n+1} = x^n \cdot x$ can be written

$$g(x) = f(x) \cdot I(x) \quad \text{for all } x;$$

thus $g = f \cdot I$. It follows from Theorem 4 that

$$\begin{aligned} g'(a) &= (f \cdot I)'(a) = f'(a) \cdot I(a) + f(a) \cdot I'(a) \\ &= na^{n-1} \cdot a + a^n \cdot 1 \\ &= na^n + a^n \\ &= (n+1)a^n, \quad \text{for all } a. \end{aligned}$$

This is precisely the case $n+1$ which we wished to prove. ■

Putting together the theorems proved so far we can now find f' for f of the form

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0.$$

We obtain

$$f'(x) = na_n x^{n-1} + (n-1)a_{n-1} x^{n-2} + \cdots + 2a_2 x + a_1.$$

We can also find f'' :

$$f''(x) = n(n-1)a_n x^{n-2} + (n-1)(n-2)a_{n-1} x^{n-3} + \cdots + 2a_2.$$

This process can be continued easily. Each differentiation reduces the highest power of x by 1, and eliminates one more a_i . It is a good idea to work out the derivatives f''' , $f^{(4)}$, and perhaps $f^{(5)}$, until the pattern becomes quite clear. The last interesting derivative is

$$f^{(n)}(x) = n! a_n;$$

for $k > n$ we have

$$f^{(k)}(x) = 0.$$

Clearly, the next step in our program is to find the derivative of a quotient f/g . It is quite a bit simpler, and, because of Theorem 4, obviously sufficient to find the derivative of $1/g$.

THEOREM 7 If g is differentiable at a , and $g(a) \neq 0$, then $1/g$ is differentiable at a , and

$$\left(\frac{1}{g}\right)'(a) = \frac{-g'(a)}{[g(a)]^2}.$$

PROOF Before we even write

$$\frac{\left(\frac{1}{g}\right)(a+h) - \left(\frac{1}{g}\right)(a)}{h}$$

we must be sure that this expression makes sense—it is necessary to check that $(1/g)(a+h)$ is defined for sufficiently small h . This requires only two observations. Since g is, by hypothesis, differentiable at a , it follows from Theorem 9-1 that g is continuous at a . Since $g(a) \neq 0$, it follows from Theorem 6-3 that there is some $\delta > 0$ such that $g(a+h) \neq 0$ for $|h| < \delta$. Therefore $(1/g)(a+h)$ does make sense for small enough h , and we can write

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{\left(\frac{1}{g}\right)(a+h) - \left(\frac{1}{g}\right)(a)}{h} &= \lim_{h \rightarrow 0} \frac{\frac{1}{g(a+h)} - \frac{1}{g(a)}}{h} \\ &= \lim_{h \rightarrow 0} \frac{g(a) - g(a+h)}{h[g(a) \cdot g(a+h)]} \\ &= \lim_{h \rightarrow 0} \frac{-[g(a+h) - g(a)]}{h} \cdot \frac{1}{g(a)g(a+h)} \\ &= \lim_{h \rightarrow 0} \frac{-[g(a+h) - g(a)]}{h} \cdot \lim_{h \rightarrow 0} \frac{1}{g(a) \cdot g(a+h)} \\ &= -g'(a) \cdot \frac{1}{[g(a)]^2}. \end{aligned}$$

(Notice that we have used continuity of g at a once again.) ■

The general formula for the derivative of a quotient is now easy to derive. Though not particularly appealing, it is important, and must simply be memorized (I always use the incantation: “bottom times derivative of top, minus top times derivative of bottom, over bottom squared.”)

THEOREM 8 If f and g are differentiable at a and $g(a) \neq 0$, then f/g is differentiable at a , and

$$\left(\frac{f}{g}\right)'(a) = \frac{g(a) \cdot f'(a) - f(a) \cdot g'(a)}{[g(a)]^2}.$$

PROOF Since $f/g = f \cdot (1/g)$ we have

$$\begin{aligned} \left(\frac{f}{g}\right)'(a) &= \left(f \cdot \frac{1}{g}\right)'(a) \\ &= f'(a) \cdot \left(\frac{1}{g}\right)(a) + f(a) \cdot \left(\frac{1}{g}\right)'(a) \\ &= \frac{f'(a)}{g(a)} + \frac{f(a)(-g'(a))}{[g(a)]^2} \\ &= \frac{f'(a) \cdot g(a) - f(a) \cdot g'(a)}{[g(a)]^2}. \blacksquare \end{aligned}$$

We can now differentiate a few more functions. For example,

$$\begin{aligned} \text{if } f(x) = \frac{x^2 - 1}{x^2 + 1}, \text{ then } f'(x) &= \frac{(x^2 + 1)(2x) - (x^2 - 1)(2x)}{(x^2 + 1)^2} = \frac{4x}{(x^2 + 1)^2}; \\ \text{if } f(x) = \frac{x}{x^2 + 1}, \text{ then } f'(x) &= \frac{(x^2 + 1) - x(2x)}{(x^2 + 1)^2} = \frac{1 - x^2}{(x^2 + 1)^2}; \\ \text{if } f(x) = \frac{1}{x}, \quad \text{then } f'(x) &= -\frac{1}{x^2} = (-1)x^{-2}. \end{aligned}$$

Notice that the last example can be generalized: if

$$f(x) = x^{-n} = \frac{1}{x^n}, \quad \text{for some natural number } n,$$

then

$$f'(x) = \frac{-nx^{n-1}}{x^{2n}} = (-n)x^{-n-1};$$

thus Theorem 6 actually holds both for positive and negative integers. If we interpret $f(x) = x^0$ to mean $f(x) = 1$, and $f'(x) = 0 \cdot x^{-1}$ to mean $f'(x) = 0$, then Theorem 6 is true for $n = 0$ also. (The word “interpret” is necessary because it is not clear how 0^0 should be defined and, in any case, $0 \cdot 0^{-1}$ is meaningless.)

Further progress in differentiation requires the knowledge of the derivatives of certain special functions to be studied later. One of these is the sine function. For the moment we shall divulge, and use, the following information, without proof:

$$\begin{aligned} \sin'(a) &= \cos a && \text{for all } a, \\ \cos'(a) &= -\sin a && \text{for all } a, \end{aligned}$$

This information allows us to differentiate many other functions. For example, if

$$f(x) = x \sin x,$$

then

$$\begin{aligned} f'(x) &= x \cos x + \sin x, \\ f''(x) &= -x \sin x + \cos x + \cos x \\ &= -x \sin x + 2 \cos x; \end{aligned}$$

if

$$g(x) = \sin^2 x = \sin x \cdot \sin x,$$

then

$$\begin{aligned} g'(x) &= \sin x \cos x + \cos x \sin x \\ &= 2 \sin x \cos x, \\ g''(x) &= 2[(\sin x)(-\sin x) + \cos x \cos x] \\ &= 2[\cos^2 x - \sin^2 x]; \end{aligned}$$

if

$$h(x) = \cos^2 x = \cos x \cdot x,$$

then

$$\begin{aligned} h'(x) &= (\cos x)(-\sin x) + (-\sin x) \cos x \\ &= -2 \sin x \cos x, \\ h''(x) &= -2[\cos^2 x - \sin^2 x]. \end{aligned}$$

Notice that

$$g'(x) + h'(x) = 0,$$

hardly surprising, since $(g + h)(x) = \sin^2 x + \cos^2 x = 1$. As we would expect, we also have $g''(x) + h''(x) = 0$.

The examples above involved only products of two functions. A function involving triple products can be handled by Theorem 4 also; in fact it can be handled in two ways. Remember that $f \cdot g \cdot h$ is an abbreviation for

$$(f \cdot g) \cdot h \quad \text{or} \quad f \cdot (g \cdot h).$$

Choosing the first of these, for example, we have

$$\begin{aligned} (f \cdot g \cdot h)'(x) &= (f \cdot g)'(x) \cdot h(x) + (f \cdot g)(x)h'(x) \\ &= [f'(x)g(x) + f(x)g'(x)]h(x) + f(x)g(x)h'(x) \\ &= f'(x)g(x)h(x) + f(x)g'(x)h(x) + f(x)g(x)h'(x). \end{aligned}$$

The choice of $f \cdot (g \cdot h)$ would, of course, have given the same result, with a different intermediate step. The final answer is completely symmetric and easily remembered:

$(f \cdot g \cdot h)'$ is the sum of the three terms obtained by differentiating each of f , g , and h and multiplying by the other two.

For example, if

$$f(x) = x^3 \sin x \cos x,$$

then

$$f'(x) = 3x^2 \sin x \cos x + x^3 \cos x \cos x + x^3(\sin x)(-\sin x).$$

Products of more than 3 functions can be handled similarly. For example, you should have little difficulty deriving the formula

$$\begin{aligned} (f \cdot g \cdot h \cdot k)'(x) &= f'(x)g(x)h(x)k(x) + f(x)g'(x)h(x)k(x) \\ &\quad + f(x)g(x)h'(x)k(x) + f(x)g(x)h(x)k'(x). \end{aligned}$$

You might even try to prove (by induction) the general formula:

$$(f_1 \cdot \dots \cdot f_n)'(x) = \sum_{i=1}^n f_1(x) \cdot \dots \cdot f_{i-1}(x) f_i'(x) f_{i+1}(x) \cdot \dots \cdot f_n(x).$$

Differentiating the most interesting functions obviously requires a formula for $(f \circ g)'(x)$ in terms of f' and g' . To ensure that $f \circ g$ be differentiable at a , one reasonable hypothesis would seem to be that g be differentiable at a . Since the behavior of $f \circ g$ near a depends on the behavior of f near $g(a)$ (not near a), it also seems reasonable to assume that f is differentiable at $g(a)$. Indeed we shall prove that if g is differentiable at a and f is differentiable at $g(a)$, then $f \circ g$ is differentiable at a , and

$$(f \circ g)'(a) = f'(g(a)) \cdot g'(a).$$

This extremely important formula is called the *Chain Rule*, presumably because a composition of functions might be called a “chain” of functions. Notice that $(f \circ g)'$ is practically the product of f' and g' , but not quite: f' must be evaluated at $g(a)$ and g' at a . Before attempting to prove this theorem we will try a few applications. Suppose

$$f(x) = \sin x^2.$$

Let us, temporarily, use S to denote the (“squaring”) function $S(x) = x^2$. Then

$$f = \sin \circ S.$$

Therefore we have

$$\begin{aligned} f'(x) &= \sin'(S(x)) \cdot S'(x) \\ &= \cos x^2 \cdot 2x. \end{aligned}$$

Quite a different result is obtained if

$$f(x) = \sin^2 x.$$

In this case

$$f = S \circ \sin,$$

so

$$\begin{aligned} f'(x) &= S'(\sin x) \cdot \sin'(x) \\ &= 2 \sin x \cdot \cos x. \end{aligned}$$

Notice that this agrees (as it should) with the result obtained by writing $f = \sin \cdot \sin$ and using the product formula.

Although we have invented a special symbol, S , to name the “squaring” function, it does not take much practice to do problems like this without bothering to write down special symbols for functions, and without even bothering to write down the particular composition which f is—one soon becomes accustomed to taking f apart in one’s head. The following differentiations may be used as practice for such mental gymnastics—if you find it necessary to work a few out on paper, by all means do so, but try to develop the knack of writing f' immediately after seeing

the definition of f ; problems of this sort are so simple that, if you just remember the Chain Rule, there is no thought necessary.

if $f(x) = \sin x^3$	then $f'(x) = \cos x^3 \cdot 3x^2$
$f(x) = \sin^3 x$	$f'(x) = 3 \sin^2 x \cdot \cos x$
$f(x) = \sin \frac{1}{x}$	$f'(x) = \cos \frac{1}{x} \cdot \left(\frac{-1}{x^2} \right)$
$f(x) = \sin(\sin x)$	$f'(x) = \cos(\sin x) \cdot \cos x$
$f(x) = \sin(x^3 + 3x^2)$	$f'(x) = \cos(x^3 + 3x^2) \cdot (3x^2 + 6x)$
$f(x) = (x^3 + 3x^2)^{53}$	$f'(x) = 53(x^3 + 3x^2)^{52} \cdot (3x^2 + 6x)$.

A function like

$$f(x) = \sin^2 x^2 = [\sin x^2]^2,$$

which is the composition of three functions,

$$f = S \circ \sin \circ S,$$

can also be differentiated by the Chain Rule. It is only necessary to remember that a triple composition $f \circ g \circ h$ means $(f \circ g) \circ h$ or $f \circ (g \circ h)$. Thus if

$$f(x) = \sin^2 x^2$$

we can write

$$\begin{aligned} f &= (S \circ \sin) \circ S, \\ f &= S \circ (\sin \circ S). \end{aligned}$$

The derivative of either expression can be found by applying the Chain Rule twice; the only doubtful point is whether the two expressions lead to equally simple calculations. As a matter of fact, as any experienced differentiator knows, it is much better to use the second:

$$f = S \circ (\sin \circ S).$$

We can now write down $f'(x)$ in one fell swoop. To begin with, note that the first function to be differentiated is S , so the formula for $f'(x)$ begins

$$f'(x) = 2(\quad) \cdot \quad.$$

Inside the parentheses we must put $\sin x^2$, the value at x of the second function, $\sin \circ S$. Thus we begin by writing

$$f'(x) = 2 \sin x^2 \cdot \quad$$

(the parentheses weren't really necessary, after all). We must now multiply this much of the answer by the derivative of $\sin \circ S$ at x ; this part is easy—it involves a composition of two functions, which we already know how to handle. We obtain, for the final answer,

$$f'(x) = 2 \sin x^2 \cdot \cos x^2 \cdot 2x.$$

The following example is handled similarly. Suppose

$$f(x) = \sin(\sin x^2).$$

Without even bothering to write down f as a composition $g \circ h \circ k$ of three functions, we can see that the left-most one will be \sin , so our expression for $f'(x)$ begins

$$f'(x) = \cos(\quad) \cdot \quad.$$

Inside the parentheses we must put the value of $h \circ k(x)$; this is simply $\sin x^2$ (what you get from $\sin(\sin x^2)$ by deleting the first \sin). So our expression for $f'(x)$ begins

$$f'(x) = \cos(\sin x^2) \cdot \quad.$$

We can now forget about the first \sin in $\sin(\sin x^2)$; we have to multiply what we have so far by the derivative of the function whose value at x is $\sin x^2$ —which is again a problem we already know how to solve:

$$f'(x) = \cos(\sin x^2) \cdot \cos x^2 \cdot 2x.$$

Finally, here are the derivatives of some other functions which are the composition of \sin and S , as well as some other triple compositions. You can probably just “see” that the answers are correct—if not, try writing out f as a composition:

if $f(x) = \sin((\sin x)^2)$	then $f'(x) = \cos((\sin x)^2) \cdot 2 \sin x \cdot \cos x$
$f(x) = [\sin(\sin x)]^2$	$f'(x) = 2 \sin(\sin x) \cdot \cos(\sin x) \cdot \cos x$
$f(x) = \sin(\sin(\sin x))$	$f'(x) = \cos(\sin(\sin x)) \cdot \cos(\sin x) \cdot \cos x$
$f(x) = \sin^2(x \sin x)$	$f'(x) = 2 \sin(x \sin x) \cdot \cos(x \sin x)$ · $[\sin x + x \cos x]$
$f(x) = \sin(\sin(x^2 \sin x))$	$f'(x) = \cos(\sin(x^2 \sin x)) \cdot \cos(x^2 \sin x)$ · $[2x \sin x + x^2 \cos x]$.

The rule for treating compositions of four (or even more) functions is easy—always (mentally) put in parentheses starting from the right,

$$f \circ (g \circ (h \circ k)),$$

and start reducing the calculation to the derivative of a composition of a smaller number of functions:

$$f'(g(h(k(x)))) \cdot \quad.$$

For example, if

$$\begin{aligned} f(x) &= \sin^2(\sin^2(x)) & [f = S \circ \sin \circ S \circ \sin \\ && = S \circ (\sin \circ (S \circ \sin))] \end{aligned}$$

then

$$f'(x) = 2 \sin(\sin^2 x) \cdot \cos(\sin^2 x) \cdot 2 \sin x \cdot \cos x;$$

if

$$f(x) = \sin((\sin x^2)^2) \quad [f = \sin \circ S \circ \sin \circ S \\ = \sin \circ (S \circ (\sin \circ S))]$$

then

$$f'(x) = \cos((\sin x^2)^2) \cdot 2 \sin x^2 \cdot \cos x^2 \cdot 2x;$$

if

$$f(x) = \sin^2(\sin(\sin x)) \quad [\text{fill in yourself, if necessary}]$$

then

$$f'(x) = 2 \sin(\sin(\sin x)) \cdot \cos(\sin(\sin x)) \cdot \cos(\sin x) \cdot \cos x.$$

With these examples as reference, you require only one thing to become a master differentiator—practice. You can be safely turned loose on the exercises at the end of the chapter, and it is now high time that we proved the Chain Rule.

The following argument, while not a proof, indicates some of the tricks one might try, as well as some of the difficulties encountered. We begin, of course, with the definition—

$$(f \circ g)'(a) = \lim_{h \rightarrow 0} \frac{(f \circ g)(a + h) - (f \circ g)(a)}{h} \\ = \lim_{h \rightarrow 0} \frac{f(g(a + h)) - f(g(a))}{h}.$$

Somewhere in here we would like the expression for $g'(a)$. One approach is to put it in by fiat:

$$\lim_{h \rightarrow 0} \frac{f(g(a + h)) - f(g(a))}{h} = \lim_{h \rightarrow 0} \frac{f(g(a + h)) - f(g(a))}{g(a + h) - g(a)} \cdot \frac{g(a + h) - g(a)}{h}.$$

This does not look bad, and it looks even better if we write

$$\lim_{h \rightarrow 0} \frac{(f \circ g)(a + h) - (f \circ g)(a)}{h} \\ = \lim_{h \rightarrow 0} \frac{f(g(a) + [g(a + h) - g(a)]) - f(g(a))}{g(a + h) - g(a)} \cdot \lim_{h \rightarrow 0} \frac{g(a + h) - g(a)}{h}.$$

The second limit is the factor $g'(a)$ which we want. If we let $g(a + h) - g(a) = k$ (to be precise we should write $k(h)$), then the first limit is

$$\lim_{h \rightarrow 0} \frac{f(g(a) + k) - f(g(a))}{k}.$$

It looks as if this limit should be $f'(g(a))$, since continuity of g at a implies that k goes to 0 as h does. In fact, one can, and we soon will, make this sort of reasoning precise. There is already a problem, however, which you will have noticed if you are the kind of person who does not divide blindly. Even for $h \neq 0$ we might have $g(a + h) - g(a) = 0$, making the division and multiplication by $g(a + h) - g(a)$ meaningless. True, we only care about small h , but $g(a + h) - g(a)$ could be 0 for arbitrarily small h . The easiest way this can happen is for g to be a constant

function, $g(x) = c$. Then $g(a+h) - g(a) = 0$ for all h . In this case, $f \circ g$ is also a constant function, $(f \circ g)(x) = f(c)$, so the Chain Rule does indeed hold:

$$(f \circ g)'(a) = 0 = f'(g(a)) \cdot g'(a).$$

However, there are also nonconstant functions g for which $g(a+h) - g(a) = 0$ for arbitrarily small h . For example, if $a = 0$, the function g might be

$$g(x) = \begin{cases} x^2 \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

In this case, $g'(0) = 0$, as we showed in Chapter 9. If the Chain Rule is correct, we must have $(f \circ g)'(0) = 0$ for any differentiable f , and this is not exactly obvious. A proof of the Chain Rule can be found by considering such recalcitrant functions separately, but it is easier simply to abandon this approach, and use a trick.

THEOREM 9 (THE CHAIN RULE)

If g is differentiable at a , and f is differentiable at $g(a)$, then $f \circ g$ is differentiable at a , and

$$(f \circ g)'(a) = f'(g(a)) \cdot g'(a).$$

PROOF

Define a function ϕ as follows:

$$\phi(h) = \begin{cases} \frac{f(g(a+h)) - f(g(a))}{g(a+h) - g(a)}, & \text{if } g(a+h) - g(a) \neq 0 \\ f'(g(a)), & \text{if } g(a+h) - g(a) = 0. \end{cases}$$

It should be intuitively clear that ϕ is continuous at 0: When h is small, $g(a+h) - g(a)$ is also small, so if $g(a+h) - g(a)$ is not zero, then $\phi(h)$ will be close to $f'(g(a))$; and if it is zero, then $\phi(h)$ actually equals $f'(g(a))$, which is even better. Since the continuity of ϕ is the crux of the whole proof we will provide a careful translation of this intuitive argument.

We know that f is differentiable at $g(a)$. This means that

$$\lim_{k \rightarrow 0} \frac{f(g(a)+k) - f(g(a))}{k} = f'(g(a)).$$

Thus, if $\varepsilon > 0$ there is some number $\delta' > 0$ such that, for all k ,

$$(1) \quad \text{if } 0 < |k| < \delta', \text{ then } \left| \frac{f(g(a)+k) - f(g(a))}{k} - f'(g(a)) \right| < \varepsilon.$$

Now g is differentiable at a , hence continuous at a , so there is a $\delta > 0$ such that, for all h ,

$$(2) \quad \text{if } |h| < \delta, \text{ then } |g(a+h) - g(a)| < \delta'.$$

Consider now any h with $|h| < \delta$. If $k = g(a+h) - g(a) \neq 0$, then

$$\phi(h) = \frac{f(g(a+h)) - f(g(a))}{g(a+h) - g(a)} = \frac{f(g(a)+k) - f(g(a))}{k};$$

it follows from (2) that $|k| < \delta'$, and hence from (1) that

$$|\phi(h) - f'(g(a))| < \varepsilon.$$

On the other hand, if $g(a+h) - g(a) = 0$, then $\phi(h) = f'(g(a))$, so it is surely true that

$$|\phi(h) - f'(g(a))| < \varepsilon.$$

We have therefore proved that

$$\lim_{h \rightarrow 0} \phi(h) = f'(g(a)),$$

so ϕ is continuous at 0. The rest of the proof is easy. If $h \neq 0$, then we have

$$\frac{f(g(a+h) - f(g(a))}{h} = \phi(h) \cdot \frac{g(a+h) - g(a)}{h}$$

even if $g(a+h) - g(a) = 0$ (because in that case both sides are 0). Therefore

$$\begin{aligned} (f \circ g)'(a) &= \lim_{h \rightarrow 0} \frac{f(g(a+h) - f(g(a))}{h} = \lim_{h \rightarrow 0} \phi(h) \cdot \lim_{h \rightarrow 0} \frac{g(a+h) - g(a)}{h} \\ &= f'(g(a)) \cdot g'(a). \blacksquare \end{aligned}$$

Now that we can differentiate so many functions so easily we can take another look at the function

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

In Chapter 9 we showed that $f'(0) = 0$, working straight from the definition (the only possible way). For $x \neq 0$ we can use the methods of this chapter. We have

$$f'(x) = 2x \sin \frac{1}{x} + x^2 \cos \frac{1}{x} \cdot \left(-\frac{1}{x^2}\right);$$

Thus

$$f'(x) = \begin{cases} 2x \sin \frac{1}{x} - \cos \frac{1}{x}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

As this formula reveals, the first derivative f' is indeed badly behaved at 0—it is not even continuous there. If we consider instead

$$f(x) = \begin{cases} x^3 \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0, \end{cases}$$

then

$$f'(x) = \begin{cases} 3x^2 \sin \frac{1}{x} - x \cos \frac{1}{x}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

In this case f' is continuous at 0, but $f''(0)$ does not exist (because the expression $3x^2 \sin 1/x$ defines a function which is differentiable at 0 but the expression $-x \cos 1/x$ does not).

As you may suspect, increasing the power of x yet again produces another improvement. If

$$f(x) = \begin{cases} x^4 \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0, \end{cases}$$

then

$$f'(x) = \begin{cases} 4x^3 \sin \frac{1}{x} - x^2 \cos \frac{1}{x}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

It is easy to compute, right from the definition, that $(f')'(0) = 0$, and $f''(x)$ is easy to find for $x \neq 0$:

$$f''(x) = \begin{cases} 12x^2 \sin \frac{1}{x} - 4x \cos \frac{1}{x} - 2x \cos \frac{1}{x} - \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

In this case, the *second* derivative f'' is not continuous at 0. By now you may have guessed the pattern, which two of the problems ask you to establish: if

$$f(x) = \begin{cases} x^{2n} \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0, \end{cases}$$

then $f'(0), \dots, f^{(n)}(0)$ exist, but $f^{(n)}$ is not continuous at 0; if

$$f(x) = \begin{cases} x^{2n+1} \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0, \end{cases}$$

then $f'(0), \dots, f^{(n)}(0)$ exist, and $f^{(n)}$ is continuous at 0, but $f^{(n)}$ is not differentiable at 0. These examples may suggest that “reasonable” functions can be characterized by the possession of higher-order derivatives—no matter how hard we try to mask the infinite oscillation of $f(x) = \sin 1/x$, a derivative of sufficiently high order seems able to reveal the underlying irregularity. Unfortunately, we will see later that much worse things can happen.

After all these involved calculations, we will bring this chapter to a close with a minor remark. It is often tempting, and seems more elegant, to write some of the theorems in this chapter as equations about functions, rather than about their values. Thus Theorem 3 might be written

$$(f + g)' = f' + g',$$

Theorem 4 might be written as

$$(f \cdot g)' = f \cdot g' + f' \cdot g,$$

and Theorem 9 often appears in the form

$$(f \circ g)' = (f' \circ g) \cdot g'.$$

Strictly speaking, these equations may be false, because the functions on the left-hand side might have a larger domain than those on the right. Nevertheless, this is hardly worth worrying about. If f and g are differentiable everywhere in their domains, then these equations, and others like them, *are* true, and this is the only case any one cares about.

PROBLEMS

1. As a warm up exercise, find $f'(x)$ for each of the following f . (Don't worry about the domain of f or f' ; just get a formula for $f'(x)$ that gives the right answer when it makes sense.)
 - (i) $f(x) = \sin(x + x^2)$.
 - (ii) $f(x) = \sin x + \sin x^2$.
 - (iii) $f(x) = \sin(\cos x)$.
 - (iv) $f(x) = \sin(\sin x)$.
 - (v) $f(x) = \sin\left(\frac{\cos x}{x}\right)$.
 - (vi) $f(x) = \frac{\sin(\cos x)}{x}$.
 - (vii) $f(x) = \sin(x + \sin x)$.
 - (viii) $f(x) = \sin(\cos(\sin x))$.
2. Find $f'(x)$ for each of the following functions f . (It took the author 20 minutes to compute the derivatives for the answer section, and it should not take you much longer. Although rapid calculation is not the goal of mathematics, if you hope to treat theoretical applications of the Chain Rule with aplomb, these concrete applications should be child's play—mathematicians like to pretend that they can't even add, but most of them can when they have to.)
 - (i) $f(x) = \sin((x + 1)^2(x + 2))$.
 - (ii) $f(x) = \sin^3(x^2 + \sin x)$.
 - (iii) $f(x) = \sin^2((x + \sin x)^2)$.
 - (iv) $f(x) = \sin\left(\frac{x^3}{\cos x^3}\right)$.
 - (v) $f(x) = \sin(x \sin x) + \sin(\sin x^2)$.
 - (vi) $f(x) = (\cos x)^{31^2}$.
 - (vii) $f(x) = \sin^2 x \sin x^2 \sin^2 x^2$.
 - (viii) $f(x) = \sin^3(\sin^2(\sin x))$.
 - (ix) $f(x) = (x + \sin^5 x)^6$.
 - (x) $f(x) = \sin(\sin(\sin(\sin(\sin x))))$.
 - (xi) $f(x) = \sin((\sin^7 x^7 + 1)^7)$.
 - (xii) $f(x) = (((x^2 + x)^3 + x)^4 + x)^5$.
 - (xiii) $f(x) = \sin(x^2 + \sin(x^2 + \sin x^2))$.
 - (xiv) $f(x) = \sin(6 \cos(6 \sin(6 \cos 6x)))$.

$$(xv) \quad f(x) = \frac{\sin x^2 \sin^2 x}{1 + \sin x}.$$

$$(xvi) \quad f(x) = \frac{1}{x - \frac{x + \sin x}{2}}.$$

$$(xvii) \quad f(x) = \sin\left(\frac{x^3}{\sin\left(\frac{x^3}{\sin x}\right)}\right).$$

$$(xviii) \quad f(x) = \sin\left(\frac{x}{x - \sin\left(\frac{x}{x - \sin x}\right)}\right).$$

3. Find the derivatives of the functions tan, cotan, sec, cosec. (You don't have to memorize these formulas, although they will be needed once in a while; if you express your answers in the right way, they will be simple and somewhat symmetrical.)

4. For each of the following functions f , find $f'(f(x))$ (not $(f \circ f)'(x)$).

$$(i) \quad f(x) = \frac{1}{1+x}.$$

$$(ii) \quad f(x) = \sin x.$$

$$(iii) \quad f(x) = x^2.$$

$$(iv) \quad f(x) = 17.$$

5. For each of the following functions f , find $f(f'(x))$.

$$(i) \quad f(x) = \frac{1}{x}.$$

$$(ii) \quad f(x) = x^2.$$

$$(iii) \quad f(x) = 17.$$

$$(iv) \quad f(x) = 17x.$$

6. Find f' in terms of g' if

$$(i) \quad f(x) = g(x + g(a)).$$

$$(ii) \quad f(x) = g(x \cdot g(a)).$$

$$(iii) \quad f(x) = g(x + g(x)).$$

$$(iv) \quad f(x) = g(x)(x - a).$$

$$(v) \quad f(x) = g(a)(x - a).$$

$$(vi) \quad f(x + 3) = g(x^2).$$

7. (a) A circular object is increasing in size in some unspecified manner, but it is known that when the radius is 6, the rate of change of the radius is 4. Find the rate of change of the area when the radius is 6. (If $r(t)$ and $A(t)$ represent the radius and the area at time t , then the functions r and A satisfy $A = \pi r^2$; a straightforward use of the Chain Rule is called for.)

- (b) Suppose that we are now informed that the circular object we have been watching is really the cross section of a spherical object. Find the rate of change of the *volume* when the radius is 6. (You will clearly need to know a formula for the volume of a sphere; in case you have forgotten, the volume is $\frac{4}{3}\pi$ times the cube of the radius.)
- (c) Now suppose that the rate of change of the area of the circular cross section is 5 when the radius is 3. Find the rate of change of the volume when the radius is 3. You should be able to do this problem in two ways: first, by using the formulas for the area and volume in terms of the radius; and then by expressing the volume in terms of the area (to use this method you will need Problem 9-3).
8. The area between two varying concentric circles is at all times 9π in². The rate of change of the area of the larger circle is 10π in²/sec. How fast is the circumference of the smaller circle changing when it has area 16π in²?
9. Particle *A* moves along the positive horizontal axis, and particle *B* along the graph of $f(x) = -\sqrt{3x}$, $x \leq 0$. At a certain time, *A* is at the point (5,0) and moving with speed 3 units/sec; and *B* is at a distance of 3 units from the origin and moving with speed 4 units/sec. At what rate is the distance between *A* and *B* changing?
10. Let $f(x) = x^2 \sin 1/x$ for $x \neq 0$, and let $f(0) = 0$. Suppose also that *h* and *k* are two functions such that

$$\begin{aligned} h'(x) &= \sin^2(\sin(x+1)) & k'(x) &= f(x+1) \\ h(0) &= 3 & k(0) &= 0. \end{aligned}$$

Find

- (i) $(f \circ h)'(0)$.
- (ii) $(k \circ f)'(0)$.
- (iii) $\alpha'(x^2)$, where $\alpha(x) = h(x^2)$. Exercise great care.

11. Find $f'(0)$ if

$$f(x) = \begin{cases} g(x) \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0, \end{cases}$$

and

$$g(0) = g'(0) = 0.$$

12. Using the derivative of $f(x) = 1/x$, as found in Problem 9-1, find $(1/g)'(x)$ by the Chain Rule.
13. (a) Using Problem 9-3, find $f'(x)$ for $-1 < x < 1$, if $f(x) = \sqrt{1-x^2}$.
- (b) Prove that the tangent line to the graph of f at $(a, \sqrt{1-a^2})$ intersects the graph only at that point (and thus show that the elementary geometry definition of the tangent line coincides with ours).

14. Prove similarly that the tangent lines to an ellipse or hyperbola intersect these sets only once.
15. If $f + g$ is differentiable at a , are f and g necessarily differentiable at a ? If $f \cdot g$ and f are differentiable at a , what conditions on f imply that g is differentiable at a ?
16. (a) Prove that if f is differentiable at a , then $|f|$ is also differentiable at a , provided that $f(a) \neq 0$.
 (b) Give a counterexample if $f(a) = 0$.
 (c) Prove that if f and g are differentiable at a , then the functions $\max(f, g)$ and $\min(f, g)$ are differentiable at a , provided that $f(a) \neq g(a)$.
 (d) Give a counterexample if $f(a) = g(a)$.
17. If f is three times differentiable and $f'(x) \neq 0$, the *Schwarzian derivative* of f at x is defined to be

$$\mathcal{D}f(x) = \frac{f'''(x)}{f'(x)} - \frac{3}{2} \left(\frac{f''(x)}{f'(x)} \right)^2.$$

- (a) Show that

$$\mathcal{D}(f \circ g) = [\mathcal{D}f \circ g] \cdot g'^2 + \mathcal{D}g.$$

- (b) Show that if $f(x) = \frac{ax+b}{cx+d}$, with $ad - bc \neq 0$, then $\mathcal{D}f = 0$. Consequently, $\mathcal{D}(f \circ g) = \mathcal{D}g$.

18. Suppose that $f^{(n)}(a)$ and $g^{(n)}(a)$ exist. Prove *Leibniz's formula*:

$$(f \cdot g)^{(n)}(a) = \sum_{k=0}^n \binom{n}{k} f^{(k)}(a) \cdot g^{n-k}(a).$$

- *19. Prove that if $f^{(n)}(g(a))$ and $g^{(n)}(a)$ both exist, then $(f \circ g)^{(n)}(a)$ exists. A little experimentation should convince you that it is unwise to seek a formula for $(f \circ g)^{(n)}(a)$. In order to prove that $(f \circ g)^{(n)}(a)$ exists you will therefore have to devise a reasonable assertion about $(f \circ g)^{(n)}(a)$ which can be proved by induction. Try something like: “ $(f \circ g)^{(n)}(a)$ exists and is a sum of terms each of which is a product of terms of the form . . .”

20. (a) If $f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0$, find a function g such that $g' = f$. Find another.

- (b) If

$$f(x) = \frac{b_2}{x^2} + \frac{b_3}{x^3} + \cdots + \frac{b_m}{x^m},$$

find a function g with $g' = f$.

- (c) Is there a function

$$f(x) = a_n x^n + \cdots + a_0 + \frac{b_1}{x} + \cdots + \frac{b_m}{x^m}$$

such that $f'(x) = 1/x$?

21. Show that there is a polynomial function f of degree n such that
- $f'(x) = 0$ for precisely $n - 1$ numbers x .
 - $f'(x) = 0$ for no x , if n is odd.
 - $f'(x) = 0$ for exactly one x , if n is even.
 - $f'(x) = 0$ for exactly k numbers x , if $n - k$ is odd.
22. (a) The number a is called a **double root** of the polynomial function f if $f(x) = (x - a)^2 g(x)$ for some polynomial function g . Prove that a is a double root of f if and only if a is a root of both f and f' .
- (b) When does $f(x) = ax^2 + bx + c$ ($a \neq 0$) have a double root? What does the condition say geometrically?
23. If f is differentiable at a , let $d(x) = f(x) - f'(a)(x - a) - f(a)$. Find $d'(a)$. In connection with Problem 22, this gives another solution for Problem 9-20.
- *24. This problem is a companion to Problem 3-6. Let a_1, \dots, a_n and b_1, \dots, b_n be given numbers.
- If x_1, \dots, x_n are distinct numbers, prove that there is a polynomial function f of degree $2n - 1$, such that $f(x_j) = f'(x_j) = 0$ for $j \neq i$, and $f(x_i) = a_i$ and $f'(x_i) = b_i$. Hint: Remember Problem 22.
 - Prove that there is a polynomial function f of degree $2n - 1$ with $f(x_i) = a_i$ and $f'(x_i) = b_i$ for all i .
- *25. Suppose that a and b are two consecutive roots of a polynomial function f , but that a and b are not double roots, so that we can write $f(x) = (x - a)(x - b)g(x)$ where $g(a) \neq 0$ and $g(b) \neq 0$.
- Prove that $g(a)$ and $g(b)$ have the same sign. (Remember that a and b are consecutive roots.)
 - Prove that there is some number x with $a < x < b$ and $f'(x) = 0$. (Also draw a picture to illustrate this fact.) Hint: Compare the sign of $f'(a)$ and $f'(b)$.
 - Now prove the same fact, even if a and b are multiple roots. Hint: If $f(x) = (x - a)^m(x - b)^n g(x)$ where $g(a) \neq 0$ and $g(b) \neq 0$, consider the polynomial function $h(x) = f'(x)/(x - a)^{m-1}(x - b)^{n-1}$.
- This theorem was proved by the French mathematician Rolle, in connection with the problem of approximating roots of polynomials, but the result was not originally stated in terms of derivatives. In fact, Rolle was one of the mathematicians who never accepted the new notions of calculus. This was not such a pigheaded attitude, in view of the fact that for one hundred years no one could define limits in terms that did not verge on the mystic, but on the whole history has been particularly kind to Rolle; his name has become attached to a much more general result, to appear in the next chapter, which forms the basis for the most important theoretical results of calculus.
26. Suppose that $f(x) = xg(x)$ for some function g which is continuous at 0. Prove that f is differentiable at 0, and find $f'(0)$ in terms of g .

- *27. Suppose f is differentiable at 0, and that $f(0) = 0$. Prove that $f(x) = xg(x)$ for some function g which is continuous at 0. Hint: What happens if you try to write $g(x) = f(x)/x$?

28. If $f(x) = x^{-n}$ for n in \mathbf{N} , prove that

$$\begin{aligned}f^{(k)}(x) &= (-1)^k \frac{(n+k-1)!}{(n-1)!} x^{-n-k} \\&= (-1)^k k! \binom{n+k-1}{k-1} x^{-n-k}, \quad \text{for } x \neq 0.\end{aligned}$$

- *29. Prove that it is impossible to write $x = f(x)g(x)$ where f and g are differentiable and $f(0) = g(0) = 0$. Hint: Differentiate.

30. What is $f^{(k)}(x)$ if

(a) $f(x) = 1/(x-a)^n$?

*(b) $f(x) = 1/(x^2-1)^n$?

- *31. Let $f(x) = x^{2n} \sin 1/x$ if $x \neq 0$, and let $f(0) = 0$. Prove that $f'(0), \dots, f^{(n)}(0)$ exist, and that $f^{(n)}$ is not continuous at 0. (You will encounter the same basic difficulty as that in Problem 19.)

- *32. Let $f(x) = x^{2n+1} \sin 1/x$ if $x \neq 0$, and let $f(0) = 0$. Prove that $f'(0), \dots, f^{(n)}(0)$ exist, that $f^{(n)}$ is continuous at 0, and that $f^{(n)}$ is not differentiable at 0.

33. In Leibnizian notation the Chain Rule ought to read:

$$\frac{df(g(x))}{dx} = \frac{df(y)}{dy} \Big|_{y=g(x)} \cdot \frac{dg(x)}{dx}.$$

Instead, one usually finds the following statement: "Let $y = g(x)$ and $z = f(y)$. Then

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}.$$

Notice that the z in dz/dx denotes the composite function $f \circ g$, while the z in dz/dy denotes the function f ; it is also understood that dz/dy will be "an expression involving y ," and that in the final answer $g(x)$ must be substituted for y . In each of the following cases, find dz/dx by using this formula; then compare with Problem 1.

- (i) $z = \sin y, \quad y = x + x^2.$
- (ii) $z = \sin y, \quad y = \cos x.$
- (iii) $z = \cos u, \quad u = \sin x.$
- (iv) $z = \sin v, \quad v = \cos u, \quad u = \sin x.$

One aim in this chapter is to justify the time we have spent learning to find the derivative of a function. As we shall see, knowing just a little about f' tells us a lot about f . Extracting information about f from information about f' requires some difficult work, however, and we shall begin with the one theorem which is really easy.

This theorem is concerned with the maximum value of a function on an interval. Although we have used this term informally in Chapter 7, it is worthwhile to be precise, and also more general.

DEFINITION

Let f be a function and A a set of numbers contained in the domain of f . A point x in A is a **maximum point** for f on A if

$$f(x) \geq f(y) \quad \text{for every } y \text{ in } A.$$

The number $f(x)$ itself is called the **maximum value** of f on A (and we also say that f “has its maximum value on A at x ”).

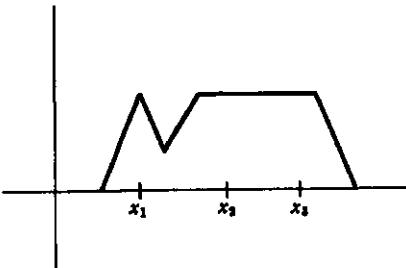


FIGURE 1

Notice that the maximum value of f on A could be $f(x)$ for several different x (Figure 1); in other words, a function f can have several different maximum points on A , although it can have at most one maximum value. Usually we shall be interested in the case where A is a closed interval $[a, b]$; if f is continuous, then Theorem 7-3 guarantees that f does indeed have a maximum value on $[a, b]$.

The definition of a minimum of f on A will be left to you. (One possible definition is the following: f has a minimum on A at x , if $-f$ has a maximum on A at x .)

We are now ready for a theorem which does not even depend upon the existence of least upper bounds.

THEOREM 1

Let f be any function defined on (a, b) . If x is a maximum (or a minimum) point for f on (a, b) , and f is differentiable at x , then $f'(x) = 0$.

(Notice that we do not assume differentiability, or even continuity, of f at other points.)

PROOF

Consider the case where f has a maximum at x . Figure 2 illustrates the simple idea behind the whole argument—secants drawn through points to the left of $(x, f(x))$ have slopes ≥ 0 , and secants drawn through points to the right of $(x, f(x))$ have slopes ≤ 0 . Analytically, this argument proceeds as follows.

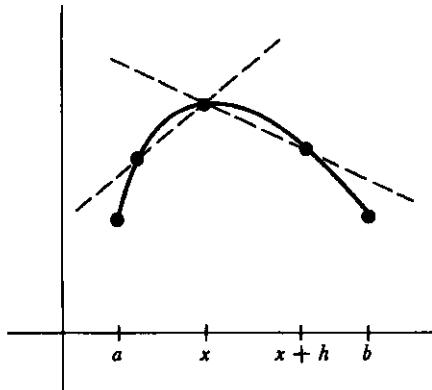


FIGURE 2

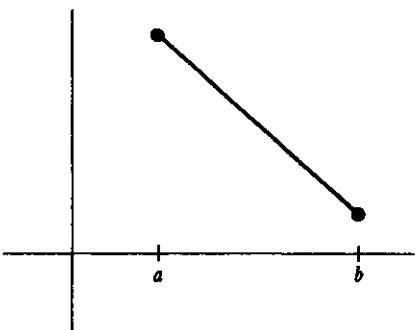


FIGURE 3

If h is any number such that $x + h$ is in (a, b) , then

$$f(x) \geq f(x + h),$$

since f has a maximum on (a, b) at x . This means that

$$f(x + h) - f(x) \leq 0.$$

Thus, if $h > 0$ we have

$$\frac{f(x + h) - f(x)}{h} \leq 0.$$

and consequently

$$\lim_{h \rightarrow 0^+} \frac{f(x + h) - f(x)}{h} \leq 0.$$

On the other hand, if $h < 0$, we have

$$\frac{f(x + h) - f(x)}{h} \geq 0,$$

so

$$\lim_{h \rightarrow 0^-} \frac{f(x + h) - f(x)}{h} \geq 0.$$

By hypothesis, f is differentiable at x , so these two limits must be equal, in fact equal to $f'(x)$. This means that

$$f'(x) \leq 0 \quad \text{and} \quad f'(x) \geq 0,$$

from which it follows that $f'(x) = 0$.

The case where f has a minimum at x is left to you (give a one-line proof). ■

Notice (Figure 3) that we cannot replace (a, b) by $[a, b]$ in the statement of the theorem (unless we add to the hypothesis the condition that x is in (a, b) .)

Since $f'(x)$ depends only on the values of f near x , it is almost obvious how to get a stronger version of Theorem 1. We begin with a definition which is illustrated in Figure 4.

DEFINITION

Let f be a function, and A a set of numbers contained in the domain of f . A point x in A is a **local maximum** [minimum] point for f on A if there is some $\delta > 0$ such that x is a maximum [minimum] point for f on $A \cap (x - \delta, x + \delta)$.

THEOREM 2 If f is defined on (a, b) and has a local maximum (or minimum) at x , and f is differentiable at x , then $f'(x) = 0$.

PROOF

You should see why this is an easy application of Theorem 1. ■

The converse of Theorem 2 is definitely not true—it is possible for $f'(x)$ to be 0 even if x is not a local maximum or minimum point for f . The simplest example is provided by the function $f(x) = x^3$; in this case $f'(0) = 0$, but f has no local maximum or minimum anywhere.

Probably the most widespread misconceptions about calculus are concerned with the behavior of a function f near x when $f'(x) = 0$. The point made in the previous paragraph is so quickly forgotten by those who want the world to be simpler than it is, that we will repeat it: the converse of Theorem 2 is *not* true—the condition $f'(x) = 0$ does *not* imply that x is a local maximum or minimum point of f . Precisely for this reason, special terminology has been adopted to describe numbers x which satisfy the condition $f'(x) = 0$.

DEFINITION

A **critical point** of a function f is a number x such that

$$f'(x) = 0.$$

The number $f(x)$ itself is called a **critical value** of f .

The critical values of f , together with a few other numbers, turn out to be the ones which must be considered in order to find the maximum and minimum of a given function f . To the uninitiated, finding the maximum and minimum value of a function represents one of the most intriguing aspects of calculus, and there is no denying that problems of this sort are fun (until you have done your first hundred or so).

Let us consider first the problem of finding the maximum or minimum of f on a closed interval $[a, b]$. (Then, if f is continuous, we can at least be sure that a maximum and minimum value exist.) In order to locate the maximum and minimum of f three kinds of points must be considered:

- (1) The critical points of f in $[a, b]$.
- (2) The end points a and b .
- (3) Points x in $[a, b]$ such that f is not differentiable at x .

If x is a maximum point or a minimum point for f on $[a, b]$, then x must be in one of the three classes listed above: for if x is not in the second or third group, then x is in (a, b) and f is differentiable at x ; consequently $f'(x) = 0$, by Theorem 1, and this means that x is in the first group.

If there are many points in these three categories, finding the maximum and minimum of f may still be a hopeless proposition, but when there are only a few critical points, and only a few points where f is not differentiable, the procedure is fairly straightforward: one simply finds $f(x)$ for each x satisfying $f'(x) = 0$, and $f(x)$ for each x such that f is not differentiable at x and, finally, $f(a)$ and $f(b)$. The biggest of these will be the maximum value of f , and the smallest will be the minimum. A simple example follows.

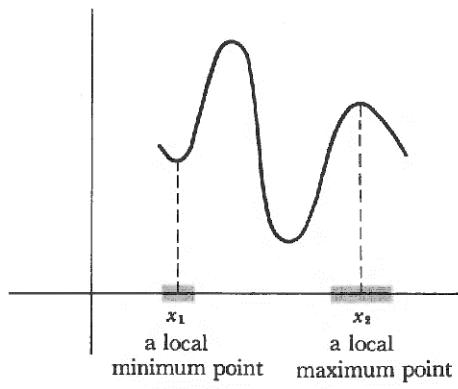


FIGURE 4

Suppose we wish to find the maximum and minimum value of the function

$$f(x) = x^3 - x$$

on the interval $[-1, 2]$. To begin with, we have

$$f'(x) = 3x^2 - 1,$$

so $f'(x) = 0$ when $3x^2 - 1 = 0$, that is, when

$$x = \sqrt{1/3} \quad \text{or} \quad -\sqrt{1/3}.$$

The numbers $\sqrt{1/3}$ and $-\sqrt{1/3}$ both lie in $[-1, 2]$, so the first group of candidates for the location of the maximum and the minimum is

$$(1) \quad \sqrt{1/3}, \quad -\sqrt{1/3}.$$

The second group contains the end points of the interval,

$$(2) \quad -1, \quad 2.$$

The third group is empty, since f is differentiable everywhere. The final step is to compute

$$f(\sqrt{1/3}) = (\sqrt{1/3})^3 - \sqrt{1/3} = \frac{1}{3}\sqrt{1/3} - \sqrt{1/3} = -\frac{2}{3}\sqrt{1/3},$$

$$f(-\sqrt{1/3}) = (-\sqrt{1/3})^3 - (-\sqrt{1/3}) = -\frac{1}{3}\sqrt{1/3} + \sqrt{1/3} = \frac{2}{3}\sqrt{1/3},$$

$$f(-1) = 0,$$

$$f(2) = 6.$$

Clearly the minimum value is $-\frac{2}{3}\sqrt{1/3}$, occurring at $\sqrt{1/3}$, and the maximum value is 6, occurring at 2.

This sort of procedure, if feasible, will always locate the maximum and minimum value of a continuous function on a closed interval. If the function we are dealing with is not continuous, however, or if we are seeking the maximum or minimum on an open interval or the whole line, then we cannot even be sure beforehand that the maximum and minimum values exist, so all the information obtained by this procedure may say nothing. Nevertheless, a little ingenuity will often reveal the nature of things. In Chapter 7 we solved just such a problem when we showed that if n is even, then the function

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$$

has a minimum value on the whole line. This proves that the minimum value must occur at some number x satisfying

$$0 = f'(x) = nx^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + a_1.$$

If we can solve this equation, and compare the values of $f(x)$ for such x , we can actually find the minimum of f . One more example may be helpful. Suppose we wish to find the maximum and minimum, if they exist, of the function

$$f(x) = \frac{1}{1-x^2}$$

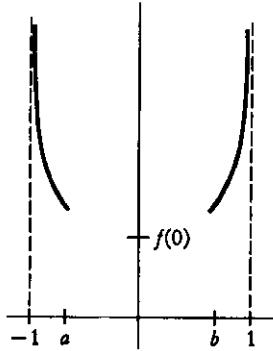


FIGURE 5

on the open interval $(-1, 1)$. We have

$$f'(x) = \frac{2x}{(1-x^2)^2}$$

so $f'(x) = 0$ only for $x = 0$. We can see immediately that for x close to 1 or -1 the values of $f(x)$ become arbitrarily large, so f certainly does not have a maximum. This observation also makes it easy to show that f has a minimum at 0. We just note (Figure 5) that there will be numbers a and b , with

$$-1 < a < 0 \quad \text{and} \quad 0 < b < 1,$$

such that $f(x) > f(0)$ for

$$-1 < x \leq a \quad \text{and} \quad b \leq x < 1.$$

This means that the minimum of f on $[a, b]$ is the minimum of f on all of $(-1, 1)$. Now on $[a, b]$ the minimum occurs either at 0 (the only place where $f' = 0$), or at a or b , and a and b have already been ruled out, so the minimum value is $f(0) = 1$.

In solving these problems we purposely did not draw the graphs of $f(x) = x^3 - x$ and $f(x) = 1/(1-x^2)$, but it is not cheating to draw the graph (Figure 6) as long as you do not rely solely on your picture to prove anything. As a matter of fact, we are now going to discuss a method of sketching the graph of a function that really gives enough information to be used in discussing maxima and minima—in fact we will be able to locate even *local* maxima and minima. This method involves consideration of the sign of $f'(x)$, and relies on some deep theorems.

The theorems about derivatives which have been proved so far, always yield information about f' in terms of information about f . This is true even of Theorem 1, although this theorem can sometimes be used to determine certain information about f , namely, the location of maxima and minima. When the derivative was first introduced, we emphasized that $f'(x)$ is not $[f(x+h) - f(x)]/h$ for any particular h , but only a limit of these numbers as h approaches 0; this fact becomes painfully relevant when one tries to extract information about f from information about f' . The simplest and most frustrating illustration of the difficulties encountered is afforded by the following question: If $f'(x) = 0$ for all x , must f be a constant function? It is impossible to imagine how f could be anything else, and this conviction is strengthened by considering the physical interpretation—if the velocity of a particle is always 0, surely the particle must be standing still! Nevertheless it is difficult even to begin a proof that only the constant functions satisfy $f'(x) = 0$ for all x . The hypothesis $f'(x) = 0$ only means that

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = 0,$$

and it is not at all obvious how one can use the information about the limit to derive information about the function.

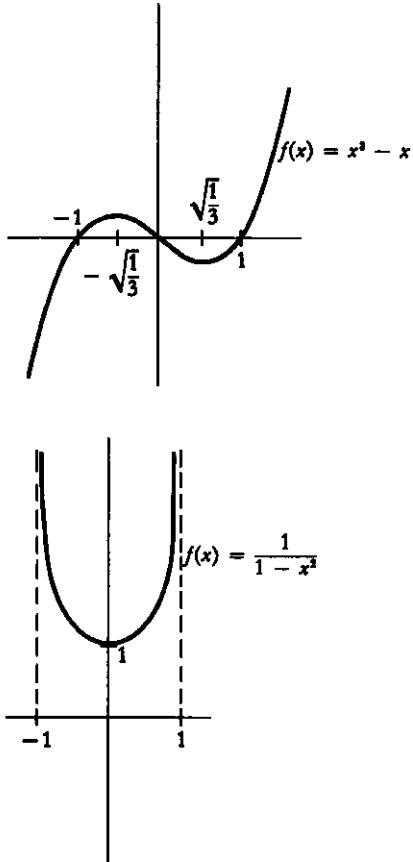


FIGURE 6

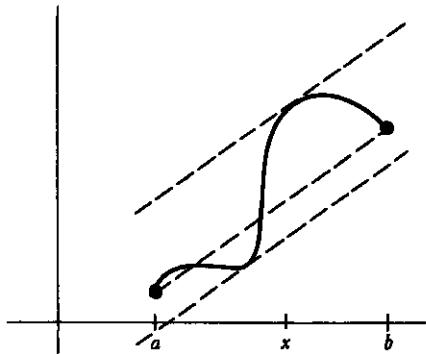


FIGURE 7

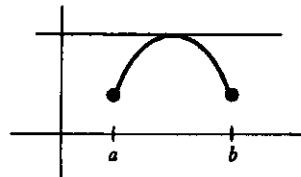


FIGURE 8

The fact that f is a constant function if $f'(x) = 0$ for all x , and many other facts of the same sort, can all be derived from a fundamental theorem, called the Mean Value Theorem, which states much stronger results. Figure 7 makes it plausible that if f is differentiable on $[a, b]$, then there is some x in (a, b) such that

$$f'(x) = \frac{f(b) - f(a)}{b - a}.$$

Geometrically this means that some tangent line is parallel to the line between $(a, f(a))$ and $(b, f(b))$. The Mean Value Theorem asserts that this is true—there is some x in (a, b) such that $f'(x)$, the instantaneous rate of change of f at x , is exactly equal to the average or “mean” change of f on $[a, b]$, this average change being $[f(b) - f(a)]/[b - a]$. (For example, if you travel 60 miles in one hour, then at some time you must have been traveling exactly 60 miles per hour.) This theorem is one of the most important theoretical tools of calculus—probably the deepest result about derivatives. From this statement you might conclude that the proof is difficult, but there you would be wrong—the hard theorems in this book have occurred long ago, in Chapter 7. It is true that if you try to prove the Mean Value Theorem yourself you will probably fail, but this is neither evidence that the theorem is hard, nor something to be ashamed of. The first proof of the theorem was an achievement, but today we can supply a proof which is quite simple. It helps to begin with a very special case.

THEOREM 3 (ROLLE'S THEOREM)

If f is continuous on $[a, b]$ and differentiable on (a, b) , and $f(a) = f(b)$, then there is a number x in (a, b) such that $f'(x) = 0$.

PROOF

It follows from the continuity of f on $[a, b]$ that f has a maximum and a minimum value on $[a, b]$.

Suppose first that the maximum value occurs at a point x in (a, b) . Then $f'(x) = 0$ by Theorem 1, and we are done (Figure 8).

Suppose next that the minimum value of f occurs at some point x in (a, b) . Then, again, $f'(x) = 0$ by Theorem 1 (Figure 9).

Finally, suppose the maximum and minimum values both occur at the end points. Since $f(a) = f(b)$, the maximum and minimum values of f are equal, so f is a constant function (Figure 10), and for a constant function we can choose any x in (a, b) . ■

Notice that we really needed the hypothesis that f is differentiable everywhere on (a, b) in order to apply Theorem 1. Without this assumption the theorem is false (Figure 11).

You may wonder why a special name should be attached to a theorem as easily proved as Rolle's Theorem. The reason is, that although Rolle's Theorem is a special case of the Mean Value Theorem, it also yields a simple proof of the Mean Value Theorem. In order to prove the Mean Value Theorem we will apply Rolle's Theorem to the function which gives the length of the vertical segment shown in

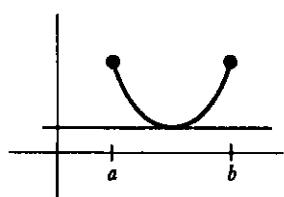


FIGURE 9

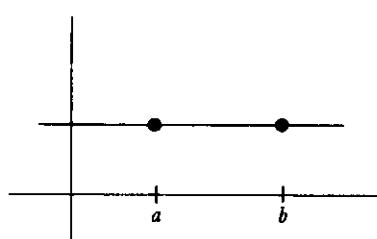


FIGURE 10

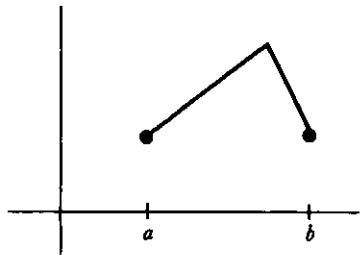


FIGURE 11

Figure 12; this is the difference between $f(x)$, and the height at x of the line L between $(a, f(a))$ and $(b, f(b))$. Since L is the graph of

$$g(x) = \left[\frac{f(b) - f(a)}{b - a} \right] (x - a) + f(a),$$

we want to look at

$$f(x) - \left[\frac{f(b) - f(a)}{b - a} \right] (x - a) - f(a).$$

As it turns out, the constant $f(a)$ is irrelevant.

THEOREM 4 (THE MEAN VALUE THEOREM)

If f is continuous on $[a, b]$ and differentiable on (a, b) , then there is a number x in (a, b) such that

$$f'(x) = \frac{f(b) - f(a)}{b - a}.$$

PROOF Let

$$h(x) = f(x) - \left[\frac{f(b) - f(a)}{b - a} \right] (x - a).$$

Clearly, h is continuous on $[a, b]$ and differentiable on (a, b) , and

$$\begin{aligned} h(a) &= f(a), \\ h(b) &= f(b) - \left[\frac{f(b) - f(a)}{b - a} \right] (x - a) \\ &= f(a). \end{aligned}$$

Consequently, we may apply Rolle's Theorem to h and conclude that there is some x in (a, b) such that

$$0 = h'(x) = f'(x) - \frac{f(b) - f(a)}{b - a},$$

so that

$$f'(x) = \frac{f(b) - f(a)}{b - a}. \blacksquare$$

Notice that the Mean Value Theorem still fits into the pattern exhibited by previous theorems—information about f yields information about f' . This information is so strong, however, that we can now go in the other direction.

COROLLARY 1

If f is defined on an interval and $f'(x) = 0$ for all x in the interval, then f is constant on the interval.

PROOF

Let a and b be any two points in the interval with $a \neq b$. Then there is some x in

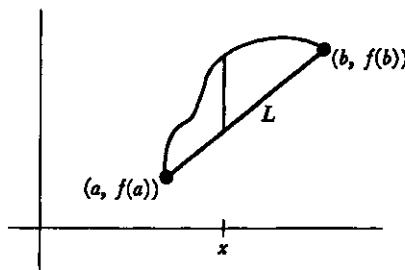


FIGURE 12

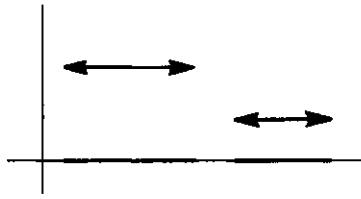


FIGURE 13

(a, b) such that

$$f'(x) = \frac{f(b) - f(a)}{b - a}.$$

But $f'(x) = 0$ for all x in the interval, so

$$0 = \frac{f(b) - f(a)}{b - a},$$

and consequently $f(a) = f(b)$. Thus the value of f at any two points in the interval is the same, i.e., f is constant on the interval. ■

Naturally, Corollary 1 does not hold for functions defined on two or more intervals (Figure 13).

COROLLARY 2 If f and g are defined on the same interval, and $f'(x) = g'(x)$ for all x in the interval, then there is some number c such that $f = g + c$.

PROOF For all x in the interval we have $(f - g)'(x) = f'(x) - g'(x) = 0$ so, by Corollary 1, there is a number c such that $f - g = c$. ■

The statement of the next corollary requires some terminology, which is illustrated in Figure 14.

DEFINITION

A function is **increasing** on an interval if $f(a) < f(b)$ whenever a and b are two numbers in the interval with $a < b$. The function f is **decreasing** on an interval if $f(a) > f(b)$ for all a and b in the interval with $a < b$. (We often say simply that f is increasing or decreasing, in which case the interval is understood to be the domain of f .)

COROLLARY 3

If $f'(x) > 0$ for all x in an interval, then f is increasing on the interval; if $f'(x) < 0$ for all x in the interval, then f is decreasing on the interval.

PROOF

Consider the case where $f'(x) > 0$. Let a and b be two points in the interval with $a < b$. Then there is some x in (a, b) with

$$f'(x) = \frac{f(b) - f(a)}{b - a}.$$

But $f'(x) > 0$ for all x in (a, b) , so

$$\frac{f(b) - f(a)}{b - a} > 0.$$

Since $b - a > 0$ it follows that $f(b) > f(a)$.

The proof when $f'(x) < 0$ for all x is left to you. ■

Notice that although the converses of Corollary 1 and Corollary 2 are true (and obvious), the converse of Corollary 3 is not true. If f is increasing, it is easy to see that $f'(x) \geq 0$ for all x , but the equality sign might hold for some x (consider $f(x) = x^3$).

Corollary 3 provides enough information to get a good idea of the graph of a function with a minimal amount of point plotting. Consider, once more, the function $f(x) = x^3 - x$. We have

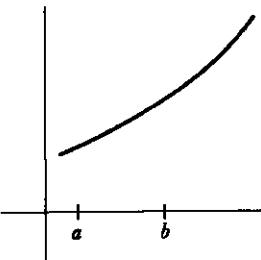
$$f'(x) = 3x^2 - 1.$$

We have already noted that $f'(x) = 0$ for $x = \sqrt{1/3}$ and $x = -\sqrt{1/3}$, and it is also possible to determine the sign of $f'(x)$ for all other x . Note that $3x^2 - 1 > 0$ precisely when

$$\begin{aligned} 3x^2 &> 1 \\ x^2 &> \frac{1}{3}, \\ x &> \sqrt{1/3} \quad \text{or} \quad x < -\sqrt{1/3}; \end{aligned}$$

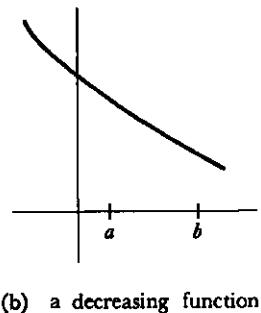
thus $3x^2 - 1 < 0$ precisely when

(a) an increasing function



$$-\sqrt{1/3} < x < \sqrt{1/3}.$$

Thus f is increasing for $x < -\sqrt{1/3}$, decreasing between $-\sqrt{1/3}$ and $\sqrt{1/3}$, and once again increasing for $x > \sqrt{1/3}$. Combining this information with the following facts



(b) a decreasing function

$$(1) f(-\sqrt{1/3}) = \frac{2}{3}\sqrt{1/3},$$

$$f(\sqrt{1/3}) = -\frac{2}{3}\sqrt{1/3},$$

$$(2) f(x) = 0 \text{ for } x = -1, 0, 1,$$

(3) $f(x)$ gets large as x gets large, and large negative as x gets large negative,

it is possible to sketch a pretty respectable approximation to the graph (Figure 15).

By the way, notice that the intervals on which f increases and decreases could have been found without even bothering to examine the sign of f' . For example, since f' is continuous, and vanishes only at $-\sqrt{1/3}$ and $\sqrt{1/3}$, we know that f' always has the same sign on the interval $(-\sqrt{1/3}, \sqrt{1/3})$. Since $f(-\sqrt{1/3}) > f(\sqrt{1/3})$, it follows that f decreases on this interval. Similarly, f' always has the same sign on $(\sqrt{1/3}, \infty)$ and $f(x)$ is large for large x , so f must be increasing on $(\sqrt{1/3}, \infty)$. Another point worth noting: If f' is continuous, then the sign of f' on the interval between two adjacent critical points can be determined simply by finding the sign of $f'(x)$ for any *one* x in this interval.

Our sketch of the graph of $f(x) = x^3 - x$ contains sufficient information to allow us to say with confidence that $-\sqrt{1/3}$ is a local maximum point, and $\sqrt{1/3}$ a local minimum point. In fact, we can give a general scheme for decid-

FIGURE 14

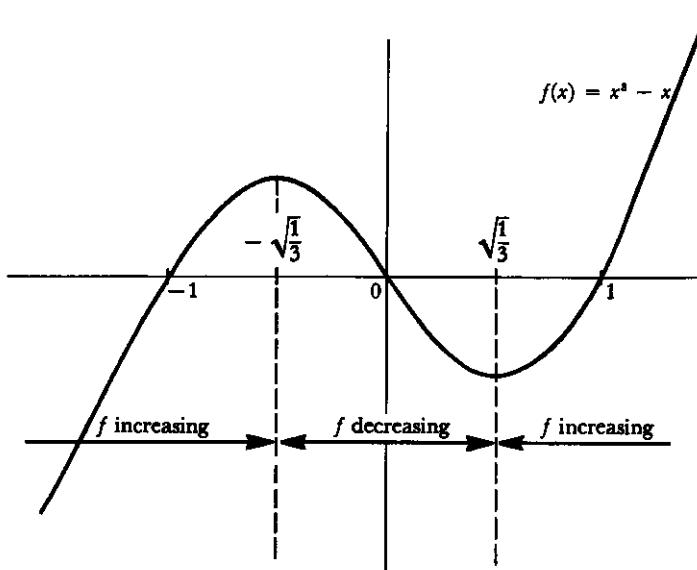


FIGURE 15

ing whether a critical point is a local maximum point, a local minimum point, or neither (Figure 16):

- (1) if $f' > 0$ in some interval to the left of x and $f' < 0$ in some interval to the right of x , then x is a local maximum point.
- (2) if $f' < 0$ in some interval to the left of x and $f' > 0$ in some interval to the right of x , then x is a local minimum point.
- (3) if f' has the same sign in some interval to the left of x as it has in some interval to the right, then x is neither a local maximum nor a local minimum point.

(There is no point in memorizing these rules—you can always draw the pictures yourself.)

The polynomial functions can all be analyzed in this way, and it is even possible to describe the general form of the graph of such functions. To begin, we need a

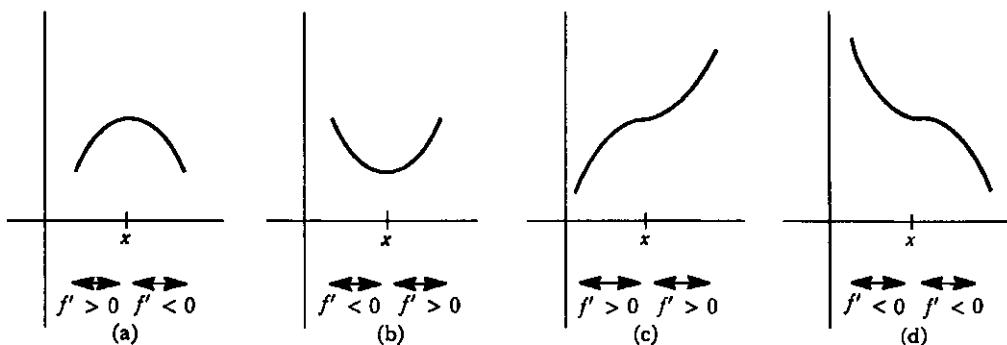


FIGURE 16

result already mentioned in Problem 3-7: If

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0,$$

then f has at most n “roots,” i.e., there are at most n numbers x such that $f(x) = 0$. Although this is really an algebraic theorem, calculus can be used to give an easy proof. Notice that if x_1 and x_2 are roots of f (Figure 17), so that $f(x_1) = f(x_2) = 0$, then by Rolle’s Theorem there is a number x between x_1 and x_2 such that $f'(x) = 0$. This means that if f has k different roots $x_1 < x_2 < \cdots < x_k$, then f' has at least $k - 1$ different roots: one between x_1 and x_2 , one between x_2 and x_3 , etc. It is now easy to prove by induction that a polynomial function

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0$$

has at most n roots: The statement is surely true for $n = 1$, and if we assume that it is true for n , then the polynomial

$$g(x) = b_{n+1} x^{n+1} + b_n x^n + \cdots + b_0$$

could not have more than $n + 1$ roots, since if it did, g' would have more than n roots.

With this information it is not hard to describe the graph of

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0.$$

The derivative, being a polynomial function of degree $n - 1$, has at most $n - 1$ roots. Therefore f has at most $n - 1$ critical points. Of course, a critical point is not necessarily a local maximum or minimum point, but at any rate, if a and b are adjacent critical points of f , then f' will remain either positive or negative on (a, b) , since f' is continuous; consequently, f will be either increasing or decreasing on (a, b) . Thus f has at most n regions of decrease or increase.

As a specific example, consider the function

$$f(x) = x^4 - 2x^2.$$

Since

$$f'(x) = 4x^3 - 4x = 4x(x - 1)(x + 1),$$

the critical points of f are $-1, 0$, and 1 , and

$$\begin{aligned} f(-1) &= -1, \\ f(0) &= 0, \\ f(1) &= -1. \end{aligned}$$

The behavior of f on the intervals between the critical points can be determined by one of the methods mentioned before. In particular, we could determine the sign of f' on these intervals simply by examining the formula for $f'(x)$. On the other hand, from the three critical values alone we can see (Figure 18) that f increases on $(-1, 0)$ and decreases on $(0, 1)$. To determine the sign of f' on

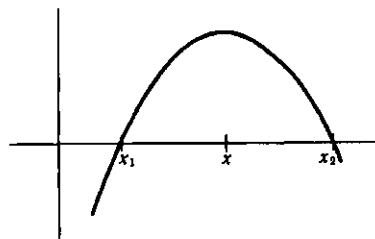


FIGURE 17

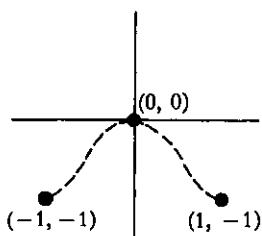


FIGURE 18

$(-\infty, -1)$ and $(1, \infty)$ we can compute

$$\begin{aligned}f'(-2) &= 4 \cdot (-2)^3 - 4 \cdot (-2) = -24, \\f'(2) &= 4 \cdot 2^3 - 4 \cdot 2 = 24,\end{aligned}$$

and conclude that f is decreasing on $(-\infty, -1)$ and increasing on $(1, \infty)$. These conclusions also follow from the fact that $f(x)$ is large for large x and for large negative x .

We can already produce a good sketch of the graph; two other pieces of information provide the finishing touches (Figure 19). First, it is easy to determine that $f(x) = 0$ for $x = 0, \pm\sqrt{2}$; second, it is clear that f is even, $f(x) = f(-x)$, so the graph is symmetric with respect to the vertical axis. The function $f(x) = x^3 - x$, already sketched in Figure 15, is odd, $f(x) = -f(-x)$, and is consequently symmetric with respect to the origin. Half the work of graph sketching may be saved by noticing these things in the beginning.

Several problems in this and succeeding chapters ask you to sketch the graphs of functions. In each case you should determine

- (1) the critical points of f ,
- (2) the value of f at the critical points,
- (3) the sign of f' in the regions between critical points (if this is not already clear),
- (4) the numbers x such that $f(x) = 0$ (if possible),
- (5) the behavior of $f(x)$ as x becomes large or large negative (if possible).

Finally, bear in mind that a quick check, to see whether the function is odd or even, may save a lot of work.

This sort of analysis, if performed with care, will usually reveal the basic shape of the graph, but sometimes there are special features which require a little more thought. It is impossible to anticipate all of these, but one piece of information is often very important. If f is not defined at certain points (for example, if f is a rational function whose denominator vanishes at some points), then the behavior of f near these points should be determined.

For example, consider the function

$$f(x) = \frac{x^2 - 2x + 2}{x - 1},$$

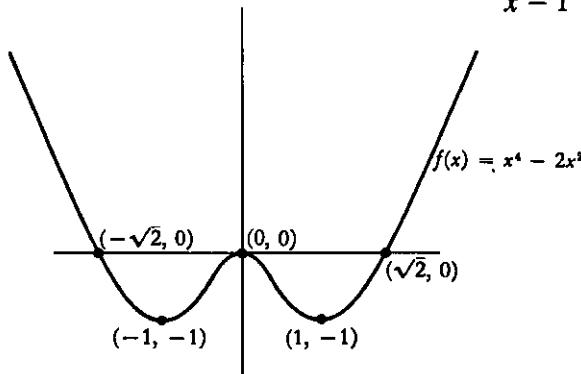


FIGURE 19

which is not defined at 1. We have

$$\begin{aligned} f'(x) &= \frac{(x-1)(2x-2) - (x^2 - 2x + 2)}{(x-1)^2} \\ &= \frac{x(x-2)}{(x-1)^2}. \end{aligned}$$

Thus

- (1) the critical points of f are 0, 2.

Moreover,

$$\begin{aligned} (2) \quad f(0) &= -2, \\ f(2) &= 2. \end{aligned}$$

Because f is not defined on the whole interval $(0, 2)$, the sign of f' must be determined separately on the intervals $(0, 1)$ and $(1, 2)$, as well as on the intervals $(-\infty, 0)$ and $(2, \infty)$. We can do this by picking particular points in each of these intervals, or simply by staring hard at the formula for f' . Either way we find that

$$\begin{aligned} (3) \quad f'(x) &> 0 \quad \text{if} \quad x < 0, \\ f'(x) &< 0 \quad \text{if} \quad 0 < x < 1, \\ f'(x) &< 0 \quad \text{if} \quad 1 < x < 2, \\ f'(x) &> 0 \quad \text{if} \quad 2 < x. \end{aligned}$$

Finally, we must determine the behavior of $f(x)$ as x becomes large or large negative, as well as when x approaches 1 (this information will also give us another way to determine the regions on which f increases and decreases). To examine the behavior as x becomes large we write

$$\frac{x^2 - 2x + 2}{x - 1} = x - 1 + \frac{1}{x-1};$$

clearly $f(x)$ is close to $x - 1$ (and slightly larger) when x is large, and $f(x)$ is close to $x - 1$ (but slightly smaller) when x is large negative. The behavior of f near 1 is also easy to determine; since

$$\lim_{x \rightarrow 1} (x^2 - 2x + 2) = 1 \neq 0,$$

the fraction

$$\frac{x^2 - 2x + 2}{x - 1}$$

becomes large as x approaches 1 from above and large negative as x approaches 1 from below.

All this information may seem a bit overwhelming, but there is only one way that it can be pieced together (Figure 20); be sure that you can account for each feature of the graph.

When this sketch has been completed, we might note that it looks like the graph

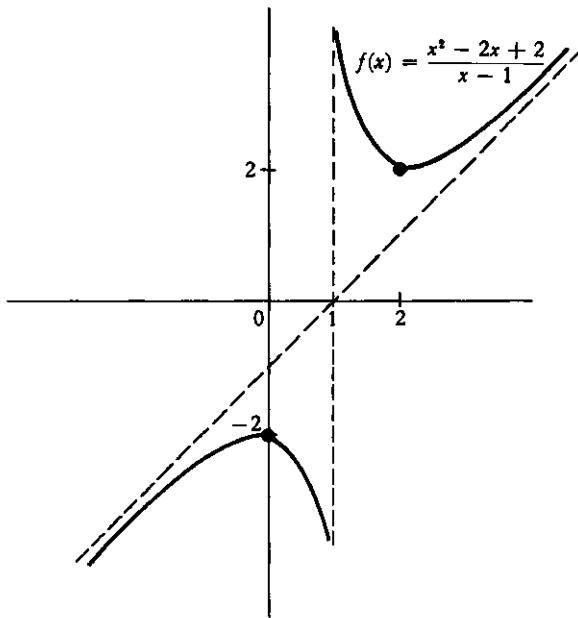


FIGURE 20

of an odd function shoved over 1 unit, and the expression

$$\frac{x^2 - 2x + 2}{x - 1} = \frac{(x - 1)^2 + 1}{x - 1}$$

shows that this is indeed the case. However, this is one of those special features which should be investigated only after you have used the other information to get a good idea of the appearance of the graph.

Although the location of local maxima and minima of a function is always revealed by a detailed sketch of its graph, it is usually unnecessary to do so much work. There is a popular test for local maxima and minima which depends on the behavior of the function only at its critical points.

THEOREM 5 Suppose $f'(a) = 0$. If $f''(a) > 0$, then f has a local minimum at a ; if $f''(a) < 0$, then f has a local maximum at a .

PROOF By definition,

$$f''(a) = \lim_{h \rightarrow 0} \frac{f'(a + h) - f'(a)}{h}.$$

Since $f'(a) = 0$, this can be written

$$f''(a) = \lim_{h \rightarrow 0} \frac{f'(a + h)}{h}.$$

Suppose now that $f''(a) > 0$. Then $f'(a + h)/h$ must be positive for sufficiently small h . Therefore:

$f'(a + h)$ must be positive for sufficiently small $h > 0$
and $f'(a + h)$ must be negative for sufficiently small $h < 0$.

This means (Corollary 3) that f is increasing in some interval to the right of a and f is decreasing in some interval to the left of a . Consequently, f has a local minimum at a .

The proof for the case $f''(a) < 0$ is similar. ■

Theorem 5 may be applied to the function $f(x) = x^3 - x$, which has already been considered. We have

$$\begin{aligned}f'(x) &= 3x^2 - 1 \\f''(x) &= 6x.\end{aligned}$$

At the critical points, $-\sqrt{1/3}$ and $\sqrt{1/3}$, we have

$$\begin{aligned}f''(-\sqrt{1/3}) &= -6\sqrt{1/3} < 0, \\f''(\sqrt{1/3}) &= 6\sqrt{1/3} > 0.\end{aligned}$$

Consequently, $-\sqrt{1/3}$ is a local maximum point and $\sqrt{1/3}$ is a local minimum point.

Although Theorem 5 will be found quite useful for polynomial functions, for many functions the second derivative is so complicated that it is easier to consider the sign of the first derivative. Moreover, if a is a critical point of f it may happen that $f''(a) = 0$. In this case, Theorem 5 provides no information: it is possible that a is a local maximum point, a local minimum point, or neither, as shown (Figure 21) by the functions

$$f(x) = -x^4, \quad f(x) = x^4, \quad f(x) = x^5;$$

in each case $f'(0) = f''(0) = 0$, but 0 is a local maximum point for the first, a local minimum point for the second, and neither a local maximum nor minimum point for the third. This point will be pursued further in Part IV.

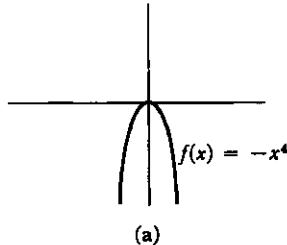
It is interesting to note that Theorem 5 automatically proves a partial converse of itself.

THEOREM 6 Suppose $f''(a)$ exists. If f has a local minimum at a , then $f''(a) \geq 0$; if f has a local maximum at a , then $f''(a) \leq 0$.

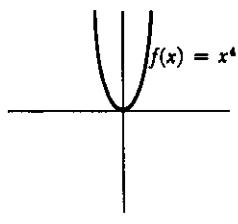
PROOF

Suppose f has local minimum at a . If $f''(a) < 0$, then f would also have a local maximum at a , by Theorem 5. Thus f would be constant in some interval containing a , so that $f''(a) = 0$, a contradiction. Thus we must have $f''(a) \geq 0$.

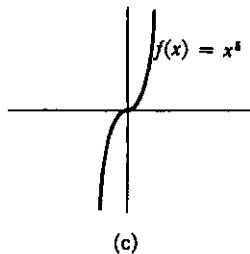
The case of a local maximum is handled similarly. ■



(a)



(b)



(c)

FIGURE 21

(This partial converse to Theorem 5 is the best we can hope for: the \geq and \leq signs cannot be replaced by $>$ and $<$, as shown by the functions $f(x) = x^4$ and $f(x) = -x^4$.)

The remainder of this chapter deals, not with graph sketching, or maxima and minima, but with three consequences of the Mean Value Theorem. The first is a simple, but very beautiful, theorem which plays an important role in Chapter 15, and which also sheds light on many examples which have occurred in previous chapters.

THEOREM 7 Suppose that f is continuous at a , and that $f'(x)$ exists for all x in some interval containing a , except perhaps for $x = a$. Suppose, moreover, that $\lim_{x \rightarrow a} f'(x)$ exists. Then $f'(a)$ also exists, and

$$f'(a) = \lim_{x \rightarrow a} f'(x).$$

PROOF By definition,

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}.$$

For sufficiently small $h > 0$ the function f will be continuous on $[a, a+h]$ and differentiable on $(a, a+h)$ (a similar assertion holds for sufficiently small $h < 0$). By the Mean Value Theorem there is a number α_h in $(a, a+h)$ such that

$$\frac{f(a+h) - f(a)}{h} = f'(\alpha_h).$$

Now α_h approaches a as h approaches 0, because α_h is in $(a, a+h)$; since $\lim_{x \rightarrow a} f'(x)$ exists, it follows that

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} = \lim_{h \rightarrow 0} f'(\alpha_h) = \lim_{x \rightarrow a} f'(x).$$

(It is a good idea to supply a rigorous ε - δ argument for this final step, which we have treated somewhat informally.) ■

Even if f is an everywhere differentiable function, it is still possible for f' to be discontinuous. This happens, for example, if

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

According to Theorem 7, however, the graph of f' can never exhibit a discontinuity of the type shown in Figure 22. Problem 55 outlines the proof of another beautiful theorem which gives further information about the function f' , and Problem 56 uses this result to strengthen Theorem 7.

The next theorem, a generalization of the Mean Value Theorem, is of interest mainly because of its applications.

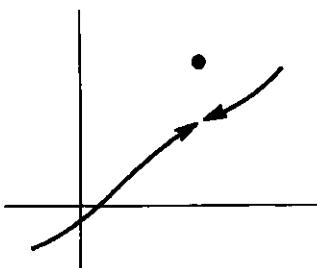


FIGURE 22

THEOREM 8 (THE CAUCHY MEAN VALUE THEOREM)

If f and g are continuous on $[a, b]$ and differentiable on (a, b) , then there is a number x in (a, b) such that

$$[f(b) - f(a)]g'(x) = [g(b) - g(a)]f'(x).$$

(If $g(b) \neq g(a)$, and $g'(x) \neq 0$, this equation can be written

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(x)}{g'(x)}.$$

Notice that if $g(x) = x$ for all x , then $g'(x) = 1$, and we obtain the Mean Value Theorem. On the other hand, applying the Mean Value Theorem to f and g separately, we find that there are x and y in (a, b) with

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(x)}{g'(y)},$$

but there is no guarantee that the x and y found in this way will be equal. These remarks may suggest that the Cauchy Mean Value Theorem will be quite difficult to prove, but actually the simplest of tricks suffices.)

PROOF

Let

$$h(x) = f(x)[g(b) - g(a)] - g(x)[f(b) - f(a)].$$

Then h is continuous on $[a, b]$, differentiable on (a, b) , and

$$h(a) = f(a)g(b) - g(a)f(b) = h(b).$$

It follows from Rolle's Theorem that $h'(x) = 0$ for some x in (a, b) , which means that

$$0 = f'(x)[g(b) - g(a)] - g'(x)[f(b) - f(a)]. \blacksquare$$

The Cauchy Mean Value Theorem is the basic tool needed to prove a theorem which facilitates evaluation of limits of the form

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)},$$

when

$$\lim_{x \rightarrow a} f(x) = 0 \quad \text{and} \quad \lim_{x \rightarrow a} g(x) = 0.$$

In this case, Theorem 5-2 is of no use. Every derivative is a limit of this form, and computing derivatives frequently requires a great deal of work. If some derivatives are known, however, many limits of this form can now be evaluated easily.

THEOREM 9 (L'HÔPITAL'S RULE)

Suppose that

$$\lim_{x \rightarrow a} f(x) = 0 \quad \text{and} \quad \lim_{x \rightarrow a} g(x) = 0,$$

and suppose also that $\lim_{x \rightarrow a} f'(x)/g'(x)$ exists. Then $\lim_{x \rightarrow a} f(x)/g(x)$ exists, and

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \lim_{x \rightarrow a} \frac{f'(x)}{g'(x)}.$$

(Notice that Theorem 7 is a special case.)

PROOF The hypothesis that $\lim_{x \rightarrow a} f'(x)/g'(x)$ exists contains two implicit assumptions:

- (1) there is an interval $(a - \delta, a + \delta)$ such that $f'(x)$ and $g'(x)$ exist for all x in $(a - \delta, a + \delta)$ except, perhaps, for $x = a$,
- (2) in this interval $g'(x) \neq 0$ with, once again, the possible exception of $x = a$.

On the other hand, f and g are not even assumed to be defined at a . If we define $f(a) = g(a) = 0$ (changing the previous values of $f(a)$ and $g(a)$, if necessary), then f and g are continuous at a . If $a < x < a + \delta$, then the Mean Value Theorem and the Cauchy Mean Value Theorem apply to f and g on the interval $[a, x]$ (and a similar statement holds for $a - \delta < x < a$). First applying the Mean Value Theorem to g , we see that $g(x) \neq 0$, for if $g(x) = 0$ there would be some x_1 in (a, x) with $g'(x_1) = 0$, contradicting (2). Now applying the Cauchy Mean Value Theorem to f and g , we see that there is a number α_x in (a, x) such that

$$[f(x) - 0]g'(\alpha_x) = [g(x) - 0]f'(\alpha_x)$$

or

$$\frac{f(x)}{g(x)} = \frac{f'(\alpha_x)}{g'(\alpha_x)}.$$

Now α_x approaches a as x approaches a , because α_x is in (a, x) ; since $\lim_{y \rightarrow a} f'(y)/g'(y)$ exists, it follows that

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \lim_{x \rightarrow a} \frac{f'(\alpha_x)}{g'(\alpha_x)} = \lim_{y \rightarrow a} \frac{f'(y)}{g'(y)}.$$

(Once again, the reader is invited to supply the details of this part of the argument.) ■

PROBLEMS

1. For each of the following functions, find the maximum and minimum values on the indicated intervals, by finding the points in the interval where the derivative is 0, and comparing the values at these points with the values at the end points.

- (i) $f(x) = x^3 - x^2 - 8x + 1$ on $[-2, 2]$.
- (ii) $f(x) = x^5 + x + 1$ on $[-1, 1]$.
- (iii) $f(x) = 3x^4 - 8x^3 + 6x^2$ on $[-\frac{1}{2}, \frac{1}{2}]$.
- (iv) $f(x) = \frac{1}{x^5 + x + 1}$ on $[-\frac{1}{2}, 1]$.
- (v) $f(x) = \frac{x+1}{x^2+1}$ on $[-1, \frac{1}{2}]$.
- (vi) $f(x) = \frac{x}{x^2-1}$ on $[0, 5]$.

2. Now sketch the graph of each of the functions in Problem 1, and find all local maximum and minimum points.

3. Sketch the graphs of the following functions.

$$(i) \quad f(x) = x + \frac{1}{x}.$$

$$(ii) \quad f(x) = x + \frac{3}{x^2}.$$

$$(iii) \quad f(x) = \frac{x^2}{x^2 - 1}.$$

$$(iv) \quad f(x) = \frac{1}{1+x^2}.$$

4. (a) If $a_1 < \dots < a_n$, find the minimum value of $f(x) = \sum_{i=1}^n (x - a_i)^2$.

- (b) Now find the minimum value of $f(x) = \sum_{i=1}^n |x - a_i|$. This is a problem where calculus won't help at all: on the intervals between the a_i 's the function f is linear, so that the minimum clearly occurs at one of the a_i , and these are precisely the points where f is not differentiable. However, the answer is easy to find if you consider how $f(x)$ changes as you pass from one such interval to another.

- (c) Let $a > 0$. Show that the maximum value of

$$f(x) = \frac{1}{1+|x|} + \frac{1}{1+|x-a|}$$

is $(2+a)/(1+a)$. (The derivative can be found on each of the intervals $(-\infty, 0)$, $(0, a)$, and (a, ∞) separately.)

5. For each of the following functions, find all local maximum and minimum points.

$$(i) \quad f(x) = \begin{cases} x, & x \neq 3, 5, 7, 9 \\ 5, & x = 3 \\ -3, & x = 5 \\ 9, & x = 7 \\ 7, & x = 9. \end{cases}$$

$$(ii) \quad f(x) = \begin{cases} 0, & x \text{ irrational} \\ 1/q, & x = p/q \text{ in lowest terms.} \end{cases}$$

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

$$(iv) \quad f(x) = \begin{cases} 1, & x = 1/n \text{ for some } n \text{ in } \mathbf{N} \\ 0, & \text{otherwise.} \end{cases}$$

$$(v) \quad f(x) = \begin{cases} 1, & \text{if the decimal expansion of } x \text{ contains a 5} \\ 0, & \text{otherwise.} \end{cases}$$

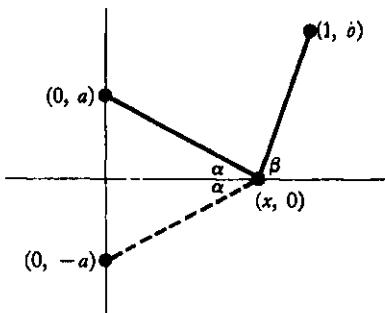


FIGURE 23

Surface area is the sum of these areas

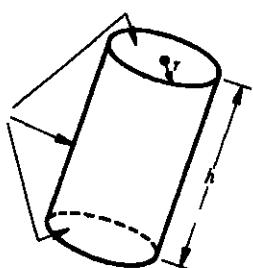


FIGURE 24

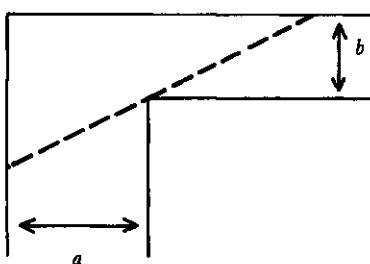


FIGURE 25

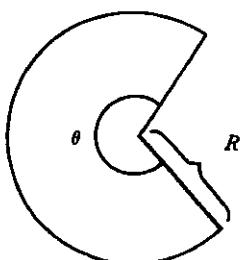


FIGURE 26

6. (a) Let (x_0, y_0) be a point of the plane, and let L be the graph of the function $f(x) = mx + b$. Find the point \bar{x} such that the distance from (x_0, y_0) to $(\bar{x}, f(\bar{x}))$ is smallest. [Notice that minimizing this distance is the same as minimizing its square. This may simplify the computations somewhat.]
 (b) Also find \bar{x} by noting that the line from (x_0, y_0) to $(\bar{x}, f(\bar{x}))$ is perpendicular to L .
 (c) Find the distance from (x_0, y_0) to L , i.e., the distance from (x_0, y_0) to $(\bar{x}, f(\bar{x}))$. [It will make the computations easier if you first assume that $b = 0$; then apply the result to the graph of $f(x) = mx$ and the point $(x_0, y_0 - b)$.] Compare with Problem 4-22.
 (d) Consider a straight line described by the equation $Ax + By + C = 0$ (Problem 4-7). Show that the distance from (x_0, y_0) to this line is $(Ax_0 + By_0 + C)/\sqrt{A^2 + B^2}$.
7. The previous Problem suggests the following question: What is the relationship between the critical points of f and those of f^2 ?
8. A straight line is drawn from the point $(0, a)$ to the horizontal axis, and then back to $(1, b)$, as in Figure 23. Prove that the total length is shortest when the angles α and β are equal. (Naturally you must bring a function into the picture: express the length in terms of x , where $(x, 0)$ is the point on the horizontal axis. The dashed line in Figure 23 suggests an alternative geometric proof; in either case the problem can be solved without actually finding the point $(x, 0)$.)
9. Prove that of all rectangles with given perimeter, the square has the greatest area.
10. Find, among all right circular cylinders of fixed volume V , the one with smallest surface area (counting the areas of the faces at top and bottom, as in Figure 24).
11. A right triangle with hypotenuse of length a is rotated about one of its legs to generate a right circular cone. Find the greatest possible volume of such a cone.
12. Two hallways, of widths a and b , meet at right angles (Figure 25). What is the greatest possible length of a ladder which can be carried horizontally around the corner?
13. A garden is to be designed in the shape of a circular sector (Figure 26), with radius R and central angle θ . The garden is to have a fixed area A . For what value of R and θ (in radians) will the length of the fencing around the perimeter be minimized?
14. Show that the sum of a positive number and its reciprocal is at least 2.
15. Find the trapezoid of largest area that can be inscribed in a semicircle of radius a , with one base lying along the diameter.

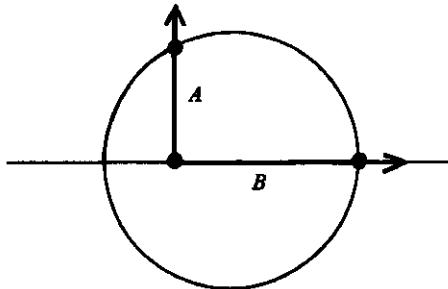


FIGURE 27

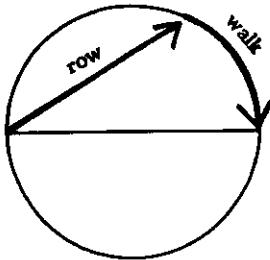


FIGURE 28

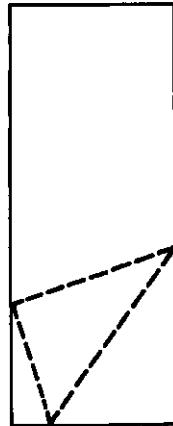


FIGURE 29

16. A right angle is moved along the diameter of a circle of radius a , as shown in Figure 27. What is the greatest possible length ($A + B$) intercepted on it by the circle?
17. Ecological Ed must cross a circular lake of radius 1 mile. He can row across at 2 mph or walk around at 4 mph, or he can row part way and walk the rest (Figure 28). What route should he take so as to
- see as much scenery as possible?
 - cross as quickly as possible?
18. The lower right-hand corner of a page is folded over so that it just touches the left edge of the paper, as in Figure 29. If the width of the paper is α and the page is very long, show that the minimum length of the crease is $3\sqrt{3}\alpha/4$.
19. Figure 30 shows the graph of the *derivative* of f . Find all local maximum and minimum points of f .

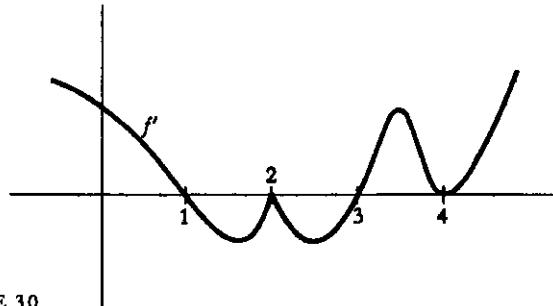


FIGURE 30

- *20. Suppose that f is a polynomial function, $f(x) = x^n + a_{n-1}x^{n-1} + \dots + a_0$, with critical points $-1, 1, 2, 3, 4$, and corresponding critical values $6, 1, 2, 4, 3$. Sketch the graph of f , distinguishing the cases n even and n odd.
- *21. (a) Suppose that the critical points of the polynomial function $f(x) = x^n + a_{n-1}x^{n-1} + \dots + a_0$ are $-1, 1, 2, 3$, and $f''(-1) = 0, f''(1) > 0, f''(2) < 0, f''(3) = 0$. Sketch the graph of f as accurately as possible on the basis of this information.
(b) Does there exist a polynomial function with the above properties, except that 3 is not a critical point?
22. Describe the graph of a rational function (in very general terms, similar to the text's description of the graph of a polynomial function).
23. (a) Prove that two polynomial functions of degree m and n , respectively, intersect in at most $\max(m, n)$ points.
(b) For each m and n exhibit two polynomial functions of degree m and n which intersect $\max(m, n)$ times.
- *24. (a) Suppose that the polynomial function $f(x) = x^n + a_{n-1}x^{n-1} + \dots + a_0$ has exactly k critical points and $f''(x) \neq 0$ for all critical points x . Show that $n - k$ is odd.

- (b) For each n , show that there is a polynomial function f of degree n with k critical points if $n - k$ is odd.

- (c) Suppose that the polynomial function $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ has k_1 local maximum points and k_2 local minimum points. Show that $k_2 = k_1 + 1$ if n is even, and $k_2 = k_1$ if n is odd.

- (d) Let n, k_1, k_2 be three integers with $k_2 = k_1 + 1$ if n is even, and $k_2 = k_1$ if n is odd, and $k_1 + k_2 < n$. Show that there is a polynomial function f of degree n , with k_1 local maximum points and k_2 local minimum points.

Hint: Pick $a_1 < a_2 < \cdots < a_{k_1+k_2}$ and try $f'(x) = \prod_{i=1}^{k_1+k_2} (x - a_i) \cdot (1+x^2)^l$ for an appropriate number l .

25. (a) Prove that if $f'(x) \geq M$ for all x in $[a, b]$, then $f(b) \geq f(a) + M(b - a)$.
 (b) Prove that if $f'(x) \leq M$ for all x in $[a, b]$, then $f(b) \leq f(a) + M(b - a)$.
 (c) Formulate a similar theorem when $|f'(x)| \leq M$ for all x in $[a, b]$.

- *26. Suppose that $f'(x) \geq M > 0$ for all x in $[0, 1]$. Show that there is an interval of length $\frac{1}{4}$ on which $|f| \geq M/4$.

27. (a) Suppose that $f'(x) > g'(x)$ for all x , and that $f(a) = g(a)$. Show that $f(x) > g(x)$ for $x > a$ and $f(x) < g(x)$ for $x < a$.
 (b) Show by an example that these conclusions do not follow without the hypothesis $f(a) = g(a)$.

28. Find all functions f such that

- (a) $f'(x) = \sin x$.
 (b) $f''(x) = x^3$.
 (c) $f'''(x) = x + x^2$.

29. Although it is true that a weight dropped from rest will fall $s(t) = 16t^2$ feet after t seconds, this experimental fact does not mention the behavior of weights which are thrown upwards or downwards. On the other hand, the law $s''(t) = 32$ is always true and has just enough ambiguity to account for the behavior of a weight released from any height, with any initial velocity. For simplicity let us agree to measure heights upwards from ground level; in this case velocities are positive for rising bodies and negative for falling bodies, and all bodies fall according to the law $s''(t) = -32$.

- (a) Show that s is of the form $s(t) = -16t^2 + \alpha t + \beta$.
 (b) By setting $t = 0$ in the formula for s , and then in the formula for s' , show that $s(t) = -16t^2 + v_0t + s_0$, where s_0 is the height from which the body is released at time 0, and v_0 is the velocity with which it is released.
 (c) A weight is thrown upwards with velocity v feet per second, at ground level. How high will it go? ("How high" means "what is the maximum height for all times".) What is its velocity at the moment it achieves its greatest height? What is its acceleration at that moment? When will it hit the ground again? What will its velocity be when it hits the ground again?

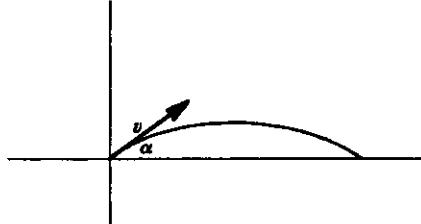


FIGURE 31

30. A cannon ball is shot from the ground with velocity v at an angle α (Figure 31) so that it has a vertical component of velocity $v \sin \alpha$ and a horizontal component $v \cos \alpha$. Its distance $s(t)$ above the ground obeys the law $s(t) = -16t^2 + (v \sin \alpha)t$, while its horizontal velocity remains constantly $v \cos \alpha$.

- (a) Show that the path of the cannon ball is a parabola (find the position at each time t , and show that these points lie on a parabola).
- (b) Find the angle α which will maximize the horizontal distance traveled by the cannon ball before striking the ground.

31. (a) Give an example of a function f for which $\lim_{x \rightarrow \infty} f(x)$ exists, but $\lim_{x \rightarrow \infty} f'(x)$ does not exist.
- (b) Prove that if $\lim_{x \rightarrow \infty} f(x)$ and $\lim_{x \rightarrow \infty} f'(x)$ both exist, then $\lim_{x \rightarrow \infty} f'(x) = 0$.
- (c) Prove that if $\lim_{x \rightarrow \infty} f(x)$ exists and $\lim_{x \rightarrow \infty} f''(x)$ exists, then $\lim_{x \rightarrow \infty} f''(x) = 0$. (See also Problem 20-15.)
32. Suppose that f and g are two differentiable functions which satisfy $fg' - f'g = 0$. Prove that if a and b are adjacent zeros of f , and $g(a)$ and $g(b)$ are not both 0, then $g(x) = 0$ for some x between a and b . (Naturally the same result holds with f and g interchanged; thus, the zeros of f and g separate each other.) Hint: Derive a contradiction from the assumption that $g(x) \neq 0$ for all x between a and b : if a number is not 0, there is a natural thing to do with it.
33. Suppose that $|f(x) - f(y)| \leq |x - y|^n$ for $n > 1$. Prove that f is constant by considering f' . Compare with Problem 3-20.
34. A function f is *Lipschitz of order α* at x if there is a constant C such that

$$(*) \quad |f(x) - f(y)| \leq C|x - y|^\alpha$$

for all y in an interval around x . The function f is *Lipschitz of order α on an interval* if $(*)$ holds for all x and y in the interval.

- (a) If f is Lipschitz of order $\alpha > 0$ at x , then f is continuous at x .
- (b) If f is Lipschitz of order $\alpha > 0$ on an interval, then f is uniformly continuous on this interval (see Chapter 8, Appendix).
- (c) If f is differentiable at x , then f is Lipschitz of order 1 at x . Is the converse true?
- (d) If f is differentiable on $[a, b]$, is f Lipschitz of order 1 on $[a, b]$?
- (e) If f is Lipschitz of order $\alpha > 1$ on $[a, b]$, then f is constant on $[a, b]$.

35. Prove that if

$$\frac{a_0}{1} + \frac{a_1}{2} + \cdots + \frac{a_n}{n+1} = 0,$$

then

$$a_0 + a_1 x + \cdots + a_n x^n = 0$$

for some x in $[0, 1]$.

36. Prove that the polynomial function $f_m(x) = x^3 - 3x + m$ never has two roots in $[0, 1]$, no matter what m may be. (This is an easy consequence of Rolle's Theorem. It is instructive, after giving an analytic proof, to graph f_0 and f_2 , and consider where the graph of f_m lies in relation to them.)

37. Suppose that f is continuous and differentiable on $[0, 1]$, that $f(x)$ is in $[0, 1]$ for each x , and that $f'(x) \neq 1$ for all x in $[0, 1]$. Show that there is exactly one number x in $[0, 1]$ such that $f(x) = x$. (Half of this problem has been done already, in Problem 7-11.)

38. (a) Prove that the function $f(x) = x^2 - \cos x$ satisfies $f(x) = 0$ for precisely two numbers x .
 (b) Prove the same for the function $f(x) = 2x^2 - x \sin x - \cos^2 x$. (Some preliminary estimates will be useful to restrict the possible location of the zeros of f .)

- *39. (a) Prove that if f is a twice differentiable function with $f(0) = 0$ and $f(1) = 1$ and $f'(0) = f'(1) = 0$, then $|f''(x)| \geq 4$ for some x in $[0, 1]$. In more picturesque terms: A particle which travels a unit distance in a unit time, and starts and ends with velocity 0, has at some time an acceleration ≥ 4 . Hint: Prove that either $f''(x) > 4$ for some x in $[0, \frac{1}{2}]$, or else $f''(x) < -4$ for some x in $[\frac{1}{2}, 1]$.
 (b) Show that in fact we must have $|f''(x)| > 4$ for some x in $[0, 1]$.

40. Suppose that f is a function such that $f'(x) = 1/x$ for all $x > 0$ and $f(1) = 0$. Prove that $f(xy) = f(x) + f(y)$ for all $x, y > 0$. Hint: Find $g'(x)$ when $g(x) = f(xy)$.

- *41. Suppose that f satisfies

$$f''(x) + f'(x)g(x) - f(x) = 0$$

for some function g . Prove that if f is 0 at two points, then f is 0 on the interval between them. Hint: Use Theorem 6.

42. Suppose that f is n -times differentiable and that $f(x) = 0$ for $n+1$ different x . Prove that $f^{(n)}(x) = 0$ for some x .
 43. Let a_1, \dots, a_{n+1} be arbitrary points in $[a, b]$, and let

$$Q(x) = \prod_{i=1}^{n+1} (x - x_i).$$

Suppose that f is $(n+1)$ -times differentiable and that P is a polynomial function of degree $\leq n$ such that $P(x_i) = f(x_i)$ for $i = 1, \dots, n+1$ (see page 49). Show that for each x in $[a, b]$ there is a number c in (a, b) such that

$$f(x) - P(x) = Q(x) \cdot \frac{f^{(n+1)}(c)}{(n+1)!}.$$

Hint: Consider the function

$$F(t) = Q(x)[f(t) - P(t)] - Q(t)[f(x) - P(x)].$$

Show that F is zero at $n+2$ different points in $[a, b]$, and use Problem 42.

- 44.** Prove that

$$\frac{1}{9} < \sqrt{66} - 8 < \frac{1}{8}$$

(without computing $\sqrt{66}$ to 2 decimal places!).

- 45.** Prove the following slight generalization of the Mean Value Theorem: If f is continuous and differentiable on (a, b) and $\lim_{y \rightarrow a^+} f(y)$ and $\lim_{y \rightarrow b^-} f(y)$ exist, then there is some x in (a, b) such that

$$f'(x) = \frac{\lim_{y \rightarrow b^-} f(y) - \lim_{y \rightarrow a^+} f(y)}{b - a}.$$

(Your proof should begin: "This is a trivial consequence of the Mean Value Theorem because . . .".)

- 46.** Prove that the conclusion of the Cauchy Mean Value Theorem can be written in the form

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(x)}{g'(x)},$$

under the additional assumptions that $g(b) \neq g(a)$ and that $f'(x)$ and $g'(x)$ are never simultaneously 0 on (a, b) .

- *47.** Prove that if f and g are continuous on $[a, b]$ and differentiable on (a, b) , and $g'(x) \neq 0$ for x in (a, b) , then there is some x in (a, b) with

$$\frac{f'(x)}{g'(x)} = \frac{f(x) - f(a)}{g(b) - g(x)}.$$

Hint: Multiply out first, to see what this really says.

- 48.** What is wrong with the following use of l'Hôpital's Rule:

$$\lim_{x \rightarrow 1} \frac{x^3 + x - 2}{x^2 - 3x + 2} = \lim_{x \rightarrow 1} \frac{3x^2 + 1}{2x - 3} = \lim_{x \rightarrow 1} \frac{6x}{2} = 3.$$

(The limit is actually -4 .)

49. Find the following limits:

$$(i) \quad \lim_{x \rightarrow 0} \frac{x}{\tan x}.$$

$$(ii) \quad \lim_{x \rightarrow 0} \frac{\cos^2 x - 1}{x^2}.$$

50. Find $f'(0)$ if

$$f(x) = \begin{cases} \frac{g(x)}{x}, & x \neq 0 \\ 0, & x = 0, \end{cases}$$

and $g(0) = g'(0) = 0$ and $g''(0) = 17$.

51. Prove the following forms of l'Hôpital's Rule (none requiring any essentially new reasoning).

(a) If $\lim_{x \rightarrow a^+} f(x) = \lim_{x \rightarrow a^+} g(x) = 0$, and $\lim_{x \rightarrow a^+} f'(x)/g'(x) = l$, then $\lim_{x \rightarrow a^+} f(x)/g(x) = l$ (and similarly for limits from below).

(b) If $\lim_{x \rightarrow a} f(x) = \lim_{x \rightarrow a} g(x) = 0$, and $\lim_{x \rightarrow a} f'(x)/g'(x) = \infty$, then $\lim_{x \rightarrow a} f(x)/g(x) = \infty$ (and similarly for $-\infty$, or if $x \rightarrow a$ is replaced by $x \rightarrow a^+$ or $x \rightarrow a^-$).

(c) If $\lim_{x \rightarrow \infty} f(x) = \lim_{x \rightarrow \infty} g(x) = 0$, and $\lim_{x \rightarrow \infty} f'(x)/g'(x) = l$, then $\lim_{x \rightarrow \infty} f(x)/g(x) = l$ (and similarly for $-\infty$). Hint: Consider $\lim_{x \rightarrow 0^+} f(1/x)/g(1/x)$.

(d) If $\lim_{x \rightarrow \infty} f(x) = \lim_{x \rightarrow \infty} g(x) = 0$, and $\lim_{x \rightarrow \infty} f'(x)/g'(x) = \infty$, then $\lim_{x \rightarrow \infty} f(x)/g(x) = \infty$.

52. There is another form of l'Hôpital's Rule which requires more than algebraic manipulations: If $\lim_{x \rightarrow \infty} f(x) = \lim_{x \rightarrow \infty} g(x) = \infty$, and $\lim_{x \rightarrow \infty} f'(x)/g'(x) = l$, then $\lim_{x \rightarrow \infty} f(x)/g(x) = l$. Prove this as follows.

(a) For every $\varepsilon > 0$ there is a number a such that

$$\left| \frac{f'(x)}{g'(x)} - l \right| < \varepsilon \quad \text{for } x > a.$$

Apply the Cauchy Mean Value Theorem to f and g on $[a, x]$ to show that

$$\left| \frac{f(x) - f(a)}{g(x) - g(a)} - l \right| < \varepsilon \quad \text{for } x > a.$$

(Why can we assume $g(x) - g(a) \neq 0$?)

(b) Now write

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(a)}{g(x) - g(a)} \cdot \frac{f(x)}{f(x) - f(a)} \cdot \frac{g(x) - g(a)}{g(x)}$$

(why can we assume that $f(x) - f(a) \neq 0$ for large x ?) and conclude that

$$\left| \frac{f(x)}{g(x)} - l \right| < 2\epsilon \quad \text{for sufficiently large } x.$$

53. To complete the orgy of variations on l'Hôpital's Rule, use Problem 52 to prove a few more cases of the following general statement (there are so many possibilities that you should select just a few, if any, that interest you):

If $\lim_{x \rightarrow []} f(x) = \lim_{x \rightarrow []} g(x) = \{ \}$ and $\lim_{x \rightarrow []} f'(x)/g'(x) = ()$, then $\lim_{x \rightarrow []} f(x)/g(x) = ()$. Here $[]$ can be a or a^+ or a^- or ∞ or $-\infty$, and $\{ \}$ can be 0 or ∞ or $-\infty$, and $()$ can be l or ∞ or $-\infty$.

- *54. (a) Suppose that f is differentiable on $[a, b]$. Prove that if the minimum of f on $[a, b]$ is at a , then $f'(a) \geq 0$, and if it is at b , then $f'(b) \leq 0$. (One half of the proof of Theorem 1 will go through.)
(b) Suppose that $f'(a) < 0$ and $f'(b) > 0$. Show that $f'(x) = 0$ for some x in (a, b) . Hint: Consider the minimum of f on $[a, b]$; why must it be somewhere in (a, b) ?
(c) Prove that if $f'(a) < c < f'(b)$, then $f'(x) = c$ for some x in (a, b) . (This result is known as Darboux's Theorem.) Hint: Cook up an appropriate function to which part (b) may be applied.
55. Suppose that f is differentiable in some interval containing a , but that f' is discontinuous at a .
(a) The one-sided limits $\lim_{x \rightarrow a^+} f'(x)$ and $\lim_{x \rightarrow a^-} f'(x)$ cannot both exist. (This is just a minor variation on Theorem 7.)
(b) Neither of these one-sided limits can exist even in the sense of being $+\infty$ or $-\infty$. Hint: Use Darboux's Theorem (Problem 54).
- *56. It is easy to find a function f such that $|f|$ is differentiable but f is not. For example, we can choose $f(x) = 1$ for x rational and $f(x) = -1$ for x irrational. In this example f is not even continuous, nor is this a mere coincidence: Prove that if $|f|$ is differentiable at a , and f is continuous at a , then f is also differentiable at a . Hint: It suffices to consider only a with $f(a) = 0$. Why? In this case, what must $|f|'(a)$ be?
- *57. (a) Let $y \neq 0$ and let n be even. Prove that $x^n + y^n = (x + y)^n$ only when $x = 0$. Hint: If $x_0^n + y^n = (x_0 + y)^n$, apply Rolle's Theorem to $f(x) = x^n + y^n - (x + y)^n$ on $[0, x_0]$.

- (b) Prove that if $y \neq 0$ and n is odd, then $x^n + y^n = (x + y)^n$ only if $x = 0$ or $x = -y$.
- *58.** Use the method of Problem 57 to prove that if n is even and $f(x) = x^n$, then every tangent line to f intersects f only once.
- *59.** Prove even more generally that if f' is increasing, then every tangent line intersects f only once.
- *60.** Suppose that $f(0) = 0$ and f' is increasing. Prove that the function $g(x) = f(x)/x$ is increasing on $(0, \infty)$. Hint: Obviously you should look at $g'(x)$. Prove that it is positive by applying the Mean Value Theorem to f on the right interval (it will help to remember that the hypothesis $f(0) = 0$ is essential, as shown by the function $f(x) = 1 + x^2$).
- *61.** Use derivatives to prove that if $n \geq 1$, then

$$(1+x)^n > 1+nx \quad \text{for } -1 < x < 0 \text{ and } 0 < x$$

(notice that equality holds for $x = 0$).

- 62.** Let $f(x) = x^4 \sin^2 1/x$ for $x \neq 0$, and let $f(0) = 0$ (Figure 32).
- (a) Prove that 0 is a local minimum point for f .
 (b) Prove that $f'(0) = f''(0) = 0$.

This function thus provides another example to show that Theorem 6 cannot be improved. It also illustrates a subtlety about maxima and minima that often goes unnoticed: a function may not be increasing in any interval to the right of a local minimum point, nor decreasing in any interval to the left.

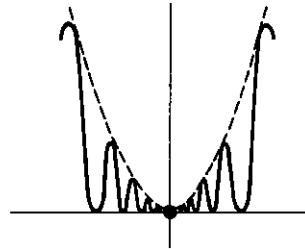


FIGURE 32

- *63.** (a) Prove that if $f'(a) > 0$ and f' is continuous at a , then f is increasing in some interval containing a .

The next two parts of this problem show that continuity of f' is essential.

- (b) If $g(x) = x^2 \sin 1/x$, show that there are numbers x arbitrarily close to 0 with $g'(x) = 1$ and also with $g'(x) = -1$.

- (c) Suppose $0 < \alpha < 1$. Let $f(x) = \alpha x + x^2 \sin 1/x$ for $x \neq 0$, and let $f(0) = 0$ (see Figure 33). Show that f is not increasing in any open interval containing 0, by showing that in any interval there are points x with $f'(x) > 0$ and also points x with $f'(x) < 0$.

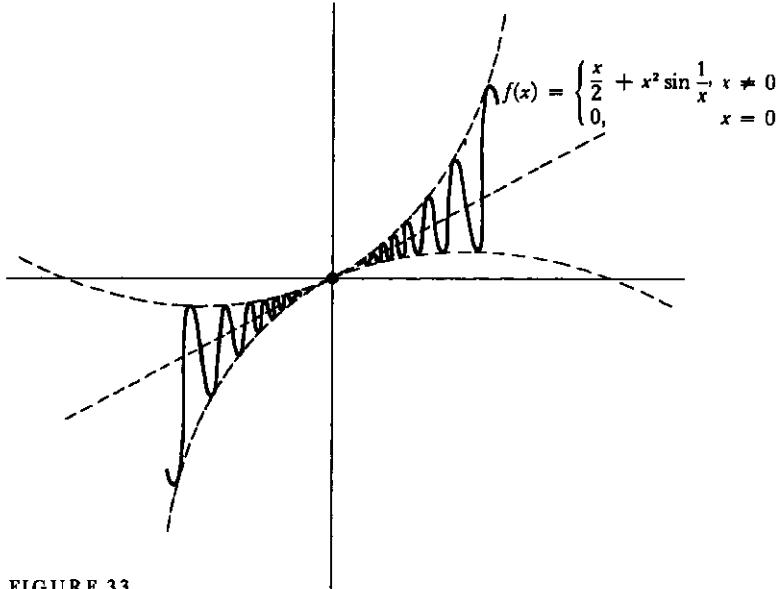


FIGURE 33

The behavior of f for $\alpha \geq 1$, which is much more difficult to analyze, is discussed in the next problem.

- **64.** Let $f(x) = \alpha x + x^2 \sin 1/x$ for $x \neq 0$, and let $f(0) = 0$. In order to find the sign of $f'(x)$ when $\alpha \geq 1$ it is necessary to decide if $2x \sin 1/x - \cos 1/x$ is < -1 for any numbers x close to 0. It is a little more convenient to consider the function $g(y) = 2(\sin y)/y - \cos y$ for $y \neq 0$; we want to know if $g(y) < -1$ for large y . This question is quite delicate; the most significant part of $g(y)$ is $-\cos y$, which does reach the value -1 , but this happens only when $\sin y = 0$, and it is not at all clear whether g itself can have values < -1 . The obvious approach to this problem is to find the local minimum values of g . Unfortunately, it is impossible to solve the equation $g'(y) = 0$ explicitly, so more ingenuity is required.

- (a) Show that if $g'(y) = 0$, then

$$\cos y = (\sin y) \left(\frac{2 - y^2}{2y} \right),$$

and conclude that

$$g(y) = (\sin y) \left(\frac{2 + y^2}{2y} \right).$$

(b) Now show that if $g'(y) = 0$, then

$$\sin^2 y = \frac{4y^2}{4 + y^4},$$

and conclude that

$$|g(y)| = \frac{2 + y^2}{\sqrt{4 + y^4}}.$$

- (c) Using the fact that $(2 + y^2)/\sqrt{4 + y^4} > 1$, show that if $\alpha = 1$, then f is not increasing in any interval around 0.
- (d) Using the fact that $\lim_{y \rightarrow \infty} (2 + y^2)/\sqrt{4 + y^4} = 1$, show that if $\alpha > 1$, then f is increasing in some interval around 0.

****65.** A function f is **increasing at a** if there is some number $\delta > 0$ such that

$$f(x) > f(a) \quad \text{if } a < x < a + \delta$$

and

$$f(x) < f(a) \quad \text{if } a - \delta < x < a.$$

Notice that this does *not* mean that f is increasing in the interval $(a - \delta, a + \delta)$; for example, the function shown in Figure 33 is increasing at 0, but is not an increasing function in any open interval containing 0.

- (a) Suppose that f is continuous on $[0, 1]$ and that f is increasing at a for every a in $[0, 1]$. Prove that f is increasing on $[0, 1]$. (First convince yourself that there is something to be proved.) Hint: For $0 < b < 1$, prove that the minimum of f on $[b, 1]$ must be at b .
- (b) Prove part (a) without the assumption that f is continuous, by considering for each b in $[0, 1]$ the set $S_b = \{x : f(y) \geq f(b) \text{ for all } y \text{ in } [b, x]\}$. (This part of the problem is not necessary for the other parts.) Hint: Prove that $S_b = \{x : b \leq x \leq 1\}$ by considering $\sup S_b$.
- (c) If f is increasing at a and f is differentiable at a , prove that $f'(a) \geq 0$ (this is easy).
- (d) If $f'(a) > 0$, prove that f is increasing at a (go right back to the definition of $f'(a)$).
- (e) Use parts (a) and (d) to show, without using the Mean Value Theorem, that if f is continuous on $[0, 1]$ and $f'(a) > 0$ for all a in $[0, 1]$, then f is increasing on $[0, 1]$.
- (f) Suppose that f is continuous on $[0, 1]$ and $f'(a) = 0$ for all a in $(0, 1)$. Apply part (e) to the function $g(x) = f(x) + \varepsilon x$ to show that $f(1) - f(0) > -\varepsilon$. Similarly, show that $f(1) - f(0) < \varepsilon$ by considering $h(x) = \varepsilon x - f(x)$. Conclude that $f(0) = f(1)$.

This particular proof that a function with zero derivative must be constant has many points in common with a proof of H. A. Schwarz, which may be the first rigorous proof ever given. Its discoverer, at least, seemed to think it was. See his exuberant letter in reference [40] of the Suggested Reading.

- **66.** (a) If f is a constant function, then every point is a local maximum point for f . It is quite possible for this to happen even if f is not a constant function: for example, if $f(x) = 0$ for $x < 0$ and $f(x) = 1$ for $x \geq 0$. But prove, using Problem 8-4, that if f is continuous on $[a, b]$ and every point of $[a, b]$ is a local maximum point, then f is a constant function. The same result holds, of course, if every point of $[a, b]$ is a local minimum point.
- (b) Suppose now that every point is either a local maximum or a local minimum point for f (but we don't preclude the possibility that some points are local maxima while others are local minima). Prove that f is constant, as follows. Suppose that $f(a_0) < f(b_0)$. We can assume that $f(a_0) < f(x) < f(b_0)$ for $a_0 < x < b_0$. (Why?) Using Theorem 1 of the Appendix to Chapter 8, partition $[a_0, b_0]$ into intervals on which $\sup f - \inf f < (f(b_0) - f(a_0))/2$; also choose the lengths of these intervals to be less than $(b_0 - a_0)/2$. Then there is one such interval $[a_1, b_1]$ with $a_0 < a_1 < b_1 < b_0$ and $f(a_1) < f(b_1)$. (Why?) Continue inductively and use the Nested Interval Theorem (Problem 8-14) to find a point x that cannot be a local maximum or minimum.
- **67.** (a) A point x is called a **strict maximum point** for f on A if $f(x) > f(y)$ for all y in A with $y \neq x$ (compare with the definition of an ordinary maximum point). A **local strict maximum point** is defined in the obvious way. Find all local strict maximum points of the function

$$f(x) = \begin{cases} 0, & x \text{ irrational} \\ \frac{1}{q}, & x = \frac{p}{q} \text{ in lowest terms.} \end{cases}$$

It seems quite unlikely that a function can have a local strict maximum at *every* point (although the above example might give one pause for thought). Prove this as follows.

- (b) Suppose that every point is a local strict maximum point for f . Let x_1 be any number and choose $a_1 < x_1 < b_1$ with $b_1 - a_1 < 1$ such that $f(x_1) > f(x)$ for all x in $[a_1, b_1]$. Let $x_2 \neq x_1$ be any point in (a_1, b_1) and choose $a_1 \leq a_2 < x_2 < b_2 \leq b_1$ with $b_2 - a_2 < \frac{1}{2}$ such that $f(x_2) > f(x)$ for all x in $[a_2, b_2]$. Continue in this way, and use the Nested Interval Theorem (Problem 8-14) to obtain a contradiction.

APPENDIX. CONVEXITY AND CONCAVITY

Although the graph of a function can be sketched quite accurately on the basis of the information provided by the derivative, some subtle aspects of the graph are revealed only by examining the second derivative. These details were purposely omitted previously because graph sketching is complicated enough without worrying about them, and the additional information obtained is often not worth the effort. Also, correct proofs of the relevant facts are sufficiently difficult to be placed in an appendix. Despite these discouraging remarks, the information presented here is well worth assimilating, because the notions of convexity and concavity are far more important than as mere aids to graph sketching. Moreover, the proofs have a pleasantly geometric flavor not often found in calculus theorems. Indeed, the basic definition is geometric in nature (see Figure 1).

DEFINITION 1

A function f is **convex** on an interval, if for all a and b in the interval, the line segment joining $(a, f(a))$ and $(b, f(b))$ lies above the graph of f .

The geometric condition appearing in this definition can be expressed in an analytic way that is sometimes more useful in proofs. The straight line between $(a, f(a))$ and $(b, f(b))$ is the graph of the function g defined by

$$g(x) = \frac{f(b) - f(a)}{b - a}(x - a) + f(a).$$

This line lies above the graph of f at x if $g(x) > f(x)$, that is, if

$$\frac{f(b) - f(a)}{b - a}(x - a) + f(a) > f(x)$$

or

$$\frac{f(b) - f(a)}{b - a}(x - a) > f(x) - f(a)$$

or

$$\frac{f(b) - f(a)}{b - a} > \frac{f(x) - f(a)}{x - a}.$$

We therefore have an equivalent definition of convexity.

DEFINITION 2

A function f is **convex** on an interval if for a , x , and b in the interval with $a < x < b$ we have

$$\frac{f(x) - f(a)}{x - a} < \frac{f(b) - f(a)}{b - a}.$$

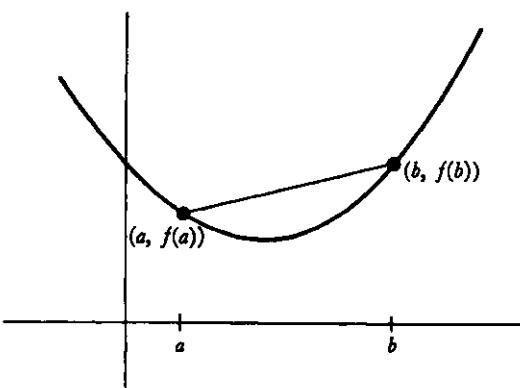


FIGURE 1

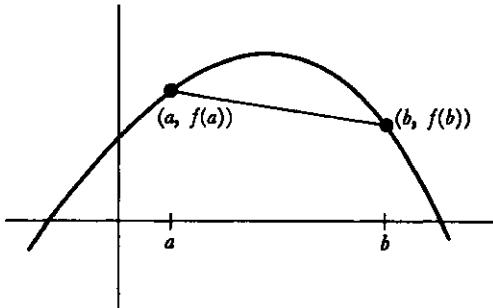


FIGURE 2

If the word “over” in Definition 1 is replaced by “under” or, equivalently, if the inequality in Definition 2 is replaced by

$$\frac{f(x) - f(a)}{x - a} < \frac{f(b) - f(a)}{b - a},$$

we obtain the definition of a **concave** function (Figure 2). It is not hard to see that the concave functions are precisely the ones of the form $-f$, where f is convex. For this reason, the next three theorems about convex functions have immediate corollaries about concave functions, so simple that we will not even bother to state them.

Figure 3 shows some tangent lines of a convex function. Two things seem to be true:

- (1) The graph of f lies above the tangent line at $(a, f(a))$ except at the point $(a, f(a))$ itself (this point is called the **point of contact** of the tangent line).
- (2) If $a < b$, then the slope of the tangent line at $(a, f(a))$ is less than the slope of the tangent line at $(b, f(b))$; that is, f' is increasing.

As a matter of fact these observations are true, and the proofs are not difficult.

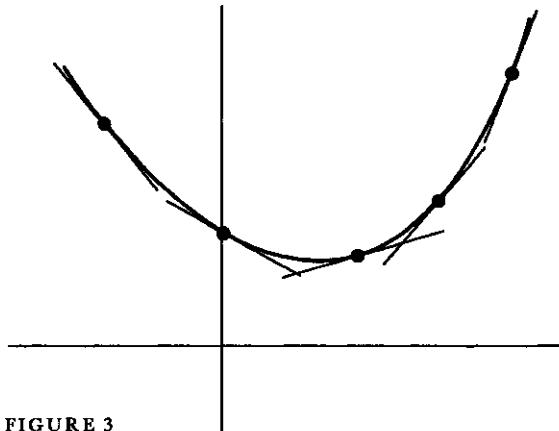


FIGURE 3

THEOREM 1 Let f be convex. If f is differentiable at a , then the graph of f lies above the tangent line through $(a, f(a))$, except at $(a, f(a))$ itself. If $a < b$ and f is differentiable at a and b , then $f'(a) < f'(b)$.

PROOF If $0 < h_1 < h_2$, then as Figure 4 indicates,

$$(1) \quad \frac{f(a + h_1) - f(a)}{h_1} < \frac{f(a + h_2) - f(a)}{h_2}.$$

A nonpictorial proof can be derived immediately from Definition 2 applied to $a < a + h_1 < a + h_2$. Inequality (1) shows that the values of

$$\frac{f(a + h) - f(a)}{h}$$

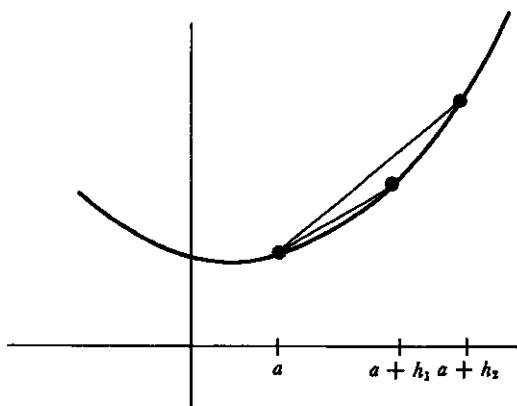


FIGURE 4

decrease as $h \rightarrow 0^+$. Consequently,

$$f'(a) < \frac{f(a+h) - f(a)}{h} \quad \text{for } h > 0$$

(in fact $f'(a)$ is the greatest lower bound of all these numbers). But this means that for $h > 0$ the secant line through $(a, f(a))$ and $(a+h, f(a+h))$ has larger slope than the tangent line, which implies that $(a+h, f(a+h))$ lies above the tangent line (an analytic translation of this argument is easily supplied).

For negative h there is a similar situation (Figure 5); if $h_2 < h_1 < 0$, then

$$\frac{f(a+h_1) - f(a)}{h_1} > \frac{f(a+h_2) - f(a)}{h_2}.$$

This shows that the slope of the tangent line is greater than

$$\frac{f(a+h) - f(a)}{h} \quad \text{for } h < 0$$

(in fact $f'(a)$ is the least upper bound of all these numbers), so that $f(a+h)$ lies above the tangent line if $h < 0$. This proves the first part of the theorem.

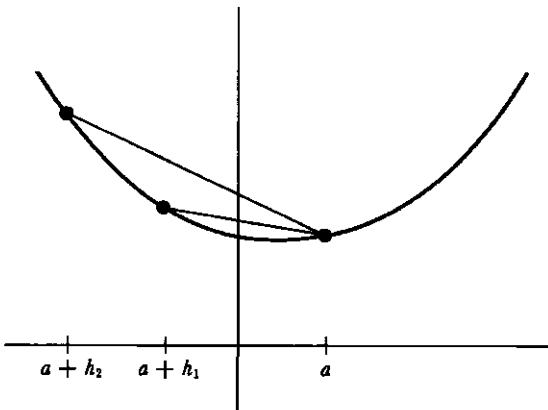


FIGURE 5

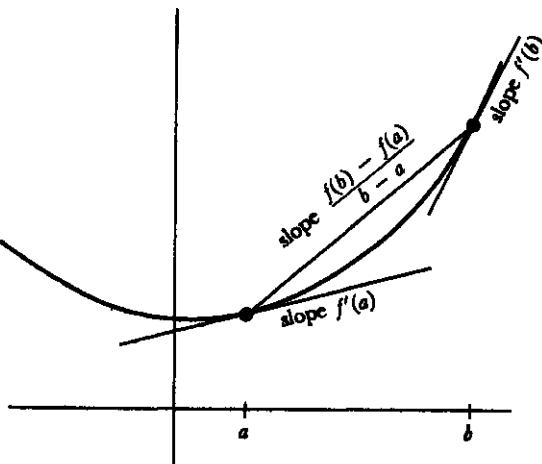


FIGURE 6

Now suppose that $a < b$. Then, as we have already seen (Figure 6),

$$\begin{aligned} f'(a) &< \frac{f(a + (b - a)) - f(a)}{b - a} \quad \text{since } b - a > 0 \\ &= \frac{f(b) - f(a)}{b - a} \end{aligned}$$

and

$$\begin{aligned} f'(b) &> \frac{f(b + (a - b)) - f(b)}{a - b} \quad \text{since } a - b < 0 \\ &= \frac{f(a) - f(b)}{a - b} = \frac{f(b) - f(a)}{b - a}. \end{aligned}$$

Combining these inequalities, we obtain $f'(a) < f'(b)$. ■

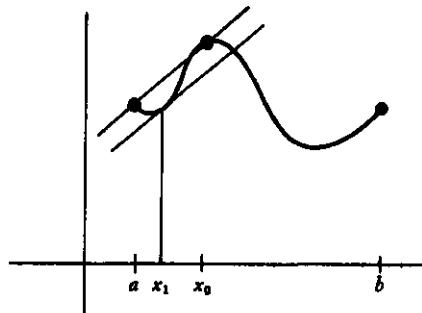
FIGURE 7

Theorem 1 has two converses. Here the proofs will be a little more difficult. We begin with a lemma that plays the same role in the next theorem that Rolle's Theorem plays in the proof of the Mean Value Theorem. It states that if f' is increasing, then the graph of f lies below any secant line which happens to be horizontal.

LEMMA Suppose f is differentiable and f' is increasing. If $a < b$ and $f(a) = f(b)$, then $f(x) < f(a) = f(b)$ for $a < x < b$.

PROOF Suppose first that $f(x) > f(a) = f(b)$ for some x in (a, b) . Then the maximum of f on $[a, b]$ occurs at some point x_0 in (a, b) with $f(x_0) > f(a)$ and, of course, $f'(x_0) = 0$ (Figure 7). On the other hand, applying the Mean Value Theorem to the interval $[a, x_0]$, we find that there is x_1 with $a < x_1 < x_0$ and

$$f'(x_1) = \frac{f(x_0) - f(a)}{x_0 - a} > 0,$$



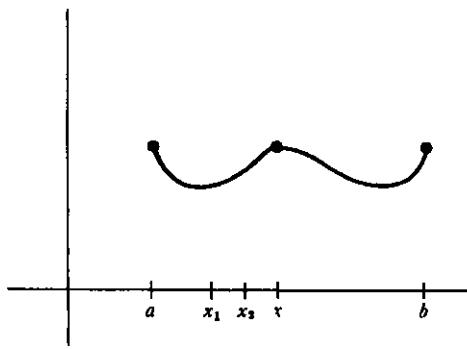


FIGURE 8

contradicting the fact that f' is increasing. This proves that $f(x) \leq f(a) = f(b)$ for $a < x < b$, and it only remains to prove that $f(x) = f(a)$ is also impossible for x in (a, b) .

Suppose $f(x) = f(a)$ for some x in (a, b) . We know that f is not constant on $[a, x]$ (if it were, f' would not be increasing on $[a, x]$) so there is (Figure 8) some x_1 with $a < x_1 < x$ and $f(x_1) < f(a)$. Applying the Mean Value Theorem to $[x_1, x]$ we conclude that there is x_2 with $x_1 < x_2 < x$ and

$$f'(x_2) = \frac{f(x) - f(x_1)}{x - x_1} > 0.$$

On the other hand, $f'(x) = 0$, since a local maximum occurs at x . Again this contradicts the hypothesis that f' is increasing. ■

We now attack the general case by the same sort of algebraic machinations that we used in the proof of the Mean Value Theorem.

THEOREM 2

If f is differentiable and f' is increasing, then f is convex.

PROOF

Let $a < b$. Define g by

$$g(x) = f(x) - \frac{f(b) - f(a)}{b - a}(x - a).$$

It is easy to see that g' is also increasing; moreover, $g(a) = g(b) = f(a)$. Applying the lemma to g we conclude that

$$g(x) < f(a) \quad \text{if } a < x < b.$$

In other words, if $a < x < b$, then

$$f(x) - \frac{f(b) - f(a)}{b - a}(x - a) < f(a)$$

or

$$\frac{f(x) - f(a)}{x - a} < \frac{f(b) - f(a)}{b - a}.$$

Hence f is convex. ■

THEOREM 3

If f is differentiable and the graph of f lies above each tangent line except at the point of contact, then f is convex.

PROOF

Let $a < b$. It is clear from Figure 9 that if $(b, f(b))$ lies above the tangent line at $(a, f(a))$, and $(a, f(a))$ lies above the tangent line at $(b, f(b))$, then the slope of the tangent line at $(b, f(b))$ must be larger than the slope of the tangent line at $(a, f(a))$. The following argument just says this with equations.

Since the tangent line at $(a, f(a))$ is the graph of the function

$$g(x) = f'(a)(x - a) + f(a),$$

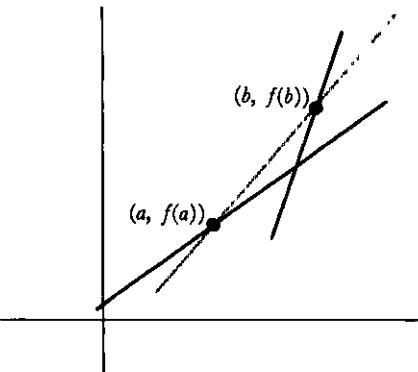


FIGURE 9

and since $(b, f(b))$ lies above the tangent line, we have

$$(1) \quad f(b) > f'(a)(b - a) + f(a).$$

Similarly, since the tangent line at $(b, f(b))$ is the graph of

$$h(x) = f'(b)(x - b) + f(b),$$

and $(a, f(a))$ lies above the tangent line at $(b, f(b))$, we have

$$(2) \quad f(a) > f'(b)(a - b) + f(b).$$

It follows from (1) and (2) that $f'(a) < f'(b)$.

It now follows from Theorem 2 that f is convex. ■

If a function f has a reasonable second derivative, the information given in these theorems can be used to discover the regions in which f is convex or concave. Consider, for example, the function

$$f(x) = \frac{1}{1+x^2}.$$

For this function,

$$f'(x) = \frac{-2x}{(1+x^2)^2}.$$

Thus $f'(x) = 0$ only for $x = 0$, and $f(0) = 1$, while

$$\begin{aligned} f'(x) &> 0 & \text{if } x < 0, \\ f'(x) &< 0 & \text{if } x > 0. \end{aligned}$$

Moreover,

$$\begin{aligned} f(x) &> 0 & \text{for all } x, \\ f(x) &\rightarrow 0 & \text{as } x \rightarrow \infty \text{ or } -\infty, \\ f &\text{ is even.} \end{aligned}$$

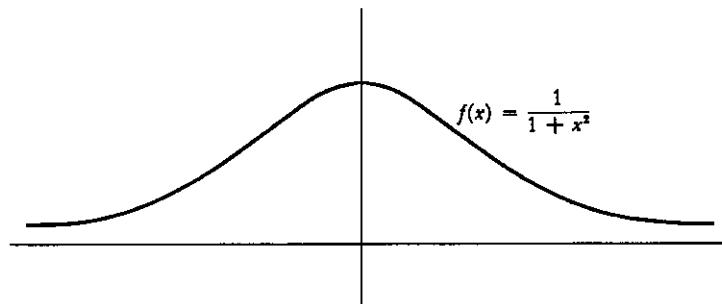


FIGURE 10

The graph of f therefore looks something like Figure 10. We now compute

$$\begin{aligned}f''(x) &= \frac{(1+x^2)^2(-2) + 2x \cdot [2(1+x^2) \cdot 2x]}{(1+x^2)^4} \\&= \frac{2(3x^2 - 1)}{(1+x^2)^3}.\end{aligned}$$

It is not hard to determine the sign of $f''(x)$. Note first that $f''(x) = 0$ only when $x = \sqrt{1/3}$ or $-\sqrt{1/3}$. Since f'' is clearly continuous, it must keep the same sign on each of the sets

$$\begin{aligned}&(-\infty, -\sqrt{1/3}), \\&(-\sqrt{1/3}, \sqrt{1/3}), \\&(\sqrt{1/3}, \infty).\end{aligned}$$

Since we easily compute, for example, that

$$\begin{aligned}f''(-1) &= \frac{1}{2} > 0, \\f''(0) &= -2 < 0, \\f''(1) &= \frac{1}{2} > 0,\end{aligned}$$

we conclude that

$$\begin{aligned}f'' &> 0 \text{ on } (-\infty, -\sqrt{1/3}) \text{ and } (\sqrt{1/3}, \infty), \\f'' &< 0 \text{ on } (-\sqrt{1/3}, \sqrt{1/3}).\end{aligned}$$

Since $f'' > 0$ means f' is increasing, it follows from Theorem 2 that f is convex on $(-\infty, -\sqrt{1/3})$ and $(\sqrt{1/3}, \infty)$, while on $(-\sqrt{1/3}, \sqrt{1/3})$ f is concave (Figure 11).

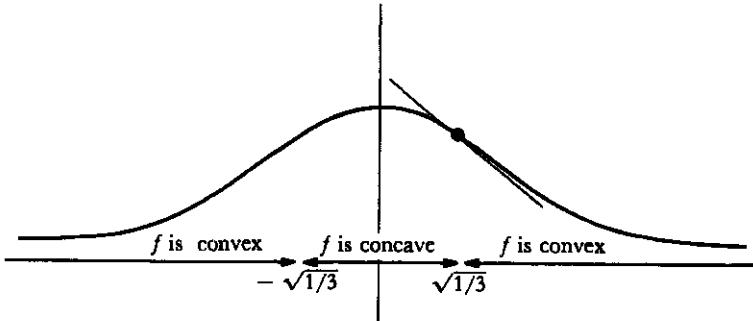


FIGURE 11

Notice that at $(\sqrt{1/3}, \frac{3}{4})$ the tangent line lies below the part of the graph to the right, since f is convex on $(\sqrt{1/3}, \infty)$, and above the part of the graph to the left, since f is concave on $(-\sqrt{1/3}, \sqrt{1/3})$; thus the tangent line crosses the graph. In general, a number a is called an **inflection point** of f if the tangent line to the graph of f at $(a, f(a))$ crosses the graph; thus $\sqrt{1/3}$ and $-\sqrt{1/3}$ are inflection points of $f(x) = 1/(1+x^2)$. Note that the condition $f''(a) = 0$ does *not* ensure that a is an inflection point of f ; for example, if $f(x) = x^4$, then $f''(0) = 0$, but f is convex, so the tangent line at $(0, 0)$ certainly doesn't cross the graph of f . In order for a to be an inflection point of a function f , it is necessary that f'' should have different signs to the left and right of a .

This example illustrates the procedure which may be used to analyze any function f . After the graph has been sketched, using the information provided by f' , the zeros of f'' are computed and the sign of f'' is determined on the intervals between consecutive zeros. On intervals where $f'' > 0$ the function is convex; on intervals where $f'' < 0$ the function is concave. Knowledge of the regions of convexity and concavity of f can often prevent absurd misinterpretation of other data about f . Several functions, which can be analyzed in this way, are given in the problems, which also contain some theoretical questions.

To round out our discussion of convexity and concavity, we will prove one further result that you may already have begun to suspect. We have seen that convex and concave functions have the property that every tangent line intersects the graph just once; a few drawings will probably convince you that no other functions have this property. The proof of this assertion is rather tricky; it is closely related to the proof of Theorem 2 of the next chapter, and is probably best deferred until after that proof has been read.

THEOREM 4

If f is differentiable on an interval and intersects each of its tangent lines just once, then f is either convex or concave on that interval.

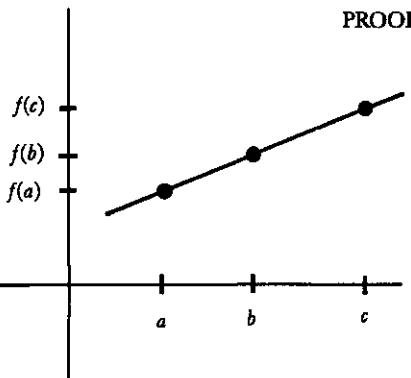
PROOF

FIGURE 12

There are two parts to the proof.

(1) First we claim that no straight line can intersect the graph of f in three different points. Suppose, on the contrary, that some straight line did intersect the graph of f at $(a, f(a))$, $(b, f(b))$ and $(c, f(c))$, with $a < b < c$ (Figure 12). Then we would have

$$(1) \quad \frac{f(b) - f(a)}{b - a} = \frac{f(c) - f(a)}{c - a}.$$

Consider the function

$$g(x) = \frac{f(x) - f(a)}{x - a} \quad \text{for } x \text{ in } [b, c].$$

Equation (1) says that $g(b) = g(c)$. So by Rolle's Theorem, there is some number x in (b, c) where $0 = g'(x)$, and thus

$$0 = (x - a)f'(x) - [f(x) - f(a)]$$

or

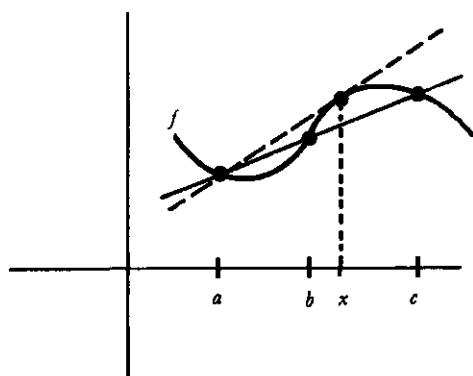
$$f'(x) = \frac{f(x) - f(a)}{x - a}.$$

But this says (Figure 13) that the tangent line at $(x, f(x))$ passes through $(a, f(a))$, contradicting the hypotheses.

(2) Suppose that $a_0 < b_0 < c_0$ and $a_1 < b_1 < c_1$ are points in the interval. Let

$$\begin{aligned} x_t &= (1 - t)a_0 + ta_1 \\ y_t &= (1 - t)b_0 + tb_1 \quad 0 \leq t \leq 1. \\ z_t &= (1 - t)c_0 + tc_1 \end{aligned}$$

FIGURE 13



Then $x_0 = a_0$ and $x_1 = a_1$ and (Problem 4-2) the points x_t all lie between a_0 and a_1 , with analogous statements for y_t and z_t . Moreover,

$$x_t < y_t < z_t \quad \text{for } 0 \leq t \leq 1.$$

Now consider the function

$$g(t) = \frac{f(y_t) - f(x_t)}{y_t - x_t} - \frac{f(z_t) - f(x_t)}{z_t - x_t} \quad \text{for } 0 \leq t \leq 1.$$

By step (1), $g(t) \neq 0$ for all t in $[0, 1]$. So either $g(t) > 0$ for all t in $[0, 1]$ or $g(t) < 0$ for all t in $[0, 1]$. Thus, either f is convex or f is concave (compare pages 231–232). ■

PROBLEMS

1. Sketch, indicating regions of convexity and concavity and points of inflection, the functions in Problem 11-1 (consider (iv) as double starred).
2. Figure 30 in Chapter 11 shows the graph of f' . Sketch the graph of f .
3. Find two convex functions f and g such that $f(x) = g(x)$ if and only if x is an integer.
4. Show that f is convex on an interval if and only if for all x and y in the interval we have

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y), \quad \text{for } 0 < t < 1.$$

(This is just a restatement of the definition, but a useful one.)

5. (a) Prove that if f and g are convex and f is increasing, then $f \circ g$ is convex. (It will be easiest to use Problem 4.)
 (b) Give an example where $g \circ f$ is not convex.
 (c) Suppose that f and g are twice differentiable. Give another proof of the result of part (a) by considering second derivatives.
6. (a) Suppose that f is differentiable and convex on an interval. Show that either f is increasing, or else f is decreasing, or else there is a number c such that f is decreasing to the left of c and increasing to the right of c .
 (b) Use this fact to give another proof of the result in Problem 5(a) when f and g are (one-time) differentiable. (You will have to be a little careful when comparing $f'(g(x))$ and $f'(g(y))$ for $x < y$).
 (c) Prove the result in part (a) without assuming f differentiable. You will have to keep track of several different cases, but no particularly clever ideas are needed. Begin by showing that if $a < b$ and $f(a) < f(b)$, then f is increasing to the right of b ; and if $f(a) > f(b)$, then f is decreasing to the left of a .
- *7. Let f be a twice-differentiable function with the following properties: $f(x) > 0$ for $x \geq 0$, and f is decreasing, and $f'(0) = 0$. Prove that

$f''(x) = 0$ for some $x > 0$ (so that in reasonable cases f will have an inflection point at x —an example is given by $f(x) = 1/(1+x^2)$). Every hypothesis in this theorem is essential, as shown by $f(x) = 1-x^2$, which is not positive for all x ; by $f(x) = x^2$, which is not decreasing; and by $f(x) = 1/(x+1)$, which does not satisfy $f'(0) = 0$. Hint: Choose $x_0 > 0$ with $f'(x_0) < 0$. We cannot have $f'(y) \leq f'(x_0)$ for all $y > x_0$. Why not? So $f'(x_1) > f'(x_0)$ for some $x_1 > x_0$. Consider f' on $[0, x_1]$.

- *8. (a) Prove that if f is convex, then $f([x+y]/2) < [f(x)+f(y)]/2$.
- (b) Suppose that f satisfies this condition. Show that $f(kx + (1-k)y) < kf(x) + (1-k)f(y)$ whenever k is a rational number, between 0 and 1, of the form $m/2^n$. Hint: Part (a) is the special case $n = 1$. Use induction, employing part (a) at each step.
- (c) Suppose that f satisfies the condition in part (a) and f is continuous. Show that f is convex.

- *9. Let p_1, \dots, p_n be *positive* numbers with $\sum_{i=1}^n p_i = 1$.

- (a) For any numbers x_1, \dots, x_n show that $\sum_{i=1}^n p_i x_i$ lies between the smallest and the largest x_i .

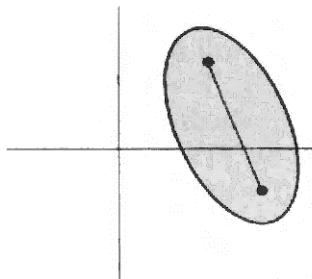
- (b) Show the same for $(1/t) \sum_{i=1}^{n-1} p_i x_i$, where $t = \sum_{i=1}^{n-1} p_i$.

- (c) Prove *Jensen's inequality*: If f is convex, then $f\left(\sum_{i=1}^n p_i x_i\right) \leq \sum_{i=1}^n p_i f(x_i)$.

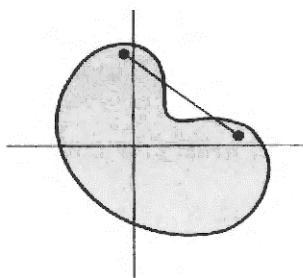
Hint: Use Problem 4, noting that $p_n = 1-t$. (Part (b) is needed to show that $(1/t) \sum_{i=1}^{n-1} p_i x_i$ is in the domain of f if x_1, \dots, x_n are.)

- *10. (a) For any function f , the right-hand derivative, $\lim_{h \rightarrow 0^+} [f(a+h) - f(a)]/h$, is denoted by $f'_+(a)$, and the left-hand derivative is denoted by $f'_-(a)$. The proof of Theorem 1 actually shows that f'_+ and f'_- always exist if f is convex. Check this assertion, and also show that f'_+ and f'_- are increasing, and that $f'_-(a) \leq f'_+(a)$.
- **(b) Show that if f is convex, then $f'_+(a) = f'_-(a)$ if and only if f'_+ is continuous at a . (Thus f is differentiable precisely when f'_+ is continuous.) Hint: $[f(b) - f(a)]/(b-a)$ is close to $f'_-(a)$ for $b < a$ close to a , and $f'_+(b)$ is less than this quotient.

- *11. (a) Prove that a convex function on \mathbf{R} , or on any open interval, must be continuous.
- (b) Give an example of a convex function on a closed interval that is *not* continuous, and explain exactly what kinds of discontinuities are possible.



(a) a convex subset of the plane



(b) a nonconvex subset of the plane

12. Call a function f *weakly convex* on an interval if for $a < b < c$ in this interval we have

$$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a}.$$

- (a) Show that a weakly convex function is convex if and only if its graph contains no straight line segments. (Sometimes a weakly convex function is simply called “convex,” while convex functions in our sense are called “strictly convex”.)
- (b) Reformulate the theorems of this section for weakly convex functions.
13. A set A of points in the plane is called *convex* if A contains the line segment joining any two points in it (Figure 14). For a function f , let A_f be the set of points (x, y) with $y \geq f(x)$, so that A_f is the set of points on or above the graph of f . Show that A is convex if and only if f is weakly convex, in the terminology of the previous problem. Further information on convex sets will be found in reference [19] of the Suggested Reading.

FIGURE 14

CHAPTER 12 INVERSE FUNCTIONS

We now have at our disposal quite powerful methods for investigating functions; what we lack is an adequate supply of functions to which these methods may be applied. We have studied various ways of forming new functions from old—addition, multiplication, division, and composition—but using these alone, we can produce only the rational functions (even the sine function, although frequently used for examples, has never been defined). In the next few chapters we will begin to construct new functions in quite sophisticated ways, but there is one important method which will practically double the usefulness of any other method we discover.

If we recall that a function is a collection of pairs of numbers, we might hit upon the bright idea of simply reversing all the pairs. Thus from the function

$$f = \{(1, 2), (3, 4), (5, 9), (13, 8)\},$$

we obtain

$$g = \{(2, 1), (4, 3), (9, 5), (8, 13)\}.$$

While $f(1) = 2$ and $f(3) = 4$, we have $g(2) = 1$ and $g(4) = 3$.

Unfortunately, this bright idea does not always work. If

$$f = \{(1, 2), (3, 4), (5, 9), (13, 4)\},$$

then the collection

$$\{(2, 1), (4, 3), (9, 5), (4, 13)\}$$

is not a function at all, since it contains both $(4, 3)$ and $(4, 13)$. It is clear where the trouble lies: $f(3) = f(13)$, even though $3 \neq 13$. This is the only sort of thing that can go wrong, and it is worthwhile giving a name to the functions for which this does not happen.

DEFINITION

A function f is **one-one** (read “one-to-one”) if $f(a) \neq f(b)$ whenever $a \neq b$.

The identity function I is obviously one-one, and so is the following modification:

$$g(x) = \begin{cases} x, & x \neq 3, 5 \\ 3, & x = 5 \\ 5, & x = 3. \end{cases}$$

The function $f(x) = x^2$ is not one-one, since $f(-1) = f(1)$, but if we define

$$g(x) = x^2, \quad x \geq 0$$

(and leave g undefined for $x < 0$), then g is one-one, because g is increasing (since $g'(x) = 2x > 0$, for $x > 0$). This observation is easily generalized: If n is a natural number and

$$f(x) = x^n, \quad x \geq 0,$$

then f is one-one. If n is odd, one can do better: the function

$$f(x) = x^n \quad \text{for all } x$$

is one-one (since $f'(x) = nx^{n-1} > 0$, for all $x \neq 0$).

It is particularly easy to decide from the graph of f whether f is one-one: the condition $f(a) \neq f(b)$ for $a \neq b$ means that no horizontal line intersects the graph of f twice (Figure 1).

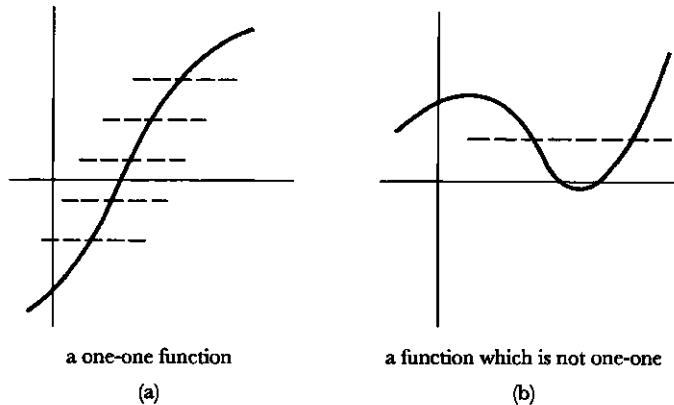


FIGURE 1

If we reverse all the pairs in (a not necessarily one-one function) f we obtain, in any case, some collection of pairs. It is popular to abstain from this procedure unless f is one-one, but there is no particular reason to do so—instead of a definition with restrictive conditions we obtain a definition and a theorem.

DEFINITION

For any function f , the **inverse** of f , denoted by f^{-1} , is the set of all pairs (a, b) for which the pair (b, a) is in f .

THEOREM 1

f^{-1} is a function if and only if f is one-one.

PROOF

Suppose first that f is one-one. Let (a, b) and (a, c) be two pairs in f^{-1} . Then (b, a) and (c, a) are in f , so $a = f(b)$ and $a = f(c)$; since f is one-one this implies that $b = c$. Thus f^{-1} is a function.

Conversely, suppose that f^{-1} is a function. If $f(b) = f(c)$, then f contains the pairs $(b, f(b))$ and $(c, f(c)) = (c, f(b))$, so $(f(b), b)$ and $(f(b), c)$ are in f^{-1} . Since f^{-1} is a function this implies that $b = c$. Thus f is one-one. ■

The graphs of f and f^{-1} are so closely related that it is possible to use the graph of f to visualize the graph of f^{-1} . Since the graph of f^{-1} consists of all pairs (a, b) with (b, a) in the graph of f , one obtains the graph of f^{-1} from the graph of f by interchanging the horizontal and vertical axes. If f has the graph shown in Figure 2(a),

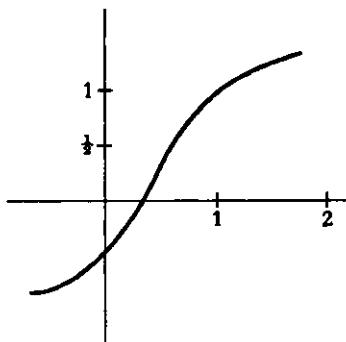
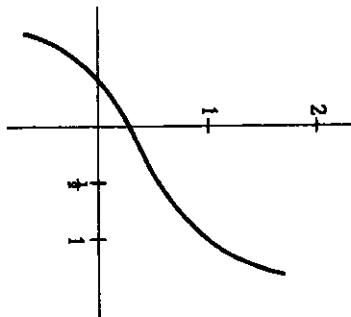


FIGURE 2(a)

and you rotate this page counter clockwise through a right angle, then the graph of f^{-1} appears on your left (Figure 2(b)). The only trouble is that the numbering on the horizontal axis goes in the wrong direction, so you must flip this picture over to get the usual picture of f^{-1} , which appears on your right (Figure 3).

FIGURE 3



This procedure is awkward with books and impossible with blackboards, so it is fortunate that there is another way of constructing the graph of f^{-1} . The points

(a, b) and (b, a) are reflections of each other through the graph of $I(x) = x$, which is called the **diagonal** (Figure 4). To obtain the graph of f^{-1} we merely reflect the graph of f through this line (Figure 5).

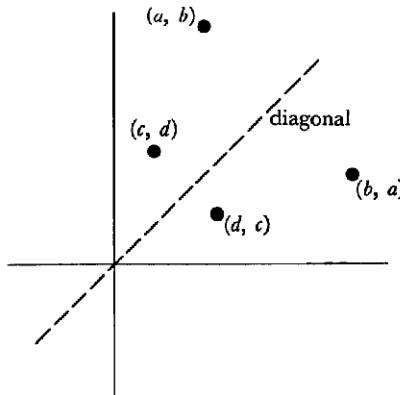


FIGURE 4

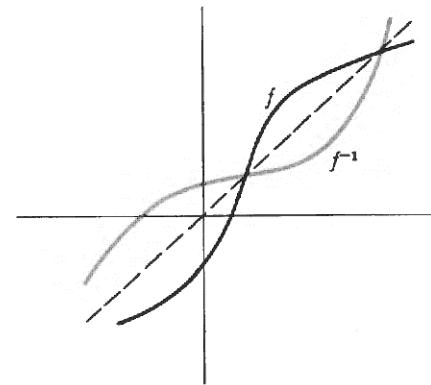


FIGURE 5

Reflecting through the diagonal twice will clearly leave us right back where we started; this means that $(f^{-1})^{-1} = f$, which is also clear from the definition. In conjunction with Theorem 1, this equation has a significant consequence: if f is a one-one function, then the function f^{-1} is also one-one (since $(f^{-1})^{-1}$ is a function).

There are a few other simple manipulations with inverse functions of which you should be aware. Since (a, b) is in f precisely when (b, a) is in f^{-1} , it follows that

$$b = f(a) \quad \text{means the same as} \quad a = f^{-1}(b).$$

Thus $f^{-1}(b)$ is the (unique) number a such that $f(a) = b$; for example, if $f(x) = x^3$, then $f^{-1}(b)$ is the unique number a such that $a^3 = b$, and this number is, by definition, $\sqrt[3]{b}$.

The fact that $f^{-1}(x)$ is the number y such that $f(y) = x$ can be restated in a much more compact form:

$$f(f^{-1}(x)) = x, \quad \text{for all } x \text{ in the domain of } f^{-1}.$$

Moreover,

$$f^{-1}(f(x)) = x, \quad \text{for all } x \text{ in the domain of } f;$$

this follows from the previous equation upon replacing f by f^{-1} . These two important equations can be written

$$\begin{aligned} f \circ f^{-1} &= I, \\ f^{-1} \circ f &= I \end{aligned}$$

(except that the right side will have a bigger domain if the domain of f or f^{-1} is not all of \mathbf{R}).

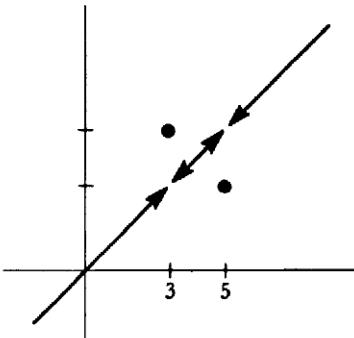


FIGURE 6

Since many standard functions will be defined as the inverses of other functions, it is quite important that we be able to tell which functions are one-one. We have already hinted which class of functions are most easily dealt with—increasing and decreasing functions are obviously one-one. Moreover, if f is increasing, then f^{-1} is also increasing, and if f is decreasing, then f^{-1} is decreasing (the proof is left to you). In addition, f is increasing if and only if $-f$ is decreasing, a very useful fact to remember.

It is certainly not true that every one-one function is either increasing or decreasing. One example has already been mentioned, and is now graphed in Figure 6:

$$g(x) = \begin{cases} x, & x \neq 3, 5 \\ 3, & x = 5 \\ 5, & x = 3. \end{cases}$$

Figure 7 shows that there are even continuous one-one functions which are neither increasing nor decreasing. But if you try drawing a few pictures you will soon agree that every one-one continuous function defined on an interval is either increasing or decreasing. It's possible to give a straightforward, but cumbersome, proof of this fact that involves keeping track of a lot of cases (very much like Problem 6(c) in the previous Appendix). The following proof dispenses with all these unpleasant details, although it is rather tricky.

THEOREM 2 If f is continuous and one-one of an interval, then f is either increasing or decreasing on that interval.

PROOF Let $a_0 < b_0$ be two numbers in the interval. Since f is one-one, we know that

$$\begin{array}{lll} \text{either} & \text{(i)} & f(b_0) - f(a_0) > 0 \\ \text{or} & \text{(ii)} & f(b_0) - f(a_0) < 0. \end{array}$$

We will assume that (i) is true, and show that the same inequality holds for any $a_1 < b_1$ in the interval, so that f is increasing. (A similar argument shows that if (ii) is true, then f is decreasing.)

Let

$$\begin{aligned} x_t &= (1-t)a_0 + ta_1 && \text{for } 0 \leq t \leq 1. \\ y_t &= (1-t)b_0 + tb_1 \end{aligned}$$

Then $x_0 = a_0$ and $x_1 = a_1$ and the points x_t all lie between a_0 and a_1 (Problem 4-2). An analogous statement holds for y_t . So x_t and y_t are all in the domain of f . Moreover, since $a_0 < b_0$ and $a_1 < b_1$, we also have

$$x_t < y_t \quad \text{for } 0 \leq t \leq 1.$$

Now consider the function

$$g(t) = f(y_t) - f(x_t) \quad \text{for } 0 \leq t \leq 1.$$

Using Theorem 6-2, it is easy to see that g is continuous on $[0, 1]$. Moreover, $g(t)$ is never 0, since $x_t < y_t$ and f is one-one. Consequently, $g(t)$ is either positive for all t in $[0, 1]$ or negative for all t in $[0, 1]$ (otherwise, by the Intermediate Value

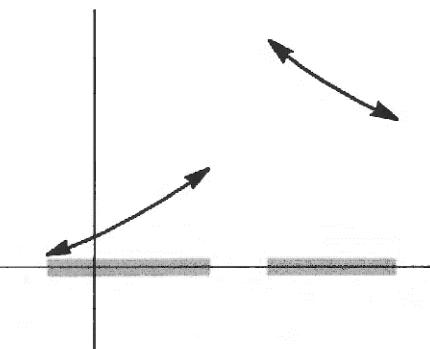


FIGURE 7

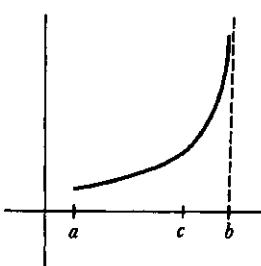


FIGURE 8

Theorem it would also be 0 somewhere in $[0, 1]$). But $g(0) > 0$ by (i). So also $g(1) > 0$, which means that (i) also holds for a_1, b_1 . ■

Henceforth we shall be concerned almost exclusively with continuous increasing or decreasing functions which are defined on an interval. If f is such a function, it is possible to say quite precisely what the domain of f^{-1} will be like.

Suppose first that f is a continuous increasing function on the closed interval $[a, b]$. Then, by the Intermediate Value Theorem, f takes on every value between $f(a)$ and $f(b)$. Therefore, the domain of f^{-1} is the closed interval $[f(a), f(b)]$. Similarly, if f is continuous and decreasing on $[a, b]$, then the domain of f^{-1} is $[f(b), f(a)]$.

If f is a continuous increasing function on an *open* interval (a, b) the analysis becomes a bit more difficult. To begin with, let us choose some point c in (a, b) . We will first decide which values $> f(c)$ are taken on by f . One possibility is that f takes on arbitrarily large values (Figure 8). In this case f takes on *all* values $> f(c)$, by the Intermediate Value Theorem. If, on the other hand, f does not take on arbitrarily large values, then $A = \{f(x) : c \leq x < b\}$ is bounded above, so A has a least upper bound α (Figure 9). Now suppose y is any number with $f(c) < y < \alpha$. Then f takes on some value $f(x) > y$ (because α is the least upper bound of A). By the Intermediate Value Theorem, f actually takes on the value y . Notice that f cannot take on the value α itself; for if $\alpha = f(x)$ for $a < x < b$ and we choose t with $x < t < b$, then $f(t) > \alpha$, which is impossible.

Precisely the same arguments work for values less than $f(c)$: either f takes on all values less than $f(c)$ or there is a number $\beta < f(c)$ such that f takes on all values between β and $f(c)$, but not β itself.

This entire argument can be repeated if f is decreasing, and if the domain of f is \mathbf{R} or (a, ∞) or $(-\infty, a)$. Summarizing: if f is a continuous increasing, or decreasing, function whose domain is an interval having one of the forms

$$(a, b), (-\infty, b), (a, \infty), \text{ or } \mathbf{R},$$

then the domain of f^{-1} is also an interval which has one of these four forms.

Now that we have completed this preliminary analysis of continuous one-one functions, it is possible to begin asking which important properties of a one-one function are inherited by its inverse. For continuity there is no problem.

THEOREM 3 If f is continuous and one-one on an interval, then f^{-1} is also continuous.

PROOF

We know by Theorem 2 that f is either increasing or decreasing. We might as well assume that f is increasing, since we can then take care of the other case by applying the usual trick of considering $-f$.

We must show that

$$\lim_{x \rightarrow b} f^{-1}(x) = f^{-1}(b)$$

for each b in the domain of f^{-1} . Such a number b is of the form $f(a)$ for some a in the domain of f . For any $\varepsilon > 0$, we want to find a $\delta > 0$ such that, for all x ,

$$\text{if } f(a) - \delta < x < f(a) + \delta, \text{ then } a - \varepsilon < f^{-1}(x) < a + \varepsilon.$$

Figure 10 suggests the way of finding δ (remember that by looking sideways you see the graph of f^{-1}): since

$$a - \varepsilon < a < a + \varepsilon,$$

it follows that

$$f(a - \varepsilon) < f(a) < f(a + \varepsilon);$$

we let δ be the smaller of $f(a + \varepsilon) - f(a)$ and $f(a) - f(a - \varepsilon)$. Figure 10 contains the entire proof that this δ works, and what follows is simply a verbal account of the information contained in this picture.

Our choice of δ ensures that

$$f(a - \varepsilon) \leq f(a) - \delta \text{ and } f(a) + \delta \leq f(a + \varepsilon).$$

Consequently, if

$$f(a) - \delta < x < f(a) + \delta,$$

then

$$f(a - \varepsilon) < x < f(a + \varepsilon).$$

Since f is increasing, f^{-1} is also increasing, and we obtain

$$f^{-1}(f(a - \varepsilon)) < f^{-1}(x) < f^{-1}(f(a + \varepsilon)),$$

i.e.,

$$a - \varepsilon < f^{-1}(x) < a + \varepsilon,$$

which is precisely what we want. ■

Having successfully investigated continuity of f^{-1} , it is only reasonable to tackle differentiability. Again, a picture indicates just what result ought to be true. Figure 11 shows the graph of a one-one function f with a tangent line L through $(a, f(a))$. If this entire picture is reflected through the diagonal, it shows the graph of f^{-1} and the tangent line L' through $(f(a), a)$. The slope of L' is the reciprocal of the slope of L . In other words, it appears that

$$(f^{-1})'(f(a)) = \frac{1}{f'(a)}.$$

This formula can equally well be written in a way which expresses $(f^{-1})'(b)$ directly, for each b in the domain of f^{-1} :

$$(f^{-1})'(b) = \frac{1}{f'(f^{-1}(b))}.$$

Unlike the argument for continuity, this pictorial “proof” becomes somewhat involved when formulated analytically. There is another approach which might

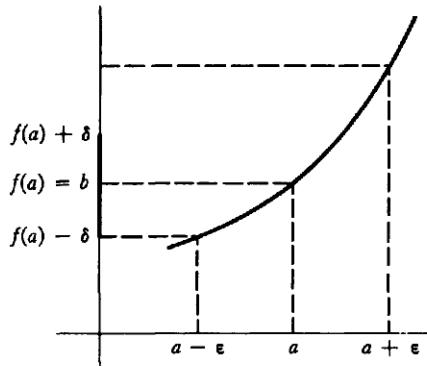


FIGURE 10

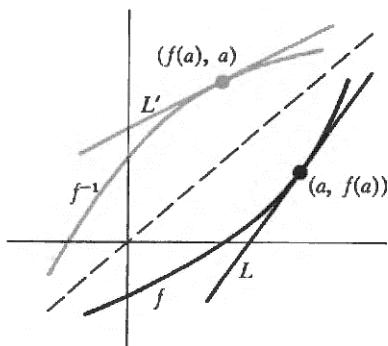


FIGURE 11

be tried. Since we know that

$$f(f^{-1}(x)) = x,$$

it is tempting to prove the desired formula by applying the Chain Rule:

$$f'(f^{-1}(x)) \cdot (f^{-1})'(x) = 1,$$

so

$$(f^{-1})'(x) = \frac{1}{f'(f^{-1}(x))}.$$

Unfortunately, this is not a proof that f^{-1} is differentiable, since the Chain Rule cannot be applied unless f^{-1} is already known to be differentiable. But this argument does show what $(f^{-1})'(x)$ will have to be if f^{-1} is differentiable, and it can also be used to obtain some important preliminary information.

THEOREM 4

PROOF

We have

$$f(f^{-1}(x)) = x.$$

If f^{-1} were differentiable at a , the Chain Rule would imply that

$$f'(f^{-1}(a)) \cdot (f^{-1})'(a) = 1,$$

hence

$$0 \cdot (f^{-1})'(a) = 1,$$

which is absurd. ■

A simple example to which Theorem 4 applies is the function $f(x) = x^3$. Since $f'(0) = 0$ and $0 = f^{-1}(0)$, the function f^{-1} is not differentiable at 0 (Figure 12).

Having decided where an inverse function cannot be differentiable, we are now ready for the rigorous proof that in all other cases the derivative is given by the formula which we have already “derived” in two different ways. Notice that the following argument uses *continuity* of f^{-1} , which we have already proved.

THEOREM 5

Let f be a continuous one-one function defined on an interval, and suppose that f is differentiable at $f^{-1}(b)$, with derivative $f'(f^{-1}(b)) \neq 0$. Then f^{-1} is differentiable at b , and

$$(f^{-1})'(b) = \frac{1}{f'(f^{-1}(b))}.$$

PROOF Let $b = f(a)$. Then

$$\begin{aligned} & \lim_{h \rightarrow 0} \frac{f^{-1}(b+h) - f^{-1}(b)}{h} \\ &= \lim_{h \rightarrow 0} \frac{f^{-1}(b+h) - a}{h} \end{aligned}$$

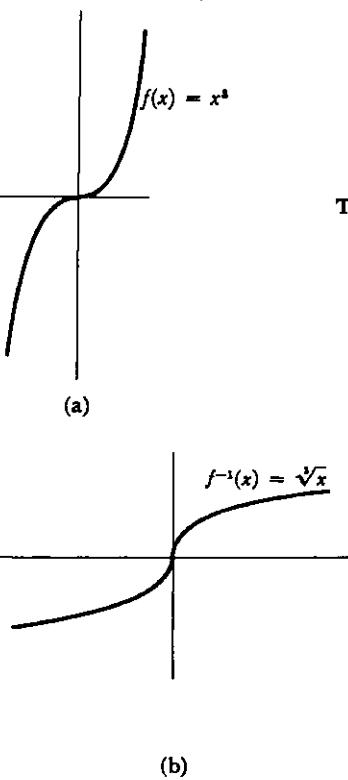


FIGURE 12

Now every number $b + h$ in the domain of f^{-1} can be written in the form

$$b + h = f(a + k)$$

for a unique k (we should really write $k(h)$, but we will stick with k for simplicity). Then

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{f^{-1}(b+h) - a}{h} &= \lim_{h \rightarrow 0} \frac{f^{-1}(f(a+k)) - a}{f(a+k) - b} \\ &= \lim_{h \rightarrow 0} \frac{k}{f(a+k) - f(a)}. \end{aligned}$$

We are clearly on the right track! It is not hard to get an explicit expression for k ; since

$$b + h = f(a + k)$$

we have

$$f^{-1}(b+h) = a + k$$

or

$$k = f^{-1}(b+h) - f^{-1}(b).$$

Now by Theorem 3, the function f^{-1} is continuous at b . This means that k approaches 0 as h approaches 0. Since

$$\lim_{k \rightarrow 0} \frac{f(a+k) - f(a)}{k} = f'(a) = f'(f^{-1}(b)) \neq 0,$$

this implies that

$$(f^{-1})'(b) = \frac{1}{f'(f^{-1}(b))}. \blacksquare$$

The work we have done on inverse functions will be amply repaid later, but here is an immediate dividend. For n odd, let

$$f_n(x) = x^n \quad \text{for all } x;$$

for n even, let

$$f_n(x) = x^n, \quad x \geq 0.$$

Then f_n is a continuous one-one function, whose inverse function is

$$g_n(x) = \sqrt[n]{x} = x^{1/n}.$$

By Theorem 5 we have, for $x \neq 0$,

$$\begin{aligned} g_n'(x) &= \frac{1}{f_n'(f_n^{-1}(x))} \\ &= \frac{1}{n(f_n^{-1}(x))^{n-1}} \\ &= \frac{1}{n(x^{1/n})^{n-1}} \\ &= \frac{1}{n} \cdot \frac{1}{x^{1-(1/n)}} \\ &= \frac{1}{n} \cdot x^{(1/n)-1}. \end{aligned}$$

Thus, if $f(x) = x^a$, and a is an integer or the reciprocal of a natural number, then $f'(x) = ax^{a-1}$. It is now easy to check that this formula is true if a is any rational number: Let $a = m/n$, where m is an integer, and n is a natural number; if

$$f(x) = x^{m/n} = (x^{1/n})^m,$$

then, by the Chain Rule,

$$\begin{aligned} f'(x) &= m(x^{1/n})^{m-1} \cdot \frac{1}{n} \cdot x^{(1/n)-1} \\ &= \frac{m}{n} \cdot x^{[(m/n)-(1/n)]+[(1/n)-1]} \\ &= \frac{m}{n} x^{(m/n)-1}. \end{aligned}$$

Although we now have a formula for $f'(x)$ when $f(x) = x^a$ and a is rational, the treatment of the function $f(x) = x^a$ for irrational a will have to be saved for later—at the moment we do not even know the *meaning* of a symbol like $x^{\sqrt{2}}$. Actually, inverse functions will be involved crucially in the definition of x^a for irrational a . Indeed, in the next few chapters several important functions will be defined in terms of their inverse functions.

PROBLEMS

1. Find f^{-1} for each of the following f .

(i) $f(x) = x^3 + 1$.
(ii) $f(x) = (x - 1)^3$.

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

(iv) $f(x) = \begin{cases} -x^2 & x \geq 0 \\ 1 - x^3, & x < 0. \end{cases}$

(v) $f(x) = \begin{cases} x, & x \neq a_1, \dots, a_n \\ a_{i+1} & x = a_i, \quad i = 1, \dots, n-1 \\ a_1, & x = a_n. \end{cases}$

(vi) $f(x) = x + [x]$.

- (vii) $f(0.a_1a_2a_3\dots) = 0.a_2a_1a_3\dots$. (Decimal representation is being used.)
- (viii) $f(x) = \frac{x}{1-x^2}$, $-1 < x < 1$.
2. Describe the graph of f^{-1} when
- f is increasing and always positive.
 - f is increasing and always negative.
 - f is decreasing and always positive.
 - f is decreasing and always negative.
3. Prove that if f is increasing, then so is f^{-1} , and similarly for decreasing functions.
4. If f and g are increasing, is $f + g$? Or $f \cdot g$? Or $f \circ g$?
5. (a) Prove that if f and g are one-one, then $f \circ g$ is also one-one. Find $(f \circ g)^{-1}$ in terms of f^{-1} and g^{-1} . Hint: The answer is *not* $f^{-1} \circ g^{-1}$.
(b) Find g^{-1} in terms of f^{-1} if $g(x) = 1 + f(x)$.
6. Show that $f(x) = \frac{ax+b}{cx+d}$ is one-one if and only if $ad - bc \neq 0$, and find f^{-1} in this case.
7. On which intervals $[a, b]$ will the following functions be one-one?
- $f(x) = x^3 - 3x^2$.
 - $f(x) = x^5 + x$.
 - $f(x) = (1+x^2)^{-1}$.
 - $f(x) = \frac{x+1}{x^2+1}$.
8. Suppose that f is differentiable with derivative $f'(x) = (1+x^3)^{-1/2}$. Show that $g = f^{-1}$ satisfies $g''(x) = \frac{3}{2}g(x)^2$.
9. Suppose that f is a one-one function and that f^{-1} has a derivative which is nowhere 0. Prove that f is differentiable. Hint: There is a one-step proof.
10. The Schwarzian derivative $\mathcal{D}f$ was defined in Problem 10-17.
- Prove that if $\mathcal{D}f(x)$ exists for all x , then $\mathcal{D}f^{-1}(x)$ also exists for all x in the domain of f^{-1} .
 - Find a formula for $\mathcal{D}f^{-1}(x)$.
- *11. (a) Prove that there is a differentiable function f such that $[f(x)]^5 + f(x) + x = 0$ for all x . Hint: Show that f can be expressed as an inverse function. The easiest way to do this is to find f^{-1} . And the easiest way to do *this* is to set $x = f^{-1}(y)$.
- (b) Find f' in terms of f , using an appropriate theorem of this chapter.
(c) Find f' in another way, by simply differentiating the equation defining f .

The function in Problem 11 is often said to be **defined implicitly** by the equation $y^5 + y + x = 0$. The situation for this equation is quite special, however. As the next problem shows, an equation does not usually define a function implicitly on the whole line, and in some regions more than one function may be defined implicitly.

12. (a) What are the two differentiable functions f which are defined implicitly on $(-1, 1)$ by the equation $x^2 + y^2 = 1$, i.e., which satisfy $x^2 + [f(x)]^2 = 1$ for all x in $(-1, 1)$? Notice that there are no solutions defined outside $[-1, 1]$.
 (b) Which functions f satisfy $x^2 + [f(x)]^2 = -1$?
 *(c) Which differentiable functions f satisfy $[f(x)]^3 - 3f(x) = x$? Hint: It will help to first draw the graph of the function $g(x) = x^3 - 3x$.

In general, determining on what intervals a differentiable function is defined implicitly by a particular equation may be a delicate affair, and is best discussed in the context of advanced calculus. If we *assume* that f is such a differentiable solution, however, then a formula for $f'(x)$ can be derived, exactly as in Problem 11(c), by differentiating both sides of the equation defining f (a process known as “implicit differentiation”):

13. (a) Apply this method to the equation $[f(x)]^2 + x^2 = 1$. Notice that your answer will involve $f(x)$; this is only to be expected, since there is more than one function defined implicitly by the equation $y^2 + x^2 = 1$.
 (b) But check that your answer works for both of the functions f found in Problem 12(a).
 (c) Apply this same method to $[f(x)]^3 - 3f(x) = x$.
 14. (a) Use implicit differentiation to find $f'(x)$ and $f''(x)$ for the functions f defined implicitly by the equation $x^3 + y^3 = 7$.
 (b) One of these functions f satisfies $f(-1) = 2$. Find $f'(-1)$ and $f''(-1)$ for this f .
 15. The collection of all points (x, y) such that $3x^3 + 4x^2y - xy^2 + 2y^3 = 4$ forms a certain curve in the plane. Find the equation of the tangent line to this curve at the point $(-1, 1)$.
 16. Leibnizian notation is particularly convenient for implicit differentiation. Because y is so consistently used as an abbreviation for $f(x)$, the equation in x and y which defines f implicitly will automatically stand for the equation which f is supposed to satisfy. How would the following computation be written in our notation?

$$\begin{aligned} y^4 + y^3 + xy &= 1, \\ 4y^3 \frac{dy}{dx} + 3y^2 \frac{dy}{dx} + y + x \frac{dy}{dx} &= 0, \\ \frac{dy}{dx} &= \frac{-y}{4y^3 + 3y^2 + x}. \end{aligned}$$

17. As long as Leibnizian notation has entered the picture, the Leibnizian notation for derivatives of inverse functions should be mentioned. If dy/dx denotes the derivative of f , then the derivative of f^{-1} is denoted by dx/dy . Write out Theorem 5 in this notation. The resulting equation will show you another reason why Leibnizian notation has such a large following. It will also explain at which point $(f^{-1})'$ is to be calculated when using the dx/dy notation. What is the significance of the following computation?

$$\begin{aligned}x &= y^n, \\y &= x^{1/n}, \\ \frac{dx^{1/n}}{dx} &= \frac{dy}{dx} = \frac{1}{\frac{dx}{dy}} = \frac{1}{ny^{n-1}}.\end{aligned}$$

18. Suppose that f is a differentiable one-one function with a nowhere zero derivative and that $f = F'$. Let $G(x) = xf^{-1}(x) - F(f^{-1}(x))$. Prove that $G'(x) = f^{-1}(x)$. (Disregarding details, this problem tells us a very interesting fact: if we know a function whose derivative is f , then we also know one whose derivative is f^{-1} . But how could anyone ever guess the function G ? Two different ways are outlined in Problems 14-17 and 19-15.)
19. Suppose h is a function such that $h'(x) = \sin^2(\sin(x + 1))$ and $h(0) = 3$. Find
- (i) $(h^{-1})'(3)$.
 - (ii) $(\beta^{-1})'(3)$, where $\beta(x) = h(x + 1)$.
20. Find a formula for $(f^{-1})''(x)$.
- *21. Prove that if $f^{(k)}(f^{-1}(x))$ exists, and is nonzero, then $(f^{-1})^{(k)}(x)$ exists.
22. (a) Prove that an increasing and a decreasing function intersect at most once.
(b) Find two continuous increasing functions f and g such that $f(x) = g(x)$ precisely when x is an integer.
(c) Find a continuous increasing function f and a continuous decreasing function g , defined on \mathbf{R} , which do not intersect at all.
- *23. (a) If f is a continuous function on \mathbf{R} and $f = f^{-1}$, prove that there is at least one x such that $f(x) = x$. (What does the condition $f = f^{-1}$ mean geometrically?)
(b) Give several examples of continuous f such that $f = f^{-1}$ and $f(x) = x$ for exactly one x . Hint: Try decreasing f , and remember the geometric interpretation. One possibility is $f(x) = -x$.
(c) Prove that if f is an increasing function such that $f = f^{-1}$, then $f(x) = x$ for all x . Hint: Although the geometric interpretation will be immediately convincing, the simplest proof (about 2 lines) is to rule out the possibilities $f(x) < x$ and $f(x) > x$.
- *24. Which functions have the property that the graph is still the graph of a function when reflected through the graph of $-I$ (the “antidiagonal”)?

25. A function f is **nondecreasing** if $f(x) \leq f(y)$ whenever $x < y$. (To be more precise we should stipulate that the domain of f be an interval.) A **nonincreasing** function is defined similarly. Caution: Some writers use “increasing” instead of “nondecreasing,” and “strictly increasing” for our “increasing.”
- Prove that if f is nondecreasing, but not increasing, then f is constant on some interval. (Beware of unintentional puns: “not increasing” is not the same as “nonincreasing.”)
 - Prove that if f is differentiable and nondecreasing, then $f'(x) \geq 0$ for all x .
 - Prove that if $f'(x) \geq 0$ for all x , then f is nondecreasing.
- *26. (a) Suppose that $f(x) > 0$ for all x , and that f is decreasing. Prove that there is a *continuous* decreasing function g such that $0 < g(x) \leq f(x)$ for all x .
- (b) Show that we can even arrange that g will satisfy $\lim_{x \rightarrow \infty} g(x)/f(x) = 0$.

APPENDIX. PARAMETRIC REPRESENTATION OF CURVES

The material in this chapter serves to emphasize something that we noticed a long time ago—a perfectly nice looking curve need not be the graph of a function (Figure 1). In other words, we may not be able to describe it as the set of all points $(x, f(x))$. Of course, we might be able to describe the curve as the set of all points $(f(x), x)$; for example, the curve in Figure 1 is the set of all points (x^2, x) . But even this trick doesn't work in most cases. It won't allow us to describe the circle, consisting of all points (x, y) with $x^2 + y^2 = 1$, or an ellipse, and it can't be used to describe a curve like the one in Figure 2.

The simplest way of describing curves in the plane in general harks back to the physical conception of a curve as the path of a particle moving in the plane. At each time t , the particle is at a certain point, which has two coordinates; to indicate the dependence of these coordinates on the time t , we can call them $u(t)$ and $v(t)$. Thus, we end up with *two* functions. Conversely, given two functions u and v , we can consider the curve consisting of all points $(u(t), v(t))$. This curve is said to be represented *parametrically* by u and v , and the pair of functions u, v is called a parametric representation of the curve. The curve represented parametrically by u and v thus consists of all pairs (x, y) with $x = u(t)$ and $y = v(t)$. It is often described briefly as “the curve $x = u(t)$, $y = v(t)$.” Notice that the graph of a function f can always be described parametrically, as the curve $x = t$, $y = f(t)$.

Instead of considering a curve in the plane as defined by two functions, we can obtain a conceptually simpler picture if we broaden our original definition of function somewhat. Instead of considering a rule which associates a number with another number, we can consider a “function c from real numbers to the plane,” i.e., a rule c that associates, to each number t , a *point in the plane*, which we can denote by $c(t)$. With this notion, a curve is just a function from some interval of real numbers to the plane.

Of course, these two different descriptions of a curve are essentially the same: A pair of (ordinary) functions u and v determines a single function c from the real numbers to the plane by the rule

$$c(t) = (u(t), v(t)),$$

and, conversely, given a function c from the real numbers to the plane, each $c(t)$ is a point in the plane, so it is a pair of numbers, which we can call $u(t)$ and $v(t)$, so that we have unique functions u and v satisfying this equation.

In Appendix 1 to Chapter 4, we used the term “vector” to describe a point in the plane. In conformity with this usage, a curve in the plane may also be called a “vector-valued function.” The conventions of that Appendix would lead us to write $c(t) = (c_1(t), c_2(t))$, but in this Appendix we'll continue to use notation like $c(t) = (u(t), v(t))$ to minimize the use of subscripts.

A simple example of a vector-valued function that is quite useful is

$$\mathbf{e}(t) = (\cos t, \sin t),$$

which goes round and round the unit circle (Figure 3).

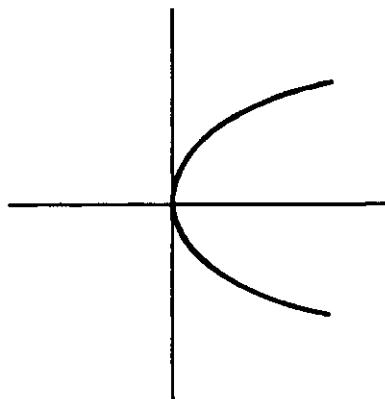


FIGURE 1

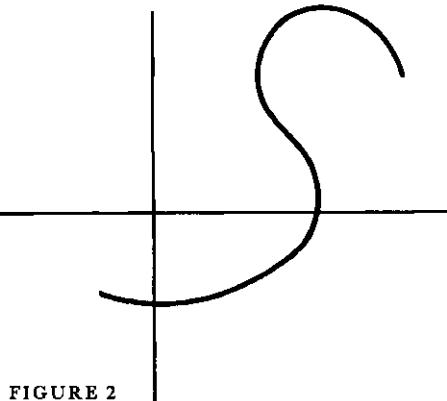


FIGURE 2

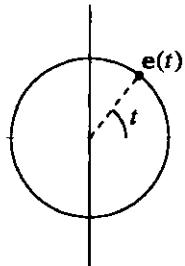


FIGURE 3

For two (ordinary) functions f and g , we defined new functions $f + g$ and $f \cdot g$ by the rules

$$(1) \quad (f + g)(x) = f(x) + g(x), \\ (2) \quad (f \cdot g)(x) = f(x) \cdot g(x).$$

Since we have defined a way of adding vectors, we can imitate the first of these definitions for vector-valued functions c and d : we define the vector-valued function $c + d$ by

$$(c + d)(t) = c(t) + d(t),$$

where the $+$ on the right-hand side is now *the sum of vectors*. This simply amounts to saying that if

$$c(t) = (u(t), v(t)), \\ d(t) = (w(t), z(t)),$$

then

$$(c + d)(t) = (u(t), v(t)) + (w(t), z(t)) = (u(t) + w(t), v(t) + z(t)).$$

Recall that we have also defined $a \cdot v$ for a number a and a vector v . To extend this to vector-valued functions, we want to consider an ordinary *function* α and a vector-valued function c , so that for each t we have a number $\alpha(t)$ and a vector $c(t)$. Then we can define a new vector-valued function $\alpha \cdot c$ by

$$(\alpha \cdot c)(t) = \alpha(t) \cdot c(t),$$

where the \cdot on the right-hand side is the product of a number and a vector. This simply amounts to saying that

$$(\alpha \cdot c)(t) = \alpha(t) \cdot (u(t), v(t)) = (\alpha(t) \cdot u(t), \alpha(t) \cdot v(t)).$$

Notice that the curve $\alpha \cdot e$,

$$(\alpha \cdot e)(t) = (\alpha(t) \cos t, \alpha(t) \sin t),$$

is already quite general (Figure 4). In the notation of Appendix 3 to Chapter 4, the point $(\alpha \cdot e)(t)$ has polar coordinates $\alpha(t)$ and t , so that $(\alpha \cdot e)(t)$ is the “graph of α in polar coordinates.”

Even more generally, given any vector-valued function c , we can define new functions r and θ by

$$c(t) = r(t) \cdot e(\theta(t)),$$

where $r(t)$ is just the distance from the origin to $c(t)$, and $\theta(t)$ is some choice of the angle of $c(t)$ (as usual, the function θ isn’t defined unambiguously, so one has to be careful when using this way of writing an arbitrary curve c).

We aren’t in a position to extend (2) to vector-valued functions in general, since we haven’t defined the product of two vectors. However, Problems 2 and 4 of Appendix 1 to Chapter 4 define two *real-valued* products $v \cdot w$ and $\det(v, w)$. It

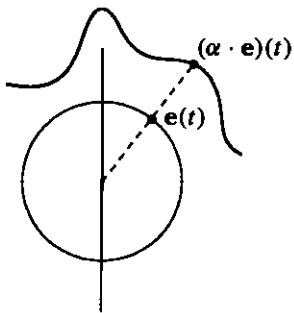


FIGURE 4

should be clear, given vector-valued functions c and d , how we would define two ordinary (real-valued) functions

$$c \cdot d \quad \text{and} \quad \det(c, d).$$

Beyond imitating simple arithmetic operations on functions, we can consider more interesting problems, like limits. For $c(t) = (u(t), v(t))$, we can define

$$(*) \quad \lim_{t \rightarrow a} c(t) = \lim_{t \rightarrow a} (u(t), v(t)) \quad \text{to be} \quad \left(\lim_{t \rightarrow a} u(t), \lim_{t \rightarrow a} v(t) \right).$$

Rules like

$$\lim_{t \rightarrow a} c + d = \lim_{t \rightarrow a} c + \lim_{t \rightarrow a} d,$$

$$\lim_{t \rightarrow a} \alpha \cdot c = \lim_{t \rightarrow a} \alpha(t) \cdot \lim_{t \rightarrow a} c$$

follow immediately. Problem 10 shows how to give an equivalent definition that imitates the basic definition of limits directly.

Limits lead us of course to derivatives. For

$$c(t) = (u(t), v(t))$$

we can define c' by the straightforward definition

$$c'(a) = (u'(a), v'(a)).$$

We could also try to imitate the basic definition:

$$c'(a) = \lim_{h \rightarrow 0} \frac{c(a+h) - c(a)}{h},$$

where the fraction on the right-hand side is understood to mean

$$\frac{1}{h} \cdot [c(a+h) - c(a)].$$

As a matter of fact, these two definitions are equivalent, because

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{c(a+h) - c(a)}{h} &= \lim_{h \rightarrow 0} \left(\frac{u(a+h) - u(a)}{h}, \frac{v(a+h) - v(a)}{h} \right) \\ &= \left(\lim_{h \rightarrow 0} \frac{u(a+h) - u(a)}{h}, \lim_{h \rightarrow 0} \frac{v(a+h) - v(a)}{h} \right) \\ &\quad \text{by our definition } (*) \text{ of limits} \\ &= (u'(a), v'(a)). \end{aligned}$$

Figure 5 shows $c(a+h)$ and $c(a)$, as well as the arrow from $c(a)$ to $c(a+h)$; as we showed in Appendix 1 to Chapter 4, this arrow is $c(a+h) - c(a)$, except moved over so that it starts at $c(a)$. As $h \rightarrow 0$, this arrow would appear to move closer and closer to the tangent of our curve, so it seems reasonable to *define* the tangent line of c at $c(a)$ to be the straight line along $c'(a)$, when $c'(a)$ is moved over so that it starts at $c(a)$. In other words, we define the tangent line of c at $c(a)$ as the set of all points

$$c(a) + s \cdot c'(a);$$

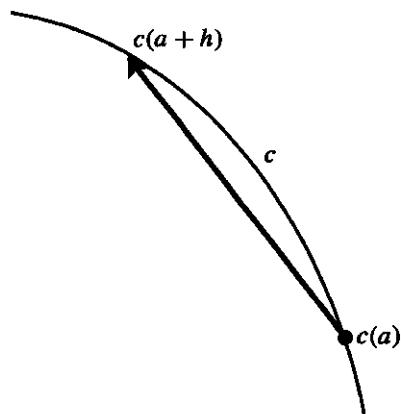


FIGURE 5

for $s = 0$ we get the point $c(a)$ itself, for $s = 1$ we get $c(a) + c'(a)$, etc. (Note, however, that this definition does not make sense when $c'(a) = 0$.) Problem 1 shows that this definition agrees with the old one when our curve c is defined by

$$c(t) = (t, f(t)),$$

so that we simply have the graph of f .

Once again, various old formulas have analogues. For example,

$$\begin{aligned}(c + d)'(a) &= c'(a) + d'(a), \\ (\alpha \cdot c)'(a) &= \alpha'(a) \cdot c(a) + \alpha(a) \cdot c'(a),\end{aligned}$$

or, as equations involving functions,

$$\begin{aligned}(c + d)' &= c' + d', \\ (\alpha \cdot c)' &= \alpha' \cdot c + \alpha \cdot c'.\end{aligned}$$

These formulas can be derived immediately from the definition in terms of the component functions. They can also be derived from the definition as a limit, by imitating previous proofs; for the second, we would of course use the standard trick of writing

$$\begin{aligned}\alpha(a + h)c(a + h) - \alpha(a)c(a) &= \\ \alpha(a + h) \cdot [c(a + h) - c(a)] + [\alpha(a + h) - \alpha(a)] \cdot c(a).\end{aligned}$$

We can also consider the function

$$d(t) = c(p(t)) = (c \circ p)(t),$$

where p is now an ordinary function, from numbers to numbers. The new curve d passes through the same points as c , except at different times; thus p corresponds to a “reparameterization” of c . For

$$\begin{aligned}c &= (u, v), \\ d &= (u \circ p, v \circ p),\end{aligned}$$

we obtain

$$\begin{aligned}d'(a) &= ((u \circ p)'(a), (v \circ p)'(a)) \\ &= (p'(a)u'(p(a)), p'(a)v'(p(a))) \\ &= p'(a) \cdot (u'(p(a)), v'(p(a))) \\ &= p'(a) \cdot c'(p(a)),\end{aligned}$$

or simply

$$d' = p' \cdot (c' \circ p).$$

Notice that if $p(a) = a$, so that d and c actually pass through the same point at time a , then $d'(a) = p'(a) \cdot c'(a)$, so that the tangent vector $d'(a)$ is just a multiple of $c'(a)$. This means that the tangent line to c at $c(a)$ is the same as the tangent line to the reparameterized curve d at $d(a) = c(a)$. The one exception occurs when $p'(a) = 0$, since the tangent line for d is then undefined, even though the

tangent line for c may be defined. For example, $d(t) = c(t^3)$ won't have a tangent line defined at $t = 0$, even though it's merely a reparameterization of c .

Finally, since we can define real-valued functions

$$(c \cdot d)(t) = c(t) \cdot d(t), \\ \det(c, d)(t) = \det(c(t), d(t)),$$

we ought to have formulas for the derivatives of these new functions. As you might guess, the proper formulas are

$$(c \cdot d)'(a) = c(a) \cdot d'(a) + c'(a) \cdot d(a), \\ [\det(c, d)]'(a) = \det(c', d)(a) + \det(c, d')(a).$$

These can be derived by straightforward calculations from the definitions in terms of the component functions. But it is more elegant to imitate the proof of the ordinary product rule, using the simple formulas in Problems 2 and 4 of Appendix 1 to Chapter 4, and, of course, the “standard trick” referred to above.

PROBLEMS

1. (a) For a function f , the “point-slope form” (Problem 4-6) of the tangent line at $(a, f(a))$ can be written as $y - f(a) = (x - a)f'(a)$, so that the tangent line consists of all points of the form

$$(x, f(a) + (x - a)f'(a)).$$

Conclude that the tangent line consists of all points of the form

$$(a + s, f(a) + sf'(a)).$$

- (b) If c is the curve $c(t) = (t, f(t))$, conclude that the tangent line of c at $(a, f(a))$ [using our new definition] is the same as the tangent line of f at $(a, f(a))$.

2. Let $c(t) = (f(t), t^2)$, where f is the function shown in Figure 21 of Chapter 9. Show that c lies along the graph of the non-differentiable function $h(x) = |x|$, but that $c'(0) = 0$. In other words, a reparameterization can “hide” a corner. For this reason, we are usually only interested in curves c with c' never 0.
3. Suppose that $x = u(t)$, $y = v(t)$ is a parametric representation of a curve, and that $u' \neq 0$ on some interval.
- Show that on this interval the curve lies along the graph of $f = v \circ u^{-1}$.
 - Show that at the point $x = u(t)$ we have

$$f'(x) = \frac{v'(t)}{u'(t)}.$$

In Leibnizian notation this is often written suggestively as

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}}.$$

(c) We also have

$$f''(x) = \frac{u'(t)v''(t) - v'(t)u''(t)}{(u'(t))^3}.$$

4. Consider a function f defined implicitly by the equation $x^{2/3} + y^{2/3} = 1$. Compute $f'(x)$ in two ways:
- (i) By implicit differentiation.
 - (ii) By considering the parametric representation $x = \cos^3 t$, $y = \sin^3 t$.
5. Let $x = u(t)$, $y = v(t)$ be the parametric representation of a curve, with u and v differentiable, and let $P = (x_0, y_0)$ be a point in the plane. Prove that if the point $Q = (u(\bar{t}), v(\bar{t}))$ on the curve is closest to (x_0, y_0) , and $u'(\bar{t})$ and $v'(\bar{t})$ are not both 0, then the line from P to Q is perpendicular to the tangent line of the curve at Q (Figure 6). The same result holds if Q is furthest from (x_0, y_0) .

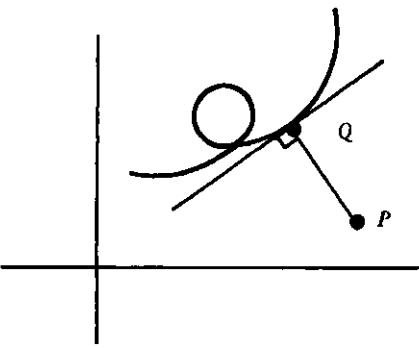


FIGURE 6

We've seen that the "graph of f in polar coordinates" is the curve

$$(f \cdot \mathbf{e})(t) = (f(t) \cos t, f(t) \sin t);$$

in other words, the graph of f in polar coordinates is the curve with the parametric representation

$$x = f(\theta) \cos \theta, \quad y = f(\theta) \sin \theta.$$

6. (a) Show that for the graph of f in polar coordinates the slope of the tangent line at the point with polar coordinates $(f(\theta), \theta)$ is

$$\frac{f(\theta) \cos \theta + f'(\theta) \sin \theta}{-f(\theta) \sin \theta + f'(\theta) \cos \theta}.$$

- (b) Show that if $f(\theta) = 0$ and f is differentiable at 0, then the line through the origin making an angle of θ with the positive horizontal axis is a tangent line of the graph of f in polar coordinates. Use this result to add some details to the graph of the Archimedean spiral in Appendix 3 of Chapter 4, and to the graphs in Problems 3 and 10 of that Appendix as well.
- (c) Suppose that the point with polar coordinates $(f(\theta), \theta)$ is further from the origin O than any other point on the graph of f . What can you say about the tangent line to the graph at this point? Compare with Problem 5.

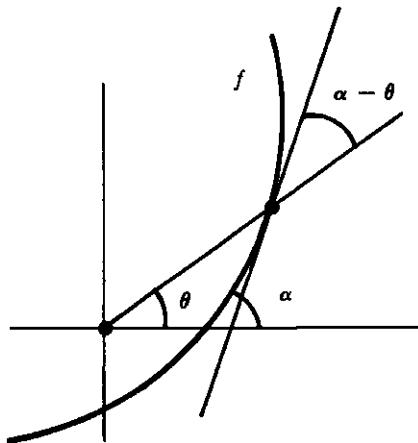


FIGURE 7

- (d) Suppose that the tangent line to the graph of f at the point with polar coordinates $(f(\theta), \theta)$ makes an angle of α with the horizontal axis (Figure 7), so that $\alpha - \theta$ is the angle between the tangent line and the ray from O to the point. Show that

$$\tan(\alpha - \theta) = \frac{f(\theta)}{f'(\theta)}.$$

7. (a) In Problem 5 of Appendix 1 to Chapter 4 we found that the cardioid $r = 1 - \sin \theta$ is also described by the equation $(x^2 + y^2 + y)^2 = x^2 + y^2$. Find the slope of the tangent line at a point on the cardioid in two ways:

- (i) By implicit differentiation.
- (ii) By using the previous problem.

- (b) Check that at the origin the tangent lines are vertical, as they appear to be in Figure 8.

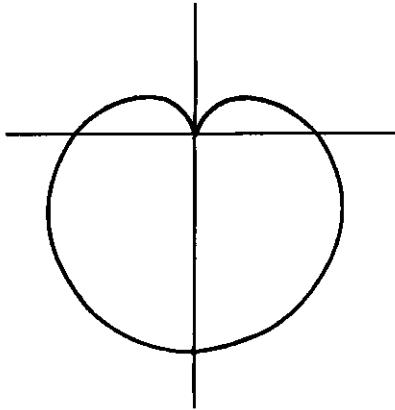


FIGURE 8

The next problem uses the material from Chapter 15, in particular, radian measure, and the inverse trigonometric functions and their properties.

8. A *cycloid* is defined as the path traced out by a point on the rim of a rolling wheel of radius a . You can see a beautiful cycloid by pasting a reflector on the edge of a bicycle wheel and having a friend ride slowly in front of the headlights of your car at night. Lacking a car, bicycle, or trusting friend, you can settle instead for Figure 9.

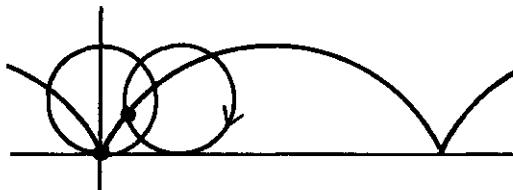


FIGURE 9

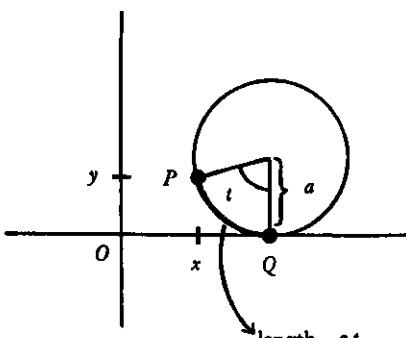


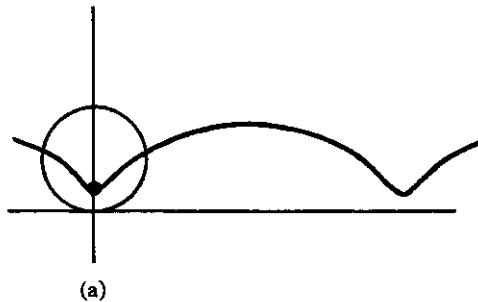
FIGURE 10

- (a) Let $u(t)$ and $v(t)$ be the coordinates of the point on the rim after the wheel has rotated through an angle of t (radians). This means that the arc of the wheel rim from P to Q in Figure 10 has length at . Since the wheel is rolling, at is also the distance from O to Q . Show that we have the parametric representation of the cycloid

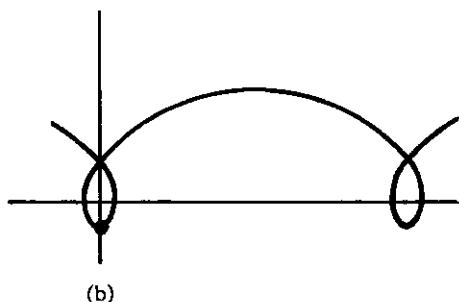
$$u(t) = a(t - \sin t)$$

$$v(t) = a(1 - \cos t).$$

Figure 11 shows the curves we obtain if the distance from the point to the center of the wheel is (a) less than the radius or (b) greater than the radius. In the latter case, the curve is not the graph of a function; at certain times the point is moving backwards, even though the wheel is moving forwards!



(a)



(b)

FIGURE 11

In Figure 9 we drew the cycloid as the graph of a function, but we really need to check that this is the case:

- (b) Compute $u'(t)$ and conclude that u is increasing. Problem 3 then shows that the cycloid is the graph of $f = v \circ u^{-1}$, and allows us to compute $f'(t)$.

It isn't possible to get an explicit formula for f , but we can come close.

- (c) Show that

$$u(t) = a \arccos \frac{a - v(t)}{a} \pm \sqrt{[2a - v(t)]v(t)}.$$

Hint: first solve for t in terms of $v(t)$.

- (d) The first half of the first arch of the cycloid is the graph of g^{-1} , where

$$g(y) = a \arccos \frac{a - y}{y} - \sqrt{(2a - y)y}.$$

9. Let u and v be continuous on $[a, b]$ and differentiable on (a, b) ; then u and v give a parametric representation of a curve from $P = (u(a), v(a))$ to $Q = (u(b), v(b))$. Geometrically, it seems clear (Figure 12) that at some point on the curve the tangent line is parallel to the line segment from P to Q . Prove this analytically. Hint: This problem will give a geometric interpretation for one of the theorems in Chapter 11.
10. The following definition of a limit for a vector-valued function is the direct analogue of the definition for ordinary functions:

$\lim_{t \rightarrow a} c(t) = l$ means that for every $\varepsilon > 0$ there is some $\delta > 0$ such that, for all t , if $0 < |t - a| < \delta$, then $\|c(t) - l\| < \varepsilon$.

Here $\| \cdot \|$ is the *norm*, defined in Problem 2 of Appendix 1 to Chapter 4. If $l = (l_1, l_2)$, then

$$\|c(t) - l\|^2 = |u(t) - l_1|^2 + |v(t) - l_2|^2.$$

- (a) Conclude that

$$|u(t) - l_1| \leq \|c(t) - l\| \quad \text{and} \quad |v(t) - l_2| \leq \|c(t) - l\|,$$

and show that if $\lim_{t \rightarrow a} c(t) = l$ according to the above definition, then we also have

$$\lim_{t \rightarrow a} u(t) = l_1 \quad \text{and} \quad \lim_{t \rightarrow a} v(t) = l_2,$$

so that $\lim_{t \rightarrow a} c(t) = l$ according to our definition (*) in terms of component functions, on page 243.

- (b) Conversely, show that if $\lim_{t \rightarrow a} c(t) = l$ according to the definition in terms of component functions, then also $\lim_{t \rightarrow a} c(t) = l$ according to the above definition.

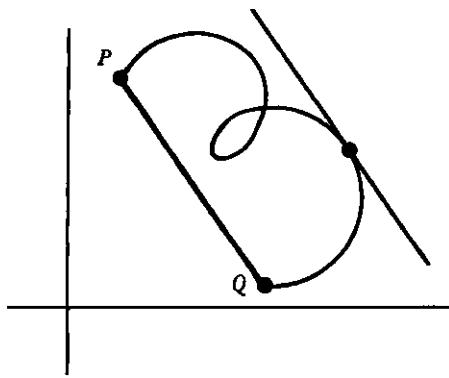


FIGURE 12

CHAPTER 13

INTEGRALS

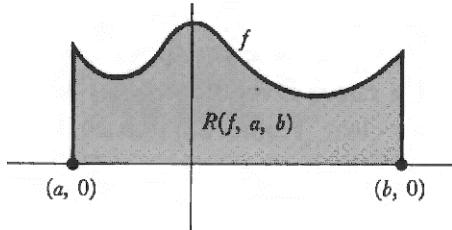


FIGURE 1

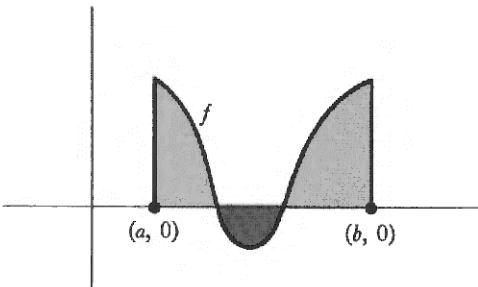


FIGURE 2

The derivative does not display its full strength until allied with the “integral,” the second main concept of Part III. At first this topic may seem to be a complete digression—in this chapter derivatives do not appear even once! The study of integrals does require a long preparation, but once this preliminary work has been completed, integrals will be an invaluable tool for creating new functions, and the derivative will reappear in Chapter 14, more powerful than ever.

Although ultimately to be defined in a quite complicated way, the integral formalizes a simple, intuitive concept—that of area. By now it should come as no surprise to learn that the definition of an intuitive concept can present great difficulties—“area” is certainly no exception.

In elementary geometry, formulas are derived for the areas of many plane figures, but a little reflection shows that an acceptable definition of area is seldom given. The area of a region is sometimes defined as the number of squares, with sides of length 1, which fit in the region. But this definition is hopelessly inadequate for any but the simplest regions. For example, a circle of radius 1 supposedly has as area the irrational number π , but it is not at all clear what “ π squares” means. Even if we consider a circle of radius $1/\sqrt{\pi}$, which supposedly has area 1, it is hard to say in what way a unit square fits in this circle, since it does not seem possible to divide the unit square into pieces which can be arranged to form a circle.

In this chapter we will only try to define the area of some very special regions (Figure 1)—those which are bounded by the horizontal axis, the vertical lines through $(a, 0)$ and $(b, 0)$, and the graph of a function f such that $f(x) \geq 0$ for all x in $[a, b]$. It is convenient to indicate this region by $R(f, a, b)$. Notice that these regions include rectangles and triangles, as well as many other important geometric figures.

The number which we will eventually assign as the area of $R(f, a, b)$ will be called the *integral* of f on $[a, b]$. Actually, the integral will be defined even for functions f which do not satisfy the condition $f(x) \geq 0$ for all x in $[a, b]$. If f is the function graphed in Figure 2, the integral will represent the difference of the area of the lightly shaded region and the area of the heavily shaded region (the “algebraic area” of $R(f, a, b)$).

The idea behind the prospective definition is indicated in Figure 3. The interval $[a, b]$ has been divided into four subintervals

$$[t_0, t_1] \quad [t_1, t_2] \quad [t_2, t_3] \quad [t_3, t_4]$$

by means of numbers t_0, t_1, t_2, t_3, t_4 with

$$a = t_0 < t_1 < t_2 < t_3 < t_4 = b$$

(the numbering of the subscripts begins with 0 so that the largest subscript will equal the number of subintervals).

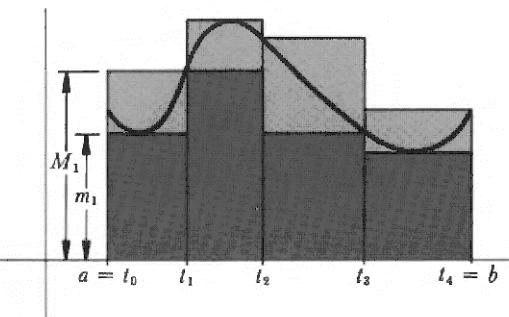


FIGURE 3

On the first interval $[t_0, t_1]$ the function f has the minimum value m_1 and the maximum value M_1 ; similarly, on the i th interval $[t_{i-1}, t_i]$ let the minimum value of f be m_i and let the maximum value be M_i . The sum

$$s = m_1(t_1 - t_0) + m_2(t_2 - t_1) + m_3(t_3 - t_2) + m_4(t_4 - t_3)$$

represents the total area of rectangles lying inside the region $R(f, a, b)$, while the sum

$$S = M_1(t_1 - t_0) + M_2(t_2 - t_1) + M_3(t_3 - t_2) + M_4(t_4 - t_3)$$

represents the total area of rectangles containing the region $R(f, a, b)$. The guiding principle of our attempt to define the area A of $R(f, a, b)$ is the observation that A should satisfy

$$s \leq A \quad \text{and} \quad A \leq S,$$

and that this should be true, *no matter how the interval $[a, b]$ is subdivided*. It is to be hoped that these requirements will determine A . The following definitions begin to formalize, and eliminate some of the implicit assumptions in, this discussion.

DEFINITION

Let $a < b$. A **partition** of the interval $[a, b]$ is a finite collection of points in $[a, b]$, one of which is a , and one of which is b .

The points in a partition can be numbered t_0, \dots, t_n so that

$$a = t_0 < t_1 < \dots < t_{n-1} < t_n = b;$$

we shall always assume that such a numbering has been assigned.

DEFINITION

Suppose f is bounded on $[a, b]$ and $P = \{t_0, \dots, t_n\}$ is a partition of $[a, b]$. Let

$$\begin{aligned} m_i &= \inf\{f(x) : t_{i-1} \leq x \leq t_i\}, \\ M_i &= \sup\{f(x) : t_{i-1} \leq x \leq t_i\}. \end{aligned}$$

The **lower sum** of f for P , denoted by $L(f, P)$, is defined as

$$L(f, P) = \sum_{i=1}^n m_i(t_i - t_{i-1}).$$

The **upper sum** of f for P , denoted by $U(f, P)$, is defined as

$$U(f, P) = \sum_{i=1}^n M_i(t_i - t_{i-1}).$$

The lower and upper sums correspond to the sums s and S in the previous example; they are supposed to represent the total areas of rectangles lying below and above the graph of f . Notice, however, that despite the geometric motivation, these sums have been defined precisely without any appeal to a concept of “area.”

Two details of the definition deserve comment. The requirement that f be bounded on $[a, b]$ is essential in order that all the m_i and M_i be defined. Note, also, that it was necessary to define the numbers m_i and M_i as inf's and sup's, rather than as minima and maxima, since f was not assumed continuous.

One thing is clear about lower and upper sums: If P is any partition, then

$$L(f, P) \leq U(f, P),$$

because

$$L(f, P) = \sum_{i=1}^n m_i(t_i - t_{i-1}),$$

$$U(f, P) = \sum_{i=1}^n M_i(t_i - t_{i-1}),$$

and for each i we have

$$m_i(t_i - t_{i-1}) \leq M_i(t_i - t_{i-1}).$$

On the other hand, something less obvious *ought* to be true: If P_1 and P_2 are any two partitions of $[a, b]$, then it should be the case that

$$L(f, P_1) \leq U(f, P_2),$$

because $L(f, P_1)$ should be \leq area $R(f, a, b)$, and $U(f, P_2)$ should be \geq area $R(f, a, b)$. This remark proves nothing (since the “area of $R(f, a, b)$ ” has not even been defined yet), but it does indicate that if there is to be any hope of defining the area of $R(f, a, b)$, a proof that $L(f, P_1) \leq U(f, P_2)$ should come first. The proof which we are about to give depends upon a lemma which concerns the behavior of lower and upper sums when more points are included in a partition. In Figure 4 the partition P contains the points in black, and Q contains both the points in black and the points in grey. The picture indicates that the rectangles drawn for the partition Q are a better approximation to the region $R(f, a, b)$ than those for the original partition P . To be precise:

LEMMA If Q contains P (i.e., if all points of P are also in Q), then

$$\begin{aligned} L(f, P) &\leq L(f, Q), \\ U(f, P) &\geq U(f, Q). \end{aligned}$$

PROOF Consider first the special case (Figure 5) in which Q contains just one more point than P :

$$\begin{aligned} P &= \{t_0, \dots, t_n\}, \\ Q &= \{t_0, \dots, t_{k-1}, u, t_k, \dots, t_n\}, \end{aligned}$$

where

$$a = t_0 < t_1 < \dots < t_{k-1} < u < t_k < \dots < t_n = b.$$

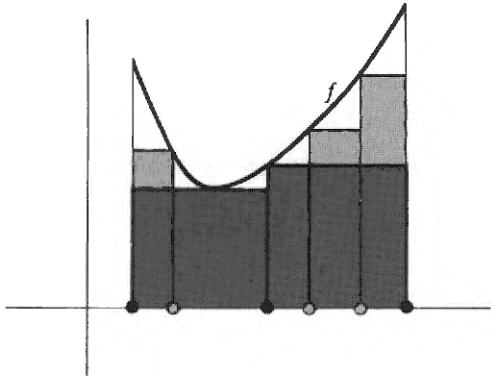


FIGURE 4

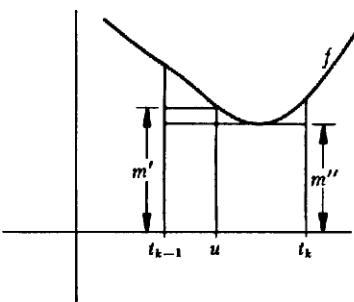


FIGURE 5

Let

$$\begin{aligned} m' &= \inf\{f(x) : t_{k-1} \leq x \leq u\}, \\ m'' &= \inf\{f(x) : u \leq x \leq t_k\}. \end{aligned}$$

Then

$$\begin{aligned} L(f, P) &= \sum_{i=1}^n m_i(t_i - t_{i-1}), \\ L(f, Q) &= \sum_{i=1}^{k-1} m_i(t_i - t_{i-1}) + m'(u - t_{k-1}) + m''(t_k - u) + \sum_{i=k+1}^n m_i(t_i - t_{i-1}). \end{aligned}$$

To prove that $L(f, P) \leq L(f, Q)$ it therefore suffices to show that

$$m_k(t_k - t_{k-1}) \leq m'(u - t_{k-1}) + m''(t_k - u).$$

Now the set $\{f(x) : t_{k-1} \leq x \leq t_k\}$ contains all the numbers in $\{f(x) : t_{k-1} \leq x \leq u\}$, and possibly some smaller ones, so the greatest lower bound of the first set is *less than or equal to* the greatest lower bound of the second; thus

$$m_k \leq m'.$$

Similarly,

$$m_k \leq m''.$$

Therefore,

$$m_k(t_k - t_{k-1}) = m_k(u - t_{k-1}) + m_k(t_k - u) \leq m'(u - t_{k-1}) + m''(t_k - u).$$

This proves, in this special case, that $L(f, P) \leq L(f, Q)$. The proof that $U(f, P) \geq U(f, Q)$ is similar, and is left to you as an easy, but valuable, exercise.

The general case can now be deduced quite easily. The partition Q can be obtained from P by adding one point at a time; in other words, there is a sequence of partitions

$$P = P_1, P_2, \dots, P_\alpha = Q$$

such that P_{j+1} contains just one more point than P_j . Then

$$L(f, P) = L(f, P_1) \leq L(f, P_2) \leq \dots \leq L(f, P_\alpha) = L(f, Q),$$

and

$$U(f, P) = U(f, P_1) \geq U(f, P_2) \geq \dots \geq U(f, P_\alpha) = U(f, Q).$$

The theorem we wish to prove is a simple consequence of this lemma.

THEOREM 1 Let P_1 and P_2 be partitions of $[a, b]$, and let f be a function which is bounded on $[a, b]$. Then

$$L(f, P_1) \leq U(f, P_2).$$

PROOF There is a partition P which contains both P_1 and P_2 (let P consist of all points in both P_1 and P_2). According to the lemma,

$$L(f, P_1) \leq L(f, P) \leq U(f, P) \leq U(f, P_2). \blacksquare$$

It follows from Theorem 1 that any upper sum $U(f, P')$ is an upper bound for the set of all lower sums $L(f, P)$. Consequently, any upper sum $U(f, P')$ is greater than or equal to the *least* upper bound of all lower sums:

$$\sup\{L(f, P) : P \text{ a partition of } [a, b]\} \leq U(f, P'),$$

for every P' . This, in turn, means that $\sup\{L(f, P)\}$ is a lower bound for the set of all upper sums of f . Consequently,

$$\sup\{L(f, P)\} \leq \inf\{U(f, P)\}.$$

It is clear that both of these numbers are between the lower sum and upper sum of f for *all* partitions:

$$\begin{aligned} L(f, P') &\leq \sup\{L(f, P)\} \leq U(f, P'), \\ L(f, P') &\leq \inf\{U(f, P)\} \leq U(f, P'), \end{aligned}$$

for all partitions P' .

It may well happen that

$$\sup\{L(f, P)\} = \inf\{U(f, P)\};$$

in this case, this is the *only* number between the lower sum and upper sum of f for all partitions, and this number is consequently an ideal candidate for the area of $R(f, a, b)$. On the other hand, if

$$\sup\{L(f, P)\} < \inf\{U(f, P)\},$$

then every number x between $\sup\{L(f, P)\}$ and $\inf\{U(f, P)\}$ will satisfy

$$L(f, P') \leq x \leq U(f, P')$$

for all partitions P' .

It is not at all clear just when such an embarrassment of riches will occur. The following two examples, although not as interesting as many which will soon appear, show that both phenomena are possible.

Suppose first that $f(x) = c$ for all x in $[a, b]$ (Figure 6). If $P = \{t_0, \dots, t_n\}$ is any partition of $[a, b]$, then

$$m_i = M_i = c,$$

so

$$L(f, P) = \sum_{i=1}^n c(t_i - t_{i-1}) = c(b - a),$$

$$U(f, P) = \sum_{i=1}^n c(t_i - t_{i-1}) = c(b - a).$$

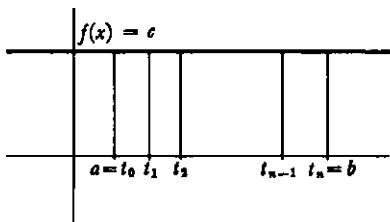


FIGURE 6

In this case, all lower sums and upper sums are equal, and

$$\sup\{L(f, P)\} = \inf\{U(f, P)\} = c(b - a).$$

Now consider (Figure 7) the function f defined by

$$f(x) = \begin{cases} 0, & x \text{ irrational} \\ 1, & x \text{ rational.} \end{cases}$$

If $P = \{t_0, \dots, t_n\}$ is any partition, then

$$m_i = 0, \text{ since there is an irrational number in } [t_{i-1}, t_i],$$

and

$$M_i = 1, \text{ since there is a rational number in } [t_{i-1}, t_i].$$

Therefore,

$$L(f, P) = \sum_{i=1}^n 0 \cdot (t_i - t_{i-1}) = 0,$$

$$U(f, P) = \sum_{i=1}^n 1 \cdot (t_i - t_{i-1}) = b - a.$$

Thus, in this case it is certainly *not* true that $\sup\{L(f, P)\} = \inf\{U(f, P)\}$. The principle upon which the definition of area was to be based provides insufficient information to determine a specific area for $R(f, a, b)$ —any number between 0 and $b - a$ seems equally good. On the other hand, the region $R(f, a, b)$ is so weird that we might with justice refuse to assign it any area at all. In fact, we can maintain, more generally, that whenever

$$\sup\{L(f, P)\} \neq \inf\{U(f, P)\},$$

the region $R(f, a, b)$ is too unreasonable to deserve having an area. As our appeal to the word “unreasonable” suggests, we are about to cloak our ignorance in terminology.

DEFINITION

A function f which is bounded on $[a, b]$ is **integrable** on $[a, b]$ if

$$\sup\{L(f, P) : P \text{ a partition of } [a, b]\} = \inf\{U(f, P) : P \text{ a partition of } [a, b]\}.$$

In this case, this common number is called the **integral** of f on $[a, b]$ and is denoted by

$$\int_a^b f.$$

(The symbol \int is called an *integral sign* and was originally an elongated *s*, for “sum;” the numbers a and b are called the *lower* and *upper limits of integration*.) The integral $\int_a^b f$ is also called the **area** of $R(f, a, b)$ when $f(x) \geq 0$ for all x in $[a, b]$.

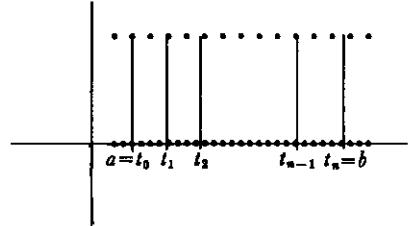


FIGURE 7

If f is integrable, then according to this definition,

$$L(f, P) \leq \int_a^b f \leq U(f, P) \quad \text{for all partitions } P \text{ of } [a, b].$$

Moreover, $\int_a^b f$ is the *unique* number with this property.

This definition merely pinpoints, and does not solve, the problem discussed before: we do not know which functions are integrable (nor do we know how to find the integral of f on $[a, b]$ when f is integrable). At present we know only two examples:

$$(1) \quad \text{if } f(x) = c, \text{ then } f \text{ is integrable on } [a, b] \text{ and } \int_a^b f = c \cdot (b - a).$$

(Notice that this integral assigns the expected area to a rectangle.)

$$(2) \quad \text{if } f(x) = \begin{cases} 0, & x \text{ irrational} \\ 1, & x \text{ rational,} \end{cases} \quad \text{then } f \text{ is not integrable on } [a, b].$$

Several more examples will be given before discussing these problems further. Even for these examples, however, it helps to have the following simple criterion for integrability stated explicitly.

THEOREM 2 If f is bounded on $[a, b]$, then f is integrable on $[a, b]$ if and only if for every $\varepsilon > 0$ there is a partition P of $[a, b]$ such that

$$U(f, P) - L(f, P) < \varepsilon.$$

PROOF Suppose first that for every $\varepsilon > 0$ there is a partition P with

$$U(f, P) - L(f, P) < \varepsilon.$$

Since

$$\inf\{U(f, P')\} \leq U(f, P), \\ \sup\{L(f, P')\} \geq L(f, P),$$

it follows that

$$\inf\{U(f, P')\} - \sup\{L(f, P')\} < \varepsilon.$$

Since this is true for all $\varepsilon > 0$, it follows that

$$\sup\{L(f, P')\} = \inf\{U(f, P')\};$$

by definition, then, f is integrable. The proof of the converse assertion is similar: If f is integrable, then

$$\sup\{L(f, P)\} = \inf\{U(f, P)\}.$$

This means that for each $\varepsilon > 0$ there are partitions P' , P'' with

$$U(f, P'') - L(f, P') < \varepsilon.$$

Let P be a partition which contains both P' and P'' . Then, according to the lemma,

$$\begin{aligned} U(f, P) &\leq U(f, P''), \\ L(f, P) &\geq L(f, P'); \end{aligned}$$

consequently,

$$U(f, P) - L(f, P) \leq U(f, P'') - L(f, P') < \varepsilon. \blacksquare$$

Although the mechanics of the proof take up a little space, it should be clear that Theorem 2 amounts to nothing more than a restatement of the definition of integrability. Nevertheless, it is a very convenient restatement because there is no mention of sup's and inf's, which are often difficult to work with. The next example illustrates this point, and also serves as a good introduction to the type of reasoning which the complicated definition of the integral necessitates, even in very simple situations.

Let f be defined on $[0, 2]$ by

$$f(x) = \begin{cases} 0, & x \neq 1 \\ 1, & x = 1. \end{cases}$$

Suppose $P = \{t_0, \dots, t_n\}$ is a partition of $[0, 2]$ with

$$t_{j-1} < 1 < t_j$$

(see Figure 8). Then

$$m_i = M_i = 0 \quad \text{if } i \neq j,$$

but

$$m_j = 0 \quad \text{and} \quad M_j = 1.$$

Since

$$\begin{aligned} L(f, P) &= \sum_{i=1}^{j-1} m_i(t_i - t_{i-1}) + m_j(t_j - t_{j-1}) + \sum_{i=j+1}^n m_i(t_i - t_{i-1}), \\ U(f, P) &= \sum_{i=1}^{j-1} M_i(t_i - t_{i-1}) + M_j(t_j - t_{j-1}) + \sum_{i=j+1}^n M_i(t_i - t_{i-1}), \end{aligned}$$

we have

$$U(f, P) - L(f, P) = t_j - t_{j-1}.$$

This certainly shows that f is integrable: to obtain a partition P with

$$U(f, P) - L(f, P) < \varepsilon,$$

it is only necessary to choose a partition with

$$t_{j-1} < 1 < t_j \quad \text{and} \quad t_j - t_{j-1} < \varepsilon.$$

Moreover, it is clear that

$$L(f, P) \leq 0 \leq U(f, P) \quad \text{for all partitions } P.$$

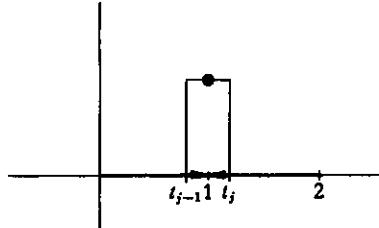


FIGURE 8

Since f is integrable, there is only *one* number between all lower and upper sums, namely, the integral of f , so

$$\int_0^2 f = 0.$$

Although the discontinuity of f was responsible for the difficulties in this example, even worse problems arise for very simple continuous functions. For example, let $f(x) = x$, and for simplicity consider an interval $[0, b]$, where $b > 0$. If $P = \{t_0, \dots, t_n\}$ is a partition of $[0, b]$, then (Figure 9)

$$m_i = t_{i-1} \quad \text{and} \quad M_i = t_i$$

and therefore

$$\begin{aligned} L(f, P) &= \sum_{i=1}^n t_{i-1}(t_i - t_{i-1}) \\ &= t_0(t_1 - t_0) + t_1(t_2 - t_1) + \cdots + t_{n-1}(t_n - t_{n-1}), \\ U(f, P) &= \sum_{i=1}^n t_i(t_i - t_{i-1}) \\ &= t_1(t_1 - t_0) + t_2(t_2 - t_1) + \cdots + t_n(t_n - t_{n-1}). \end{aligned}$$

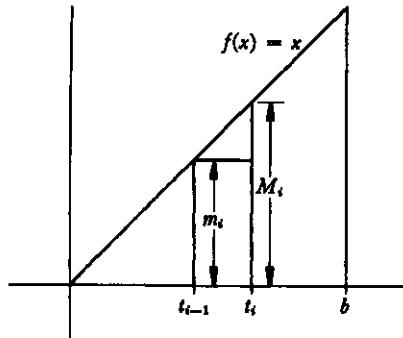


FIGURE 9

Neither of these formulas is particularly appealing, but both simplify considerably for partitions $P_n = \{t_0, \dots, t_n\}$ into n *equal* subintervals. In this case, the length $t_i - t_{i-1}$ of each subinterval is b/n , so

$$\begin{aligned} t_0 &= 0, \\ t_1 &= \frac{b}{n}, \\ t_2 &= \frac{2b}{n}, \text{ etc;} \end{aligned}$$

in general

$$t_i = \frac{ib}{n}.$$

Then

$$\begin{aligned} L(f, P_n) &= \sum_{i=1}^n t_{i-1}(t_i - t_{i-1}) \\ &= \sum_{i=1}^n \left\{ \frac{(i-1)b}{n} \right\} \cdot \frac{b}{n} \\ &= \left[\sum_{i=1}^n (i-1) \right] \frac{b^2}{n^2} \\ &= \left(\sum_{j=0}^{n-1} j \right) \frac{b^2}{n^2}. \end{aligned}$$

Remembering the formula

$$1 + \cdots + k = \frac{k(k+1)}{2},$$

this can be written

$$\begin{aligned} L(f, P_n) &= \frac{(n-1)(n)}{2} \cdot \frac{b^2}{n^2} \\ &= \frac{n-1}{n} \cdot \frac{b^2}{2}. \end{aligned}$$

Similarly,

$$\begin{aligned} U(f, P_n) &= \sum_{i=1}^n t_i(t_i - t_{i-1}) \\ &= \sum_{i=1}^n \frac{ib}{n} \cdot \frac{b}{n} \\ &= \frac{n(n+1)}{2} \cdot \frac{b^2}{n^2} \\ &= \frac{n+1}{n} \cdot \frac{b^2}{2}. \end{aligned}$$

If n is very large, both $L(f, P_n)$ and $U(f, P_n)$ are close to $b^2/2$, and this remark makes it easy to show that f is integrable. Notice first that

$$U(f, P_n) - L(f, P_n) = \frac{2}{n} \cdot \frac{b^2}{2}.$$

This shows that there are partitions P_n with $U(f, P_n) - L(f, P_n)$ as small as desired. By Theorem 2 the function f is integrable. Moreover, $\int_0^b f$ may now be found with only a little work. It is clear, first of all, that

$$L(f, P_n) \leq \frac{b^2}{2} \leq U(f, P_n) \quad \text{for all } n.$$

This inequality shows only that $b^2/2$ lies between certain special upper and lower sums, but we have just seen that $U(f, P_n) - L(f, P_n)$ can be made as small as desired, so there is *only one* number with this property. Since the integral certainly has this property, we can conclude that

$$\int_0^b f = \frac{b^2}{2}.$$

Notice that this equation assigns area $b^2/2$ to a right triangle with base and altitude b (Figure 10). Using more involved calculations, or appealing to Theorem 4, it can be shown that

$$\int_a^b f = \frac{b^2}{2} - \frac{a^2}{2}.$$

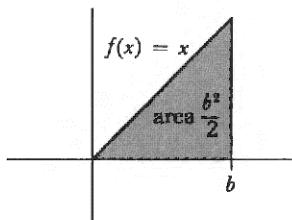


FIGURE 10

The function $f(x) = x^2$ presents even greater difficulties. In this case (Figure 11), if $P = \{t_0, \dots, t_n\}$ is a partition of $[0, b]$, then

$$m_i = f(t_{i-1}) = (t_{i-1})^2 \quad \text{and} \quad M_i = f(t_i) = t_i^2.$$

Choosing, once again, a partition $P_n = \{t_0, \dots, t_n\}$ into n equal parts, so that

$$t_i = \frac{i \cdot b}{n}$$

the lower and upper sums become

$$\begin{aligned} L(f, P_n) &= \sum_{i=1}^n (t_{i-1})^2 \cdot (t_i - t_{i-1}) \\ &= \sum_{i=1}^n (i-1)^2 \frac{b^2}{n^2} \cdot \frac{b}{n} \\ &= \frac{b^3}{n^3} \cdot \sum_{j=0}^{n-1} j^2, \end{aligned}$$

$$\begin{aligned} U(f, P_n) &= \sum_{i=1}^n t_i^2 \cdot (t_i - t_{i-1}) \\ &= \sum_{i=1}^n i^2 \frac{b^2}{n^2} \cdot \frac{b}{n} \\ &= \frac{b^3}{n^3} \sum_{j=1}^n j^2. \end{aligned}$$

Recalling the formula

$$1^2 + \dots + k^2 = \frac{1}{6}k(k+1)(2k+1)$$

from Problem 2-1, these sums can be written as

$$\begin{aligned} L(f, P_n) &= \frac{b^3}{n^3} \cdot \frac{1}{6}(n-1)n(2n-1), \\ U(f, P_n) &= \frac{b^3}{n^3} \cdot \frac{1}{6}(n+1)(2n+1). \end{aligned}$$

It is not too hard to show that

$$L(f, P_n) \leq \frac{b^3}{3} \leq U(f, P_n),$$

and that $U(f, P_n) - L(f, P_n)$ can be made as small as desired, by choosing n sufficiently large. The same sort of reasoning as before then shows that

$$\int_0^b f = \frac{b^3}{3}.$$

This calculation already represents a nontrivial result—the area of the region bounded by a parabola is not usually derived in elementary geometry. Nevertheless, the result was known to Archimedes, who derived it in essentially the same

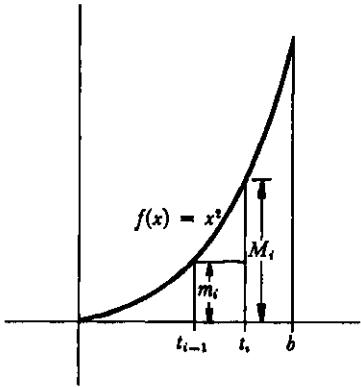


FIGURE 11

way. The only superiority we can claim is that in the next chapter we will discover a much simpler way to arrive at this result.

Some of our investigations can be summarized as follows:

$$\int_a^b f = c \cdot (b - a) \quad \text{if } f(x) = c \text{ for all } x,$$

$$\int_a^b f = \frac{b^2}{2} - \frac{a^2}{2} \quad \text{if } f(x) = x \text{ for all } x,$$

$$\int_a^b f = \frac{b^3}{3} - \frac{a^3}{3} \quad \text{if } f(x) = x^2 \text{ for all } x.$$

This list already reveals that the notation $\int_a^b f$ suffers from the lack of a convenient notation for naming functions defined by formulas. For this reason an alternative notation,* analogous to the notation $\lim_{x \rightarrow a} f(x)$, is also useful:

$$\int_a^b f(x) dx \quad \text{means precisely the same as} \quad \int_a^b f.$$

Thus

$$\int_a^b c dx = c \cdot (b - a),$$

$$\int_a^b x dx = \frac{b^2}{2} - \frac{a^2}{2},$$

$$\int_a^b x^2 dx = \frac{b^3}{3} - \frac{a^3}{3}.$$

Notice that, as in the notation $\lim_{x \rightarrow a} f(x)$, the symbol x can be replaced by any other letter (except f , a , or b , of course):

$$\int_a^b f(x) dx = \int_a^b f(t) dt = \int_a^b f(\alpha) d\alpha = \int_a^b f(y) dy = \int_a^b f(c) dc.$$

The symbol dx has no meaning in isolation, any more than the symbol $x \rightarrow$ has any meaning, except in the context $\lim_{x \rightarrow a} f(x)$. In the equation

$$\int_a^b x^2 dx = \frac{b^3}{3} - \frac{a^3}{3},$$

*The notation $\int_a^b f(x) dx$ is actually the older, and was for many years the only, symbol for the integral. Leibniz used this symbol because he considered the integral to be the sum (denoted by \int) of infinitely many rectangles with height $f(x)$ and “infinitely small” width dx . Later writers used x_0, \dots, x_n to denote the points of a partition, and abbreviated $x_i - x_{i-1}$ by Δx_i . The integral was defined as the limit as Δx_i approaches 0 of the sums $\sum_{i=1}^n f(x_i) \Delta x_i$ (analogous to lower and upper sums). The fact that the limit is obtained by changing \sum to \int , $f(x_i)$ to $f(x)$, and Δx_i to dx , delights many people.

the *entire* symbol $x^2 dx$ may be regarded as an abbreviation for:

the function f such that $f(x) = x^2$ for all x .

This notation for the integral is as flexible as the notation $\lim_{x \rightarrow a} f(x)$. Several examples may aid in the interpretation of various types of formulas which frequently appear; we have made use of Theorems 5 and 6.*

$$(1) \quad \int_a^b (x + y) dx = \int_a^b x dx + \int_a^b y dx = \frac{b^2}{2} - \frac{a^2}{2} + y(b - a).$$

$$(2) \quad \int_a^x (y + t) dy = \int_a^x y dy + \int_a^x t dy = \frac{x^2}{2} - \frac{a^2}{2} + t(x - a).$$

$$\begin{aligned} (3) \quad \int_a^b \left(\int_a^x (1 + t) dz \right) dx &= \int_a^b (1 + t)(x - a) dx \\ &= (1 + t) \int_a^b (x - a) dx \\ &= (1 + t) \left[\frac{b^2}{2} - \frac{a^2}{2} - a(b - a) \right]. \end{aligned}$$

$$\begin{aligned} (4) \quad \int_a^b \left(\int_c^d (x + y) dy \right) dx &= \int_a^b \left[x(d - c) + \frac{d^2}{2} - \frac{c^2}{2} \right] dx \\ &= \left(\frac{d^2}{2} - \frac{c^2}{2} \right) (b - a) + (d - c) \int_a^b x dx \\ &= \left(\frac{d^2}{2} - \frac{c^2}{2} \right) (b - a) + (d - c) \left(\frac{b^2}{2} - \frac{a^2}{2} \right). \end{aligned}$$

The computations of $\int_a^b x dx$ and $\int_a^b x^2 dx$ may suggest that evaluating integrals is generally difficult or impossible. As a matter of fact, the integrals of most functions *are* impossible to determine exactly (*although they may be computed to any degree of accuracy desired by calculating lower and upper sums*). Nevertheless, as we shall see in the next chapter, the integral of many functions can be computed very easily.

Even though most integrals cannot be computed exactly, it is important at least to know when a function f is integrable on $[a, b]$. Although it is possible to say precisely which functions are integrable, the criterion for integrability is a little too difficult to be stated here, and we will have to settle for partial results. The next Theorem gives the most useful result, but the proof given here uses material from the Appendix to Chapter 8. If you prefer, you can wait until the end of the next chapter, when a totally different proof will be given.

*Lest chaos overtake the reader when consulting other books, equation (1) requires an important qualification. This equation interprets $\int_a^b y dx$ to mean the integral of the function f such that each value $f(x)$ is the number y . But classical notation often uses y for $y(x)$, so $\int_a^b y dx$ might mean the integral of some arbitrary *function* y .

THEOREM 3 If f is continuous on $[a, b]$, then f is integrable on $[a, b]$.

PROOF Notice, first, that f is bounded on $[a, b]$, because it is continuous on $[a, b]$. To prove that f is integrable on $[a, b]$, we want to use Theorem 2, and show that for every $\varepsilon > 0$ there is a partition P of $[a, b]$ such that

$$U(f, P) - L(f, P) < \varepsilon.$$

Now we know, by Theorem 1 of the Appendix to Chapter 8, that f is uniformly continuous on $[a, b]$. So there is some $\delta > 0$ such that for all x and y in $[a, b]$,

$$\text{if } |x - y| < \delta, \text{ then } |f(x) - f(y)| < \frac{\varepsilon}{2(b - a)}.$$

The trick is simply to choose a partition $P = \{t_0, \dots, t_n\}$ such that each $|t_i - t_{i-1}| < \delta$. Then for each i we have

$$|f(x) - f(y)| < \frac{\varepsilon}{2(b - a)} \quad \text{for all } x, y \text{ in } [t_{i-1}, t_i],$$

and it follows easily that

$$M_i - m_i \leq \frac{\varepsilon}{2(b - a)} < \frac{\varepsilon}{b - a}.$$

Since this is true for all i , we then have

$$\begin{aligned} U(f, P) - L(f, P) &= \sum_{i=1}^n (M_i - m_i)(t_i - t_{i-1}) \\ &< \frac{\varepsilon}{b - a} \sum_{i=1}^n t_i - t_{i-1} \\ &= \frac{\varepsilon}{b - a} \cdot b - a \\ &= \varepsilon, \end{aligned}$$

which is what we wanted. ■

Although this theorem will provide all the information necessary for the use of integrals in this book, it is more satisfying to have a somewhat larger supply of integrable functions. Several problems treat this question in detail. It will help to know the following three theorems, which show that f is integrable on $[a, b]$, if it is integrable on $[a, c]$ and $[c, b]$; that $f + g$ is integrable if f and g are; and that $c \cdot f$ is integrable if f is integrable and c is any number.

As a simple application of these theorems, recall that if f is 0 except at one point, where its value is 1, then f is integrable. Multiplying this function by c , it follows that the same is true if the value of f at the exceptional point is c . Adding such a function to an integrable function, we see that the value of an integrable function may be changed arbitrarily at one point without destroying integrability. By breaking up the interval into many subintervals, we see that the value can be changed at finitely many points.

The proofs of these theorems usually use the alternative criterion for integrability in Theorem 2; as some of our previous demonstrations illustrate, the details of the

argument often conspire to obscure the point of the proof. It is a good idea to attempt proofs of your own, consulting those given here as a last resort, or as a check. This will probably clarify the proofs, and will certainly give good practice in the techniques used in some of the problems.

THEOREM 4

Let $a < c < b$. If f is integrable on $[a, b]$, then f is integrable on $[a, c]$ and on $[c, b]$. Conversely, if f is integrable on $[a, c]$ and on $[c, b]$, then f is integrable on $[a, b]$. Finally, if f is integrable on $[a, b]$, then

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

PROOF

Suppose f is integrable on $[a, b]$. If $\varepsilon > 0$, there is a partition $P = \{t_0, \dots, t_n\}$ of $[a, b]$ such that

$$U(f, P) - L(f, P) < \varepsilon.$$

We might as well assume that $c = t_j$ for some j . (Otherwise, let Q be the partition which contains t_0, \dots, t_n and c ; then Q contains P , so $U(f, Q) - L(f, Q) \leq U(f, P) - L(f, P) < \varepsilon$.)

Now $P' = \{t_0, \dots, t_j\}$ is a partition of $[a, c]$ and $P'' = \{t_j, \dots, t_n\}$ is a partition of $[c, b]$ (Figure 12). Since

$$\begin{aligned} L(f, P) &= L(f, P') + L(f, P''), \\ U(f, P) &= U(f, P') + U(f, P''), \end{aligned}$$

we have

$$[U(f, P') - L(f, P')] + [U(f, P'') - L(f, P'')] = U(f, P) - L(f, P) < \varepsilon.$$

Since each of the terms in brackets is nonnegative, each is less than ε . This shows that f is integrable on $[a, c]$ and $[c, b]$. Note also that

$$L(f, P') \leq \int_a^c f \leq U(f, P'),$$

$$L(f, P'') \leq \int_c^b f \leq U(f, P''),$$

so that

$$L(f, P) \leq \int_a^c f + \int_c^b f \leq U(f, P).$$

Since this is true for any P , this proves that

$$\int_a^c f + \int_c^b f = \int_a^b f.$$

Now suppose that f is integrable on $[a, c]$ and on $[c, b]$. If $\varepsilon > 0$, there is a partition P' of $[a, c]$ and a partition P'' of $[c, b]$ such that

$$\begin{aligned} U(f, P') - L(f, P') &< \varepsilon/2, \\ U(f, P'') - L(f, P'') &< \varepsilon/2. \end{aligned}$$

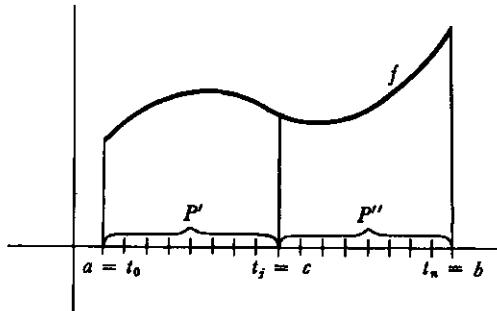


FIGURE 12

If P is the partition of $[a, b]$ containing all the points of P' and P'' , then

$$\begin{aligned} L(f, P) &= L(f, P') + L(f, P''), \\ U(f, P) &= U(f, P') + U(f, P''); \end{aligned}$$

consequently,

$$U(f, P) - L(f, P) = [U(f, P') - L(f, P')] + [U(f, P'') - L(f, P'')] < \varepsilon. \blacksquare$$

Theorem 4 is the basis for some minor notational conventions. The integral $\int_a^b f$ was defined only for $a < b$. We now add the definitions

$$\int_a^a f = 0 \quad \text{and} \quad \int_a^b f = - \int_b^a f \quad \text{if } a > b.$$

With these definitions, the equation $\int_a^c f + \int_c^b f = \int_a^b f$ holds for all a, c, b even if $a < c < b$ is not true (the proof of this assertion is a rather tedious case-by-case check).

THEOREM 5 If f and g are integrable on $[a, b]$, then $f + g$ is integrable on $[a, b]$ and

$$\int_a^b (f + g) = \int_a^b f + \int_a^b g.$$

PROOF Let $P = \{t_0, \dots, t_n\}$ be any partition of $[a, b]$. Let

$$\begin{aligned} m_i &= \inf\{(f + g)(x) : t_{i-1} \leq x \leq t_i\}, \\ m'_i &= \inf\{f(x) : t_{i-1} \leq x \leq t_i\}, \\ m''_i &= \inf\{g(x) : t_{i-1} \leq x \leq t_i\}, \end{aligned}$$

and define M_i, M'_i, M''_i similarly. It is not necessarily true that

$$m_i = m'_i + m''_i,$$

but it is true (Problem 10) that

$$m_i \geq m'_i + m''_i.$$

Similarly,

$$M_i \leq M'_i + M''_i.$$

Therefore,

$$L(f, P) + L(g, P) \leq L(f + g, P)$$

and

$$U(f + g, P) \leq U(f, P) + U(g, P).$$

Thus,

$$L(f, P) + L(g, P) \leq L(f + g, P) \leq U(f + g, P) \leq U(f, P) + U(g, P).$$

Since f and g are integrable, there are partitions P', P'' with

$$\begin{aligned} U(f, P') - L(f, P') &< \varepsilon/2, \\ U(g, P'') - L(g, P'') &< \varepsilon/2. \end{aligned}$$

If P contains both P' and P'' , then

$$U(f, P) + U(g, P) - [L(f, P) + L(g, P)] < \varepsilon,$$

and consequently

$$U(f + g, P) - L(f + g, P) < \varepsilon.$$

This proves that $f + g$ is integrable on $[a, b]$. Moreover,

$$\begin{aligned} (1) \quad L(f, P) + L(g, P) &\leq L(f + g, P) \\ &\leq \int_a^b (f + g) \\ &\leq U(f + g, P) \leq U(f, P) + U(g, P); \end{aligned}$$

and also

$$(2) \quad L(f, P) + L(g, P) \leq \int_a^b f + \int_a^b g \leq U(f, P) + U(g, P).$$

Since $U(f, P) - L(f, P)$ and $U(g, P) - L(g, P)$ can both be made as small as desired, it follows that

$$U(f, P) + U(g, P) - [L(f, P) + L(g, P)]$$

can also be made as small as desired; it therefore follows from (1) and (2) that

$$\int_a^b (f + g) = \int_a^b f + \int_a^b g. \blacksquare$$

THEOREM 6 If f is integrable on $[a, b]$, then for any number c , the function cf is integrable on $[a, b]$ and

$$\int_a^b cf = c \cdot \int_a^b f.$$

PROOF The proof (which is much easier than that of Theorem 5) is left to you. It is a good idea to treat separately the cases $c \geq 0$ and $c \leq 0$. Why? \blacksquare

(Theorem 6 is just a special case of the more general theorem that $f \cdot g$ is integrable on $[a, b]$, if f and g are, but this result is quite hard to prove (see Problem 38).)

In this chapter we have acquired only one complicated definition, a few simply theorems with intricate proofs, and one theorem which required material from the Appendix to Chapter 8. This is not because integrals constitute a more difficult topic than derivatives, but because powerful tools developed in previous chapters have been allowed to remain dormant. The most significant discovery of calculus is the fact that the integral and the derivative are intimately related—once we learn the connection, the integral will become as useful as the derivative, and as easy to use. The connection between derivatives and integrals deserves a separate chapter, but the preparations which we will make in this chapter may serve as a hint. We first state a simple inequality concerning integrals, which plays a role in many important theorems.

THEOREM 7 Suppose f is integrable on $[a, b]$ and that

$$m \leq f(x) \leq M \quad \text{for all } x \text{ in } [a, b].$$

Then

$$m(b-a) \leq \int_a^b f \leq M(b-a).$$

PROOF It is clear that

$$m(b-a) \leq L(f, P) \quad \text{and} \quad U(f, P) \leq M(b-a)$$

for every partition P . Since $\int_a^b f = \sup\{L(f, P)\} = \inf\{U(f, P)\}$, the desired inequality follows immediately. ■

Suppose now that f is integrable on $[a, b]$. We can define a new function F on $[a, b]$ by

$$F(x) = \int_a^x f = \int_a^x f(t) dt.$$

(This depends on Theorem 4.) We have seen that f may be integrable even if it is not continuous, and the Problems give examples of integrable functions which are quite pathological. The behavior of F is therefore a very pleasant surprise.

THEOREM 8 If f is integrable on $[a, b]$ and F is defined on $[a, b]$ by

$$F(x) = \int_a^x f,$$

then F is continuous on $[a, b]$.

PROOF Suppose c is in $[a, b]$. Since f is integrable on $[a, b]$ it is, by definition, bounded on $[a, b]$; let M be a number such that

$$|f(x)| \leq M \quad \text{for all } x \text{ in } [a, b].$$

If $h > 0$, then (Figure 13)

$$F(c+h) - F(c) = \int_a^{c+h} f - \int_a^c f = \int_c^{c+h} f.$$

Since

$$-M \leq f(x) \leq M \quad \text{for all } x,$$

it follows from Theorem 7 that

$$-M \cdot h \leq \int_c^{c+h} f \leq Mh;$$

in other words,

$$(1) \quad -M \cdot h \leq F(c+h) - F(c) \leq M \cdot h.$$

If $h < 0$, a similar inequality can be derived: Note that

$$F(c+h) - F(c) = \int_c^{c+h} f = - \int_{c+h}^c f.$$

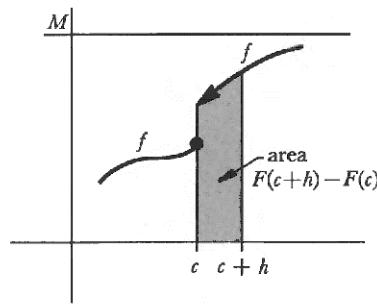


FIGURE 13

Applying Theorem 7 to the interval $[c+h, c]$, of length $-h$, we obtain

$$Mh \leq \int_{c+h}^c f \leq -Mh;$$

multiplying by -1 , which reverses all the inequalities, we have

$$(2) \quad Mh \leq F(c+h) - F(c) \leq -Mh.$$

Inequalities (1) and (2) can be combined:

$$|F(c+h) - F(c)| \leq M \cdot |h|.$$

Therefore, if $\varepsilon > 0$, we have

$$|F(c+h) - F(c)| < \varepsilon,$$

provided that $|h| < \varepsilon/M$. This proves that

$$\lim_{h \rightarrow 0} F(c+h) = F(c);$$

in other words F is continuous at c . ■

Figure 14 compares f and $F(x) = \int_a^x f$ for various functions f ; it appears that F is always better behaved than f . In the next chapter we will see how true this is.

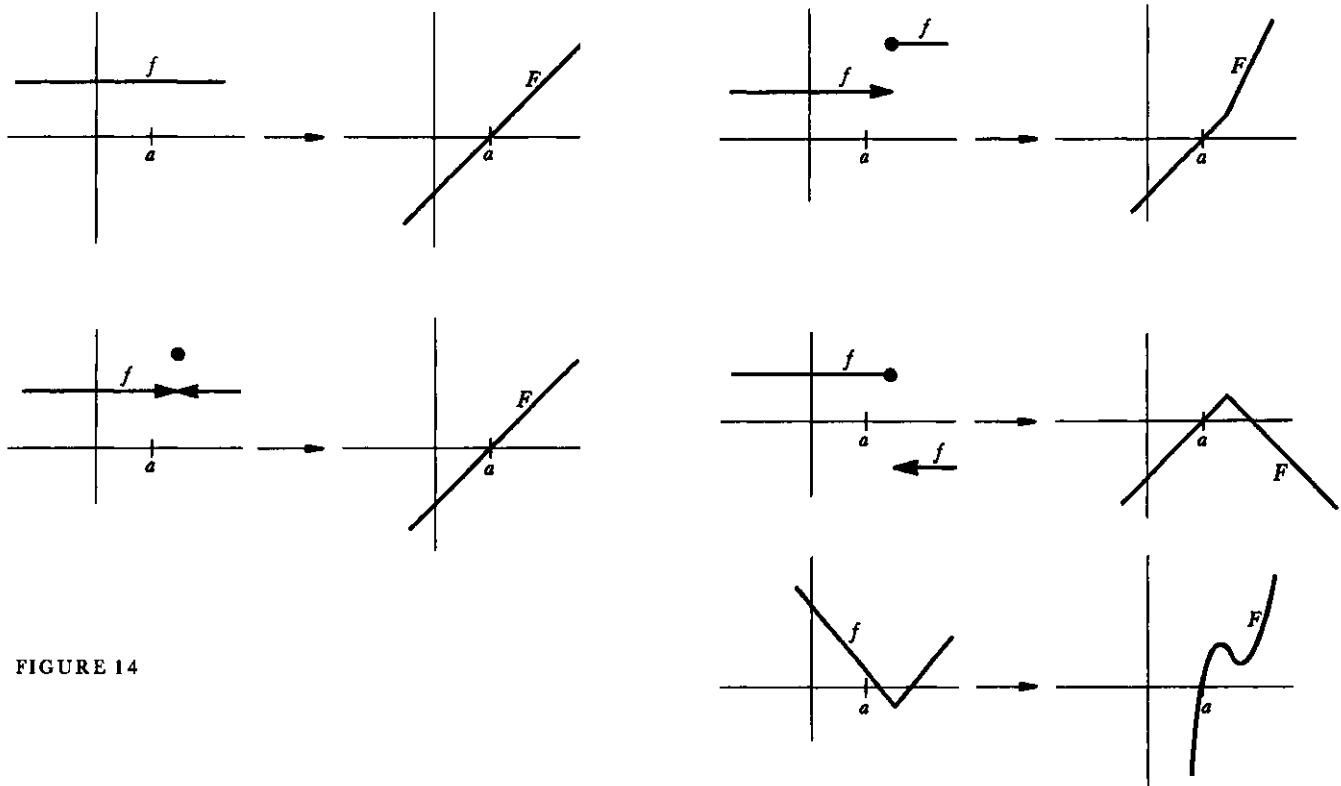


FIGURE 14

PROBLEMS

1. Prove that $\int_0^b x^3 dx = b^4/4$, by considering partitions into n equal subintervals, using the formula for $\sum_{i=1}^n i^3$ which was found in Problem 2-6. This problem requires only a straightforward imitation of calculations in the text, but you should write it up as a formal proof to make certain that all the fine points of the argument are clear.
2. Prove, similarly, that $\int_0^b x^4 dx = b^5/5$.
- *3. (a) Using Problem 2-7, show that the sum $\sum_{k=1}^n k^p/n^{p+1}$ can be made as close to $1/(p+1)$ as desired, by choosing n large enough.
 (b) Prove that $\int_0^b x^p dx = b^{p+1}/(p+1)$.
- *4. This problem outlines a clever way to find $\int_a^b x^p dx$ for $0 < a < b$. (The result for $a = 0$ will then follow by continuity.) The trick is to use partitions $P = \{t_0, \dots, t_n\}$ for which all ratios $r = t_i/t_{i-1}$ are equal, instead of using partitions for which all differences $t_i - t_{i-1}$ are equal.
 - (a) Show that for such a partition P we have

$$t_i = a \cdot c^{i/n} \quad \text{for } c = \frac{b}{a}.$$
 - (b) If $f(x) = x^p$, show, using the formula in Problem 2-5, that

$$\begin{aligned} U(f, P) &= a^{p+1}(1 - c^{-1/n}) \sum_{i=1}^n (c^{(p+1)/n})^i \\ &= (a^{p+1} - b^{p+1})c^{(p+1)/n} \frac{1 - c^{-1/n}}{1 - c^{(p+1)/n}} \\ &= (b^{p+1} - a^{p+1})c^{p/n} \cdot \frac{1}{1 + c^{1/n} + \dots + c^{p/n}} \end{aligned}$$
 and find a similar formula for $L(f, P)$.
 - (c) Conclude that

$$\int_a^b x^p dx = \frac{b^{p+1} - a^{p+1}}{p+1}.$$
5. Evaluate without doing any computations:
 - (i) $\int_{-1}^1 x^3 \sqrt{1 - x^2} dx.$
 - (ii) $\int_{-1}^1 (x^5 + 3)\sqrt{1 - x^2} dx.$

6. Prove that

$$\int_0^x \frac{\sin t}{t+1} dt > 0$$

for all $x > 0$.

7. Decide which of the following functions are integrable on $[0, 2]$, and calculate the integral when you can.

(i) $f(x) = \begin{cases} x, & 0 \leq x < 1 \\ x - 2, & 1 \leq x \leq 2. \end{cases}$

(ii) $f(x) = \begin{cases} x, & 0 \leq x \leq 1 \\ x - 2, & 1 < x \leq 2. \end{cases}$

(iii) $f(x) = x + [x]$.

(iv) $f(x) = \begin{cases} x + [x], & x \text{ rational} \\ 0, & x \text{ irrational.} \end{cases}$

(v) $f(x) = \begin{cases} 1, & x \text{ of the form } a + b\sqrt{2} \text{ for rational } a \text{ and } b \\ 0, & x \text{ not of this form.} \end{cases}$

(vi) $f(x) = \begin{cases} \frac{1}{\left[\frac{1}{x} \right]}, & 0 < x \leq 1 \\ 0, & x = 0 \text{ or } x > 1. \end{cases}$

(vii) f is the function shown in Figure 15.

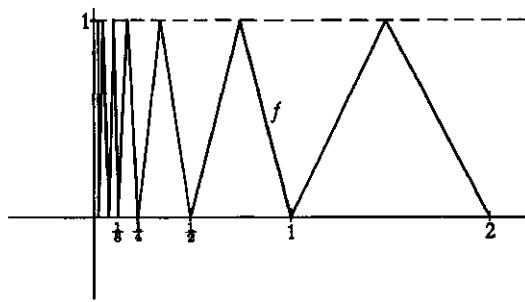


FIGURE 15

8. Find the areas of the regions bounded by

(i) the graphs of $f(x) = x^2$ and $g(x) = \frac{x^2}{2} + 2$.

(ii) the graphs of $f(x) = x^2$ and $g(x) = -x^2$ and the vertical lines through $(-1, 0)$ and $(1, 0)$.

(iii) the graphs of $f(x) = x^2$ and $g(x) = 1 - x^2$.

(iv) the graphs of $f(x) = x^2$ and $g(x) = 1 - x^2$ and $h(x) = 2$.

(v) the graphs of $f(x) = x^2$ and $g(x) = x^2 - 2x + 4$ and the vertical axis.

- (vi) the graph of $f(x) = \sqrt{x}$, the horizontal axis, and the vertical line through $(2, 0)$. (Don't try to find $\int_0^2 \sqrt{x} dx$; you should see a way of guessing the answer, using only integrals that you already know how to evaluate. The questions that this example should suggest are considered in Problem 21.)
- 9.** Find
- $$\int_a^b \left(\int_c^d f(x)g(y) dy \right) dx$$
- in terms of $\int_a^b f$ and $\int_c^d g$. (This problem is an exercise in notation, with a vengeance; it is crucial that you recognize a constant when it appears.)
- 10.** Prove, using the notation of Theorem 5, that
- $$m_i' + m_i'' = \inf\{f(x_1) + g(x_2) : t_{i-1} \leq x_1, x_2 \leq t_i\} \leq m_i.$$
- 11.** (a) Which functions have the property that every lower sum equals every upper sum?
 (b) Which functions have the property that some upper sum equals some (other) lower sum?
 (c) Which continuous functions have the property that all lower sums are equal?
 *(d) Which integrable functions have the property that all lower sums are equal? (Bear in mind that one such function is $f(x) = 0$ for x irrational, $f(x) = 1/q$ for $x = p/q$ in lowest terms.) Hint: You will need the notion of a dense set, introduced in Problem 8-6, as well as the results of Problem 30.
- 12.** If $a < b < c < d$ and f is integrable on $[a, d]$, prove that f is integrable on $[b, c]$. (Don't work hard.)
- 13.** (a) Prove that if f is integrable on $[a, b]$ and $f(x) \geq 0$ for all x in $[a, b]$, then $\int_a^b f \geq 0$.
 (b) Prove that if f and g are integrable on $[a, b]$ and $f(x) \geq g(x)$ for all x in $[a, b]$, then $\int_a^b f \geq \int_a^b g$. (By now it should be unnecessary to warn that if you work hard on part (b) you are wasting time.)
- 14.** Prove that

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

(The geometric interpretation should make this very plausible.) Hint: Every partition $P = \{t_0, \dots, t_n\}$ of $[a, b]$ gives rise to a partition $P' = \{t_0 + c, \dots, t_n + c\}$ of $[a + c, b + c]$, and conversely.

- *15. Prove that

$$\int_1^a \frac{1}{t} dt + \int_1^b \frac{1}{t} dt = \int_1^{ab} \frac{1}{t} dt.$$

Hint: This can be written $\int_1^a 1/t dt = \int_b^{ab} 1/t dt$. Every partition $P = \{t_0, \dots, t_n\}$ of $[1, a]$ gives rise to a partition $P' = \{bt_0, \dots, bt_n\}$ of $[b, ab]$, and conversely.

- *16. Prove that

$$\int_{ca}^{cb} f(t) dt = c \int_a^b f(ct) dt.$$

(Notice that Problem 15 is a special case.)

17. Given that the area enclosed by the unit circle, described by the equation $x^2 + y^2 = 1$, is π , use Problem 16 to show that the area enclosed by the ellipse described by the equation $x^2/a^2 + y^2/b^2 = 1$ is πab .

18. This problem outlines yet another way to compute $\int_a^b x^n dx$; it was used by Cavalieri, one of the mathematicians working just before the invention of calculus.

- (a) Let $c_n = \int_0^1 x^n dx$. Use Problem 16 to show that $\int_0^a x^n dx = c_n a^{n+1}$.
(b) Problem 14 shows that

$$\int_0^{2a} x^n dx = \int_{-a}^a (x+a)^n dx.$$

Use this formula to prove that

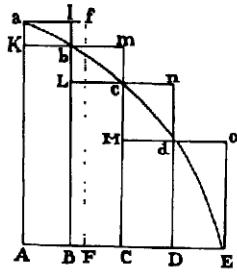
$$2^{n+1} c_n a^{n+1} = 2a^{n+1} \sum_{k \text{ even}} \binom{n}{k} c_k.$$

- (c) Now use Problem 2-3 to prove that $c_n = 1/(n+1)$.
19. Suppose that f is bounded on $[a, b]$ and that f is continuous at each point in $[a, b]$ with the exception of x_0 in (a, b) . Prove that f is integrable on $[a, b]$. Hint: Imitate one of the examples in the text.
20. Suppose that f is nondecreasing on $[a, b]$. Notice that f is automatically bounded on $[a, b]$, because $f(a) \leq f(x) \leq f(b)$ for x in $[a, b]$.

- (a) If $P = \{t_0, \dots, t_n\}$ is a partition of $[a, b]$, what is $L(f, P)$ and $U(f, P)$?
(b) Suppose that $t_i - t_{i-1} = \delta$ for each i . Prove that $U(f, P) - L(f, P) = \delta[f(b) - f(a)]$.
(c) Prove that f is integrable.
(d) Give an example of a nondecreasing function on $[0, 1]$ which is discontinuous at infinitely many points.

It might be of interest to compare this problem with the following extract from Newton's *Principia*.*

LEMMA II



If in any figure AacE, terminated by the right lines Aa, AE, and the curve acE, there be inscribed any number of parallelograms Ab, Bc, Cd, &c., comprehended under equal bases AB, BC, CD, &c., and the sides, Bb, Cc, Dd, &c., parallel to one side Aa of the figure; and the parallelograms aKbl, bLcm, cMdD, &c., are completed: then if the breadth of those parallelograms be supposed to be diminished, and their number to be augmented in infinitum, I say, that the ultimate ratios which the inscribed figure AKbLcMdD, the circumscribed figure AalbmndoE, and curvilinear figure AabcdE, will have to one another, are ratios of equality.

For the difference of the inscribed and circumscribed figures is the sum of the parallelograms Kl, Lm, Mn, Do, that is (from the equality of all their bases), the rectangle under one of their bases Kb and the sum of their altitudes Aa, that is, the rectangle ABla. But this rectangle, because its breadth AB is supposed diminished *in infinitum*, becomes less than any given space. And therefore (by Lem. 1) the figures inscribed and circumscribed become ultimately equal one to the other; and much more will the intermediate curvilinear figure be ultimately equal to either. Q.E.D.

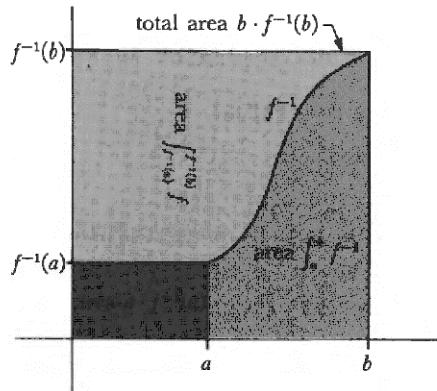


FIGURE 16

- *21. Suppose that f is increasing. Figure 16 suggests that

$$\int_a^b f^{-1} = bf^{-1}(b) - af^{-1}(a) - \int_{f^{-1}(a)}^{f^{-1}(b)} f.$$

- (a) If $P = \{t_0, \dots, t_n\}$ is a partition of $[a, b]$, let $P' = \{f^{-1}(t_0), \dots, f^{-1}(t_n)\}$. Prove that, as suggested in Figure 17,

$$L(f^{-1}, P) + U(f, P') = bf^{-1}(b) - af^{-1}(a).$$

- (b) Now prove the formula stated above.

(c) Find $\int_a^b \sqrt[3]{x} dx$ for $0 \leq a < b$.

22. Suppose that f is a continuous increasing function with $f(0) = 0$. Prove that for $a, b > 0$ we have *Young's inequality*,

$$ab \leq \int_0^a f(x) dx + \int_0^b f^{-1}(x) dx,$$

and that equality holds if and only if $b = f(a)$. Hint: Draw a picture like Figure 16!

*Newton's *Principia*, A Revision of Mott's Translation, by Florian Cajori. University of California Press, Berkeley, California, 1946.

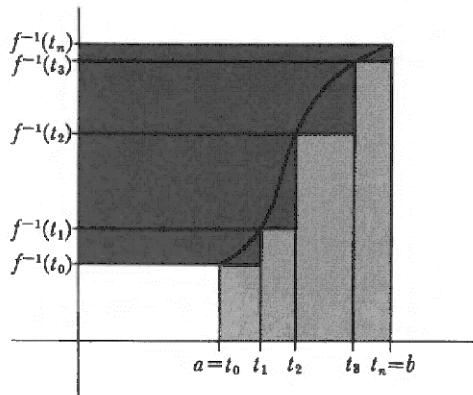


FIGURE 17

23. (a) Prove that if f is integrable on $[a, b]$ and $m \leq f(x) \leq M$ for all x in $[a, b]$, then

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

for some number μ with $m \leq \mu \leq M$.

- (b) Prove that if f is continuous on $[a, b]$, then

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

for some ξ in $[a, b]$; and show by an example that continuity is essential.

- (c) More generally, suppose that f is continuous on $[a, b]$ and that g is integrable and nonnegative on $[a, b]$. Prove that

$$\int_a^b f(x)g(x) dx = f(\xi) \int_a^b g(x) dx$$

for some ξ in $[a, b]$. This result is called the Mean Value Theorem for Integrals.

- (d) Deduce the same result if g is integrable and nonpositive on $[a, b]$.
(e) Show that one of these two hypotheses for g is essential.

24. In this problem we consider the graph of a function in polar coordinates (Chapter 4, Appendix 3). Figure 18 shows a sector of a circle, with central angle θ . When θ is measured in radians (Chapter 15), the area of this sector is $r^2 \cdot \frac{\theta}{2}$. Now consider the region A shown in Figure 19, where the curve is the graph in polar coordinates of the continuous function f . Show that

$$\text{area } A = \frac{1}{2} \int_{\theta_0}^{\theta_1} f(\theta)^2 d\theta.$$

- *25. Let f be a continuous function on $[a, b]$. If $P = \{t_0, \dots, t_n\}$ is a partition of $[a, b]$, define

$$\ell(f, P) = \sum_{i=1}^n \sqrt{(t_i - t_{i-1})^2 + [f(t_i) - f(t_{i-1})]^2}.$$

FIGURE 18

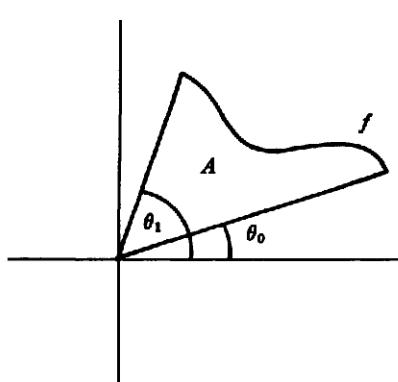


FIGURE 19

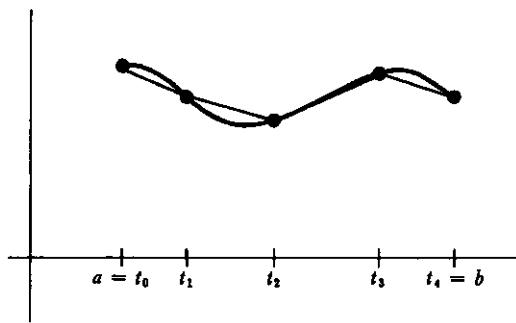


FIGURE 20

The number $\ell(f, P)$ represents the length of a polygonal curve inscribed in the graph of f (see Figure 20). We define the **length** of f on $[a, b]$ to be the least upper bound of all $\ell(f, P)$ for all partitions P (provided that the set of all such $\ell(f, P)$ is bounded above).

- If f is a linear function on $[a, b]$, prove that the length of f is the distance from $(a, f(a))$ to $(b, f(b))$.
- If f is not linear, prove that there is a partition $P = \{a, t, b\}$ of $[a, b]$ such that $\ell(f, P)$ is greater than the distance from $(a, f(a))$ to $(b, f(b))$. (You will need Problem 4-9.)
- Conclude that of all functions f on $[a, b]$ with $f(a) = c$ and $f(b) = d$, the length of the linear function is less than the length of any other. (Or, in conventional but hopelessly muddled terminology: "A straight line is the shortest distance between two points".)
- Suppose that f' is bounded on $[a, b]$. If P is any partition of $[a, b]$ show that

$$L(\sqrt{1 + (f')^2}, P) \leq \ell(f, P) \leq U(\sqrt{1 + (f')^2}, P).$$

Hint: Use the Mean Value Theorem.

- Why is $\sup\{L(\sqrt{1 + (f')^2}, P)\} \leq \sup\{\ell(f, P)\}$? (This is easy.)
- Now show that $\sup\{\ell(f, P)\} \leq \inf\{U(\sqrt{1 + (f')^2}, P)\}$, thereby proving that the length of f on $[a, b]$ is $\int_a^b \sqrt{1 + (f')^2}$, if $\sqrt{1 + (f')^2}$ is integrable on $[a, b]$. Hint: It suffices to show that if P' and P'' are any two partitions, then $\ell(f, P') \leq U(\sqrt{1 + (f')^2}, P'')$. If P contains the points of both P' and P'' , how does $\ell(f, P')$ compare to $\ell(f, P)$?
- Let $\mathcal{L}(x)$ be the length of the graph of f on $[a, x]$, and let $d(x)$ be the length of the straight line segment from $(a, f(a))$ to $(x, f(x))$. Show that

$$\lim_{x \rightarrow a} \frac{\mathcal{L}(x)}{d(x)} = 1.$$

Hint: It will help to use a couple of Mean Value Theorems.

- A function s defined on $[a, b]$ is called a **step function** if there is a partition $P = \{t_0, \dots, t_n\}$ of $[a, b]$ such that s is a constant on each (t_{i-1}, t_i) (the values of s at t_i may be arbitrary).
 - Prove that if f is integrable on $[a, b]$, then for any $\varepsilon > 0$ there is a step function $s_1 \leq f$ with $\int_a^b f - \int_a^b s_1 < \varepsilon$, and also a step function $s_2 \geq f$ with $\int_a^b s_2 - \int_a^b f < \varepsilon$.
 - Suppose that for all $\varepsilon > 0$ there are step functions $s_1 \leq f$ and $s_2 \geq f$ such that $\int_a^b s_2 - \int_a^b s_1 < \varepsilon$. Prove that f is integrable.

- (c) Find a function f which is not a step function, but which satisfies $\int_a^b f = L(f, P)$ for some partition P of $[a, b]$.
- *27. Prove that if f is integrable on $[a, b]$, then for any $\varepsilon > 0$ there are continuous functions $g \leq f \leq h$ with $\int_a^b h - \int_a^b g < \varepsilon$. Hint: First get step functions with this property, and then continuous ones. A picture will help immensely.
28. (a) Show that if s_1 and s_2 are step functions on $[a, b]$, then $s_1 + s_2$ is also.
 (b) Prove, without using Theorem 5, that $\int_a^b (s_1 + s_2) = \int_a^b s_1 + \int_a^b s_2$.
 (c) Use part (b) (and Problem 26) to give an alternative proof of Theorem 5.
29. Suppose that f is integrable on $[a, b]$. Prove that there is a number x in $[a, b]$ such that $\int_a^x f = \int_x^b f$. Show by example that it is *not* always possible to choose x to be in (a, b) .
- *30. The purpose of this problem is to show that if f is integrable on $[a, b]$, then f must be continuous at many points in $[a, b]$.
- (a) Let $P = \{t_0, \dots, t_n\}$ be a partition of $[a, b]$ with $U(f, P) - L(f, P) < b - a$. Prove that for some i we have $M_i - m_i < 1$.
 (b) Prove that there are numbers a_1 and b_1 with $a < a_1 < b_1 < b$ and $\sup\{f(x) : a_1 \leq x \leq b_1\} - \inf\{f(x) : a_1 \leq x \leq b_1\} < 1$. (You can choose $[a_1, b_1] = [t_{i-1}, t_i]$ from part (a) unless $i = 1$ or n ; and in these two cases a very simple device solves the problem.)
 (c) Prove that there are numbers a_2 and b_2 with $a_1 < a_2 < b_2 < b_1$ and $\sup\{f(x) : a_2 \leq x \leq b_2\} - \inf\{f(x) : a_2 \leq x \leq b_2\} < \frac{1}{2}$.
 (d) Continue in this way to find a sequence of intervals $I_n = [a_n, b_n]$ such that $\sup\{f(x) : x \text{ in } I_n\} - \inf\{f(x) : x \text{ in } I_n\} < 1/n$. Apply the Nested Intervals Theorem (Problem 8-14) to find a point x at which f is continuous.
 (e) Prove that f is continuous at infinitely many points in $[a, b]$.
- *31. Recall, from Problem 13, that $\int_a^b f \geq 0$ if $f(x) \geq 0$ for all x in $[a, b]$.
- (a) Give an example where $f(x) \geq 0$ for all x , and $f(x) > 0$ for some x in $[a, b]$, and $\int_a^b f = 0$.
 (b) Suppose $f(x) \geq 0$ for all x in $[a, b]$ and f is continuous at x_0 in $[a, b]$ and $f(x_0) > 0$. Prove that $\int_a^b f > 0$. Hint: It suffices to find one lower sum $L(f, P)$ which is positive.
 (c) Suppose f is integrable on $[a, b]$ and $f(x) > 0$ for all x in $[a, b]$. Prove that $\int_a^b f > 0$. Hint: You will need Problem 30; indeed that was one reason for including Problem 30.

- *32. (a) Suppose that f is continuous on $[a, b]$ and $\int_a^b fg = 0$ for all continuous functions g on $[a, b]$. Prove that $f = 0$. (This is easy; there is an obvious g to choose.)
- (b) Suppose f is continuous on $[a, b]$ and that $\int_a^b fg = 0$ for those continuous functions g on $[a, b]$ which satisfy the extra conditions $g(a) = g(b) = 0$. Prove that $f = 0$. (This innocent looking fact is an important lemma in the calculus of variations; see the Suggested Reading for references.) Hint: Derive a contradiction from the assumption $f(x_0) > 0$ or $f(x_0) < 0$; the g you pick will depend on the behavior of f near x_0 .
33. Let $f(x) = x$ for x rational and $f(x) = 0$ for x irrational.
- (a) Compute $L(f, P)$ for all partitions P of $[0, 1]$.
 (b) Find $\inf\{U(f, P) : P \text{ a partition of } [0, 1]\}$.
- *34. Let $f(x) = 0$ for irrational x , and $1/q$ if $x = p/q$ in lowest terms. Show that f is integrable on $[0, 1]$ and that $\int_0^1 f = 0$. (Every lower sum is clearly 0; you must figure out how to make upper sums small.)
- *35. Find two functions f and g which are integrable, but whose composition $g \circ f$ is not. Hint: Problem 34 is relevant.
- *36. Let f be a bounded function on $[a, b]$ and let P be a partition of $[a, b]$. Let M_i and m_i have their usual meanings, and let M'_i and m'_i have the corresponding meanings for the function $|f|$.
- (a) Prove that $M'_i - m'_i \leq M_i - m_i$.
 (b) Prove that if f is integrable on $[a, b]$, then so is $|f|$.
 (c) Prove that if f and g are integrable on $[a, b]$, then so are $\max(f, g)$ and $\min(f, g)$.
 (d) Prove that f is integrable on $[a, b]$ if and only if its “positive part” $\max(f, 0)$ and its “negative part” $\min(f, 0)$ are integrable on $[a, b]$.
37. Prove that if f is integrable on $[a, b]$, then
- $$\left| \int_a^b f(t) dt \right| \leq \int_a^b |f(t)| dt.$$
- Hint: This follows easily from a certain string of inequalities; Problem 1-14 is relevant.
- *38. Suppose f and g are integrable on $[a, b]$ and $f(x), g(x) \geq 0$ for all x in $[a, b]$. Let P be a partition of $[a, b]$. Let M'_i and m'_i denote the appropriate sup's and inf's for f , define M''_i and m''_i similarly for g , and define M_i and m_i similarly for fg .
- (a) Prove that $M_i \leq M'_i M''_i$ and $m_i \geq m'_i m''_i$.

(b) Show that

$$U(fg, P) - L(fg, P) \leq \sum_{i=1}^n [M_i' M_i'' - m_i' m_i''](t_i - t_{i-1}).$$

(c) Using the fact that f and g are bounded, so that $|f(x)|, |g(x)| \leq M$ for x in $[a, b]$, show that

$$U(fg, P) - L(fg, P)$$

$$\leq M \left\{ \sum_{i=1}^n [M_i' - m_i'](t_i - t_{i-1}) + \sum_{i=1}^n [M_i'' - m_i''](t_i - t_{i-1}) \right\}.$$

(d) Prove that fg is integrable.

(e) Now eliminate the restriction that $f(x), g(x) \geq 0$ for x in $[a, b]$.

39. Suppose that f and g are integrable on $[a, b]$. The *Cauchy-Schwarz inequality* states that

$$\left(\int_a^b fg \right)^2 \leq \left(\int_a^b f^2 \right) \left(\int_a^b g^2 \right).$$

(a) Show that the Schwarz inequality is a special case of the Cauchy-Schwarz inequality.

(b) Give three proofs of the Cauchy-Schwarz inequality by imitating the proofs of the Schwarz inequality in Problem 2-21. (The last one will take some imagination.)

(c) If equality holds, is it necessarily true that $f = \lambda g$ for some λ ? What if f and g are continuous?

(d) Prove that $\left(\int_0^1 f \right)^2 \leq \left(\int_0^1 f^2 \right)$. Is this result true if 0 and 1 are replaced by a and b ?

- *40. Suppose that f is continuous and $\lim_{x \rightarrow \infty} f(x) = a$. Prove that

$$\lim_{x \rightarrow \infty} \frac{1}{x} \int_0^x f(t) dt = a.$$

Hint: The condition $\lim_{x \rightarrow \infty} f(x) = a$ implies that $f(t)$ is close to a for $t \geq$ some N . This means that $\int_N^{N+M} f(t) dt$ is close to Ma . If M is large in comparison to N , then $Ma/(N+M)$ is close to a .

APPENDIX. RIEMANN SUMS

Suppose that $P = \{t_0, \dots, t_n\}$ is a partition of $[a, b]$, and that for each i we choose some point x_i in $[t_{i-1}, t_i]$. Then we clearly have

$$L(f, P) \leq \sum_{i=1}^n f(x_i)(t_i - t_{i-1}) \leq U(f, P).$$

Any sum $\sum_{i=1}^n f(x_i)(t_i - t_{i-1})$ is called a *Riemann sum* of f for P . Figure 1 shows the geometric interpretation of a Riemann sum; it is the total area of n rectangles that lie partly below the graph of f and partly above it. Because of the arbitrary way in which the heights of the rectangles have been picked, we can't say for sure whether a particular Riemann sum is less than or greater than the integral $\int_a^b f(x) dx$. But it does seem that the overlaps shouldn't matter too much; if the bases of all the rectangles are narrow enough, then the Riemann sum ought to be close to the integral. The following theorem states this precisely.

THEOREM 1

Suppose that f is integrable on $[a, b]$. Then for every $\varepsilon > 0$ there is some $\delta > 0$ such that, if $P = \{t_0, \dots, t_n\}$ is any partition of $[a, b]$ with all lengths $t_i - t_{i-1} < \delta$, then

$$\left| \sum_{i=1}^n f(x_i)(t_i - t_{i-1}) - \int_a^b f(x) dx \right| < \varepsilon,$$

for any Riemann sum formed by choosing x_i in $[t_{i-1}, t_i]$.

PROOF

First we will prove the theorem when f is continuous. As in the proof that a continuous function is integrable (Theorem 13-3), we will use Theorem 1 from the Appendix to Chapter 8, so you might want to skip it. But if you've already read the proof of Theorem 13-3, this part of the proof will be a snap—in fact, it's practically the same.

Given $\varepsilon > 0$, choose $\delta > 0$ so that for all x and y in $[a, b]$

$$\text{if } |x - y| < \delta, \text{ then } |f(x) - f(y)| < \frac{\varepsilon}{2(b-a)}.$$

Now consider any partition $P = \{t_0, \dots, t_n\}$ with each $t_i - t_{i-1} < \delta$, and any x_i in $[t_{i-1}, t_i]$. Then, as we saw in the proof of Theorem 13-3, we have

$$(1) \quad U(f, P) - L(f, P) < \varepsilon.$$

But we also have

$$(2) \quad L(f, P) \leq \sum_{i=1}^n f(x_i)(t_i - t_{i-1}) \leq U(f, P)$$

and

$$(3) \quad L(f, P) \leq \int_a^b f(x) dx \leq U(f, P).$$

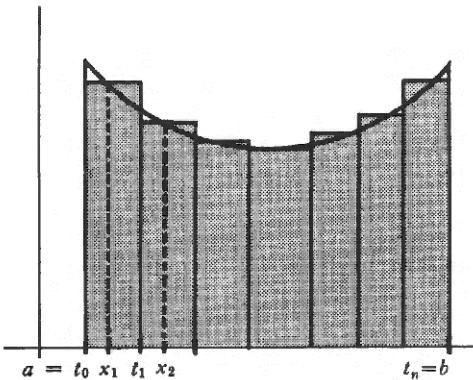


FIGURE 1

The desired inequality, for our continuous function f , follows immediately from (1), (2) and (3).

The argument in the general case is simple (though perhaps a bit messy), using Problem 13-27, which says that there are continuous functions $g \leq f \leq h$ satisfying

$$(4) \quad \int_a^b g \leq \int_a^b f \leq \int_a^b h,$$

with

$$\int_a^b h - \int_a^b g < \varepsilon.$$

We have

$$\sum_{i=1}^n g(x_i)(t_i - t_{i-1}) \leq \sum_{i=1}^n f(x_i)(t_i - t_{i-1}) \leq \sum_{i=1}^n h(x_i)(t_i - t_{i-1}),$$

and since the theorem holds for continuous functions, we know that for $t_i - t_{i-1} < \delta$, the left- and right-hand sides of this inequality are close to the left- and right-hand sides of (4). This implies that the two middle terms,

$$\int_a^b f \quad \text{and} \quad \sum_{i=1}^n f(x_i)(t_i - t_{i-1}),$$

must be close to $\int_a^b h - \int_a^b g$, which is small. Detailed inequalities are left to the skeptical reader. ■

The moral of this tale is that anything which looks like a good approximation to an integral really is, provided that all the lengths $t_i - t_{i-1}$ of the intervals in the partition are small enough. Some of the following problems should bring home this message with even greater force.

PROBLEMS

- Suppose that f and g are continuous functions on $[a, b]$. For a partition $P = \{t_0, \dots, t_n\}$ of $[a, b]$ choose a set of points x_i in $[t_{i-1}, t_i]$ and another set of points u_i in $[t_{i-1}, t_i]$. Consider the sum

$$\sum_{i=1}^n f(x_i)g(u_i)(t_i - t_{i-1}).$$

Notice that this is *not* a Riemann sum of fg for P . Nevertheless, show that all such sums will be within ε of $\int_a^b fg$ provided that the partition P has all lengths $t_i - t_{i-1}$ small enough. Hint: Estimate the difference between such a sum and a Riemann sum; you will need to use uniform continuity.

2. This problem is similar to, but somewhat harder than, the previous one. Suppose that f and g are continuous nonnegative functions on $[a, b]$. For a partition P , consider sums

$$\sum_{i=1}^n \sqrt{f(x_i) + g(u_i)} (t_i - t_{i-1}).$$

Show that these sums will be within ε of $\int_a^b \sqrt{f+g}$ if all $t_i - t_{i-1}$ are small enough. Hint: Use the fact that the square-root function is uniformly continuous on a closed interval $[0, M]$.

3. Finally, we're ready to tackle something big! (Compare Problem 13-25.) Consider a curve c given parametrically by two functions u and v on $[a, b]$. For a partition $P = \{t_0, \dots, t_n\}$ of $[a, b]$ we define

$$\ell(c, P) = \sum_{i=1}^n \sqrt{[u(t_i) - u(t_{i-1})]^2 + [v(t_i) - v(t_{i-1})]^2};$$

this represents the length of an inscribed polygonal curve (Figure 2). We define the length of c to be the least upper bound of all $\ell(f, P)$, if it exists. Prove that if u' and v' are continuous on $[a, b]$, then the length of c is

$$\int_a^b \sqrt{u'^2 + v'^2}.$$

4. Let f' be continuous on the interval $[\theta_0, \theta_1]$. Show that the graph of f in polar coordinates on this interval has the length

$$\int_{\theta_0}^{\theta_1} \sqrt{f^2 + f'^2}.$$

5. Using Theorem 1, show that the Cauchy-Schwarz inequality (Problem 13-39) is a consequence of the Schwarz inequality.

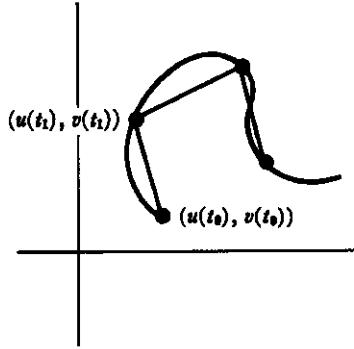


FIGURE 2

CHAPTER 14

THE FUNDAMENTAL THEOREM OF CALCULUS

From the hints given in the previous chapter you may have already guessed the first theorem of this chapter. We know that if f is integrable, then $F(x) = \int_a^x f$ is continuous; it is only fitting that we ask what happens when the original function f is continuous. It turns out that F is differentiable (and its derivative is especially simple).

THEOREM 1 (THE FIRST FUNDAMENTAL THEOREM OF CALCULUS)

Let f be integrable on $[a, b]$, and define F on $[a, b]$ by

$$F(x) = \int_a^x f.$$

If f is continuous at c in $[a, b]$, then F is differentiable at c , and

$$F'(c) = f(c).$$

(If $c = a$ or b , then $F'(c)$ is understood to mean the right- or left-hand derivative of F .)

PROOF We will assume that c is in (a, b) ; the easy modifications for $c = a$ or b may be supplied by the reader. By definition,

$$F'(c) = \lim_{h \rightarrow 0} \frac{F(c + h) - F(c)}{h}.$$

Suppose first that $h > 0$. Then

$$F(c + h) - F(c) = \int_c^{c+h} f.$$

Define m_h and M_h as follows (Figure 1):

$$\begin{aligned} m_h &= \inf\{f(x) : c \leq x \leq c + h\}, \\ M_h &= \sup\{f(x) : c \leq x \leq c + h\}. \end{aligned}$$

It follows from Theorem 13-7 that

$$m_h \cdot h \leq \int_c^{c+h} f \leq M_h \cdot h.$$

Therefore

$$m_h \leq \frac{F(c + h) - F(c)}{h} \leq M_h.$$

If $h < 0$, only a few details of the argument have to be changed. Let

$$\begin{aligned} m_h &= \inf\{f(x) : c + h \leq x \leq c\}, \\ M_h &= \sup\{f(x) : c + h \leq x \leq c\}. \end{aligned}$$

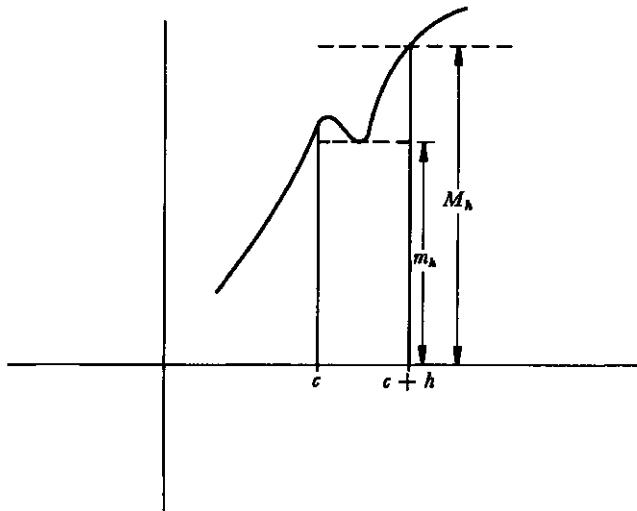


FIGURE 1

Then

$$m_h \cdot (-h) \leq \int_{c+h}^c f \leq M_h \cdot (-h).$$

Since

$$F(c+h) - F(c) = \int_c^{c+h} f = - \int_{c+h}^c f$$

this yields

$$m_h \cdot h \geq F(c+h) - F(c) \geq M_h \cdot h.$$

Since $h < 0$, dividing by h reverses the inequality again, yielding the same result as before:

$$m_h \leq \frac{F(c+h) - F(c)}{h} \leq M_h.$$

This inequality is true for any integrable function, continuous or not. Since f is continuous at c , however,

$$\lim_{h \rightarrow 0} m_h = \lim_{h \rightarrow 0} M_h = f(c),$$

and this proves that

$$F'(c) = \lim_{h \rightarrow 0} \frac{F(c+h) - F(c)}{h} = f(c). \blacksquare$$

Although Theorem 1 deals only with the function obtained by varying the upper limit of integration, a simple trick shows what happens when the lower limit is varied. If G is defined by

$$G(x) = \int_x^b f,$$

then

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

Consequently, if f is continuous at c , then

$$G'(c) = -f(c).$$

The minus sign appearing here is very fortunate, and allows us to extend Theorem 1 to the situation where the function

$$F(x) = \int_a^x f$$

is defined even for $x < a$. In this case we can write

$$F(x) = - \int_x^a f,$$

so if $c < a$ we have

$$F'(c) = -(-f(c)) = f(c),$$

exactly as before.

Notice that in either case, differentiability of F at c is ensured by continuity of f at c alone. Nevertheless, Theorem 1 is most interesting when f is continuous at all points in $[a, b]$. In this case F is differentiable at all points in $[a, b]$ and

$$F' = f.$$

In general, it is extremely difficult to decide whether a given function f is the derivative of some other function; for this reason Theorem 11-7 and Problems 11-54 and 11-55 are particularly interesting, since they reveal certain properties which f must have. If f is continuous, however, there is no problem at all—according to Theorem 1, f is the derivative of some function, namely the function

$$F(x) = \int_a^x f.$$

Theorem 1 has a simple corollary which frequently reduces computations of integrals to a triviality.

COROLLARY If f is continuous on $[a, b]$ and $f = g'$ for some function g , then

$$\int_a^b f = g(b) - g(a).$$

PROOF Let

$$F(x) = \int_a^x f.$$

Then $F' = f = g'$ on $[a, b]$. Consequently, there is a number c such that

$$F = g + c.$$

The number c can be evaluated easily: note that

$$0 = F(a) = g(a) + c,$$

so $c = -g(a)$; thus

$$F(x) = g(x) - g(a).$$

This is true, in particular, for $x = b$. Thus

$$\int_a^b f = F(b) = g(b) - g(a). \blacksquare$$

The proof of this corollary tends, at first sight, to make the corollary seem useless: after all, what good is it to know that

$$\int_a^b f = g(b) - g(a)$$

if g is, for example, $g(x) = \int_a^x f$? The point, of course, is that one might happen to know a quite different function g with this property. For example, if

$$g(x) = \frac{x^3}{3} \quad \text{and} \quad f(x) = x^2,$$

then $g'(x) = f(x)$ so we obtain, without ever computing lower and upper sums:

$$\int_a^b x^2 dx = \frac{b^3}{3} - \frac{a^3}{3}.$$

One can treat other powers similarly; if n is a natural number and $g(x) = x^{n+1}/(n+1)$, then $g'(x) = x^n$, so

$$\int_a^b x^n dx = \frac{b^{n+1}}{n+1} - \frac{a^{n+1}}{n+1}.$$

For any natural number n , the function $f(x) = x^{-n}$ is not bounded on any interval containing 0, but if a and b are both positive or both negative, then

$$\int_a^b x^{-n} dx = \frac{b^{-n+1}}{-n+1} - \frac{a^{-n+1}}{-n+1}.$$

Naturally this formula is only true for $n \neq -1$. *We do not know a simple expression for*

$$\int_a^b \frac{1}{x} dx.$$

The problem of computing this integral is discussed later, but it provides a good opportunity to warn against a serious error. The conclusion of Corollary 1 is often confused with the definition of integrals—many students think that $\int_a^b f$ is defined as: “ $g(b) - g(a)$, where g is a function whose derivative is f .” This “definition” is not only wrong—it is useless. One reason is that a function f may be integrable without being the derivative of another function. For example, if $f(x) = 0$ for $x \neq 1$ and $f(1) = 1$, then f is integrable, but f cannot be a derivative (why not?). There is also another reason that is much more important: If f is continuous,

then we know that $f = g'$ for some function g ; but we know this *only because of Theorem 1*. The function $f(x) = 1/x$ provides an excellent illustration: if $x > 0$, then $f(x) = g'(x)$, where

$$g(x) = \int_1^x \frac{1}{t} dt,$$

and we know of no simpler function g with this property.

The corollary to Theorem 1 is so useful that it is frequently called the Second Fundamental Theorem of Calculus. In this book, that name is reserved for a somewhat stronger result (which in practice, however, is not much more useful). As we have just mentioned, a function f might be of the form g' even if f is not continuous. If f is integrable, then it is still true that

$$\int_a^b f = g(b) - g(a).$$

The proof, however, must be entirely different—we cannot use Theorem 1, so we must return to the definition of integrals.

THEOREM 2 (THE SECOND FUNDAMENTAL THEOREM OF CALCULUS)

If f is integrable on $[a, b]$ and $f = g'$ for some function g , then

$$\int_a^b f = g(b) - g(a).$$

PROOF Let $P = \{t_0, \dots, t_n\}$ be any partition of $[a, b]$. By the Mean Value Theorem there is a point x_i in $[t_{i-1}, t_i]$ such that

$$\begin{aligned} g(t_i) - g(t_{i-1}) &= g'(x_i)(t_i - t_{i-1}) \\ &= f(x_i)(t_i - t_{i-1}). \end{aligned}$$

If

$$\begin{aligned} m_i &= \inf\{f(x) : t_{i-1} \leq x \leq t_i\}, \\ M_i &= \sup\{f(x) : t_{i-1} \leq x \leq t_i\}, \end{aligned}$$

then clearly

$$m_i(t_i - t_{i-1}) \leq f(x_i)(t_i - t_{i-1}) \leq M_i(t_i - t_{i-1}),$$

that is,

$$m_i(t_i - t_{i-1}) \leq g(t_i) - g(t_{i-1}) \leq M_i(t_i - t_{i-1}).$$

Adding these equations for $i = 1, \dots, n$ we obtain

$$\sum_{i=1}^n m_i(t_i - t_{i-1}) \leq g(b) - g(a) \leq \sum_{i=1}^n M_i(t_i - t_{i-1})$$

so that

$$L(f, P) \leq g(b) - g(a) \leq U(f, P)$$

for every partition P . But this means that

$$g(b) - g(a) = \int_a^b f. \blacksquare$$

We have already used the corollary to Theorem 1 (or, equivalently, Theorem 2) to find the integrals of a few elementary functions:

$$\int_a^b x^n dx = \frac{b^{n+1}}{n+1} - \frac{a^{n+1}}{n+1}, \quad n \neq -1. \quad (a \text{ and } b \text{ both positive or both negative if } n > 0).$$

As we pointed out in Chapter 13, this integral does not always represent the area bounded by the graph of the function, the horizontal axis, and the vertical lines through $(a, 0)$ and $(b, 0)$. For example, if $a < 0 < b$, then

$$\int_a^b x^3 dx$$

does not represent the area of the region shown in Figure 2, which is given instead by

$$\begin{aligned} -\left(\int_a^0 x^3 dx\right) + \int_0^b x^3 dx &= -\left(\frac{0^4}{4} - \frac{a^4}{4}\right) + \left(\frac{b^4}{4} - \frac{0^4}{4}\right) \\ &= \frac{a^4}{4} + \frac{b^4}{4}. \end{aligned}$$

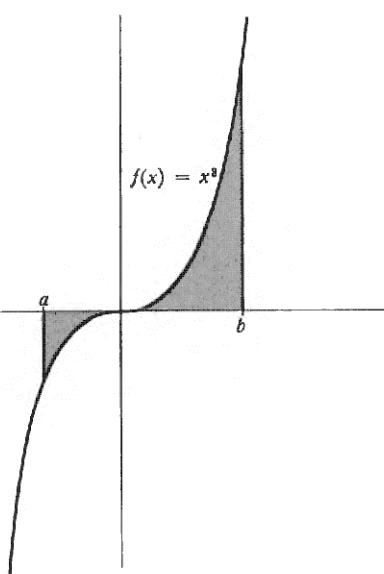


FIGURE 2

Similar care must be exercised in finding the areas of regions which are bounded by the graphs of more than one function—a problem which may frequently involve considerable ingenuity in any case. Suppose, to take a simple example first, that we wish to find the area of the region, shown in Figure 3, between the graphs of the functions

$$f(x) = x^2 \quad \text{and} \quad g(x) = x^3$$

on the interval $[0, 1]$. If $0 \leq x \leq 1$, then $0 \leq x^3 \leq x^2$, so that the graph of g lies below that of f . The area of the region of interest to us is therefore

$$\text{area } R(f, 0, 1) - \text{area } R(g, 0, 1),$$

which is

$$\int_0^1 x^2 dx - \int_0^1 x^3 dx = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

This area could have been expressed as

$$\int_a^b (f - g).$$

If $g(x) \leq f(x)$ for all x in $[a, b]$, then this integral always gives the area bounded by f and g , even if f and g are sometimes negative. The easiest way to see this is shown in Figure 4. If c is a number such that $f + c$ and $g + c$ are nonnegative on $[a, b]$, then the region R_1 , bounded by f and g , has the same area as the region R_2 ,

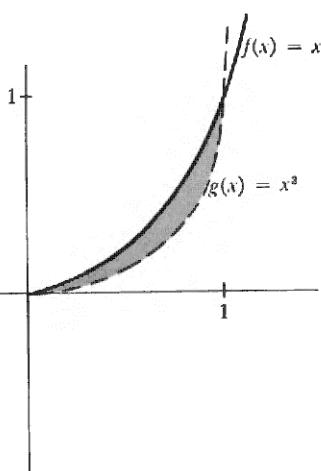


FIGURE 3

bounded by $f + c$ and $g + c$. Consequently,

$$\begin{aligned}\text{area } R_1 = \text{area } R_2 &= \int_a^b (f + c) - \int_a^b (g + c) \\ &= \int_a^b [(f + c) - (g + c)] \\ &= \int_a^b (f - g).\end{aligned}$$

This observation is useful in the following problem: Find the area of the region bounded by the graphs of

$$f(x) = x^3 - x \quad \text{and} \quad g(x) = x^2.$$

The first necessity is to determine this region more precisely. The graphs of f and g intersect when

$$\begin{aligned}x^3 - x &= x^2, \\ \text{or } x^3 - x^2 - x &= 0, \\ \text{or } x(x^2 - x - 1) &= 0, \\ \text{or } x &= 0, \frac{1+\sqrt{5}}{2}, \frac{1-\sqrt{5}}{2}.\end{aligned}$$

On the interval $([1 - \sqrt{5}]/2, 0)$ we have $x^3 - x \geq x^2$ and on the interval $(0, [1 + \sqrt{5}]/2)$ we have $x^2 \geq x^3 - x$. These assertions are apparent from the graphs (Figure 5), but they can also be checked easily, as follows. Since $f(x) = g(x)$ only if $x = 0, [1 + \sqrt{5}]/2$, or $[1 - \sqrt{5}]/2$, the function $f - g$ does not change sign on the intervals $([1 - \sqrt{5}]/2, 0)$ and $(0, [1 + \sqrt{5}]/2)$; it is therefore only necessary to observe, for example, that

$$\begin{aligned}(-\tfrac{1}{2})^3 - (-\tfrac{1}{2}) - (-\tfrac{1}{2})^2 &= \tfrac{1}{8} > 0, \\ 1^3 - 1 - 1^2 &= -1 < 0,\end{aligned}$$

to conclude that

$$\begin{aligned}f - g &\geq 0 \quad \text{on } ([1 - \sqrt{5}]/2, 0), \\ f - g &\leq 0 \quad \text{on } (0, [1 + \sqrt{5}]/2).\end{aligned}$$

The area of the region in question is thus

$$\int_{\frac{1-\sqrt{5}}{2}}^0 (x^3 - x - x^2) dx + \int_0^{\frac{1+\sqrt{5}}{2}} [x^2 - (x^3 - x)] dx.$$

As this example reveals, one of the major problems involved in finding the areas of a region may be the exact determination of the region. There are, however, more substantial problems of a logical nature—we have thus far defined the areas of some very special regions only, which do not even include some of the regions whose areas have just been computed! We have simply assumed that area made

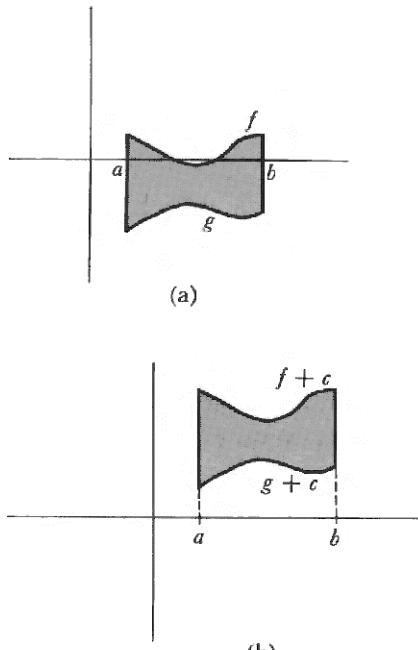


FIGURE 4

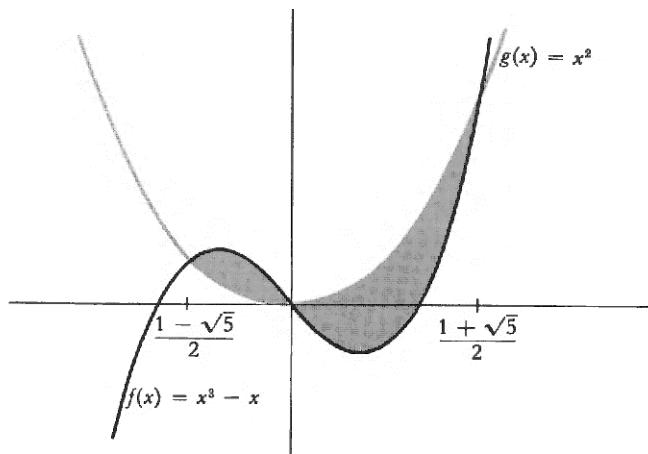


FIGURE 5

sense for these regions, and that certain reasonable properties of “area” do hold. These remarks are not meant to suggest that you should regard exercising ingenuity to compute areas as beneath you, but are meant to indicate that a better approach to the definition of area is available, although its proper place is somewhere in advanced calculus. The desire to define area was the motivation, both in this book and historically, for the definition of the integral, but the integral does not really provide the best method of *defining* areas, although it is frequently the proper tool for *computing* them.

It may be discouraging to learn that integrals are not suitable for the very purpose for which they were invented, but we will soon see how essential they are for other purposes. The most important use of integrals has already been emphasized: if f is continuous, the integral provides a function y such that

$$y'(x) = f(x).$$

This equation is the simplest example of a “differential equation” (an equation for a function y which involves derivatives of y). The Fundamental Theorem of Calculus says that this differential equation has a solution, if f is continuous. In succeeding chapters, and in various problems, we will solve more complicated equations, but the solution almost always depends somehow on the integral; in order to solve a differential equation it is necessary to construct a new function, and the integral is one of the best ways of doing this.

Since the differentiable functions provided by the Fundamental Theorem of Calculus will play such a prominent role in later work, it is very important to realize that these functions may be combined, like less esoteric functions, to yield still more functions, whose derivatives can be found by the Chain Rule.

Suppose, for example, that

$$f(x) = \int_a^{x^3} \frac{1}{1 + \sin^2 t} dt.$$

Although the notation tends to disguise the fact somewhat, f is the composition of the functions

$$C(x) = x^3 \quad \text{and} \quad F(x) = \int_a^x \frac{1}{1 + \sin^2 t} dt.$$

In fact, $f(x) = F(C(x))$; in other words, $f = F \circ C$. Therefore, by the Chain Rule,

$$\begin{aligned} f'(x) &= F'(C(x)) \cdot C'(x) \\ &= F'(x^3) \cdot 3x^2 \\ &= \frac{1}{1 + \sin^2 x^3} \cdot 3x^2. \end{aligned}$$

If f is defined, instead, as

$$f(x) = \int_{x^3}^a \frac{1}{1 + \sin^2 t} dt,$$

then

$$f'(x) = -\frac{1}{1 + \sin^2 x^3} \cdot 3x^2.$$

If f is defined as the *reverse* composition,

$$f(x) = \left(\int_a^x \frac{1}{1 + \sin^2 t} dt \right)^3,$$

then

$$\begin{aligned} f'(x) &= C'(F(x)) \cdot F'(x) \\ &= 3 \left(\int_a^x \frac{1}{1 + \sin^2 t} dt \right)^2 \cdot \frac{1}{1 + \sin^2 x}. \end{aligned}$$

Similarly, if

$$f(x) = \int_a^{\sin x} \frac{1}{1 + \sin^2 t} dt,$$

$$g(x) = \int_{\sin x}^a \frac{1}{1 + \sin^2 t} dt,$$

$$h(x) = \sin \left(\int_a^x \frac{1}{1 + \sin^2 t} dt \right),$$

then

$$f'(x) = \frac{1}{1 + \sin^2(\sin x)} \cdot \cos x,$$

$$g'(x) = \frac{-1}{1 + \sin^2(\sin x)} \cdot \cos x,$$

$$h'(x) = \cos \left(\int_a^x \frac{1}{1 + \sin^2 t} dt \right) \cdot \frac{1}{1 + \sin^2 x}.$$

The formidable appearing function

$$f(x) = \int_a^{\left(\int_a^x \frac{1}{1+\sin^2 t} dt\right)} \frac{1}{1+\sin^2 t} dt$$

is also a composition; in fact, $f = F \circ F$. Therefore

$$\begin{aligned} f'(x) &= F'(F(x)) \cdot F'(x) \\ &= \frac{1}{1+\sin^2 \left(\int_a^x \frac{1}{1+\sin^2 t} dt \right)} \cdot \frac{1}{1+\sin^2 x}. \end{aligned}$$

As these examples reveal, the expression occurring above (or below) the integral sign indicates the function which will appear on the *right* when f is written as a composition. As a final example, consider the triple compositions

$$f(x) = \int_a^{\left(\int_a^{x^3} \frac{1}{1+\sin^2 t} dt\right)} \frac{1}{1+\sin^2 t} dt, \quad g(x) = \int_a^{\left[\int_a^{\left(\int_a^x \frac{1}{1+\sin^2 t} dt\right)} \frac{1}{1+\sin^2 t} dt \right]} \frac{1}{1+\sin^2 t} dt,$$

which can be written

$$f = F \circ F \circ C \quad \text{and} \quad g = F \circ F \circ F.$$

Omitting the intermediate steps (which you may supply, if you still feel insecure), we obtain

$$\begin{aligned} f'(x) &= \frac{1}{1+\sin^2 \left(\int_a^{x^3} \frac{1}{1+\sin^2 t} dt \right)} \cdot \frac{1}{1+\sin^2 x^3} \cdot 3x^2, \\ g'(x) &= \frac{1}{1+\sin^2 \left[\int_a^{\left(\int_a^{x^3} \frac{1}{1+\sin^2 t} dt\right)} \frac{1}{1+\sin^2 t} dt \right]} \cdot \frac{1}{1+\sin^2 \left(\int_a^x \frac{1}{1+\sin^2 t} dt \right)} \\ &\quad \cdot \frac{1}{1+\sin^2 x}. \end{aligned}$$

Like the simpler differentiations of Chapter 10, these manipulations should become much easier after the practice provided by some of the problems, and, like the problems of Chapter 10, these differentiations are simply a test of your understanding of the Chain Rule, in the somewhat unfamiliar context provided by the Fundamental Theorem of Calculus.

The powerful uses to which the integral will be put in the following chapters all depend on the Fundamental Theorem of Calculus, yet the proof of that theorem was quite easy—it seems that all the real work went into the definition of the integral. Actually, this is not quite true. In order to apply Theorem 1 to a continuous function we need to know that if f is continuous on $[a, b]$, then f is integrable on $[a, b]$. Although we've already offered one proof of this result, there

is a more elementary argument that you might prefer. Like most “elementary” arguments, it’s quite tricky, but it has the virtue that it will force a review of the proof of Theorem 1.

If f is any bounded function on $[a, b]$, then

$$\sup\{L(f, P)\} \quad \text{and} \quad \inf\{U(f, P)\}$$

will both exist, even if f is not integrable. These numbers are called the **lower integral** of f on $[a, b]$ and the **upper integral** of f on $[a, b]$, respectively, and will be denoted by

$$L \int_a^b f \quad \text{and} \quad U \int_a^b f.$$

The lower and upper integrals both have several properties which the integral possesses. In particular, if $a < c < b$, then

$$L \int_a^b f = L \int_a^c f + L \int_c^b f \quad \text{and} \quad U \int_a^b f = U \int_a^c f + U \int_c^b f,$$

and if $m \leq f(x) \leq M$ for all x in $[a, b]$, then

$$m(b-a) \leq L \int_a^b f \leq U \int_a^b f \leq M(b-a).$$

The proofs of these facts are left as an exercise, since they are quite similar to the corresponding proofs for integrals. The results for integrals are actually a corollary of the results for upper and lower integrals, because f is integrable precisely when

$$L \int_a^b f = U \int_a^b f.$$

We will prove that a continuous function f is integrable by showing that this equality always holds for continuous functions. It is actually easier to show that

$$L \int_a^x f = U \int_a^x f$$

for all x in $[a, b]$; the trick is to note that most of the proof of Theorem 1 didn’t even depend on the fact that f was integrable!

THEOREM 13-3 If f is continuous on $[a, b]$, then f is integrable on $[a, b]$.

PROOF Define functions L and U on $[a, b]$ by

$$L(x) = L \int_a^x f \quad \text{and} \quad U(x) = U \int_a^x f.$$

Let x be in (a, b) . If $h > 0$ and

$$m_h = \inf\{f(t) : x \leq t \leq x+h\}, \\ M_h = \sup\{f(t) : x \leq t \leq x+h\},$$

then

$$m_h \cdot h \leq \mathbf{L} \int_x^{x+h} f \leq \mathbf{U} \int_x^{x+h} f \leq M_h \cdot h,$$

so

$$m_h \cdot h \leq L(x+h) - L(x) \leq U(x+h) - U(x) \leq M_h \cdot h$$

or

$$m_h \leq \frac{L(x+h) - L(x)}{h} \leq \frac{U(x+h) - U(x)}{h} \leq M_h.$$

If $h < 0$ and

$$\begin{aligned} m_h &= \inf\{f(t) : x+h \leq t \leq x\}, \\ M_h &= \sup\{f(t) : x+h \leq t \leq x\}, \end{aligned}$$

one obtains the same inequality, precisely as in the proof of Theorem 1.

Since f is continuous at x , we have

$$\lim_{h \rightarrow 0} m_h = \lim_{h \rightarrow 0} M_h = f(x),$$

and this proves that

$$L'(x) = U'(x) = f(x) \quad \text{for } x \text{ in } (a, b).$$

This means that there is a number c such that

$$U(x) = L(x) + c \quad \text{for all } x \text{ in } [a, b].$$

Since

$$U(a) = L(a) = 0,$$

the number c must equal 0, so

$$U(x) = L(x) \quad \text{for all } x \text{ in } [a, b].$$

In particular,

$$\mathbf{U} \int_a^b f = U(b) = L(b) = \mathbf{L} \int_a^b f,$$

and this means that f is integrable on $[a, b]$. ■

PROBLEMS

1. Find the derivatives of each of the following functions.

$$(i) \quad F(x) = \int_a^{x^3} \sin^3 t dt.$$

$$(ii) \quad F(x) = \int_3^{\left(\int_1^x \sin^3 t dt\right)} \frac{1}{1 + \sin^6 t + t^2} dt$$

$$(iii) \quad F(x) = \int_{15}^x \left(\int_8^y \frac{1}{1 + t^2 + \sin^2 t} dt \right) dy.$$

$$(iv) \quad F(x) = \int_x^b \frac{1}{1 + t^2 + \sin^2 t} dt.$$

- (v) $F(x) = \int_a^b \frac{x}{1+t^2 + \sin^2 t} dt.$
- (vi) $F(x) = \sin \left(\int_0^x \sin \left(\int_0^y \sin^3 t dt \right) dy \right).$
- (vii) F^{-1} , where $F(x) = \int_1^x \frac{1}{t} dt.$
- (viii) F^{-1} , where $F(x) = \int_0^x \frac{1}{\sqrt{1-t^2}} dt.$
- } (Find $(F^{-1})'(x)$ in terms of $F^{-1}(x).$)

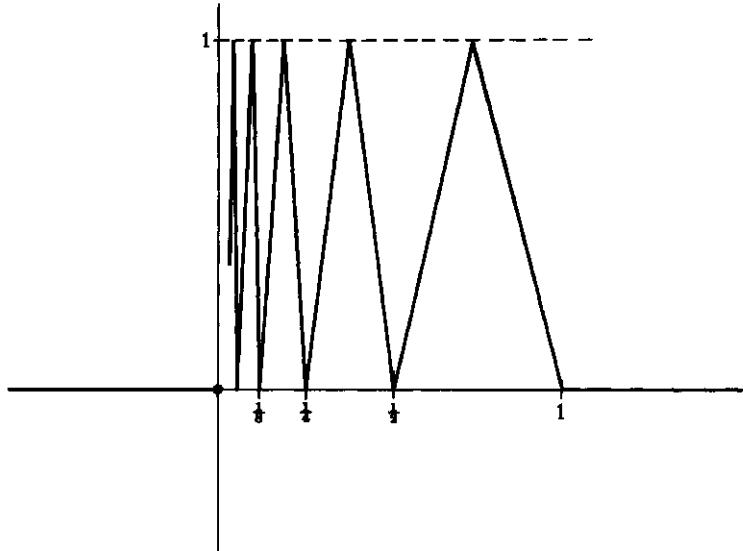


FIGURE 6

2. For each of the following f , if $F(x) = \int_0^x f$, at which points x is $F'(x) = f(x)$? (Caution: it might happen that $F'(x) = f(x)$, even if f is not continuous at x .)
- (i) $f(x) = 0$ if $x \leq 1$, $f(x) = 1$ if $x > 1$.
 - (ii) $f(x) = 0$ if $x < 1$, $f(x) = 1$ if $x \geq 1$.
 - (iii) $f(x) = 0$ if $x \neq 1$, $f(x) = 1$ if $x = 1$.
 - (iv) $f(x) = 0$ if x is irrational, $f(x) = 1/q$ if $x = p/q$ in lowest terms.
 - (v) $f(x) = 0$ if $x \leq 0$, $f(x) = x$ if $x \geq 0$.
 - (vi) $f(x) = 0$ if $x \leq 0$ or $x > 1$, $f(x) = 1/[1/x]$ if $0 < x \leq 1$.
 - (vii) f is the function shown in Figure 6.
 - (viii) $f(x) = 1$ if $x = 1/n$ for some n in \mathbb{N} , $f(x) = 0$ otherwise.
3. Let f be integrable on $[a, b]$, let c be in (a, b) , and let

$$F(x) = \int_a^x f, \quad a \leq x \leq b.$$

For each of the following statements, give either a proof or a counterexample.

- (a) If f is differentiable at c , then F is differentiable at c .

- (b) If f is differentiable at c , then F' is continuous at c .
 (c) If f' is continuous at c , then F' is continuous at c .

4. Show that the values of the following expressions do not depend on x :

$$\begin{aligned} \text{(i)} \quad & \int_0^x \frac{1}{1+t^2} dt + \int_0^{1/x} \frac{1}{1+t^2} dt. \\ \text{(ii)} \quad & \int_{-\cos x}^{\sin x} \frac{1}{\sqrt{1-t^2}} dt, \quad x \in [0, \pi/2]. \end{aligned}$$

5. Find $(f^{-1})'(0)$ if

$$\begin{aligned} \text{(i)} \quad f(x) &= \int_0^x 1 + \sin(\sin t) dt. \\ \text{(ii)} \quad f(x) &= \int_1^x \cos(\cos t) dt. \end{aligned}$$

(Don't try to evaluate f explicitly.)

6. Find a function g such that

$$\begin{aligned} \text{(i)} \quad & \int_0^x tg(t) dt = x + x^2. \\ \text{(ii)} \quad & \int_0^{x^2} tg(t) dt = x + x^2. \end{aligned}$$

(Notice that g is not assumed continuous at 0.)

7. Find all continuous functions f satisfying

$$\int_0^x f = (f(x))^2 + C.$$

for some constant C .

- *8. Suppose that f is a differentiable function with $f(0) = 0$ and $0 < f' \leq 1$. Prove that for all $x \geq 0$ we have

$$\int_0^x f^3 \leq \left(\int_0^x f \right)^2.$$

- *9. Let

$$f(x) = \begin{cases} \cos \frac{1}{x}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

Is the function $F(x) = \int_0^x f$ differentiable at 0? Hint: Stare at page 177.

10. Use Problem 13-23 to prove that

$$\begin{aligned} \text{(i)} \quad & \frac{1}{7\sqrt{2}} \leq \int_0^1 \frac{x^6}{\sqrt{1+x^2}} dx \leq \frac{1}{7}. \\ \text{(ii)} \quad & \frac{3}{8} \leq \int_0^{1/2} \sqrt{\frac{1-x}{1+x}} dx \leq \frac{\sqrt{3}}{4}. \end{aligned}$$

11. Find $F'(x)$ if $F(x) = \int_0^x xf(t) dt$. (The answer is *not* $xf(x)$; you should perform an obvious manipulation on the integral before trying to find F' .)

12. Prove that if f is continuous, then

$$\int_0^x f(u)(x-u) du = \int_0^x \left(\int_0^u f(t) dt \right) du.$$

Hint: Differentiate both sides, making use of Problem 11.

- *13. Use Problem 12 to prove that

$$\int_0^x f(u)(x-u)^2 du = 2 \int_0^x \left(\int_0^{u_2} \left(\int_0^{u_1} f(t) dt \right) du_1 \right) du_2.$$

14. Find a function f such that $f'''(x) = 1 / \sqrt{1 + \sin^2 x}$. (This problem is supposed to be easy; don't misinterpret the word "find.")

- *15. A function f is **periodic**, with **period a** , if $f(x+a) = f(x)$ for all x .

- (a) If f is periodic with period a and integrable on $[0, a]$, show that

$$\int_0^a f = \int_b^{b+a} f \quad \text{for all } b.$$

- (b) Find a function f such that f is not periodic, but f' is. Hint: Choose a periodic g for which it can be guaranteed that $f(x) = \int_0^x g$ is not periodic.

- (c) Suppose that f' is periodic with period a . Prove that f is periodic if and only if $f(a) = f(0)$.

16. Find $\int_0^b \sqrt[3]{x} dx$, by simply guessing a function f with $f'(x) = \sqrt[3]{x}$, and using the Second Fundamental Theorem of Calculus. Then check with Problem 13-21.

- *17. Use the Fundamental Theorem of Calculus and Problem 13-21 to derive the result stated in Problem 12-18.

18. Let C_1 , C and C_2 be curves passing through the origin, as shown in Figure 7. Each point on C can be joined to a point of C_1 with a vertical line segment and to a point of C_2 with a horizontal line segment. We will say that C bisects C_1 and C_2 if the regions A and B have equal areas for every point on C .

- (a) If C_1 is the graph of $f(x) = x^2$, $x \geq 0$ and C is the graph of $f(x) = 2x^2$, $x \geq 0$, find C_2 so that C bisects C_1 and C_2 .

- (b) More generally, find C_2 if C_1 is the graph of $f(x) = x^m$, and C is the graph of $f(x) = cx^m$ for some $c > 1$.

19. (a) Find the derivatives of $F(x) = \int_1^x 1/t dt$ and $G(x) = \int_b^{bx} 1/t dt$.
(b) Now give a new proof for Problem 13-15.

- *20. Use the Fundamental Theorem of Calculus and Darboux's Theorem (Problem 11-54) to give another proof of the Intermediate Value Theorem.

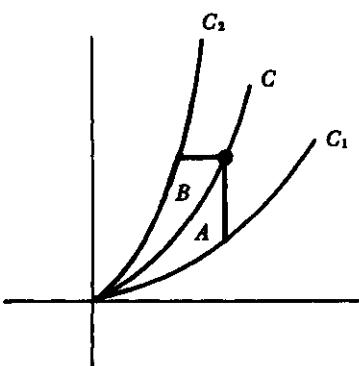


FIGURE 7

21. Prove that if h is continuous, f and g are differentiable, and

$$F(x) = \int_{f(x)}^{g(x)} h(t) dt,$$

then $F'(x) = h(g(x)) \cdot g'(x) - h(f(x)) \cdot f'(x)$. Hint: Try to reduce this to the two cases you can already handle, with a constant either as the lower or the upper limit of integration.

22. Suppose that f' is integrable on $[0, 1]$ and $f(0) = 0$. Prove that for all x in $[0, 1]$ we have

$$|f(x)| \leq \sqrt{\int_0^1 |f'|^2}.$$

Show also that the hypothesis $f(0) = 0$ is needed. Hint: Problem 13-39.

- *23. (a) Suppose $G' = g$ and $F' = f$. Prove that if the function y satisfies the differential equation

$$(*) \quad g(y(x)) \cdot y'(x) = f(x) \quad \text{for all } x \text{ in some interval},$$

then there is a number c such that

$$(**) \quad G(y(x)) = F(x) + c \quad \text{for all } x \text{ in this interval.}$$

- (b) Show, conversely, that if y satisfies (**), then y is a solution of (*).
(c) Find what condition y must satisfy if

$$y'(x) = \frac{1+x^2}{1+y(x)}.$$

(In this case $g(t) = 1+t$ and $f(t) = 1+t^2$.) Then “solve” the resulting equations to find all possible solutions y (no solution will have \mathbf{R} as its domain).

- (d) Find what condition y must satisfy if

$$y'(x) = \frac{-1}{1+5[y(x)]^4}.$$

(An appeal to Problem 12-11 will show that there *are* functions satisfying the resulting equation.)

- (e) Find all functions y satisfying

$$y(x)y'(x) = -x.$$

Find the solution y satisfying $y(0) = -1$.

24. In Problem 10-17 we found that the Schwarzian derivative

$$\frac{f'''(x)}{f'(x)} - \frac{3}{2} \left(\frac{f''(x)}{f'(x)} \right)^2$$

was 0 for $f(x) = (ax+b)/(cx+d)$. Now suppose that f is any function whose Schwarzian derivative is 0.

(a) f''^2/f'^3 is a constant function.

(b) f is the form $f(x) = (ax + b)/(cx + d)$. Hint: Consider $u = f'$ and apply the previous problem.

- *25. The limit $\lim_{N \rightarrow \infty} \int_a^N f$, if it exists, is denoted by $\int_a^\infty f$ (or $\int_a^\infty f(x) dx$), and called an “improper integral.”

(a) Determine $\int_1^\infty x^r dx$, if $r < -1$.

(b) Use Problem 13-15 to show that $\int_1^\infty 1/x dx$ does not exist. Hint: What can you say about $\int_1^{2^n} 1/x dx$?

(c) Suppose that $f(x) \geq 0$ for $x \geq 0$ and that $\int_0^\infty f$ exists. Prove that if $0 \leq g(x) \leq f(x)$ for all $x \geq 0$, and g is integrable on each interval $[0, N]$, then $\int_0^\infty g$ also exists.

(d) Explain why $\int_0^\infty 1/(1+x^2) dx$ exists. Hint: Split this integral up at 1.

26. Decide whether or not the following improper integrals exist.

$$(i) \quad \int_0^\infty \frac{1}{\sqrt{1+x^3}} dx.$$

$$(ii) \quad \int_0^\infty \frac{x}{1+x^{3/2}} dx.$$

$$(iii) \quad \int_0^\infty \frac{1}{x\sqrt{1+x}} dx.$$

- *27. The improper integral $\int_{-\infty}^a f$ is defined in the obvious way, as $\lim_{N \rightarrow -\infty} \int_N^a f$. But another kind of improper integral $\int_{-\infty}^\infty f$ is defined in a nonobvious way: it is $\int_0^\infty f + \int_{-\infty}^0 f$, provided these improper integrals both exist.

(a) Explain why $\int_{-\infty}^\infty 1/(1+x^2) dx$ exists.

(b) Explain why $\int_{-\infty}^\infty x dx$ does not exist. (But notice that $\lim_{N \rightarrow \infty} \int_{-N}^N x dx$ does exist.)

(c) Prove that if $\int_{-\infty}^\infty f$ exists, then $\lim_{N \rightarrow \infty} \int_{-N}^N f$ exists and equals $\int_{-\infty}^\infty f$. Show moreover, that $\lim_{N \rightarrow \infty} \int_{-N}^{N+1} f$ and $\lim_{N \rightarrow \infty} \int_{-N}^{N^2} f$ both exist and equal $\int_{-\infty}^\infty f$. Can you state a reasonable generalization of these facts? (If you can't, you will have a miserable time trying to do these special cases!)

- *28. There is another kind of “improper integral” in which the interval is bounded, but the *function* is unbounded:

(a) If $a > 0$, find $\lim_{\epsilon \rightarrow 0^+} \int_\epsilon^a 1/\sqrt{x} dx$. This limit is denoted by $\int_0^a 1/\sqrt{x} dx$, even though the function $f(x) = 1/\sqrt{x}$ is not bounded on $[0, a]$, no matter how we define $f(0)$.

(b) Find $\int_0^a x^r dx$ if $-1 < r < 0$.

(c) Use Problem 13-15 to show that $\int_0^a x^{-1} dx$ does not make sense, even as a limit.

- (d) Invent a reasonable definition of $\int_a^0 |x|^r dx$ for $a < 0$ and compute it for $-1 < r < 0$.
- (e) Invent a reasonable definition of $\int_{-1}^1 (1 - x^2)^{-1/2} dx$, as a sum of two limits, and show that the limits exist. Hint: Why does $\int_{-1}^0 (1 + x)^{-1/2} dx$ exist? How does $(1 + x)^{-1/2}$ compare with $(1 - x^2)^{-1/2}$ for $-1 < x < 0$?
29. (a) If f is continuous on $[0, 1]$, compute $\lim_{x \rightarrow 0^+} x \int_x^1 \frac{f(t)}{t} dt$.
- (b) If f is integrable on $[0, 1]$ and $\lim_{x \rightarrow 0^+} f(x)$ exists, compute $\lim_{x \rightarrow 0^+} x \int_x^1 \frac{f(t)}{t^2} dt$.
- *30. It is possible, finally, to combine the two possible extensions of the notion of the integral.
- (a) If $f(x) = 1/\sqrt{x}$ for $0 \leq x \leq 1$ and $f(x) = 1/x^2$ for $x \geq 1$, find $\int_0^\infty f(x) dx$ (after deciding what this should mean).
- (b) Show that $\int_0^\infty x^r dx$ never makes sense. (Distinguish the cases $-1 < r < 0$ and $r < -1$. In one case things go wrong at 0, in the other case at ∞ ; for $r = -1$ things go wrong at both places.)

CHAPTER 15

THE TRIGONOMETRIC FUNCTIONS

The definitions of the functions \sin and \cos are considerably more subtle than one might suspect. For this reason, this chapter begins with some informal and intuitive definitions, which should not be scrutinized too carefully, as they shall soon be replaced by the formal definitions which we really intend to use.

In elementary geometry an angle is simply the union of two half-lines with a common initial point (Figure 1).

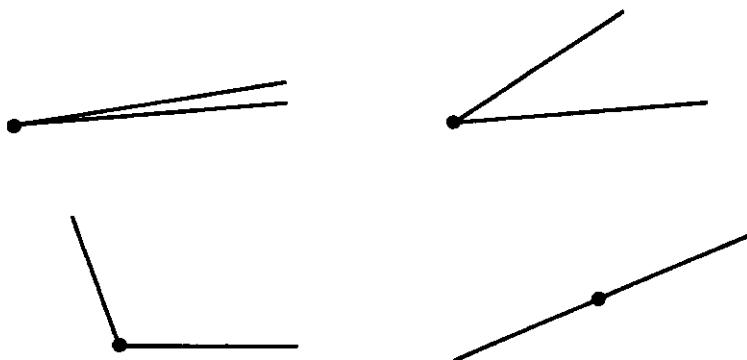


FIGURE 1

More useful for trigonometry are “directed angles,” which may be regarded as pairs (l_1, l_2) of half-lines with the same initial point, visualized as in Figure 2.

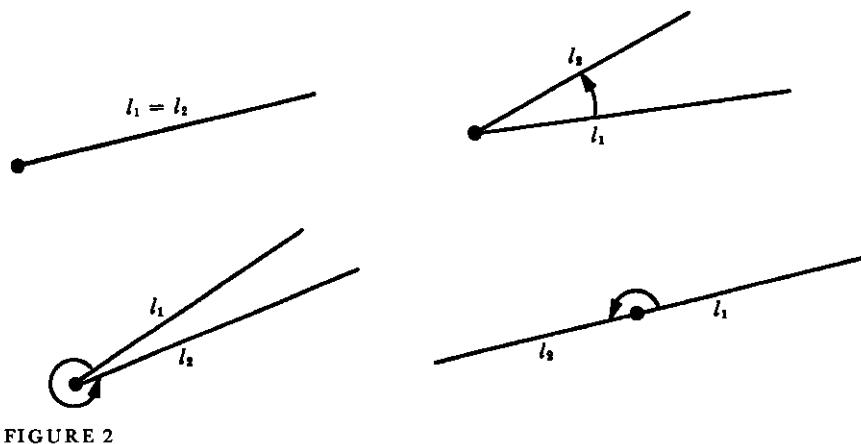


FIGURE 2

If for l_1 we always choose the positive half of the horizontal axis, a directed angle is described completely by the second half-line (Figure 3).

Since each half-line intersects the unit circle precisely once, a directed angle is described, even more simply, by a point on the unit circle (Figure 4), that is, by a point (x, y) with $x^2 + y^2 = 1$.

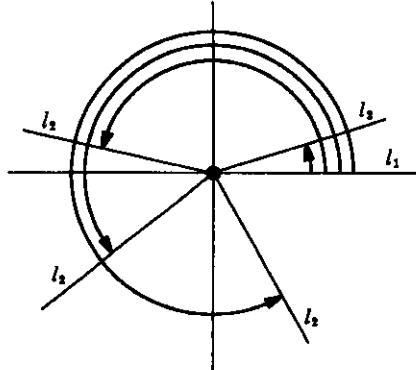


FIGURE 3

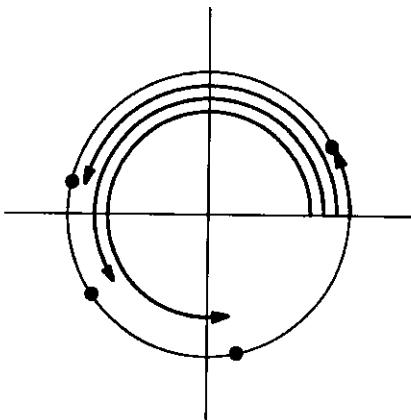


FIGURE 4

The sine and cosine of a directed angle can now be defined as follows (Figure 5): a directed angle is determined by a point (x, y) with $x^2 + y^2 = 1$; the sine of the angle is defined as y , and the cosine as x .

Despite the aura of precision surrounding the previous paragraph, we are not yet finished with the definitions of \sin and \cos . Indeed, we have barely begun. What we have defined is the sine and cosine of a directed angle; what we *want* to define is $\sin x$ and $\cos x$ for each *number* x . The usual procedure for doing this depends on associating an angle to every number. The oldest method is to "measure angles in degrees." An angle "all the way around" is associated to 360, an angle "half-way around" is associated to 180, an angle "a quarter way around" to 90, etc. (Figure 6). The angle associated, in this manner, to the number x , is called "the angle of x degrees." The angle of 0 degrees is the same as the angle of 360 degrees, and this ambiguity is purposely extended further, so that an angle of 90 degrees is also an angle of 360 + 90 degrees, etc. One can now define a function, which we will denote by \sin° , as follows:

$$\sin^\circ(x) = \text{sine of the angle of } x \text{ degrees.}$$

There are two difficulties with this approach. Although it may be clear what we mean by an angle of 90 or 45 degrees, it is not quite clear what an angle of $\sqrt{2}$ degrees is, for example. Even if this difficulty could be circumvented, it is unlikely that this system, depending as it does on the arbitrary choice of 360, will lead to elegant results—it would be sheer luck if the function \sin° had mathematically pleasing properties.

"Radian measure" appears to offer a remedy for both these defects. Given any number x , choose a point P on the unit circle such that x is the length of the arc of the circle beginning at $(1, 0)$ and running counterclockwise to P (Figure 7). The directed angle determined by P is called "the angle of x radians." Since the length of the whole circle is 2π , the angle of x radians and the angle of $2\pi + x$ radians are identical. A function \sin' can now be defined as follows:

$$\sin'(x) = \text{sine of the angle of } x \text{ radians.}$$

This same method can easily be adopted to define \sin° ; since we want to have $\sin^\circ 360 = \sin' 2\pi$, we can define

$$\sin^\circ x = \sin' \frac{2\pi x}{360} = \sin' \frac{\pi x}{180}.$$

We shall soon drop the superscript r in \sin' , since \sin' (and not \sin°) is the only function which will interest us; before we do, a few words of warning are advisable.

The expressions $\sin^\circ x$ and $\sin' x$ are sometimes written

$$\begin{aligned} &\sin x^\circ \\ &\sin x \text{ radians,} \end{aligned}$$

but this notation is quite misleading; a number x is simply a number—it does not carry a banner indicating that it is "in degrees" or "in radians." If the meaning

FIGURE 5

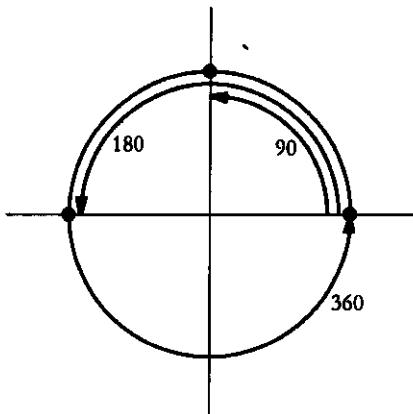


FIGURE 6

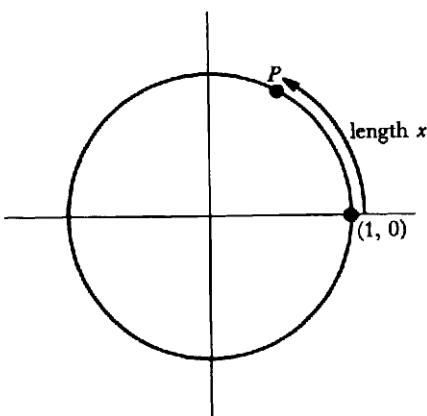


FIGURE 7

of the notation “ $\sin x$ ” is in doubt one usually asks:

“Is x in degrees or radians?”

but what one means is:

“Do you mean ‘ \sin ’ or ‘ \sin' ?’”

Even for mathematicians, addicted to precision, these remarks might be dispensable, were it not for the fact that failure to take them into account will lead to incorrect answers to certain problems (an example is given in Problem 19).

Although the function \sin' is the function which we wish to denote simply by \sin (and use exclusively henceforth), there is a difficulty involved even in the definition of \sin' . Our proposed definition depends on the concept of the length of a curve. Although the length of a curve has been defined in several problems, it is also easy to reformulate the definition in terms of areas. (A treatment in terms of length is outlined in Problem 28.)

Suppose that x is the length of the arc of the unit circle from $(1, 0)$ to P ; this arc thus contains $x/2\pi$ of the total length 2π of the circumference of the unit circle. Let S denote the “sector” shown in Figure 8; S is bounded by the unit circle, the horizontal axis, and the half-line through $(0, 0)$ and P . The area of S should be $x/2\pi$ times the area inside the unit circle, which we expect to be π ; thus S should have area

$$\frac{x}{2\pi} \cdot \pi = \frac{x}{2}.$$

We can therefore define $\cos x$ and $\sin x$ as the coordinates of the point P which determines a sector of area $x/2$.

With these remarks as background, the rigorous definition of the functions \sin and \cos now begins. The first definition identifies π as the area of the unit circle—more precisely, as twice the area of a semicircle (Figure 9).

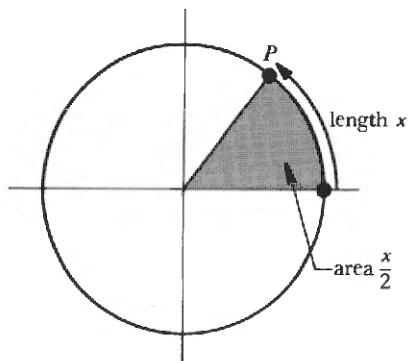


FIGURE 8

DEFINITION

$$\pi = 2 \cdot \int_{-1}^1 \sqrt{1 - x^2} dx.$$

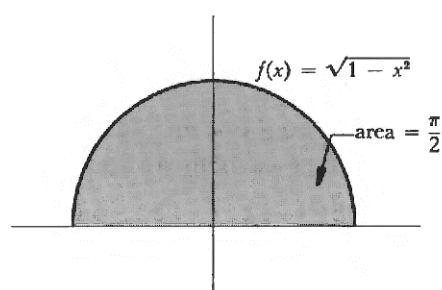


FIGURE 9

(This definition is not offered simply as an embellishment; to define the trigonometric functions it will be necessary to first define $\sin x$ and $\cos x$ only for $0 \leq x \leq \pi$.)

The second definition is meant to describe, for $-1 \leq x \leq 1$, the area $A(x)$ of the sector bounded by the unit circle, the horizontal axis, and the half-line through $(x, \sqrt{1 - x^2})$. If $0 \leq x \leq 1$, this area can be expressed (Figure 10) as the sum of the area of a triangle and the area of a region under the unit circle:

$$\frac{x\sqrt{1 - x^2}}{2} + \int_x^1 \sqrt{1 - t^2} dt.$$

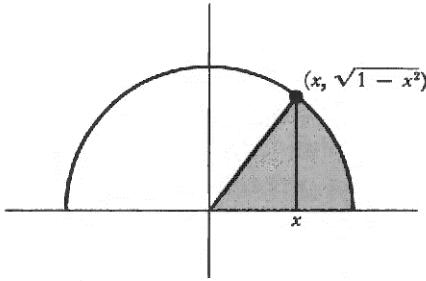


FIGURE 10

This same formula happens to work for $-1 \leq x \leq 0$ also. In this case (Figure 11), the term

$$\frac{x\sqrt{1-x^2}}{2}$$

is negative, and represents the area of the triangle which must be subtracted from the term

$$\int_x^1 \sqrt{1-t^2} dt.$$

DEFINITION

If $-1 \leq x \leq 1$, then

$$A(x) = \frac{x\sqrt{1-x^2}}{2} + \int_x^1 \sqrt{1-t^2} dt.$$

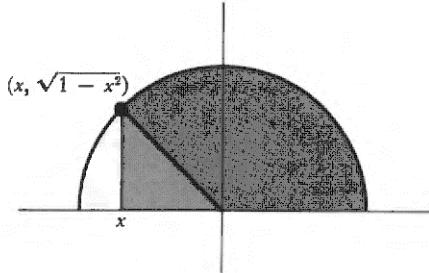


FIGURE 11

Notice that if $-1 < x < 1$, then A is differentiable at x and (using the Fundamental Theorem of Calculus),

$$\begin{aligned} A'(x) &= \frac{1}{2} \left[x \cdot \frac{-2x}{2\sqrt{1-x^2}} + \sqrt{1-x^2} \right] - \sqrt{1-x^2} \\ &= \frac{1}{2} \left[\frac{-x^2 + (1-x^2)}{\sqrt{1-x^2}} \right] - \sqrt{1-x^2} \\ &= \frac{1-2x^2}{2\sqrt{1-x^2}} - \sqrt{1-x^2} \\ &= \frac{1-2x^2-2(1-x^2)}{2\sqrt{1-x^2}} \\ &= \frac{-1}{2\sqrt{1-x^2}}. \end{aligned}$$

Notice also (Figure 12) that on the interval $[-1, 1]$ the function A decreases from

$$A(-1) = 0 + \int_{-1}^1 \sqrt{1-t^2} dt = \frac{\pi}{2}$$

to $A(1) = 0$. This follows directly from the definition of A , and also from the fact that its derivative is negative on $(-1, 1)$.

For $0 \leq x \leq \pi$ we wish to define $\cos x$ and $\sin x$ as the coordinates of a point $P = (\cos x, \sin x)$ on the unit circle which determines a sector whose area is $x/2$ (Figure 13). In other words:

DEFINITION

If $0 \leq x \leq \pi$, then **cos** x is the unique number in $[-1, 1]$ such that

$$A(\cos x) = \frac{x}{2};$$

and

$$\sin x = \sqrt{1 - (\cos x)^2}.$$

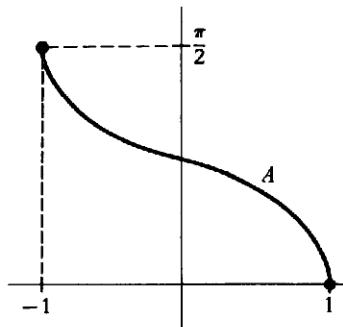


FIGURE 12

This definition actually requires a few words of justification. In order to know that there *is* a number y satisfying $A(y) = x/2$, we use the fact that A is continuous, and that A takes on the values 0 and $\pi/2$. This tacit appeal to the Intermediate Value Theorem is crucial, if we want to make our preliminary definition precise. Having made, and justified, our definition, we can now proceed quite rapidly.

THEOREM 1 If $0 < x < \pi$, then

$$\begin{aligned}\cos'(x) &= -\sin x, \\ \sin'(x) &= \cos x.\end{aligned}$$

PROOF If $B = 2A$, then the definition $A(\cos x) = x/2$ can be written

$$B(\cos x) = x;$$

in other words, \cos is just the inverse of B . We have already computed that

$$A'(x) = -\frac{1}{2\sqrt{1-x^2}},$$

from which we conclude that

$$B'(x) = -\frac{1}{\sqrt{1-x^2}}.$$

Consequently,

$$\begin{aligned}\cos'(x) &= (B^{-1})'(x) \\ &= \frac{1}{B'(B^{-1}(x))} \\ &= \frac{1}{-\frac{1}{\sqrt{1-[B^{-1}(x)]^2}}} \\ &= -\sqrt{1-(\cos x)^2} \\ &= -\sin x.\end{aligned}$$

Since

$$\sin x = \sqrt{1-(\cos x)^2},$$

we also obtain

$$\begin{aligned}\sin'(x) &= \frac{1}{2} \cdot \frac{-2\cos x \cdot \cos'(x)}{\sqrt{1-(\cos x)^2}} \\ &= \frac{\cos x \sin x}{\sin x} \\ &= \cos x. \blacksquare\end{aligned}$$

The information contained in Theorem 1 can be used to sketch the graphs of

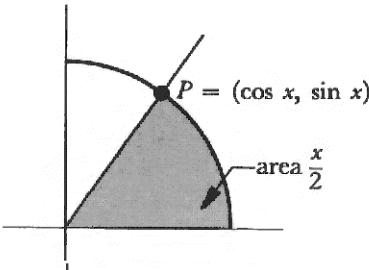


FIGURE 13

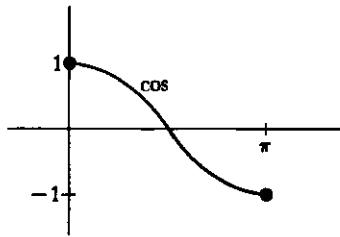


FIGURE 14

\sin and \cos on the interval $[0, \pi]$. Since

$$\cos'(x) = -\sin x < 0, \quad 0 < x < \pi,$$

the function \cos decreases from $\cos 0 = 1$ to $\cos \pi = -1$ (Figure 14). Consequently, $\cos y = 0$ for a unique y in $[0, \pi]$. To find y , we note that the definition of \cos ,

$$A(\cos x) = \frac{x}{2},$$

means that

$$A(0) = \frac{y}{2},$$

so

$$y = 2 \int_0^1 \sqrt{1-t^2} dt.$$

It is easy to see that

$$\int_{-1}^0 \sqrt{1-t^2} dt = \int_0^1 \sqrt{1-t^2} dt$$

so we can also write

$$y = \int_{-1}^1 \sqrt{1-t^2} dt = \frac{\pi}{2}.$$

Now we have

$$\sin'(x) = \cos x \begin{cases} > 0, & 0 < x < \pi/2 \\ < 0, & \pi/2 < x < \pi, \end{cases}$$

so \sin increases on $[0, \pi/2]$ from $\sin 0 = 0$ to $\sin \pi/2 = 1$, and then decreases on $[\pi/2, \pi]$ to $\sin \pi = 0$ (Figure 15).

The values of $\sin x$ and $\cos x$ for x not in $[0, \pi]$ are most easily defined by a two-step piecing together process:

- (1) If $\pi \leq x \leq 2\pi$, then

$$\begin{aligned} \sin x &= -\sin(2\pi - x), \\ \cos x &= \cos(2\pi - x). \end{aligned}$$

Figure 16 shows the graphs of \sin and \cos on $[0, 2\pi]$.

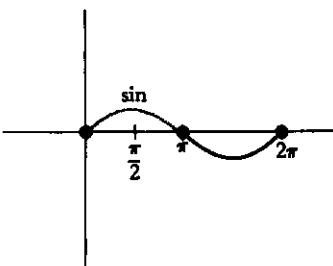
- (2) If $x = 2\pi k + x'$ for some integer k , and some x' in $[0, 2\pi]$, then

$$\begin{aligned} \sin x &= \sin x', \\ \cos x &= \cos x'. \end{aligned}$$

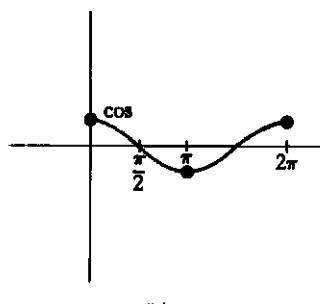
Figure 17 shows the graphs of \sin and \cos , now defined on all of \mathbf{R} .

Having extended the functions \sin and \cos to \mathbf{R} , we must now check that the basic properties of these functions continue to hold. In most cases this is easy. For example, it is clear that the equation

$$\sin^2 x + \cos^2 x = 1$$



(a)



(b)

FIGURE 16

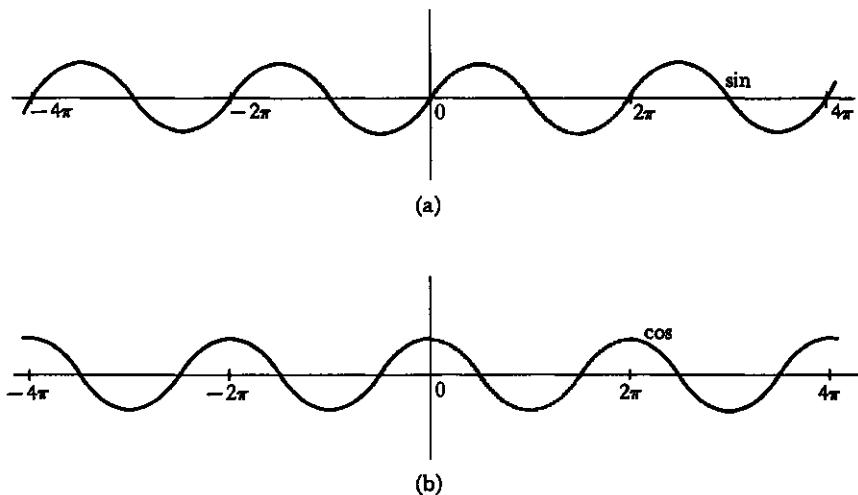


FIGURE 17

holds for all x . It is also not hard to prove that

$$\begin{aligned}\sin'(x) &= \cos x, \\ \cos'(x) &= -\sin x,\end{aligned}$$

if x is not a multiple of π . For example, if $\pi < x < 2\pi$, then

$$\sin x = -\sin(2\pi - x),$$

so

$$\begin{aligned}\sin'(x) &= -\sin'(2\pi - x) \cdot (-1) \\ &= \cos(2\pi - x) \\ &= \cos x.\end{aligned}$$

If x is a multiple of π we resort to a trick; it is only necessary to apply Theorem 11-7 to conclude that the same formulas are true in this case also.

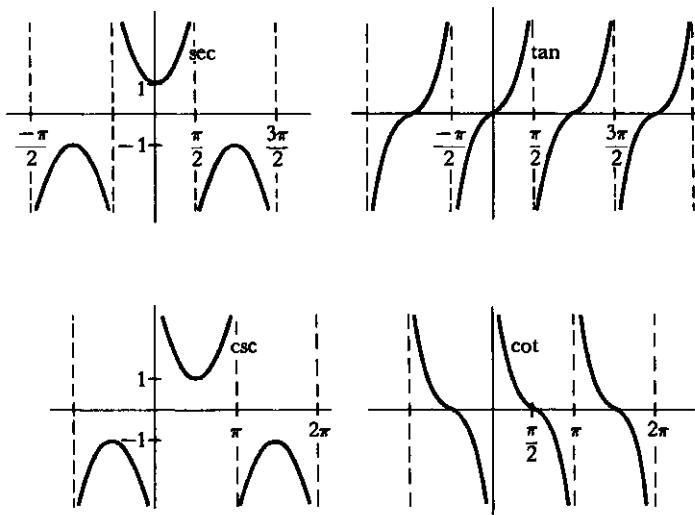


FIGURE 18

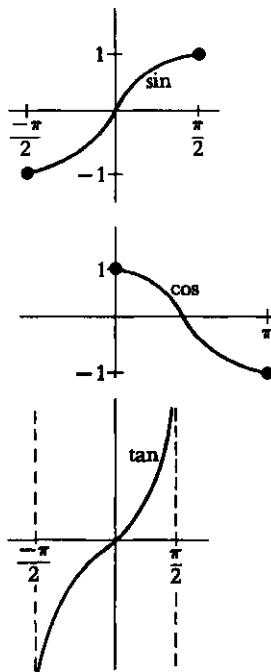


FIGURE 19

The other standard trigonometric functions present no difficulty at all. We define

$$\left. \begin{array}{l} \sec x = \frac{1}{\cos x} \\ \tan x = \frac{\sin x}{\cos x} \end{array} \right\} x \neq k\pi + \pi/2,$$

$$\left. \begin{array}{l} \csc x = \frac{1}{\sin x} \\ \cot x = \frac{\cos x}{\sin x} \end{array} \right\} x \neq k\pi.$$

The graphs are sketched in Figure 18. It is a good idea to convince yourself that the general features of these graphs can be predicted from the derivatives of these functions, which are listed in the next theorem (there is no need to memorize the statement of the theorem, since the results can be rederived whenever needed.)

THEOREM 2 If $x \neq k\pi + \pi/2$, then

$$\begin{aligned} \sec'(x) &= \sec x \tan x, \\ \tan'(x) &= \sec^2 x. \end{aligned}$$

If $x \neq k\pi$, then

$$\begin{aligned} \csc'(x) &= -\csc x \cot x, \\ \cot'(x) &= -\csc^2 x. \end{aligned}$$

PROOF Left to you (a straightforward computation). ■

The inverses of the trigonometric functions are also easily differentiated. The trigonometric functions are not one-one, so it is first necessary to restrict them to suitable intervals; the largest possible length obtainable is π , and the intervals usually chosen are (Figure 19)

$[-\pi/2, \pi/2]$ $[0, \pi]$ $(-\pi/2, \pi/2)$	for \sin , for \cos , for \tan .
--	--

(The inverses of the other trigonometric functions are so rarely used that they will not even be discussed here.)

The inverse of the function

$$f(x) = \sin x, \quad -\pi/2 \leq x \leq \pi/2$$

is denoted by **arcsin** (Figure 20); the domain of arcsin is $[-1, 1]$. The notation \sin^{-1} has been avoided because arcsin is not the inverse of sin (which is not one-one), but of the restricted function f ; sometimes this function f is denoted by Sin, and arcsin by Sin^{-1} .

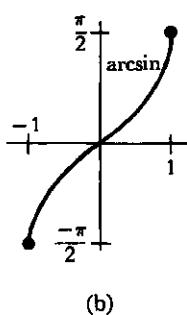


FIGURE 20

The inverse of the function

$$g(x) = \cos x, \quad 0 \leq x \leq \pi$$

is denoted by **arccos** (Figure 21); the domain of arccos is $[-1, 1]$. Sometimes g is denoted by Cos , and arccos by Cos^{-1} .

The inverse of the function

$$h(x) = \tan x, \quad -\pi/2 < x < \pi/2$$

is denoted by **arctan** (Figure 22); arctan is one of the simplest examples of a differentiable function which is bounded even though it is one-one on all of \mathbf{R} . Sometimes the function h is denoted by Tan , and arctan by Tan^{-1} .

The derivatives of the inverse trigonometric functions are surprisingly simple, and do not involve trigonometric functions at all. Finding the derivatives is a simple matter, but to express them in a suitable form we will have to simplify expressions like

$$\cos(\arcsin x), \quad \sec(\arctan x).$$

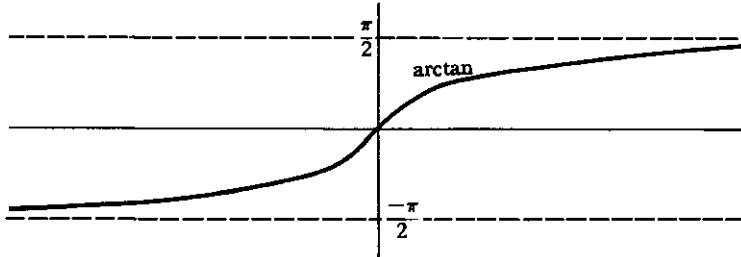


FIGURE 22

A little picture is the best way to remember the correct simplifications. For example, Figure 23 shows a directed angle whose sine is x —the angle shown is thus an angle of $(\arcsin x)$ radians; consequently $\cos(\arcsin x)$ is the length of the other side, namely, $\sqrt{1 - x^2}$. However, in the proof of the next theorem we will not resort to such pictures.

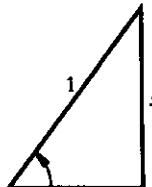


FIGURE 23

THEOREM 3 If $-1 < x < 1$, then

$$\begin{aligned}\arcsin'(x) &= \frac{1}{\sqrt{1-x^2}}, \\ \arccos'(x) &= \frac{-1}{\sqrt{1-x^2}}.\end{aligned}$$

Moreover, for all x we have

$$\arctan'(x) = \frac{1}{1+x^2}.$$

PROOF

$$\begin{aligned}\arcsin'(x) &= (f^{-1})'(x) \\ &= \frac{1}{f'(f^{-1}(x))} \\ &= \frac{1}{\sin'(\arcsin x)} \\ &= \frac{1}{\cos(\arcsin x)}.\end{aligned}$$

Now

$$[\sin(\arcsin x)]^2 + [\cos(\arcsin x)]^2 = 1,$$

that is,

$$x^2 + [\cos(\arcsin x)]^2 = 1;$$

therefore,

$$\cos(\arcsin x) = \sqrt{1 - x^2}.$$

(The positive square root is to be taken because $\arcsin x$ is in $(-\pi/2, \pi/2)$, so $\cos(\arcsin x) > 0$.) This proves the first formula.

The second formula has already been established (in the proof of Theorem 1). It is also possible to imitate the proof for the first formula, a valuable exercise if that proof presented any difficulties. The third formula is proved as follows.

$$\begin{aligned}\arctan'(x) &= (h^{-1})'(x) \\ &= \frac{1}{h'(h^{-1}(x))} \\ &= \frac{1}{\tan'(\arctan x)} \\ &= \frac{1}{\sec^2(\arctan x)}\end{aligned}$$

Dividing both sides of the identity

$$\sin^2 a + \cos^2 a = 1$$

by $\cos^2 a$ yields

$$\tan^2 a + 1 = \sec^2 a.$$

It follows that

$$[\tan(\arctan x)]^2 + 1 = \sec^2(\arctan x),$$

or

$$x^2 + 1 = \sec^2(\arctan x),$$

which proves the third formula. ■

The traditional proof of the formula $\sin'(x) = \cos x$ (quite different from the one given here) is outlined in Problem 27. This proof depends upon first establishing

the limit

$$\lim_{h \rightarrow 0} \frac{\sin h}{h} = 1,$$

and the “addition formula”

$$\sin(x + y) = \sin x \cos y + \cos x \sin y.$$

Both of these formulas can be derived easily now that the derivative of \sin and \cos are known. The first is just the special case $\sin'(0) = \cos 0$. The second depends on a beautiful characterization of the functions \sin and \cos . In order to derive this result we need a lemma whose proof involves a clever trick; a more straightforward proof will be supplied in Part IV.

LEMMA Suppose f has a second derivative everywhere and that

$$\begin{aligned} f'' + f &= 0, \\ f(0) &= 0, \\ f'(0) &= 0. \end{aligned}$$

Then $f = 0$.

PROOF Multiplying both sides of the first equation by f' yields

$$f'f'' + ff' = 0.$$

Thus

$$[(f')^2 + f^2]' = 2(f'f'' + ff') = 0,$$

so $(f')^2 + f^2$ is a constant function. From $f(0) = 0$ and $f'(0) = 0$ it follows that the constant is 0; thus

$$[f'(x)]^2 + [f(x)]^2 = 0 \quad \text{for all } x.$$

This implies that

$$f(x) = 0 \quad \text{for all } x. \blacksquare$$

THEOREM 4 If f has a second derivative everywhere and

$$\begin{aligned} f'' + f &= 0, \\ f(0) &= a, \\ f'(0) &= b, \end{aligned}$$

then

$$f = b \cdot \sin x + a \cdot \cos x.$$

(In particular, if $f(0) = 0$ and $f'(0) = 1$, then $f = \sin$; if $f(0) = 1$ and $f'(0) = 0$, then $f = \cos$.)

PROOF Let

$$g(x) = f(x) - b \sin x - a \cos x.$$

Then

$$\begin{aligned} g'(x) &= f'(x) - b \cos x + a \sin x, \\ g''(x) &= f''(x) + b \sin x + a \cos x. \end{aligned}$$

Consequently,

$$\begin{aligned} g'' + g &= 0, \\ g(0) &= 0, \\ g'(0) &= 0, \end{aligned}$$

which shows that

$$0 = g(x) = f(x) - b \sin x - a \cos x, \quad \text{for all } x. \blacksquare$$

THEOREM 5 If x and y are any two numbers, then

$$\begin{aligned} \sin(x + y) &= \sin x \cos y + \cos x \sin y, \\ \cos(x + y) &= \cos x \cos y - \sin x \sin y. \end{aligned}$$

PROOF For any particular number y we can define a function f by

$$f(x) = \sin(x + y).$$

Then

$$\begin{aligned} f'(x) &= \cos(x + y) \\ f''(x) &= -\sin(x + y). \end{aligned}$$

Consequently,

$$\begin{aligned} f'' + f &= 0, \\ f(0) &= \sin y, \\ f'(0) &= \cos y. \end{aligned}$$

It follows from Theorem 4 that

$$f = (\cos y) \cdot \sin + (\sin y) \cdot \cos;$$

that is,

$$\sin(x + y) = \cos y \sin x + \sin y \cos x, \quad \text{for all } x.$$

Since any number y could have been chosen to begin with, this proves the first formula for all x and y .

The second formula is proved similarly. \blacksquare

As a conclusion to this chapter, and as a prelude to Chapter 18, we will mention an alternative approach to the definition of the function \sin . Since

$$\arcsin'(x) = \frac{1}{\sqrt{1-x^2}} \quad \text{for } -1 < x < 1,$$

it follows from the Second Fundamental Theorem of Calculus that

$$\arcsin x = \arcsin x - \arcsin 0 = \int_0^x \frac{1}{\sqrt{1-t^2}} dt.$$

This equation could have been taken as the *definition* of \arcsin . It would follow immediately that

$$\arcsin'(x) = \frac{1}{\sqrt{1-x^2}};$$

the function \sin could then be defined as $(\arcsin)^{-1}$ and the formula for the derivative of an inverse function would show that

$$\sin'(x) = \sqrt{1-\sin^2 x},$$

which could be defined as $\cos x$. Eventually, one could show that $A(\cos x) = x/2$, recovering at the very end of the development the definition with which we started. While much of this presentation would proceed more rapidly, the definition would be utterly unmotivated; the reasonableness of the definitions would be known to the author, but not to the student, for whom it was intended! Nevertheless, as we shall see in Chapter 18, an approach of this sort is sometimes very reasonable indeed.

PROBLEMS

1. Differentiate each of the following functions.

$$\begin{aligned} \text{(i)} \quad f(x) &= \arctan(\arctan(\arctan x)). \\ \text{(ii)} \quad f(x) &= \arcsin(\arctan(\arccos x)). \\ \text{(iii)} \quad f(x) &= \arctan(\tan x \arctan x). \\ \text{(iv)} \quad f(x) &= \arcsin\left(\frac{1}{\sqrt{1+x^2}}\right). \end{aligned}$$

2. Find the following limits by l'Hôpital's Rule.

$$\begin{aligned} \text{(i)} \quad \lim_{x \rightarrow 0} \frac{\sin x - x + x^3/6}{x^3}. \\ \text{(ii)} \quad \lim_{x \rightarrow 0} \frac{\sin x - x + x^3/6}{x^4}. \\ \text{(iii)} \quad \lim_{x \rightarrow 0} \frac{\cos x - 1 + x^2/2}{x^2}. \\ \text{(iv)} \quad \lim_{x \rightarrow 0} \frac{\cos x - 1 + x^2/2}{x^4}. \\ \text{(v)} \quad \lim_{x \rightarrow 0} \frac{\arctan x - x + x^3/3}{x^3}. \\ \text{(vi)} \quad \lim_{x \rightarrow 0} \left(\frac{1}{x} - \frac{1}{\sin x} \right). \end{aligned}$$

3. Let $f(x) = \begin{cases} \frac{\sin x}{x}, & x \neq 0 \\ 1, & x = 0. \end{cases}$

- (a) Find $f'(0)$.
- (b) Find $f''(0)$.

At this point, you will almost certainly have to use l'Hôpital's Rule, but in Chapter 24 we will be able to find $f^{(k)}(0)$ for all k , with almost no work at all.

4. Graph the following functions.

- (a) $f(x) = \sin 2x$.
- (b) $f(x) = \sin(x^2)$. (A pretty respectable sketch of this graph can be obtained using only a picture of the graph of \sin . Indeed, pure thought is your only hope in this problem, because determining the sign of the derivative $f'(x) = \cos(x^2) \cdot 2x$ is no easier than determining the behavior of f directly. The formula for $f'(x)$ does indicate one important fact, however— $f'(0) = 0$, which must be true since f is even, and which should be clear in your graph.)
- (c) $f(x) = \sin x + \sin 2x$. (It will probably be instructive to first draw the graphs of $g(x) = \sin x$ and $h(x) = \sin 2x$ carefully on the same set of axes, from 0 to 2π , and guess what the sum will look like. You can easily find out how many critical points f has on $[0, 2\pi]$ by considering the derivative of f . You can then determine the nature of these critical points by finding out the sign of f at each point; your sketch will probably suggest the answer.)
- (d) $f(x) = \tan x - x$. (First determine the behavior of f in $(-\pi/2, \pi/2)$; in the intervals $(k\pi - \pi/2, k\pi + \pi/2)$ the graph of f will look exactly the same, except moved up a certain amount. Why?)
- (e) $f(x) = \sin x - x$. (The material in the Appendix to Chapter 11 will be particularly helpful for this function.)

(f) $f(x) = \begin{cases} \frac{\sin x}{x}, & x \neq 0 \\ 1, & x = 0. \end{cases}$

(Part (d) should enable you to determine approximately where the zeros of f' are located. Notice that f is even and continuous at 0; also consider the size of f for large x .)

- (g) $f(x) = x \sin x$.

- *5. The *hyperbolic spiral* is the graph of the function $f(\theta) = a/\theta$ in polar coordinates (Chapter 4, Appendix 3). Sketch this curve, paying particular attention to its behavior for θ close to 0.
- 6. Prove the addition formula for \cos .

7. (a) From the addition formula for sin and cos derive formulas for $\sin 2x$, $\cos 2x$, $\sin 3x$, and $\cos 3x$.
 (b) Use these formulas to find the following values of the trigonometric functions (usually deduced by geometric arguments in elementary trigonometry):

$$\sin \frac{\pi}{4} = \cos \frac{\pi}{4} = \frac{\sqrt{2}}{2},$$

$$\tan \frac{\pi}{4} = 1,$$

$$\sin \frac{\pi}{6} = \frac{1}{2},$$

$$\cos \frac{\pi}{6} = \frac{\sqrt{3}}{2}.$$

8. (a) Show that $A \sin(x + B)$ can be written as $a \sin x + b \cos x$ for suitable a and b . (One of the theorems in this chapter provides a one-line proof. You should also be able to figure out what a and b are.)
 (b) Conversely, given a and b , find numbers A and B such that $a \sin x + b \cos x = A \sin(x + B)$ for all x .
 (c) Use part (b) to graph $f(x) = \sqrt{3} \sin x + \cos x$.

9. (a) Prove that

$$\tan(x + y) = \frac{\tan x + \tan y}{1 - \tan x \tan y}$$

provided that x , y , and $x + y$ are not of the form $k\pi + \pi/2$. (Use the addition formulas for sin and cos.)

- (b) Prove that

$$\arctan x + \arctan y = \arctan \left(\frac{x + y}{1 - xy} \right),$$

indicating any necessary restrictions on x and y . Hint: Replace x by $\arctan x$ and y by $\arctan y$ in part (a).

10. Prove that

$$\arcsin \alpha + \arcsin \beta = \arcsin(\alpha \sqrt{1 - \beta^2} + \beta \sqrt{1 - \alpha^2}),$$

indicating any restrictions on α and β .

- 11.

Prove that if m and n are any numbers, then

$$\sin mx \sin nx = \frac{1}{2}[\cos(m - n)x - \cos(m + n)x],$$

$$\sin mx \cos nx = \frac{1}{2}[\sin(m + n)x + \sin(m - n)x],$$

$$\cos mx \cos nx = \frac{1}{2}[\cos(m + n)x + \cos(m - n)x].$$

12. Prove that if m and n are natural numbers, then

$$\begin{aligned}\int_{-\pi}^{\pi} \sin mx \sin nx dx &= \begin{cases} 0, & m \neq n \\ \pi, & m = n, \end{cases} \\ \int_{-\pi}^{\pi} \cos mx \cos nx dx &= \begin{cases} 0, & m \neq n \\ \pi, & m = n, \end{cases} \\ \int_{-\pi}^{\pi} \sin mx \cos nx dx &= 0.\end{aligned}$$

These relations are particularly important in the theory of Fourier series. Although this topic will receive serious attention only in the Suggested Reading, the next problem provides a hint as to their importance.

13. (a) If f is integrable on $[-\pi, \pi]$, show that the minimum value of

$$\int_{-\pi}^{\pi} (f(x) - a \cos nx)^2 dx$$

occurs when

$$a = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx dx,$$

and the minimum value of

$$\int_{-\pi}^{\pi} (f(x) - a \sin nx)^2 dx$$

when

$$a = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx dx.$$

(In each case, bring a outside the integral sign, obtaining a quadratic expression in a .)

- (b) Define

$$\begin{aligned}a_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx dx, \quad n = 0, 1, 2, \dots, \\ b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx dx, \quad n = 1, 2, 3, \dots.\end{aligned}$$

Show that if c_i and d_i are any numbers, then

$$\begin{aligned}&\int_{-\pi}^{\pi} \left(f(x) - \left[\frac{c_0}{2} + \sum_{n=1}^N c_n \cos nx + d_n \sin nx \right] \right)^2 dx \\ &= \int_{-\pi}^{\pi} [f(x)]^2 dx - 2\pi \left(\frac{a_0 c_0}{2} + \sum_{n=1}^N a_n c_n + b_n d_n \right) + \pi \left(\frac{c_0^2}{2} + \sum_{n=1}^N c_n^2 + d_n^2 \right) \\ &= \int_{-\pi}^{\pi} [f(x)]^2 dx - \pi \left(\frac{a_0^2}{2} + \sum_{n=1}^N a_n^2 + b_n^2 \right) \\ &\quad + \pi \left(\left(\frac{c_0}{\sqrt{2}} - \frac{a_0}{\sqrt{2}} \right)^2 + \sum_{n=1}^N (c_n - a_n)^2 + (d_n - b_n)^2 \right),\end{aligned}$$

thus showing that the first integral is smallest when $a_i = c_i$ and $b_i = d_i$. In other words, among all “linear combinations” of the functions $s_n(x) = \sin nx$ and $c_n(x) = \cos nx$ for $1 \leq n \leq N$, the particular function

$$g(x) = \frac{a_0}{2} + \sum_{n=1}^N a_n \cos nx + b_n \sin nx$$

has the “closest fit” to f on $[-\pi, \pi]$.

14. (a) Find a formula for $\sin x + \sin y$. (Notice that this also gives a formula for $\sin x - \sin y$.) Hint: First find a formula for $\sin(a+b) + \sin(a-b)$. What good does that do?
 (b) Also find a formula for $\cos x + \cos y$ and $\cos x - \cos y$.

15. (a) Starting from the formula for $\cos 2x$, derive formulas for $\sin^2 x$ and $\cos^2 x$ in terms of $\cos 2x$.
 (b) Prove that

$$\cos \frac{x}{2} = \sqrt{\frac{1 + \cos x}{2}} \quad \text{and} \quad \sin \frac{x}{2} = \sqrt{\frac{1 - \cos x}{2}}$$

for $0 \leq x \leq \pi/2$.

- (c) Use part (a) to find $\int_a^b \sin^2 x \, dx$ and $\int_a^b \cos^2 x \, dx$.
 (d) Graph $f(x) = \sin^2 x$.

16. Find $\sin(\arctan x)$ and $\cos(\arctan x)$ as expressions not involving trigonometric functions. Hint: $y = \arctan x$ means that $x = \tan y = \sin y / \cos y = \sin y / \sqrt{1 - \sin^2 y}$.

17. If $x = \tan u/2$, express $\sin u$ and $\cos u$ in terms of x . (Use Problem 16; the answers should be very simple expressions.)

18. (a) Prove that $\sin(x + \pi/2) = \cos x$. (All along we have been drawing the graphs of sin and cos as if this were the case.)
 (b) What is $\arcsin(\cos x)$ and $\arccos(\sin x)$?

19. (a) Find $\int_0^1 \frac{1}{1+t^2} dt$. Hint: The answer is not 45.
 (b) Find $\int_0^\infty \frac{1}{1+t^2} dt$.

20. Find $\lim_{x \rightarrow \infty} x \sin \frac{1}{x}$.

21. (a) Define functions \sin° and \cos° by $\sin^\circ(x) = \sin(\pi x/180)$ and $\cos^\circ(x) = \cos(\pi x/180)$. Find $(\sin^\circ)'$ and $(\cos^\circ)'$ in terms of these same functions.
 (b) Find $\lim_{x \rightarrow 0} \frac{\sin^\circ x}{x}$ and $\lim_{x \rightarrow \infty} x \sin^\circ \frac{1}{x}$.

22. Prove that every point on the unit circle is of the form $(\cos \theta, \sin \theta)$ for at least one (and hence for infinitely many) numbers θ .

23. (a) Prove that π is the maximum possible length of an interval on which \sin is one-one, and that such an interval must be of the form $[2k\pi - \pi/2, 2k\pi + \pi/2]$ or $[2k\pi + \pi/2, 2(k+1)\pi - \pi/2]$.
 (b) Suppose we let $g(x) = \sin x$ for x in $(2k\pi - \pi/2, 2k\pi + \pi/2)$. What is $(g^{-1})'$?
24. Let $f(x) = \sec x$ for $0 \leq x \leq \pi$. Find the domain of f^{-1} and sketch its graph.
25. Prove that $|\sin x - \sin y| < |x - y|$ for all numbers $x \neq y$. Hint: The same statement, with $<$ replaced by \leq , is a very straightforward consequence of a well-known theorem; simple supplementary considerations then allow \leq to be improved to $<$.
- *26. It is an excellent test of intuition to predict the value of

$$\lim_{\lambda \rightarrow \infty} \int_a^b f(x) \sin \lambda x \, dx.$$

Continuous functions should be most accessible to intuition, but once you get the right idea for a proof the limit can easily be established for any integrable f .

- (a) Show that $\lim_{\lambda \rightarrow \infty} \int_c^d \sin \lambda x \, dx = 0$, by computing the integral explicitly.
 (b) Show that if s is a step function on $[a, b]$ (terminology from Problem 13-26), then $\lim_{\lambda \rightarrow \infty} \int_a^b s(x) \sin \lambda x \, dx = 0$.
 (c) Finally, use Problem 13-26 to show that $\lim_{\lambda \rightarrow \infty} \int_a^b f(x) \sin \lambda x \, dx = 0$ for any function f which is integrable on $[a, b]$. This result, like Problem 12, plays an important role in the theory of Fourier series; it is known as the Riemann-Lebesgue Lemma.

27. This problem outlines the classical approach to the trigonometric functions. The shaded sector in Figure 24 has area $x/2$.
 (a) By considering the triangles OAB and OCB prove that if $0 < x < \pi/4$, then

$$\frac{\sin x}{2} < \frac{x}{2} < \frac{\sin x}{2 \cos x}.$$

- (b) Conclude that

$$\cos x < \frac{\sin x}{x} < 1,$$

and prove that

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1.$$

- (c) Use this limit to find

$$\lim_{x \rightarrow 0} \frac{1 - \cos x}{x}.$$

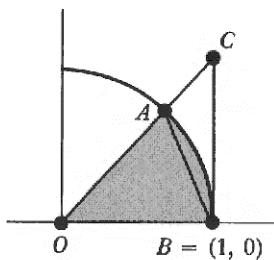


FIGURE 24

- (d) Using parts (b) and (c), and the addition formula for sin, find $\sin'(x)$, starting from the definition of the derivative.
- *28. This problem gives a treatment of the trigonometric functions in terms of length, and uses Problem 13-25. Let $f(x) = \sqrt{1-x^2}$ for $-1 \leq x \leq 1$. Define $\mathcal{L}(x)$ to be the length of f on $[x, 1]$.
- Show that
$$\mathcal{L}(x) = \int_x^1 \frac{1}{\sqrt{1-t^2}} dt.$$

(This is actually an improper integral, as defined in Problem 14-28.)
 - Show that
$$\mathcal{L}'(x) = -\frac{1}{\sqrt{1-x^2}} \quad \text{for } -1 < x < 1.$$
 - Define π as $\mathcal{L}(-1)$. For $0 \leq x \leq \pi$, define $\cos x$ by $\mathcal{L}(\cos x) = x$, and define $\sin x = \sqrt{1-\cos^2 x}$. Prove that $\cos'(x) = -\sin x$ and $\sin'(x) = \cos x$ for $0 < x < \pi$.
- *29. Yet another development of the trigonometric functions was briefly mentioned in the text—starting with inverse functions defined by integrals. It is convenient to begin with \arctan , since this function is defined for all x . To do this problem, pretend that you have never heard of the trigonometric functions.
- Let $\alpha(x) = \int_0^x (1+t^2)^{-1} dt$. Prove that α is odd and increasing, and that $\lim_{x \rightarrow \infty} \alpha(x)$ and $\lim_{x \rightarrow -\infty} \alpha(x)$ both exist, and are negatives of each other. If we define $\pi = 2 \lim_{x \rightarrow \infty} \alpha(x)$, then α^{-1} is defined on $(-\pi/2, \pi/2)$.
 - Show that $(\alpha^{-1})'(x) = 1 + [\alpha^{-1}(x)]^2$.
 - For $x = k\pi + x'$ with $x' \neq \pi/2$ or $-\pi/2$, define $\tan x = \alpha^{-1}(x')$. Then define $\cos x = 1/\sqrt{1+\tan^2 x}$, for x not of the form $k\pi + \pi/2$ or $k\pi - \pi/2$, and $\cos(k\pi \pm \pi/2) = 0$. Prove first that $\cos'(x) = -\tan x \cos x$, and then that $\cos''(x) = -\cos x$ for all x .
- *30. If we are willing to assume that certain differential equations have solutions, another approach to the trigonometric functions is possible. Suppose, in particular, that there is some function y_0 which is not always 0 and which satisfies $y_0'' + y_0 = 0$.
- Prove that $y_0^2 + (y_0')^2$ is constant, and conclude that either $y_0(0) \neq 0$ or $y_0'(0) \neq 0$.
 - Prove that there is a function s satisfying $s'' + s = 0$ and $s(0) = 0$ and $s'(0) = 1$. Hint: Try s of the form $ay_0 + by_0'$.

If we define $\sin = s$ and $\cos = s'$, then almost all facts about trigonometric functions become trivial. There is one point which requires work,

however—producing the number π . This is most easily done using an exercise from the Appendix to Chapter 11:

- (c) Use Problem 7 of the Appendix to Chapter 11 to prove that $\cos x$ cannot be positive for all $x > 0$. It follows that there is a smallest $x_0 > 0$ with $\cos x_0 = 0$, and we can define $\pi = 2x_0$.
 - (d) Prove that $\sin \pi/2 = 1$. (Since $\sin^2 + \cos^2 = 1$, we have $\sin \pi/2 = \pm 1$; the problem is to decide why $\sin \pi/2$ is positive.)
 - (e) Find $\cos \pi$, $\sin \pi$, $\cos 2\pi$, and $\sin 2\pi$. (Naturally you may use any addition formulas, since these can be derived once we know that $\sin' = \cos$ and $\cos' = -\sin$.)
 - (f) Prove that \cos and \sin are periodic with period 2π .
31. (a) After all the work involved in the definition of \sin , it would be disconcerting to find that \sin is actually a rational function. Prove that it isn't. (There is a simple property of \sin which a rational function cannot possibly have.)
- (b) Prove that \sin isn't even defined implicitly by an algebraic equation; that is, there do not exist rational functions f_0, \dots, f_{n-1} such that

$$(\sin x)^n + f_{n-1}(x)(\sin x)^{n-1} + \cdots + f_0(x) = 0 \quad \text{for all } x.$$

Hint: Prove that $f_0 = 0$, so that $\sin x$ can be factored out. The remaining factor is 0 except perhaps at multiples of 2π . But this implies that it is 0 for all x . (Why?) You are now set up for a proof by induction.

- *32. Suppose that ϕ_1 and ϕ_2 satisfy

$$\begin{aligned}\phi_1'' + g_1\phi_1 &= 0, \\ \phi_2'' + g_2\phi_2 &= 0,\end{aligned}$$

and that $g_2 > g_1$.

- (a) Show that

$$\phi_1''\phi_2 - \phi_2''\phi_1 - (g_2 - g_1)\phi_1\phi_2 = 0.$$

- (b) Show that if $\phi_1(x) > 0$ and $\phi_2(x) > 0$ for all x in (a, b) , then

$$\int_a^b [\phi_1''\phi_2 - \phi_2''\phi_1] > 0,$$

and conclude that

$$[\phi_1'(b)\phi_2(b) - \phi_1'(a)\phi_2(a)] + [\phi_1(b)\phi_2'(b) - \phi_1(a)\phi_2'(a)] > 0.$$

- (c) Show that in this case we cannot have $\phi_1(a) = \phi_1(b) = 0$. Hint: Consider the sign of $\phi_1'(a)$ and $\phi_1'(b)$.
- (d) Show that the equations $\phi_1(a) = \phi_1(b) = 0$ are also impossible if $\phi_1 > 0$, $\phi_2 < 0$ or $\phi_1 < 0$, $\phi_2 > 0$, or $\phi_1 < 0$, $\phi_2 < 0$ on (a, b) . (You should be able to do this with almost no extra work.)

The net result of this problem may be stated as follows: if a and b are consecutive zeros of ϕ_1 , then ϕ_2 must have a zero somewhere between a and b . This result, in a slightly more general form, is known as the Sturm Comparison Theorem. As a particular example, any solution of the differential equation

$$y'' + (x + 1)y = 0$$

must have zeros on the positive horizontal axis which are within π of each other.

33. (a) Using the formula for $\sin x - \sin y$ derived in Problem 14, show that

$$\sin(k + \frac{1}{2})x - \sin(k - \frac{1}{2})x = 2 \sin \frac{x}{2} \cos kx.$$

- (b) Conclude that

$$\frac{1}{2} + \cos x + \cos 2x + \cdots + \cos nx = \frac{\sin(n + \frac{1}{2})x}{2 \sin \frac{x}{2}}.$$

Like two other results in this problem set, this equation is very important in the study of Fourier series, and we also make use of it in Problems 19-42 and 23-19.

- (c) Similarly, derive the formula

$$\sin x + \sin 2x + \cdots + \sin nx = \frac{\sin\left(\frac{n+1}{2}x\right) \sin\left(\frac{n}{2}x\right)}{\sin \frac{x}{2}}.$$

(A more natural derivation of these formulas will be given in Problem 27-14.)

- (d) Use parts (b) and (c) to find $\int_0^b \sin x \, dx$ and $\int_0^b \cos x \, dx$ directly from the definition of the integral.

16

*CHAPTER π IS IRRATIONAL

This short chapter, diverging from the main stream of the book, is included to demonstrate that we are already in a position to do some sophisticated mathematics. This entire chapter is devoted to an elementary proof that π is irrational. Like many “elementary” proofs of deep theorems, the motivation for many steps in our proof cannot be supplied; nevertheless, it is still quite possible to follow the proof step-by-step.

Two observations must be made before the proof. The first concerns the function

$$f_n(x) = \frac{x^n(1-x)^n}{n!},$$

which clearly satisfies

$$0 < f_n(x) < \frac{1}{n!} \quad \text{for } 0 < x < 1.$$

An important property of the function f_n is revealed by considering the expression obtained by actually multiplying out $x^n(1-x)^n$. The lowest power of x appearing will be n and the highest power will be $2n$. Thus f_n can be written in the form

$$f_n(x) = \frac{1}{n!} \sum_{i=n}^{2n} c_i x^i,$$

where the numbers c_i are integers. It is clear from this expression that

$$f_n^{(k)}(0) = 0 \quad \text{if } k < n \text{ or } k > 2n.$$

Moreover,

$$\begin{aligned} f_n^{(n)}(x) &= \frac{1}{n!} [n! c_n + \text{terms involving } x] \\ f_n^{(n+1)}(x) &= \frac{1}{n!} [(n+1)! c_{n+1} + \text{terms involving } x] \\ &\vdots \\ &\vdots \\ f_n^{(2n)}(x) &= \frac{1}{n!} [(2n)! c_{2n}]. \end{aligned}$$

This means that

$$\begin{aligned} f_n^{(n)}(0) &= c_n, \\ f_n^{(n+1)}(0) &= (n+1)c_{n+1} \\ &\quad \vdots \\ f_n^{(2n)}(0) &= (2n)(2n-1) \cdots (n+1)c_{2n}, \end{aligned}$$

where the numbers on the right are all integers. Thus

$$f_n^{(k)}(0) \text{ is an integer for all } k.$$

The relation

$$f_n(x) = f_n(1-x)$$

implies that

$$f_n^{(k)}(x) = (-1)^k f_n^{(k)}(1-x);$$

therefore,

$$f_n^{(k)}(1) \text{ is also an integer for all } k.$$

The proof that π is irrational requires one further observation: if a is any number, and $\varepsilon > 0$, then for sufficiently large n we will have

$$\frac{a^n}{n!} < \varepsilon.$$

To prove this, notice that if $n \geq 2a$, then

$$\frac{a^{n+1}}{(n+1)!} = \frac{a}{n+1} \cdot \frac{a^n}{n!} < \frac{1}{2} \cdot \frac{a^n}{n!}.$$

Now let n_0 be any natural number with $n_0 \geq 2a$. Then, whatever value

$$\frac{a^{n_0}}{(n_0)!}$$

may have, the succeeding values satisfy

$$\begin{aligned} \frac{a^{(n_0+1)}}{(n_0+1)!} &< \frac{1}{2} \cdot \frac{a^{n_0}}{(n_0)!} \\ \frac{a^{(n_0+2)}}{(n_0+2)!} &< \frac{1}{2} \cdot \frac{a^{(n_0+1)}}{(n_0+1)!} < \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{a^{n_0}}{(n_0)!} \\ &\quad \vdots \\ \frac{a^{(n_0+k)}}{(n_0+k)!} &< \frac{1}{2^k} \cdot \frac{a^{n_0}}{(n_0)!}. \end{aligned}$$

If k is so large that $\frac{a^{n_0}}{(n_0)! \varepsilon} < 2^k$, then

$$\frac{a^{(n_0+k)}}{(n_0+k)!} < \varepsilon,$$

which is the desired result. Having made these observations, we are ready for the one theorem in this chapter.

THEOREM 1 The number π is irrational; in fact, π^2 is irrational. (Notice that the irrationality of π^2 implies the irrationality of π , for if π were rational, then π^2 certainly would be.)

PROOF Suppose π^2 were rational, so that

$$\pi^2 = \frac{a}{b}$$

for some positive integers a and b . Let

$$(1) \quad G(x) = b^n [\pi^{2n} f_n(x) - \pi^{2n-2} f_n''(x) + \pi^{2n-4} f_n^{(4)}(x) - \cdots + (-1)^n f_n^{(2n)}(x)].$$

Notice that each of the factors

$$b^n \pi^{2n-2k} = b^n (\pi^2)^{n-k} = b^n \left(\frac{a}{b}\right)^{n-k} = a^{n-k} b^k$$

is an integer. Since $f_n^{(k)}(0)$ and $f_n^{(k)}(1)$ are integers, this shows that

$G(0)$ and $G(1)$ are integers.

Differentiating G twice yields

$$(2) \quad G''(x) = b^n [\pi^{2n} f_n''(x) - \pi^{2n-2} f_n^{(4)}(x) + \cdots + (-1)^n f_n^{(2n+2)}(x)].$$

The last term, $(-1)^n f_n^{(2n+2)}(x)$, is zero. Thus, adding (1) and (2) gives

$$(3) \quad G''(x) + \pi^2 G(x) = b^n \pi^{2n+2} f_n(x) = \pi^2 a^n f_n(x).$$

Now let

$$H(x) = G'(x) \sin \pi x - \pi G(x) \cos \pi x.$$

Then

$$\begin{aligned} H'(x) &= \pi G'(x) \cos \pi x + G''(x) \sin \pi x - \pi G'(x) \cos \pi x + \pi^2 G(x) \sin \pi x \\ &= [G''(x) + \pi^2 G(x)] \sin \pi x \\ &= \pi^2 a^n f_n(x) \sin \pi x, \text{ by (3).} \end{aligned}$$

By the Second Fundamental Theorem of Calculus,

$$\begin{aligned} \pi^2 \int_0^1 a^n f_n(x) \sin \pi x \, dx &= H(1) - H(0) \\ &= G'(1) \sin \pi - \pi G(1) \cos \pi - G'(0) \sin 0 + \pi G(0) \cos 0 \\ &= \pi [G(1) + G(0)]. \end{aligned}$$

Thus

$$\pi \int_0^1 a^n f_n(x) \sin \pi x \, dx \text{ is an integer.}$$

On the other hand, $0 < f_n(x) < 1/n!$ for $0 < x < 1$, so

$$0 < \pi a^n f_n(x) \sin \pi x < \frac{\pi a^n}{n!} \quad \text{for } 0 < x < 1.$$

Consequently,

$$0 < \pi \int_0^1 a^n f_n(x) \sin \pi x \, dx < \frac{\pi a^n}{n!}.$$

This reasoning was completely independent of the value of n . Now if n is large enough, then

$$0 < \pi \int_0^1 a^n f_n(x) \sin \pi x \, dx < \frac{\pi a^n}{n!} < 1.$$

But this is absurd, because the integral is an integer, and there is no integer between 0 and 1. Thus our original assumption must have been incorrect: π^2 is irrational. ■

This proof is admittedly mysterious; perhaps most mysterious of all is the way that π enters the proof—it almost looks as if we have proved π irrational without ever mentioning a definition of π . A close reexamination of the proof will show that precisely one property of π is essential—

$$\sin(\pi) = 0.$$

The proof really depends on the properties of the function \sin , and proves the irrationality of the smallest positive number x with $\sin x = 0$. In fact, very few properties of \sin are required, namely,

$$\begin{aligned}\sin' &= \cos, \\ \cos' &= -\sin, \\ \sin(0) &= 0, \\ \cos(0) &= 1.\end{aligned}$$

Even this list could be shortened; as far as the proof is concerned, \cos might just as well be defined as \sin' . The properties of \sin required in the proof may then be written

$$\begin{aligned}\sin'' + \sin &= 0, \\ \sin(0) &= 0, \\ \sin'(0) &= 1.\end{aligned}$$

Of course, this is not really very surprising at all, since, as we have seen in the previous chapter, these properties characterize the function \sin completely.

PROBLEMS

1. (a) Prove that the areas of triangles OAB and OAC in Figure 1 are related by the equation

$$\text{area } OAC = \frac{1}{2} \sqrt{\frac{1 - \sqrt{1 - 16(\text{area } OAB)^2}}{2}}.$$

Hint: Solve the equations $xy = 2(\text{area } OAB)$, $x^2 + y^2 = 1$, for y .

- (b) Let P_m be the regular polygon of m sides inscribed in the unit circle. If A_m is the area of P_m show that

$$A_{2m} = \frac{m}{2} \sqrt{2 - 2\sqrt{1 - (2A_m/m)^2}}.$$

This result allows one to obtain (more and more complicated) expressions for A_{2^n} , starting with $A_4 = 2$, and thus to compute π as accurately as desired (according to Problem 8-11). Although better methods will appear in Chapter 20, a slight variant of this approach yields a very interesting expression for π :

2. (a) Using the fact that

$$\frac{\text{area}(OAB)}{\text{area}(OAC)} = OB,$$

show that if α_m is the distance from O to one side of P_m , then

$$\frac{A_m}{A_{2m}} = \alpha_m.$$

- (b) Show that

$$\frac{2}{A_{2^k}} = \alpha_4 \cdot \alpha_8 \cdot \dots \cdot \alpha_{2^{k-1}}.$$

- (c) Using the fact that

$$\alpha_m = \cos \frac{\pi}{m},$$

and the formula $\cos x/2 = \sqrt{\frac{1 + \cos x}{2}}$ (Problem 15-15), prove that

$$\alpha_4 = \sqrt{\frac{1}{2}}$$

$$\alpha_8 = \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}},$$

$$\alpha_{16} = \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}}},$$

etc.

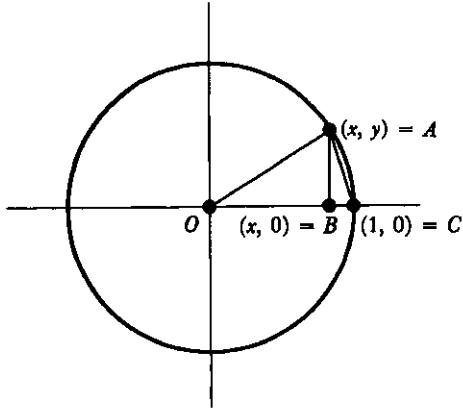


FIGURE 1

Together with part (b), this shows that $2/\pi$ can be written as an “infinite product”

$$\frac{2}{\pi} = \sqrt{\frac{1}{2}} \cdot \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}} \cdot \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}}} \cdots ;$$

to be precise, this equation means that the product of the first n factors can be made as close to $2/\pi$ as desired, by choosing n sufficiently large. This product was discovered by François Viète in 1579, and is only one of many fascinating expressions for π , some of which are mentioned later.

17

*CHAPTER PLANETARY MOTION

Nature and Nature's Laws lay hid in night
God said "Let Newton be," and all was light.

Alexander Pope

Unlike Chapter 16, a short chapter diverging from the main stream of the book, this long chapter diverges from the main stream of the book to demonstrate that we are already in a position to do some real physics.

In 1609 Kepler published his first two laws of planetary motion. The first law describes the shape of planetary orbits:

The planets move in ellipses, with the sun at one focus.

The second law involves the area swept out by the segment from the sun to the planet (the 'radius vector from the sun to the planet') in various time intervals (Figure 1):

Equal areas are swept out by the radius vector in equal times. (Equivalently, the area swept out in time t is proportional to t .)

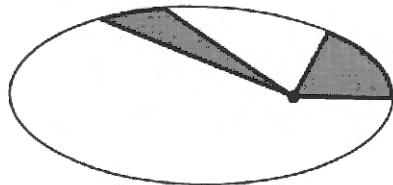


FIGURE 1

Kepler's third law, published in 1619, relates the motions of different planets. If a is the major axis of a planet's elliptical orbit and T is its period, the time it takes the planet to return to a given position, then:

The ratio a^3/T^2 is the same for all planets.

Newton's great accomplishment was to show (using his general law that the force on a body is its mass times its acceleration) that Kepler's laws follow from the assumption that the planets are attracted to the sun by a force (the gravitational force of the sun) always directed toward the sun, proportional to the mass of the planet, and satisfying an inverse square law; that is, by a force directed toward the sun whose magnitude varies inversely with the square of the distance from the sun to the planet and directly with the mass of the planet. Since force is mass times acceleration, this is equivalent simply to saying that the magnitude of the acceleration is a constant divided by the square of the distance from the sun.

Newton's analysis actually established three results that correlate with Kepler's individual laws. The first of Newton's results concerns Kepler's second law (which was actually discovered first, nicely preserving the symmetry of the situation):

Kepler's second law is true precisely for 'central forces', i.e., if and only the force between the sun and the planet always lies along the line between the sun and the planet.

Although Newton is revered as the discoverer of calculus, and indeed invented calculus precisely in order to treat such problems, his derivation hardly seems to use calculus at all. Instead of considering a force that varies continuously as the planet moves, Newton first considers short equal time intervals and assumes that a momentary force is exerted at the ends of each of these intervals.

To be specific, let us imagine that during the first time interval the planet moves along the line P_1P_2 , with uniform velocity (Figure 2a). If, during the next equal time interval, the planet continued to move along this line, it would end up at P_3 , where the length of P_1P_2 equals the length of P_2P_3 . This would imply that the triangle SP_1P_2 has the same area as the triangle SP_2P_3 (since they have equal bases, and the same height)—this just says that Kepler's law holds in the special case where the force is 0.

Now suppose (Figure 2b) that at the moment the planet arrives at P_2 it experiences a force exerted *along the line from S to P_2* , which by itself would cause the planet to move to the point Q . Combined with the motion that the planet already has, this causes the planet to move to R , the vertex opposite P_2 in the parallelogram whose sides are P_2P_3 and P_2Q .

Thus, the area swept out in the second time interval is actually the triangle SP_2R . But the area of triangle SP_2R is equal to the area of triangle SP_3P_2 , since they have the same base SP_2 , and the same heights (since RP_3 is parallel to SP_2). Hence, finally, the area of triangle SP_2R is the same as the area of the original triangle SP_1P_2 ! Conversely, if the triangle SP_2R has the same area as SP_1P_2 , and hence the same area as SP_3P_2 , then RP_3 must be parallel to SP_2 , and this implies that Q must lie along SP_2 .

Of course, this isn't quite the sort of argument one would expect to find in a modern book, but in its own charming way it shows physically just *why* the result should be true.

To analyze planetary motion we will be using the material in the Appendix to Chapter 12, and the "determinant" det defined in Problem 4 of Appendix 1 to Chapter 4. We describe the motion of the planet by the parameterized curve

$$c(t) = r(t)(\cos \theta(t), \sin \theta(t)),$$

so that r always gives the length of the line from the sun to the planet, while θ gives the angle. It will be convenient to write this also as

$$(1) \quad c(t) = r(t) \cdot \mathbf{e}(\theta(t)),$$

where

$$\mathbf{e}(t) = (\cos t, \sin t)$$

is just the parameterized curve that runs along the unit circle. Note that

$$\mathbf{e}'(t) = (-\sin t, \cos t)$$

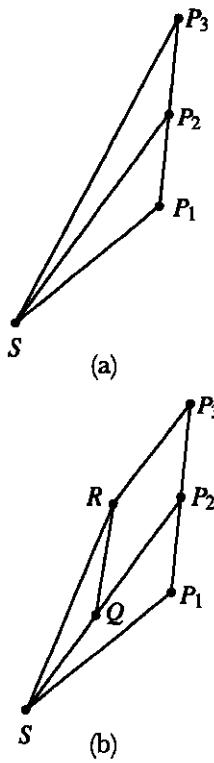


FIGURE 2

is also a vector of unit length, but perpendicular to $\mathbf{e}(t)$, and that we also have

$$(2) \quad \det(\mathbf{e}(t), \mathbf{e}'(t)) = 1.$$

Differentiating (1), using the formulas on page 244, we obtain

$$(3) \quad \mathbf{c}'(t) = \mathbf{r}'(t) \cdot \mathbf{e}(\theta(t)) + \mathbf{r}(t)\theta'(t) \cdot \mathbf{e}'(\theta(t)),$$

and combining with (1), together with the formulas in Problem 6 of Appendix 1 to Chapter 4, we get

$$\begin{aligned} \det(\mathbf{c}(t), \mathbf{c}'(t)) &= \mathbf{r}(t)\mathbf{r}'(t) \det(\mathbf{e}(\theta(t)), \mathbf{e}(\theta(t))) + \mathbf{r}(t)^2\theta'(t) \det(\mathbf{e}(\theta(t)), \mathbf{e}'(\theta(t))) \\ &= \mathbf{r}(t)^2\theta'(t) \det(\mathbf{e}(\theta(t)), \mathbf{e}'(\theta(t))), \end{aligned}$$

since $\det(v, v)$ is always 0. Using (2) we then get

$$(4) \quad \det(\mathbf{c}, \mathbf{c}') = \mathbf{r}^2\theta'.$$

As we will see, $\mathbf{r}^2\theta'$ turns out to have another important interpretation.

Suppose that $A(t)$ is the area swept out from time 0 to t (Figure 3). We want to get a formula for $A'(t)$, and, in the spirit of Newton, we'll begin by making an educated guess. Figure 4 shows $A(t+h) - A(t)$, together with a straight line segment between $c(t)$ and $c(t+h)$. It is easy to write down a formula for the area of the triangle $\Delta(h)$ with vertices O , $c(t)$, and $c(t+h)$: according to Problems 4 and 5 of Appendix 1 to Chapter 4, the area is

$$\text{area}(\Delta(h)) = \frac{1}{2} \det(\mathbf{c}(t), \mathbf{c}(t+h) - \mathbf{c}(t)).$$

Since the triangle $\Delta(h)$ has practically the same area as the region $A(t+h) - A(t)$, this shows (or practically shows) that

$$\begin{aligned} A'(t) &= \lim_{h \rightarrow 0} \frac{A(t+h) - A(t)}{h} \\ &= \lim_{h \rightarrow 0} \frac{\text{area } \Delta(h)}{h} \\ &= \frac{1}{2} \det \left(\mathbf{c}(t), \lim_{h \rightarrow 0} \frac{\mathbf{c}(t+h) - \mathbf{c}(t)}{h} \right) \\ &= \frac{1}{2} \det(\mathbf{c}(t), \mathbf{c}'(t)). \end{aligned}$$

A rigorous derivation, establishing more in the process, can be made using Problem 13-24, which gives a formula for the area of a region determined by the graph of a function in polar coordinates. According to this Problem, we can write

$$(*) \quad A(t) = \frac{1}{2} \int_0^{\theta(t)} \rho(\phi)^2 d\phi$$

if our parameterized curve $\mathbf{c}(t) = \mathbf{r}(t) \cdot \mathbf{e}(\theta(t))$ is the graph of the function ρ in polar coordinates (here we've used ϕ for the angular polar coordinate, to avoid confusion with the function θ used to describe the curve c).

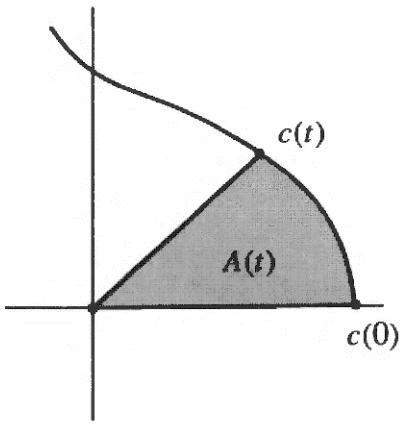


FIGURE 3

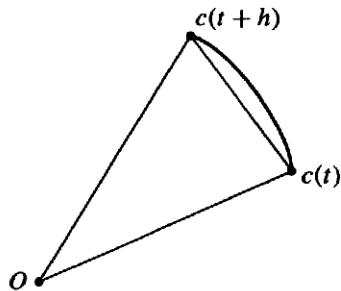


FIGURE 4

Now the function ρ is just

$$\rho = r \circ \theta^{-1}$$

[for any particular angle ϕ , $\theta^{-1}(\phi)$ is the time at which the curve c has angular polar coordinate ϕ , so $r(\theta^{-1}(t))$ is the radius coordinate corresponding to ϕ]. Although the presence of the inverse function might look a bit forbidding, it's actually quite innocent: Applying the First Fundamental Theorem of Calculus and the Chain Rule to (*) we immediately get

$$\begin{aligned} A'(t) &= \frac{1}{2}\rho(\theta(t))^2 \cdot \theta'(t) \\ &= \frac{1}{2}r(t)^2\theta'(t), \quad \text{since } \rho = r \circ \theta^{-1}. \end{aligned}$$

Briefly,

$$A' = \frac{1}{2}r^2\theta'.$$

Combining with (4), we thus have

(5)

$$A' = \frac{1}{2}\det(c, c') = \frac{1}{2}r^2\theta'.$$

Now we're ready to consider Kepler's second law. Notice that *Kepler's second law is equivalent to saying that A' is constant*, and thus it is equivalent to $A'' = 0$. But

$$\begin{aligned} A'' &= \frac{1}{2}[\det(c, c')]' = \frac{1}{2}\det(c', c') + \frac{1}{2}\det(c, c'') \\ &= \frac{1}{2}\det(c, c''). \end{aligned} \quad (\text{see page 245})$$

So

Kepler's second law is equivalent to $\det(c, c'') = 0$.

Putting this all together we have:

THEOREM 1 Kepler's second law is true if and only if the force is central, and in this case each planetary path $c(t) = r(t) \cdot \mathbf{e}(\theta(t))$ satisfies the equation

$$(K_2) \quad r^2\theta' = \det(c, c') = \text{constant.}$$

PROOF Saying that the force is central just means that it always points along $c(t)$. Since $c''(t)$ is in the direction of the force, that is equivalent to saying that $c''(t)$ always points along $c(t)$. And this is equivalent to saying that we always have

$$\det(c, c'') = 0.$$

We've just seen that this is equivalent to Kepler's second law.

Moreover, this equation implies that $[\det(c, c')]' = 0$, which by (5) gives (K_2) . ■

Newton next showed that if the gravitational force of the sun is a central force and also satisfies an inverse square law, then the path of any object in it will be a conic section having the sun at one focus. Planets, of course, correspond to the case where the conic section is an ellipse, and this is also true for comets that visit the sun periodically; parabolas and hyperbolas represent objects that come from outside the solar system, and eventually continue on their merry way back outside the system.

THEOREM 2

If the gravitational force of the sun is a central force that satisfies an inverse square law, then the path of any body in it will be a conic section having the sun at one focus.

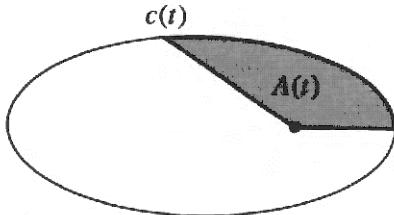
PROOF

FIGURE 5

Notice that our conclusion specifies the shape of the path, not a particular parameterization. But this parameterization is essentially determined by Theorem 1: the hypothesis of a central force implies that the area $A(t)$ (Figure 5) is proportional to t , so determining $c(t)$ is essentially equivalent to determining A for arbitrary points on the ellipse. Unfortunately, the areas of such segments cannot be determined explicitly.¹ This means that we have to determine the *shape* of the path $c = r(t) \cdot \mathbf{e}(\theta(t))$ without finding its parameterization! Since it is the function $r \circ \theta^{-1}$ which actually describes the shape of the path in polar coordinates, we shouldn't be surprised to find θ^{-1} entering into the proof.

By Theorem 1, the hypothesis of a central force implies that

$$(K_2) \quad r^2 \theta' = \det(c, c') = M$$

for some constant M . The hypothesis of an inverse square law can be written

$$(*) \quad c''(t) = -\frac{H}{r(t)^2} \mathbf{e}(\theta(t))$$

for some constant H . Using (K_2) , this can be written

$$\frac{c''(t)}{\theta'(t)} = -\frac{H}{M} \mathbf{e}(\theta(t)).$$

Notice that the left-hand side of this equation is

$$[c' \circ \theta^{-1}]'(\theta(t)).$$

So if we let

$$D = c' \circ \theta^{-1}$$

(this is the main trick—"we consider c' as a function of θ "), then the equation can be written as

$$D'(\theta(t)) = -\frac{H}{M} \mathbf{e}(\theta(t)) = -\frac{H}{M} (\cos \theta(t), \sin \theta(t)),$$

¹ More precisely, we can't write down a solution in terms of familiar "standard functions," like \sin , \arcsin , etc.

and we can write this simply as

$$D'(u) = -\frac{H}{M}(\cos u, \sin u) = \left(-\frac{H}{M} \cos u, -\frac{H}{M} \sin u\right)$$

[for all u of the form $\theta(t)$ for some t , which happens to be all u], completely eliminating θ .

The equation that we have just obtained is simply a pair of equations, for the components of D , each of which we can easily solve individually; we thus find that

$$D(u) = \left(\frac{H \cdot \sin u}{-M} + A, \frac{H \cdot \cos u}{M} + B\right)$$

for two constants A and B . Letting $u = \theta(t)$ again we thus have an explicit formula for c' :

$$c' = \left(\frac{H \cdot \sin \theta}{-M} + A, \frac{H \cdot \cos \theta}{M} + B\right).$$

[Here $\sin \theta$ really stands for $\sin \circ \theta$, etc., abbreviations that we will use throughout.]

Although we can't get an explicit formula for c itself, if we substitute this equation, together with $c = r(\cos \theta, \sin \theta)$, into the equation

$$\det(c, c') = M \quad (\text{equation } (K_2)),$$

we get

$$r \left[\frac{H}{M} \cos^2 \theta + B \cos \theta + \frac{H}{M} \sin^2 \theta - A \sin \theta \right] = M,$$

which simplifies to

$$r \left[\frac{H}{M^2} + \frac{B}{M} \cos \theta - \frac{A}{M} \sin \theta \right] = 1.$$

Problem 15-8 shows that this can be written in the form

$$r(t) \left[\frac{H}{M^2} + C \cos(\theta(t) + D) \right] = 1,$$

for some constants C and D . We can let $D = 0$, since this simply amounts to rotating our polar coordinate system (choosing which ray corresponds to $\theta = 0$), so we can write, finally,

$$r[1 + \varepsilon \cos \theta] = \frac{M^2}{H} = \Lambda.$$

But this is the formula for a conic section derived in Appendix 3 of Chapter 4. ■

In terms of the constant M in the equation

$$r^2 \theta' = M$$

and the constant Λ in the equation of the orbit

$$r[1 + \varepsilon \cos \theta] = \Lambda$$

the last equation in our proof shows that we can rewrite (*) as

$$(**) \quad c''(t) = -\frac{M^2}{\Lambda} \cdot \frac{1}{r^2} \mathbf{e}(\theta(t)).$$

Recall (page 87) that the major axis a of the ellipse is given by

$$(a) \quad a = \frac{\Lambda}{1 - \varepsilon^2},$$

while the minor axis b is given by

$$(b) \quad b = \frac{\Lambda}{\sqrt{1 - \varepsilon^2}}.$$

Consequently,

$$(c) \quad \frac{b^2}{\Lambda} = a.$$

Remember that equation (5) gives

$$A'(t) = \frac{1}{2} r^2 \theta' = \frac{1}{2} M,$$

and thus

$$A(t) = \frac{1}{2} Mt.$$

We can therefore interpret M in terms of the period T of the orbit. This period T is, by definition, the value of t for which we have $\theta(t) = 2\pi$, so that we obtain the complete ellipse. Hence

$$\text{area of the ellipse} = A(T) = \frac{1}{2} MT,$$

or

$$M = \frac{2(\text{area of the ellipse})}{T} = \frac{2\pi ab}{T} \quad \text{by Problem 13-17.}$$

Hence the constant M^2/Λ in (**) is

$$\begin{aligned} \frac{M^2}{\Lambda} &= \frac{4\pi^2 a^2 b^2}{T^2 \Lambda} \\ &= \frac{4\pi^2 a^3}{T^2}, \quad \text{using (c).} \end{aligned}$$

This completes the final step of Newton's analysis:

THEOREM 3 Kepler's third law is true if and only if the acceleration $c''(t)$ of any planet, moving on an ellipse, satisfies

$$c''(t) = -G \cdot \frac{1}{r^2} \mathbf{e}(\theta(t))$$

for a constant G that does not depend on the planet.

It should be mentioned that the converse of Theorem 2 is also true. To prove this, we first want to establish one further consequence of Kepler's second law. Recall that for

$$\mathbf{e}(t) = (\cos t, \sin t)$$

we have

$$\mathbf{e}'(t) = (-\sin t, \cos t).$$

Consequently,

$$\mathbf{e}''(t) = (-\cos t, -\sin t) = -\mathbf{e}(t).$$

Now differentiating (3) gives

$$\begin{aligned} c''(t) &= r''(t) \cdot \mathbf{e}(\theta(t)) + r'(t)\theta'(t) \cdot \mathbf{e}'(\theta(t)) \\ &\quad + r'(t)\theta'(t) \cdot \mathbf{e}'(\theta(t)) + r(t)\theta''(t) \cdot \mathbf{e}'(\theta(t)) + r(t)\theta'(t)\theta'(t) \cdot \mathbf{e}''(\theta(t)). \end{aligned}$$

Using $\mathbf{e}''(t) = -\mathbf{e}(t)$ we get

$$c''(t) = [r''(t) - r(t)\theta'(t)^2] \cdot \mathbf{e}(\theta(t)) + [2r'(t)\theta'(t) + r(t)\theta''(t)] \cdot \mathbf{e}'(\theta(t)).$$

Since Kepler's second law implies central forces, hence that $c''(t)$ is always a multiple of $c(t)$, and thus always a multiple of $\mathbf{e}(\theta(t))$, the coefficient of $\mathbf{e}'(\theta(t))$ must be 0 [as a matter of fact, we can see this directly by taking the derivative of formula (K_2)]. Thus Kepler's second law implies that

$$(6) \quad c''(t) = [r''(t) - r(t)\theta'(t)^2] \cdot \mathbf{e}(\theta(t)).$$

THEOREM 4 If the path of a planet moving under a central gravitational force is an ellipse with the sun as focus, then the force must satisfies an inverse square law.

PROOF

As in Theorem 2, notice that the hypothesis on the shape of the path, together with the hypothesis of a central force, which is equivalent to Kepler's second law, essentially determines the parameterization. But we can't write down an explicit solution, so we have to obtain information about the acceleration without actually knowing what it is.

Once again, the hypothesis of a central force implies that

$$(K_2) \quad r^2\theta' = M,$$

for some constant M , and the hypothesis that the path is an ellipse with the sun as focus implies that it satisfies the equation

$$(A) \quad r[1 + \varepsilon \cos \theta] = \Lambda,$$

for some ε and Λ . For our (not especially illuminating) proof, we will keep differentiating and substituting from these two equations.

First we differentiate (A) to obtain

$$r'[1 + \varepsilon \cos \theta] - \varepsilon r\theta' \sin \theta = 0.$$

Multiplying by r this becomes

$$rr'[1 + \varepsilon \cos \theta] - \varepsilon r^2\theta' \sin \theta = 0.$$

Using both (A) and (K_2), this becomes

$$\Lambda r' - \varepsilon M \sin \theta = 0.$$

Differentiating again, we get

$$\Lambda r'' - \varepsilon M \theta' \cos \theta = 0.$$

Using (K_2) we get

$$\Lambda r'' - \frac{\varepsilon M^2}{r^2} \cos \theta = 0,$$

and then using (A) we get

$$\Lambda r'' - \frac{M^2}{r^2} \left[\frac{\Lambda}{r} - 1 \right] = 0.$$

Substituting from (K_2) yet again, we get

$$\Lambda [r'' - r(\theta')^2] + \frac{M^2}{r^2} = 0,$$

or

$$r'' - r(\theta')^2 = -\frac{M^2}{\Lambda r^2}.$$

Comparing with (6), we obtain

$$\mathbf{c}''(t) = -\frac{M^2}{\Lambda r^2} \mathbf{e}(\theta(t)),$$

which is precisely what we wanted to show: the force is inversely proportional to the square of the distance from the sun to the planet. ■

CHAPTER 18

THE LOGARITHM AND EXPONENTIAL FUNCTIONS

In Chapter 15 the integral provided a rigorous formulation for a preliminary definition of the functions \sin and \cos . In this chapter the integral plays a more essential role. For certain functions even a preliminary definition presents difficulties. For example, consider the function

$$f(x) = 10^x.$$

This function is assumed to be defined for all x and to have an inverse function, defined for positive x , which is the “logarithm to the base 10,”

$$f^{-1}(x) = \log_{10} x.$$

In algebra, 10^x is usually defined only for *rational* x , while the definition for irrational x is quietly ignored. A brief review of the definition for rational x will not only explain this omission, but also recall an important principle behind the definition of 10^x .

The symbol 10^n is first defined for natural numbers n . This notation turns out to be extremely convenient, especially for multiplying very large numbers, because

$$10^n \cdot 10^m = 10^{n+m}.$$

The extension of the definition of 10^x to rational x is motivated by the desire to preserve this equation; this requirement actually forces upon us the customary definition. Since we want the equation

$$10^0 \cdot 10^n = 10^{0+n} = 10^n$$

to be true, we must define $10^0 = 1$; since we want the equation

$$10^{-n} \cdot 10^n = 10^0 = 1$$

to be true, we must define $10^{-n} = 1/10^n$; since we want the equation

$$\underbrace{10^{1/n} \cdot \dots \cdot 10^{1/n}}_{n \text{ times}} = 10^{\underbrace{1/n + \dots + 1/n}_{n \text{ times}}} = 10^1 = 10$$

to be true, we must define $10^{1/n} = \sqrt[n]{10}$; and since we want the equation

$$\underbrace{10^{1/n} \cdot \dots \cdot 10^{1/n}}_{m \text{ times}} = 10^{\underbrace{1/n + \dots + 1/n}_{m \text{ times}}} = 10^{m/n}$$

to be true, we must define $10^{m/n} = (\sqrt[n]{10})^m$.

Unfortunately, at this point the program comes to a dead halt. We have been guided by the principle that 10^x should be defined so as to ensure that $10^{x+y} = 10^x 10^y$; but this principle does not suggest any simple algebraic way of defining

10^x for irrational x . For this reason we will try some more sophisticated ways of finding a function f such that

$$(*) \quad f(x+y) = f(x) \cdot f(y) \quad \text{for all } x \text{ and } y.$$

Of course, we are interested in a function which is not always zero, so we might add the condition $f(1) \neq 0$. If we add the more specific condition $f(1) = 10$, then $(*)$ will imply that $f(x) = 10^x$ for rational x , and 10^x could be *defined* as $f(x)$ for other x ; in general $f(x)$ will equal $[f(1)]^x$ for rational x .

One way to find such a function is suggested if we try to solve an apparently more difficult problem: find a *differentiable* function f such that

$$\begin{aligned} f(x+y) &= f(x) \cdot f(y) \quad \text{for all } x \text{ and } y, \\ f(1) &= 10. \end{aligned}$$

Assuming that such a function exists, we can try to find f' —knowing the derivative of f might provide a clue to the definition of f itself. Now

$$\begin{aligned} f'(x) &= \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(x) \cdot f(h) - f(x)}{h} \\ &= f(x) \cdot \lim_{h \rightarrow 0} \frac{f(h) - 1}{h}. \end{aligned}$$

The answer thus depends on

$$f'(0) = \lim_{h \rightarrow 0} \frac{f(h) - 1}{h};$$

for the moment assume this limit exists, and denote it by α . Then

$$f'(x) = \alpha \cdot f(x) \quad \text{for all } x.$$

Even if α could be computed, this approach seems self-defeating. The derivative of f has been expressed in terms of f again.

If we examine the inverse function $f^{-1} = \log_{10}$, the whole situation appears in a new light:

$$\begin{aligned} \log_{10}'(x) &= \frac{1}{f'(f^{-1}(x))} \\ &= \frac{1}{\alpha \cdot f(f^{-1}(x))} = \frac{1}{\alpha x}. \end{aligned}$$

The derivative of f^{-1} is about as simple as one could ask! And, what is even more interesting, of all the integrals $\int_a^b x^n dx$ examined previously, the integral $\int_a^b x^{-1} dx$ is the only one which we cannot evaluate. Since $\log_{10} 1 = 0$ we should have

$$\frac{1}{\alpha} \int_1^x \frac{1}{t} dt = \log_{10} x - \log_{10} 1 = \log_{10} x.$$

This suggests that we define $\log_{10} x$ as $(1/\alpha) \int_1^x t^{-1} dt$. The difficulty is that α is unknown. One way of evading this difficulty is to define

$$\log x = \int_1^x \frac{1}{t} dt,$$

and hope that this integral will be the logarithm to *some* base, which might be determined later. In any case, the function defined in this way is surely more reasonable, from a mathematical point of view, than \log_{10} . The usefulness of \log_{10} depends on the important role of the number 10 in arabic notation (and thus ultimately on the fact that we have ten fingers), while the function \log provides a notation for an extremely simple integral which cannot be evaluated in terms of any functions already known to us.

DEFINITION

If $x > 0$, then

$$\log x = \int_1^x \frac{1}{t} dt.$$

The graph of \log is shown in Figure 1. Notice that if $x > 1$, then $\log x > 0$, and if $0 < x < 1$, then $\log x < 0$, since, by our conventions,

$$\int_1^x \frac{1}{t} dt = - \int_x^1 \frac{1}{t} dt < 0.$$

For $x \leq 0$, a number $\log x$ cannot be defined in this way, because $f(t) = 1/t$ is not bounded on $[x, 1]$.

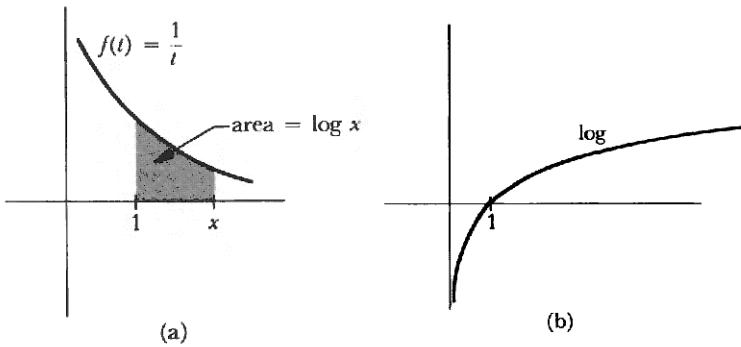


FIGURE 1

The justification for the notation “log” comes from the following theorem.

THEOREM 1 If $x, y > 0$, then

$$\log(xy) = \log x + \log y.$$

PROOF Notice first that $\log'(x) = 1/x$, by the Fundamental Theorem of Calculus. Now choose a number $y > 0$ and let

$$f(x) = \log(xy).$$

Then

$$f'(x) = \log'(xy) \cdot y = \frac{1}{xy} \cdot y = \frac{1}{x}.$$

Thus $f' = \log'$. This means that there is a number c such that

$$f(x) = \log x + c \quad \text{for all } x > 0,$$

that is,

$$\log(xy) = \log x + c \quad \text{for all } x > 0.$$

The number c can be evaluated by noting that when $x = 1$ we obtain

$$\begin{aligned} \log(1 \cdot y) &= \log 1 + c \\ &= c. \end{aligned}$$

Thus

$$\log(xy) = \log x + \log y \quad \text{for all } x.$$

Since this is true for all $y > 0$, the theorem is proved. ■

COROLLARY 1 If n is a natural number and $x > 0$, then

$$\log(x^n) = n \log x.$$

PROOF Let to you (use induction). ■

COROLLARY 2 If $x, y > 0$, then

$$\log\left(\frac{x}{y}\right) = \log x - \log y.$$

PROOF This follows from the equations

$$\log x = \log\left(\frac{x}{y} \cdot y\right) = \log\left(\frac{x}{y}\right) + \log y. \blacksquare$$

Theorem 1 provides some important information about the graph of \log . The function \log is clearly increasing, but since $\log'(x) = 1/x$, the derivative becomes very small as x becomes large, and \log consequently grows more and more slowly. It is not immediately clear whether \log is bounded or unbounded on \mathbf{R} . Observe, however, that for a natural number n ,

$$\log(2^n) = n \log 2 \quad (\text{and } \log 2 > 0);$$

it follows that \log is, in fact, not bounded above. Similarly,

$$\log\left(\frac{1}{2^n}\right) = \log 1 - \log 2^n = -n \log 2;$$

therefore \log is not bounded below on $(0, 1)$. Since \log is continuous, it actually takes on all values. Therefore \mathbf{R} is the domain of the function \log^{-1} . This important function has a special name, whose appropriateness will soon become clear.

DEFINITION

The “exponential function,” **exp**, is defined as \log^{-1} .

The graph of \exp is shown in Figure 2. Since $\log x$ is defined only for $x > 0$, we always have $\exp(x) > 0$. The derivative of the function \exp is easy to determine.

THEOREM 2 For all numbers x ,

$$\exp'(x) = \exp(x).$$

PROOF

$$\begin{aligned}\exp'(x) &= (\log^{-1})'(x) = \frac{1}{\log'(\log^{-1}(x))} \\ &= \frac{1}{\frac{1}{\log^{-1}(x)}} \\ &= \log^{-1}(x) = \exp(x).\blacksquare\end{aligned}$$

A second important property of \exp is an easy consequence of Theorem 1.

THEOREM 3 If x and y are any two numbers, then

$$\exp(x + y) = \exp(x) \cdot \exp(y).$$

PROOF Let $x' = \exp(x)$ and $y' = \exp(y)$, so that

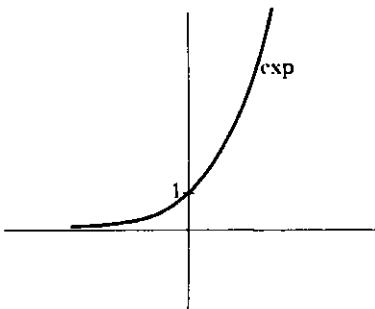
$$\begin{aligned}x &= \log x', \\ y &= \log y'.\end{aligned}$$

Then

$$x + y = \log x' + \log y' = \log(x'y').$$

This means that

$$\exp(x + y) = x'y' = \exp(x) \cdot \exp(y).\blacksquare$$



This theorem, and the discussion at the beginning of this chapter, suggest that $\exp(1)$ is particularly important. There is, in fact, a special symbol for this number.

DEFINITION

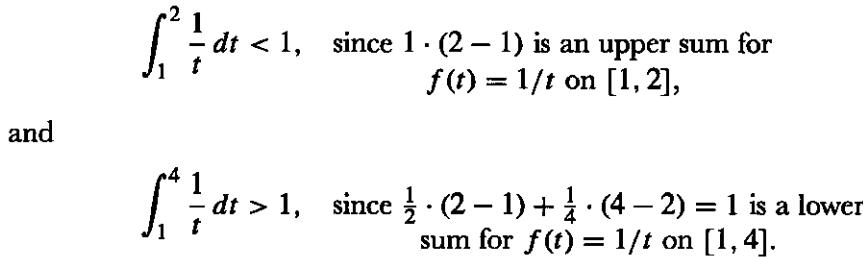
$e = \exp(1)$.

FIGURE 2

This definition is equivalent to the equation

$$1 = \log e = \int_1^e \frac{1}{t} dt.$$

As illustrated in Figure 3,



Thus

$$\int_1^2 \frac{1}{t} dt < \int_1^e \frac{1}{t} dt < \int_1^4 \frac{1}{t} dt,$$

which shows that

$$2 < e < 4.$$

In Chapter 20 we will find much better approximations for e , and also prove that e is irrational (the proof is much easier than the proof that π is irrational!).

As we remarked at the beginning of the chapter, the equation

$$\exp(x + y) = \exp(x) \cdot \exp(y)$$

implies that

$$\begin{aligned}\exp(x) &= [\exp(1)]^x \\ &= e^x, \text{ for all rational } x.\end{aligned}$$

Since \exp is defined for all x and $\exp(x) = e^x$ for rational x , it is consistent with our earlier use of the exponential notation to define e^x as $\exp(x)$ for all x .

DEFINITION

For any number x ,

$$e^x = \exp(x).$$

The terminology “exponential function” should now be clear. We have succeeded in defining e^x for an arbitrary (even irrational) exponent x . We have not yet defined a^x , if $a \neq e$, but there is a reasonable principle to guide us in the attempt. If x is rational, then

$$a^x = (e^{\log a})^x = e^{x \log a}.$$

But the last expression is defined for all x , so we can use it to define a^x .

DEFINITION

If $a > 0$, then, for any real number x ,

$$a^x = e^{x \log a}.$$

(If $a = e$ this definition clearly agrees with the previous one.)

The requirement $a > 0$ is necessary, in order that $\log a$ be defined. This is not unduly restrictive since, for example, we would not even expect

$$(-1)^{1/2} = \sqrt{-1}$$

to be defined. (Of course, for certain rational x , the symbol a^x will make sense, according to the old definition; for example,

$$(-1)^{1/3} = \sqrt[3]{-1} = -1.$$

Our definition of a^x was designed to ensure that

$$(e^x)^y = e^{xy} \quad \text{for all } x \text{ and } y.$$

As we would hope, this equation turns out to be true when e is replaced by any number $a > 0$. The proof is a moderately involved unraveling of terminology. At the same time we will prove the other important properties of a^x .

THEOREM 4

If $a > 0$, then

$$(1) \quad (a^b)^c = a^{bc} \quad \text{for all } b, c.$$

(Notice that a^b will automatically be positive, so $(a^b)^c$ will be defined);

$$(2) \quad a^1 = a \text{ and } a^{x+y} = a^x \cdot a^y \quad \text{for all } x, y.$$

(Notice that (2) implies that this definition of a^x agrees with the old one for all rational x .)

PROOF

$$(1) \quad (a^b)^c = e^{c \log(a^b)} = e^{c \log(e^{b \log a})} = e^{c(b \log a)} = e^{cb \log a} = a^{bc}.$$

(Each of the steps in this string of equalities depends upon our last definition, or the fact that $\exp = \log^{-1}$.)

$$(2) \quad a^1 = e^{1 \log a} = e^{\log a} = a,$$

$$a^{x+y} = e^{(x+y) \log a} = e^{x \log a + y \log a} = e^{x \log a} \cdot e^{y \log a} = a^x \cdot a^y. \blacksquare$$

Figure 4 shows the graphs of $f(x) = a^x$ for several different a . The behavior of the function depends on whether $a < 1$, $a = 1$, or $a > 1$. If $a = 1$, then

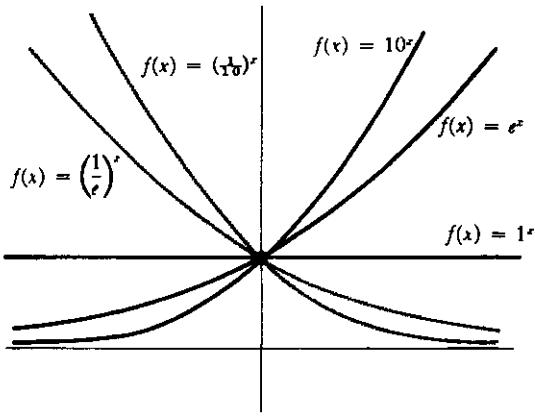


FIGURE 4

$f(x) = 1^x = 1$. Suppose $a > 1$. In this case $\log a > 0$. Thus,

$$\begin{aligned} \text{if } & x < y, \\ \text{then } & x \log a < y \log a, \\ \text{so } & e^{x \log a} < e^{y \log a}, \\ \text{i.e., } & a^x < a^y. \end{aligned}$$

Thus the function $f(x) = a^x$ is increasing. On the other hand, if $0 < a < 1$, so that $\log a < 0$, the same sort of reasoning shows that the function $f(x) = a^x$ is decreasing. In either case, if $a > 0$ and $a \neq 1$, then $f(x) = a^x$ is one-one. Since \exp takes on every positive value it is also easy to see that a^x takes on every positive value. Thus the inverse function is defined for all positive numbers, and takes on all values. If $f(x) = a^x$, then f^{-1} is the function usually denoted by \log_a (Figure 5).

Just as a^x can be expressed in terms of \exp , so \log_a can be expressed in terms of \log . Indeed,

$$\begin{aligned} \text{if } & y = \log_a x, \\ \text{then } & x = a^y = e^{y \log a}, \\ \text{so } & \log x = y \log a, \\ \text{or } & y = \frac{\log x}{\log a}. \end{aligned}$$

In other words,

$$\log_a x = \frac{\log x}{\log a}.$$

The derivatives of $f(x) = a^x$ and $g(x) = \log_a x$ are both easy to find:

$$\begin{aligned} f(x) &= e^{x \log a}, \quad \text{so } f'(x) = \log a \cdot e^{x \log a} = \log a \cdot a^x, \\ g(x) &= \frac{\log x}{\log a}, \quad \text{so } g'(x) = \frac{1}{x \log a}. \end{aligned}$$

A more complicated function like

$$f(x) = g(x)^{h(x)}$$

is also easy to differentiate, if you remember that, *by definition*,

$$f(x) = e^{h(x) \log g(x)};$$

it follows from the Chain Rule that

$$\begin{aligned} f'(x) &= e^{h(x) \log g(x)} \cdot \left[h'(x) \log g(x) + h(x) \frac{g'(x)}{g(x)} \right] \\ &= g(x)^{h(x)} \cdot \left[h'(x) \log g(x) + h(x) \frac{g'(x)}{g(x)} \right]. \end{aligned}$$

There is no point in remembering this formula—simply apply the principle behind it in any specific case that arises; it does help, however, to remember that the first factor in the derivative will be $g(x)^{h(x)}$.

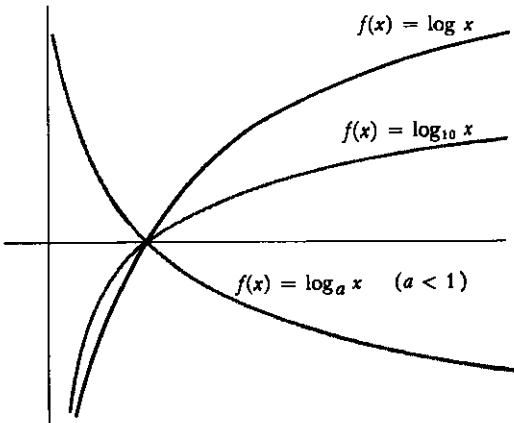


FIGURE 5

There is one special case of the above formula which is worth remembering. The function $f(x) = x^a$ was previously defined only for rational a . We can now define and find the derivative of the function $f(x) = x^a$ for any number a ; the result is just what we would expect:

$$f(x) = x^a = e^{a \log x}$$

so

$$f'(x) = \frac{a}{x} \cdot e^{a \log x} = \frac{a}{x} \cdot x^a = ax^{a-1}.$$

Algebraic manipulations with the exponential functions will become second nature after a little practice—just remember that all the rules which ought to work actually do. The basic properties of \exp are still those stated in Theorems 2 and 3:

$$\begin{aligned}\exp'(x) &= \exp(x), \\ \exp(x+y) &= \exp(x) \cdot \exp(y).\end{aligned}$$

In fact, each of these properties comes close to characterizing the function \exp . Naturally, \exp is not the only function f satisfying $f' = f$, for if $f = ce^x$, then $f'(x) = ce^x = f(x)$; these functions are the only ones with this property, however.

THEOREM 5 If f is differentiable and

$$f'(x) = f(x) \quad \text{for all } x,$$

then there is a number c such that

$$f(x) = ce^x \quad \text{for all } x.$$

PROOF Let

$$g(x) = \frac{f(x)}{e^x}.$$

(This is permissible, since $e^x \neq 0$ for all x .) Then

$$g'(x) = \frac{e^x f'(x) - f(x)e^x}{(e^x)^2} = 0.$$

Therefore there is a number c such that

$$g(x) = \frac{f(x)}{e^x} = c \quad \text{for all } x. \blacksquare$$

The second basic property of \exp requires a more involved discussion. The function \exp is clearly not the only function f which satisfies

$$f(x+y) = f(x) \cdot f(y).$$

In fact, $f(x) = 0$ or any function of the form $f(x) = a^x$ also satisfies this equation. But the true story is much more complex than this—there are infinitely many other functions which satisfy this property, but it is impossible, without appealing to more advanced mathematics, to prove that there is even one function other than those

already mentioned! It is for this reason that the definition of 10^x is so difficult: there are infinitely many functions f which satisfy

$$\begin{aligned} f(x+y) &= f(x) \cdot f(y), \\ f(1) &= 10, \end{aligned}$$

but which are *not* the function $f(x) = 10^x$! One thing is true however—any *continuous* function f satisfying

$$f(x+y) = f(x) \cdot f(y)$$

must be of the form $f(x) = a^x$ or $f(x) = 0$. (Problem 38 indicates the way to prove this, and also has a few words to say about discontinuous functions with this property.)

In addition to the two basic properties stated in Theorems 2 and 3, the function \exp has one further property which is very important— \exp “grows faster than any polynomial.” In other words,

THEOREM 6 For any natural number n ,

$$\lim_{x \rightarrow \infty} \frac{e^x}{x^n} = \infty.$$

PROOF The proof consists of several steps.

Step 1. $e^x > x$ for all x , and consequently $\lim_{x \rightarrow \infty} e^x = \infty$ (this may be considered to be the case $n = 0$).

To prove this statement (which is clear for $x \leq 0$) it suffices to show that

$$x > \log x \quad \text{for all } x > 0.$$

If $x < 1$ this is clearly true, since $\log x < 0$. If $x > 1$, then (Figure 6) $x - 1$ is an upper sum for $f(t) = 1/t$ on $[1, x]$, so $\log x < x - 1 < x$.

Step 2. $\lim_{x \rightarrow \infty} \frac{e^x}{x} = \infty$.

To prove this, note that

$$\frac{e^x}{x} = \frac{e^{x/2} \cdot e^{x/2}}{\frac{x}{2} \cdot 2} = \frac{1}{2} \left(\frac{e^{x/2}}{\frac{x}{2}} \right) \cdot e^{x/2}.$$

By Step 1, the expression in parentheses is greater than 1, and $\lim_{x \rightarrow \infty} e^{x/2} = \infty$; this shows that $\lim_{x \rightarrow \infty} e^x/x = \infty$.

Step 3. $\lim_{x \rightarrow \infty} \frac{e^x}{x^n} = \infty$.

Note that

$$\frac{e^x}{x^n} = \frac{(e^{x/n})^n}{\left(\frac{x}{n}\right)^n \cdot n^n} = \frac{1}{n^n} \cdot \left(\frac{e^{x/n}}{\frac{x}{n}}\right)^n.$$

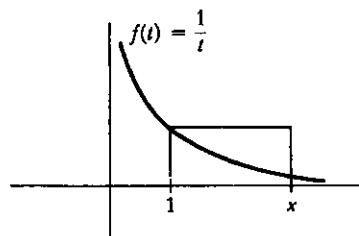


FIGURE 6

The expression in parentheses becomes arbitrarily large, by Step 2, so the n th power certainly becomes arbitrarily large. ■

It is now possible to examine carefully the following very interesting function: $f(x) = e^{-1/x^2}$, $x \neq 0$. We have

$$f'(x) = e^{-1/x^2} \cdot \frac{2}{x^3}.$$

Therefore,

$$\begin{aligned} f'(x) &< 0 & \text{for } x < 0, \\ f'(x) &> 0 & \text{for } x > 0, \end{aligned}$$

so f is decreasing for negative x and increasing for positive x . Moreover, if $|x|$ is large, then x^2 is large, so $-1/x^2$ is close to 0, so e^{-1/x^2} is close to 1 (Figure 7).

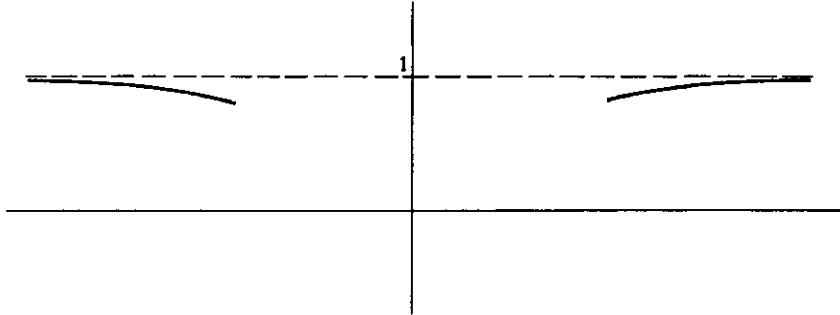


FIGURE 7

The behavior of f near 0 is more interesting. If x is small, then $1/x^2$ is large, so e^{1/x^2} is large, so $e^{-1/x^2} = 1/(e^{1/x^2})$ is small. This argument, suitably stated with ε 's and δ 's, shows that

$$\lim_{x \rightarrow 0} e^{-1/x^2} = 0.$$

Therefore, if we define

$$f(x) = \begin{cases} e^{-1/x^2}, & x \neq 0 \\ 0, & x = 0, \end{cases}$$

then the function f is continuous (Figure 8). In fact, f is actually differentiable

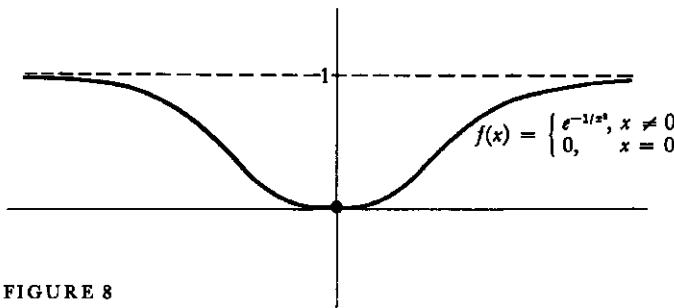


FIGURE 8

at 0: Indeed

$$f'(0) = \lim_{h \rightarrow 0} \frac{e^{-1/h^2}}{h} = \lim_{h \rightarrow 0} \frac{1/h}{e^{(1/h)^2}},$$

and

$$\lim_{h \rightarrow 0^+} \frac{1/h}{e^{(1/h)^2}} = \lim_{x \rightarrow \infty} \frac{x}{e^{(x^2)}}, \quad \text{while} \quad \lim_{h \rightarrow 0^-} \frac{1/h}{e^{(1/h)^2}} = -\lim_{x \rightarrow \infty} \frac{x}{e^{(x^2)}}.$$

We already know that

$$\lim_{x \rightarrow \infty} \frac{e^x}{x} = \infty;$$

it is all the more true that

$$\lim_{x \rightarrow \infty} \frac{e^{(x^2)}}{x} = \infty,$$

and this means that

$$\lim_{x \rightarrow \infty} \frac{x}{e^{(x^2)}} = 0.$$

Thus

$$f'(x) = \begin{cases} e^{-1/x^2} \cdot \frac{2}{x^3}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

We can now compute that

$$\begin{aligned} f''(0) &= \lim_{h \rightarrow 0} \frac{f'(h) - f'(0)}{h} \\ &= \lim_{h \rightarrow 0} \frac{e^{-1/h^2} \cdot \frac{2}{h^3}}{h} \\ &= \lim_{h \rightarrow 0} \frac{2 \cdot e^{-1/h^2}}{h^4} = \lim_{h \rightarrow 0} \frac{2 \cdot \frac{1}{h^4}}{e^{1/h^2}} = \lim_{x \rightarrow \infty} \frac{2x^4}{e^{(x^2)}}, \end{aligned}$$

an argument similar to the one above shows that $f''(0) = 0$. Thus

$$f''(x) = \begin{cases} e^{-1/x^2} \cdot \frac{-6}{x^4} + e^{-1/x^2} \cdot \frac{4}{x^6}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

This argument can be continued. In fact, using induction it can be shown (Problem 40) that $f^{(k)}(0) = 0$ for every k . The function f is *extremely* flat at 0, and approaches 0 so quickly that it can mask many irregularities of other functions. For example (Figure 9), suppose that

$$f(x) = \begin{cases} e^{-1/x^2} \cdot \sin \frac{1}{x}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

It can be shown (Problem 41) that for this function it is also true that $f^{(k)}(0) = 0$ for all k . This example shows, perhaps more strikingly than any other, just how bad a function can be, and still be infinitely differentiable. In Part IV we will investigate even more restrictive conditions on a function, which will finally rule out behavior of this sort.

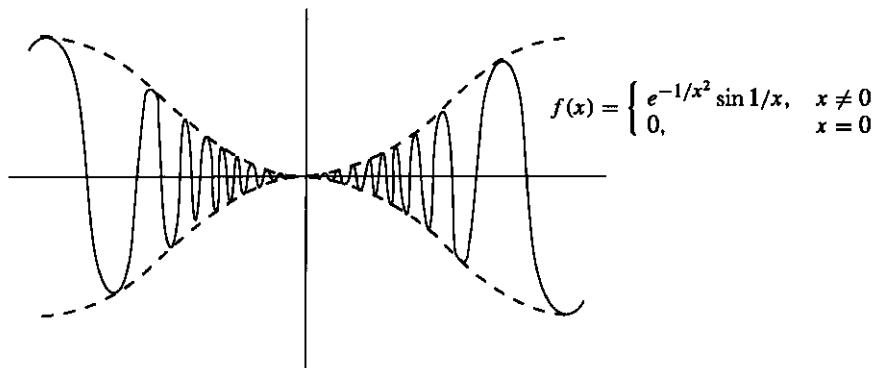


FIGURE 9

PROBLEMS

1. Differentiate each of the following functions (remember that a^{bc} always denotes $a^{(bc)}$).

- (i) $f(x) = e^{e^{e^x}}$.
- (ii) $f(x) = \log(1 + \log(1 + \log(1 + e^{1+e^{1+x}})))$.
- (iii) $f(x) = (\sin x)^{\sin(\sin x)}$.
- (iv) $f(x) = e^{\left(\int_0^x e^{-t^2} dt\right)}$.
- (v) $f(x) = \sin x^{\sin x^{\sin x}}$.
- (vi) $f(x) = \log_{(e^x)} \sin x$.
- (vii) $f(x) = \left[\arcsin \left(\frac{x}{\sin x} \right) \right]^{\log(\sin e^x)}$.
- (viii) $f(x) = (\log(3 + e^4))e^{4x} + (\arcsin x)^{\log 3}$.
- (ix) $f(x) = (\log x)^{\log x}$.
- (x) $f(x) = x^x$.

2. (a) The derivative of $\log \circ f$ is f'/f .

This expression is called the *logarithmic derivative* of f . It is often easier to compute than f' , since products and powers in the expression for f become sums and products in the expression for $\log \circ f$. The derivative f' can then be recovered simply by multiplying by f ; this process is called *logarithmic differentiation*.

- (b) Use logarithmic differentiation to find $f'(x)$ for each of the following.

(i) $f(x) = (1 + x)(1 + e^{x^2})$.

- (ii) $f(x) = \frac{(3-x)^{1/3}x^2}{(1-x)(3+x)^{2/3}}.$
 (iii) $f(x) = (\sin x)^{\cos x} + (\cos x)^{\sin x}.$
 (iv) $f(x) = \frac{e^x - e^{-x}}{e^{2x}(1+x^3)}.$

3. Find

$$\int_a^b \frac{f'(t)}{f(t)} dt$$

(for $f > 0$ on $[a, b]$).

4. Graph each of the following functions.

- (a) $f(x) = e^{x+1}.$
 (b) $f(x) = e^{\sin x}.$
 (c) $f(x) = e^x + e^{-x}.$ } (Compare the graph with the graphs of \exp and
 (d) $f(x) = e^x - e^{-x}.$ } $1/\exp.)$
 (e) $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1} = 1 - \frac{2}{e^{2x} + 1}.$

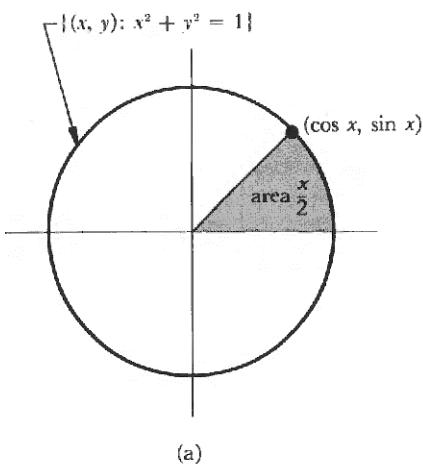
5. Find the following limits by l'Hôpital's Rule.

- (i) $\lim_{x \rightarrow 0} \frac{e^x - 1 - x - x^2/2}{x^2}.$
 (ii) $\lim_{x \rightarrow 0} \frac{e^x - 1 - x - x^2/2 - x^3/6}{x^3}.$
 (iii) $\lim_{x \rightarrow 0} \frac{e^x - 1 - x - x^2/2}{x^3}.$
 (iv) $\lim_{x \rightarrow 0} \frac{\log(1+x) - x + x^2/2}{x^2}.$
 (v) $\lim_{x \rightarrow 0} \frac{\log(1+x) - x + x^2/2}{x^3}.$
 (vi) $\lim_{x \rightarrow 0} \frac{\log(1+x) - x + x^2/2 - x^3/3}{x^3}.$

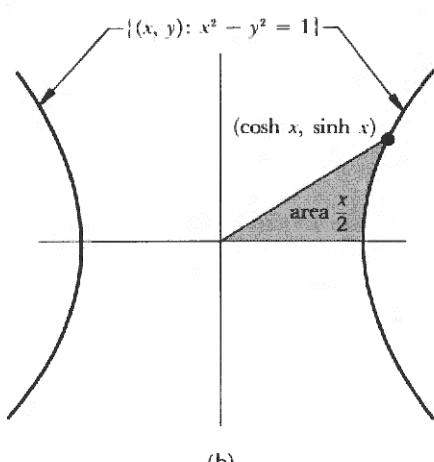
6. The functions

$$\begin{aligned}\sinh x &= \frac{e^x - e^{-x}}{2}, \\ \cosh x &= \frac{e^x + e^{-x}}{2}, \\ \tanh x &= \frac{e^x - e^{-x}}{e^x + e^{-x}} = 1 - \frac{2}{e^{2x} + 1},\end{aligned}$$

are called the **hyperbolic sine**, **hyperbolic cosine**, and **hyperbolic tangent**, respectively (but usually read 'sinch,' 'cosh,' and 'tanh'). There are many analogies between these functions and their ordinary trigonometric counterparts. One analogy is illustrated in Figure 10; a proof that the region



(a)



(b)

shown in Figure 10(b) really has area $x/2$ is best deferred until the next chapter, when we will develop methods of computing integrals. Other analogies are discussed in the following three problems, but the deepest analogies must wait until Chapter 27. If you have not already done Problem 4, graph the functions sinh, cosh, and tanh.

7. Prove that

- (a) $\cosh^2 - \sinh^2 = 1$.
- (b) $\tanh^2 + 1/\cosh^2 = 1$.
- (c) $\sinh(x+y) = \sinh x \cosh y + \cosh x \sinh y$.
- (d) $\cosh(x+y) = \cosh x \cosh y + \sinh x \sinh y$.
- (e) $\sinh' = \cosh$.
- (f) $\cosh' = \sinh$.
- (g) $\tanh' = \frac{1}{\cosh^2}$.

8. The functions sinh and tanh are one-one; their inverses \sinh^{-1} and \tanh^{-1} , are defined on \mathbf{R} and $(-1, 1)$, respectively. These inverse functions are sometimes denoted by $\arg \sinh$ and $\arg \tanh$ (the “argument” of the hyperbolic sine and tangent). If cosh is restricted to $[0, \infty)$ it has an inverse, denoted by $\arg \cosh$, or simply \cosh^{-1} , which is defined on $[1, \infty)$. Prove, using the information in Problem 7, that

- (a) $\sinh(\cosh^{-1} x) = \sqrt{x^2 - 1}$.
- (b) $\cosh(\sinh^{-1} x) = \sqrt{1+x^2}$.
- (c) $(\sinh^{-1})'(x) = \frac{1}{\sqrt{1+x^2}}$.
- (d) $(\cosh^{-1})'(x) = \frac{1}{\sqrt{x^2 - 1}}$ for $x > 1$.
- (e) $(\tanh^{-1})'(x) = \frac{1}{1-x^2}$ for $|x| < 1$.

9. (a) Find an explicit formula for \sinh^{-1} , \cosh^{-1} , and \tanh^{-1} (by solving the equation $y = \sinh^{-1} x$ for x in terms of y , etc.).
 (b) Find

$$\int_a^b \frac{1}{\sqrt{1+x^2}} dx,$$

$$\int_a^b \frac{1}{\sqrt{x^2 - 1}} dx \quad \text{for } a, b > 1 \text{ or } a, b < 1,$$

$$\int_a^b \frac{1}{1-x^2} dx \quad \text{for } |a|, |b| < 1.$$

Compare your answer for the third integral with that obtained by writing

$$\frac{1}{1-x^2} = \frac{1}{2} \left[\frac{1}{1-x} + \frac{1}{1+x} \right].$$

10. Show that

$$F(x) = \int_2^x \frac{1}{\log t} dt$$

is not bounded on $[2, \infty)$.

11. Let f be a nondecreasing function on $[1, \infty)$, and define

$$F(x) = \int_1^x \frac{f(t)}{t} dt, \quad x \geq 1.$$

Prove that f is bounded on $[1, \infty)$ if and only if F/\log is bounded on $[1, \infty)$.

12. Find

(a) $\lim_{x \rightarrow \infty} a^x$ for $0 < a < 1$. (Remember the definition!)

(b) $\lim_{x \rightarrow \infty} \frac{x}{(\log x)^n}$.

(c) $\lim_{x \rightarrow \infty} \frac{(\log x)^n}{x}$.

(d) $\lim_{x \rightarrow 0^+} x(\log x)^n$. Hint: $x(\log x)^n = \frac{(-1)^n \left(\log \frac{1}{x}\right)^n}{\frac{1}{x}}$.

(e) $\lim_{x \rightarrow 0^+} x^x$.

13. Graph $f(x) = x^x$ for $x > 0$. (Use Problem 12(e).)

14. (a) Find the minimum value of $f(x) = e^x/x^n$ for $x > 0$, and conclude that $f(x) > e^n/n^n$ for $x > n$.

- (b) Using the expression $f'(x) = e^x(x - n)/x^{n+1}$, prove that $f'(x) > e^{n+1}/(n + 1)^{n+1}$ for $x > n + 1$, and thus obtain another proof that $\lim_{x \rightarrow \infty} f(x) = \infty$.

15. Graph $f(x) = e^x/x^n$.

16. (a) Find $\lim_{y \rightarrow 0} \log(1 + y)/y$. (You can use l'Hôpital's Rule, but that would be silly.)

- (b) Find $\lim_{x \rightarrow \infty} x \log(1 + 1/x)$.

- (c) Prove that $e = \lim_{x \rightarrow \infty} (1 + 1/x)^x$.

- (d) Prove that $e^a = \lim_{x \rightarrow \infty} (1 + a/x)^x$. (It is possible to derive this from part (c) with just a little algebraic fiddling.)

- *(e) Prove that $\log b = \lim_{x \rightarrow \infty} x(b^{1/x} - 1)$.

17. Graph $f(x) = (1 + 1/x)^x$ for $x > 0$. (Use Problem 16(c).)

18. If a bank gives a percent interest per annum, then an initial investment I yields $I(1 + a/100)$ after 1 year. If the bank compounds the interest (counts the accrued interest as part of the capital for computing interest the next

year), then the initial investment grows to $I(1 + a/100)^n$ after n years. Now suppose that interest is given twice a year. The final amount after n years is, alas, not $I(1 + a/100)^{2n}$, but merely $I(1 + a/200)^{2n}$ —although interest is awarded twice as often, the interest must be halved in each calculation, since the interest is $a/2$ per half year. This amount is larger than $I(1 + a/100)^n$, but not that much larger. Suppose that the bank now compounds the interest continuously, i.e., the bank considers what the investment would yield when compounding k times a year, and then takes the least upper bound of all these numbers. How much will an initial investment of 1 dollar yield after 1 year?

19. (a) Let $f(x) = \log|x|$ for $x \neq 0$. Prove that $f'(x) = 1/x$ for $x \neq 0$.
 (b) If $f(x) \neq 0$ for all x , prove that $(\log|f|)' = f'/f$.
20. Suppose that on some interval the function f satisfies $f' = cf$ for some number c .
 - (a) Assuming that f is never 0, use Problem 19(b) to prove that $|f(x)| = le^{cx}$ for some number $l (> 0)$. It follows that $f(x) = ke^{cx}$ for some k .
 - (b) Show that this result holds without the added assumption that f is never 0. Hint: Show that f can't be 0 at the endpoint of an open interval on which it is nowhere 0.
 - (c) Give a simpler proof that $f(x) = ke^{cx}$ for some k by considering the function $g(x) = f(x)/e^{cx}$.
 - (d) Suppose that $f' = fg'$ for some g . Show that $f(x) = ke^{g(x)}$ for some k .
- *21. A radioactive substance diminishes at a rate proportional to the amount present (since all atoms have equal probability of disintegrating, the total disintegration is proportional to the number of atoms remaining). If $A(t)$ is the amount at time t , this means that $A'(t) = cA(t)$ for some c (which represents the probability that an atom will disintegrate).
 - (a) Find $A(t)$ in terms of the amount $A_0 = A(0)$ present at time 0.
 - (b) Show that there is a number τ (the “half-life” of the radioactive element) with the property that $A(t + \tau) = A(t)/2$.
- *22. *Newton's law of cooling* states that an object cools at a rate proportional to the difference of its temperature and the temperature of the surrounding medium. Find the temperature $T(t)$ of the object at time t , in terms of its temperature T_0 at time 0, assuming that the temperature of the surrounding medium is kept at a constant, M . Hint: To solve the differential equation expressing Newton's law, remember that $T' = (T - M)'$.
- *23. Prove that if $f(x) = \int_0^x f(t) dt$, then $f = 0$.
24. Find all continuous functions f satisfying
 - (i) $\int_0^x f = e^x$.

$$(ii) \quad \int_0^{x^2} f = 1 - e^{2x^2}.$$

25. Given a differentiable function f , find all continuous functions g satisfying

$$\int_0^{f(x)} fg = g(f(x)) - 1.$$

- *26. Find all functions f satisfying $f'(t) = f(t) + \int_0^1 f(t) dt$.

27. Find all continuous functions f which satisfy the equation

$$(f(x))^2 = \int_0^x f(t) \frac{t}{1+t^2} dt.$$

28. (a) Let f and g be continuous nonnegative functions on $[a, b]$, and let $C > 0$. Suppose that

$$f(x) \leq C + \int_a^x fg \quad a \leq x \leq b.$$

Prove *Gronwall's inequality*:

$$f(x) \leq Ce^{\int_a^x g}.$$

Hint: Consider the derivative of the function $h(x) = C + \int_a^x fg$.

- (b) Use a limiting argument to show that this result holds even for $C = 0$.
(c) Suppose that $f'(x) = g(x)f(x)$ for some continuous function g , and that $f(0) = 0$. Then $f = 0$. (Compare Problem 20.)

29. (a) Prove that

$$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} \leq e^x \quad \text{for } x \geq 0.$$

Hint: Use induction on n , and compare derivatives.

- (b) Give a new proof that $\lim_{x \rightarrow \infty} e^x/x^n = \infty$.

30. Give yet another proof of this fact, using the appropriate form of l'Hôpital's Rule. (See Problem 11-53.)

31. (a) Evaluate $\lim_{x \rightarrow \infty} e^{-x^2} \int_0^x e^{t^2} dt$. (You should be able to make an educated guess before doing any calculations.)

- (b) Evaluate the following limits.

$$(i) \quad \lim_{x \rightarrow \infty} e^{-x^2} \int_x^{x+(1/x)} e^{t^2} dt.$$

$$(ii) \quad \lim_{x \rightarrow \infty} e^{-x^2} \int_x^{x+(\log x/x)} e^{t^2} dt.$$

$$(iii) \quad \lim_{x \rightarrow \infty} e^{-x^2} \int_x^{x+(\log x/2x)} e^{t^2} dt.$$

32. This problem outlines the classical approach to logarithms and exponentials. To begin with, we will simply assume that the function $f(x) = a^x$, defined in an elementary way for rational x , can somehow be extended to a continuous one-one function, obeying the same algebraic rules, on the whole line. (See Problem 22-29 for a direct proof of this.) The inverse of f will then be denoted by \log_a .

- (a) Show, directly from the definition, that

$$\begin{aligned}\log_a'(x) &= \lim_{h \rightarrow 0} \log_a \left(1 + \frac{h}{x} \right)^{1/h} \\ &= \frac{1}{x} \cdot \log_a \left(\lim_{k \rightarrow 0} (1+k)^{1/k} \right).\end{aligned}$$

Thus, the whole problem has been reduced to the determination of $\lim_{h \rightarrow 0} (1+h)^{1/h}$. If we can show that this has a limit e , then $\log_e'(x) = \frac{1}{x} \cdot \log_e e = \frac{1}{x}$, and consequently $\exp = \log_e^{-1}$ has derivative $\exp'(x) = \exp(x)$.

- (b) Let $a_n = \left(1 + \frac{1}{n}\right)^n$ for natural numbers n . Using the binomial theorem, show that

$$a_n = 2 + \sum_{k=2}^n \frac{1}{k!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right).$$

Conclude that $a_n < a_{n+1}$.

- (c) Using the fact that $1/k! \leq 1/2^{k-1}$ for $k \geq 2$, show that all $a_n < 3$. Thus, the set of numbers $\{a_1, a_2, a_3, \dots\}$ is bounded, and therefore has a least upper bound e . Show that for any $\varepsilon > 0$ we have $e - a_n < \varepsilon$ for large enough n .
- (d) If $n \leq x \leq n+1$, then

$$\left(1 + \frac{1}{n+1}\right)^n \leq \left(1 + \frac{1}{x}\right)^x \leq \left(1 + \frac{1}{n+1}\right)^{n+1}.$$

Conclude that $\lim_{x \rightarrow \infty} \left(1 + \frac{1}{x}\right)^x = e$. Also show that $\lim_{x \rightarrow -\infty} \left(1 + \frac{1}{x}\right)^x = e$, and conclude that $\lim_{h \rightarrow 0} (1+h)^{1/h} = e$.

- *33. A point P is moving along a line segment AB of length 10^7 while another point Q moves along an infinite ray (Figure 11). The velocity of P is always equal to the distance from P to B (in other words, if $P(t)$ is the position of P at time t , then $P'(t) = 10^7 - P(t)$), while Q moves with constant velocity $Q'(t) = 10^7$. The distance traveled by Q after time t is defined to be the *Napierian logarithm* of the distance from P to B at time t . Thus

$$10^7 t = \text{Nap log}[10^7 - P(t)].$$

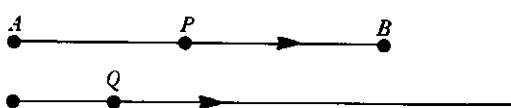


FIGURE 11

This was the definition of logarithms given by Napier (1550–1617) in his publication of 1614, *Mirifici logarithmonum canonis description* (A Description of the Wonderful Law of Logarithms); work which was done *before* the use of exponents was invented! The number 10^7 was chosen because Napier's tables (intended for astronomical and navigational calculations), listed the logarithms of sines of angles, for which the best possible available tables extended to seven decimal places, and Napier wanted to avoid fractions. Prove that

$$\text{Nap log } x = 10^7 \log \frac{10^7}{x}.$$

Hint: Use the same trick as in Problem 22 to solve the equation for P .

- *34. (a) Sketch the graph of $f(x) = (\log x)/x$ (paying particular attention to the behavior near 0 and ∞).
- (b) Which is larger, e^π or π^e ?
- (c) Prove that if $0 < x \leq 1$, or $x = e$, then the only number y satisfying $x^y = y^x$ is $y = x$; but if $x > 1$, $x \neq e$, then there is precisely one number $y \neq x$ satisfying $x^y = y^x$; moreover, if $x < e$, then $y > e$, and if $x > e$, then $y < e$. (Interpret these statements in terms of the graph in part (a)!)
- (d) Prove that if x and y are natural numbers and $x^y = y^x$, then $x = y$ or $x = 2$, $y = 4$, or $x = 4$, $y = 2$.
- (e) Show that the set of all pairs (x, y) with $x^y = y^x$ consists of a curve and a straight line which intersect; find the intersection and draw a rough sketch.
- **(f) For $1 < x < e$ let $g(x)$ be the unique number $> e$ with $x^{g(x)} = g(x)^x$. Prove that g is differentiable. (It is a good idea to consider separate functions,

$$f_1(x) = \frac{\log x}{x}, \quad 0 < x < e$$

$$f_2(x) = \frac{\log x}{x}, \quad e < x$$

and write g in terms of f_1 and f_2 . You should be able to show that

$$g'(x) = \frac{[g(x)]^2}{1 - \log g(x)} \cdot \frac{1 - \log x}{x^2}$$

if you do this part properly.)

- *35. This problem uses the material from the Appendix to Chapter 11.

- (a) Prove that \exp is convex and \log is concave.
- (b) Prove that if $\sum_{i=1}^n p_i = 1$ and all $p_i > 0$, then

$$z_1^{p_1} \cdots z_n^{p_n} < p_1 z_1 + \cdots + p_n z_n.$$

(Use Problem 9 from the Appendix to Chapter 11.)

- (c) Deduce another proof that $G_n \leq A_n$ (Problem 2-22).

36. (a) Let f be a positive function on $[a, b]$, and let P_n be the partition of $[a, b]$ into n equal intervals. Use Problem 2-22 to show that

$$\frac{1}{b-a} L(\log f, P_n) \leq \log \left(\frac{1}{b-a} L(f, P_n) \right).$$

- (b) Use the Appendix to Chapter 13 to conclude that for all integrable $f > 0$ we have

$$\frac{1}{b-a} \int_a^b \log f \leq \log \left(\frac{1}{b-a} \int_a^b f \right).$$

A more direct approach is illustrated in the next part:

- (c) In Problem 35, Problem 2-22 was deduced as a special case of the inequality

$$g \left(\sum_{i=1}^n p_i x_i \right) \leq \sum_{i=1}^n p_i g(x_i)$$

for $p_i > 0$, $\sum_{i=1}^n p_i = 1$ and g convex. For g concave we have the reverse inequality

$$\sum_{i=1}^n p_i g(x_i) \leq g \left(\sum_{i=1}^n p_i x_i \right).$$

Apply this with $g = \log$ to prove the result of part (b) directly for any integrable f .

- (d) State a general theorem of which part (b) is just a special case.

37. Suppose f satisfies $f' = f$ and $f(x+y) = f(x)f(y)$ for all x and y . Prove that $f = \exp$ or $f = 0$.

- *38. Prove that if f is continuous and $f(x+y) = f(x)f(y)$ for all x and y , then either $f = 0$ or $f(x) = [f(1)]^x$ for all x . Hint: Show that $f(x) = [f(1)]^x$ for rational x , and then use Problem 8-6. This problem is closely related to Problem 8-7, and the information mentioned at the end of Problem 8-7 can be used to show that there are discontinuous functions f satisfying $f(x+y) = f(x)f(y)$.

- *39. Prove that if f is a continuous function defined on the positive real numbers, and $f(xy) = f(x) + f(y)$ for all positive x and y , then $f = 0$ or $f(x) = f(e) \log x$ for all $x > 0$. Hint: Consider $g(x) = f(e^x)$.

- *40. Prove that if $f(x) = e^{-1/x^2}$ for $x \neq 0$, and $f(0) = 0$, then $f^{(k)}(0) = 0$ for all k .

- *41. Prove that if $f(x) = e^{-1/x^2} \sin 1/x$ for $x \neq 0$, and $f(0) = 0$, then $f^{(k)}(0) = 0$ for all k .

- 42.** (a) Prove that if α is a root of the equation

$$(*) \quad a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 = 0,$$

then the function $y(x) = e^{\alpha x}$ satisfies the differential equation

$$(**) \quad a_n y^{(n)} + a_{n-1} y^{(n-1)} + \cdots + a_1 y' + a_0 y = 0.$$

- *(b) Prove that if α is a double root of (*), then $y(x) = x e^{\alpha x}$ also satisfies (**).
Hint: Remember that if α is a double root of a polynomial equation $f(x) = 0$, then $f'(x) = 0$.
(c) Prove that if α is a root of (*) of order r , then $y(x) = x^k e^{\alpha x}$ is a solution for $0 \leq k \leq r - 1$.

If (*) has n real numbers as roots (counting multiplicities), part (c) gives n solutions y_1, \dots, y_n of (**).

- (d) Prove that in this case the function $c_1 y_1 + \cdots + c_n y_n$ also satisfies (**).

It is a theorem that in this case these are the only solutions of (**). Problem 20 and the next two problems prove special cases of this theorem, and the general case is considered in Problem 20-19. In Chapter 27 we will see what to do when (*) does not have n real numbers as roots.

- *43.** Suppose that f satisfies $f'' - f = 0$ and $f(0) = f'(0) = 0$. Prove that $f = 0$ as follows.

- (a) Show that $f^2 - (f')^2 = 0$.
 - (b) Suppose that $f(x) \neq 0$ for all x in some interval (a, b) . Show that either $f(x) = ce^x$ or else $f(x) = ce^{-x}$ for all x in (a, b) , for some constant c .
 - **(c) If $f(x_0) \neq 0$ for $x_0 > 0$, say, then there would be a number a such that $0 \leq a < x_0$ and $f(a) = 0$, while $f(x) \neq 0$ for $a < x < x_0$. Why? Use this fact and part (b) to deduce a contradiction.
- *44.** (a) Show that if f satisfies $f'' - f = 0$, then $f(x) = ae^x + be^{-x}$ for some a and b . (First figure out what a and b should be in terms of $f(0)$ and $f'(0)$, and then use Problem 43.)
(b) Show also that $f = a \sinh x + b \cosh x$ for some (other) a and b .

- 45.** Find all functions f satisfying

- (a) $f^{(n)} = f^{(n-1)}$.
- (b) $f^{(n)} = f^{(n-2)}$.

- *46.** This problem, a companion to Problem 15-30, outlines a treatment of the exponential function starting from the assumption that the differential equation $f' = f$ has a nonzero solution.
- (a) Suppose there is a function $f \neq 0$ with $f' = f$. Prove that $f(x) \neq 0$ for each x by considering the function $g(x) = f(x_0 + x)f(x - x_0)$, where $f(x_0) \neq 0$.

- (b) Show that there is a function f satisfying $f' = f$ and $f(0) = 1$.
 (c) For this f show that $f(x+y) = f(x) \cdot f(y)$ by considering the function $g(x) = f(x+y)/f(x)$.
 (d) Prove that f is one-one and that $(f^{-1})'(x) = 1/x$.
47. Let f and g be continuous functions such that $\lim_{x \rightarrow \infty} f(x) = \lim_{x \rightarrow \infty} g(x) = \infty$. We say that f grows faster than g ($f \gg g$) if
- $$\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = \infty,$$
- and we say that f and g grow at the same rate ($f \sim g$) if
- $$\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} \text{ exists and is } \neq 0, \infty.$$
- For example, $\exp \gg P$ for any polynomial function P , and $P \gg \log^n$ for any positive integer n .
- (a) Given f and g , with $\lim_{x \rightarrow \infty} f(x) = \lim_{x \rightarrow \infty} g(x) = \infty$, is it necessarily true that one of the three conditions $f \gg g$ or $g \gg f$ or $f \sim g$ holds?
 (b) If $f \gg g$, then $f + g \sim f$.
 (c) If
- $$\frac{\log f}{\log g} \geq c > 1$$
- for sufficiently large x , then $f \gg g$.
- (d) If $f \gg g$ and $F(x) = \int_0^x f$, $G(x) = \int_0^x g$, does it necessarily follow that $F \gg G$?
 (e) Arrange each of the following sets of functions in increasing order of growth (for convenience, we indicate each function simply by giving its value at x):
- (i) $x^3, e^x, x^3 + \log(x^3), \log 4x, (\log x)^x, x^x, x + e^{-5x}, x^3 \log x$.
 - (ii) $x \log^2 x, e^{5x}, \log(x^x), e^{x^2}, x^x, x^{\log x}, (\log x)^x$.
 - (iii) $e^x, x^e, x^x, e^{x^2}, 2^x, e^{x/2}, (\log x)^{2x}$.
48. Suppose that g_1, g_2, g_3, \dots are continuous functions. Show that there is a continuous function f which grows faster than each g_i .
49. Prove that $\log_{10} 2$ is irrational.

CHAPTER 19

INTEGRATION IN ELEMENTARY TERMS

Every computation of a derivative yields, according to the Second Fundamental Theorem of Calculus, a formula about integrals. For example,

$$\text{if } F(x) = x(\log x) - x \quad \text{then } F'(x) = \log x;$$

consequently,

$$\int_a^b \log x \, dx = F(b) - F(a) = b(\log b) - b - [a(\log a) - a], \quad 0 < a, b.$$

Formulas of this sort are simplified considerably if we adopt the notation

$$F(x) \Big|_a^b = F(b) - F(a).$$

We may then write

$$\int_a^b \log x \, dx = x(\log x) - x \Big|_a^b.$$

This evaluation of $\int_a^b \log x \, dx$ depended on the lucky guess that \log is the derivative of the function $F(x) = x(\log x) - x$. In general, a function F satisfying $F' = f$ is called a **primitive** of f . Of course, a **continuous function f always has a primitive**, namely,

$$F(x) = \int_a^x f,$$

but in this chapter we will try to find a primitive which can be written in terms of familiar functions like \sin , \log , etc. A function which can be written in this way is called an **elementary function**. To be precise,* an **elementary function** is one which can be obtained by addition, multiplication, division, and composition from the rational functions, the trigonometric functions and their inverses, and the functions \log and \exp .

It should be stated at the very outset that elementary primitives usually cannot be found. For example, there is no *elementary* function F such that

$$F'(x) = e^{-x^2} \quad \text{for all } x$$

(this is not merely a report on the present state of mathematical ignorance; it is a (difficult) theorem that no such function exists). And, what is even worse, you

*The definition which we will give is precise, but not really accurate, or at least not quite standard. Usually the elementary functions are defined to include “algebraic” functions, that is, functions g satisfying an equation

$$(g(x))^n + f_{n-1}(x)(g(x))^{n-1} + \cdots + f_0(x) = 0,$$

where the f_i are rational functions. But for our purposes these functions can be ignored.

will have no way of knowing whether or not an elementary primitive *can* be found (you will just have to hope that the problems for this chapter contain no misprints). Because the search for elementary primitives is so uncertain, finding one is often peculiarly satisfying. If we observe that the function

$$F(x) = x \arctan x - \frac{\log(1+x^2)}{2}$$

satisfies

$$F'(x) = \arctan x$$

(just how we would ever be led to such an observation is quite another matter), so that

$$\int_a^b \arctan x \, dx = x \arctan x - \frac{\log(1+x^2)}{2} \Big|_a^b,$$

then we may feel that we have “really” evaluated $\int_a^b \arctan x \, dx$.

This chapter consists of little more than methods for finding elementary primitives of given elementary functions (a process known simply as “integration”), together with some notation, abbreviations, and conventions designed to facilitate this procedure. This preoccupation with elementary functions can be justified by three considerations:

- (1) Integration is a standard topic in calculus, and everyone should know about it.
- (2) Every once in a while you might actually need to evaluate an integral, under conditions which do not allow you to consult any of the standard integral tables (for example, you might take a (physics) course in which you are expected to be able to integrate).
- (3) The most useful “methods” of integration are actually very important theorems (that apply to all functions, not just elementary ones).

Naturally, the last reason is the crucial one. Even if you intend to forget how to integrate (and you probably will forget some details the first time through), you must never forget the basic methods.

These basic methods are theorems which allow us to express primitives of one function in terms of primitives of other functions. To begin integrating we will therefore need a list of primitives for *some* functions; such a list can be obtained simply by differentiating various well-known functions. The list given below makes use of a standard symbol which requires some explanation. The symbol

$$\int f \quad \text{or} \quad \int f(x) \, dx$$

means “a primitive of f ” or, more precisely, “the collection of all primitives of f .“ The symbol $\int f$ will often be used in stating theorems, while $\int f(x) \, dx$ is most useful in formulas like the following:

$$\int x^3 \, dx = \frac{x^4}{4}.$$

This “equation” means that the function $F(x) = x^4/4$ satisfies $F'(x) = x^3$. It cannot be interpreted literally because the right side is a number, not a function, but in this one context we will allow such discrepancies; our aim is to make the integration process as mechanical as possible, and we will resort to any possible device. Another feature of the equation deserves mention. Most people write

$$\int x^3 dx = \frac{x^4}{4} + C$$

to emphasize that the primitives of $f(x) = x^3$ are precisely the functions of the form $F(x) = x^4/4 + C$ for some number C . Although it is possible (Problem 13) to obtain contradictions if this point is disregarded, in practice such difficulties do not arise, and concern for this constant is merely an annoyance.

There is one important convention accompanying this notation: the letter appearing on the right side of the equation should match with the letter appearing after the “ d ” on the left side—thus

$$\begin{aligned}\int u^3 du &= \frac{u^4}{4}, \\ \int tx dx &= \frac{tx^2}{2}, \\ \int tx dt &= \frac{xt^2}{2}.\end{aligned}$$

A function in $\int f(x) dx$, i.e., a primitive of f , is often called an “indefinite integral” of f , while $\int_a^b f(x) dx$ is called, by way of contrast, a “definite integral.” This suggestive notation works out quite well in practice, but it is important not to be led astray. At the risk of boring you, the following fact is emphasized once again: the integral $\int_a^b f(x) dx$ is *not* defined as $F(b) - F(a)$, where F is an indefinite integral of f ” (if you do not find this statement repetitious, it is time to reread Chapter 13).

We can verify the formulas in the following short table of indefinite integrals simply by differentiating the functions indicated on the right side.

$$\int a dx = ax$$

$$\int x^n dx = \frac{x^{n+1}}{n+1}, \quad n \neq -1$$

$$\int \frac{1}{x} dx = \log x \quad (\int \frac{1}{x} dx \text{ is often written } \int \frac{dx}{x} \text{ for convenience; similar abbreviations are used in the last two examples of this table.})$$

$$\int e^x dx = e^x$$

$$\int \sin x dx = -\cos x$$

$$\begin{aligned}\int \cos x \, dx &= \sin x \\ \int \sec^2 x \, dx &= \tan x \\ \int \sec x \tan x \, dx &= \sec x \\ \int \frac{dx}{1+x^2} &= \arctan x \\ \int \frac{dx}{\sqrt{1-x^2}} &= \arcsin x\end{aligned}$$

Two general formulas of the same nature are consequences of theorems about differentiation:

$$\begin{aligned}\int [f(x) + g(x)] \, dx &= \int f(x) \, dx + \int g(x) \, dx, \\ \int c \cdot f(x) \, dx &= c \cdot \int f(x) \, dx.\end{aligned}$$

These equations should be interpreted as meaning that a primitive of $f + g$ can be obtained by adding a primitive of f to a primitive of g , while a primitive of $c \cdot f$ can be obtained by multiplying a primitive of f by c .

Notice the consequences of these formulas for definite integrals: If f and g are continuous, then

$$\begin{aligned}\int_a^b [f(x) + g(x)] \, dx &= \int_a^b f(x) \, dx + \int_a^b g(x) \, dx, \\ \int_a^b c \cdot f(x) \, dx &= c \cdot \int_a^b f(x) \, dx.\end{aligned}$$

These follow from the previous formulas, since each definite integral may be written as the difference of the values at a and b of a corresponding primitive. Continuity is required in order to know that these primitives exist. (Of course, the formulas are also true when f and g are merely integrable, but recall how much more difficult the proofs are in this case.)

The product formula for the derivative yields a more interesting theorem, which will be written in several different ways.

THEOREM 1 (INTEGRATION BY PARTS)

If f' and g' are continuous, then

$$\begin{aligned}\int f g' &= f g - \int f' g, \\ \int f(x) g'(x) \, dx &= f(x) g(x) - \int f'(x) g(x) \, dx, \\ \int_a^b f(x) g'(x) \, dx &= f(x) g(x) \Big|_a^b - \int_a^b f'(x) g(x) \, dx.\end{aligned}$$

(Notice that in the second equation $f(x)g(x)$ denotes the function $f \cdot g$.)

PROOF The formula

$$(fg)' = f'g + fg'$$

can be written

$$fg' = (fg)' - f'g.$$

Thus

$$\int fg' = \int (fg)' - \int f'g,$$

and fg can be chosen as one of the functions denoted by $\int (fg)'$. This proves the first formula.

The second formula is merely a restatement of the first, and the third formula follows immediately from either of the first two. ■

As the following examples illustrate, integration by parts is useful when the function to be integrated can be considered as a product of a function f , whose derivative is simpler than f , and another function which is obviously of the form g' .

$$\begin{aligned} \int xe^x dx &= xe^x - \int 1 \cdot e^x dx \\ &\quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\ &f g' \quad f g \quad f' g \\ &= xe^x - e^x \\ \int x \sin x dx &= x \cdot (-\cos x) - \int 1 \cdot (-\cos x) dx \\ &\quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\ &f g' \quad f \quad g \quad f' \quad g \\ &= -x \cos x + \sin x \end{aligned}$$

There are two special tricks which often work with integration by parts. The first is to consider the function g' to be the factor 1, which can always be written in.

$$\begin{aligned} \int \log x dx &= \int 1 \cdot \log x dx = x \log x - \int x \cdot (1/x) dx \\ &\quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\ &g' \quad f \quad g \quad f \quad g \quad f' \\ &= x(\log x) - x. \end{aligned}$$

The second trick is to use integration by parts to find $\int h$ in terms of $\int h$ again, and then solve for $\int h$. A simple example is the calculation

$$\int (1/x) \cdot \log x dx = \log x \cdot \log x - \int (1/x) \cdot \log x dx,$$

$$\quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow$$

$$g' \quad f \quad g \quad f \quad f' \quad g$$

which implies that

$$2 \int \frac{1}{x} \log x dx = (\log x)^2$$

or

$$\int \frac{1}{x} \log x \, dx = \frac{(\log x)^2}{2}.$$

A more complicated calculation is often required:

$$\begin{aligned} \int e^x \sin x \, dx &= e^x \cdot (-\cos x) - \int e^x \cdot (-\cos x) \, dx \\ &\quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\ f & g' & f & g & f' & g \\ & & & & & \\ & = -e^x \cos x + \int e^x \cos x \, dx \\ & \quad \downarrow \quad \downarrow \\ u & v' \\ & = -e^x \cos x + [e^x \cdot (\sin x) - \int e^x (\sin x) \, dx]; \\ & \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\ u & v & u' & v \end{aligned}$$

therefore,

$$2 \int e^x \sin x \, dx = e^x (\sin x - \cos x)$$

or

$$\int e^x \sin x \, dx = \frac{e^x (\sin x - \cos x)}{2}.$$

Since integration by parts depends upon recognizing that a function is of the form g' , the more functions you can already integrate, the greater your chances for success. It is frequently reasonable to do a preliminary integration before tackling the main problem. For example, we can use parts to integrate

$$\int (\log x)^2 \, dx = \int (\log x)(\log x) \, dx$$

$$\quad \downarrow \quad \downarrow$$

$$f \quad g'$$

if we recall that $\int \log x \, dx = x(\log x) - x$ (this formula was itself derived by integration by parts); we have

$$\begin{aligned} \int (\log x)(\log x) \, dx &= (\log x)[x(\log x) - x] - \int (1/x)[x(\log x) - x] \, dx \\ &\quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \\ f & g' & f & g & f' & g \\ & & & & & \\ & = (\log x)[x(\log x) - x] - \int [\log x - 1] \, dx \\ & = (\log x)[x(\log x) - x] - \int \log x \, dx + \int 1 \, dx \\ & = (\log x)[x(\log x) - x] - [x(\log x) - x] + x \\ & = x(\log x)^2 - 2x(\log x) + 2x. \end{aligned}$$

The most important method of integration is a consequence of the Chain Rule. The use of this method requires considerably more ingenuity than integrating by parts, and even the explanation of the method is more difficult. We will therefore

develop this method in stages, stating the theorem for definite integrals first, and saving the treatment of indefinite integrals for later.

THEOREM 2 If f and g' are continuous, then

(THE SUBSTITUTION FORMULA)

$$\int_{g(a)}^{g(b)} f = \int_a^b (f \circ g) \cdot g' \\ \int_{g(a)}^{g(b)} f(u) du = \int_a^b f(g(x)) \cdot g'(x) dx.$$

PROOF If F is a primitive of f , then the left side is $F(g(b)) - F(g(a))$. On the other hand,

$$(F \circ g)' = (F' \circ g) \cdot g' = (f \circ g) \cdot g',$$

so $F \circ g$ is a primitive of $(f \circ g) \cdot g'$ and the right side is

$$(F \circ g)(b) - (F \circ g)(a) = F(g(b)) - F(g(a)). \blacksquare$$

The simplest uses of the substitution formula depend upon recognizing that a given function is of the form $(f \circ g) \cdot g'$. For example, the integration of

$$\int_a^b \sin^5 x \cos x dx \quad \left(= \int_a^b (\sin x)^5 \cos x dx \right)$$

is facilitated by the appearance of the factor $\cos x$, which will be the factor $g'(x)$ for $g(x) = \sin x$; the remaining expression, $(\sin x)^5$, can be written as $(g(x))^5 = f(g(x))$, for $f(u) = u^5$. Thus

$$\begin{aligned} & \int_a^b \sin^5 x \cos x dx \quad \left[\begin{array}{l} g(x) = \sin x \\ f(u) = u^5 \end{array} \right] \\ &= \int_a^b f(g(x))g'(x) dx = \int_{g(a)}^{g(b)} f(u) du \\ &= \int_{\sin a}^{\sin b} u^5 du = \frac{\sin^6 b}{6} - \frac{\sin^6 a}{6}. \end{aligned}$$

The integration of $\int_a^b \tan x dx$ can be treated similarly if we write

$$\int_a^b \tan x dx = - \int_a^b \frac{-\sin x}{\cos x} dx.$$

In this case the factor $-\sin x$ is $g'(x)$, where $g(x) = \cos x$; the remaining factor $1/\cos x$ can then be written $f(\cos x)$ for $f(u) = 1/u$. Hence

$$\begin{aligned} & \int_a^b \tan x dx \quad \left[\begin{array}{l} g(x) = \cos x \\ f(u) = \frac{1}{u} \end{array} \right] \\ &= - \int_a^b f(g(x))g'(x) dx = - \int_{g(a)}^{g(b)} f(u) du \\ &= - \int_{\cos a}^{\cos b} \frac{1}{u} du = \log(\cos a) - \log(\cos b). \end{aligned}$$

Finally, to find

$$\int_a^b \frac{1}{x \log x} dx,$$

notice that $1/x = g'(x)$ where $g(x) = \log x$, and that $1/\log x = f(g(x))$ for $f(u) = 1/u$. Thus

$$\begin{aligned} & \int_a^b \frac{1}{x \log x} dx \quad \left[\begin{array}{l} g(x) = \log x \\ f(u) = \frac{1}{u} \end{array} \right] \\ &= \int_a^b f(g(x))g'(x) dx = \int_{g(a)}^{g(b)} f(u) du \\ &= \int_{\log a}^{\log b} \frac{1}{u} du = \log(\log b) - \log(\log a). \end{aligned}$$

Fortunately, these uses of the substitution formula can be shortened considerably. The intermediate steps, which involve writing

$$\int_a^b f(g(x))g'(x) dx = \int_{g(a)}^{g(b)} f(u) du,$$

can easily be eliminated by noticing the following: To go from the left side to the right side,

substitute $\begin{cases} u \text{ for } g(x) \\ du \text{ for } g'(x) dx \end{cases}$
(and change the limits of integration);

the substitutions can be performed directly on the original function (accounting for the name of this theorem). For example,

$$\int_a^b \sin^5 x \cos x dx \left[\text{substitute } \begin{array}{l} u \text{ for } \sin x \\ du \text{ for } \cos x dx \end{array} \right] = \int_{\sin a}^{\sin b} u^5 du,$$

and similarly

$$\int_a^b \frac{-\sin x}{\cos x} dx \left[\text{substitute } \begin{array}{l} u \text{ for } \cos x \\ du \text{ for } -\sin x dx \end{array} \right] = \int_{\cos a}^{\cos b} \frac{1}{u} du.$$

Usually we abbreviate this method even more, and say simply:

“Let $u = g(x)$
 $du = g'(x) dx$.”

Thus

$$\int_a^b \frac{1}{x \log x} dx \left[\begin{array}{l} \text{let } u = \log x \\ du = \frac{1}{x} dx \end{array} \right] = \int_{\log a}^{\log b} \frac{1}{u} du.$$

In this chapter we are usually interested in primitives rather than definite integrals, but if we can find $\int_a^b f(x) dx$ for all a and b , then we can certainly find

$\int f(x) dx$. For example, since

$$\int_a^b \sin^5 x \cos x dx = \frac{\sin^6 b}{6} - \frac{\sin^6 a}{6},$$

it follows that

$$\int \sin^5 x \cos x dx = \frac{\sin^6 x}{6}.$$

Similarly,

$$\begin{aligned}\int \tan x dx &= -\log \cos x, \\ \int \frac{1}{x \log x} dx &= \log(\log x).\end{aligned}$$

It is quite uneconomical to obtain primitives from the substitution formula by first finding definite integrals. Instead, the two steps can be combined, to yield the following procedure:

(1) Let

$$\begin{aligned}u &= g(x), \\ du &= g'(x) dx;\end{aligned}$$

(after this manipulation only the letter u should appear, *not* the letter x).

- (2) Find a primitive (as an expression involving u).
- (3) Substitute $g(x)$ back for u .

Thus, to find

$$\int \sin^5 x \cos x dx,$$

(1) let

$$\begin{aligned}u &= \sin x, \\ du &= \cos x dx\end{aligned}$$

so that we obtain

$$\int u^5 du;$$

(2) evaluate

$$\int u^5 du = \frac{u^6}{6};$$

(3) remember to substitute $\sin x$ back for u , so that

$$\int \sin^5 x \cos x dx = \frac{\sin^6 x}{6}.$$

Similarly, if

$$\begin{aligned} u &= \log x, \\ du &= \frac{1}{x} dx, \end{aligned}$$

then

$$\int \frac{1}{x \log x} dx \quad \text{becomes} \quad \int \frac{1}{u} du = \log u,$$

so that

$$\int \frac{1}{x \log x} dx = \log(\log x).$$

To evaluate

$$\int \frac{x}{1+x^2} dx,$$

let

$$\begin{aligned} u &= 1 + x^2, \\ du &= 2x dx; \end{aligned}$$

the factor 2 which has just popped up causes no problem—the integral becomes

$$\frac{1}{2} \int \frac{1}{u} du = \frac{1}{2} \log u,$$

so

$$\int \frac{x}{1+x^2} dx = \frac{1}{2} \log(1+x^2).$$

(This result may be combined with integration by parts to yield

$$\begin{aligned} \int 1 \cdot \arctan x dx &= x \arctan x - \int \frac{x}{1+x^2} dx \\ &= x \arctan x - \frac{1}{2} \log(1+x^2), \end{aligned}$$

a formula that has already been mentioned.)

These applications of the substitution formula* illustrate the most straightforward and least interesting types—once the suitable factor $g'(x)$ is recognized, the whole problem may even become simple enough to do mentally. The following three problems require only the information provided by the short table of indefinite integrals at the beginning of the chapter and, of course, the right substitution

*The substitution formula is often written in the form

$$\int f(u) du = \int f(g(x))g'(x) dx, \quad u = g(x).$$

This formula cannot be taken literally (after all, $\int f(u) du$ should mean a primitive of f and the symbol $\int f(g(x))g'(x) dx$ should mean a primitive of $(f \circ g) \cdot g'$; these are certainly not equal). However, it may be regarded as a symbolic summary of the procedure which we have developed. If we use Leibniz's notation, and a little fudging, the formula reads particularly well:

$$\int f(u) du = \int f(u) \frac{du}{dx} dx.$$

(the third problem has been disguised a little by some algebraic chicanery).

$$\begin{aligned} & \int \sec^2 x \tan^5 x \, dx, \\ & \int (\cos x) e^{\sin x} \, dx, \\ & \int \frac{e^x}{\sqrt{1 - e^{2x}}} \, dx. \end{aligned}$$

If you have not succeeded in finding the right substitutions, you should be able to guess them from the answers, which are $(\tan^6 x)/6$, $e^{\sin x}$, and $\arcsin e^x$. At first you may find these problems too hard to do in your head, but at least when g is of the very simple form $g(x) = ax + b$ you should not have to waste time writing out the substitution. The following integrations should all be clear. (The only worrisome detail is the proper positioning of the constant—should the answer to the second be $e^{3x}/3$ or $3e^{3x}$? I always take care of these problems as follows. Clearly $\int e^{3x} \, dx = e^{3x}$. (something). Now if I differentiate $F(x) = e^{3x}$, I get $F'(x) = 3e^{3x}$, so the “something” must be $\frac{1}{3}$, to cancel the 3.)

$$\begin{aligned} & \int \frac{dx}{x+3} = \log(x+3), \\ & \int e^{3x} \, dx = \frac{e^{3x}}{3}, \\ & \int \cos 4x \, dx = \frac{\sin 4x}{4}, \\ & \int \sin(2x+1) \, dx = \frac{-\cos(2x+1)}{2}, \\ & \int \frac{dx}{1+4x^2} = \frac{\arctan 2x}{2}. \end{aligned}$$

More interesting uses of the substitution formula occur when the factor $g'(x)$ does *not* appear. There are two main types of substitutions where this happens. Consider first

$$\int \frac{1+e^x}{1-e^x} \, dx.$$

The prominent appearance of the expression e^x suggests the simplifying substitution

$$\begin{aligned} u &= e^x, \\ du &= e^x \, dx. \end{aligned}$$

Although the expression $e^x \, dx$ does not appear, it can always be put in:

$$\int \frac{1+e^x}{1-e^x} \, dx = \int \frac{1+e^x}{1-e^x} \cdot \frac{1}{e^x} \cdot e^x \, dx.$$

We therefore obtain

$$\int \frac{1+u}{1-u} \cdot \frac{1}{u} \, du,$$

which can be evaluated by the algebraic trick

$$\int \frac{1+u}{1-u} \cdot \frac{1}{u} du = \int \frac{2}{1-u} + \frac{1}{u} du = -2 \log(1-u) + \log u,$$

so that

$$\int \frac{1+e^x}{1-e^x} dx = -2 \log(1-e^x) + \log e^x = -2 \log(1-e^x) + x.$$

There is an alternative and preferable way of handling this problem, which does not require multiplying and dividing by e^x . If we write

$$\begin{aligned} u &= e^x, & x &= \log u, \\ dx &= \frac{1}{u} du, \end{aligned}$$

then

$$\int \frac{1+e^x}{1-e^x} dx \text{ immediately becomes } \int \frac{1+u}{1-u} \cdot \frac{1}{u} du.$$

Most substitution problems are much easier if one resorts to this trick of expressing x in terms of u , and dx in terms of du , instead of vice versa. It is not hard to see why this trick always works (as long as the function expressing u in terms of x is one-one for all x under consideration): If we apply the substitution

$$\begin{aligned} u &= g(x), & x &= g^{-1}(u) \\ dx &= (g^{-1})'(u) du \end{aligned}$$

to the integral

$$\int f(g(x)) dx,$$

we obtain

$$(1) \quad \int f(u)(g^{-1})'(u) du.$$

On the other hand, if we apply the straightforward substitution

$$\begin{aligned} u &= g(x) \\ du &= g'(x) dx \end{aligned}$$

to the same integral,

$$\int f(g(x)) dx = \int f(g(x)) \cdot \frac{1}{g'(x)} \cdot g'(x) dx,$$

we obtain

$$(2) \quad \int f(u) \cdot \frac{1}{g'(g^{-1}(u))} du.$$

The integrals (1) and (2) are identical, since $(g^{-1})'(u) = 1/g'(g^{-1}(u))$.

As another concrete example, consider

$$\int \frac{e^{2x}}{\sqrt{e^x + 1}} dx.$$

In this case we will go the whole hog and replace the entire expression $\sqrt{e^x + 1}$ by one letter. Thus we choose the substitution

$$\begin{aligned} u &= \sqrt{e^x + 1}, \\ u^2 &= e^x + 1, \\ u^2 - 1 &= e^x, \quad x = \log(u^2 - 1), \\ dx &= \frac{2u}{u^2 - 1} du. \end{aligned}$$

The integral then becomes

$$\int \frac{(u^2 - 1)^2}{u} \cdot \frac{2u}{u^2 - 1} du = 2 \int u^2 - 1 du = \frac{2u^3}{3} - 2u.$$

Thus

$$\int \frac{e^{2x}}{\sqrt{e^x + 1}} dx = \frac{2}{3}(e^x + 1)^{3/2} - 2(e^x + 1)^{1/2}.$$

Another example, which illustrates the second main type of substitution that can occur, is the integral

$$\int \sqrt{1 - x^2} dx.$$

In this case, instead of replacing a complicated expression by a simpler one, we will replace x by $\sin u$, because $\sqrt{1 - \sin^2 u} = \cos u$. This really means that we are using the substitution $u = \arcsin x$, but it is the expression for x in terms of u which helps us find the expression to be substituted for dx . Thus,

$$\begin{aligned} \text{let } x &= \sin u, \quad [u = \arcsin x] \\ dx &= \cos u du; \end{aligned}$$

then the integral becomes

$$\int \sqrt{1 - \sin^2 u} \cos u du = \int \cos^2 u du.$$

The evaluation of this integral depends on the equation

$$\cos^2 u = \frac{1 + \cos 2u}{2}$$

(see the discussion of trigonometric functions below) so that

$$\int \cos^2 u du = \int \frac{1 + \cos 2u}{2} du = \frac{u}{2} + \frac{\sin 2u}{4},$$

and

$$\begin{aligned} \int \sqrt{1 - x^2} dx &= \frac{\arcsin x}{2} + \frac{\sin(2 \arcsin x)}{4} \\ &= \frac{\arcsin x}{2} + \frac{1}{2} \sin(\arcsin x) \cdot \cos(\arcsin x) \\ &= \frac{\arcsin x}{2} + \frac{1}{2} x \sqrt{1 - x^2}. \end{aligned}$$

Substitution and integration by parts are the only fundamental methods which you have to learn; with their aid primitives can be found for a large number of functions. Nevertheless, as some of our examples reveal, success often depends upon some additional tricks. The most important are listed below. Using these you should be able to integrate all the functions in Problems 1 to 9 (a few other interesting tricks are explained in some of the remaining problems).

1. TRIGONOMETRIC FUNCTIONS

Since

$$\sin^2 x + \cos^2 x = 1$$

and

$$\cos 2x = \cos^2 x - \sin^2 x,$$

we obtain

$$\begin{aligned}\cos 2x &= \cos^2 x - (1 - \cos^2 x) = 2\cos^2 x - 1, \\ \cos 2x &= (1 - \sin^2 x) - \sin^2 x = 1 - 2\sin^2 x,\end{aligned}$$

or

$$\begin{aligned}\sin^2 x &= \frac{1 - \cos 2x}{2}, \\ \cos^2 x &= \frac{1 + \cos 2x}{2}.\end{aligned}$$

These formulas may be used to integrate

$$\int \sin^n x \, dx,$$

$$\int \cos^n x \, dx,$$

if n is even. Substituting

$$\frac{(1 - \cos 2x)}{2} \quad \text{or} \quad \frac{(1 + \cos 2x)}{2}$$

for $\sin^2 x$ or $\cos^2 x$ yields a sum of terms involving lower powers of cos. For example,

$$\int \sin^4 x \, dx = \int \left(\frac{1 - \cos 2x}{2} \right)^2 \, dx = \int \frac{1}{4} \, dx - \frac{1}{2} \int \cos 2x \, dx + \frac{1}{4} \int \cos^2 2x \, dx$$

and

$$\int \cos^2 2x \, dx = \int \frac{1 + \cos 4x}{2} \, dx.$$

If n is odd, $n = 2k + 1$, then

$$\int \sin^n x \, dx = \int \sin x (1 - \cos^2 x)^k \, dx;$$

the latter expression, multiplied out, involves terms of the form $\sin x \cos^l x$, all of which can be integrated easily. The integral for $\cos^n x$ is treated similarly. An integral

$$\int \sin^n x \cos^m x \, dx$$

is handled the same way if n or m is odd. If n and m are both even, use the formulas for $\sin^2 x$ and $\cos^2 x$.

A final important trigonometric integral is

$$\int \frac{1}{\cos x} \, dx = \int \sec x \, dx = \log(\sec x + \tan x).$$

Although there are several ways of “deriving” this result, by means of the methods already at our disposal (Problem 12), it is simplest to check this formula by differentiating the right side, and to memorize it.

2. REDUCTION FORMULAS

Integration by parts yields (Problem 20)

$$\begin{aligned}\int \sin^n x \, dx &= -\frac{1}{n} \sin^{n-1} x \cos x + \frac{n-1}{n} \int \sin^{n-2} x \, dx, \\ \int \cos^n x \, dx &= \frac{1}{n} \cos^{n-1} x \sin x + \frac{n-1}{n} \int \cos^{n-2} x \, dx, \\ \int \frac{1}{(x^2 + 1)^n} \, dx &= \frac{1}{2n-2} \frac{x}{(x^2 + 1)^{n-1}} + \frac{2n-3}{2n-2} \int \frac{1}{(x^2 + 1)^{n-1}} \, dx\end{aligned}$$

and many similar formulas. The first two, used repeatedly, give a different method for evaluating primitives of \sin^n or \cos^n . The third is very important for integrating a large general class of functions, which will complete our discussion.

3. RATIONAL FUNCTIONS

Consider a rational function p/q where

$$\begin{aligned}p(x) &= a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0, \\ q(x) &= b_m x^m + b_{m-1} x^{m-1} + \cdots + b_0.\end{aligned}$$

We might as well assume that $a_n = b_m = 1$. Moreover, we can assume that $n < m$, for otherwise we may express p/q as a polynomial function plus a rational function which is of this form by dividing (the calculation

$$\frac{u^2}{u-1} = u + 1 + \frac{1}{u-1}$$

is a simple example). The integration of an arbitrary rational function depends on two facts; the first follows from the “Fundamental Theorem of Algebra” (see Chapter 26, Theorem 2 and Problem 26-3), but the second will not be proved in this book.

THEOREM Every polynomial function

$$q(x) = x^m + b_{m-1}x^{m-1} + \cdots + b_0$$

can be written as a product

$$q(x) = (x - \alpha_1)^{r_1} \cdots (x - \alpha_k)^{r_k} (x^2 + \beta_1 x + \gamma_1)^{s_1} \cdots (x^2 + \beta_l x + \gamma_l)^{s_l}$$

(where $r_1 + \cdots + r_k + 2(s_1 + \cdots + s_l) = m$).

(In this expression, identical factors have been collected together, so that all $x - \alpha_i$ and $x^2 + \beta_i x + \gamma_i$ may be assumed distinct. Moreover, we assume that each quadratic factor cannot be factored further. This means that

$$\beta_i^2 - 4\gamma_i < 0,$$

since otherwise we can factor

$$x^2 + \beta_i x + \gamma_i = \left[x - \left(\frac{-\beta_i + \sqrt{\beta_i^2 - 4\gamma_i}}{2} \right) \right] \cdot \left[x - \left(\frac{-\beta_i - \sqrt{\beta_i^2 - 4\gamma_i}}{2} \right) \right]$$

into linear factors.)

THEOREM If $n < m$ and

$$\begin{aligned} p(x) &= x^n + a_{n-1}x^{n-1} + \cdots + a_0, \\ q(x) &= x^m + b_{m-1}x^{m-1} + \cdots + b_0 \\ &= (x - \alpha_1)^{r_1} \cdots (x - \alpha_k)^{r_k} (x^2 + \beta_1 x + \gamma_1)^{s_1} \cdots (x^2 + \beta_l x + \gamma_l)^{s_l}, \end{aligned}$$

then $p(x)/q(x)$ can be written in the form

$$\begin{aligned} \frac{p(x)}{q(x)} &= \left[\frac{a_{1,1}}{(x - \alpha_1)} + \cdots + \frac{a_{1,r_1}}{(x - \alpha_1)^{r_1}} \right] + \cdots \\ &\quad + \left[\frac{\alpha_{k,1}}{(x - \alpha_k)} + \cdots + \frac{\alpha_{k,r_k}}{(x - \alpha_k)^{r_k}} \right] \\ &\quad + \left[\frac{b_{1,1}x + c_{1,1}}{(x^2 + \beta_1 x + \gamma_1)} + \cdots + \frac{b_{1,s_1}x + c_{1,s_1}}{(x^2 + \beta_1 x + \gamma_1)^{s_1}} \right] + \cdots \\ &\quad + \left[\frac{b_{l,1}x + c_{l,1}}{(x^2 + \beta_l x + \gamma_l)} + \cdots + \frac{b_{l,s_l}x + c_{l,s_l}}{(x^2 + \beta_l x + \gamma_l)^{s_l}} \right]. \end{aligned}$$

This expression, known as the “partial fraction decomposition” of $p(x)/q(x)$, is so complicated that it is simpler to examine the following example, which illustrates such an expression and shows how to find it. According to the theorem, it is possible to write

$$\begin{aligned} &\frac{2x^7 + 8x^6 + 13x^5 + 20x^4 + 15x^3 + 16x^2 + 7x + 10}{(x^2 + x + 1)^2(x^2 + 2x + 2)(x - 1)^2} \\ &= \frac{a}{x - 1} + \frac{b}{(x - 1)^2} + \frac{cx + d}{x^2 + 2x + 2} + \frac{ex + f}{x^2 + x + 1} + \frac{gx + h}{(x^2 + x + 1)^2}. \end{aligned}$$

To find the numbers a, b, c, d, e, f, g , and h , write the right side as a polynomial over the common denominator $(x^2 + x + 1)^2(x^2 + 2x + 3)(x - 1)^2$; the numerator becomes

$$\begin{aligned} & a(x-1)(x^2+2x+2)(x^2+x+1)^2 + b(x^2+2x+2)(x^2+x+1)^2 \\ & + (cx+d)(x-1)^2(x^2+x+1)^2 + (ex+f)(x-1)^2(x^2+2x+2)(x^2+x+1) \\ & \quad + (gx+h)(x-1)^2(x^2+2x+2). \end{aligned}$$

Actually multiplying this out (!) we obtain a polynomial of degree 8, whose coefficients are combinations of a, \dots, h . Equating these coefficients with the coefficients of $2x^7 + 8x^6 + 13x^5 + 20x^4 + 15x^3 + 16x^2 + 7x + 10$ (the coefficient of x^8 is 0) we obtain 8 equations in the eight unknowns a, \dots, h . After heroic calculations these can be solved to give

$$\begin{aligned} a &= 1, & b &= 2, & c &= 1, & d &= 3, \\ e &= 0, & f &= 0, & g &= 0, & h &= 1. \end{aligned}$$

Thus

$$\begin{aligned} & \int \frac{2x^7 + 8x^6 + 13x^5 + 20x^4 + 15x^3 + 16x^2 + 7x + 10}{(x^2 + x + 1)^2(x^2 + 2x + 3)(x - 1)^2} dx \\ & = \int \frac{1}{(x-1)} dx + \int \frac{2}{(x-1)^2} dx + \int \frac{1}{(x^2+x+1)^2} dx + \int \frac{x+3}{x^2+2x+2} dx. \end{aligned}$$

(In simpler cases the requisite calculations may actually be feasible. I obtained this particular example by *starting* with the partial fraction decomposition and converting it into one fraction.)

We are already in a position to find each of the integrals appearing in the above expression; the calculations will illustrate all the difficulties which arise in integrating rational functions.

The first two integrals are simple:

$$\begin{aligned} \int \frac{1}{x-1} dx &= \log(x-1), \\ \int \frac{2}{(x-1)^2} dx &= \frac{-2}{x-1}. \end{aligned}$$

The third integration depends on “completing the square”:

$$\begin{aligned} x^2 + x + 1 &= (x + \frac{1}{2})^2 + \frac{3}{4} \\ &= \frac{3}{4} \left[\left(\frac{x + \frac{1}{2}}{\sqrt{\frac{3}{4}}} \right)^2 + 1 \right]. \end{aligned}$$

(If we had obtained $-\frac{3}{4}$ instead of $\frac{3}{4}$ we could not take the square root, but in this case our original quadratic factor could have been factored into linear factors.) We

can now write

$$\int \frac{1}{(x^2 + x + 1)^2} dx = \frac{16}{9} \int \frac{1}{\left[\left(\frac{x + \frac{1}{2}}{\sqrt{\frac{3}{4}}} \right) + 1 \right]^2} dx.$$

The substitution

$$u = \frac{x + \frac{1}{2}}{\sqrt{\frac{3}{4}}},$$

$$du = \frac{1}{\sqrt{\frac{3}{4}}} dx,$$

changes this integral to

$$\frac{16}{9} \int \frac{\sqrt{\frac{3}{4}}}{(u^2 + 1)^2} du,$$

which can be computed using the third reduction formula given above.

Finally, to evaluate

$$\int \frac{x + 3}{(x^2 + 2x + 2)} dx$$

we write

$$\int \frac{x + 3}{x^2 + 2x + 2} dx = \frac{1}{2} \int \frac{2x + 2}{x^2 + 2x + 2} dx + \int \frac{2}{(x + 1)^2 + 1} dx.$$

The first integral on the right side has been purposely constructed so that we can evaluate it by using the substitution

$$u = x^2 + 2x + 2,$$

$$du = (2x + 2) dx$$

The second integral on the right, which is just the difference of the other two, is simply $2 \arctan(x + 1)$. If the original integral were

$$\int \frac{x + 3}{(x^2 + 2x + 2)^n} dx = \frac{1}{2} \int \frac{2x + 2}{(x^2 + 2x + 2)^n} dx + \int \frac{2}{[(x + 1)^2 + 1]^n} dx,$$

the first integral on the right would still be evaluated by the same substitution. The second integral would be evaluated by means of a reduction formula.

This example has probably convinced you that integration of rational functions is a theoretical curiosity only, especially since it is necessary to find the factorization of $q(x)$ before you can even begin. This is only partly true. We have already seen that simple rational functions sometimes arise, as in the integration

$$\int \frac{1 + e^x}{1 - e^x} dx;$$

another important example is the integral

$$\int \frac{1}{x^2 - 1} dx = \int \frac{\frac{1}{2}}{x - 1} - \frac{\frac{1}{2}}{x + 1} dx = \frac{1}{2} \log(x - 1) - \frac{1}{2} \log(x + 1).$$

Moreover, if a problem has been reduced to the integration of a rational function, it is then certain that an elementary primitive exists, even when the difficulty or impossibility of finding the factors of the denominator may preclude writing this primitive explicitly.

PROBLEMS

- This problem contains some integrals which require little more than algebraic manipulation, and consequently test your ability to discover algebraic tricks, rather than your understanding of the integration processes. Nevertheless, any one of these tricks might be an important preliminary step in an honest integration problem. Moreover, you want to have some feel for which integrals are easy, so that you can see when the end of an integration process is in sight. The answer section, if you resort to it, will only reveal what algebra you should have used.

$$(i) \int \frac{\sqrt[5]{x^3} + \sqrt[5]{x}}{\sqrt{x}} dx.$$

$$(ii) \int \frac{dx}{\sqrt{x-1} + \sqrt{x+1}}.$$

$$(iii) \int \frac{e^x + e^{2x} + e^{3x}}{e^{4x}} dx.$$

$$(iv) \int \frac{a^x}{b^x} dx.$$

$$(v) \int \tan^2 x dx. \text{ (Trigonometric integrals are always very touchy, because there are so many trigonometric identities that an easy problem can easily look hard.)}$$

$$(vi) \int \frac{dx}{a^2 + x^2}.$$

$$(vii) \int \frac{dx}{\sqrt{a^2 - x^2}}.$$

$$(viii) \int \frac{dx}{1 + \sin x}.$$

$$(ix) \int \frac{8x^2 + 6x + 4}{x + 1} dx.$$

$$(x) \int \frac{1}{\sqrt{2x - x^2}} dx.$$

- The following integrations involve simple substitutions, most of which you should be able to do in your head.

$$(i) \int e^x \sin e^x dx.$$

- (ii) $\int xe^{-x^2} dx.$
- (iii) $\int \frac{\log x}{x} dx.$ (In the text this was done by parts.)
- (iv) $\int \frac{e^x dx}{e^{2x} + 2e^x + 1}.$
- (v) $\int e^{e^x} e^x dx.$
- (vi) $\int \frac{x dx}{\sqrt{1 - x^4}}.$
- (vii) $\int \frac{e^{\sqrt{x}}}{\sqrt{x}} dx.$
- (viii) $\int x \sqrt{1 - x^2} dx.$
- (ix) $\int \log(\cos x) \tan x dx.$
- (x) $\int \frac{\log(\log x)}{x \log x} dx.$

3. Integration by parts.

- (i) $\int x^2 e^x dx.$
- (ii) $\int x^3 e^{x^2} dx.$
- (iii) $\int e^{ax} \sin bx dx.$
- (iv) $\int x^2 \sin x dx.$
- (v) $\int (\log x)^3 dx.$
- (vi) $\int \frac{\log(\log x)}{x} dx.$
- (vii) $\int \sec^3 x dx.$ (This is a tricky and important integral that often comes up. If you do not succeed in evaluating it, be sure to consult the answers.)
- (viii) $\int \cos(\log x) dx.$
- (ix) $\int \sqrt{x} \log x dx.$
- (x) $\int x(\log x)^2 dx.$

4. The following integrations can all be done with substitutions of the form $x = \sin u$, $x = \cos u$, etc. To do some of these you will need to remember that

$$\int \sec x \, dx = \log(\sec x + \tan x)$$

as well as the following formula, which can also be checked by differentiation:

$$\int \csc x \, dx = -\log(\csc x + \cot x).$$

In addition, at this point the derivatives of all the trigonometric functions should be kept handy.

- (i) $\int \frac{dx}{\sqrt{1-x^2}}$. (You already know this integral, but use the substitution $x = \sin u$ anyway, just to see how it works out.)
- (ii) $\int \frac{dx}{\sqrt{1+x^2}}$. (Since $\tan^2 u + 1 = \sec^2 u$, you want to use the substitution $x = \tan u$.)
- (iii) $\int \frac{dx}{\sqrt{x^2-1}}$.
- (iv) $\int \frac{dx}{x\sqrt{x^2-1}}$. (The answer will be a certain inverse function that was given short shrift in the text.)
- (v) $\int \frac{dx}{x\sqrt{1-x^2}}$.
- (vi) $\int \frac{dx}{x\sqrt{1+x^2}}$.
- (vii) $\int x^3 \sqrt{1-x^2} \, dx.$
- (viii) $\int \sqrt{1-x^2} \, dx.$
- (ix) $\int \sqrt{1+x^2} \, dx.$
- (x) $\int \sqrt{x^2-1} \, dx.$

You will need to remember the methods for integrating powers of sin and cos.

5. The following integrations involve substitutions of various types. There is no substitute for cleverness, but there is a general rule to follow: substitute for an expression which appears frequently or prominently; if two different troublesome expressions appear, try to express them both in terms of some new expression. And don't forget that it usually helps to express x directly in terms of u , to find out the proper expression to substitute for dx .

- (i) $\int \frac{dx}{1+\sqrt{x+1}}$.
- (ii) $\int \frac{dx}{1+e^x}$.

(iii) $\int \frac{dx}{\sqrt{x} + \sqrt[3]{x}}.$

(iv) $\int \frac{dx}{\sqrt{1 + e^x}}.$ (The substitution $u = e^x$ leads to an integral requiring yet another substitution; this is all right, but both substitutions can be done at once.)

(v) $\int \frac{dx}{2 + \tan x}.$

(vi) $\int \frac{dx}{\sqrt{\sqrt{x} + 1}}.$ (Another place where one substitution can be made to do the work of two.)

(vii) $\int \frac{4^x + 1}{2^x + 1} dx.$

(viii) $\int e^{\sqrt{x}} dx.$

(ix) $\int \frac{\sqrt{1-x}}{1-\sqrt{x}} dx.$ (In this case two successive substitutions work out best; there are two obvious candidates for the first substitution, and either will work.)

*(x) $\int \sqrt{\frac{x-1}{x+1}} \cdot \frac{1}{x^2} dx.$

6. The previous problem provided gratis a haphazard selection of rational functions to be integrated. Here is a more systematic selection.

(i) $\int \frac{2x^2 + 7x - 1}{x^3 + x^2 - x - 1} dx.$

(ii) $\int \frac{2x + 1}{x^3 - 3x^2 + 3x - 1} dx.$

(iii) $\int \frac{x^3 + 7x^2 - 5x + 5}{(x-1)^2(x+1)^3} dx.$

(iv) $\int \frac{2x^2 + x + 1}{(x+3)(x-1)^2} dx.$

(v) $\int \frac{x+4}{x^2+1} dx.$

(vi) $\int \frac{x^3 + x + 2}{x^4 + 2x^2 + 1} dx.$

(vii) $\int \frac{3x^2 + 3x + 1}{x^3 + 2x^2 + 2x + 1} dx.$

(viii) $\int \frac{dx}{x^4 + 1}.$

(ix) $\int \frac{2x}{(x^2 + x + 1)^2} dx.$

(x) $\int \frac{3x}{(x^2 + x + 1)^3} dx.$

***7.** Potpourri. (No holds barred.) The following integrations involve all the methods of the previous problems

- (i) $\int \frac{\arctan x}{1+x^2} dx.$
- (ii) $\int \frac{x \arctan x}{(1+x^2)^3} dx.$
- (iii) $\int \log \sqrt{1+x^2} dx.$
- (iv) $\int x \log \sqrt{1+x^2} dx.$
- (v) $\int \frac{x^2-1}{x^2+1} \cdot \frac{1}{\sqrt{1+x^4}} dx.$
- (vi) $\int \arcsin \sqrt{x} dx.$
- (vii) $\int \frac{x}{1+\sin x} dx.$
- (viii) $\int e^{\sin x} \cdot \frac{x \cos^3 x - \sin x}{\cos^2 x} dx.$
- (ix) $\int \sqrt{\tan x} dx.$
- (x) $\int \frac{dx}{x^6+1}. \text{(To factor } x^6+1, \text{ first factor } y^3+1, \text{ using Problem 1-1.)}$

The following two problems provide still more practice at integration, if you need it (and can bear it). Problem 8 involves algebraic and trigonometric manipulations and integration by parts, while Problem 9 involves substitutions. (Of course, in many cases the resulting integrals will require still further manipulations.)

8. Find the following integrals.

- (i) $\int \log(a^2+x^2) dx.$
- (ii) $\int \frac{1+\cos x}{\sin^2 x} dx.$
- (iii) $\int \frac{x+1}{\sqrt{4-x^2}} dx.$
- (iv) $\int x \arctan x dx.$
- (v) $\int \sin^3 x dx.$
- (vi) $\int \frac{\sin^3 x}{\cos^2 x} dx.$

(vii) $\int x^2 \arctan x \, dx.$

(viii) $\int \frac{x \, dx}{\sqrt{x^2 - 2x + 2}}.$

(ix) $\int \sec^3 x \tan x \, dx.$

(x) $\int x \tan^2 x \, dx.$

9. Find the following integrals.

(i) $\int \frac{dx}{(a^2 + x^2)^2}.$

(ii) $\int \sqrt{1 - \sin x} \, dx.$

(iii) $\int \arctan \sqrt{x} \, dx.$

(iv) $\int \sin \sqrt{x+1} \, dx.$

(v) $\int \frac{\sqrt{x^3 - 2}}{x} \, dx.$

(vi) $\int \log(x + \sqrt{x^2 - 1}) \, dx.$

(vii) $\int \log(x + \sqrt{x}) \, dx.$

(viii) $\int \frac{dx}{x - x^{3/5}}.$

(ix) $\int (\arcsin x)^2 \, dx.$

(x) $\int x^5 \arctan(x^2) \, dx.$

10. If you have done Problem 18-9, the integrals (ii) and (iii) in Problem 4 will look very familiar. In general, the substitution $x = \cosh u$ often works for integrals involving $\sqrt{x^2 - 1}$, while $x = \sinh u$ is the thing to try for integrals involving $\sqrt{x^2 + 1}$. Try these substitutions on the other integrals in Problem 4. (The method is not really recommended; it is easier to stick with trigonometric substitutions.)
- *11. The world's sneakiest substitution is undoubtedly

$$t = \tan \frac{x}{2}, \quad x = 2 \arctan t,$$

$$dx = \frac{2}{1+t^2} dt.$$

As we found in Problem 15-17, this substitution leads to the expressions

$$\sin x = \frac{2t}{1+t^2}, \quad \cos x = \frac{1-t^2}{1+t^2}.$$

This substitution thus transforms any integral which involves only sin and cos, combined by addition, multiplication, and division, into the integral of a rational function. Find

- (i) $\int \frac{dx}{1+\sin x}$. (Compare your answer with Problem 1(viii).)
- (ii) $\int \frac{dx}{1-\sin^2 x}$. (In this case it is better to let $t = \tan x$. Why?)
- (iii) $\int \frac{dx}{a \sin x + b \cos x}$. (There is also another way to do this, using Problem 15-8.)
- (iv) $\int \sin^2 x \, dx$. (An exercise to convince you that this substitution should be used only as a last resort.)
- (v) $\int \frac{dx}{3+5 \sin x}$. (A last resort.)

*12. Derive the formula for $\int \sec x \, dx$ in the following two ways:

- (a) By writing

$$\begin{aligned}\frac{1}{\cos x} &= \frac{\cos x}{\cos^2 x} \\ &= \frac{\cos x}{1 - \sin^2 x} \\ &= \frac{1}{2} \left[\frac{\cos x}{1 + \sin x} + \frac{\cos x}{1 - \sin x} \right],\end{aligned}$$

an expression obviously inspired by partial fraction decompositions. Be sure to note that $\int \cos x / (1 - \sin x) \, dx = -\log(1 - \sin x)$; the minus sign is very important. And remember that $\frac{1}{2} \log \alpha = \log \sqrt{\alpha}$. From there on, keep doing algebra, and trust to luck.

- (b) By using the substitution $t = \tan x/2$. One again, quite a bit of manipulation is required to put the answer in the desired form; the expression $\tan x/2$ can be attacked by using Problem 15-9, or both answers can be expressed in terms of t . There is another expression for $\int \sec x \, dx$, which is less cumbersome than $\log(\sec x + \tan x)$; using Problem 15-9, we obtain

$$\int \sec x \, dx = \log \left(\frac{1 + \tan \frac{x}{2}}{1 - \tan \frac{x}{2}} \right) = \log \left(\tan \left(\frac{x}{2} + \frac{\pi}{4} \right) \right).$$

This last expression was actually the one first discovered, and was due, not to any mathematician's cleverness, but to a curious historical acci-

dent: In 1599 Wright computed nautical tables that amounted to definite integrals of sec. When the first tables for the logarithms of tangents were produced, the correspondence between the two tables was immediately noticed (but remained unexplained until the invention of calculus).

13. The derivation of $\int e^x \sin x \, dx$ given in the text seems to prove that the only primitive of $f(x) = e^x \sin x$ is $F(x) = e^x(\sin x - \cos x)/2$, whereas $F(x) = e^x(\sin x - \cos x)/2 + C$ is also a primitive for any number C . Where does C come from? (What is the meaning of the equation

$$\int e^x \sin x \, dx = e^x \sin x - e^x \cos x - \int e^x \sin x \, dx?$$

14. Suppose that f'' is continuous and that

$$\int_0^\pi [f(x) + f''(x)] \sin x \, dx = 2.$$

Given that $f(\pi) = 1$, compute $f(0)$.

15. (a) Find $\int \arcsin x \, dx$, using the same trick that worked for log and arctan.
 *(b) Generalize this trick: Find $\int f^{-1}(x) \, dx$ in terms of $\int f(x) \, dx$. Compare with Problems 12-18 and 14-17.
16. (a) Find $\int \sin^4 x \, dx$ in two different ways: first using the reduction formula, and then using the formula for $\sin^2 x$.
 (b) Combine your answers to obtain an impressive trigonometric identity.
17. Express $\int \log(\log x) \, dx$ in terms of $\int (\log x)^{-1} \, dx$. (Neither is expressible in terms of elementary functions.)
18. Express $\int x^2 e^{-x^2} \, dx$ in terms of $\int e^{-x^2} \, dx$.
19. Prove that the function $f(x) = e^x/(e^{5x} + e^x + 1)$ has an elementary primitive. (Do not try to find it!)
20. Prove the reduction formulas in the text. For the third one write

$$\int \frac{dx}{(1+x^2)^n} = \int \frac{dx}{(1+x^2)^{n-1}} - \int \frac{x^2 \, dx}{(1+x^2)^n}$$

and work on the last integral. (Another possibility is to use the substitution $x = \tan u$.)

21. Find a reduction formula for

$$(a) \int x^n e^x \, dx$$

$$(b) \int (\log x)^n \, dx.$$

- *22. Prove that

$$\int_1^{\cosh x} \sqrt{t^2 - 1} \, dt = \frac{\cosh x \sinh x}{2} - \frac{x}{2}.$$

(See Problem 18-6 for the significance of this computation.)

23. Prove that

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

(A geometric interpretation makes this clear, but it is also a good exercise in the handling of limits of integration during a substitution.)

24. Prove that the area of a circle of radius r is πr^2 . (Naturally you must remember that π is defined as the area of the unit circle.)

25. Let ϕ be a nonnegative integrable function such that $\phi(x) = 0$ for $|x| \geq 1$ and such that $\int_{-1}^1 \phi = 1$. For $h > 0$, let

$$\phi_h(x) = \frac{1}{h} \phi(x/h).$$

- (a) Show that $\phi_h(x) = 0$ for $|x| \geq h$ and that $\int_{-h}^h \phi_h = 1$.

- (b) Let f be integrable on $[-1, 1]$ and continuous at 0. Show that

$$\lim_{h \rightarrow 0^+} \int_{-1}^1 \phi_h f = \lim_{h \rightarrow 0^+} \int_{-h}^h \phi_h f = f(0).$$

- (c) Show that

$$\lim_{h \rightarrow 0^+} \int_{-1}^1 \frac{h}{h^2 + x^2} dx = \pi.$$

The final part of this problem might appear, at first sight, to be an exact analogue of part (b), but it actually requires more careful argument.

- (d) Let f be integrable on $[-1, 1]$ and continuous at 0. Show that

$$\lim_{h \rightarrow 0^+} \int_{-1}^1 \frac{h}{h^2 + x^2} f(x) dx = \pi f(0).$$

Hint: If h is small, then $h/(h^2 + x^2)$ will be small on most of $[-1, 1]$.

The next two problems use the formula

$$\frac{1}{2} \int_{\theta_0}^{\theta_1} f(\theta)^2 d\theta,$$

derived in Problem 13-24, for the area of a region bounded by the graph of f in polar coordinates.

26. For each of the following functions, find the area bounded by the graphs in polar coordinates. (Be careful about the proper range for θ , or you will get nonsensical results!)

- (i) $f(\theta) = a \sin \theta$.
- (ii) $f(\theta) = 2 + \cos \theta$.
- (iii) $f(\theta)^2 = 2a^2 \cos 2\theta$.
- (iv) $f(\theta) = a \cos 2\theta$.

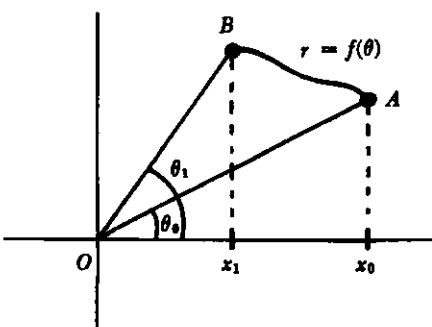


FIGURE 1

27. Figure 1 shows the graph of f in polar coordinates; the region OAB thus has area $\frac{1}{2} \int_{\theta_0}^{\theta_1} f(\theta)^2 d\theta$. Now suppose that this graph also happens to be the ordinary graph of some function g . Then the region OAB also has area

$$\text{area } \Delta Ox_1B + \int_{x_1}^{x_0} g - \text{area } \Delta Ox_0A.$$

Prove analytically that these two numbers are indeed the same. Hint: The function g is determined by the equations

$$x = f(\theta) \cos \theta, \quad g(x) = f(\theta) \sin \theta.$$

The next four problems use the formulas, derived in Problems 3 and 4 of the Appendix to Chapter 13, for the length of a curve represented parametrically (and, in particular, as the graph of a function in polar coordinates).

28. Let c be a curve represented parametrically by u and v on $[a, b]$, and let h be an increasing function with $h(\bar{a}) = a$ and $h(\bar{b}) = b$. Then on $[\bar{a}, \bar{b}]$ the functions $\bar{u} = u \circ h$, $\bar{v} = v \circ h$ give a parametric representation of another curve \bar{c} ; intuitively, \bar{c} is just the same curve c traversed at a different rate.

- (a) Show, directly from the definition of length, that the length of c on $[a, b]$ equals the length of \bar{c} on $[\bar{a}, \bar{b}]$.
- (b) Assuming differentiability of any functions required, show that the lengths are equal by using the integral formula for length, and the appropriate substitution.

29. Find the length of the following curves, all described as the graphs of functions, except for (iii), which is represented parametrically.

(i) $f(x) = \frac{1}{3}(x^2 + 2)^{3/2}$, $0 \leq x \leq 1$.

(ii) $f(x) = x^3 + \frac{1}{12x}$, $1 \leq x \leq 2$.

(iii) $x = a^3 \cos^3 t$, $y = a^3 \sin^3 t$, $0 \leq t \leq 2\pi$.

(iv) $f(x) = \log(\cos x)$, $0 \leq x \leq \pi/6$.

(v) $f(x) = \log x$, $1 \leq x \leq e$.

(vi) $f(x) = \arcsin e^x$, $-\log 2 \leq x \leq 0$.

30. For the following functions, find the length of the graph in polar coordinates.

(i) $f(\theta) = a \cos \theta$.

(ii) $f(\theta) = a(1 - \cos \theta)$.

(iii) $f(\theta) = a \sin^2 \theta/2$.

(iv) $f(\theta) = \theta$, $0 \leq \theta \leq 2\pi$.

(v) $f(\theta) = 3 \sec \theta$, $0 \leq \theta \leq \pi/3$.

31. In Problem 8 of the Appendix to Chapter 12 we described the cycloid, which has the parametric representation

$$x = u(t) = a(t - \sin t), \quad y = v(t) = a(1 - \cos t).$$

- (a) Find the length of one arch of the cycloid. [Answer: $8a$.]
- (b) Recall that the cycloid is the graph of $v \circ u^{-1}$. Find the area under one arch of the cycloid by using the appropriate substitution in $\int f$ and evaluating the resultant integral. [Answer: $3\pi a^2$.]

32. Use induction and integration by parts to generalize Problem 14-13:

$$\int_0^x \frac{f(u)(x-u)^n}{n!} du = \int_0^x \left(\int_0^{u_n} \left(\dots \left(\int_0^{u_1} f(t) dt \right) du_1 \right) \dots \right) du_n.$$

33. If f' is continuous on $[a, b]$, use integration by parts to prove the Riemann-Lebesgue Lemma for f :

$$\lim_{\lambda \rightarrow \infty} \int_a^b f(t) \sin(\lambda t) dt = 0.$$

This result is just a special case of Problem 15-26, but it can be used to prove the general case (in much the same way that the Riemann-Lebesgue Lemma was derived in Problem 15-26 from the special case in which f is a step function).

34. The Mean Value Theorem for Integrals was introduced in Problem 13-23. The “Second Mean Value Theorem for Integrals” states the following. Suppose that f is integrable on $[a, b]$ and that ϕ is either nondecreasing or nonincreasing on $[a, b]$. Then there is a number ξ in $[a, b]$ such that

$$\int_a^b f(x)\phi(x) dx = \phi(a) \int_a^\xi f(x) dx + \phi(b) \int_\xi^b f(x) dx.$$

In this problem, we will assume that f is continuous and that ϕ is differentiable, with a continuous derivative ϕ' .

- (a) Prove that if the result is true for nonincreasing ϕ , then it is also true for nondecreasing ϕ .
- (b) Prove that if the result is true for nonincreasing ϕ satisfying $\phi(b) = 0$, then it is true for all nonincreasing ϕ .

Thus, we can assume that ϕ is nonincreasing and $\phi(b) = 0$. In this case, we have to prove that

$$\int_a^b f(x)\phi(x) dx = \phi(a) \int_a^\xi f(x) dx.$$

- (c) Prove this by using integration by parts.
- (d) Show that the hypothesis that ϕ is either nondecreasing or nonincreasing is needed.

From this special case of the Second Mean Value Theorem for Integrals, the general case could be derived by some approximation arguments, just as in the case of the Riemann-Lebesgue Lemma. But there is a more instructive way, outlined in the next problem.

35. (a) Given a_1, \dots, a_n and b_1, \dots, b_n , let $s_k = a_1 + \dots + a_k$. Show that

$$(*) \quad a_1 b_1 + \dots + a_n b_n = s_1(b_1 - b_2) + s_2(b_2 - b_3) + \dots + s_{n-1}(b_{n-1} - b_n) + s_n b_n$$

This disarmingly simple formula is sometimes called “Abel’s formula for summation by parts.” It may be regarded as an analogue for sums of the integration by parts formula

$$\int_a^b f'(x)g(x) dx = f(b)g(b) - f(a)g(a) - \int_a^b f(x)g'(x) dx,$$

especially if we use Riemann sums (Chapter 13, Appendix). In fact, for a partition $P = \{t_0, \dots, t_n\}$ of $[a, b]$, the left side is approximately

$$(1) \quad \sum_{k=1}^n f'(t_k)g(t_{k-1})(t_k - t_{k-1}),$$

while the right side is approximately

$$f(b)g(b) - f(a)g(a) - \sum_{k=1}^n f(t_k)g'(t_k)(t_k - t_{k-1})$$

which is approximately

$$\begin{aligned} & f(b)g(b) - f(a)g(a) - \sum_{k=1}^n f(t_k) \frac{g(t_k) - g(t_{k-1})}{t_k - t_{k-1}} (t_k - t_{k-1}) \\ &= f(b)g(b) - f(a)g(a) + \sum_{k=1}^n f(t_k)[g(t_{k-1}) - g(t_k)] \\ &= f(b)g(b) - f(a)g(a) + \sum_{k=1}^n [f(t_k) - f(a)] \cdot [g(t_{k-1}) - g(t_k)] \\ & \quad + f(a) \sum_{k=1}^n g(t_{k-1}) - g(t_k). \end{aligned}$$

Since the right-most sum is just $g(a) - g(b)$, this works out to be

$$(2) \quad [f(b) - f(a)]g(b) + \sum_{k=1}^n [f(t_k) - f(a)] \cdot [g(t_{k-1}) - g(t_k)].$$

If we choose

$$a_k = f'(t_k)(t_k - t_{k-1}), \quad b_k = g(t_{k-1})$$

then

$$(1) \quad \text{is} \quad \sum_{k=1}^n a_k b_k,$$

which is the left side of $(*)$, while

$$s_k = \sum_{i=1}^k f'(t_i)(t_i - t_{i-1}) \quad \text{is approximately} \quad \sum_{i=1}^k f(t_i) - f(t_{i-1}) = f(t_k) - f(a),$$

so

$$(2) \quad \text{is approximately} \quad s_n b_n + \sum_{k=1}^n s_k (b_k - b_{k-1}),$$

which is the right side of $(*)$.

This discussion is not meant to suggest that Abel's formula can actually be derived from the formula for integration by parts, or *vice versa*. But, as we shall see, Abel's formula can often be used as a substitute for integration by parts in situations where the functions in question aren't differentiable.

- (b) Suppose that $\{b_n\}$ is nonincreasing, with $b_n \geq 0$ for each n , and that

$$m \leq a_1 + \cdots + a_n \leq M$$

for all n . Prove Abel's Lemma:

$$b_1 m \leq a_1 b_1 + \cdots + a_n b_n \leq b_1 M.$$

(And, moreover,

$$b_k m \leq a_k b_k + \cdots + a_n b_n \leq b_k M,$$

a formula which only looks more general, but really isn't.)

- (c) Let f be integrable on $[a, b]$ and let ϕ be nonincreasing on $[a, b]$ with $\phi(b) = 0$. Let $P = \{t_0, \dots, t_n\}$ be a partition of $[a, b]$. Show that the sum

$$\sum_{i=1}^n f(t_{i-1})\phi(t_{i-1})(t_i - t_{i-1})$$

lies between the smallest and the largest of the sums

$$\phi(a) \sum_{i=1}^k f(t_{i-1})(t_i - t_{i-1}).$$

Conclude that

$$\int_a^b f(x)\phi(x) dx$$

lies between the minimum and the maximum of

$$\phi(a) \int_a^x f(t) dt,$$

and that it therefore equals $\phi(a) \int_a^\xi f(t) dt$ for some ξ in $[a, b]$.

36. (a) Show that the following improper integrals both converge.

$$(i) \quad \int_0^1 \sin\left(x + \frac{1}{x}\right) dx.$$

$$(ii) \quad \int_0^1 \sin^2\left(x + \frac{1}{x}\right) dx.$$

- (b) Decide which of the following improper integrals converge.

$$(i) \quad \int_1^\infty \sin\left(\frac{1}{x}\right) dx.$$

$$(ii) \quad \int_1^\infty \sin^2\left(\frac{1}{x}\right) dx.$$

37. (a) Compute the (improper) integral $\int_0^1 \log x \, dx$.

(b) Show that the improper integral $\int_0^\pi \log(\sin x) \, dx$ converges.

(c) Use the substitution $x = 2u$ to show that

$$\int_0^\pi \log(\sin x) \, dx = 2 \int_0^{\pi/2} \log(\sin x) \, dx + 2 \int_0^{\pi/2} \log(\cos x) \, dx + \pi \log 2.$$

(d) Compute $\int_0^{\pi/2} \log(\cos x) \, dx$.

(e) Using the relation $\cos x = \sin(\pi/2 - x)$, compute $\int_0^\pi \log(\sin x) \, dx$.

38. Prove the following version of integration by parts for improper integrals:

$$\int_a^\infty u'(x)v(x) \, dx = u(x)v(x) \Big|_a^\infty - \int_a^\infty u(x)v'(x) \, dx.$$

The first symbol on the right side means, of course,

$$\lim_{x \rightarrow \infty} u(x)v(x) - u(a)v(a).$$

***39.** One of the most important functions in analysis is the gamma function,

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

(a) Prove that the improper integral $\Gamma(x)$ is defined if $x > 0$.

(b) Use integration by parts (more precisely, the improper integral version in the previous problem) to prove that

$$\Gamma(x+1) = x\Gamma(x).$$

(c) Show that $\Gamma(1) = 1$, and conclude that $\Gamma(n) = (n-1)!$ for all natural numbers n .

The gamma function thus provides a simple example of a continuous function which “interpolates” the values of $n!$ for natural numbers n . Of course there are infinitely many continuous functions f with $f(n) = (n-1)!$; there are even infinitely many continuous functions f with $f(x+1) = xf(x)$ for all $x > 0$. However, the gamma function has the important additional property that $\log \circ \Gamma$ is convex, a condition which expresses the extreme smoothness of this function. A beautiful theorem due to Harold Bohr and Johannes Mollerup states that Γ is the only function f with $\log \circ f$ convex, $f(1) = 1$ and $f(x+1) = xf(x)$. See the Suggested Reading for a reference.

***40.** (a) Use the reduction formula for $\int \sin^n x \, dx$ to show that

$$\int_0^{\pi/2} \sin^n x \, dx = \frac{n-1}{n} \int_0^{\pi/2} \sin^{n-2} x \, dx.$$

(b) Now show that

$$\int_0^{\pi/2} \sin^{2n+1} x \, dx = \frac{2}{3} \cdot \frac{4}{5} \cdot \frac{6}{7} \cdots \frac{2n}{2n+1},$$

$$\int_0^{\pi/2} \sin^{2n} x \, dx = \frac{\pi}{2} \cdot \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{5}{6} \cdots \frac{2n-1}{2n},$$

and conclude that

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdots \frac{2n}{2n-1} \cdot \frac{2n}{2n+1} \frac{\int_0^{\pi/2} \sin^{2n} x \, dx}{\int_0^{\pi/2} \sin^{2n+1} x \, dx}.$$

(c) Show that the quotient of the two integrals in this expression is between 1 and $1 + 1/(2n)$, starting with the inequalities

$$0 < \sin^{2n+1} x \leq \sin^{2n} x \leq \sin^{2n-1} x \quad \text{for } 0 < x < \pi/2.$$

This result, which shows that the products

$$\frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdots \frac{2n}{2n-1} \cdot \frac{2n}{2n+1}$$

can be made as close to $\pi/2$ as desired, is usually written as an infinite product, known as Wallis' product:

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdots$$

(d) Show also that the products

$$\frac{1}{\sqrt{n}} \frac{2 \cdot 4 \cdot 6 \cdots 2n}{1 \cdot 3 \cdot 5 \cdots (2n-1)}$$

can be made as close to $\sqrt{\pi}$ as desired. (This fact is used in the next problem and in Problem 27-19.)

- **41.** It is an astonishing fact that improper integrals $\int_0^\infty f(x) \, dx$ can often be computed in cases where ordinary integrals $\int_a^b f(x) \, dx$ cannot. There is no elementary formula for $\int_a^b e^{-x^2} \, dx$, but we can find the value of $\int_0^\infty e^{-x^2} \, dx$ precisely! There are many ways of evaluating this integral, but most require some advanced techniques; the following method involves a fair amount of work, but no facts that you do not already know.

(a) Show that

$$\int_0^1 (1-x^2)^n dx = \frac{2}{3} \cdot \frac{4}{5} \cdot \dots \cdot \frac{2n}{2n+1},$$

$$\int_0^\infty \frac{1}{(1+x^2)^n} dx = \frac{\pi}{2} \cdot \frac{1}{2} \cdot \frac{3}{4} \cdot \dots \cdot \frac{2n-3}{2n-2}.$$

(This can be done using reduction formulas, or by appropriate substitutions, combined with the previous problem.)

(b) Prove, using the derivative, that

$$\begin{aligned} 1-x^2 &\leq e^{-x^2} \quad \text{for } 0 \leq x \leq 1. \\ e^{-x^2} &\leq \frac{1}{1+x^2} \quad \text{for } 0 \leq x. \end{aligned}$$

(c) Integrate the n th powers of these inequalities from 0 to 1 and from 0 to ∞ , respectively. Then use the substitution $y = \sqrt{n}x$ to show that

$$\begin{aligned} \sqrt{n} \frac{2}{3} \cdot \frac{4}{5} \cdot \dots \cdot \frac{2n}{2n+1} \\ \leq \int_0^{\sqrt{n}} e^{-y^2} dy \leq \int_0^\infty e^{-y^2} dy \\ \leq \frac{\pi}{2} \sqrt{n} \frac{1}{2} \cdot \frac{3}{4} \cdot \dots \cdot \frac{2n-3}{2n-2}. \end{aligned}$$

(d) Now use Problem 40(d) to show that

$$\int_0^\infty e^{-y^2} dy = \frac{\sqrt{\pi}}{2}.$$

****42.** (a) Use integration by parts to show that

$$\int_a^b \frac{\sin x}{x} dx = \frac{\cos a}{a} - \frac{\cos b}{b} - \int_a^b \frac{\cos x}{x^2} dx,$$

and conclude that $\int_0^\infty (\sin x)/x dx$ exists. (Use the left side to investigate the limit as $a \rightarrow 0^+$ and the right side for the limit as $b \rightarrow \infty$.)

(b) Use Problem 15-33 to show that

$$\int_0^\pi \frac{\sin(n + \frac{1}{2})t}{\sin \frac{t}{2}} dt = \pi$$

for any natural number n .

(c) Prove that

$$\lim_{\lambda \rightarrow \pi} \int_0^\pi \sin(\lambda + \frac{1}{2})t \left[\frac{2}{t} - \frac{1}{\sin \frac{t}{2}} \right] dt = 0.$$

Hint: The term in brackets is bounded by Problem 15-2(vi); the Riemann-Lebesgue Lemma then applies.

- (d) Use the substitution $u = (\lambda + \frac{1}{2})t$ and part (b) to show that

$$\int_0^\infty \frac{\sin x}{x} dx = \frac{\pi}{2}.$$

43. Given the value of $\int_0^\infty (\sin x)/x dx$ from Problem 42, compute

$$\int_0^\infty \left(\frac{\sin x}{x} \right)^2 dx$$

by using integration by parts. (As in Problem 37, the formula for $\sin 2x$ will play an important role.)

- *44. (a) Use the substitution $u = t^x$ to show that

$$\Gamma(x) = \frac{1}{x} \int_0^\infty e^{-u^{1/x}} du.$$

- (b) Find $\Gamma(\frac{1}{2})$.

- *45. (a) Suppose that $\frac{f(x)}{x}$ is integrable on every interval $[a, b]$ for $0 < a < b$, and that $\lim_{x \rightarrow 0} f(x) = A$ and $\lim_{x \rightarrow \infty} f(x) = B$. Prove that for all $\alpha, \beta > 0$ we have

$$\int_0^\infty \frac{f(ax) - f(\beta x)}{x} dx = (A - B) \log \frac{\beta}{\alpha}.$$

Hint: To estimate $\int_s^N \frac{f(\alpha x) - f(\beta x)}{x} dx$ use two different substitutions.

- (b) Now suppose instead that $\int_a^\infty \frac{f(x)}{x} dx$ converges for all $a > 0$ and that $\lim_{x \rightarrow 0} f(x) = A$. Prove that

$$\int_0^\infty \frac{f(\alpha x) - f(\beta x)}{x} dx = A \log \frac{\beta}{\alpha}.$$

- (c) Compute the following integrals:

$$(i) \quad \int_0^\infty \frac{e^{-\alpha x} - e^{-\beta x}}{x} dx.$$

$$(ii) \quad \int_0^\infty \frac{\cos(\alpha x) - \cos(\beta x)}{x} dx.$$

In Chapter 13 we said, rather blithely, that integrals may be computed to any degree of accuracy desired by calculating lower and upper sums. But an applied mathematician, who really has to do the calculation, rather than just talking about doing it, may not be overjoyed at the prospect of computing lower sums to evaluate an integral to three decimal places, say (a degree of accuracy that might easily be needed in certain circumstances). The next three problems show how more refined methods can make the calculations much more efficient.

We ought to mention at the outset that computing upper and lower sums might not even be practical, since it might not be possible to compute the quantities m_i and M_i for each interval $[t_{i-1}, t_i]$. It is far more reasonable simply to pick points x_i in $[t_{i-1}, t_i]$ and consider $\sum_{i=1}^n f(x_i) \cdot (t_i - t_{i-1})$. This represents the sum of the areas of certain rectangles which partially overlap the graph of f —see Figure 1 in the Appendix to Chapter 13. But we will get a much better result if we instead choose the trapezoids shown in Figure 2.

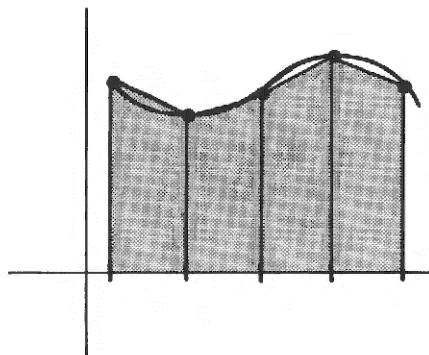


FIGURE 2

Suppose, in particular, that we divide $[a, b]$ into n equal intervals, by means of the points

$$t_i = a + i \left(\frac{b-a}{n} \right) = a + ih.$$

Then the trapezoid with base $[t_{i-1}, t_i]$ has area

$$\frac{f(t_{i-1}) + f(t_i)}{2} \cdot (t_i - t_{i-1})$$

and the sum of all these areas is simply

$$\begin{aligned} \Sigma_n &= h \left[\frac{f(t_1) + f(a)}{2} + \frac{f(t_2) + f(t_1)}{2} + \dots + \frac{f(b) + f(t_{n-1})}{2} \right] \\ &= \frac{h}{2} \left[f(a) + 2 \sum_{i=1}^{n-1} f(a + ih) + f(b) \right], \quad h = \frac{b-a}{n}. \end{aligned}$$

This method of approximating an integral is called the *trapezoid rule*. Notice that to obtain Σ_{2n} from Σ_n it isn't necessary to recompute the old $f(t_i)$; their contribution to Σ_{2n} is just $\frac{1}{2}\Sigma_n$. So in practice it is best to compute $\Sigma_2, \Sigma_4, \Sigma_8, \dots$ to get approximations to $\int_a^b f$. In the next problem we will estimate $\int_a^b f - \Sigma_n$.

46. (a) Suppose that f'' is continuous. Let P_i be the linear function which agrees with f at t_{i-1} and t_i . Using Problem 11-43, show that if n_i and N_i are

the minimum and maximum of f'' on $[t_{i-1}, t_i]$ and

$$I = \int_{t_{i-1}}^{t_i} (x - t_{i-1})(x - t_i) dx$$

then

$$\frac{n_i I}{2} \geq \int_{t_{i-1}}^{t_i} (f - P_i) \geq \frac{N_i I}{2}.$$

(b) Evaluate I to get

$$\frac{n_i h^3}{12} \leq \int_{t_{i-1}}^{t_i} (f - P_i) \leq \frac{N_i h^3}{12}.$$

(c) Conclude that there is some c in $[a, b]$ with

$$\int_a^b f = \Sigma_n - \frac{(b-a)^3}{12n^2} f''(c).$$

Notice that the “error term” $(b-a)^3 f''(c)/12n^2$ varies as $1/n^2$ (while the error obtained using ordinary sums varies as $1/n$).

We can obtain still more accurate results if we approximate f by quadratic functions rather than by linear functions. We first consider what happens when the interval $[a, b]$ is divided into two equal intervals (Figure 3).

47. (a) Suppose first that $a = 0$ and $b = 2$. Let P be the second degree polynomial function which agrees with f at 0, 1, and 2 (Problem 3-6). Show that

$$\int_0^2 P = \frac{1}{3}[f(0) + 4f(1) + f(2)].$$

(b) Conclude that in the general case

$$\int_a^b P = \frac{b-a}{6} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right].$$

- (c) Naturally $\int_a^b P = \int_a^b f$ when f is a quadratic polynomial. But, remarkably enough, this same relation holds when f is a cubic polynomial! Prove this, using Problem 11-43; note that f''' is a constant.

The previous problem shows that we do not have to do any new calculations to compute $\int_a^b Q$ when Q is a *cubic* polynomial which agrees with f at a , b , and $\frac{a+b}{2}$: we still have

$$\int_a^b Q = \frac{b-a}{6} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right].$$

But there is much more lee-way in choosing Q , which we can use to our advantage:

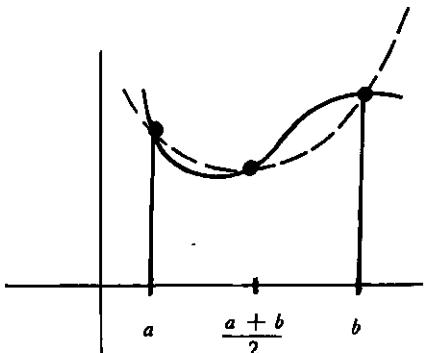


FIGURE 3

48. (a) Show that there is a cubic polynomial function Q satisfying

$$Q(a) = f(a), \quad Q(b) = f(b), \quad Q\left(\frac{a+b}{2}\right) = f\left(\frac{a+b}{2}\right)$$

$$Q'\left(\frac{a+b}{2}\right) = f'\left(\frac{a+b}{2}\right).$$

Hint: Clearly $Q(x) = P(x) + A(x-a)(x-b)\left(x - \frac{a+b}{2}\right)$ for some A .

- (b) Prove that for every x we have

$$f(x) - Q(x) = (x-a)\left(x - \frac{a+b}{2}\right)^2 (x-b) \frac{f^{(4)}(\xi)}{4!}$$

for some ξ in $[a, b]$. Hint: Imitate the proof of Problem 11-43.

- (c) Conclude that if $f^{(4)}$ is continuous, then

$$\int_a^b f = \frac{b-a}{6} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right] - \frac{(b-a)^5}{2880} f^{(4)}(c)$$

for some c in $[a, b]$.

- (d) Now divide $[a, b]$ into $2n$ intervals by means of the points

$$t_i = a + ih, \quad h = \frac{b-a}{2n}.$$

Prove *Simpson's rule*:

$$\int_a^b f = \frac{b-a}{n} \left(f(a) + 4 \sum_{i=1}^n f(t_{2i-1}) + 2 \sum_{i=1}^{n-1} f(t_{2i}) + f(b) \right)$$

$$- \frac{(b-a)^5}{2880n^4} f^{(4)}(\bar{c})$$

for some \bar{c} in $[a, b]$.

APPENDIX. THE COSMOPOLITAN INTEGRAL

We originally introduced integrals in order to find the area under the graph of a function, but the integral is considerably more versatile than that. For example, Problem 13-24 used the integral to express the area of a region of quite another sort. Moreover, Problem 13-25 showed that the integral can also be used to express the lengths of curves—though, as we've seen in Appendix to Chapter 13, a lot of work may be necessary to consider the general case! This result was probably a little more surprising, since the integral seems, at first blush, to be a very two-dimensional creature. Actually, the integral makes its appearance in quite a few geometric formulas, which we will present in this Appendix. To derive these formulas we will assume some results from elementary geometry (and allow a little fudging).

Instead of going down to one-dimensional objects, we'll begin by tackling some three-dimensional ones. There are some very special solids whose volumes can be expressed by integrals. The simplest such solid V is a “volume of revolution,” obtained by revolving the region under the graph of $f \geq 0$ on $[a, b]$ around the horizontal axis, when we regard the plane as situated in space (Figure 1). If $P = \{t_0, \dots, t_n\}$ is any partition of $[a, b]$, and m_i and M_i have their usual meanings, then

$$\pi m_i^2(t_i - t_{i-1})$$

is the volume of a disc that lies inside the solid V (Figure 2). Similarly, $\pi M_i^2(t_i - t_{i-1})$ is the volume of a disc that contains the part of V between t_{i-1} and t_i . Consequently,

$$\pi \sum_{i=1}^n m_i^2(t_i - t_{i-1}) \leq \text{volume } V \leq \pi \sum_{i=1}^n M_i^2(t_i - t_{i-1}).$$

But the sums on the ends of this inequality are just the lower and upper sums for f^2 on $[a, b]$:

$$\pi \cdot L(f^2, P) \leq \text{volume } V \leq \pi \cdot U(f^2, P).$$

Consequently, the volume of V must be given by

$$\text{volume } V = \pi \int_a^b f(x)^2 dx.$$

This method of finding volumes is affectionately referred to as the “disc method.”

Figure 3 shows a more complicated solid V obtained by revolving the region under the graph of f around the vertical axis (V is the solid left over when we start with the big cylinder of radius b and take away both the small cylinder of radius a and the solid V_1 sitting right on top of it). In this case we assume $a \geq 0$ as well

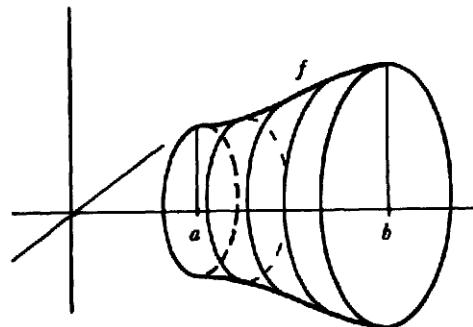


FIGURE 1

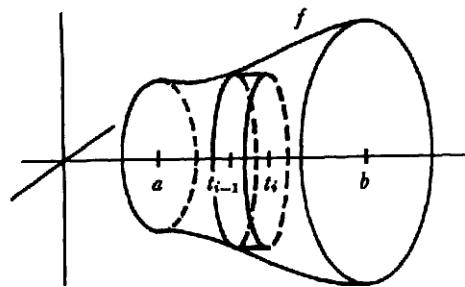


FIGURE 2

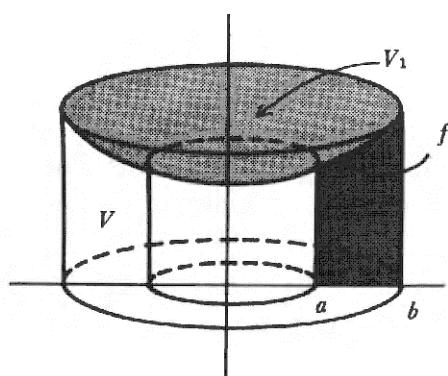


FIGURE 3

as $f \geq 0$. Figures 4 and 5 indicate some other possible shapes for V .

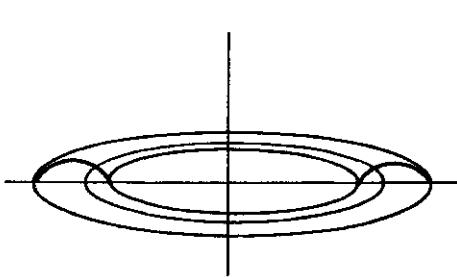


FIGURE 4

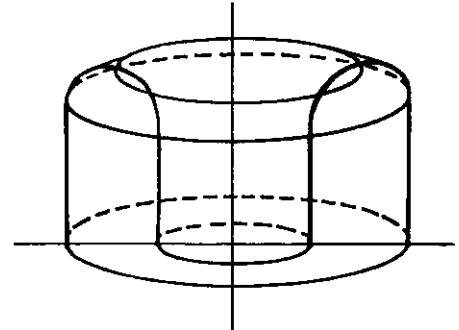


FIGURE 5

For a partition $P = \{t_0, \dots, t_n\}$ we consider the “shells” obtained by rotating the rectangle with base $[t_{i-1}, t_i]$ and height m_i or M_i (Figure 6). Adding the volumes of these shells we obtain

$$\pi \sum_{i=1}^n m_i(t_i^2 - t_{i-1}^2) \leq \text{volume } V \leq \pi \sum_{i=1}^n M_i(t_i^2 - t_{i-1}^2),$$

which we can write as

$$\pi \sum_{i=1}^n m_i(t_i + t_{i-1})(t_i - t_{i-1}) \leq \text{volume } V \leq \pi \sum_{i=1}^n M_i(t_i + t_{i-1})(t_i - t_{i-1}).$$

Now these sums are not lower or upper sums of anything. But Problem 1 of the Appendix to Chapter 13 shows that each sum

$$\sum_{i=1}^n m_i t_i (t_i - t_{i-1}) \quad \text{and} \quad \sum_{i=1}^n m_i t_{i-1} (t_i - t_{i-1})$$

can be made as close as desired to $\int_a^b xf(x) dx$ by choosing the lengths $t_i - t_{i-1}$ small enough. The same is true of the sums on the right, so we find that

$$\text{volume } V = 2\pi \int_a^b xf(x) dx;$$

this is the so-called “shell method” of finding volumes.

The surface area of certain curved regions can also be expressed in terms of integrals. Before we tackle complicated regions, a little review of elementary geometric formulas may be appreciated here.

Figure 7 shows a right pyramid made up of triangles with bases of length l and altitude s . The total surface area of the sides of the pyramid is thus

$$\frac{1}{2} ps,$$

where p is the perimeter of the base. By choosing the base to be a regular polygon

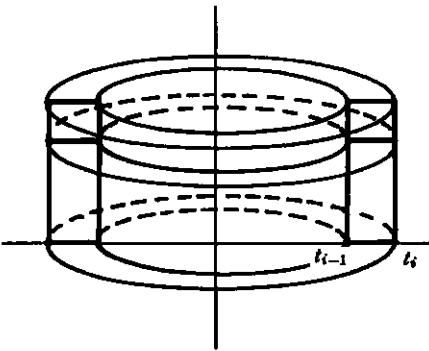


FIGURE 6

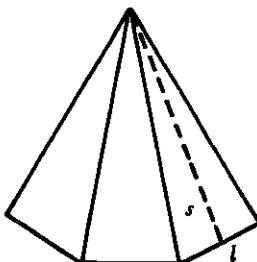


FIGURE 7

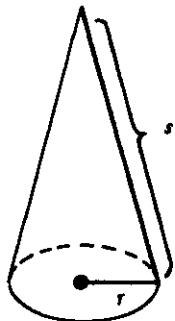
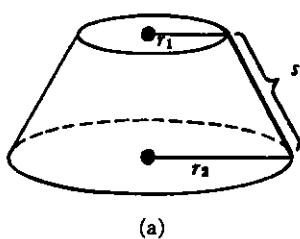
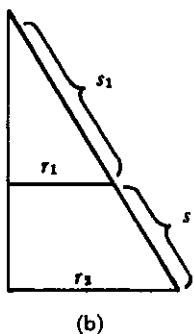


FIGURE 8



(a)



(b)

FIGURE 9

with a large number of sides we see that the area of a right circular cone (Figure 8) must be

$$\frac{1}{2}(2\pi r)s = \pi rs,$$

where s is the “slant height.” Finally, consider the frustum of a cone with slant height s and radii r_1 and r_2 shown in Figure 9(a). Completing this to a cone, as in Figure 9(b), we have

$$\frac{s_1}{r_1} = \frac{s_1 + s}{r_2},$$

so

$$s_1 = \frac{r_1 s}{r_2 - r_1}, \quad s_1 + s = \frac{r_2 s}{r_2 - r_1}.$$

Consequently, the surface area is

$$\pi r_2(s_1 + s) - \pi r_1 s_1 = \pi s \frac{r_2^2 - r_1^2}{r_2 - r_1} = \pi s(r_1 + r_2).$$

Now consider the surface formed by revolving the graph of f around the horizontal axis. For a partition $P = \{t_0, \dots, t_n\}$ we can inscribe a series of frusta of cones, as in Figure 10. The total surface area of these frusta is

$$\begin{aligned} \pi \sum_{i=1}^n [f(t_{i-1}) + f(t_i)] \sqrt{(t_i - t_{i-1})^2 + [f(t_i) - f(t_{i-1})]^2} \\ = \pi \sum_{i=1}^n [f(t_{i-1}) + f(t_i)] \sqrt{1 + \left(\frac{f(t_i) - f(t_{i-1})}{t_i - t_{i-1}} \right)^2} (t_i - t_{i-1}). \end{aligned}$$

By the Mean Value Theorem, this is

$$\pi \sum_{i=1}^n [f(t_{i-1}) + f(t_i)] \sqrt{1 + f'(x_i)^2} (t_i - t_{i-1})$$

for some x_i in (t_{i-1}, t_i) . Appealing to Problem 1 of the Appendix to Chapter 13, we conclude that the surface area is

$$2\pi \int_a^b f(x) \sqrt{1 + f'(x)^2} dx.$$

PROBLEMS

1. (a) Find the volume of the solid obtained by revolving the region bounded by the graphs of $f(x) = x$ and $f(x) = x^2$ around the horizontal axis.
(b) Find the volume of the solid obtained by revolving this same region around the vertical axis.
2. Find the volume of a sphere of radius r .

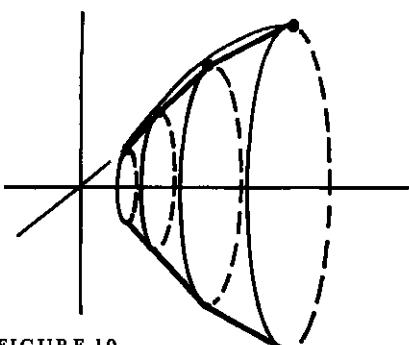


FIGURE 10

3. When the ellipse consisting of all points (x, y) with $x^2/a^2 + y^2/b^2 = 1$ is rotated around the horizontal axis we obtain an “ellipsoid of revolution” (Figure 11). Find the volume of the enclosed solid.

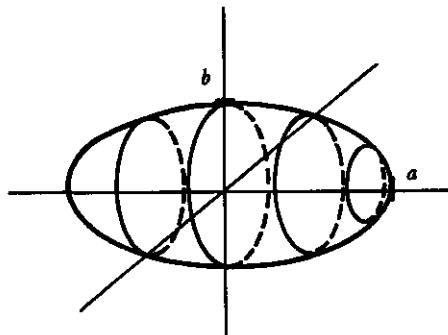


FIGURE 11

4. Find the volume of the “torus” (Figure 12), obtained by rotating the circle $(x - a)^2 + y^2 = b^2$ ($a > b$) around the vertical axis.

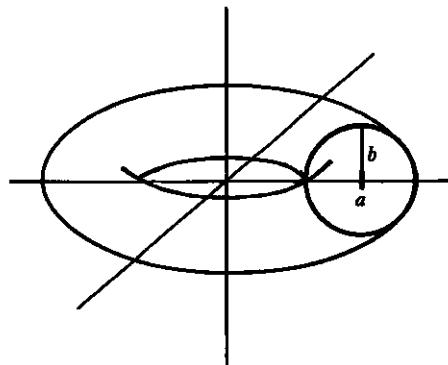


FIGURE 12

5. A cylindrical hole of radius a is bored through the center of a sphere of radius $2a$ (Figure 13). Find the volume of the remaining solid.

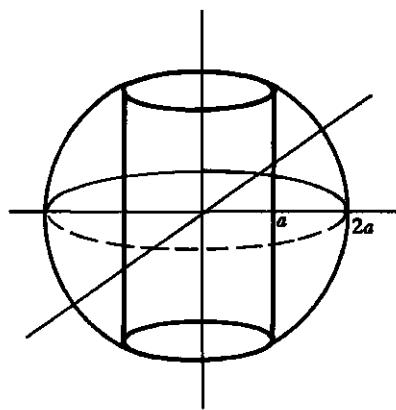


FIGURE 13

6. (a) For the solid shown in Figure 14, find the volume by the shell method.

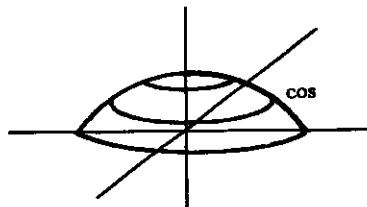


FIGURE 14

- (b) This volume can also be evaluated by the disc method. Write down the integral which must be evaluated in this case; notice that it is more complicated. The next problem takes up a question which this might suggest.
7. Figure 15 shows a cylinder of height b and radius $f(b)$, divided into three solids, one of which, V_1 , is a cylinder of height a and radius $f(a)$. If f is one-one, then a comparison of the disk method and the shell method of computing volumes leads us to believe that

$$\begin{aligned} \pi b f(b)^2 - \pi a f(a)^2 - \pi \int_a^b f(x)^2 dx &= \text{volume } V_2 \\ &= 2\pi \int_{f(a)}^{f(b)} y f^{-1}(y) dy. \end{aligned}$$

Prove this analytically, using the formula for $\int f^{-1}$ from Problem 19-15.

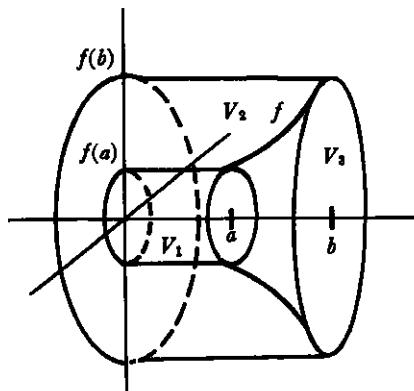


FIGURE 15

8. (a) Figure 16 shows a solid with a circular base of radius a . Each plane perpendicular to the diameter AB intersects the solid in a square. Using arguments similar to those already used in this Appendix, express the volume of the solid as an integral, and evaluate it.
- (b) Same problem if each plane intersects the solid in an equilateral triangle.

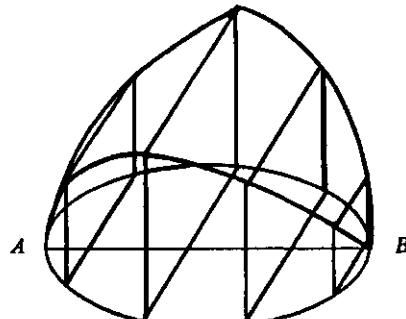


FIGURE 16

9. Find the volume of a pyramid (Figure 17) in terms of its height h and the area A of its base.

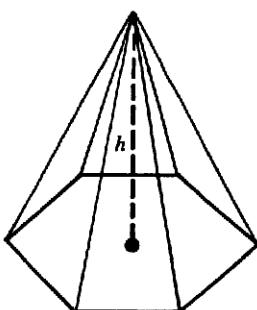


FIGURE 17

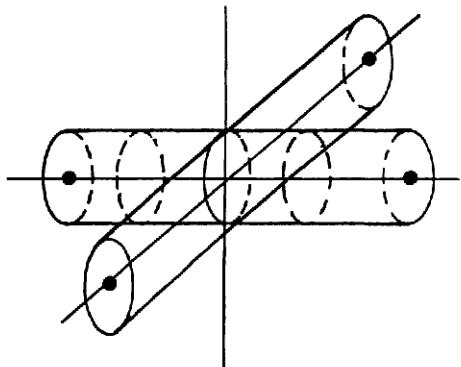


FIGURE 18

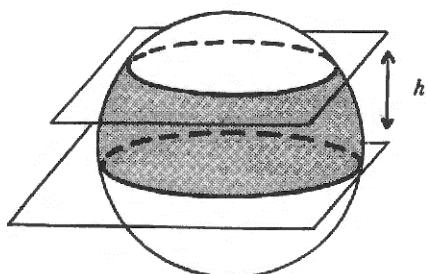


FIGURE 19

10. Find the volume of the solid which is the intersection of the two cylinders in Figure 18. Hint: Find the intersection of this solid with each horizontal plane.
11. (a) Prove that the surface area of a sphere of radius r is $4\pi r^2$.
 (b) Prove, more generally, that the area of the portion of the sphere shown in Figure 19 is $2\pi rh$. (Notice that this depends only on h , not on the position of the planes!)
12. (a) Find the surface area of the ellipsoid of revolution in Problem 19-3.
 (b) Find the surface area of the torus in Problem 19-4.
13. The graph of $f(x) = 1/x$, $x \geq 1$ is revolved around the horizontal axis (Figure 20).
- (a) Find the volume of the enclosed “infinite trumpet.”
 (b) Show that the surface area is infinite.
 (c) Suppose that we fill up the trumpet with the finite amount of paint found in part (a). It would seem that we have thereby coated the infinite inside surface area with only a finite amount of paint! How is this possible?

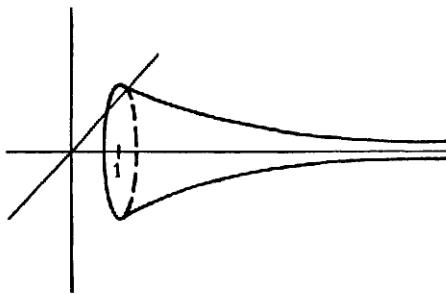


FIGURE 20

PART **4**
INFINITE
SEQUENCES
AND
INFINITE
SERIES

One of the most remarkable series of algebraic analysis is the following:

$$\begin{aligned}1 + \frac{m}{1}x + \frac{m(m-1)}{1 \cdot 2}x^2 \\+ \frac{m(m-1)(m-2)}{1 \cdot 2 \cdot 3}x^3 + \dots \\+ \frac{m(m-1) \dots [m - (n-1)]}{1 \cdot 2 \cdot \dots \cdot n}x^n \\+ \dots\end{aligned}$$

*When m is a positive whole number
the sum of the series,
which is then finite, can be expressed,
as is known, by $(1+x)^m$.*

*When m is not an integer,
the series goes on to infinity, and it will
converge or diverge according
as the quantities
 m and x have this or that value.*

In this case, one writes the same equality

$$\begin{aligned}(1+x)^m = 1 + \frac{m}{1}x \\+ \frac{m(m-1)}{1 \cdot 2}x^2 + \dots \text{etc.}\end{aligned}$$

*. . . It is assumed that
the numerical equality will always occur
whenever the series is convergent, but
this has never yet been proved.*

NIELS HENRIK ABEL

CHAPTER

20

APPROXIMATION BY
POLYNOMIAL FUNCTIONS

There is one sense in which the “elementary functions” are not elementary at all. If p is a polynomial function,

$$p(x) = a_0 + a_1x + \cdots + a_nx^n,$$

then $p(x)$ can be computed easily for any number x . This is not at all true for functions like sin, log, or exp. At present, to find $\log x = \int_1^x 1/t dt$ approximately, we must compute some upper or lower sums, and make certain that the error involved in accepting such a sum for $\log x$ is not too great. Computing $e^x = \log^{-1}(x)$ would be even more difficult: we would have to compute $\log a$ for many values of a until we found a number a such that $\log a$ is approximately x —then a would be approximately e^x .

In this chapter we will obtain important theoretical results which reduce the computation of $f(x)$, for many functions f , to the evaluation of polynomial functions. The method depends on finding polynomial functions which are close approximations to f . In order to guess a polynomial which is appropriate, it is useful to first examine polynomial functions themselves more thoroughly.

Suppose that

$$p(x) = a_0 + a_1x + \cdots + a_nx^n.$$

It is interesting, and for our purposes very important, to note that the coefficients a_i can be expressed in terms of the value of p and its various derivatives at 0. To begin with, note that

$$p(0) = a_0.$$

Differentiating the original expression for $p(x)$ yields

$$p'(x) = a_1 + 2a_2x + \cdots + na_nx^{n-1}.$$

Therefore,

$$p'(0) = p^{(1)}(0) = a_1.$$

Differentiating again we obtain

$$p''(x) = 2a_2 + 3 \cdot 2 \cdot a_3x + \cdots + n(n-1) \cdot a_nx^{n-2}.$$

Therefore,

$$p''(0) = p^{(2)}(0) = 2a_2.$$

In general, we will have

$$p^{(k)}(0) = k! a_k \quad \text{or} \quad a_k = \frac{p^{(k)}(0)}{k!}.$$

If we agree to define $0! = 1$, and recall the notation $p^{(0)} = p$, then this formula holds for $k = 0$ also.

If we had begun with a function p that was written as a “polynomial in $(x - a)$,”

$$p(x) = a_0 + a_1(x - a) + \cdots + a_n(x - a)^n,$$

then a similar argument would show that

$$a_k = \frac{p^{(k)}(a)}{k!}.$$

Suppose now that f is a function (not necessarily a polynomial) such that

$$f^{(1)}(a), \dots, f^{(n)}(a)$$

all exist. Let

$$a_k = \frac{f^{(k)}(a)}{k!}, \quad 0 \leq k \leq n,$$

and define

$$P_{n,a}(x) = a_0 + a_1(x - a) + \cdots + a_n(x - a)^n.$$

The polynomial $P_{n,a}$ is called the **Taylor polynomial of degree n for f at a** . (Strictly speaking, we should use an even more complicated expression, like $P_{n,a,f}$, to indicate the dependence on f ; at times this more precise notation will be useful.) The Taylor polynomial has been defined so that

$$P_{n,a}^{(k)}(a) = f^{(k)}(a) \quad \text{for } 0 \leq k \leq n;$$

in fact, it is clearly the only polynomial of degree $\leq n$ with this property.

Although the coefficients of $P_{n,a,f}$ seem to depend upon f in a fairly complicated way, the most important elementary functions have extremely simple Taylor polynomials. Consider first the function \sin . We have

$$\begin{aligned}\sin(0) &= 0, \\ \sin'(0) &= \cos 0 = 1, \\ \sin''(0) &= -\sin 0 = 0, \\ \sin'''(0) &= -\cos 0 = -1, \\ \sin^{(4)}(0) &= \sin 0 = 0.\end{aligned}$$

From this point on, the derivatives repeat in a cycle of 4. The numbers

$$a_k = \frac{\sin^{(k)}(0)}{k!}$$

are

$$0, 1, 0, -\frac{1}{3!}, 0, \frac{1}{5!}, 0, -\frac{1}{7!}, 0, \frac{1}{9!}, \dots$$

Therefore the Taylor polynomial $P_{2n+1,0}$ of degree $2n + 1$ for \sin at 0 is

$$P_{2n+1,0}(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!}.$$

(Of course, $P_{2n+1,0} = P_{2n+2,0}$).

The Taylor polynomial $P_{2n,0}$ of degree $2n$ for \cos at 0 is (the computations are left to you)

$$P_{2n,0}(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots + (-1)^n \frac{x^{2n}}{(2n)!}.$$

The Taylor polynomial for \exp is especially easy to compute. Since $\exp^{(k)}(0) = \exp(0) = 1$ for all k , the Taylor polynomial of degree n at 0 is

$$P_{n,0}(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots + \frac{x^n}{n!}.$$

The Taylor polynomial for \log must be computed at some point $a \neq 0$, since \log is not even defined at 0. The standard choice is $a = 1$. Then

$$\begin{aligned}\log'(x) &= \frac{1}{x}, & \log'(1) &= 1; \\ \log''(x) &= -\frac{1}{x^2}, & \log''(1) &= -1; \\ \log'''(x) &= \frac{2}{x^3}, & \log'''(1) &= 2;\end{aligned}$$

in general

$$\log^{(k)}(x) = \frac{(-1)^{k-1}(k-1)!}{x^k}, \quad \log^{(k)}(1) = (-1)^{k-1}(k-1)!.$$

Therefore the Taylor polynomial of degree n for \log at 1 is

$$P_{n,1}(x) = (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} + \cdots + \frac{(-1)^{n-1}(x-1)^n}{n}.$$

It is often more convenient to consider the function $f(x) = \log(1+x)$. In this case we can choose $a = 0$. We have

$$f^{(k)}(x) = \log^{(k)}(1+x),$$

so

$$f^{(k)}(0) = \log^{(k)}(1) = (-1)^{k-1}(k-1)!.$$

Therefore the Taylor polynomial of degree n for f at 0 is

$$P_{n,0}(x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots + \frac{(-1)^{n-1}x^n}{n}.$$

There is one other elementary function whose Taylor polynomial is important—the arctan. The computations of the derivatives begin

$$\arctan'(x) = \frac{1}{1+x^2} \quad \arctan'(0) = 1;$$

$$\arctan''(x) = \frac{-2x}{(1+x^2)^2}, \quad \arctan''(0) = 0;$$

$$\arctan'''(x) = \frac{(1+x^2)^2 \cdot (-2) + 2x \cdot 2(1+x^2) \cdot 2x}{(1+x^2)^4}, \quad \arctan'''(0) = -2.$$

It is clear that this brute force computation will never do. However, the Taylor polynomials of arctan will be easy to find after we have examined the properties of Taylor polynomials more closely—although the Taylor polynomial $P_{n,a,f}$ was simply defined so as to have the same first n derivatives at a as f , the connection between f and $P_{n,a,f}$ will actually turn out to be much deeper.

One line of evidence for a closer connection between f and the Taylor polynomials for f may be uncovered by examining the Taylor polynomial of degree 1, which is

$$P_{1,a}(x) = f(a) + f'(a)(x - a).$$

Notice that

$$\frac{f(x) - P_{1,a}(x)}{x - a} = \frac{f(x) - f(a)}{x - a} - f'(a).$$

Now, by the definition of $f'(a)$ we have

$$\lim_{x \rightarrow a} \frac{f(x) - P_{1,a}(x)}{x - a} = 0.$$

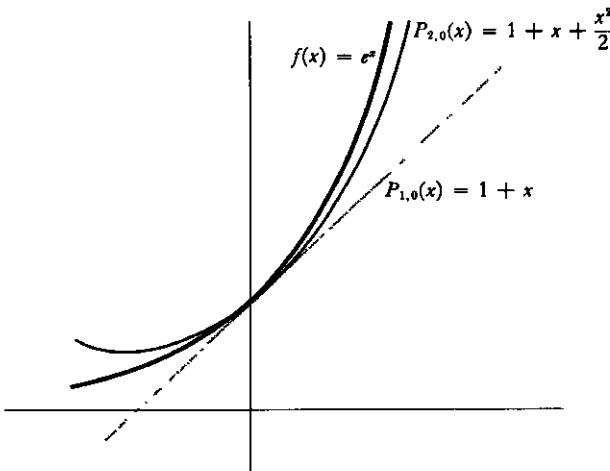


FIGURE 1

In other words, as x approaches a the difference $f(x) - P_{1,a}(x)$ not only becomes small, but actually becomes small even compared to $x - a$. Figure 1 illustrates the graph of $f(x) = e^x$ and of

$$P_{1,0}(x) = f(0) + f'(0)x = 1 + x,$$

which is the Taylor polynomial of degree 1 for f at 0. The diagram also shows the graph of

$$P_{2,0}(x) = f(0) + f'(0) + \frac{f''(0)}{2!}x^2 = 1 + x + \frac{x^2}{2},$$

which is the Taylor polynomial of degree 2 for f at 0. As x approaches 0, the difference $f(x) - P_{2,0}(x)$ seems to be getting small even faster than the difference

$f(x) - P_{1,0}(x)$. As it stands, this assertion is not very precise, but we are now prepared to give it a definite meaning. We have just noted that in general

$$\lim_{x \rightarrow a} \frac{f(x) - P_{1,a}(x)}{x - a} = 0.$$

For $f(x) = e^x$ and $a = 0$ this means that

$$\lim_{x \rightarrow 0} \frac{f(x) - P_{1,0}(x)}{x} = \lim_{x \rightarrow 0} \frac{e^x - 1 - x}{x} = 0.$$

On the other hand, an easy double application of l'Hôpital's Rule shows that

$$\lim_{x \rightarrow 0} \frac{e^x - 1 - x}{x^2} = \frac{1}{2} \neq 0.$$

Thus, although $f(x) - P_{1,0}(x)$ becomes small compared to x , as x approaches 0, it does *not* become small compared to x^2 . For $P_{2,0}(x)$ the situation is quite different; the extra term $x^2/2$ provides just the right compensation:

$$\begin{aligned} \lim_{x \rightarrow 0} \frac{e^x - 1 - x - \frac{x^2}{2}}{x^2} &= \lim_{x \rightarrow 0} \frac{e^x - 1 - x}{2x} \\ &= \lim_{x \rightarrow 0} \frac{e^x - 1}{2} = 0. \end{aligned}$$

This result holds in general—if $f'(a)$ and $f''(a)$ exist, then

$$\lim_{x \rightarrow a} \frac{f(x) - P_{2,a}(x)}{(x - a)^2} = 0;$$

in fact, the analogous assertion for $P_{n,a}$ is also true.

THEOREM 1 Suppose that f is a function for which

$$f'(a), \dots, f^{(n)}(a)$$

all exist. Let

$$a_k = \frac{f^{(k)}(a)}{k!}, \quad 0 \leq k \leq n,$$

and define

$$P_{n,a}(x) = a_0 + a_1(x - a) + \dots + a_n(x - a)^n.$$

Then

$$\lim_{x \rightarrow a} \frac{f(x) - P_{n,a}(x)}{(x - a)^n} = 0.$$

PROOF Writing out $P_{n,a}(x)$ explicitly, we obtain

$$\frac{f(x) - P_{n,a}(x)}{(x-a)^n} = \frac{f(x) - \sum_{i=0}^{n-1} \frac{f^{(i)}(a)}{i!}(x-a)^i}{(x-a)^n} - \frac{f^{(n)}(a)}{n!}.$$

It will help to introduce the new functions

$$Q(x) = \sum_{i=0}^{n-1} \frac{f^{(i)}(a)}{i!}(x-a)^i \quad \text{and} \quad g(x) = (x-a)^n;$$

now we must prove that

$$\lim_{x \rightarrow a} \frac{f(x) - Q(x)}{g(x)} = \frac{f^{(n)}(a)}{n!}.$$

Notice that

$$\begin{aligned} Q^{(k)}(a) &= f^{(k)}(a), \quad k \leq n-1, \\ g^{(k)}(x) &= n!(x-a)^{n-k}/(n-k)! . \end{aligned}$$

Thus

$$\lim_{x \rightarrow a} [f(x) - Q(x)] = f(a) - Q(a) = 0,$$

$$\lim_{x \rightarrow a} [f'(x) - Q'(x)] = f'(a) - Q'(a) = 0,$$

.

.

$$\lim_{x \rightarrow a} [f^{(n-2)}(x) - Q^{(n-2)}(x)] = f^{(n-2)}(a) - Q^{(n-2)}(a) = 0.$$

and

$$\lim_{x \rightarrow a} g(x) = \lim_{x \rightarrow a} g'(x) = \cdots = \lim_{x \rightarrow a} g^{(n-2)}(x) = 0.$$

We may therefore apply l'Hôpital's Rule $n-1$ times to obtain

$$\lim_{x \rightarrow a} \frac{f(x) - Q(x)}{(x-a)^n} = \lim_{x \rightarrow a} \frac{f^{(n-1)}(x) - Q^{(n-1)}(x)}{n!(x-a)}.$$

Since Q is a polynomial of degree $n-1$, its $(n-1)$ st derivative is a constant; in fact, $Q^{(n-1)}(x) = f^{(n-1)}(a)$. Thus

$$\lim_{x \rightarrow a} \frac{f(x) - Q(x)}{(x-a)^n} = \lim_{x \rightarrow a} \frac{f^{(n-1)}(x) - f^{(n-1)}(a)}{n!(x-a)}$$

and this last limit is $f^{(n)}(a)/n!$ by definition of $f^{(n)}(a)$. ■

One simple consequence of Theorem 1 allows us to perfect the test for local maxima and minima which was developed in Chapter 11. If a is a critical point of f , then, according to Theorem 11-5, the function f has a local minimum at a if $f''(a) > 0$, and a local maximum at a if $f''(a) < 0$. If $f''(a) = 0$ no conclusion was possible, but it is conceivable that the sign of $f'''(a)$ might give

further information; and if $f'''(a) = 0$, then the sign of $f^{(4)}(a) = 0$ might be significant. Even more generally, we can ask what happens when

$$(*) \quad f'(a) = f''(a) = \cdots = f^{(n-1)}(a) = 0, \\ f^{(n)}(a) \neq 0.$$

The situation in this case can be guessed by examining the functions

$$f(x) = (x - a)^n, \\ g(x) = -(x - a)^n,$$

which satisfy (*). Notice (Figure 2) that if n is odd, then a is neither a local maximum nor a local minimum point for f or g . On the other hand, if n is even, then f , with a positive n th derivative, has a local minimum at a , while g , with a negative n th derivative, has a local maximum at a . Of all functions satisfying (*), these are about the simplest available; nevertheless they indicate the general situation exactly. In fact, the whole point of the next proof is that any function satisfying (*) looks very much like one of these functions, in a sense that is made precise by Theorem 1.

THEOREM 2 Suppose that

$$f'(a) = \cdots = f^{(n-1)}(a) = 0, \\ f^{(n)}(a) \neq 0.$$

- (1) If n is even and $f^{(n)}(a) > 0$, then f has a local minimum at a .
- (2) If n is even and $f^{(n)}(a) < 0$, then f has a local maximum at a .
- (3) If n is odd, then f has neither a local maximum nor a local minimum at a .

PROOF There is clearly no loss of generality in assuming that $f(a) = 0$, since neither the hypotheses nor the conclusion are affected if f is replaced by $f - f(a)$. Then, since the first $n - 1$ derivatives of f at a are 0, the Taylor polynomial $P_{n,a}$ of f is

$$P_{n,a}(x) = f(a) + \frac{f'(a)}{1!}(x - a) + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n \\ = \frac{f^{(n)}(a)}{n!}(x - a)^n.$$

Thus, Theorem 1 states that

$$0 = \lim_{x \rightarrow a} \frac{f(x) - P_{n,a}(x)}{(x - a)^n} = \lim_{x \rightarrow a} \left[\frac{f(x)}{(x - a)^n} - \frac{f^{(n)}(a)}{n!} \right].$$

Consequently, if x is sufficiently close to a , then

$$\frac{f(x)}{(x - a)^n} \text{ has the same sign as } \frac{f^{(n)}(a)}{n!}.$$

Suppose now that n is even. In this case $(x - a)^n > 0$ for all $x \neq a$. Since $f(x)/(x - a)^n$ has the same sign as $f^{(n)}(a)/n!$ for x sufficiently close to a , it follows

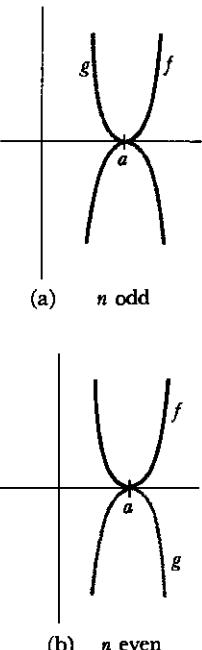


FIGURE 2

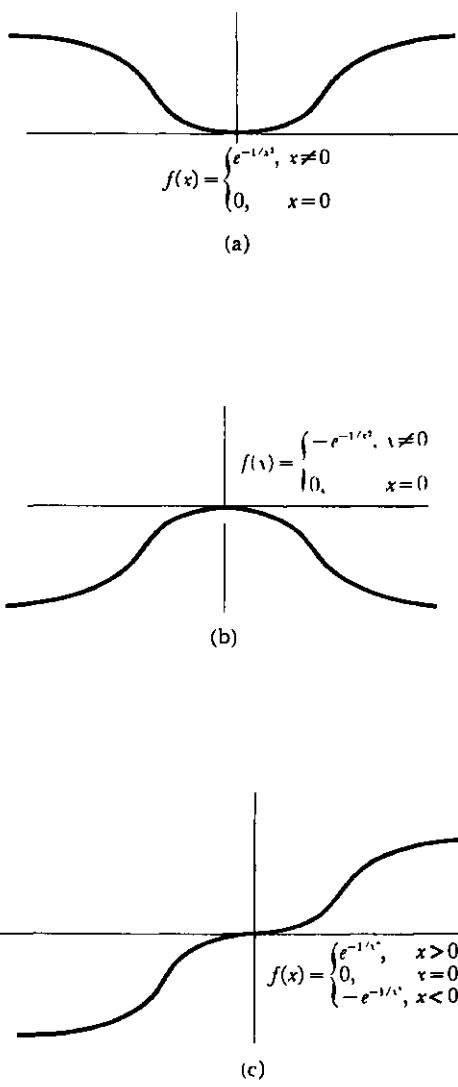


FIGURE 3

that $f(x)$ itself has the same sign as $f^n(a)/n!$ for x sufficiently close to a . If $f^{(n)}(a) > 0$, this means that

$$f(x) > 0 = f(a)$$

for x close to a . Consequently, f has a local minimum at a . A similar proof works for the case $f^{(n)}(a) < 0$.

Now suppose that n is odd. The same argument as before shows that if x is sufficiently close to a , then

$$\frac{f(x)}{(x-a)^n} \text{ always has the same sign.}$$

But $(x-a)^n > 0$ for $x > a$ and $(x-a)^n < 0$ for $x < a$. Therefore $f(x)$ has *different* signs for $x > a$ and $x < a$. This proves that f has neither a local maximum nor a local minimum at a . ■

Although Theorem 2 will settle the question of local maxima and minima for just about any function which arises in practice, it does have some theoretical limitations, because $f^{(k)}(a)$ may be 0 for all k . This happens (Figure 3(a)) for the function

$$f(x) = \begin{cases} e^{-1/x^2}, & x \neq 0 \\ 0, & x = 0, \end{cases}$$

which has a minimum at 0, and also for the negative of this function (Figure 3(b)), which has a maximum at 0. Moreover (Figure 3(c)), if

$$f(x) = \begin{cases} e^{-1/x^2}, & x > 0 \\ 0, & x = 0 \\ -e^{-1/x^2}, & x < 0, \end{cases}$$

then $f^{(k)}(0) = 0$ for all k , but f has neither a local minimum nor a local maximum at 0.

The conclusion of Theorem 1 is often expressed in terms of an important concept of “order of equality.” Two functions f and g are **equal up to order n at a** if

$$\lim_{x \rightarrow a} \frac{f(x) - g(x)}{(x-a)^n} = 0.$$

In the language of this definition, Theorem 1 says that the Taylor polynomial $P_{n,a,f}$ equals f up to order n at a . The Taylor polynomial might very well have been designed to make this fact true, because there is at most one polynomial of degree $\leq n$ with this property. This assertion is a consequence of the following elementary theorem.

THEOREM 3

Let P and Q be two polynomials in $(x-a)$, of degree $\leq n$, and suppose that P and Q are equal up to order n at a . Then $P = Q$.

PROOF

Let $R = P - Q$. Since R is a polynomial of degree $\leq n$, it is only necessary to

prove that if

$$R(x) = b_0 + \cdots + b_n(x - a)^n$$

satisfies

$$\lim_{x \rightarrow a} \frac{R(x)}{(x - a)^n} = 0,$$

then $R = 0$. Now the hypotheses on R surely imply that

$$\lim_{x \rightarrow a} \frac{R(x)}{(x - a)^i} = 0 \quad \text{for } 0 \leq i \leq n.$$

For $i = 0$ this condition reads simply $\lim_{x \rightarrow a} R(x) = 0$; on the other hand,

$$\begin{aligned} \lim_{x \rightarrow a} R(x) &= \lim_{x \rightarrow a} [b_0 + b_1(x - a) + \cdots + b_n(x - a)^n] \\ &= b_0. \end{aligned}$$

Thus $b_0 = 0$ and

$$R(x) = b_1(x - a) + \cdots + b_n(x - a)^n.$$

Therefore,

$$\frac{R(x)}{x - a} = b_1 + b_2(x - a) + \cdots + b_n(x - a)^{n-1}$$

and

$$\lim_{x \rightarrow a} \frac{R(x)}{x - a} = b_1.$$

Thus $b_1 = 0$ and

$$R(x) = b_2(x - a)^2 + \cdots + b_n(x - a)^n.$$

Continuing in this way, we find that

$$b_0 = \cdots = b_n = 0. \blacksquare$$

COROLLARY Let f be n -times differentiable at a , and suppose that P is a polynomial in $(x - a)$ of degree $\leq n$, which equals f up to order n at a . Then $P = P_{n,a,f}$.

PROOF Since P and $P_{n,a,f}$ both equal f up to order n at a , it is easy to see that P equals $P_{n,a,f}$ up to order n at a . Consequently, $P = P_{n,a,f}$ by the Theorem. \blacksquare

At first sight this corollary appears to have unnecessarily complicated hypotheses; it might seem that the existence of the polynomial P would automatically imply that f is sufficiently differentiable for $P_{n,a,f}$ to exist. But in fact this is not so. For example (Figure 4), suppose that

$$f(x) = \begin{cases} x^{n+1}, & x \text{ irrational} \\ 0, & x \text{ rational.} \end{cases}$$

If $P(x) = 0$, then P is certainly a polynomial of degree $\leq n$ which equals f up to order n at 0. On the other hand, $f'(a)$ does not exist for any $a \neq 0$, so $f''(0)$ is undefined.

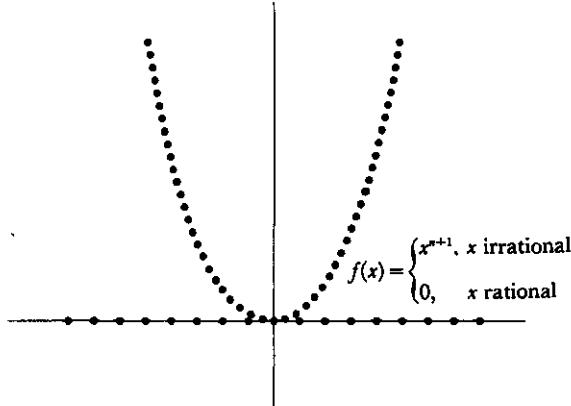


FIGURE 4

When f does have n derivatives at a , however, the corollary may provide a useful method for finding the Taylor polynomial of f . In particular, remember that our first attempt to find the Taylor polynomial for arctan ended in failure. The equation

$$\arctan x = \int_0^x \frac{1}{1+t^2} dt$$

suggests a promising method of finding a polynomial close to arctan—divide 1 by $1+t^2$, to obtain a polynomial plus a remainder:

$$\frac{1}{1+t^2} = 1 - t^2 + t^4 - t^6 + \cdots + (-1)^n t^{2n} + \frac{(-1)^{n+1} t^{2n+2}}{1+t^2}.$$

This formula, which can be checked easily by multiplying both sides by $1+t^2$, shows that

$$\begin{aligned} \arctan x &= \int_0^x 1 - t^2 + t^4 - \cdots + (-1)^n t^{2n} dt + (-1)^{n+1} \int_0^x \frac{t^{2n+2}}{1+t^2} dt \\ &= x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^n \frac{x^{2n+1}}{2n+1} + (-1)^{n+1} \int_0^x \frac{t^{2n+2}}{1+t^2} dt. \end{aligned}$$

According to our corollary, the polynomial which appears here will be the Taylor polynomial of degree $2n+1$ for arctan at 0, provided that

$$\lim_{x \rightarrow 0} \frac{\int_0^x \frac{t^{2n+2}}{1+t^2} dt}{x^{2n+1}} = 0.$$

Since

$$\left| \int_0^x \frac{t^{2n+2}}{1+t^2} dt \right| \leq \left| \int_0^x t^{2n+2} dt \right| = \frac{|x|^{2n+3}}{2n+3},$$

this is clearly true. Thus we have found that the Taylor polynomial of degree $2n+1$ for arctan at 0 is

$$P_{2n+1,0}(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^n \frac{x^{2n+1}}{2n+1}.$$

By the way, now that we have discovered the Taylor polynomials of \arctan , it is possible to work backwards and find $\arctan^{(k)}(0)$ for all k : Since

$$P_{2n+1,0}(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^n \frac{x^{2n+1}}{2n+1},$$

and since this polynomial is, by definition,

$$\arctan^{(0)}(0) + \arctan^{(1)}(0) + \frac{\arctan^{(2)}(0)}{2!} x^2 + \cdots + \frac{\arctan^{(2n+1)}(0)}{(2n+1)!} x^{2n+1},$$

we can find $\arctan^{(k)}(0)$ by simply equating the coefficients of x^k in these two polynomials:

$$\begin{aligned} \frac{\arctan^{(k)}(0)}{k!} &= 0 \quad \text{if } k \text{ is even,} \\ \frac{\arctan^{(2l+1)}(0)}{(2l+1)!} &= \frac{(-1)^l}{2l+1} \quad \text{or} \quad \arctan^{(2l+1)}(0) = (-1)^l \cdot (2l+1)!. \end{aligned}$$

A much more interesting fact emerges if we go back to the original equation

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^n \frac{x^{2n+1}}{2n+1} + (-1)^{n+1} \int_0^x \frac{t^{2n+2}}{1+t^2} dt,$$

and remember the estimate

$$\left| \int_0^x \frac{t^{2n+2}}{1+t^2} dt \right| \leq \frac{|x|^{2n+3}}{2n+3}.$$

When $|x| \leq 1$, this expression is at most $1/(2n+3)$, and we can make this as small as we like simply by choosing n large enough. In other words, for $|x| \leq 1$ we can use the Taylor polynomials for \arctan to compute $\arctan x$ as accurately as we like. The most important theorems about Taylor polynomials extend this isolated result to other functions, and the Taylor polynomials will soon play quite a new role. The theorems proved so far have always examined the behavior of the Taylor polynomial $P_{n,a}$ for fixed n , as x approaches a . Henceforth we will compare Taylor polynomials $P_{n,a}$ for fixed x , and different n . In anticipation of the coming theorem we introduce some new notation.

If f is a function for which $P_{n,a}(x)$ exists, we define the **remainder term** $R_{n,a}(x)$ by

$$\begin{aligned} f(x) &= P_{n,a}(x) + R_{n,a}(x) \\ &= f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_{n,a}(x). \end{aligned}$$

We would like to have an expression for $R_{n,a}(x)$ whose size is easy to estimate. There is such an expression, involving an integral, just as in the case for \arctan . One way to guess this expression is to begin with the case $n = 0$:

$$f(x) = f(a) + R_{0,a}(x).$$

416 Infinite Sequences and Infinite Series

The Fundamental Theorem of Calculus enables us to write

$$f(x) = f(a) + \int_a^x f'(t) dt,$$

so that

$$R_{0,a}(x) = \int_a^x f'(t) dt.$$

A similar expression for $R_{1,a}(x)$ can be derived from this formula using integration by parts in a rather tricky way: Let

$$u(t) = f'(t) \quad \text{and} \quad v(t) = t - x$$

(notice that x represents some fixed number in the expression for $v(t)$, so $v'(t) = 1$); then

$$\begin{aligned} \int_a^x f'(t) dt &= \int_a^x f'(t) \cdot 1 dt \\ &\quad \downarrow \quad \downarrow \\ u(t) \quad v'(t) &= u(t)v(t) \Big|_a^x - \int_a^x f''(t)(t-x) dt. \\ &\quad \downarrow \quad \downarrow \\ u'(t) \quad v(t) \end{aligned}$$

Since $v(x) = 0$, we obtain

$$\begin{aligned} f(x) &= f(a) + \int_a^x f'(t) dt \\ &= f(a) - u(a)v(a) + \int_a^x f''(t)(x-t) dt \\ &= f(a) + f'(a)(x-a) + \int_a^x f''(t)(x-t) dt. \end{aligned}$$

Thus

$$R_{1,a}(x) = \int_a^x f''(t)(x-t) dt.$$

It is hard to give any motivation for choosing $v(t) = t - x$, rather than $v(t) = t$. It just happens to be the choice which works out, the sort of thing one might discover after sufficiently many similar but futile manipulations. However, it is now easy to guess the formula for $R_{2,a}(x)$. If

$$u(t) = f''(t) \quad \text{and} \quad v(t) = \frac{-(x-t)^2}{2},$$

then $v'(t) = (x-t)$, so

$$\begin{aligned} \int_a^x f''(t)(x-t) dt &= u(t)v(t) \Big|_a^x - \int_a^x f'''(t) \cdot \frac{-(x-t)^2}{2} dt \\ &= \frac{f''(a)(x-a)^2}{2} + \int_a^x \frac{f'''(t)}{2}(x-t)^2 dt. \end{aligned}$$

This shows that

$$R_{2,a}(x) = \int_a^x \frac{f^{(3)}(t)}{2}(x-t)^2 dt.$$

You should now have little difficulty giving a rigorous proof, by induction, that

if $f^{(n+1)}$ is continuous on $[a, x]$, then

$$R_{n,a}(x) = \int_a^x \frac{f^{(n+1)}(t)}{n!} (x-t)^n dt.$$

From this formula, which is called the integral form of the remainder, it is possible (Problem 15) to derive two other important expressions for $R_{n,a}(x)$: the Cauchy form of the remainder,

$$R_{n,a}(x) = \frac{f^{(n+1)}(t)}{n!} (x-t)^n (x-a) \quad \text{for some } t \text{ in } (a, x),$$

and the Lagrange form of the remainder,

$$R_{n,a}(x) = \frac{f^{(n+1)}(t)}{(n+1)!} (x-a)^{n+1} \quad \text{for some } t \text{ in } (a, x).$$

In the proof of the next theorem (Taylor's Theorem) we will derive all three forms of the remainder in an entirely different way. One virtue of this proof (aside from its cleverness) is the fact that the Cauchy and Lagrange forms of the remainder will be proved without assuming the extra hypothesis that $f^{(n+1)}$ is continuous. In this way Taylor's Theorem appears as a direct generalization of the Mean Value Theorem, to which it reduces for $n = 0$, and which is the crucial tool used in the proof.

These remarks may suggest a strategy for proving Taylor's Theorem. Since $R_{n,a}(a) = 0$, we might try to apply the Mean Value Theorem to the expression

$$\frac{R_{n,a}(x)}{x-a} = \frac{R_{n,a}(x) - R_{n,a}(a)}{x-a}.$$

On second thought, however, this idea does not look very promising, since it is not at all clear how $f^{(n+1)}(t)$ is ever going to be involved in the answer. Indeed, if we take the most straightforward route, and differentiate both sides of the equation which defines $R_{n,a}$, we obtain

$$f'(x) = f'(a) + f''(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{(n-1)!} (x-a)^{n-1} + R_{n,a}'(x),$$

which is useless. The proper application of the Mean Value Theorem has a lot in common with the integration by parts proof outlined above. This proof involved the derivative of a function in which x denoted a number which was fixed. This is just how x will be treated in the following proof.

THEOREM 4 (TAYLOR'S THEOREM) Suppose that $f', \dots, f^{(n+1)}$ are defined on $[a, x]$, and that $R_{n,a}(x)$ is defined by

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!} (x-a)^n + R_{n,a}(x).$$

Then

$$(1) \quad R_{n,a}(x) = \frac{f^{(n+1)}(t)}{n!} (x-t)^n (x-a) \quad \text{for some } t \text{ in } (a, x).$$

$$(2) \quad R_{n,a}(x) = \frac{f^{(n+1)}(t)}{(n+1)!} (x-a)^{n+1} \quad \text{for some } t \text{ in } (a, x).$$

Moreover, if $f^{(n+1)}$ is integrable on $[a, x]$, then

$$(3) \quad R_{n,a}(x) = \int_a^x \frac{f^{(n+1)}(t)}{n!} (x-t)^n dt.$$

(If $x < a$, then the hypothesis should state that f is $(n+1)$ -times differentiable on $[x, a]$; the number t in (1) and (2) will then be in (x, a) , while (3) will remain true as stated, provided that $f^{(n+1)}$ is integrable on $[x, a]$.)

PROOF For each number t in $[a, x]$ we have

$$f(x) = f(t) + f'(t)(x-t) + \cdots + \frac{f^{(n)}(t)}{n!} (x-t)^n + R_{n,t}(x).$$

Let us denote the number $R_{n,t}(x)$ by $S(t)$; the function S is defined on $[a, x]$, and we have

$$(*) \quad f(x) = f(t) + f'(t)(x-t) + \cdots + \frac{f^{(n)}(t)}{n!} (x-t)^n + S(t) \quad \text{for all } t \text{ in } [a, x].$$

We will now differentiate both sides of this equation, which asserts the equality of two functions: the one whose value at t is $f(x)$, and the one whose value at t is

$$f(t) + \cdots + \frac{f^{(n)}(t)}{n!} (x-t)^n + S(t).$$

(In common parlance we are considering both sides of $(*)$ "as a function of t ".) Just to make sure that the letter x causes no confusion, notice that if

$$g(t) = f(x) \quad \text{for all } t,$$

then

$$g'(t) = 0 \quad \text{for all } t;$$

and if

$$g(t) = \frac{f^{(k)}(t)}{k!} (x-t)^k,$$

then

$$\begin{aligned} g'(t) &= \frac{f^{(k)}(t)}{k!} k(x-t)^{k-1} (-1) + \frac{f^{(k+1)}(t)}{k!} (x-t)^k \\ &= -\frac{f^{(k)}(t)}{(k-1)!} (x-t)^{k-1} + \frac{f^{(k+1)}(t)}{k!} (x-t)^k. \end{aligned}$$

Applying these formulas to each term of (*), we obtain

$$\begin{aligned} 0 = f'(t) + \left[-f'(t) + \frac{f''(t)}{1!}(x-t) \right] + \left[\frac{-f''(t)}{1!}(x-t) + \frac{f^{(3)}(t)}{2!}(x-t)^2 \right] \\ + \cdots + \left[\frac{-f^{(n)}(t)}{(n-1)!}(x-t)^{n-1} + \frac{f^{(n+1)}(t)}{n!}(x-t)^n \right] + S'(t). \end{aligned}$$

In this beautiful formula practically everything in sight cancels out, and we obtain

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

Now we can apply the Mean Value Theorem to the function S on $[a, x]$: there is some t in (a, x) such that

$$\frac{S(x) - S(a)}{x - a} = S'(t) = -\frac{f^{(n+1)}(t)}{n!}(x-t)^n.$$

Remember that

$$S(t) = R_{n,t}(x);$$

this means in particular that

$$\begin{aligned} S(x) &= R_{n,x}(x) = 0, \\ S(a) &= R_{n,a}(x). \end{aligned}$$

Thus

$$\frac{0 - R_{n,a}(x)}{x - a} = -\frac{f^{(n+1)}(t)}{n!}(x-t)^n$$

or

$$R_{n,a}(x) = \frac{f^{(n+1)}(t)}{n!}(x-t)^n(x-a);$$

this is the Cauchy form of the remainder.

To derive the Lagrange form we apply the Cauchy Mean Value Theorem to the functions S and $g(t) = (x-t)^{n+1}$: there is some t in (a, x) such that

$$\frac{S(x) - S(a)}{g(x) - g(a)} = \frac{\frac{S'(t)}{g'(t)}}{\frac{-(n+1)(x-t)^n}{-(n+1)(x-t)^n}} = \frac{-\frac{f^{(n+1)}(t)}{n!}(x-t)^n}{-(n+1)(x-t)^n}.$$

Thus

$$\frac{R_{n,a}(x)}{(x-a)^{n+1}} = \frac{f^{(n+1)}(t)}{(n+1)!}$$

or

$$R_{n,a}(x) = \frac{f^{(n+1)}(t)}{(n+1)!}(x-a)^{n+1},$$

which is the Lagrange form.

Finally, if $f^{(n+1)}$ is integrable on $[a, x]$, then

$$S(x) - S(a) = \int_a^x S'(t) dt = - \int_a^x \frac{f^{(n+1)}(t)}{n!} (x-t)^n dt$$

or

$$R_{n,a}(x) = \int_a^x \frac{f^{(n+1)}(t)}{n!} (x-t)^n dt. \blacksquare$$

Although the Lagrange and Cauchy forms of the remainder are more than theoretical curiosities (see, e.g., Problem 23-18), the integral form of the remainder will usually be quite adequate. If this form is applied to the functions \sin , \cos , and e^x , with $a = 0$, Taylor's Theorem yields the following formulas:

$$\begin{aligned} \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} \\ &\quad + \int_0^x \frac{\sin^{(2n+2)}(t)}{(2n+1)!} (x-t)^{2n+1} dt, \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots + (-1)^n \frac{x^{2n}}{(2n)!} + \int_0^x \frac{\cos^{(2n+1)}(t)}{(2n)!} (x-t)^{2n} dt, \\ e^x &= 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \int_0^x \frac{e^t}{n!} (x-t)^n dt. \end{aligned}$$

To evaluate any of these integrals explicitly would be supreme foolishness—the answer of course will be exactly the difference of the left side and all the other terms on the right side! To *estimate* these integrals, however, is both easy and worthwhile.

The first two integrals are especially easy. Since

$$|\sin^{(2n+2)}(t)| \leq 1 \quad \text{for all } t,$$

we have

$$\left| \int_0^x \frac{\sin^{(2n+2)}(t)}{(2n+1)!} (x-t)^{2n+1} dt \right| \leq \frac{1}{(2n+1)!} \left| \int_0^x (x-t)^{2n+1} dt \right|.$$

Since

$$\begin{aligned} \int_0^x (x-t)^{2n+1} dt &= \frac{-(x-t)^{2n+2}}{2n+2} \Big|_{t=0}^{t=x} \\ &= \frac{x^{2n+2}}{2n+2} \end{aligned}$$

we conclude that

$$\left| \int_0^x \frac{\sin^{(2n+2)}(t)}{(2n+1)!} (x-t)^{2n+1} dt \right| \leq \frac{|x|^{2n+2}}{(2n+2)!}.$$

Similarly, we can show that

$$\left| \int_0^x \frac{\cos^{(2n+1)}(t)}{(2n)!} (x-t)^{2n} dt \right| \leq \frac{|x|^{2n+1}}{(2n+1)!}.$$

These estimates are particularly interesting, because (as proved in Chapter 16) for any $\varepsilon > 0$ we can make

$$\frac{x^n}{n!} < \varepsilon$$

by choosing n large enough (how large n must be will depend on x). This enables us to compute $\sin x$ to any degree of accuracy desired simply by evaluating the proper Taylor polynomial $P_{n,0}(x)$. For example, suppose we wish to compute $\sin 2$ with an error of less than 10^{-4} . Since

$$\sin 2 = P_{2n+1,0}(2) + R, \quad \text{where } |R| \leq \frac{2^{2n+2}}{(2n+2)!},$$

we can use $P_{2n+1,0}(2)$ as our answer, provided that

$$\frac{2^{2n+2}}{(2n+2)!} < 10^{-4}.$$

A number n with this property can be found by a straightforward search—it obviously helps to have a table of values for $n!$ and 2^n (see page 428). In this case it happens that $n = 5$ works, so that

$$\begin{aligned} \sin 2 &= P_{11,0}(2) + R \\ &= 2 - \frac{2^3}{3!} + \frac{2^5}{5!} - \frac{2^7}{7!} + \frac{2^9}{9!} - \frac{2^{11}}{11!} + R, \\ &\quad \text{where } |R| < 10^{-4}. \end{aligned}$$

It is even easier to calculate $\sin 1$ approximately, since

$$\sin 1 = P_{2n+1,0}(1) + R, \quad \text{where } |R| < \frac{1}{(2n+2)!}.$$

To obtain an error less than ε we need only find an n such that

$$\frac{1}{(2n+2)!} < \varepsilon,$$

and this requires only a brief glance at a table of factorials. (Moreover, the individual terms of $P_{2n+1,0}(1)$ will be easier to handle.)

For very small x the estimates will be even easier. For example,

$$\sin \frac{1}{10} = P_{2n+1,0}\left(\frac{1}{10}\right) + R, \quad \text{where } |R| < \frac{1}{10^{2n+2}(2n+2)!}.$$

To obtain $|R| < 10^{-10}$ we can clearly take $n = 4$ (and we could even get away with $n = 3$). These methods are actually used to compute tables of \sin and \cos . A high-speed computer can compute $P_{2n+1,0}(x)$ for many different x in almost no time at all.

Estimating the remainder for e^x is only slightly harder. For simplicity assume that $x \geq 0$ (the estimates for $x \leq 0$ are obtained in Problem 10). On the interval $[0, x]$ the maximum value of e^t is e^x , since \exp is increasing, so

$$\int_0^x \frac{e^t}{n!} (x-t)^n dt \leq \frac{e^x}{n!} \int_0^x (x-t)^n dt = \frac{e^x x^{n+1}}{(n+1)!}.$$

Since we already know that $e < 4$, we have

$$\frac{e^x x^{n+1}}{(n+1)!} < \frac{4^x x^{n+1}}{(n+1)!},$$

which can be made as small as desired by choosing n sufficiently large. How large n must be will depend on x (and the factor 4^x will make things more difficult). Once again, the estimates are easier for small x . If $0 \leq x \leq 1$, then

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + R, \quad \text{where } 0 < R < \frac{4}{(n+1)!}.$$

(The inequality $0 < R$ follows immediately from the integral form for R .) In particular, if $n = 4$, then

$$0 < R < \frac{4}{5!} < \frac{1}{10},$$

so

$$\begin{aligned} e = e^1 &= 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + R, \quad \text{where } 0 < R < \frac{1}{10} \\ &= \frac{65}{24} + R \\ &= 2 + \frac{17}{24} + R, \end{aligned}$$

which shows that

$$2 < e < 3.$$

(This then shows that

$$0 < R < \frac{3^x x^{n+1}}{(n+1)!},$$

allowing us to improve our estimate of R slightly.) By taking $n = 7$ you can compute that the first 3 decimals for e are

$$e = 2.718 \dots$$

(you should check that $n = 7$ does give this degree of accuracy, but it would be cruel to insist that you actually do the computations).

The function \arctan is also important but, as you may recall, an expression for $\arctan^{(k)}(x)$ is hopelessly complicated, so that the integral form of the remainder is useless. On the other hand, our derivation of the Taylor polynomial for \arctan automatically provided a formula for the remainder:

$$\arctan x = x - \frac{x^3}{3} + \cdots + \frac{(-1)^n x^{2n+1}}{2n+1} + \int_0^x \frac{(-1)^{n+1} t^{2n+2}}{1+t^2} dt.$$

As we have already estimated,

$$\left| \int_0^x \frac{(-1)^{n+1} t^{2n+2}}{1+t^2} dt \right| \leq \left| \int_0^x t^{2n+2} dt \right| = \frac{|x|^{2n+3}}{2n+3}.$$

For the moment we will consider only numbers x with $|x| \leq 1$. In this case, the remainder term can clearly be made as small as desired by choosing n sufficiently large. In particular,

$$\arctan 1 = 1 - \frac{1}{3} + \frac{1}{5} - \cdots + \frac{(-1)^n}{2n+1} + R, \quad \text{where } |R| < \frac{1}{2n+3}.$$

With this estimate it is easy to find an n which will make the remainder less than any preassigned number; on the other hand, n will usually have to be so large as to make computations hopelessly long. To obtain a remainder $< 10^{-4}$, for example, we must take $n > (10^4 - 3)/2$. This is really a shame, because $\arctan 1 = \pi/4$, so the Taylor polynomial for \arctan should allow us to compute π . Fortunately, there are some clever tricks which enable us to surmount these difficulties. Since

$$|R_{2n+1,0}(x)| < \frac{|x|^{2n+3}}{2n+3},$$

much smaller n 's will work for only somewhat smaller x 's. The trick for computing π is to express $\arctan 1$ in terms of $\arctan x$ for smaller x ; Problem 6 shows how this can be done in a convenient way.

The Taylor polynomial for the function $f(x) = \log(x+1)$ at $a = 1$ is best handled in the same manner as the Taylor polynomial for \arctan . Although the integral form of the remainder for f is not hard to write down, it is difficult to estimate. On the other hand, we obtain a simple formula if we begin with the equation

$$\frac{1}{1+t} = 1 - t + t^2 - \cdots + (-1)^{n-1} t^{n-1} + \frac{(-1)^n t^n}{1+t};$$

this implies that

$$\begin{aligned} \log(1+x) &= \int_0^x \frac{1}{1+t} dt = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + (-1)^{n-1} \frac{x^n}{n} \\ &\quad + (-1)^n \int_0^x \frac{t^n}{1+t} dt, \end{aligned}$$

for all $x > -1$. If $x \geq 0$, then

$$\int_0^x \frac{t^n}{t+1} dt \leq \int_0^x t^n dt = \frac{x^{n+1}}{n+1},$$

and there is a slightly more complicated estimate when $-1 < x < 0$ (Problem 11). For this function the remainder term can be made as small as desired by choosing n sufficiently large, provided that $-1 < x \leq 1$.

The behavior of the remainder terms for arctan and $f(x) = \log(x+1)$ is quite another matter when $|x| > 1$. In this case, the estimates

$$|R_{2n+1,0}(x)| < \frac{|x|^{2n+3}}{2n+3} \quad \text{for arctan,}$$

$$|R_{n,0}(x)| < \frac{x^{n+1}}{n+1} \quad (x > 0) \text{ for } f,$$

are of no use, because when $|x| > 1$ the bounds x^m/m become large as m becomes large. This predicament is unavoidable, and is not just a deficiency of our estimates. It is easy to get estimates in the other direction which show that the remainders actually do remain large. To obtain such an estimate for arctan, note that if t is in $[0, x]$ (or in $[x, 0]$ if $x < 0$), then

$$1 + t^2 \leq 1 + x^2 \leq 2x^2, \quad \text{if } |x| \geq 1,$$

so

$$\left| \int_0^x \frac{t^{2n+2}}{1+t^2} dt \right| \geq \frac{1}{2x^2} \left| \int_0^x t^{2n+2} dt \right| = \frac{|x|^{2n+1}}{4n+6}.$$

Similarly, if $x > 0$, then for t in $[0, x]$ we have

$$1 + t \leq 1 + x \leq 2x, \quad \text{if } x \geq 1,$$

so

$$\int_0^x \frac{t^n}{t+1} dt \geq \frac{1}{2x} \int_0^x t^n dt = \frac{x^n}{2n+2}.$$

These estimates show that if $|x| > 1$, then the remainder terms become large as n becomes large. In other words, for $|x| > 1$, the Taylor polynomials for arctan and f are of no use whatsoever in computing $\arctan x$ and $\log(x+1)$. This is no tragedy, because the values of these functions can be found for any x once they are known for all x with $|x| < 1$.

This same situation occurs in a spectacular way for the function

$$f(x) = \begin{cases} e^{-1/x^2}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

We have already seen that $f^{(k)}(0) = 0$ for every natural number k . This means that the Taylor polynomial $P_{n,0}$ for f is

$$\begin{aligned} P_{n,0}(x) &= f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n \\ &= 0. \end{aligned}$$

In other words, the remainder term $R_{n,0}(x)$ always equals $f(x)$, and the Taylor polynomial is useless for computing $f(x)$, except for $x = 0$. Eventually we will be able to offer some explanation for the behavior of this function, which is such a disconcerting illustration of the limitations of Taylor's Theorem.

The word "compute" has been used so often in connection with our estimates for the remainder term, that the significance of Taylor's Theorem might be misconstrued. It is true that Taylor's Theorem is an almost ideal computational aid

(despite its ignominious failure in the previous example), but it has equally important theoretical consequences. Most of these will be developed in succeeding chapters, but two proofs will illustrate some ways in which Taylor's Theorem may be used. The first illustration will be particularly impressive to those who have waded through the proof, in Chapter 16, that π is irrational.

THEOREM 5 e is irrational.

PROOF We know that, for any n ,

$$e = e^1 = 1 + \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{n!} + R_n, \quad \text{where } 0 < R_n < \frac{3}{(n+1)!}.$$

Suppose that e were rational, say $e = a/b$, where a and b are positive integers. Choose $n > b$ and also $n > 3$. Then

$$\frac{a}{b} = 1 + 1 + \frac{1}{2!} + \cdots + \frac{1}{n!} + R_n,$$

so

$$\frac{n!a}{b} = n! + n! + \frac{n!}{2!} + \cdots + \frac{n!}{n!} + n!R_n.$$

Every term in this equation other than $n!R_n$ is an integer (the left side is an integer because $n > b$). Consequently, $n!R_n$ must be an integer also. But

$$0 < R_n < \frac{3}{(n+1)!},$$

so

$$0 < n!R_n < \frac{3}{n+1} < \frac{3}{4} < 1,$$

which is impossible for an integer. ■

The second illustration is merely a straightforward demonstration of a fact proved in Chapter 15: If

$$\begin{aligned} f'' + f &= 0, \\ f(0) &= 0, \\ f'(0) &= 0, \end{aligned}$$

then $f = 0$. To prove this, observe first that $f^{(k)}$ exists for every k ; in fact

$$\begin{aligned} f^{(3)} &= (f'')' = -f', \\ f^{(4)} &= (f^3)' = (-f')' = -f'' = f, \\ f^{(5)} &= (f^{(4)})' = f', \\ &\text{etc.} \end{aligned}$$

This shows, not only that all $f^{(k)}$ exist, but also that there are at most 4 different ones: f , f' , $-f$, $-f'$. Since $f(0) = f'(0) = 0$, all $f^{(k)}(0)$ are 0. Now Taylor's Theorem states, for any n , that

$$f(x) = \int_0^x \frac{f^{(n+1)}(t)}{n!} (x-t)^n dt.$$

Each function $f^{(n+1)}$ is continuous (since $f^{(n+2)}$ exists), so for any particular x there is a number M such that

$$|f^{(n+1)}(t)| \leq M \quad \text{for } 0 \leq t \leq x, \text{ and all } n$$

(we can add the phrase "and all n " because there are only four different $f^{(k)}$). Thus

$$|f(x)| \leq M \left| \int_0^x \frac{(x-t)^n}{n!} dt \right| = \frac{M|x|^{n+1}}{(n+1)!}.$$

Since this is true for every n , and since $|x|^n/n!$ can be made as small as desired by choosing n sufficiently large, this shows that $|f(x)| \leq \varepsilon$ for any $\varepsilon > 0$; consequently, $f(x) = 0$.

The other uses to which Taylor's Theorem will be put in succeeding chapters are closely related to the computational considerations which have concerned us for much of this chapter. If the remainder term $R_{n,a}(x)$ can be made as small as desired by choosing n sufficiently large, then $f(x)$ can be computed to any degree of accuracy desired by using the polynomials $P_{n,a}(x)$. As we require greater and greater accuracy we must add on more and more terms. If we are willing to add up infinitely many terms (in theory at least!), then we ought to be able to ignore the remainder completely. There should be "infinite sums" like

$$\begin{aligned} \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots, \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \dots, \\ e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots, \\ \arctan x &= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots \quad \text{if } |x| \leq 1, \\ \log(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \quad \text{if } -1 < x \leq 1. \end{aligned}$$

We are almost completely prepared for this step. Only one obstacle remains—we have never even defined an infinite sum. Chapters 22 and 23 contain the necessary definitions.

PROBLEMS

1. Find the Taylor polynomials (of the indicated degree, and at the indicated point) for the following functions.

- (i) $f(x) = e^{ex}$; degree 3, at 0.
- (ii) $f(x) = e^{\sin x}$ degree 3, at 0.
- (iii) \sin ; degree $2n$, at $\frac{\pi}{2}$.
- (iv) \cos ; degree $2n$, at π .
- (v) \exp ; degree n , at 1.
- (vi) \log ; degree n , at 2.
- (vii) $f(x) = x^5 + x^3 + x$; degree 4, at 0.
- (viii) $f(x) = x^5 + x^3 + x$; degree 4, at 1.
- (ix) $f(x) = \frac{1}{1+x^2}$; degree $2n+1$, at 0.
- (x) $f(x) = \frac{1}{1+x}$; degree n , at 0.

2. Write each of the following polynomials in x as a polynomial in $(x - 3)$. (It is only necessary to compute the Taylor polynomial at 3, of the same degree as the original polynomial. Why?)

- (i) $x^2 - 4x - 9$.
- (ii) $x^4 - 12x^3 + 44x^2 + 2x + 1$.
- (iii) x^5 .
- (iv) $ax^2 + bx + c$.

3. Write down a sum (using \sum notation) which equals each of the following numbers to within the specified accuracy. To minimize needless computation, consult the tables for 2^n and $n!$ on the next page.

- (i) $\sin 1$; error $< 10^{-17}$.
- (ii) $\sin 2$; error $< 10^{-12}$.
- (iii) $\sin \frac{1}{2}$; error $< 10^{-20}$.
- (iv) e ; error $< 10^{-4}$.
- (v) e^2 ; error $< 10^{-5}$.

n	2^n	$n!$
1	2	1
2	4	2
3	8	6
4	16	24
5	32	120
6	64	720
7	128	5,040
8	256	40,430
9	512	362,880
10	1,024	3,628,800
11	2,048	39,916,800
12	4,096	479,001,600
13	8,192	6,227,020,800
14	16,384	87,178,291,200
15	32,768	1,307,674,368,000
16	65,536	20,922,789,888,000
17	131,072	355,687,428,096,000
18	262,144	6,402,373,705,728,000
19	524,888	121,645,100,408,832,000
20	1,048,576	2,432,902,008,176,640,000

- *4. This problem is similar to the previous one, except that the errors demanded are so small that the tables cannot be used. You will have to do a little thinking, and in some cases it may be necessary to consult the proof, in Chapter 16, that $x^n/n!$ can be made small by choosing n large—the proof actually provides a method for finding the appropriate n . In the previous problem it was possible to find rather short sums; in fact, it was possible to find the smallest n which makes the estimate of the remainder given by Taylor's Theorem less than the desired error. But in this problem finding *any* specific sum is a moral victory (provided you can demonstrate that the sum works).
- (i) $\sin 1$; error $< 10^{-(10^{10})}$.
(ii) e ; error $< 10^{-1,000}$.
(iii) $\sin 10$; error $< 10^{-20}$.
(iv) e^{10} ; error $< 10^{-30}$.
(v) $\arctan \frac{1}{10}$; error $< 10^{-(10^{10})}$.
5. (a) In Problem 11-38 you showed that the equation $x^2 = \cos x$ has precisely two solutions. Use the Taylor polynomial of \cos to show that the solutions are approximately $\pm\sqrt{2/3}$, and find bounds on the error.
(b) Similarly, estimate the solutions of the equation $2x^2 = x \sin x + \cos^2 x$.

6. (a) Prove, using Problem 15-9, that

$$\frac{\pi}{4} = \arctan \frac{1}{2} + \arctan \frac{1}{3},$$

$$\frac{\pi}{4} = 4 \arctan \frac{1}{5} - \arctan \frac{1}{239}.$$

- (b) Show that $\pi = 3.14159 \dots$. (Every budding mathematician should verify a few decimals of π , but the purpose of this exercise is not to set you off on an immense calculation. If the second expression in part (a) is used, the first 5 decimals for π can be computed with remarkably little work.)
7. For every number α , and every natural number n , we define the “binomial coefficient”

$$\binom{\alpha}{n} = \frac{\alpha(\alpha-1) \cdot \dots \cdot (\alpha-n+1)}{n!},$$

and we define $\binom{\alpha}{0} = 1$, as usual. If α is not an integer, then $\binom{\alpha}{n}$ is never 0, and alternates in sign for $n > \alpha$. Show that the Taylor polynomial of degree n for $f(x) = (1+x)^\alpha$ at 0 is $P_{n,0}(x) = \sum_{k=0}^n \binom{\alpha}{k} x^k$, and that the Cauchy and Lagrange forms of the remainder are the following:

Cauchy form:

$$\begin{aligned} R_{n,0}(x) &= \frac{\alpha(\alpha-1) \cdot \dots \cdot (\alpha-n)}{n!} x(x-t)^n (1+t)^{\alpha-n-1} \\ &= \frac{\alpha(\alpha-1) \cdot \dots \cdot (\alpha-n)}{n!} x(1+t)^{\alpha-1} \left(\frac{x-t}{1+t} \right)^n \\ &= (n+1) \binom{\alpha}{n+1} x(1+t)^{\alpha-1} \left(\frac{x-t}{1+t} \right)^n, \quad t \text{ in } [0, x] \text{ or } [x, 0]. \end{aligned}$$

Lagrange form:

$$\begin{aligned} R_{n,0}(x) &= \frac{\alpha(\alpha-1) \cdot \dots \cdot (\alpha-n)}{(n+1)!} x^{n+1} (1+t)^{\alpha-n-1} \\ &= \binom{\alpha}{n+1} x^{n+1} (1+t)^{\alpha-n-1}, \quad t \text{ in } [0, x] \text{ or } [x, 0]. \end{aligned}$$

Estimates for these remainder terms are rather difficult to handle, and are postponed to Problem 23-18.

8. Suppose that a_i and b_i are the coefficients in the Taylor polynomials at a of f and g , respectively. In other words, $a_i = f^{(i)}(a)/i!$ and $b_i = g^{(i)}(a)/i!$. Find the coefficients c_i of the Taylor polynomials at a of the following functions, in terms of the a_i 's and b_i 's.

- (i) $f + g$.
(ii) fg .

(iii) f' .

(iv) $h(x) = \int_a^x f(t) dt.$

(v) $k(x) = \int_0^x f(t) dt.$

9. (a) Prove that the Taylor polynomial of $f(x) = \sin(x^2)$ of degree $4n+2$ at 0 is

$$x^2 - \frac{x^6}{3!} + \frac{x^{10}}{5!} - \cdots + (-1)^n \frac{x^{4n+2}}{(2n+1)!}.$$

Hint: If P is the Taylor polynomial of degree $2n+1$ for \sin at 0, then $\sin x = P(x) + R(x)$, where $\lim_{x \rightarrow 0} R(x)/x^{2n+1} = 0$. What does this imply about $\lim_{x \rightarrow 0} R(x^2)/x^{4n+2}$?

- (b) Find $f^{(k)}(0)$ for all k .
 (c) In general, if $f(x) = g(x^n)$, find $f^{(k)}(0)$ in terms of the derivatives of g at 0.

10. Prove that if $x \leq 0$, then

$$\left| \int_0^x \frac{e^t}{n!} (x-t)^n dt \right| \leq \frac{|x|^{n+1}}{(n+1)!}.$$

11. Prove that if $-1 < x \leq 0$, then

$$\left| \int_0^x \frac{t^n}{1+t} dt \right| \leq \frac{|x|^{n+1}}{(1+x)(n+1)}.$$

- *12. (a) Show that if $|g'(x)| \leq M|x-a|^n$ for $|x-a| < \delta$, then $|g(x) - g(a)| \leq M|x-a|^{n+1}/(n+1)$ for $|x-a| < \delta$.
 (b) Use part (a) to show that if $\lim_{x \rightarrow a} g'(x)/(x-a)^n = 0$, then

$$\lim_{x \rightarrow a} \frac{g(x)}{(x-a)^{n+1}} = 0.$$

- (c) Show that if $g(x) = f(x) - P_{n,a,f}(x)$, then $g'(x) = f'(x) - P_{n-1,a,f'}(x)$.
 (d) Give an inductive proof of Theorem 1, without using l'Hôpital's Rule.

13. Deduce Theorem 1 as a corollary of Taylor's Theorem, with any form of the remainder. (The catch is that it will be necessary to assume one more derivative than in the hypotheses for Theorem 1.)
14. Deduce the Cauchy and Lagrange forms of the remainder from the integral form, using Problem 13-23. There will be the same catch as in Problem 13.

15. (a) Suppose that f is twice differentiable on $(0, \infty)$ and that $|f(x)| \leq M_0$ for all $x > 0$, while $|f''(x)| \leq M_2$ for all $x > 0$. Prove that for all $x > 0$ we have

$$|f'(x)| \leq \frac{2}{h} M_0 + \frac{h}{2} M_2 \quad \text{for all } h > 0.$$

- (b) Show that for all $x > 0$ we have

$$|f'(x)| \leq 2\sqrt{M_0 M_2}.$$

- (c) If f is twice differentiable on $(0, \infty)$, f'' is bounded, and $f(x)$ approaches 0 as $x \rightarrow \infty$, then also $f'(x)$ approaches 0 as $x \rightarrow \infty$.
- (d) If $\lim_{x \rightarrow \infty} f(x)$ exists and $\lim_{x \rightarrow \infty} f''(x)$ exists, then $\lim_{x \rightarrow \infty} f''(x) = \lim_{x \rightarrow \infty} f'(x) = 0$. (Compare Problem 11-31.)

16. (a) Prove that if $f''(a)$ exists, then

$$f''(a) = \lim_{h \rightarrow 0} \frac{f(a+h) + f(a-h) - 2f(a)}{h^2}.$$

The limit on the right is called the *Schwarz second derivative* of f at a . Hint: Use the Taylor polynomial $P_{2,a}(x)$ with $x = a + h$ and with $x = a - h$.

- (b) Let $f(x) = x^2$ for $x \geq 0$, and $-x^2$ for $x \leq 0$. Show that

$$\lim_{h \rightarrow 0} \frac{f(0+h) + f(0-h) - 2f(0)}{h^2}$$

exists, even though $f''(0)$ does not.

- (c) Prove that if f has a local maximum at a , and the Schwarz second derivative of f at a exists, then it is ≤ 0 .
- (d) Prove that if $f'''(a)$ exists, then

$$\frac{f'''(a)}{3} = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a-h) - 2hf'(x)}{h^3}.$$

17. Use the Taylor polynomial $P_{1,a,f}$, together with the remainder, to prove a weak form of Theorem 2 of the Appendix to Chapter 11: If $f'' > 0$, then the graph of f always lies above the tangent line of f , except at the point of contact.

- *18. Problem 18-43 presented a rather complicated proof that $f = 0$ if $f'' - f = 0$ and $f(0) = f'(0) = 0$. Give another proof, using Taylor's Theorem. (This problem is really a preliminary skirmish before doing battle with the general case in Problem 19, and is meant to convince you that Taylor's Theorem is a good tool for tackling such problems, even though tricks work out more neatly for special cases.)

- **19. Consider a function f which satisfies the differential equation

$$f^{(n)} = \sum_{j=0}^{n-1} a_j f^{(j)},$$

for certain numbers a_0, \dots, a_{n-1} . Several special cases have already received detailed treatment, either in the text or in other problems; in particular, we have found all functions satisfying $f' = f$, or $f'' + f = 0$, or $f'' - f = 0$. The trick in Problem 18-42 enables us to find many solutions for such equations, but doesn't say whether these are the only solutions. This requires a *uniqueness* result, which will be supplied by this problem. At the end you will find some (necessarily sketchy) remarks about the general solution.

- (a) Derive the following formula for $f^{(n+1)}$ (let us agree that " a_{-1} " will be 0):

$$f^{(n+1)} = \sum_{j=0}^{n-1} (a_{j-1} + a_{n-1}a_j) f^{(j)}.$$

- (b) Deduce a formula for $f^{(n+2)}$.

The formula in part (b) is not going to be used; it was inserted only to convince you that a general formula for $f^{(n+k)}$ is out of the question. On the other hand, as part (c) shows, it is not very hard to obtain estimates on the size of $f^{(n+k)}(x)$.

- (c) Let $N = \max(1, |a_0|, \dots, |a_{n-1}|)$. Then $|a_{j-1} + a_{n-1}a_j| \leq 2N^2$; this means that

$$f^{(n+1)} = \sum_{j=0}^{n-1} b_j^{-1} f^{(j)}, \quad \text{where } |b_j^{-1}| \leq 2N^2.$$

Show that

$$f^{(n+2)} = \sum_{j=0}^{n-1} b_j^{-2} f^{(j)}, \quad \text{where } |b_j^{-2}| \leq 4N^3,$$

and, more generally,

$$f^{(n+k)} = \sum_{j=0}^{n-1} b_j^{-k} f^{(j)}, \quad \text{where } |b_j^{-k}| \leq 2N^{k+1}.$$

- (d) Conclude from part (c) that, for any particular number x , there is a number M such that

$$|f^{(n+k)}(x)| \leq M \cdot 2^k N^{k+1} \quad \text{for all } k.$$

- (e) Now suppose that $f(0) = f'(0) = \dots = f^{(n-1)}(0) = 0$. Show that

$$|f(x)| \leq \frac{M \cdot 2^{k+1} N^{k+2} |x|^{n+k+1}}{(n+k+1)!} \leq \frac{M \cdot |2Nx|^{n+k+1}}{(n+k+1)!},$$

and conclude that $f = 0$.

- (f) Show that if f_1 and f_2 are both solutions of the differential equation

$$f^{(n)} = \sum_{j=0}^{n-1} a_j f^{(j)},$$

and $f_1^{(j)}(0) = f_2^{(j)}(0)$ for $0 \leq j \leq n - 1$, then $f_1 = f_2$.

In other words, the solutions of this differential equation are determined by the “initial conditions” (the values $f^{(j)}(0)$ for $0 \leq j \leq n - 1$). This means that we can find *all* solutions once we can find enough solutions to obtain any given set of initial conditions. If the equation

$$x^n - a_{n-1}x^{n-1} - \cdots - a_0 = 0$$

has n distinct roots $\alpha_1, \dots, \alpha_n$, then any function of the form

$$f(x) = c_1 e^{\alpha_1 x} + \cdots + c_n e^{\alpha_n x}$$

is a solution, and

$$\begin{aligned} f(0) &= c_1 + \cdots + c_n, \\ f'(0) &= \alpha_1 c_1 + \cdots + \alpha_n c_n, \\ &\vdots \\ f^{(n-1)}(0) &= \alpha_1^{n-1} c_1 + \cdots + \alpha_n^{n-1} c_n. \end{aligned}$$

As a matter of fact, every solution is of this form, because we can obtain any set of numbers on the left side by choosing the c 's properly, but we will not try to prove this last assertion. (It is a purely algebraic fact, which you can easily check for $n = 2$ or 3.) These remarks are also true if some of the roots are multiple roots, and even in the more general situation considered in Chapter 27.

- **20.** (a) Suppose that f is a continuous function on $[a, b]$ with $f(a) = f(b)$ and that for all x in (a, b) the Schwarz second derivative of f at x is 0 (Problem 16). Show that f is constant on $[a, b]$. Hint: Suppose that $f(x) > f(a)$ for some x in (a, b) . Consider the function

$$g(x) = f(x) - \varepsilon(x - a)(b - x)$$

with $g(a) = g(b) = f(a)$. For sufficiently small $\varepsilon > 0$ we will have $g(x) > g(a)$, so g will have a maximum point y in (a, b) . Now use Problem 16(c) (the Schwarz second derivative of $(x - a)(b - x)$ is simply its ordinary second derivative).

- (b) If f is a continuous function on $[a, b]$ whose Schwarz second derivative is 0 at all points of (a, b) , then f is linear.

- *21. (a) Let $f(x) = x^4 \sin 1/x^2$ for $x \neq 0$, and $f(0) = 0$. Show that $f = 0$ up to order 2 at 0, even though $f''(0)$ does not exist.

This example is slightly more complex, but also slightly more impressive, than the example in the text, because both $f'(a)$ and $f''(a)$ exist for $a \neq 0$. Thus, for each number a there is another number $m(a)$ such that

$$(*) \quad f(x) = f(a) + f'(a)(x - a) + \frac{m(a)}{2}(x - a)^2 + R_a(x),$$

$$\text{where } \lim_{x \rightarrow a} \frac{R_a(x)}{(x - a)^2} = 0;$$

namely, $m(a) = f''(a)$ for $a \neq 0$, and $m(0) = 0$. Notice that the function m defined in this way is not continuous.

- (b) Suppose that f is a differentiable function such that $(*)$ holds for all a , with $m(a) = 0$. Use Problem 20 to show that $f''(a) = m(a) = 0$ for all a .
- (c) Now suppose that $(*)$ holds for all a , and that m is continuous. Prove that for all a the second derivative $f''(a)$ exists and equals $m(a)$.

*CHAPTER

21

e IS TRANSCENDENTAL

The irrationality of e was so easy to prove that in this optional chapter we will attempt a more difficult feat, and prove that the number e is not merely irrational, but actually much worse. Just how a number might be even worse than irrational is suggested by a slight rewording of definitions. A number x is irrational if it is not possible to write $x = a/b$ for any integers a and b , with $b \neq 0$. This is the same as saying that x does not satisfy any equation

$$bx - a = 0$$

for integers a and b , except for $a = 0, b = 0$. Viewed in this light, the irrationality of $\sqrt{2}$ does not seem to be such a terrible deficiency; rather, it appears that $\sqrt{2}$ just barely manages to be irrational—although $\sqrt{2}$ is not the solution of an equation

$$a_1x + a_0 = 0,$$

it is the solution of the equation

$$x^2 - 2 = 0,$$

of one higher degree. Problem 2-18 shows how to produce many irrational numbers x which satisfy higher-degree equations

$$a_nx^n + a_{n-1}x^{n-1} + \cdots + a_0 = 0,$$

where the a_i are integers and $a_0 \neq 0$ (this condition rules out the possibility that all $a_i = 0$). A number which satisfies an “algebraic” equation of this sort is called an **algebraic number**, and practically every number we have ever encountered is defined in terms of solutions of algebraic equations (π and e are the great exceptions in our limited mathematical experience). All roots, such as

$$\sqrt{2}, \quad \sqrt[10]{3}, \quad \sqrt[4]{7},$$

are clearly algebraic numbers, and even complicated combinations, like

$$\sqrt[3]{3 + \sqrt{5}} + \sqrt[4]{1 + \sqrt{2}} + \sqrt[5]{6}$$

are algebraic (although we will not try to prove this). Numbers which cannot be obtained by the process of solving algebraic equations are called **transcendental**; the main result of this chapter states that e is a number of this anomalous sort.

The proof that e is transcendental is well within our grasp, and was theoretically possible even before Chapter 20. Nevertheless, with the inclusion of this proof, we can justifiably classify ourselves as something more than novices in the study of higher mathematics; while many irrationality proofs depend only on elementary properties of numbers, the proof that a number is transcendental usually involves

some really high-powered mathematics. Even the dates connected with the transcendence of e are impressively recent—the first proof that e is transcendental, due to Hermite, dates from 1873. The proof that we will give is a simplification, due to Hilbert.

Before tackling the proof itself, it is a good idea to map out the strategy, which depends on an idea used even in the proof that e is irrational. Two features of the expression

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{n!} + R_n$$

were important for the proof that e is irrational: On the one hand, the number

$$1 + \frac{1}{1!} + \cdots + \frac{1}{n!}$$

can be written as a fraction p/q with $q \leq n!$ (so that $n!(p/q)$ is an integer); on the other hand, $0 < R_n < 3/(n+1)!$ (so $n!R_n$ is not an integer). These two facts show that e can be approximated particularly well by rational numbers. Of course, every number x can be approximated arbitrarily closely by rational numbers—if $\varepsilon > 0$ there is a rational number r with $|x - r| < \varepsilon$; the catch, however, is that it may be necessary to allow a very large denominator for r , as large as $1/\varepsilon$ perhaps. For e we are assured that this is not the case: there is a fraction p/q within $3/(n+1)!$ of e , whose denominator q is at most $n!$. If you look carefully at the proof that e is irrational, you will see that only this fact about e is ever used. The number e is by no means unique in this respect: generally speaking, the *better* a number can be approximated by rational numbers, the *worse* it is (some evidence for this assertion is presented in Problem 3). The proof that e is transcendental depends on a natural extension of this idea: not only e , but any finite number of powers e, e^2, \dots, e^n , can be simultaneously approximated especially well by rational numbers. In our proof we will begin by assuming that e is algebraic, so that

$$(*) \quad a_n e^n + \cdots + a_1 e + a_0 = 0, \quad a_0 \neq 0$$

for some integers a_0, \dots, a_n . In order to reach a contradiction we will then find certain integers M, M_1, \dots, M_n and certain “small” numbers $\epsilon_1, \dots, \epsilon_n$ such that

$$\begin{aligned} e^1 &= \frac{M_1 + \epsilon_1}{M}, \\ e^2 &= \frac{M_2 + \epsilon_2}{M}, \\ &\vdots \\ e^n &= \frac{M_n + \epsilon_n}{M}. \end{aligned}$$

Just how small the ϵ ’s must be will appear when these expressions are substituted into the assumed equation (*). After multiplying through by M we obtain

$$[a_0 M + a_1 M_1 + \cdots + a_n M_n] + [\epsilon_1 a_1 + \cdots + \epsilon_n a_n] = 0.$$

The first term in brackets is an integer, and we will choose the M 's so that it will necessarily be a *nonzero* integer. We will also manage to find ϵ 's so small that

$$|\epsilon_1 a_1 + \cdots + \epsilon_n a_n| < \frac{1}{2};$$

this will lead to the desired contradiction—the sum of a nonzero integer and a number of absolute value less than $\frac{1}{2}$ cannot be zero!

As a basic strategy this is all very reasonable and quite straightforward. The remarkable part of the proof will be the way that the M 's and ϵ 's are defined. In order to read the proof you will need to know about the gamma function! (This function was introduced in Problem 19-39.)

THEOREM 1 e is transcendental.

PROOF Suppose there were integers a_0, \dots, a_n , with $a_0 \neq 0$, such that

$$(*) \quad a_n e^n + a_{n-1} e^{n-1} + \cdots + a_0 = 0.$$

Define numbers M, M_1, \dots, M_n and $\epsilon_1, \dots, \epsilon_n$ as follows:

$$\begin{aligned} M &= \int_0^\infty \frac{x^{p-1}[(x-1) \cdot \dots \cdot (x-n)]^p e^{-x}}{(p-1)!} dx, \\ M_k &= e^k \int_k^\infty \frac{x^{p-1}[(x-1) \cdot \dots \cdot (x-n)]^p e^{-x}}{(p-1)!} dx, \\ \epsilon_k &= e^k \int_0^k \frac{x^{p-1}[(x-1) \cdot \dots \cdot (x-n)]^p e^{-x}}{(p-1)!} dx. \end{aligned}$$

The unspecified number p represents a prime number* which we will choose later. Despite the forbidding aspect of these three expressions, with a little work they will appear much more reasonable. We concentrate on M first. If the expression in brackets,

$$[(x-1) \cdot \dots \cdot (x-n)],$$

is actually multiplied out, we obtain a polynomial

$$x^n + \cdots \pm n!$$

*The term “prime number” was defined in Problem 2-17. An important fact about prime numbers will be used in the proof, although it is not proved in this book: If p is a prime number which does not divide the integer a , and which does not divide the integer b , then p also does not divide ab . The Suggested Reading mentions references for this theorem (which is crucial in proving that the factorization of an integer into primes is unique). We will also use the result of Problem 2-17(d), that there are infinitely many primes—the reader is asked to determine at precisely which points this information is required.

with integer coefficients. When raised to the p th power this becomes an even more complicated polynomial

$$x^{np} + \cdots \pm (n!)^p.$$

Thus M can be written in the form

$$M = \sum_{\alpha=0}^{np} \frac{1}{(p-1)!} C_\alpha \int_0^\infty x^{p-1+\alpha} e^{-x} dx,$$

where the C_α are certain integers, and $C_0 = \pm(n!)^p$. But

$$\int_0^\infty x^k e^{-x} dx = k!.$$

Thus

$$M = \sum_{\alpha=0}^{np} C_\alpha \frac{(p-1+\alpha)!}{(p-1)!}.$$

Now, for $\alpha = 0$ we obtain the term

$$\pm(n!)^p \frac{(p-1)!}{(p-1)!} = \pm(n!)^p.$$

We will now consider only primes $p > n$; then this term is an integer which is *not* divisible by p . On the other hand, if $\alpha > 0$, then

$$C_\alpha \frac{(p-1+\alpha)!}{(p-1)!} = C_\alpha (p+\alpha-1)(p+\alpha-2) \cdots p,$$

which is divisible by p . Therefore M itself is an integer which is *not* divisible by p .

Now consider M_k . We have

$$\begin{aligned} M_k &= e^k \int_k^\infty \frac{x^{p-1}[(x-1) \cdots (x-n)]^p e^{-x}}{(p-1)!} dx \\ &= \int_k^\infty \frac{x^{p-1}[(x-1) \cdots (x-n)]^p e^{-(x-k)}}{(p-1)!} dx. \end{aligned}$$

This can be transformed into an expression looking very much like M by the substitution

$$\begin{aligned} u &= x - k \\ du &= dx. \end{aligned}$$

The limits of integration are changed to 0 and ∞ , and

$$M_k = \int_0^\infty \frac{(u+k)^{p-1}[(u+k-1) \cdots u \cdots (u+k-n)]^p e^{-u}}{(p-1)!} du.$$

There is one very significant difference between this expression and that for M . The term in brackets contains the factor u in the k th place. Thus the p th power contains the factor u^p . This means that the entire expression

$$(u+k)^{p-1}[(u+k-1) \cdots (u+k-n)]^p$$

is a polynomial with integer coefficients, *every term of which* has degree at least p . Thus

$$M_k = \sum_{\alpha=1}^{np} \frac{1}{(p-1)!} D_\alpha \int_0^\infty u^{p-1+\alpha} e^{-u} du = \sum_{\alpha=1}^{np} D_\alpha \frac{(p-1+\alpha)!}{(p-1)!},$$

where the D_α are certain integers. Notice that the summation begins with $\alpha = 1$; in this case *every* term in the sum is divisible by p . Thus each M_k is an integer which is divisible by p .

Now it is clear that

$$e^k = \frac{M_k + \epsilon_k}{M}, \quad k = 1, \dots, n.$$

Substituting into (*) and multiplying by M we obtain

$$[a_0 M + a_1 M + \dots + a_n M_n] + [a_1 \epsilon_1 + \dots + a_n \epsilon_n] = 0.$$

In addition to requiring that $p > n$ let us also stipulate that $p > |a_0|$. This means that both M and a_0 are not divisible by p , so $a_0 M$ is also not divisible by p . Since each M_k is divisible by p , it follows that

$$a_0 M + a_1 M_1 + \dots + a_n M_n$$

is *not* divisible by p . In particular it is a *nonzero* integer.

In order to obtain a contradiction to the assumed equation (*), and thereby prove that e is transcendental, it is only necessary to show that

$$|a_1 \epsilon_1 + \dots + a_n \epsilon_n|$$

can be made as small as desired, by choosing p large enough; it is clearly sufficient to show that each $|\epsilon_k|$ can be made as small as desired. This requires nothing more than some simple estimates; for the remainder of the argument remember that n is a certain fixed number (the degree of the assumed polynomial equation (*)). To begin with, if $1 \leq k \leq n$, then

$$\begin{aligned} |\epsilon_k| &\leq e^k \int_0^k \frac{|x^{p-1}[(x-1) \cdot \dots \cdot (x-n)]^p| e^{-x}}{(p-1)!} dx \\ &\leq e^n \int_0^n \frac{n^{p-1}|[(x-1) \cdot \dots \cdot (x-n)]^p| e^{-x}}{(p-1)!} dx. \end{aligned}$$

Now let A be the maximum of $|(x-1) \cdot \dots \cdot (x-n)|$ for x in $[0, n]$. Then

$$\begin{aligned} |\epsilon_k| &\leq \frac{e^n n^{p-1} A^p}{(p-1)!} \int_0^n e^{-x} dx \\ &\leq \frac{e^n n^{p-1} A^p}{(p-1)!} \int_0^\infty e^{-x} dx \\ &= \frac{e^n n^{p-1} A^p}{(p-1)!} \\ &\leq \frac{e^n n^p A^p}{(p-1)!} = \frac{e^n (nA)^p}{(p-1)!}. \end{aligned}$$

But n and A are fixed; thus $(nA)^p/(p - 1)!$ can be made as small as desired by making p sufficiently large. ■

This proof, like the proof that π is irrational, deserves some philosophic afterthoughts. At first sight, the argument seems quite “advanced”—after all, we use integrals, and integrals from 0 to ∞ at that. Actually, as many mathematicians have observed, integrals can be eliminated from the argument completely; the only integrals essential to the proof are of the form

$$\int_0^\infty x^k e^{-x} dx$$

for integral k , and these integrals can be replaced by $k!$ whenever they occur. Thus M , for example, could have been defined initially as

$$M = \sum_{\alpha=0}^{np} C_\alpha \frac{(p-1+\alpha)!}{(p-1)!},$$

where C_α are the coefficients of the polynomial

$$[(x-1) \cdot \dots \cdot (x-n)]^p.$$

If this idea is developed consistently, one obtains a “completely elementary” proof that e is transcendental, depending only on the fact that

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots$$

Unfortunately, this “elementary” proof is harder to understand than the original one—the whole structure of the proof must be hidden just to eliminate a few integral signs! This situation is by no means peculiar to this specific theorem—“elementary” arguments are frequently more difficult than “advanced” ones. Our proof that π is irrational is a case in point. You probably remember nothing about this proof except that it involves quite a few complicated functions. There is actually a more advanced, but much more conceptual proof, which shows that π is *transcendental*, a fact which is of great historical, as well as intrinsic, interest. One of the classical problems of Greek mathematics was to construct, with compass and straightedge alone, a square whose area is that of a circle of radius 1. This requires the construction of a line segment whose length is $\sqrt{\pi}$, which can be accomplished if a line segment of length π is constructible. The Greeks were totally unable to decide whether such a line segment could be constructed, and even the full resources of modern mathematics were unable to settle this question until 1882. In that year Lindemann proved that π is transcendental; since the length of any segment that can be constructed with straightedge and compass can be written in terms of $+$, \cdot , $-$, \div , and $\sqrt{}$, and is therefore algebraic, this proves that a line segment of length π cannot be constructed.

The proof that π is transcendental requires a sizable amount of mathematics which is too advanced to be reached in this book. Nevertheless, the proof is not much more difficult than the proof that e is transcendental. In fact, the proof for π is practically the same as the proof for e . This last statement should certainly

surprise you. The proof that e is transcendental seems to depend so thoroughly on particular properties of e that it is almost inconceivable how any modifications could ever be used for π ; after all, what does e have to do with π ? Just wait and see!

PROBLEMS

1. (a) Prove that if $\alpha > 0$ is algebraic, then $\sqrt{\alpha}$ is algebraic.
 (b) Prove that if α is algebraic and r is rational, then $\alpha + r$ and αr are algebraic.

Part (b) can actually be strengthened considerably: the sum, product, and quotient of algebraic numbers is algebraic. This fact is too difficult for us to prove here, but some special cases can be examined:

2. Prove that $\sqrt{2} + \sqrt{3}$ and $\sqrt{2}(1 + \sqrt{3})$ are algebraic, by actually finding algebraic equations which they satisfy. (You will need equations of degree 4.)
- *3. (a) Let α be an algebraic number which is not rational. Suppose that α satisfies the polynomial equation

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 = 0,$$

and that no polynomial function of lower degree has this property. Show that $f(p/q) \neq 0$ for any rational number p/q . Hint: Use Problem 3-7(b).

- (b) Now show that $|f(p/q)| \geq 1/q^n$ for all rational numbers p/q with $q > 0$. Hint: Write $f(p/q)$ as a fraction over the common denominator q^n .
- (c) Let $M = \sup\{|f'(x)| : |x - \alpha| < 1\}$. Use the Mean Value Theorem to prove that if p/q is a rational number with $|\alpha - p/q| < 1$, then $|\alpha - p/q| > 1/Mq^n$. (It follows that for $c = \max(1, 1/M)$ we have $|\alpha - p/q| > c/q^n$ for all rational p/q .)

- *4. Let

$$\alpha = 0.11000100000000000000001000\dots,$$

where the 1's occur in the $n!$ place, for each n . Use Problem 3 to prove that α is transcendental. (For each n , show that α is not the root of an equation of degree n .)

Although Problem 4 mentions only one specific transcendental number, it should be clear that one can easily construct infinitely many other numbers α which do not satisfy $|\alpha - p/q| > c/q^n$ for any c and n . Such numbers were first considered by Liouville (1809–1882), and the inequality in Problem 3 is often called Liouville's inequality. None of the transcendental numbers constructed in this way happens to be particularly interesting, but for a long time Liouville's transcendental numbers were the only ones known. This situation was changed quite radically by the work of Cantor (1845–1918), who showed, without exhibiting a single transcendental number, that *most* numbers are transcendental. The next two problems provide an

introduction to the ideas that allow us to make sense of such statements. The basic definition with which we must work is the following: A set A is called **countable** if its elements can be arranged in a sequence

$$a_1, a_2, a_3, a_4, \dots$$

The obvious example (in fact, more or less the Platonic ideal of) a countable set is \mathbf{N} , the set of natural numbers; clearly the set of even natural numbers is also countable:

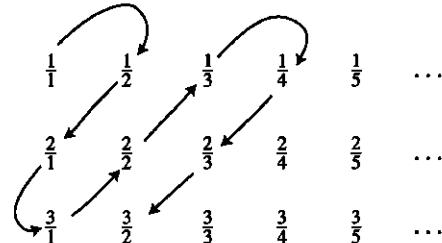
$$2, 4, 6, 8, \dots$$

It is a little more surprising to learn that \mathbf{Z} , the set of all integers (positive, negative and 0) is also countable, but seeing is believing:

$$0, 1, -1, 2, -2, 3, -3, \dots$$

The next two problems, which outline the basic features of countable sets, are really a series of examples to show that (1) a lot more sets are countable than one might think and (2) nevertheless, some sets are not countable.

- *5. (a) Show that if A and B are countable, then so is $A \cup B = \{x : x \text{ is in } A \text{ or } x \text{ is in } B\}$. Hint: Use the same trick that worked for \mathbf{Z} .
- (b) Show that the set of positive rational numbers is countable. (This is really quite startling, but the figure below indicates the path to enlightenment.)



- (c) Show that the set of all pairs (m, n) of integers is countable. (This is practically the same as part (b).)
- (d) If A_1, A_2, A_3, \dots are each countable, prove that

$$A_1 \cup A_2 \cup A_3 \cup \dots$$

is also countable. (Again use the same trick as in part (b).)

- (e) Prove that the set of all triples (l, m, n) of integers is countable. (A triple (l, m, n) can be described by a pair (l, m) and a number n .)
- (f) Prove that the set of all n -tuples (a_1, a_2, \dots, a_n) is countable. (If you have done part (e), you can do this, using induction.)
- (g) Prove that the set of all roots of polynomial functions of degree n with integer coefficients is countable. (Part (f) shows that the set of all these

polynomial functions can be arranged in a sequence, and each has at most n roots.)

- (h) Now use parts (d) and (g) to prove that the set of all algebraic numbers is countable.

- *6. Since so many sets turn out to be countable, it is important to note that the set of all real numbers between 0 and 1 is *not* countable. In other words, there is no way of listing all these real numbers in a sequence

$$\alpha_1 = 0.a_{11}a_{12}a_{13}a_{14}\dots$$

$$\alpha_2 = 0.a_{21}a_{22}a_{23}a_{24}\dots$$

$$\alpha_3 = 0.a_{31}a_{32}a_{33}a_{34}\dots$$

...

(decimal notation is being used on the right). To prove that this is so, suppose such a list were possible and consider the decimal

$$0.\bar{a}_{11}\bar{a}_{22}\bar{a}_{33}\bar{a}_{44}\dots,$$

where $\bar{a}_{nn} = 5$ if $a_{nn} \neq 5$ and $\bar{a}_{nn} = 6$ if $a_{nn} = 5$. Show that this number cannot possibly be in the list, thus obtaining a contradiction.

Problems 5 and 6 can be summed up as follows. The set of algebraic numbers is countable. If the set of transcendental numbers were also countable, then the set of all real numbers would be countable, by Problem 5(a), and consequently the set of real numbers between 0 and 1 would be countable. But this is false. Thus, the set of algebraic numbers is countable and the set of transcendental numbers is not ("there are more transcendental numbers than algebraic numbers"). The remaining two problems illustrate further how important it can be to distinguish between sets which are countable and sets which are not.

- *7. Let f be a nondecreasing function on $[0, 1]$. Recall (Problem 8-8) that $\lim_{x \rightarrow a^+} f(x)$ and $\lim_{x \rightarrow a^-} f(x)$ both exist.

- (a) For any $\varepsilon > 0$ prove that there are only finitely many numbers a in $[0, 1]$ with $\lim_{x \rightarrow a^+} f(x) - \lim_{x \rightarrow a^-} f(x) > \varepsilon$. Hint: There are, in fact, at most $[f(1) - f(0)]/\varepsilon$ of them.
- (b) Prove that the set of points at which f is discontinuous is countable. Hint: If $\lim_{x \rightarrow a^+} f(x) - \lim_{x \rightarrow a^-} f(x) > 0$, then it is $> 1/n$ for some natural number n .

This problem shows that a nondecreasing function is automatically continuous at most points. For differentiability the situation is more difficult to analyze and also more interesting. A nondecreasing function can fail to be differentiable at a set of points which is not countable, but it is still true that nondecreasing functions are differentiable at most points (in a different sense of the word "most"). Reference [32] of the Suggested Reading gives a beautiful proof, using the Rising Sun Lemma of

Problem 8-20. For those who have done Problem 10 of the Appendix to Chapter 11, it is possible to provide at least one application to differentiability of the ideas already developed in this problem set: If f is convex, then f is differentiable except at those points where its right-hand derivative f'_+ is discontinuous; but the function f'_+ is increasing, so a convex function is automatically differentiable except at a countable set of points.

- *8. (a) Problem 11-66 showed that if every point is a local maximum point for a *continuous* function f , then f is a constant function. Suppose now that the hypothesis of continuity is dropped. Prove that f takes on only a countable set of values. Hint: For each x choose *rational* numbers a_x and b_x such that $a_x < x < b_x$ and x is a maximum point for f on (a_x, b_x) . Then every value $f(x)$ is the maximum value of f on some interval (a_x, b_x) . How many such intervals are there?
(b) Deduce Problem 11-66(a) as a corollary.
(c) Prove the result of Problem 11-66(b) similarly.

CHAPTER 22 INFINITE SEQUENCES

The idea of an infinite sequence is so natural a concept that it is tempting to dispense with a definition altogether. One frequently writes simply “an infinite sequence

$$a_1, a_2, a_3, a_4, a_5, \dots,$$

the three dots indicating that the numbers a_i continue to the right “forever.” A rigorous definition of an infinite sequence is not hard to formulate, however; the important point about an infinite sequence is that for each natural number, n , there is a real number a_n . This sort of correspondence is precisely what functions are meant to formalize.

DEFINITION

An **infinite sequence** of real numbers is a function whose domain is \mathbf{N} .

From the point of view of this definition, a sequence should be designated by a single letter like a , and particular values by

$$a(1), a(2), a(3), \dots,$$

but the subscript notation

$$a_1, a_2, a_3, \dots$$

is almost always used instead, and the sequence itself is usually denoted by a symbol like $\{a_n\}$. Thus $\{n\}$, $\{(-1)^n\}$, and $\{1/n\}$ denote the sequences α , β , and γ defined by

$$\begin{aligned}\alpha_n &= n, \\ \beta_n &= (-1)^n, \\ \gamma_n &= \frac{1}{n}.\end{aligned}$$

A sequence, like any function, can be graphed (Figure 1) but the graph is usually rather unrevealing, since most of the function cannot be fit on the page.

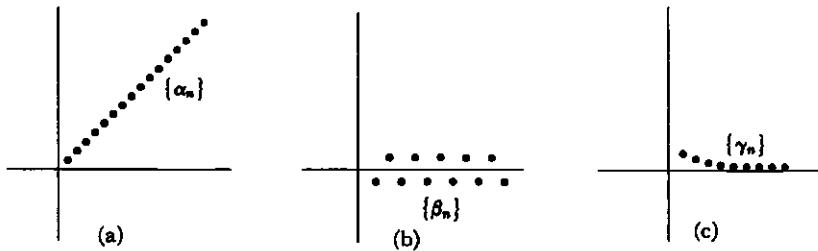


FIGURE 1

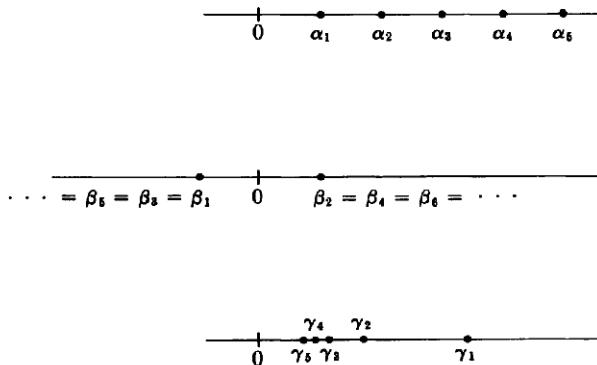


FIGURE 2

A more convenient representation of a sequence is obtained by simply labeling the points a_1, a_2, a_3, \dots on a line (Figure 2). This sort of picture shows where the sequence “is going.” The sequence $\{a_n\}$ “goes out to infinity,” the sequence $\{\beta_n\}$ “jumps back and forth between -1 and 1 ,” and the sequence $\{\gamma_n\}$ “converges to 0 .” Of the three phrases in quotation marks, the last is the crucial concept associated with sequences, and will be defined precisely (the definition is illustrated in Figure 3).



FIGURE 3

DEFINITION

A sequence $\{a_n\}$ **converges to l** (in symbols $\lim_{n \rightarrow \infty} a_n = l$) if for every $\varepsilon > 0$ there is a natural number N such that, for all natural numbers n ,

$$\text{if } n > N, \text{ then } |a_n - l| < \varepsilon.$$

In addition to the terminology introduced in this definition, we sometimes say that the sequence $\{a_n\}$ **approaches l** or has the **limit l** . A sequence $\{a_n\}$ is said to **converge** if it converges to l for some l , and to **diverge** if it does not converge.

To show that the sequence $\{\gamma_n\}$ converges to 0 , it suffices to observe the following. If $\varepsilon > 0$, there is a natural number N such that $1/N < \varepsilon$. Then, if $n > N$ we have

$$\gamma_n = \frac{1}{n} < \frac{1}{N} < \varepsilon, \quad \text{so } |\gamma_n - 0| < \varepsilon.$$

The limit

$$\lim_{n \rightarrow \infty} \sqrt{n+1} - \sqrt{n} = 0$$

will probably seem reasonable after a little reflection (it just says that $\sqrt{n+1}$ is practically the same as \sqrt{n} for large n), but a mathematical proof might not be so

obvious. To estimate $\sqrt{n+1} - \sqrt{n}$ we can use an algebraic trick:

$$\begin{aligned}\sqrt{n+1} - \sqrt{n} &= \frac{(\sqrt{n+1} - \sqrt{n})(\sqrt{n+1} + \sqrt{n})}{\sqrt{n+1} + \sqrt{n}} \\ &= \frac{n+1-n}{\sqrt{n+1} + \sqrt{n}} \\ &= \frac{1}{\sqrt{n+1} + \sqrt{n}}.\end{aligned}$$

It is also possible to estimate $\sqrt{n+1} - \sqrt{n}$ by applying the Mean Value Theorem to the function $f(x) = \sqrt{x}$ on the interval $[n, n+1]$. We obtain

$$\begin{aligned}\frac{\sqrt{n+1} - \sqrt{n}}{1} &= f'(x) \\ &= \frac{1}{2\sqrt{x}}, \quad \text{for some } x \text{ in } (n, n+1) \\ &< \frac{1}{2\sqrt{n}}.\end{aligned}$$

Either of these estimates may be used to prove the above limit; the detailed proof is left to you, as a simple but valuable exercise.

The limit

$$\lim_{n \rightarrow \infty} \frac{3n^3 + 7n^2 + 1}{4n^3 - 8n + 63} = \frac{3}{4}$$

should also seem reasonable, because the terms involving n^3 are the most important when n is large. If you remember the proof of Theorem 7-9 you will be able to guess the trick that translates this idea into a proof—dividing top and bottom by n^3 yields

$$\frac{3n^3 + 7n^2 + 1}{4n^3 - 8n + 63} = \frac{3 + \frac{7}{n} + \frac{1}{n^3}}{4 - \frac{8}{n^2} + \frac{63}{n^3}}.$$

Using this expression, the proof of the above limit is not difficult, especially if one uses the following facts:

If $\lim_{n \rightarrow \infty} a_n$ and $\lim_{n \rightarrow \infty} b_n$ both exist, then

$$\begin{aligned}\lim_{n \rightarrow \infty} (a_n + b_n) &= \lim_{n \rightarrow \infty} a_n + \lim_{n \rightarrow \infty} b_n, \\ \lim_{n \rightarrow \infty} (a_n \cdot b_n) &= \lim_{n \rightarrow \infty} a_n \cdot \lim_{n \rightarrow \infty} b_n;\end{aligned}$$

moreover, if $\lim_{n \rightarrow \infty} b_n \neq 0$, then $b_n \neq 0$ for all n greater than some N , and

$$\lim_{n \rightarrow \infty} a_n/b_n = \lim_{n \rightarrow \infty} a_n / \lim_{n \rightarrow \infty} b_n.$$

(If we wanted to be utterly precise, the third statement would have to be even more complicated. As it stands, we are considering the limit of the sequence $\{c_n\} = \{a_n/b_n\}$, where the numbers c_n might not even be defined for certain $n < N$. This doesn't really matter—we could define c_n any way we liked for such n —because the limit of a sequence is not changed if we change the sequence at a finite number of points.)

Although these facts are very useful, we will not bother stating them as a theorem—you should have no difficulty proving these results for yourself, because the definition of $\lim_{n \rightarrow \infty} a_n = l$ is so similar to previous definitions of limits, especially $\lim_{x \rightarrow \infty} f(x) = l$.

The similarity between the definitions of $\lim_{n \rightarrow \infty} a_n = l$ and $\lim_{x \rightarrow \infty} f(x) = l$ is actually closer than mere analogy; it is possible to define the first in terms of the second. If f is the function whose graph (Figure 4) consists of line segments joining

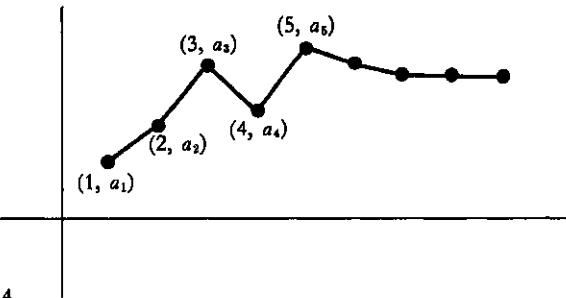


FIGURE 4

the points in the graph of the sequence $\{a_n\}$, so that

$$f(x) = (a_{n+1} - a_n)(x - n) + a_n \quad n \leq x \leq n + 1,$$

then

$$\lim_{n \rightarrow \infty} a_n = l \quad \text{if and only if} \quad \lim_{x \rightarrow \infty} f(x) = l.$$

Conversely, if f satisfies $\lim_{x \rightarrow \infty} f(x) = l$, and we set $a_n = f(n)$, then $\lim_{n \rightarrow \infty} a_n = l$.

This second observation is frequently very useful. For example, suppose that $0 < a < 1$. Then

$$\lim_{n \rightarrow \infty} a^n = 0.$$

To prove this we note that

$$\lim_{x \rightarrow \infty} a^x = \lim_{x \rightarrow \infty} e^{x \log a} = 0,$$

since $\log a < 0$, so that $x \log a$ is a negative and large in absolute value for large x . Notice that we actually have

$$\lim_{n \rightarrow \infty} a^n = 0 \quad \text{if } |a| < 1;$$

for if $a < 0$ we can write

$$\lim_{n \rightarrow \infty} a^n = \lim_{n \rightarrow \infty} (-1)^n |a|^n = 0.$$

The behavior of the logarithm function also shows that if $a > 1$, then a^n becomes arbitrarily large as n becomes large. This assertion is often written

$$\lim_{n \rightarrow \infty} a^n = \infty, \quad a > 1,$$

and it is sometimes even said that $\{a^n\}$ approaches ∞ . We also write equations like

$$\lim_{n \rightarrow \infty} -a^n = -\infty,$$

and say that $\{-a^n\}$ approaches $-\infty$. Notice, however, that if $a < -1$, then $\lim_{n \rightarrow \infty} a^n$ does not exist, even in this extended sense.

Despite this connection with a familiar concept, it is more important to visualize convergence in terms of the picture of a sequence as points on a line (Figure 3). There is another connection between limits of functions and limits of sequences which is related to *this* picture. This connection is somewhat less obvious, but considerably more interesting, than the one previously mentioned—instead of defining limits of sequences in terms of limits of functions, it is possible to reverse the procedure.

THEOREM 1 Let f be a function defined in an open interval containing c , except perhaps at c itself, with

$$\lim_{x \rightarrow c} f(x) = l.$$

Suppose that $\{a_n\}$ is a sequence such that

- (1) each a_n is in the domain of f ,
- (2) each $a_n \neq c$,
- (3) $\lim_{n \rightarrow \infty} a_n = c$.

Then the sequence $\{f(a_n)\}$ satisfies

$$\lim_{n \rightarrow \infty} f(a_n) = l.$$

Conversely, if this is true for every sequence $\{a_n\}$ satisfying the above conditions, then $\lim_{x \rightarrow c} f(x) = l$.

PROOF Suppose first that $\lim_{x \rightarrow c} f(x) = l$. Then for every $\varepsilon > 0$ there is a $\delta > 0$ such that, for all x ,

$$\text{if } 0 < |x - c| < \delta, \text{ then } |f(x) - l| < \varepsilon.$$

If the sequence $\{a_n\}$ satisfies $\lim_{n \rightarrow \infty} a_n = c$, then (Figure 3) there is a natural number N such that,

$$\text{if } n > N, \text{ then } |a_n - c| < \delta.$$

By our choice of δ , this means that

$$|f(a_n) - l| < \varepsilon,$$

showing that

$$\lim_{n \rightarrow \infty} f(a_n) = l.$$

Suppose, conversely, that $\lim_{n \rightarrow \infty} f(a_n) = l$ for every sequence $\{a_n\}$ with $\lim_{n \rightarrow \infty} a_n = c$. If $\lim_{x \rightarrow c} f(x) = l$ were not true, there would be some $\varepsilon > 0$ such that for every $\delta > 0$ there is an x with

$$0 < |x - c| < \delta \quad \text{but} \quad |f(x) - l| > \varepsilon.$$

In particular, for each n there would be a number x_n such that

$$0 < |x_n - c| < \frac{1}{n} \quad \text{but} \quad |f(x_n) - l| > \varepsilon.$$

Now the sequence $\{x_n\}$ clearly converges to c but, since $|f(x_n) - l| > \varepsilon$ for all n , the sequence $\{f(x_n)\}$ does not converge to l . This contradicts the hypothesis, so $\lim_{x \rightarrow c} f(x) = l$ must be true. ■

Theorem 1 provides many examples of convergent sequences. For example, the sequences $\{a_n\}$ and $\{b_n\}$ defined by

$$\begin{aligned} a_n &= \sin\left(13 + \frac{1}{n^2}\right) \\ b_n &= \cos\left(\sin\left(1 + (-1)^n \cdot \frac{1}{n}\right)\right), \end{aligned}$$

clearly converge to $\sin(13)$ and $\cos(\sin(1))$, respectively. It is important, however, to have some criteria guaranteeing convergence of sequences which are not obviously of this sort. There is one important criterion which is very easy to prove, but which is the basis for all other results. This criterion is stated in terms of concepts defined for functions, which therefore apply also to sequences: a sequence $\{a_n\}$ is **increasing** if $a_{n+1} > a_n$ for all n , **nondecreasing** if $a_{n+1} \geq a_n$ for all n , and **bounded above** if there is a number M such that $a_n \leq M$ for all n ; there are similar definitions for sequences which are decreasing, nonincreasing, and bounded below.

THEOREM 2

If $\{a_n\}$ is nondecreasing and bounded above, then $\{a_n\}$ converges (a similar statement is true if $\{a_n\}$ is nonincreasing and bounded below).

PROOF

The set A consisting of all numbers a_n is, by assumption, bounded above, so A has a least upper bound α . We claim that $\lim_{n \rightarrow \infty} a_n = \alpha$ (Figure 5). In fact, if $\varepsilon > 0$, there is some a_N satisfying $\alpha - a_N < \varepsilon$, since α is the least upper bound of A . Then if $n > N$ we have

$$a_n \geq a_N, \quad \text{so} \quad \alpha - a_n \leq \alpha - a_N < \varepsilon.$$

This proves that $\lim_{n \rightarrow \infty} a_n = \alpha$. ■

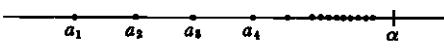


FIGURE 5

The hypothesis that $\{a_n\}$ is bounded above is clearly essential in Theorem 2: if $\{a_n\}$ is not bounded above, then (whether or not $\{a_n\}$ is nondecreasing) $\{a_n\}$ clearly diverges. Upon first consideration, it might appear that there should be little trouble deciding whether or not a given nondecreasing sequence $\{a_n\}$ is bounded above, and consequently whether or not $\{a_n\}$ converges. In the next chapter such sequences will arise very naturally and, as we shall see, deciding whether or not they converge is hardly a trivial matter. For the present, you might try to decide whether or not the following (obviously increasing) sequence is bounded above:

$$1, 1 + \frac{1}{2}, 1 + \frac{1}{2} + \frac{1}{3}, 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}, \dots$$

Although Theorem 2 treats only a very special class of sequences, it is more useful than might appear at first, because it is always possible to extract from an arbitrary sequence $\{a_n\}$ another sequence which is either nonincreasing or else nondecreasing. To be precise, let us define a **subsequence** of the sequence $\{a_n\}$ to be a sequence of the form

$$a_{n_1}, a_{n_2}, a_{n_3}, \dots,$$

where the n_j are natural numbers with

$$n_1 < n_2 < n_3 \dots.$$

Then every sequence contains a subsequence which is either nondecreasing or nonincreasing. It is possible to become quite befuddled trying to prove this assertion, although the proof is very short if you think of the right idea; it is worth recording as a lemma.

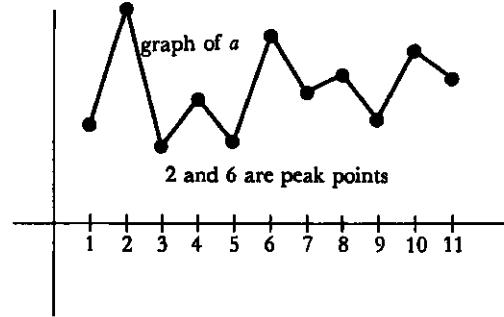


FIGURE 6

LEMMA Any sequence $\{a_n\}$ contains a subsequence which is either nondecreasing or nonincreasing.

PROOF Call a natural number n a “peak point” of the sequence $\{a_n\}$ if $a_m < a_n$ for all $m > n$ (Figure 6).

Case 1. The sequence has infinitely many peak points. In this case, if $n_1 < n_2 < n_3 < \dots$ are the peak points, then $a_{n_1} > a_{n_2} > a_{n_3} > \dots$, so $\{a_{n_k}\}$ is the desired (nonincreasing) subsequence.

Case 2. The sequence has only finitely many peak points. In this case, let n_1 be greater than all peak points. Since n_1 is not a peak point, there is some $n_2 > n_1$ such that $a_{n_2} \geq a_{n_1}$. Since n_2 is not a peak point (it is greater than n_1 , and hence greater than all peak points) there is some $n_3 > n_2$ such that $a_{n_3} \geq a_{n_2}$. Continuing in this way we obtain the desired (nondecreasing) subsequence. ■

If we assume that our original sequence $\{a_n\}$ is bounded, we can pick up an extra corollary along the way.

COROLLARY (THE BOLZANO-WEIERSTRASS THEOREM)

Every bounded sequence has a convergent subsequence.

Without some additional assumptions this is as far as we can go: it is easy to construct sequences having many, even infinitely many, subsequences converging to different numbers (see Problem 3). There is a reasonable assumption to add, which yields a necessary and sufficient condition for convergence of any sequence. Although this condition will not be crucial for our work, it does simplify many proofs. Moreover, this condition plays a fundamental role in more advanced investigations, and for this reason alone it is worth stating now.

If a sequence converges, so that the individual terms are eventually all close to the same number, then the difference of any two such individual terms should be very small. To be precise, if $\lim_{n \rightarrow \infty} a_n = l$ for some l , then for any $\varepsilon > 0$ there is an N such that $|a_n - l| < \varepsilon/2$ for $n > N$; now if both $n > N$ and $m > N$, then

$$|a_n - a_m| \leq |a_n - l| + |l - a_m| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This final inequality, $|a_n - a_m| < \varepsilon$, which eliminates mention of the limit l , can be used to formulate a condition (the Cauchy condition) which is clearly necessary for convergence of a sequence.

DEFINITION

A sequence $\{a_n\}$ is a **Cauchy sequence** if for every $\varepsilon > 0$ there is a natural number N such that, for all m and n ,

$$\text{if } m, n > N, \text{ then } |a_n - a_m| < \varepsilon.$$

(This condition is usually written $\lim_{m,n \rightarrow \infty} |a_m - a_n| = 0$.)

The beauty of the Cauchy condition is that it is also sufficient to ensure convergence of a sequence. After all our preliminary work, there is very little left to do in order to prove this.

THEOREM 3 A sequence $\{a_n\}$ converges if and only if it is a Cauchy sequence.

PROOF

We have already shown that $\{a_n\}$ is a Cauchy sequence if it converges. The proof of the converse assertion contains only one tricky feature: showing that every Cauchy sequence $\{a_n\}$ is bounded. If we take $\varepsilon = 1$ in the definition of a Cauchy sequence we find that there is some N such that

$$|a_m - a_n| < 1 \quad \text{for } m, n > N.$$

In particular, this means that

$$|a_m - a_{N+1}| < 1 \quad \text{for all } m > N.$$

Thus $\{a_m : m > N\}$ is bounded; since there are only finitely many other a_i 's the whole sequence is bounded.

The corollary to the Lemma thus implies that some subsequence of $\{a_n\}$ converges.

Only one point remains, whose proof will be left to you: if a subsequence of a Cauchy sequence converges, then the Cauchy sequence itself converges. ■

PROBLEMS

1. Verify each of the following limits.

$$(i) \lim_{n \rightarrow \infty} \frac{n}{n+1} = 1.$$

$$(ii) \lim_{n \rightarrow \infty} \frac{n+3}{n^3+4} = 0.$$

$$(iii) \lim_{n \rightarrow \infty} \sqrt[8]{n^2+1} - \sqrt[4]{n+1} = 0. \text{ Hint: You should at least be able to prove that } \lim_{n \rightarrow \infty} \sqrt[8]{n^2+1} - \sqrt[8]{n^2} = 0.$$

$$(iv) \lim_{n \rightarrow \infty} \frac{n!}{n^n} = 0. \text{ Hint: } n! = n(n-1) \cdot \dots \cdot k! \text{ for } k < n, \text{ in particular, for } k < n/2.$$

$$(v) \lim_{n \rightarrow \infty} \sqrt[p]{a} = 1, \quad a > 0.$$

$$(vi) \lim_{n \rightarrow \infty} \sqrt[3]{n} = 1.$$

$$(vii) \lim_{n \rightarrow \infty} \sqrt[5]{n^2+n} = 1.$$

$$(viii) \lim_{n \rightarrow \infty} \sqrt[3]{a^n+b^n} = \max(a, b), \quad a, b \geq 0.$$

$$(ix) \lim_{n \rightarrow \infty} \frac{\alpha(n)}{n} = 0, \text{ where } \alpha(n) \text{ is the number of primes which divide } n. \\ \text{Hint: The fact that each prime is } \geq 2 \text{ gives a very simple estimate of how small } \alpha(n) \text{ must be.}$$

$$*(x) \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n k^p}{n^{p+1}} = \frac{1}{p+1}.$$

2. Find the following limits.

$$(i) \lim_{n \rightarrow \infty} \frac{n}{n+1} - \frac{n+1}{n}.$$

$$(ii) \lim_{n \rightarrow \infty} n - \sqrt{n+a} \sqrt{n+b}.$$

$$(iii) \lim_{n \rightarrow \infty} \frac{2^n + (-1)^n}{2^{n+1} + (-1)^{n+1}}.$$

$$(iv) \lim_{n \rightarrow \infty} \frac{(-1)^n \sqrt{n} \sin(n^n)}{n+1}.$$

$$(v) \lim_{n \rightarrow \infty} \frac{a^n - b^n}{a^n + b^n}.$$

$$(vi) \lim_{n \rightarrow \infty} nc^n, \quad |c| < 1.$$

$$(vii) \lim_{n \rightarrow \infty} \frac{2^{n^2}}{n!}.$$

3. (a) What can be said about the sequence $\{a_n\}$ if it converges and each a_n is an integer?
 (b) Find all convergent subsequences of the sequence $1, -1, 1, -1, 1, -1, \dots$ (There are infinitely many, although there are only two limits which such subsequences can have.)
 (c) Find all convergent subsequences of the sequence $1, 2, 1, 2, 3, 1, 2, 3, 4, 1, 2, 3, 4, 5, \dots$ (There are infinitely many limits which such subsequences can have.)
 (d) Consider the sequence

$$\frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \frac{1}{6}, \dots$$

For which numbers α is there a subsequence converging to α ?

4. (a) Prove that if a subsequence of a Cauchy sequence converges, then so does the original Cauchy sequence.
 (b) Prove that any subsequence of a convergent sequence converges.
 5. (a) Prove that if $0 < a < 2$, then $a < \sqrt{2a} < 2$.
 (b) Prove that the sequence

$$\sqrt{2}, \sqrt{2\sqrt{2}}, \sqrt{2\sqrt{2\sqrt{2}}}, \dots$$

converges.

- (c) Find the limit. Hint: Notice that if $\lim_{n \rightarrow \infty} a_n = l$, then $\lim_{n \rightarrow \infty} \sqrt{2a_n} = \sqrt{2l}$, by Theorem 1.

6. Let $0 < a_1 < b_1$ and define

$$a_{n+1} = \sqrt{a_n b_n}, \quad b_{n+1} = \frac{a_n + b_n}{2}.$$

- (a) Prove that the sequences $\{a_n\}$ and $\{b_n\}$ each converge.
 (b) Prove that they have the same limit.
 7. In Problem 2-16 we saw that any rational approximation m/n to $\sqrt{2}$ can be replaced by a better approximation $(m+2n)/(m+n)$. In particular, starting with $m = n = 1$, we obtain

$$1, \frac{3}{2}, \frac{7}{5}, \dots$$

- (a) Prove that this sequence is given recursively by

$$a_1 = 1, \quad a_{n+1} = 1 + \frac{1}{1 + a_n}.$$

- (b) Prove that $\lim_{n \rightarrow \infty} a_n = \sqrt{2}$. This gives the so-called *continued fraction expansion*

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \dots}}.$$

Hint: Consider separately the subsequences $\{a_{2n}\}$ and $\{a_{2n+1}\}$.

- (c) Prove that for any natural numbers a and b ,

$$\sqrt{a^2 + b} = a + \cfrac{b}{2a + \cfrac{b}{2a + \dots}}.$$

8. Identify the function $f(x) = \lim_{n \rightarrow \infty} (\lim_{k \rightarrow \infty} (\cos n! \pi x)^{2k})$. (It has been mentioned many times in this book.)
9. Many impressive looking limits can be evaluated easily (especially by the person who makes them up), because they are really lower or upper sums in disguise. With this remark as hint, evaluate each of the following. (Warning: the list contains one red herring which can be evaluated by elementary considerations.)

$$\begin{aligned} \text{(i)} \quad & \lim_{n \rightarrow \infty} \frac{\sqrt[n]{e} + \sqrt[n]{e^2} + \dots + \sqrt[n]{e^n}}{n}. \\ \text{(ii)} \quad & \lim_{n \rightarrow \infty} \frac{\sqrt[n]{e} + \sqrt[n]{e^2} + \dots + \sqrt[n]{e^{2n}}}{n}. \\ \text{(iii)} \quad & \lim_{n \rightarrow \infty} \left(\frac{1}{n+1} + \dots + \frac{1}{2n} \right). \\ \text{(iv)} \quad & \lim_{n \rightarrow \infty} \left(\frac{1}{n^2} + \frac{1}{(n+1)^2} + \dots + \frac{1}{(2n)^2} \right). \\ \text{(v)} \quad & \lim_{n \rightarrow \infty} \left(\frac{n}{(n+1)^2} + \frac{n}{(n+2)^2} + \dots + \frac{n}{(n+n)^2} \right). \\ \text{(vi)} \quad & \lim_{n \rightarrow \infty} \left(\frac{n}{n^2+1} + \frac{n}{n^2+2^2} + \dots + \frac{n}{n^2+n^2} \right). \end{aligned}$$

10. Although limits like $\lim_{n \rightarrow \infty} \sqrt[n]{n}$ and $\lim_{n \rightarrow \infty} a^n$ can be evaluated using facts about the behavior of the logarithm and exponential functions, this approach is vaguely dissatisfying, because integral roots and powers can be defined without using the exponential function. Some of the standard “elementary” arguments for such limits are outlined here; the basic tools are inequalities derived from the binomial theorem, notably

$$(1+h)^n \geq 1 + nh, \quad \text{for } h > 0;$$

and, for part (e),

$$(1+h)^n \geq 1 + nh + \frac{n(n-1)}{2}h^2 \geq \frac{n(n-1)}{2}h^2, \quad \text{for } h > 0.$$

- (a) Prove that $\lim_{n \rightarrow \infty} a^n = \infty$ if $a > 1$, by setting $a = 1 + h$, where $h > 0$.
 (b) Prove that $\lim_{n \rightarrow \infty} a^n = 0$ if $0 < a < 1$.
 (c) Prove that $\lim_{n \rightarrow \infty} \sqrt[n]{a} = 1$ if $a > 1$, by setting $\sqrt[n]{a} = 1 + h$ and estimating h .
 (d) Prove that $\lim_{n \rightarrow \infty} \sqrt[n]{a} = 1$ if $0 < a < 1$.
 (e) Prove that $\lim_{n \rightarrow \infty} \sqrt[n]{n} = 1$.
11. (a) Prove that a convergent sequence is always bounded.
 (b) Suppose that $\lim_{n \rightarrow \infty} a_n = 0$, and that each $a_n > 0$. Prove that the set of all numbers a_n actually has a maximum member.
12. (a) Prove that

$$\frac{1}{n+1} < \log(n+1) - \log n < \frac{1}{n}.$$

(b) If

$$a_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \log n,$$

show that the sequence $\{a_n\}$ is decreasing, and that each $a_n \geq 0$. It follows that there is a number

$$\gamma = \lim_{n \rightarrow \infty} \left(1 + \cdots + \frac{1}{n} - \log n \right).$$

This number, known as Euler's number, has proved to be quite refractory; it is not even known whether γ is rational.

13. (a) Suppose that f is increasing on $[1, \infty)$. Show that

$$f(1) + \cdots + f(n-1) < \int_1^n f(x) dx < f(2) + \cdots + f(n).$$

- (b) Now choose $f = \log$ and show that

$$\frac{n^n}{e^{n-1}} < n! < \frac{(n+1)^{n+1}}{e^n};$$

it follows that

$$\lim_{n \rightarrow \infty} \frac{\sqrt[n]{n!}}{n} = \frac{1}{e}.$$

This result shows that $\sqrt[n]{n!}$ is approximately n/e , in the sense that the ratio of these two quantities is close to 1 for large n . But we cannot conclude that $n!$ is close to $(n/e)^n$ in this sense; in fact, this is false. An estimate for $n!$ is very desirable, even for concrete computations, because $n!$ cannot be calculated easily even with logarithm tables. The standard (and difficult) theorem which provides the right information will be found in Problem 27-19.

14. (a) Show that the tangent line to the graph of f at $(x_0, f(x_0))$ intersects the horizontal axis at $(x_1, 0)$, where

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

This intersection point may be regarded as a rough approximation to the point where the graph of f intersects the horizontal axis. If we now start at x_1 and repeat the process to get x_2 , then use x_2 to get x_3 , etc., we have a sequence $\{x_n\}$ defined inductively by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Figure 7 suggests that $\{x_n\}$ will converge to a number c with $f(c) = 0$; this is called *Newton's method* for finding a zero of f . In the remainder of this problem we will establish some conditions under which Newton's method works (Figures 8 and 9 show two cases where it doesn't). A few facts about convexity may be found useful; see Chapter 11, Appendix.

- (b) Suppose that $f', f'' > 0$, and that we choose x_0 with $f(x_0) > 0$. Show that $x_0 \geq x_1 \geq x_2 \geq \dots \geq c$.
- (c) Let $\delta_k = x_k - c$. Then

$$\delta_k = \frac{f(x_k)}{f'(\xi_k)}$$

for some ξ_k in (c, x_k) . Show that

$$\delta_{k+1} = \frac{f(x_k)}{f'(\xi_k)} - \frac{f(x_k)}{f'(x_k)}.$$

Conclude that

$$\delta_{k+1} = \frac{f(x_k)}{f'(\xi_k)f'(x_k)} \cdot f''(\eta)(x_k - \xi_k)$$

for some η in (c, x_k) , and then that

$$(*) \quad \delta_{k+1} \leq \frac{f''(\eta)}{f'(x_k)} \delta_k^2.$$

- (d) Let $m = \inf f'$ on $[c, x_1]$ and let $M = \sup |f''|$ on $[c, x_1]$. Show that Newton's method works if $x_0 - c < m/M$.
- (e) What is the formula for x_{n+1} when $f(x) = x^2 - A$?

If we take $A = 2$ and $x_0 = 1.4$ we get

$$\begin{aligned} x_0 &= 1.4 \\ x_1 &= 1.4142857 \\ x_2 &= 1.4142136 \\ x_3 &= 1.4142136, \end{aligned}$$

which is already correct to 7 decimals! Notice that the number of correct decimals at least doubled each time. This is essentially guaranteed by the inequality $(*)$ when $M/m < 1$.

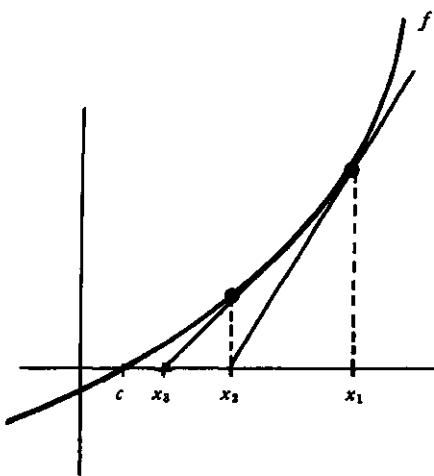


FIGURE 7

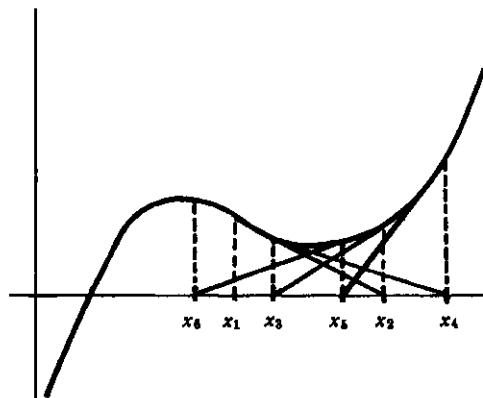


FIGURE 8

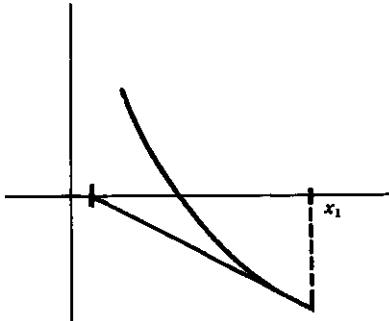


FIGURE 9

15. Use Newton's method to estimate the zeros of the following functions.

- (i) $f(x) = \tan x - \cos^2 x$ near 0.
- (ii) $f(x) = \cos x - x^2$ near 0.
- (iii) $f(x) = x^3 + x - 1$ on $[0, 1]$.
- (iv) $f(x) = x^3 - 3x^2 + 1$ on $[0, 1]$.

- *16. Prove that if $\lim_{n \rightarrow \infty} a_n = l$, then

$$\lim_{n \rightarrow \infty} \frac{(a_1 + \cdots + a_n)}{n} = l.$$

Hint: This problem is very similar to (in fact it is a special case of) Problem 13-40.

17. Suppose that f is continuous and $\lim_{x \rightarrow \infty} f(x+1) - f(x) = 0$. Prove that $\lim_{x \rightarrow \infty} f(x)/x = 0$. Hint: See the previous problem.

- *18. Suppose that $a_n > 0$ for each n and that $\lim_{n \rightarrow \infty} a_{n+1}/a_n = l$. Prove that $\lim_{n \rightarrow \infty} \sqrt[n]{a_n} = l$. Hint: This requires the same sort of argument that works in Problem 16, together with the fact that $\lim_{n \rightarrow \infty} \sqrt[n]{a} = 1$, for $a > 0$.

19. (a) Suppose that $\{a_n\}$ is a convergent sequence of points all in $[0, 1]$. Prove that $\lim_{n \rightarrow \infty} a_n$ is also in $[0, 1]$.
(b) Find a convergent sequence $\{a_n\}$ of points all in $(0, 1)$ such that $\lim_{n \rightarrow \infty} a_n$ is not in $(0, 1)$.

20. Suppose that f is continuous and that the sequence

$$x, f(x), f(f(x)), f(f(f(x))), \dots$$

converges to l . Prove that l is a "fixed point" for f , i.e., $f(l) = l$. Hint: Two special cases have occurred already.

21. (a) Suppose that f is continuous on $[0, 1]$ and that $0 \leq f(x) \leq 1$ for all x in $[0, 1]$. Problem 7-11 shows that f has a fixed point (in the terminology of Problem 20). If f is increasing, a much stronger statement can be made: For any x in $[0, 1]$, the sequence

$$x, f(x), f(f(x)), \dots$$

has a limit (which is necessarily a fixed point, by Problem 20). Prove this assertion, by examining the behavior of the sequence for $f(x) > x$ and $f(x) < x$, or by looking at Figure 10. A diagram of this sort is used in Littlewood's *Mathematician's Miscellany* to preach the value of drawing pictures: "For the professional the only proof needed is [this Figure]."

- *(b) Suppose that f and g are two continuous functions on $[0, 1]$, with $0 \leq f(x) \leq 1$ and $0 \leq g(x) \leq 1$ for all x in $[0, 1]$, which satisfy $f \circ g = g \circ f$. Suppose, moreover, that f is increasing. Show that f and g have

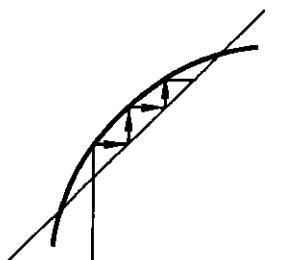


FIGURE 10

a common fixed point; in other words, there is a number l such that $f(l) = l = g(l)$. Hint: Begin by choosing a fixed point for g .

For a long time mathematicians amused themselves by asking whether the conclusion of part (b) holds without the assumption that f is increasing, but two independent announcements in the *Notices* of the American Mathematical Society, Volume 14, Number 2 give counterexamples, so it was probably a pretty silly problem all along.

The trick in Problem 20 is really much more valuable than Problem 20 might suggest, and some of the most important “fixed point theorems” depend upon looking at sequences of the form $x, f(x), f(f(x)), \dots$. A special, but representative, case of one such theorem is treated in Problem 23 (for which the next problem is preparation).

- 22.** (a) Use Problem 2-5 to show that if $c \neq 1$, then

$$c^m + c^{m+1} + \cdots + c^n = \frac{c^m - c^{n+1}}{1 - c}.$$

- (b) Suppose that $|c| < 1$. Prove that

$$\lim_{m,n \rightarrow \infty} c^m + \cdots + c^n = 0.$$

- (c) Suppose that $\{x_n\}$ is a sequence with $|x_n - x_{n+1}| \leq c^n$, where $c < 1$. Prove that $\{x_n\}$ is a Cauchy sequence.

- *23.** Suppose that f is a function on \mathbf{R} such that

$$(*) \quad |f(x) - f(y)| \leq c|x - y|, \quad \text{for all } x \text{ and } y,$$

where $c < 1$. (Such a function is called a *contraction*.)

- (a) Prove that f is continuous.
- (b) Prove that f has at most one fixed point.
- (c) By considering the sequence

$$x, f(x), f(f(x)), \dots,$$

for any x , prove that f does have a fixed point. (This result, in a more general setting, is known as the “contraction lemma.”)

- 24.** (a) Prove that if f is differentiable and $|f'| < 1$, then f has at most one fixed point.
- (b) Prove that if $|f'(x)| \leq c < 1$ for all x , then f has a fixed point.
- (c) Give an example to show that the hypothesis $|f'(x)| \leq 1$ is not sufficient to insure that f has a fixed point.
- 25.** This problem is a sort of converse to the previous problem. Let b_n be a sequence defined by $b_1 = a$, $b_{n+1} = f(b_n)$. Prove that if $b = \lim_{n \rightarrow \infty} b_n$ exists and f' is continuous at b , then $|f'(b)| \leq 1$. Hint: If $|f'(b)| > 1$, then

$|f'(x)| > 1$ for all x in an interval around b , and b_n will be in this interval for large enough n . Now consider f on the interval $[b, b_n]$.

26. This problem investigates for which $a > 0$ the symbol

$$a^{a^{\dots}}$$

makes sense. In other words, if we define $b_1 = a$, $b_{n+1} = a^{b_n}$, when does $b = \lim_{n \rightarrow \infty} b_n$ exist?

- (a) Prove that if b exists, then $a^b = b$. (The situation is similar to that in Problem 5.)
- (b) According to part (a), if b exists, then a can be written in the form $y^{1/y}$ for some y . Describe the graph of $g(y) = y^{1/y}$ and conclude that $0 < a \leq e^{1/e}$.
- (c) Suppose that $1 \leq a \leq e^{1/e}$. Show that $\{b_n\}$ is increasing and also $b_n \leq e$. This proves that b exists (and also that $b \leq e$).

The analysis for $a < 1$ is more difficult.

- (d) Using Problem 25, show that if b exists, then $e^{-1} \leq b \leq e$. Then show that $e^{-e} \leq a \leq e^{1/e}$.

From now on we will suppose that $e^{-e} \leq a < 1$.

- (e) Show that the function

$$f(x) = \frac{a^x}{\log x}$$

is decreasing on the interval $(0, 1)$.

- (f) Let b be the unique number such that $a^b = b$. Show that $a < b < 1$. Using part (e), show that if $0 < x < b$, then $x < a^{a^x} < b$. Conclude that $l = \lim_{n \rightarrow \infty} a_{2n+1}$ exists and that $a^{a^l} = l$.
- (g) Using part (e) again, show that $l = b$.
- (h) Finally, show that $\lim_{n \rightarrow \infty} a_{2n+2} = b$, so that $\lim_{n \rightarrow \infty} b_n = b$.

27. Let $\{x_n\}$ be a sequence which is bounded, and let

$$y_n = \sup\{x_n, x_{n+1}, x_{n+2}, \dots\}.$$

- (a) Prove that the sequence $\{y_n\}$ converges. The limit $\lim_{n \rightarrow \infty} y_n$ is denoted by $\overline{\lim}_{n \rightarrow \infty} x_n$ or $\limsup_{n \rightarrow \infty} x_n$, and called the **limit superior**, or **upper limit**, of the sequence $\{x_n\}$.
- (b) Find $\overline{\lim}_{n \rightarrow \infty} x_n$ for each of the following:

$$(i) \quad x_n = \frac{1}{n}.$$

$$(ii) \quad x_n = (-1)^n \frac{1}{n}.$$

$$(iii) \quad x_n = (-1)^n \left[1 + \frac{1}{n} \right].$$

$$(iv) \quad x_n = \sqrt[n]{n}.$$

- (c) Define $\underline{\lim}_{n \rightarrow \infty} x_n$ (or $\liminf_{n \rightarrow \infty} x_n$) and prove that

$$\underline{\lim}_{n \rightarrow \infty} x_n \leq \overline{\lim}_{n \rightarrow \infty} x_n.$$

- (d) Prove that $\underline{\lim}_{n \rightarrow \infty} x_n$ exists if and only if $\overline{\lim}_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} x_n$ and that in this case $\lim_{n \rightarrow \infty} x_n = \overline{\lim}_{n \rightarrow \infty} x_n = \underline{\lim}_{n \rightarrow \infty} x_n$.
- (e) Recall the definition, in Problem 8-18, of $\overline{\lim} A$ for a bounded set A . Prove that if the numbers x_n are distinct, then $\overline{\lim}_{n \rightarrow \infty} x_n = \overline{\lim} A$, where $A = \{x_n : n \in \mathbb{N}\}$.

28. In the Appendix to Chapter 8 we defined uniform continuity of a function on an interval. If $f(x)$ is defined only for rational x , this concept still makes sense: we say that f is uniformly continuous on an interval if for every $\varepsilon > 0$ there is some $\delta > 0$ such that, if x and y are rational numbers in the interval and $|x - y| < \delta$, then $|f(x) - f(y)| < \varepsilon$.
- (a) Let x be any (rational or irrational) point in the interval, and let $\{x_n\}$ be a sequence of *rational* points in the interval such that $\lim_{n \rightarrow \infty} x_n = x$. Show that the sequence $\{f(x_n)\}$ converges.
- (b) Prove that the limit of the sequence $\{f(x_n)\}$ doesn't depend on the choice of the sequence $\{x_n\}$.

We will denote this limit by $\bar{f}(x)$, so that \bar{f} is an extension of f to the whole interval.

- (c) Prove that the extended function \bar{f} is uniformly continuous on the interval.

29. Let $a > 0$, and for rational x let $f(x) = a^x$, as defined in the usual elementary algebraic way. This problem shows directly that f can be extended to a continuous function \bar{f} on the whole line. Problem 28 provides the necessary machinery.
- (a) Show that $a^x < a^y$ for rational $x < y$.
- (b) Using Problem 10, show that for any $\varepsilon > 0$ we have $|a^x - 1| < \varepsilon$ for rational numbers x close enough to 0.
- (c) Using the equation $a^x - a^y = a^y(a^{x-y} - 1)$, prove that on any closed interval f is uniformly continuous, in the sense of Problem 28.
- (d) Show that the extended function \bar{f} of Problem 28 is increasing and satisfies $\bar{f}(x+y) = \bar{f}(x)\bar{f}(y)$.

- *30. The Bolzano-Weierstrass Theorem is usually stated, and also proved, quite differently than in the text—the classical statement uses the notion of limit

points. A point x is a **limit point** of the set A if for every $\varepsilon > 0$ there is a point a in A with $|x - a| < \varepsilon$ but $x \neq a$.

- (a) Find all limit points of the following sets.

$$(i) \quad \left\{ \frac{1}{n} : n \text{ in } \mathbf{N} \right\}.$$

$$(ii) \quad \left\{ \frac{1}{n} + \frac{1}{m} : n \text{ and } m \text{ in } \mathbf{N} \right\}.$$

$$(iii) \quad \left\{ (-1)^n \left[1 + \frac{1}{n} \right] : n \text{ in } \mathbf{N} \right\}.$$

$$(iv) \quad \mathbf{Z}.$$

$$(v) \quad \mathbf{Q}.$$

- (b) Prove that x is a limit point of A if and only if for every $\varepsilon > 0$ there are infinitely many points a of A satisfying $|x - a| < \varepsilon$.
(c) Prove that $\overline{\lim} A$ is the largest limit point of A , and $\underline{\lim} A$ the smallest.

The usual form of the Bolzano-Weierstrass Theorem states that if A is an infinite set of numbers contained in a closed interval $[a, b]$, then some point of $[a, b]$ is a limit point of A . Prove this in two ways:

- (d) Using the form already proved in the text. Hint: Since A is infinite, there are distinct numbers x_1, x_2, x_3, \dots in A .
(e) Using the Nested Intervals Theorem. Hint: If $[a, b]$ is divided into two intervals, at least one must contain infinitely many points of A .
31. (a) Use the Bolzano-Weierstrass Theorem to prove that if f is continuous on $[a, b]$, then f is bounded above on $[a, b]$. Hint: If f is not bounded above, then there are points x_n in $[a, b]$ with $f(x_n) > n$.
(b) Also use the Bolzano-Weierstrass Theorem to prove that if f is continuous on $[a, b]$, then f is uniformly continuous on $[a, b]$ (see Chapter 8, Appendix).

- **32. (a) Let $\{a_n\}$ be the sequence

$$\frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \frac{1}{6}, \frac{2}{6}, \dots$$

Suppose that $0 \leq a < b \leq 1$. Let $N(n; a, b)$ be the number of integers $j \leq n$ such that a_j is in $[a, b]$. (Thus $N(2; \frac{1}{3}, \frac{2}{3}) = 2$, and $N(4; \frac{1}{3}, \frac{2}{3}) = 3$.) Prove that

$$\lim_{n \rightarrow \infty} \frac{N(n; a, b)}{n} = b - a.$$

- (b) A sequence $\{a_n\}$ of numbers in $[0, 1]$ is called **uniformly distributed** in $[0, 1]$ if

$$\lim_{n \rightarrow \infty} \frac{N(n; a, b)}{n} = b - a$$

for all a and b with $0 \leq a < b \leq 1$. Prove that if s is a step function defined on $[0, 1]$, and $\{a_n\}$ is uniformly distributed in $[0, 1]$, then

$$\int_0^1 s = \lim_{n \rightarrow \infty} \frac{s(a_1) + \cdots + s(a_n)}{n}.$$

- (c) Prove that if $\{a_n\}$ is uniformly distributed in $[0, 1]$ and f is integrable on $[0, 1]$, then

$$\int_0^1 f = \lim_{n \rightarrow \infty} \frac{f(a_1) + \cdots + f(a_n)}{n}.$$

- **33.** (a) Let f be a function defined on $[0, 1]$ such that $\lim_{y \rightarrow a} f(y)$ exists for all a in $[0, 1]$. For any $\varepsilon > 0$ prove that there are only finitely many points a in $[0, 1]$ with $|\lim_{y \rightarrow a} f(y) - f(a)| > \varepsilon$. Hint: Show that the set of such points cannot have a limit point x , by showing that $\lim_{y \rightarrow x} f(y)$ could not exist.
 (b) Prove that, in the terminology of Problem 21-5, the set of points where f is discontinuous is countable. This finally answers the question of Problem 6-16: If f has only removable discontinuities, then f is continuous except at a countable set of points, and in particular, f cannot be discontinuous everywhere.

CHAPTER 23

INFINITE SERIES

Infinite sequences were introduced in the previous chapter with the specific intention of considering their “sums”

$$a_1 + a_2 + a_3 + \dots$$

in this chapter. This is not an entirely straightforward matter, for the sum of infinitely many numbers is as yet completely undefined. What can be defined are the “partial sums”

$$s_n = a_1 + \dots + a_n,$$

and the infinite sum must presumably be defined in terms of these partial sums. Fortunately, the mechanism for formulating this definition has already been developed in the previous chapter. If there is to be any hope of computing the infinite sum $a_1 + a_2 + a_3 + \dots$, the partial sums s_n should represent closer and closer approximations as n is chosen larger and larger. This last assertion amounts to little more than a sloppy definition of limits: the “infinite sum” $a_1 + a_2 + a_3 + \dots$ ought to be $\lim_{n \rightarrow \infty} s_n$. This approach will necessarily leave the “sum” of many sequences undefined, since the sequence $\{s_n\}$ may easily fail to have a limit. For example, the sequence

$$1, -1, 1, -1, \dots$$

with $a_n = (-1)^{n+1}$ yields the new sequence

$$\begin{aligned}s_1 &= a_1 = 1, \\s_2 &= a_1 + a_2 = 0, \\s_3 &= a_1 + a_2 + a_3 = 1, \\s_4 &= a_1 + a_2 + a_3 + a_4 = 0, \\&\dots\end{aligned}$$

for which $\lim_{n \rightarrow \infty} s_n$ does not exist. Although there happen to be some clever extensions of the definition suggested here (see Problems 9 and 24-20) it seems unavoidable that some sequences will have no sum. For this reason, an acceptable definition of the sum of a sequence should contain, as an essential component, terminology which distinguishes sequences for which sums can be defined from less fortunate sequences.

DEFINITION

The sequence $\{a_n\}$ is **summable** if the sequence $\{s_n\}$ converges, where

$$s_n = a_1 + \cdots + a_n.$$

In this case, $\lim_{n \rightarrow \infty} s_n$ is denoted by

$$\sum_{n=1}^{\infty} a_n \quad (\text{or, less formally, } a_1 + a_2 + a_3 + \cdots)$$

and is called the **sum** of the sequence $\{a_n\}$.

The terminology introduced in this definition is usually replaced by less precise expressions; indeed the title of this chapter is derived from such everyday language.

An infinite sum $\sum_{n=1}^{\infty} a_n$ is usually called an *infinite series*, the word “series” emphasizing the connection with the infinite sequence $\{a_n\}$. The statement that $\{a_n\}$ is, or is not, summable is conventionally replaced by the statement that the series $\sum_{n=1}^{\infty} a_n$ does, or does not, converge. This terminology is somewhat peculiar, because at best the symbol $\sum_{n=1}^{\infty} a_n$ denotes a number (so it can’t “converge”), and it doesn’t denote anything at all unless $\{a_n\}$ is summable. Nevertheless, this informal language is convenient, standard, and unlikely to yield to attacks on logical grounds.

Certain elementary arithmetical operations on infinite series are direct consequences of the definition. It is a simple exercise to show that if $\{a_n\}$ and $\{b_n\}$ are summable, then

$$\begin{aligned} \sum_{n=1}^{\infty} (a_n + b_n) &= \sum_{n=1}^{\infty} a_n + \sum_{n=1}^{\infty} b_n, \\ \sum_{n=1}^{\infty} c \cdot a_n &= c \cdot \sum_{n=1}^{\infty} a_n. \end{aligned}$$

As yet these equations are not very interesting, since we have no examples of summable sequences (except for the trivial examples in which the terms are eventually all 0). Before we actually exhibit a summable sequence, some general conditions for summability will be recorded.

There is one necessary and sufficient condition for summability which can be stated immediately. The sequence $\{a_n\}$ is summable if and only if the sequence $\{s_n\}$ converges, which happens, according to Theorem 22-3, if and only if $\lim_{m,n \rightarrow \infty} s_m - s_n = 0$; this condition can be rephrased in terms of the original sequence as follows.

THE CAUCHY CRITERION

The sequence $\{a_n\}$ is summable if and only if

$$\lim_{m,n \rightarrow \infty} a_{n+1} + \cdots + a_m = 0.$$

Although the Cauchy criterion is of theoretical importance, it is not very useful for deciding the summability of any particular sequence. However, one simple consequence of the Cauchy criterion provides a *necessary* condition for summability which is too important not to be mentioned explicitly.

THE VANISHING CONDITION

If $\{a_n\}$ is summable, then

$$\lim_{n \rightarrow \infty} a_n = 0.$$

This condition follows from the Cauchy criterion by taking $m = n + 1$; it can also be proved directly as follows. If $\lim_{n \rightarrow \infty} s_n = l$, then

$$\begin{aligned}\lim_{n \rightarrow \infty} a_n &= \lim_{n \rightarrow \infty} (s_n - s_{n-1}) = \lim_{n \rightarrow \infty} s_n - \lim_{n \rightarrow \infty} s_{n-1} \\ &= l - l = 0.\end{aligned}$$

Unfortunately, this condition is far from sufficient. For example, $\lim_{n \rightarrow \infty} 1/n = 0$, but the sequence $\{1/n\}$ is not summable; in fact, the following grouping of the numbers $1/n$ shows that the sequence $\{s_n\}$ is not bounded:

$$\begin{aligned}1 + \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{\geq \frac{1}{2}} + \underbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}_{\geq \frac{1}{2}} + \underbrace{\frac{1}{9} + \cdots + \frac{1}{16}}_{\geq \frac{1}{2}} + \cdots\end{aligned}$$

(2 terms, each $\geq \frac{1}{4}$) (4 terms, each $\geq \frac{1}{8}$) (8 terms, each $\geq \frac{1}{16}$)

The method of proof used in this example, a clever trick which one might never see, reveals the need for some more standard methods for attacking these problems. These methods shall be developed soon (one of them will give an alternate proof that $\sum_{n=1}^{\infty} 1/n$ does not converge) but it will be necessary to first procure a few examples of convergent series.

The most important of all infinite series are the “geometric series”

$$\sum_{n=0}^{\infty} r^n = 1 + r + r^2 + r^3 + \cdots$$

Only the cases $|r| < 1$ are interesting, since the individual terms do not approach 0 if $|r| \geq 1$. These series can be managed because the partial sums

$$s_n = 1 + r + \cdots + r^n$$

can be evaluated in simple terms. The two equations

$$\begin{aligned}s_n &= 1 + r + r^2 + \cdots + r^n \\ rs_n &= \quad r + r^2 + \cdots + r^n + r^{n+1}\end{aligned}$$

lead to

$$s_n(1 - r) = 1 - r^{n+1}$$

or

$$s_n = \frac{1 - r^{n+1}}{1 - r}$$

(division by $1 - r$ is valid since we are not considering the case $r = 1$). Now $\lim_{n \rightarrow \infty} r^n = 0$, since $|r| < 1$. It follows that

$$\sum_{n=0}^{\infty} r^n = \lim_{n \rightarrow \infty} \frac{1 - r^{n+1}}{1 - r} = \frac{1}{1 - r}, \quad |r| < 1.$$

In particular,

$$\sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^n = \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n - 1 = \frac{1}{1 - \frac{1}{2}} - 1 = 1,$$

that is,

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \dots = 1,$$

an infinite sum which can always be remembered from the picture in Figure 1.

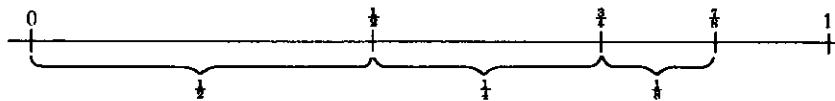


FIGURE 1

Special as they are, geometric series are standard examples from which important tests for summability will be derived.

For a while we shall consider only sequences $\{a_n\}$ with each $a_n \geq 0$; such sequences are called **nonnegative**. If $\{a_n\}$ is a nonnegative sequence, then the sequence $\{s_n\}$ is clearly nondecreasing. This remark, combined with Theorem 22-2, provides a simple-minded test for summability:

THE BOUNDEDNESS CRITERION

A nonnegative sequence $\{a_n\}$ is summable if and only if the set of partial sums s_n is bounded.

By itself, this criterion is not very helpful—deciding whether or not the set of all s_n is bounded is just what we are unable to do. On the other hand, if some convergent series are already available for comparison, this criterion can be used to obtain a result whose simplicity belies its importance (it is the basis for almost all other tests).

THEOREM 1 (THE COMPARISON TEST)

Suppose that

$$0 \leq a_n \leq b_n \quad \text{for all } n.$$

Then if $\sum_{n=1}^{\infty} b_n$ converges, so does $\sum_{n=1}^{\infty} a_n$.

PROOF If

$$\begin{aligned}s_n &= a_1 + \cdots + a_n, \\ t_n &= b_1 + \cdots + b_n,\end{aligned}$$

then

$$0 \leq s_n \leq t_n \quad \text{for all } n.$$

Now $\{t_n\}$ is bounded, since $\sum_{n=1}^{\infty} b_n$ converges. Therefore $\{s_n\}$ is bounded; consequently, by the boundedness criterion $\sum_{n=1}^{\infty} a_n$ converges. ■

Quite frequently the comparison test can be used to analyze very complicated looking series in which most of the complication is irrelevant. For example,

$$\sum_{n=1}^{\infty} \frac{2 + \sin^3(n+1)}{2^n + n^2}$$

converges because

$$0 \leq \frac{2 + \sin^3(n+1)}{2^n + n^2} < \frac{3}{2^n},$$

and

$$\sum_{n=1}^{\infty} \frac{3}{2^n} = 3 \sum_{n=1}^{\infty} \frac{1}{2^n}$$

is a convergent (geometric) series.

Similarly, we would expect the series

$$\sum_{n=1}^{\infty} \frac{1}{2^n - 1 + \sin^2 n^3}$$

to converge, since the n th term of the series is practically $1/2^n$ for large n , and we would expect the series

$$\sum_{n=1}^{\infty} \frac{n+1}{n^2 + 1}$$

to diverge, since $(n+1)/(n^2 + 1)$ is practically $1/n$ for large n . These facts can be derived immediately from the following theorem, another kind of “comparison test.”

THEOREM 2 If $a_n, b_n > 0$ and $\lim_{n \rightarrow \infty} a_n/b_n = c \neq 0$, then $\sum_{n=1}^{\infty} a_n$ converges if and only if $\sum_{n=1}^{\infty} b_n$ converges.

PROOF Suppose $\sum_{n=1}^{\infty} b_n$ converges. Since $\lim_{n \rightarrow \infty} a_n/b_n = c$, there is some N such that

$$a_n \leq 2cb_n \quad \text{for } n \geq N.$$

But the sequence $2c \sum_{n=N}^{\infty} b_n$ certainly converges. Then Theorem 1 shows that $\sum_{n=N}^{\infty} a_n$ converges, and this implies convergence of the whole series $\sum_{n=1}^{\infty} a_n$, which has only finitely many additional terms.

The converse follows immediately, since we also have $\lim_{n \rightarrow \infty} b_n/a_n = 1/c \neq 0$. ■

The comparison test yields other important tests when we use previously analyzed series as catalysts. Choosing the geometric series $\sum_{n=0}^{\infty} r^n$, the convergent series *par excellence*, we obtain the most important of all tests for summability.

THEOREM 3 (THE RATIO TEST)

Let $a_n > 0$ for all n , and suppose that

$$\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = r.$$

Then $\sum_{n=1}^{\infty} a_n$ converges if $r < 1$. On the other hand, if $r > 1$, then the terms a_n do not approach 0, so $\sum_{n=1}^{\infty} a_n$ diverges. (Notice that it is therefore essential to compute $\lim_{n \rightarrow \infty} a_{n+1}/a_n$ and not $\lim_{n \rightarrow \infty} a_n/a_{n+1}$!)

PROOF Suppose first that $r < 1$. Choose any number s with $r < s < 1$. The hypothesis

$$\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = r < 1$$

implies that there is some N such that

$$\frac{a_{n+1}}{a_n} \leq s \quad \text{for } n \geq N.$$

This can be written

$$a_{n+1} \leq sa_n \quad \text{for } n \geq N.$$

Thus

$$\begin{aligned} a_{N+1} &\leq sa_N, \\ a_{N+2} &\leq sa_{N+1} \leq s^2 a_N, \\ &\vdots \\ &\vdots \\ a_{N+k} &\leq s^k a_N. \end{aligned}$$

Since $\sum_{k=0}^{\infty} a_N s^k = a_N \sum_{k=0}^{\infty} s^k$ converges, the comparison test shows that

$$\sum_{n=N}^{\infty} a_n = \sum_{k=0}^{\infty} a_{N+k}$$

converges. This implies the convergence of the whole series $\sum_{n=1}^{\infty} a_n$.

The case $r > 1$ is even easier. If $1 < s < r$, then there is a number N such that

$$\frac{a_{n+1}}{a_n} \geq s \quad \text{for } n \geq N,$$

which means that

$$a_{N+k} \geq a_N s^k \geq a_N \quad k = 0, 1, \dots$$

This shows that the individual terms of $\{a_n\}$ do not approach 0, so $\{a_n\}$ is not summable. ■

As a simple application of the ratio test, consider the series $\sum_{n=1}^{\infty} 1/n!$. Letting $a_n = 1/n!$ we obtain

$$\frac{a_{n+1}}{a_n} = \frac{\frac{1}{(n+1)!}}{\frac{1}{n!}} = \frac{n!}{(n+1)!} = \frac{1}{n+1}.$$

Thus

$$\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} = 0,$$

which shows that the series $\sum_{n=1}^{\infty} 1/n!$ converges. If we consider instead the series

$\sum_{n=1}^{\infty} r^n/n!$, where r is some fixed positive number, then

$$\lim_{n \rightarrow \infty} \frac{\frac{r^{n+1}}{(n+1)!}}{\frac{r^n}{n!}} = \lim_{n \rightarrow \infty} \frac{r}{n+1} = 0,$$

so $\sum_{n=1}^{\infty} r^n/n!$ converges. It follows that

$$\lim_{n \rightarrow \infty} \frac{r^n}{n!} = 0,$$

a result already proved in Chapter 16 (the proof given there was based on the same ideas as those used in the ratio test). Finally, if we consider the series $\sum_{n=1}^{\infty} nr^n$ we have

$$\lim_{n \rightarrow \infty} \frac{(n+1)r^{n+1}}{nr^n} = \lim_{n \rightarrow \infty} r \cdot \frac{n+1}{n} = r,$$

since $\lim_{n \rightarrow \infty} (n+1)/n = 1$. This proves that if $0 \leq r < 1$, then $\sum_{n=1}^{\infty} nr^n$ converges, and consequently

$$\lim_{n \rightarrow \infty} nr^n = 0.$$

(This result clearly holds for $-1 < r \leq 0$, also.) It is a useful exercise to provide a direct proof of this limit, without using the ratio test as an intermediary.

Although the ratio test will be of the utmost theoretical importance, as a practical tool it will frequently be found disappointing. One drawback of the ratio test is the fact that $\lim_{n \rightarrow \infty} a_{n+1}/a_n$ may be quite difficult to determine, and may not even exist. A more serious deficiency, which appears with maddening regularity, is the fact that the limit might equal 1. The case $\lim_{n \rightarrow \infty} a_{n+1}/a_n = 1$ is precisely the one which is inconclusive: $\{a_n\}$ might not be summable (for example, if $a_n = 1/n$), but then again it might be. In fact, our very next test will show that $\sum_{n=1}^{\infty} (1/n)^2$ converges, even though

$$\lim_{n \rightarrow \infty} \frac{\left(\frac{1}{n+1}\right)^2}{\left(\frac{1}{n}\right)^2} = 1.$$

This test provides a quite different method for determining convergence or divergence of infinite series—like the ratio test, it is an immediate consequence of the comparison test, but the series chosen for comparison is quite novel.

THEOREM 4 (THE INTEGRAL TEST)

Suppose that f is positive and decreasing on $[1, \infty)$, and that $f(n) = a_n$ for all n .

Then $\sum_{n=1}^{\infty} a_n$ converges if and only if the limit

$$\int_1^{\infty} f = \lim_{A \rightarrow \infty} \int_1^A f$$

exists

PROOF The existence of $\lim_{A \rightarrow \infty} \int_1^A f$ is equivalent to convergence of the series

$$\int_1^2 f + \int_2^3 f + \int_3^4 f + \dots$$

Now, since f is decreasing we have (Figure 2)

$$f(n+1) < \int_n^{n+1} f < f(n).$$

The first half of this double inequality shows that the series $\sum_{n=1}^{\infty} a_{n+1}$ may be compared to the series $\sum_{n=1}^{\infty} \int_n^{n+1} f$, proving that $\sum_{n=1}^{\infty} a_{n+1}$ (and hence $\sum_{n=1}^{\infty} a_n$) converges if $\lim_{A \rightarrow \infty} \int_1^A f$ exists.

The second half of the inequality shows that the series $\sum_{n=1}^{\infty} \int_n^{n+1} f$ may be compared to the series $\sum_{n=1}^{\infty} a_n$, proving that $\lim_{A \rightarrow \infty} \int_1^A f$ must exist if $\sum_{n=1}^{\infty} a_n$ converges. ■

Only one example using the integral test will be given here, but it settles the question of convergence for infinitely many series at once. If $p > 0$, the convergence of $\sum_{n=1}^{\infty} 1/n^p$ is equivalent, by the integral test, to the existence of

$$\int_1^{\infty} \frac{1}{x^p} dx.$$

Now

$$\int_1^A \frac{1}{x^p} dx = \begin{cases} -\frac{1}{(p-1)} \cdot \frac{1}{A^{p-1}} + \frac{1}{p-1}, & p \neq 1 \\ \log A, & p = 1. \end{cases}$$

This shows that $\lim_{A \rightarrow \infty} \int_1^A 1/x^p dx$ exists if $p > 1$, but not if $p \leq 1$. Thus $\sum_{n=1}^{\infty} 1/n^p$ converges precisely for $p > 1$. In particular, $\sum_{n=1}^{\infty} 1/n$ diverges.

The tests considered so far apply only to nonnegative sequences, but nonpositive sequences may be handled in precisely the same way. In fact, since

$$\sum_{n=1}^{\infty} a_n = -\left(\sum_{n=1}^{\infty} -a_n\right),$$

all considerations about nonpositive sequences can be reduced to questions involving nonnegative sequences. Sequences which contain both positive and negative terms are quite another story.

If $\sum_{n=1}^{\infty} a_n$ is a sequence with both positive and negative terms, one can consider instead the sequence $\sum_{n=1}^{\infty} |a_n|$, all of whose terms are nonnegative. Cheerfully

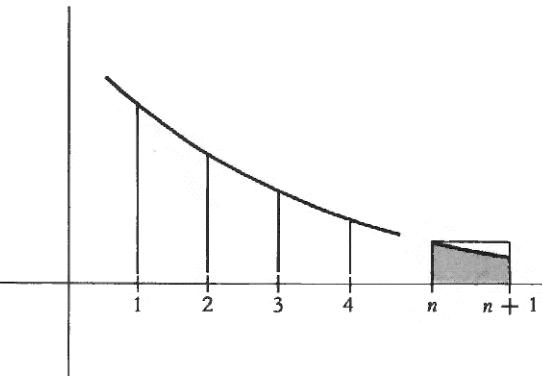


FIGURE 2

ignoring the possibility that we may have thrown away all the interesting information about the original sequence, we proceed to eulogize those sequences which are converted by this procedure into convergent sequences.

DEFINITION

The series $\sum_{n=1}^{\infty} a_n$ is **absolutely convergent** if the series $\sum_{n=1}^{\infty} |a_n|$ is convergent.

(In more formal language, the sequence $\{a_n\}$ is **absolutely summable** if the sequence $\{|a_n|\}$ is summable.)

Although we have no right to expect this definition to be of any interest, it turns out to be exceedingly important. The following theorem shows that the definition is at least not entirely useless.

THEOREM 5

Every absolutely convergent series is convergent. Moreover, a series is absolutely convergent if and only if the series formed from its positive terms and the series formed from its negative terms both converge.

PROOF

If $\sum_{n=1}^{\infty} |a_n|$ converges, then, by the Cauchy criterion,

$$\lim_{m,n \rightarrow \infty} |a_{n+1}| + \cdots + |a_m| = 0.$$

Since

$$|a_{n+1} + \cdots + a_m| \leq |a_{n+1}| + \cdots + |a_m|,$$

it follows that

$$\lim_{m,n \rightarrow \infty} a_{n+1} + \cdots + a_m = 0,$$

which shows that $\sum_{n=1}^{\infty} a_n$ converges.

To prove the second part of the theorem, let

$$a_n^+ = \begin{cases} a_n, & \text{if } a_n \geq 0 \\ 0, & \text{if } a_n \leq 0, \end{cases}$$

$$a_n^- = \begin{cases} a_n, & \text{if } a_n \leq 0 \\ 0, & \text{if } a_n \geq 0, \end{cases}$$

so that $\sum_{n=1}^{\infty} a_n^+$ is the series formed from the positive terms of $\sum_{n=1}^{\infty} a_n$, and $\sum_{n=1}^{\infty} a_n^-$ is the series formed from the negative terms.

If $\sum_{n=1}^{\infty} a_n^+$ and $\sum_{n=1}^{\infty} a_n^-$ both converge, then

$$\sum_{n=1}^{\infty} |a_n| = \sum_{n=1}^{\infty} [a_n^+ - (a_n^-)] = \sum_{n=1}^{\infty} a_n^+ - \sum_{n=1}^{\infty} a_n^-$$

also converges, so $\sum_{n=1}^{\infty} a_n$ converges absolutely.

On the other hand, if $\sum_{n=1}^{\infty} |a_n|$ converges, then, as we have just shown, $\sum_{n=1}^{\infty} a_n$ also converges. Therefore

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

and

$$\sum_{n=1}^{\infty} a_n^- = \frac{1}{2} \left(\sum_{n=1}^{\infty} a_n - \sum_{n=1}^{\infty} |a_n| \right)$$

both converge. ■

It follows from Theorem 5 that every convergent series with positive terms can be used to obtain infinitely many other convergent series, simply by putting in minus signs at random. Not every convergent series can be obtained in this way, however—there are series which are convergent but not absolutely convergent (such series are called **conditionally convergent**). In order to prove this statement we need a test for convergence which applies specifically to series with positive and negative terms.

THEOREM 6 (LEIBNIZ'S THEOREM)

Suppose that

$$a_1 \geq a_2 \geq a_3 \geq \dots \geq 0,$$

and that

$$\lim_{n \rightarrow \infty} a_n = 0.$$

Then the series

$$\sum_{n=1}^{\infty} (-1)^{n+1} a_n = a_1 - a_2 + a_3 - a_4 + a_5 - \dots$$

converges.

PROOF Figure 3 illustrates relationships between the partial sums which we will establish:

- (1) $s_2 \leq s_4 \leq s_6 \leq \dots$,
- (2) $s_1 \geq s_3 \geq s_5 \geq \dots$,
- (3) $s_k \leq s_l \quad \text{if } k \text{ is even and } l \text{ is odd.}$



FIGURE 3

To prove the first two inequalities, observe that

$$\begin{aligned} (1) \quad s_{2n+2} &= s_{2n} + a_{2n+1} - a_{2n+2} \\ &\geq s_{2n}, \quad \text{since } a_{2n+1} \geq a_{2n+2} \\ (2) \quad s_{2n+3} &= s_{2n+1} - a_{2n+2} + a_{2n+3} \\ &\geq s_{2n+1}, \quad \text{since } a_{2n+2} \geq a_{2n+3}. \end{aligned}$$

To prove the third inequality, notice first that

$$\begin{aligned} s_{2n} &= s_{2n_1} - a_{2n} \\ &\leq s_{2n-1} \quad \text{since } a_{2n} \geq 0. \end{aligned}$$

This proves only a special case of (3), but in conjunction with (1) and (2) the general case is easy: if k is even and l is odd, choose n such that

$$2n \geq k \quad \text{and} \quad 2n-1 \geq l;$$

then

$$s_k \leq s_{2n} \leq s_{2n-1} \leq s_l,$$

which proves (3).

Now, the sequence $\{s_{2n}\}$ converges, because it is nondecreasing and is bounded above (by s_l for any odd l). Let

$$\alpha = \sup\{s_{2n}\} = \lim_{n \rightarrow \infty} s_{2n}.$$

Similarly, let

$$\beta = \inf\{s_{2n+1}\} = \lim_{n \rightarrow \infty} s_{2n+1}.$$

It follows from (3) that $\alpha \leq \beta$; since

$$s_{2n+1} - s_{2n} = a_{2n+1} \quad \text{and} \quad \lim_{n \rightarrow \infty} a_n = 0$$

it is actually the case that $\alpha = \beta$. This proves that $\alpha = \beta = \lim_{n \rightarrow \infty} s_n$. ■

The standard example derived from Theorem 6 is the series

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots,$$

which is convergent, but *not* absolutely convergent (since $\sum_{n=1}^{\infty} 1/n$ does not converge). If the sum of this series is denoted by x , the following manipulations lead to quite a paradoxical result:

$$\begin{aligned} x &= 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \cdots \\ &= 1 - \frac{1}{2} - \frac{1}{4} + \frac{1}{3} - \frac{1}{6} - \frac{1}{8} + \frac{1}{5} - \frac{1}{10} - \frac{1}{12} + \frac{1}{7} - \frac{1}{14} - \frac{1}{16} + \cdots \\ &\quad (\text{the pattern here is one positive term followed by two negative ones}) \\ &= (1 - \frac{1}{2}) - \frac{1}{4} + (\frac{1}{3} - \frac{1}{6}) - \frac{1}{8} + (\frac{1}{5} - \frac{1}{10}) - \frac{1}{12} + (\frac{1}{7} - \frac{1}{14}) - \frac{1}{16} + \cdots \\ &= \frac{1}{2} - \frac{1}{4} + \frac{1}{6} - \frac{1}{8} + \frac{1}{10} - \frac{1}{12} + \frac{1}{14} - \frac{1}{16} + \cdots \\ &= \frac{1}{2}(1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \cdots) \\ &= \frac{1}{2}x, \end{aligned}$$

so $x = x/2$, implying that $x = 0$. On the other hand, it is easy to see that $x \neq 0$: the partial sum s_2 equals $\frac{1}{2}$, and the proof of Leibniz's Theorem shows that $x \geq s_2$.

This contradiction depends on a step which takes for granted that operations valid for finite sums necessarily have analogues for infinite sums. It is true that the sequence

$$\{b_n\} = 1, -\frac{1}{2}, -\frac{1}{4}, \frac{1}{3}, -\frac{1}{6}, -\frac{1}{8}, \frac{1}{5}, -\frac{1}{10}, -\frac{1}{12}, \dots$$

contains all the numbers in the sequence

$$\{a_n\} = 1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \frac{1}{5}, -\frac{1}{6}, \frac{1}{7}, -\frac{1}{8}, \frac{1}{9}, -\frac{1}{10}, \frac{1}{11}, -\frac{1}{12}, \dots$$

In fact, $\{b_n\}$ is a **rearrangement** of $\{a_n\}$ in the following precise sense: each $b_n = a_{f(n)}$ where f is a certain function which "permutes" the natural numbers, that is, every natural number m is $f(n)$ for precisely one n . In our example

$$\begin{aligned} f(2m+1) &= 3m+1 && (\text{the terms } 1, \frac{1}{3}, \frac{1}{5}, \dots \text{ go into the 1st, 4th, 7th, \dots places}), \\ f(4m) &= 3m && (\text{the terms } -\frac{1}{4}, -\frac{1}{8}, -\frac{1}{12}, \dots \text{ go into the 3rd, 6th, 9th, \dots places}), \\ f(4m+2) &= 3m+2 && (\text{the terms } -\frac{1}{2}, -\frac{1}{6}, -\frac{1}{10}, \dots \text{ go into the 2nd, 5th, 8th, \dots places}). \end{aligned}$$

Nevertheless, there is no reason to assume that $\sum_{n=1}^{\infty} b_n$ should equal $\sum_{n=1}^{\infty} a_n$: these sums are, by definition, $\lim_{n \rightarrow \infty} b_1 + \dots + b_n$ and $\lim_{n \rightarrow \infty} a_1 + \dots + a_n$, so the particular order of the terms can quite conceivably matter. The series $\sum_{n=1}^{\infty} (-1)^{n+1}/n$ is not special in this regard; indeed, its behavior is typical of series which are not absolutely convergent—the following result (really more of a grand counterexample than a theorem) shows how bad conditionally convergent series are.

THEOREM 7 If $\sum_{n=1}^{\infty} a_n$ converges, but does not converge absolutely, then for any number α there is a rearrangement $\{b_n\}$ of $\{a_n\}$ such that $\sum_{n=1}^{\infty} b_n = \alpha$.

PROOF Let $\sum_{n=1}^{\infty} p_n$ denote the series formed from the positive terms of $\{a_n\}$ and let $\sum_{n=1}^{\infty} q_n$ denote the series of negative terms. It follows from Theorem 5 that at least one of these series does not converge. As a matter of fact, both must fail to converge, for if one had bounded partial sums, and the other had unbounded partial sums, then

the original series $\sum_{n=1}^{\infty} a_n$ would also have unbounded partial sums, contradicting the assumption that it converges.

Now let α be any number. Assume, for simplicity, that $\alpha > 0$ (the proof for $\alpha < 0$ will be a simple modification). Since the series $\sum_{n=1}^{\infty} p_n$ is not convergent, there is a number N such that

$$\sum_{n=1}^N p_n > \alpha.$$

We will choose N_1 to be the *smallest* N with this property. This means that

$$(1) \quad \sum_{n=1}^{N_1-1} p_n \leq \alpha,$$

but (2) $\sum_{n=1}^{N_1} p_n > \alpha.$

Then if

$$S_1 = \sum_{n=1}^{N_1} p_n,$$

we have

$$S_1 - \alpha \leq p_{N_1}.$$

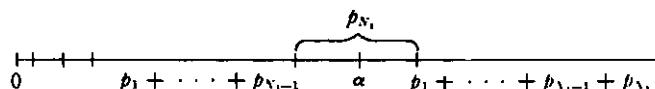


FIGURE 4

This relation, which is clear from Figure 4, follows immediately from equation (1):

$$S_1 - \alpha \leq S_1 - \sum_{n=1}^{N_1-1} p_n = p_{N_1}.$$

To the sum S_1 we now add on just enough negative terms to obtain a new sum T_1 which is less than α . In other words, we choose the smallest integer M_1 for which

$$T_1 = S_1 + \sum_{n=1}^{M_1} q_n < \alpha.$$

As before, we have

$$\alpha - T_1 \leq -q_{M_1}.$$

We now continue this procedure indefinitely, obtaining sums alternately larger and smaller than α , each time choosing the smallest N_k or M_k possible. The

sequence

$$p_1, \dots, p_{N_1}, q_1, \dots, q_{M_1}, p_{N_1+1}, \dots, p_{N_2}, \dots$$

is a rearrangement of $\{a_n\}$. The partial sums of this rearrangement increase to S_1 , then decrease to T_1 , then increase to S_2 , then decrease to T_2 , etc. To complete the proof we simply note that $|S_k - \alpha|$ and $|T_k - \alpha|$ are less than or equal to p_{N_k} or $-q_{M_k}$, respectively, and that these terms, being members of the original sequence $\{a_n\}$, must decrease to 0, since $\sum_{n=1}^{\infty} a_n$ converges. ■

Together with Theorem 7, the next theorem establishes conclusively the distinction between conditionally convergent and absolutely convergent series.

THEOREM 8 If $\sum_{n=1}^{\infty} a_n$ converges absolutely, and $\{b_n\}$ is any rearrangement of $\{a_n\}$, then $\sum_{n=1}^{\infty} b_n$ also converges (absolutely), and

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} b_n.$$

PROOF Let us denote the partial sums of $\{a_n\}$ by s_n , and the partial sums of $\{b_n\}$ by t_n .

Suppose that $\varepsilon > 0$. Since $\sum_{n=1}^{\infty} a_n$ converges, there is some N such that

$$\left| \sum_{n=1}^{\infty} a_n - s_N \right| < \varepsilon.$$

Moreover, since $\sum_{n=1}^{\infty} |a_n|$ converges, we can also choose N so that

$$\sum_{n=1}^{\infty} |a_n| - (|a_1| + \dots + |a_N|) < \varepsilon,$$

i.e., so that

$$|a_{N+1}| + |a_{N+2}| + |a_{N+3}| + \dots < \varepsilon.$$

Now choose M so large that each of a_1, \dots, a_N appear among b_1, \dots, b_M . Then whenever $m > M$, the difference $t_m - s_N$ is the sum of certain a_i , where a_1, \dots, a_N are definitely excluded. Consequently,

$$|t_m - s_N| \leq |a_{N+1}| + |a_{N+2}| + |a_{N+3}| + \dots$$

Thus, if $m > M$, then

$$\begin{aligned} \left| \sum_{n=1}^{\infty} a_n - t_m \right| &= \left| \sum_{n=1}^{\infty} a_n - s_N - (t_m - s_N) \right| \\ &\leq \left| \sum_{n=1}^{\infty} a_n - s_N \right| + |t_m - s_N| \\ &< \varepsilon + \varepsilon. \end{aligned}$$

Since this is true for every $\varepsilon > 0$, the series $\sum_{n=1}^{\infty} b_n$ converges to $\sum_{n=1}^{\infty} a_n$.

To show that $\sum_{n=1}^{\infty} b_n$ converges absolutely, note that $\{|b_n|\}$ is a rearrangement of $\{|a_n|\}$; since $\sum_{n=1}^{\infty} |a_n|$ converges absolutely, $\sum_{n=1}^{\infty} |b_n|$ converges by the first part of the theorem. ■

Absolute convergence is also important when we want to multiply two infinite series. Unlike the situation for addition, where we have the simple formula

$$\sum_{n=1}^{\infty} a_n + \sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty} (a_n + b_n),$$

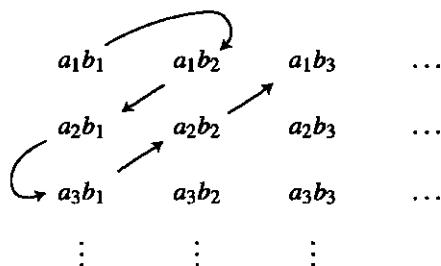
there isn't quite so obvious a candidate for the product

$$\left(\sum_{n=1}^{\infty} a_n \right) \cdot \left(\sum_{n=1}^{\infty} b_n \right) = (a_1 + a_2 + \dots) \cdot (b_1 + b_2 + \dots).$$

It would seem that we ought to sum all the products $a_i b_j$. The trouble is that these form a two-dimensional array, rather than a sequence:

$$\begin{array}{ccccccc} a_1 b_1 & a_1 b_2 & a_1 b_3 & \dots \\ a_2 b_1 & a_2 b_2 & a_2 b_3 & \dots \\ a_3 b_1 & a_3 b_2 & a_3 b_3 & \dots \\ \vdots & \vdots & \vdots & & & & \end{array}$$

Nevertheless, all the elements of this array can be arranged in a sequence. The picture below shows one way of doing this, and of course, there are (infinitely) many other ways.



Suppose that $\{c_n\}$ is some sequence of this sort, containing each product $a_i b_j$ just once. Then we might naively expect to have

$$\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{\infty} a_n \cdot \sum_{n=1}^{\infty} b_n.$$

But this *isn't* true (see Problem 8), nor is this really so surprising, since we've said nothing about the specific arrangement of the terms. The next theorem shows that the result does hold when the arrangement of terms is irrelevant.

THEOREM 9 If $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ converge absolutely, and $\{c_n\}$ is any sequence containing the products $a_i b_j$ for each pair (i, j) , then

$$\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{\infty} a_n \cdot \sum_{n=1}^{\infty} b_n.$$

PROOF Notice first that the sequence

$$p_L = \sum_{i=1}^L |a_i| \cdot \sum_{j=1}^L |b_j|$$

converges, since $\{a_n\}$ and $\{b_n\}$ are absolutely convergent, and since the limit of a product is the product of the limits. So $\{p_L\}$ is a Cauchy sequence, which means that for any $\varepsilon > 0$, if L and L' are large enough, then

$$\left| \sum_{i=1}^{L'} |a_i| \cdot \sum_{j=1}^{L'} |b_j| - \sum_{i=1}^L |a_i| \cdot \sum_{j=1}^L |b_j| \right| < \frac{\varepsilon}{2}.$$

It follows that

$$(1) \quad \sum_{i \text{ or } j > L} |a_i| \cdot |b_j| \leq \frac{\varepsilon}{2} < \varepsilon.$$

Now suppose that N is any number so large that the terms c_n for $n \leq N$ include every term $a_i b_j$ for $i, j \leq L$. Then the difference

$$\sum_{n=1}^N c_n - \sum_{i=1}^L a_i \cdot \sum_{j=1}^L b_j$$

consists of terms $a_i b_j$ with $i > L$ or $j > L$, so

$$(2) \quad \left| \sum_{n=1}^N c_n - \sum_{i=1}^L a_i \cdot \sum_{j=1}^L b_j \right| \leq \sum_{i \text{ or } j > L} |a_i| \cdot |b_j| < \varepsilon \quad \text{by (1).}$$

But since the limit of a product is the product of the limits, we also have

$$(3) \quad \left| \sum_{i=1}^{\infty} a_i \cdot \sum_{j=1}^{\infty} b_j - \sum_{i=1}^L a_i \cdot \sum_{j=1}^L b_j \right| < \varepsilon$$

for large enough L . Consequently, if we choose L , and then N , large enough, we will have

$$\begin{aligned} \left| \sum_{i=1}^{\infty} a_i \cdot \sum_{j=1}^{\infty} b_j - \sum_{i=1}^N c_n \right| &\leq \left| \sum_{i=1}^{\infty} a_i \cdot \sum_{j=1}^{\infty} b_j - \sum_{i=1}^L a_i \cdot \sum_{j=1}^L b_j \right| \\ &\quad + \left| \sum_{i=1}^L a_i \cdot \sum_{j=1}^L b_j - \sum_{n=1}^N c_n \right| \\ &< 2\epsilon \quad \text{by (2) and (3),} \end{aligned}$$

which proves the theorem. ■

Unlike our previous theorems, which were merely concerned with summability, this result says something about the actual sums. Generally speaking, there is no reason to presume that a given infinite sum can be “evaluated” in any simpler terms. However, many simple expressions can be equated to infinite sums by using Taylor’s Theorem. Chapter 20 provides many examples of functions for which

$$f(x) = \sum_{i=0}^n \frac{f^{(i)}(a)}{i!} (x-a)^i + R_{n,a}(x),$$

where $\lim_{n \rightarrow \infty} R_{n,a}(x) = 0$. This is precisely equivalent to

$$f(x) = \lim_{n \rightarrow \infty} \sum_{i=0}^n \frac{f^{(i)}(a)}{i!} (x-a)^i,$$

which means, in turn, that

$$f(x) = \sum_{i=0}^{\infty} \frac{f^{(i)}(a)}{i!} (x-a)^i.$$

As particular examples we have

$$\begin{aligned} \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots, \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots, \\ e^x &= 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots, \\ \arctan x &= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots, \quad |x| \leq 1, \\ \log(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} + \dots, \quad -1 < x \leq 1. \end{aligned}$$

(Notice that the series for $\arctan x$ and $\log(1+x)$ do not even converge for $|x| > 1$; in addition, when $x = -1$, the series for $\log(1+x)$ becomes

$$-1 - \frac{1}{2} - \frac{1}{3} - \frac{1}{4} - \dots$$

which does not converge.)

Some pretty impressive results are obtained with particular values of x :

$$0 = \pi - \frac{\pi^3}{3!} + \frac{\pi^5}{5!} - \frac{\pi^7}{7!} + \dots,$$

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots,$$

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots,$$

$$\log 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots.$$

More significant developments may be anticipated if we compare the series for $\sin x$ and $\cos x$ a little more carefully. The series for $\cos x$ is just the one we would have obtained if we had enthusiastically differentiated both sides of the equation

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots$$

term-by-term, ignoring the fact that we have never proved anything about the derivatives of infinite sums. Likewise, if we differentiate both sides of the formula for $\cos x$ formally (i.e., without justification) we obtain the formula $\cos'(x) = -\sin x$, and if we differentiate the formula for e^x we obtain $\exp'(x) = \exp(x)$. In the next chapter we shall see that such term-by-term differentiation of infinite sums is indeed valid in certain important cases.

PROBLEMS

- Decide whether each of the following infinite series is convergent or divergent. The tools which you will need are Leibniz's Theorem and the comparison, ratio, and integral tests. A few examples have been picked with malice aforethought; two series which look quite similar may require different tests (and then again, they may not). The hint below indicates which tests may be used.

(i) $\sum_{n=1}^{\infty} \frac{\sin n\theta}{n^2}$.

(ii) $1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots$.

(iii) $1 - \frac{1}{2} + \frac{2}{3} - \frac{1}{3} + \frac{2}{4} - \frac{1}{4} + \frac{2}{5} - \frac{1}{5} + \dots$.

(iv) $\sum_{n=1}^{\infty} (-1)^n \frac{\log n}{n}$.

(v) $\sum_{n=2}^{\infty} \frac{1}{\sqrt[3]{n^2 - 1}}$. (The summation begins with $n = 2$ simply to avoid the meaningless term obtained for $n = 1$).

$$(vi) \sum_{n=1}^{\infty} \frac{1}{\sqrt[3]{n^2 + 1}}.$$

$$(vii) \sum_{n=1}^{\infty} \frac{n^2}{n!}.$$

$$(viii) \sum_{n=1}^{\infty} \frac{\log n}{n}.$$

$$(ix) \sum_{n=2}^{\infty} \frac{1}{\log n}.$$

$$(x) \sum_{n=2}^{\infty} \frac{1}{(\log n)^k}.$$

$$(xi) \sum_{n=2}^{\infty} \frac{1}{(\log n)^n}.$$

$$(xii) \sum_{n=2}^{\infty} (-1)^n \frac{1}{(\log n)^n}.$$

$$(xiii) \sum_{n=1}^{\infty} \frac{n^2}{n^3 + 1}.$$

$$(xiv) \sum_{n=1}^{\infty} \sin \frac{1}{n}.$$

$$(xv) \sum_{n=2}^{\infty} \frac{1}{n \log n}.$$

$$(xvi) \sum_{n=2}^{\infty} \frac{1}{n (\log n)^2}.$$

$$(xvii) \sum_{n=2}^{\infty} \frac{1}{n^2 (\log n)}.$$

$$(xviii) \sum_{n=1}^{\infty} \frac{n!}{n^n}.$$

$$(xix) \sum_{n=1}^{\infty} \frac{2^n n!}{n^n}.$$

$$(xx) \sum_{n=1}^{\infty} \frac{3^n n!}{n^n}.$$

Hint: Use the comparison test for (i), (ii), (v), (vi), (ix), (x), (xi), (xiii), (xiv), (xvii); the ratio test for (vii), (xviii), (xix), (xx); the integral test for (viii), (xv), (xvi).

The next two problems examine, with hints, some infinite series that require more delicate analysis than those in Problem 1.

- *2. (a) If you have successfully solved examples (xix) and (xx) from Problem 1, it should be clear that $\sum_{n=1}^{\infty} a^n n! / n^n$ converges for $a < e$ and diverges for $a > e$. For $a = e$ the ratio test fails; show that $\sum_{n=1}^{\infty} e^n n! / n^n$ actually diverges, by using Problem 22-13.
- (b) Decide when $\sum_{n=1}^{\infty} n^n / a^n n!$ converges, again resorting to Problem 22-13 when the ratio test fails.
- *3. Problem 1 presented the two series $\sum_{n=2}^{\infty} (\log n)^{-k}$ and $\sum_{n=2}^{\infty} (\log n)^{-n}$, of which the first diverges while the second converges. The series

$$\sum_{n=2}^{\infty} \frac{1}{(\log n)^{\log n}},$$

which lies between these two, is analyzed in parts (a) and (b).

- (a) Show that $\int_0^{\infty} e^y / y^y dy$ exists, by considering the series $\sum_{n=1}^{\infty} (e/n)^n$.

(b) Show that

$$\sum_{n=2}^{\infty} \frac{1}{(\log n)^{\log n}}$$

converges, by using the integral test. Hint: Use an appropriate substitution and part (a).

(c) Show that

$$\sum_{n=2}^{\infty} \frac{1}{(\log n)^{\log(\log n)}}$$

diverges, by using the integral test. Hint: Use the same substitution as in part (b), and show directly that the resulting integral diverges.

4. Decide whether or not $\sum_{n=1}^{\infty} \frac{1}{n^{1+1/n}}$ converges.
5. (a) Let $\{a_n\}$ be a sequence of integers with $0 \leq a_n \leq 9$. Prove that $\sum_{n=1}^{\infty} a_n 10^{-n}$ exists (and lies between 0 and 1). (This, of course, is the number which we usually denote by $0.a_1a_2a_3a_4\dots$)
- (b) Suppose that $0 \leq x \leq 1$. Prove that there is a sequence of integers $\{a_n\}$ with $0 \leq a_n \leq 9$ and $\sum_{n=1}^{\infty} a_n 10^{-n} = x$. Hint: For example, $a_1 = [10x]$ (where $[y]$ denotes the greatest integer which is $\leq y$).
- (c) Show that if $\{a_n\}$ is repeating, i.e., is of the form $a_1, a_2, \dots, a_k, a_1, a_2, \dots, a_k, a_1, a_2, \dots$, then $\sum_{n=1}^{\infty} a_n 10^{-n}$ is a rational number (and find it). The same result naturally holds if $\{a_n\}$ is eventually repeating, i.e., if the sequence $\{a_{N+k}\}$ is repeating for some N .
- (d) Prove that if $x = \sum_{n=1}^{\infty} a_n 10^{-n}$ is rational, then $\{a_n\}$ is eventually repeating. (Just look at the process of finding the decimal expansion of p/q —dividing q into p by long division.)
6. Suppose that $\{a_n\}$ satisfies the hypothesis of Leibniz's Theorem. Use the proof of Leibniz's Theorem to obtain the following estimate:

$$\left| \sum_{n=1}^{\infty} (-1)^{n+1} a_n - [a_1 - a_2 + \dots \pm a_N] \right| < a_N.$$

7. Prove that if $a_n \geq 0$ and $\lim_{n \rightarrow \infty} \sqrt[n]{a_n} = r$, then $\sum_{n=1}^{\infty} a_n$ converges if $r < 1$, and diverges if $r > 1$. (The proof is very similar to that of the ratio test.) This result is known as the “root test.” It is easy to construct series for which the ratio test fails, while the root test works. For example, the root test shows that the series

$$\frac{1}{2} + \frac{1}{3} + (\frac{1}{2})^2 + (\frac{1}{3})^2 + (\frac{1}{2})^3 + (\frac{1}{3})^3 + \dots$$

converges, even though the ratios of successive terms do not approach a limit. Most examples are of this rather artificial nature, but the root test is nevertheless quite an important theoretical tool, and if the ratio test works the root test will also (by Problem 22-18). It is possible to eliminate limits from the root test; a simple modification of the proof shows that $\sum_{n=1}^{\infty} a_n$ converges if there is some $s < 1$ such that all but finitely many $\sqrt[n]{a_n}$ are $\leq s$, and that $\sum_{n=1}^{\infty} a_n$ diverges if infinitely many $\sqrt[n]{a_n}$ are ≥ 1 . This result is known as the

“delicate root test” (there is a similar delicate ratio test). It follows, using the notation of Problem 22-27, that $\sum_{n=1}^{\infty} a_n$ converges if $\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{a_n} < 1$ and diverges if $\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{a_n} > 1$; no conclusion is possible if $\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{a_n} = 1$.

8. For two sequences $\{a_n\}$ and $\{b_n\}$, let $c_n = \sum_{k=1}^n a_k b_{n+1-k}$. (Then c_n is the sum of the terms on the n th diagonal in the picture on page 479.) The series $\sum_{n=1}^{\infty} c_n$ is called the *Cauchy product* of $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$. If $a_n = b_n = (-1)^n/\sqrt{n}$, show that $|c_n| \geq 1$, so that the Cauchy product does not converge.

9. A sequence $\{a_n\}$ is called **Cesaro summable**, with Cesaro sum l , if

$$\lim_{n \rightarrow \infty} \frac{s_1 + \cdots + s_n}{n} = l$$

(where $s_k = a_1 + \cdots + a_k$). Problem 22-16 shows that a summable sequence is automatically Cesaro summable, with sum equal to its Cesaro sum. Find a sequence which is *not* summable, but which *is* Cesaro summable.

10. Suppose that $a_n > 0$ and $\{a_n\}$ is Cesaro summable. Suppose also that the sequence $\{na_n\}$ is bounded. Prove that the series $\sum_{n=1}^{\infty} a_n$ converges. Hint: If $s_n = \sum_{i=1}^n a_i$ and $\sigma_n = \frac{1}{n} \sum_{i=1}^n s_i$, prove that $s_n - \frac{n}{n+1} \sigma_n$ is bounded.
11. This problem outlines an alternative proof of Theorem 8 which does not rely on the Cauchy criterion.
- (a) Suppose that $a_n \geq 0$ for each n . Let $\{b_n\}$ be a rearrangement of $\{a_n\}$, and let $s_n = a_1 + \cdots + a_n$ and $t_n = b_1 + \cdots + b_n$. Show that for each n there is some m with $s_n \leq t_m$.
 - (b) Show that $\sum_{n=1}^{\infty} a_n \leq \sum_{n=1}^{\infty} b_n$ if $\sum_{n=1}^{\infty} b_n$ exists.
 - (c) Show that $\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} b_n$.
 - (d) Now replace the condition $a_n \geq 0$ by the hypothesis that $\sum_{n=1}^{\infty} a_n$ converges absolutely, using the second part of Theorem 5.

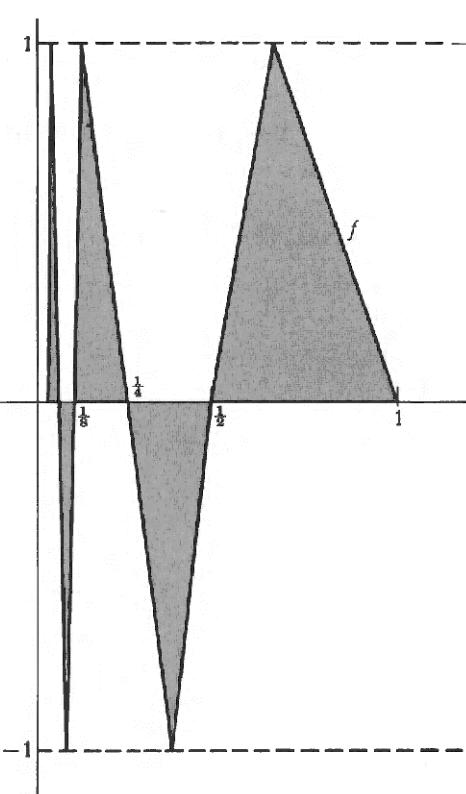


FIGURE 5

12. (a) Prove that if $\sum_{n=1}^{\infty} a_n$ converges absolutely, and $\{b_n\}$ is any subsequence of $\{a_n\}$, then $\sum_{n=1}^{\infty} b_n$ converges (absolutely).
- (b) Show that this is false if $\sum_{n=1}^{\infty} a_n$ does not converge absolutely.
- *(c) Prove that if $\sum_{n=1}^{\infty} a_n$ converges absolutely, then
- $$\sum_{n=1}^{\infty} a_n = (a_1 + a_3 + a_5 + \dots) + (a_2 + a_4 + a_6 + \dots).$$
13. Prove that if $\sum_{n=1}^{\infty} a_n$ is absolutely convergent, then $\left| \sum_{n=1}^{\infty} a_n \right| \leq \sum_{n=1}^{\infty} |a_n|$.
- *14. Problem 19-42 shows that $\int_0^\infty (\sin x)/x \, dx$ converges. Prove that $\int_0^\infty |(\sin x)/x| \, dx$ diverges.
- *15. Find a continuous function f with $f(x) \geq 0$ for all x such that $\int_0^\infty f(x) \, dx$ exists, but $\lim_{x \rightarrow \infty} f(x)$ does not exist.
- *16. Let $f(x) = x \sin 1/x$ for $0 < x \leq 1$, and let $f(0) = 0$. Recall the definition of $\ell(f, P)$ from Problem 13-25. Show that the set of all $\ell(f, P)$ for P a partition of $[0, 1]$ is not bounded (thus f has “infinite length”). Hint: Try partitions of the form

$$P = \left\{ 0, \frac{2}{(2n+1)\pi}, \dots, \frac{2}{7\pi}, \frac{2}{5\pi}, \frac{2}{3\pi}, \frac{2}{\pi}, 1 \right\}.$$

17. Let f be the function shown in Figure 5. Find $\int_0^1 f$, and also the area of the shaded region in Figure 5.
- *18. In this problem we will establish the “binomial series”

$$(1+x)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k, \quad |x| < 1,$$

for any α , by showing that $\lim_{n \rightarrow \infty} R_{n,0}(x) = 0$. The proof is in several steps, and uses the Cauchy and Lagrange forms as found in Problem 20-7.

- (a) Use the ratio test to show that the series $\sum_{k=0}^{\infty} \binom{\alpha}{k} r^k$ does indeed converge for $|r| < 1$ (this is not to say that it necessarily converges to $(1+r)^\alpha$). It follows in particular that $\lim_{n \rightarrow \infty} \binom{\alpha}{n} r^n = 0$ for $|r| < 1$.

- (b) Suppose first that $0 \leq x < 1$. Show that $\lim_{n \rightarrow \infty} R_{n,0}(x) = 0$, by using Lagrange's form of the remainder, noticing that $(1+t)^{\alpha-n-1} \leq 1$ for $n+1 > \alpha$.
- (c) Now suppose that $-1 < x < 0$; the number t in Cauchy's form of the remainder satisfies $-1 < x < t \leq 0$. Show that

$$|x(1+t)^{\alpha-1}| \leq |x|M, \quad \text{where } M = \max(1, (1+x)^{\alpha-1}),$$

and

$$\left| \frac{x-t}{1+t} \right| = |x| \left(\frac{1-t/x}{1+t} \right) \leq |x|.$$

Using Cauchy's form of the remainder, and the fact that

$$(n+1) \binom{\alpha}{n+1} = \alpha \binom{\alpha-1}{n},$$

show that $\lim_{n \rightarrow \infty} R_{n,0}(x) = 0$.

- 19.** (a) Suppose that the partial sums of the sequence $\{a_n\}$ are bounded and that $\{b_n\}$ is a sequence with $b_n \geq b_{n+1}$ and $\lim_{n \rightarrow \infty} b_n = 0$. Prove that $\sum_{n=1}^{\infty} a_n b_n$ converges. This is known as *Dirichlet's test*. Hint: Use Abel's Lemma (Problem 19-35) to check the Cauchy criterion.

- (b) Derive Leibniz's Theorem from this result.

- (c) Prove, using Problem 15-33, that the series $\sum_{n=1}^{\infty} (\cos nx)/n$ converges if x is not of the form $2k\pi$ for any integer k (in which case it clearly diverges).

- (d) Prove *Abel's test*: If $\sum_{n=1}^{\infty} a_n$ converges and $\{b_n\}$ is a sequence which is either nondecreasing or nonincreasing and which is bounded, then $\sum_{n=1}^{\infty} a_n b_n$ converges. Hint: Consider $b_n - b$, where $b = \lim_{n \rightarrow \infty} b_n$.

- *20. Suppose $\{a_n\}$ is decreasing and $\lim_{n \rightarrow \infty} a_n = 0$. Prove that if $\sum_{n=1}^{\infty} a_n$ converges, then $\sum_{n=1}^{\infty} 2^n a_{2^n}$ also converges (the "Cauchy Condensation Theorem"). Notice that the divergence of $\sum_{n=1}^{\infty} 1/n$ is a special case, for if $\sum_{n=1}^{\infty} 1/n$ converged, then $\sum_{n=1}^{\infty} 2^n (1/2^n)$ would also converge; this remark may serve as a hint.

- *21. (a) Prove that if $\sum_{n=1}^{\infty} a_n^2$ and $\sum_{n=1}^{\infty} b_n^2$ converge, then $\sum_{n=1}^{\infty} a_n b_n$ converges.
- (b) Prove that if $\sum_{n=1}^{\infty} a_n^2$ converges, then $\sum_{n=1}^{\infty} a_n/n^{\alpha}$ converges for any $\alpha > \frac{1}{2}$.
- *22. Suppose $\{a_n\}$ is decreasing and each $a_n \geq 0$. Prove that if $\sum_{n=1}^{\infty} a_n$ converges, then $\lim_{n \rightarrow \infty} n a_n = 0$. Hint: Write down the Cauchy criterion and be sure to use the fact that $\{a_n\}$ is decreasing.
- *23. If $\sum_{n=1}^{\infty} a_n$ converges, then the partial sums s_n are bounded, and $\lim_{n \rightarrow \infty} a_n = 0$. It is tempting to conjecture that boundedness of the partial sums, together with the condition $\lim_{n \rightarrow \infty} a_n = 0$, implies convergence of $\sum_{n=1}^{\infty} a_n$. This is *not* true, but finding a counterexample requires a little ingenuity. As a hint, notice that some *subsequence* of the partial sums will have to converge; you must somehow allow this to happen, without letting the sequence itself converge.
24. Prove that if $a_n \geq 0$ and $\sum_{n=1}^{\infty} a_n$ diverges, then $\sum_{n=1}^{\infty} \frac{a_n}{1+a_n}$ also diverges. Hint: Compare the partial sums. Does the converse hold?
25. Let $b_n \neq 0$. We say that the infinite product $\prod_{n=1}^{\infty} b_n$ converges if the sequence $p_n = \prod_{i=1}^n b_i$ converges, and also $\lim_{n \rightarrow \infty} p_n \neq 0$.
- (a) Prove that if $\prod_{n=1}^{\infty} (1 + a_n)$ converges, then a_n approaches 0.
- (b) Prove that $\prod_{n=1}^{\infty} (1 + a_n)$ converges if and only if $\sum_{n=1}^{\infty} \log(1 + a_n)$ converges.
- (c) For $a_n \geq 0$, prove that $\prod_{n=1}^{\infty} (1 + a_n)$ converges if and only if $\sum_{n=1}^{\infty} a_n$ converges. Hint: Use Problem 24 for one implication, and a simple estimate for $\log(1 + a)$ for the reverse implication.
26. (a) Compute $\prod_{n=2}^{\infty} \left(1 - \frac{1}{n^2}\right)$.
- (b) Compute $\prod_{n=1}^{\infty} (1 + x^{2^n})$ for $|x| < 1$.

27. The divergence of $\sum_{n=1}^{\infty} 1/n$ is related to the following remarkable fact: Any positive rational number x can be written as a *finite* sum of *distinct* numbers of the form $1/n$. The idea of the proof is shown by the following calculation for $\frac{27}{31}$: Since

$$\begin{aligned}\frac{27}{31} - \frac{1}{2} &= \frac{23}{62} \\ \frac{23}{62} - \frac{1}{3} &= \frac{7}{186} \\ \frac{7}{186} &< \frac{1}{4}, \dots, \frac{1}{26} \\ \frac{7}{186} - \frac{1}{27} &= \frac{1}{1674}\end{aligned}$$

we have.

$$\frac{27}{31} = \frac{1}{2} + \frac{1}{3} + \frac{1}{27} + \frac{1}{1674}.$$

Notice that the numerators 23, 7, 1 of the differences are decreasing.

- (a) Prove that if $1/(n+1) < x < 1/n$ for some n , then the numerator in this sort of calculation must always decrease; conclude that x can be written as a finite sum of distinct numbers $1/k$.
- (b) Now prove the result for all x by using the divergence of $\sum_{n=1}^{\infty} 1/n$.

CHAPTER 24

UNIFORM CONVERGENCE AND POWER SERIES

The considerations at the end of the previous chapter suggest an entirely new way of looking at infinite series. Our attention will shift from particular infinite sums to equations like

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \dots$$

which concern sums of quantities that depend on x . In other words, we are interested in *functions* defined by equations of the form

$$f(x) = f_1(x) + f_2(x) + f_3(x) + \dots$$

(in the previous example $f_n(x) = x^{n-1}/(n-1)!$). In such a situation $\{f_n\}$ will be some sequence of functions; for each x we obtain a sequence of numbers $\{f_n(x)\}$, and $f(x)$ is the sum of this sequence. In order to analyze such functions it will certainly be necessary to remember that each sum

$$f_1(x) + f_2(x) + f_3(x) + \dots$$

is, by definition, the limit of the sequence

$$f_1(x), \quad f_1(x) + f_2(x), \quad f_1(x) + f_2(x) + f_3(x), \quad \dots$$

If we define a new sequence of functions $\{s_n\}$ by

$$s_n = f_1 + \dots + f_n,$$

then we can express this fact more succinctly by writing

$$f(x) = \lim_{n \rightarrow \infty} s_n(x).$$

For some time we shall therefore concentrate on functions defined as limits,

$$f(x) = \lim_{n \rightarrow \infty} f_n(x),$$

rather than on functions defined as infinite sums. The total body of results about such functions can be summed up very easily: nothing one would hope to be true actually is—instead we have a splendid collection of counterexamples. The first of these shows that even if each f_n is continuous, the function f may not be! Contrary to what you may expect, the functions f_n will be very simple. Figure 1 shows the graphs of the functions

$$f_n(x) = \begin{cases} x^n, & 0 \leq x \leq 1 \\ 1, & x \geq 1. \end{cases}$$

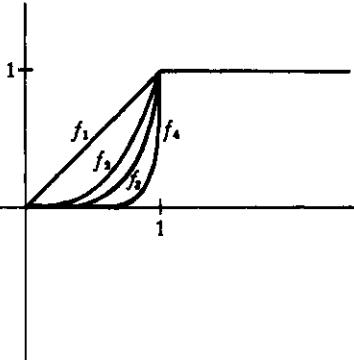


FIGURE 1

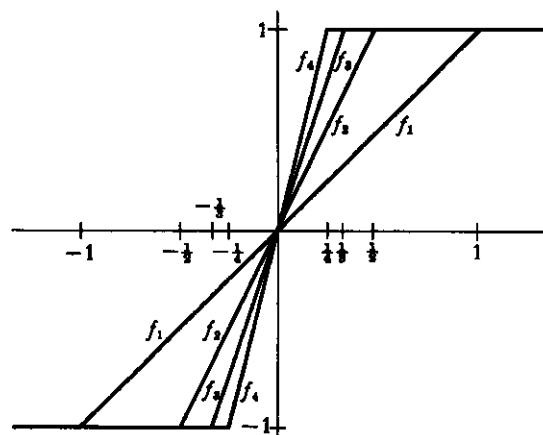


FIGURE 2

These functions are all continuous, but the function $f(x) = \lim_{n \rightarrow \infty} f_n(x)$ is not continuous; in fact,

$$\lim_{n \rightarrow \infty} f_n(x) = \begin{cases} 0, & 0 \leq x < 1 \\ 1, & x \geq 1. \end{cases}$$

Another example of this same phenomenon is illustrated in Figure 2; the functions f_n are defined by

$$f_n(x) = \begin{cases} -1, & x \leq -\frac{1}{n} \\ nx, & -\frac{1}{n} \leq x \leq \frac{1}{n} \\ 1, & \frac{1}{n} \leq x. \end{cases}$$

In this case, if $x < 0$, then $f_n(x)$ is eventually (i.e., for large enough n) equal to -1 , and if $x > 0$, then $f_n(x)$ is eventually 1 , while $f_n(0) = 0$ for all n . Thus

$$\lim_{n \rightarrow \infty} f_n(x) = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0; \end{cases}$$

so, once again, the function $f(x) = \lim_{n \rightarrow \infty} f_n(x)$ is not continuous.

By rounding off the corners in the previous examples it is even possible to produce a sequence of *differentiable* functions $\{f_n\}$ for which the function $f(x) = \lim_{n \rightarrow \infty} f_n(x)$ is not continuous. One such sequence is easy to define explicitly:

$$f_n(x) = \begin{cases} -1, & x \leq -\frac{1}{n} \\ \sin\left(\frac{n\pi x}{2}\right), & -\frac{1}{n} \leq x \leq \frac{1}{n} \\ 1, & \frac{1}{n} \leq x. \end{cases}$$

These functions are differentiable (Figure 3), but we still have

$$\lim_{n \rightarrow \infty} f_n(x) = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0. \end{cases}$$

Continuity and differentiability are, moreover, not the only properties for which problems arise. Another difficulty is illustrated by the sequence $\{f_n\}$ shown in Figure 4; on the interval $[0, 1/n]$ the graph of f_n forms an isosceles triangle of altitude n , while $f_n(x) = 0$ for $x \geq 1/n$. These functions may be defined explicitly as follows:

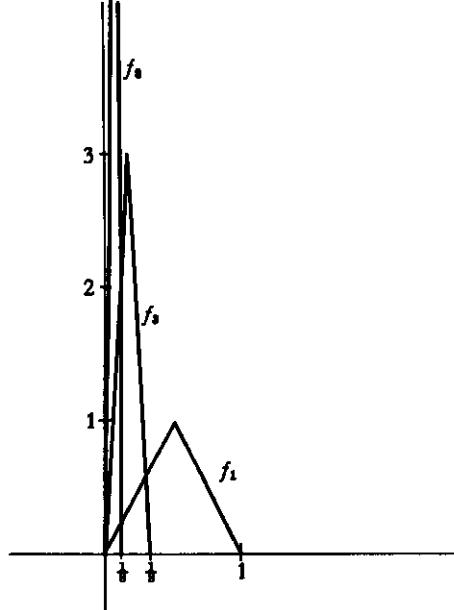


FIGURE 4

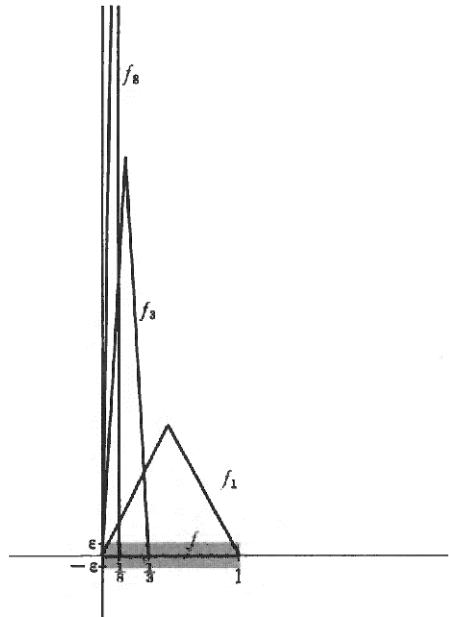


FIGURE 5

$$f_n(x) = \begin{cases} 2n^2x, & 0 \leq x \leq \frac{1}{2n} \\ 2n - 2n^2x, & \frac{1}{2n} \leq x \leq \frac{1}{n} \\ 0, & \frac{1}{n} \leq x \leq 1. \end{cases}$$

Because this sequence varies so erratically near 0, our primitive mathematical instincts might suggest that $\lim_{n \rightarrow \infty} f_n(x)$ does not always exist. Nevertheless, this limit does exist for all x , and the function $f(x) = \lim_{n \rightarrow \infty} f_n(x)$ is even continuous. In fact, if $x > 0$, then $f_n(x)$ is eventually 0, so $\lim_{n \rightarrow \infty} f_n(x) = 0$; moreover, $f_n(0) = 0$ for all n , so that we certainly have $\lim_{n \rightarrow \infty} f_n(0) = 0$. In other words, $f(x) = \lim_{n \rightarrow \infty} f_n(x) = 0$ for all x . On the other hand, the integral quickly reveals the strange behavior of this sequence; we have

$$\int_0^1 f_n(x) dx = \frac{1}{2},$$

but

$$\int_0^1 f(x) dx = 0.$$

Thus,

$$\lim_{n \rightarrow \infty} \int_0^1 f_n(x) dx \neq \int_0^1 \lim_{n \rightarrow \infty} f_n(x) dx.$$

This particular sequence of functions behaves in a way that we really never imagined when we first considered functions defined by limits. Although it is true that

$$f(x) = \lim_{n \rightarrow \infty} f_n(x) \quad \text{for each } x \text{ in } [0, 1],$$

the graphs of the functions f_n do not “approach” the graph of f in the sense of lying close to it—if, as in Figure 5, we draw a strip around f of total width 2ϵ (allowing a width of ϵ above and below), then the graphs of f_n do not lie completely within this strip, no matter how large an n we choose. Of course, for each x there is some N such that the point $(x, f_n(x))$ lies in this strip for $n > N$; this assertion just amounts to the fact that $\lim_{n \rightarrow \infty} f_n(x) = f(x)$. But it is necessary to choose larger and larger N 's as x is chosen closer and closer to 0, and no one N will work for all x at once.

The same situation actually occurs, though less blatantly, for each of the other examples given previously. Figure 6 illustrates this point for the sequence

$$f_n(x) = \begin{cases} x^n, & 0 \leq x \leq 1 \\ 1, & x \geq 1. \end{cases}$$

A strip of total width 2ϵ has been drawn around the graph of $f(x) = \lim_{n \rightarrow \infty} f_n(x)$. If $\epsilon < \frac{1}{2}$, this strip consists of two pieces, which contain no points with second

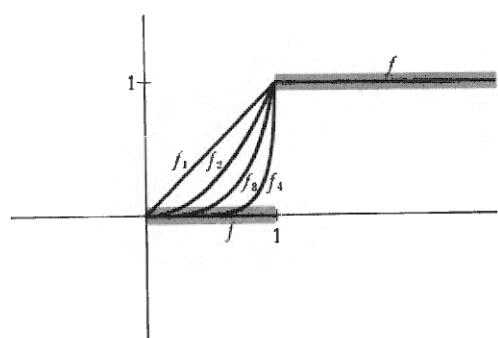


FIGURE 6

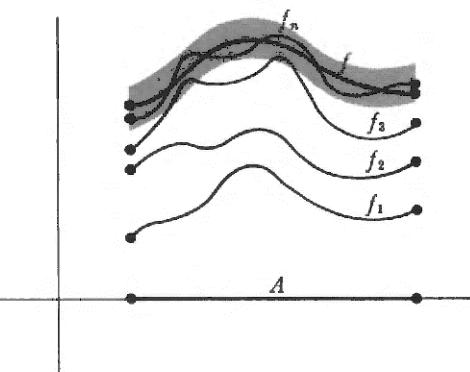


FIGURE 7

coordinate equal to $\frac{1}{2}$; since each function f_n takes on the value $\frac{1}{2}$, the graph of each f_n fails to lie within this strip. Once again, for each point x there is some N such that $(x, f_n(x))$ lies in the strip for $n > N$; but it is not possible to pick one N which works for all x at once.

It is easy to check that precisely the same situation occurs for each of the other examples. In each case we have a function f , and a sequence of functions $\{f_n\}$, all defined on some set A , such that

$$f(x) = \lim_{n \rightarrow \infty} f_n(x) \quad \text{for all } x \text{ in } A.$$

This means that

for all $\varepsilon > 0$, and for all x in A , there is some N such that if $n > N$, then $|f(x) - f_n(x)| < \varepsilon$.

But in each case different N 's must be chosen for different x 's, and in each case it is *not* true that

for all $\varepsilon > 0$ there is some N such that for all x in A , if $n > N$, then $|f(x) - f_n(x)| < \varepsilon$.

Although this condition differs from the first only by a minor displacement of the phrase “for all x in A ,” it has a totally different significance. If a sequence $\{f_n\}$ satisfies this second condition, then the graphs of f_n eventually lie close to the graph of f , as illustrated in Figure 7. This condition turns out to be just the one which makes the study of limit functions feasible.

DEFINITION

Let $\{f_n\}$ be a sequence of functions defined on A , and let f be a function which is also defined on A . Then f is called the **uniform limit of $\{f_n\}$ on A** if for every $\varepsilon > 0$ there is some N such that for all x in A ,

$$\text{if } n > N, \text{ then } |f(x) - f_n(x)| < \varepsilon.$$

We also say that $\{f_n\}$ **converges uniformly to f on A** , or that f_n **approaches f uniformly on A** .

As a contrast to this definition, if we know only that

$$f(x) = \lim_{n \rightarrow \infty} f_n(x) \quad \text{for each } x \text{ in } A,$$

then we say that $\{f_n\}$ **converges pointwise to f on A** . Clearly, uniform convergence implies pointwise convergence (but not conversely!).

Evidence for the usefulness of uniform convergence is not at all difficult to amass. Integrals represent a particularly easy topic; Figure 7 makes it almost obvious that if $\{f_n\}$ converges uniformly to f , then the integral of f_n can be made as close to the integral of f as desired. Expressed more precisely, we have the following theorem.

THEOREM 1 Suppose that $\{f_n\}$ is a sequence of functions which are integrable on $[a, b]$, and that $\{f_n\}$ converges uniformly on $[a, b]$ to a function f which is integrable on $[a, b]$. Then

$$\int_a^b f = \lim_{n \rightarrow \infty} \int_a^b f_n.$$

PROOF Let $\varepsilon > 0$. There is some N such that for all $n > N$ we have

$$|f(x) - f_n(x)| < \varepsilon \quad \text{for all } x \text{ in } [a, b].$$

Thus, if $n > N$ we have

$$\begin{aligned} \left| \int_a^b f(x) dx - \int_a^b f_n(x) dx \right| &= \left| \int_a^b [f(x) - f_n(x)] dx \right| \\ &\leq \int_a^b |f(x) - f_n(x)| dx \\ &\leq \int_a^b \varepsilon dx \\ &= \varepsilon(b - a). \end{aligned}$$

Since this is true for any $\varepsilon > 0$, it follows that

$$\int_a^b f = \lim_{n \rightarrow \infty} \int_a^b f_n. \blacksquare$$

The treatment of continuity is only a little more difficult, involving an “ $\varepsilon/3$ -argument,” a three-step estimate of $|f(x) - f(x + h)|$. If $\{f_n\}$ is a sequence of continuous functions which converges uniformly to f , then there is some n such that

$$(1) \quad |f(x) - f_n(x)| < \frac{\varepsilon}{3},$$

$$(2) \quad |f(x + h) - f_n(x + h)| < \frac{\varepsilon}{3}.$$

Moreover, since f_n is continuous, for sufficiently small h we have

$$(3) \quad |f_n(x) - f_n(x + h)| < \frac{\varepsilon}{3}.$$

It will follow from (1), (2), and (3) that $|f(x) - f(x + h)| < \varepsilon$. In order to obtain (3), however, we must restrict the size of $|h|$ in a way that cannot be predicted until n has already been chosen; it is therefore quite essential that there be some fixed n which makes (2) true, no matter how small $|h|$ may be—it is precisely at this point that uniform convergence enters the proof.

THEOREM 2 Suppose that $\{f_n\}$ is a sequence of functions which are continuous on $[a, b]$, and that $\{f_n\}$ converges uniformly on $[a, b]$ to f . Then f is also continuous on $[a, b]$.

PROOF For each x in $[a, b]$ we must prove that f is continuous at x . We will deal only with x in (a, b) ; the cases $x = a$ and $x = b$ require the usual simple modifications.

Let $\varepsilon > 0$. Since $\{f_n\}$ converges uniformly to f on $[a, b]$, there is some n such that

$$|f(y) - f_n(y)| < \frac{\varepsilon}{3} \quad \text{for all } y \text{ in } [a, b].$$

In particular, for all h such that $x + h$ is in $[a, b]$, we have

$$(1) \quad |f(x) - f_n(x)| < \frac{\varepsilon}{3},$$

$$(2) \quad |f(x + h) - f_n(x + h)| < \frac{\varepsilon}{3}.$$

Now f_n is continuous, so there is some $\delta > 0$ such that for $|h| < \delta$ we have

$$(3) \quad |f_n(x) - f_n(x + h)| < \frac{\varepsilon}{3}.$$

Thus, if $|h| < \delta$, then

$$\begin{aligned} & |f(x + h) - f(x)| \\ &= |f(x + h) - f_n(x + h) + f_n(x + h) - f_n(x) + f_n(x) - f(x)| \\ &\leq |f(x + h) - f_n(x + h)| + |f_n(x + h) - f_n(x)| + |f_n(x) - f(x)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \\ &= \varepsilon. \end{aligned}$$

This proves that f is continuous at x . ■

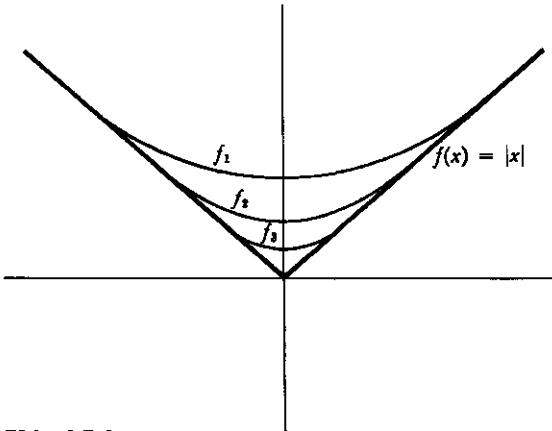


FIGURE 8

After the two noteworthy successes provided by Theorem 1 and Theorem 2, the situation for differentiability turns out to be very disappointing. If each f_n is differentiable, and if $\{f_n\}$ converges uniformly to f , it is still not necessarily true that f is differentiable. For example, Figure 8 shows that there is a sequence of differentiable functions $\{f_n\}$ which converges uniformly to the function $f(x) = |x|$.

Even if f is differentiable, it may not be true that

$$f'(x) = \lim_{n \rightarrow \infty} f_n'(x);$$

this is not at all surprising if we reflect that a smooth function can be approximated by very rapidly oscillating functions. For example (Figure 9), if

$$f_n(x) = \frac{1}{n} \sin(n^2 x),$$

then $\{f_n\}$ converges uniformly to the function $f(x) = 0$, but

$$f_n'(x) = n \cos(n^2 x),$$

and $\lim_{n \rightarrow \infty} n \cos(n^2 x)$ does not always exist (for example, it does not exist if $x = 0$).

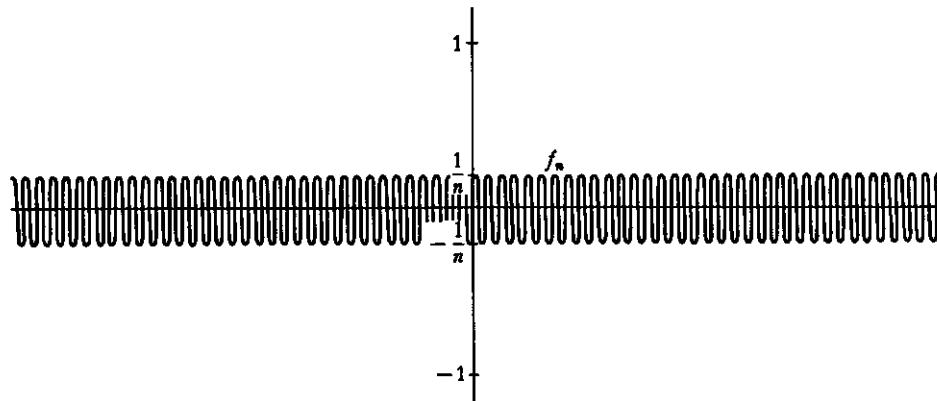


FIGURE 9

Despite such examples, the Fundamental Theorem of Calculus practically guarantees that some sort of theorem about derivatives will be a consequence of Theorem 1; the crucial hypothesis is that $\{f_n'\}$ converge uniformly (to *some* continuous function).

THEOREM 3 Suppose that $\{f_n\}$ is a sequence of functions which are differentiable on $[a, b]$, with integrable derivatives f_n' , and that $\{f_n\}$ converges (pointwise) to f . Suppose, moreover, that $\{f_n'\}$ converges uniformly on $[a, b]$ to some continuous function g . Then f is differentiable and

$$f'(x) = \lim_{n \rightarrow \infty} f_n'(x).$$

PROOF Applying Theorem 1 to the interval $[a, x]$, we see that for each x we have

$$\begin{aligned} \int_a^x g &= \lim_{n \rightarrow \infty} \int_a^x f_n' \\ &= \lim_{n \rightarrow \infty} [f_n(x) - f_n(a)] \\ &= f(x) - f(a). \end{aligned}$$

Since g is continuous, it follows that $f'(x) = g(x) = \lim_{n \rightarrow \infty} f_n'(x)$ for all x in the interval $[a, b]$. ■

Now that the basic facts about uniform limits have been established, it is clear how to treat functions defined as infinite sums,

$$f(x) = f_1(x) + f_2(x) + f_3(x) + \dots$$

This equation means that

$$f(x) = \lim_{n \rightarrow \infty} f_1(x) + \dots + f_n(x);$$

our previous theorems apply when the new sequence

$$f_1, f_1 + f_2, f_1 + f_2 + f_3, \dots$$

converges uniformly to f . Since this is the only case we shall ever be interested in, we single it out with a definition.

DEFINITION

The series $\sum_{n=1}^{\infty} f_n$ **converges uniformly** (more formally: the sequence $\{f_n\}$ is **uniformly summable**) to f on A , if the sequence

$$f_1, f_1 + f_2, f_1 + f_2 + f_3, \dots$$

converges uniformly to f on A .

We can now apply each of Theorems 1, 2, and 3 to uniformly convergent series; the results may be stated in one common corollary.

COROLLARY

Let $\sum_{n=1}^{\infty} f_n$ converge uniformly to f on $[a, b]$.

- (1) If each f_n is continuous on $[a, b]$, then f is continuous on $[a, b]$.
- (2) If f and each f_n is integrable on $[a, b]$, then

$$\int_a^b f = \sum_{n=1}^{\infty} \int_a^b f_n.$$

Moreover, if $\sum_{n=1}^{\infty} f_n$ converges (pointwise) to f on $[a, b]$, each f_n has an integrable derivative f_n' and $\sum_{n=1}^{\infty} f_n'$ converges uniformly on $[a, b]$ to some continuous function, then

$$(3) \quad f'(x) = \sum_{n=1}^{\infty} f_n'(x) \quad \text{for all } x \text{ in } [a, b].$$

PROOF

(1) If each f_n is continuous, then so is each $f_1 + \dots + f_n$, and f is the uniform limit of the sequence $f_1, f_1 + f_2, f_1 + f_2 + f_3, \dots$, so f is continuous by Theorem 2.

(2) Since $f_1, f_1 + f_2, f_1 + f_2 + f_3, \dots$ converges uniformly to f , it follows from Theorem 1 that

$$\begin{aligned}\int_a^b f &= \lim_{n \rightarrow \infty} \int_a^b (f_1 + \dots + f_n) \\ &= \lim_{n \rightarrow \infty} \left(\int_a^b f_1 + \dots + \int_a^b f_n \right) \\ &= \sum_{n=1}^{\infty} \int_a^b f_n.\end{aligned}$$

(3) Each function $f_1 + \dots + f_n$ is differentiable, with derivative $f_1' + \dots + f_n'$, and $f_1', f_1' + f_2', f_1' + f_2' + f_3', \dots$ converges uniformly to a continuous function, by hypothesis. It follows from Theorem 3 that

$$\begin{aligned}f'(x) &= \lim_{n \rightarrow \infty} [f_1'(x) + \dots + f_n'(x)] \\ &= \sum_{n=1}^{\infty} f_n'(x).\blacksquare\end{aligned}$$

At the moment this corollary is not very useful, since it seems quite difficult to predict when the sequence $f_1, f_1 + f_2, f_1 + f_2 + f_3, \dots$ will converge uniformly. The most important condition which ensures such uniform convergence is provided by the following theorem; the proof is almost a triviality because of the cleverness with which the very simple hypotheses have been chosen.

**THEOREM 4
(THE WEIERSTRASS M-TEST)**

Let $\{f_n\}$ be a sequence of functions defined on A , and suppose that $\{M_n\}$ is a sequence of numbers such that

$$|f_n(x)| \leq M_n \quad \text{for all } x \text{ in } A.$$

Suppose moreover that $\sum_{n=1}^{\infty} M_n$ converges. Then for each x in A the series $\sum_{n=1}^{\infty} f_n(x)$ converges (in fact, it converges absolutely), and $\sum_{n=1}^{\infty} f_n$ converges uniformly on A to the function

$$f(x) = \sum_{n=1}^{\infty} f_n(x).$$

PROOF For each x in A the series $\sum_{n=1}^{\infty} |f_n(x)|$ converges, by the comparison test; consequently $\sum_{n=1}^{\infty} f_n(x)$ converges (absolutely). Moreover, for all x in A we have ~

$$\begin{aligned}|f(x) - [f_1(x) + \cdots + f_n(x)]| &= \left| \sum_{n=N+1}^{\infty} f_n(x) \right| \\ &\leq \sum_{n=N+1}^{\infty} |f_n(x)| \\ &\leq \sum_{n=N+1}^{\infty} M_n.\end{aligned}$$

Since $\sum_{n=1}^{\infty} M_n$ converges, the number $\sum_{n=N+1}^{\infty} M_n$ can be made as small as desired, by choosing N sufficiently large. ■

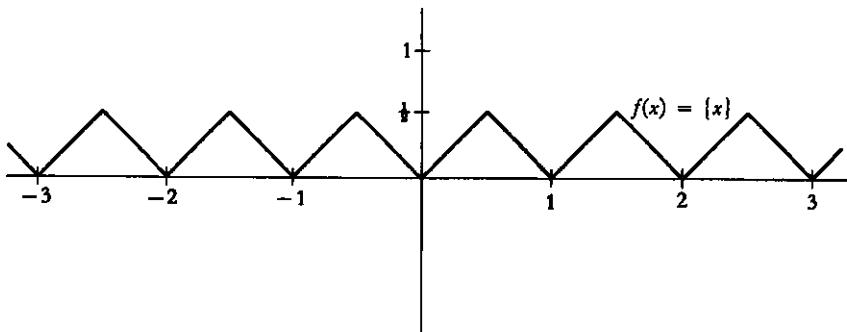


FIGURE 10

The following sequence $\{f_n\}$ illustrates a simple application of the Weierstrass M -test. Let $\{x\}$ denote the distance from x to the nearest integer (the graph of $f(x) = \{x\}$ is illustrated in Figure 10). Now define

$$f_n(x) = \frac{1}{10^n} \{10^n x\}.$$

The functions f_1 and f_2 are shown in Figure 11 (but to make the drawings simpler, 10^n has been replaced by 2^n). This sequence of functions has been defined so that the Weierstrass M -test automatically applies: clearly

$$|f_n(x)| \leq \frac{1}{10^n} \quad \text{for all } x,$$

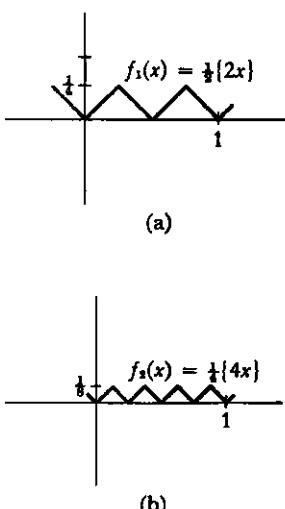


FIGURE 11

and $\sum_{n=1}^{\infty} 1/10^n$ converges. Thus $\sum_{n=1}^{\infty} f_n$ converges uniformly; since each f_n is continuous, the corollary implies that the function

$$f(x) = \sum_{n=1}^{\infty} f_n(x) = \sum_{n=1}^{\infty} \frac{1}{10^n} \{10^n x\}$$

is also continuous. Figure 12 shows the graph of the first few partial sums $f_1 + \dots + f_n$. As n increases, the graphs become harder and harder to draw, and the infinite sum $\sum_{n=1}^{\infty} f_n$ is quite undrawable, as shown by the following theorem (included mainly as an interesting sidelight, to be skipped if you find the going too rough).

THEOREM 5 The function

$$f(x) = \sum_{n=1}^{\infty} \frac{1}{10^n} \{10^n x\}$$

is continuous everywhere and differentiable nowhere!

PROOF We have just shown that f is continuous; this is the only part of the proof which uses uniform convergence. We will prove that f is not differentiable at a , for any a , by the straightforward method of exhibiting a particular sequence $\{h_m\}$ approaching 0 for which

$$\lim_{m \rightarrow \infty} \frac{f(a + h_m) - f(a)}{h_m}$$

does not exist. It obviously suffices to consider only those numbers a satisfying $0 < a \leq 1$.

Suppose that the decimal expansion of a is

$$a = 0.a_1 a_2 a_3 a_4 \dots$$

Let $h_m = 10^{-m}$ if $a_m \neq 4$ or 9, but let $h_m = -10^{-m}$ if $a_m = 4$ or 9 (the reason for these two exceptions will appear soon). Then

$$\begin{aligned} \frac{f(a + h_m) - f(a)}{h_m} &= \sum_{n=1}^{\infty} \frac{1}{10^n} \cdot \frac{\{10^n(a + h_m)\} - \{10^n a\}}{\pm 10^{-m}} \\ &= \sum_{n=1}^{\infty} \pm 10^{m-n} [\{10^n(a + h_m)\} - \{10^n a\}]. \end{aligned}$$

This infinite series is really a finite sum, because if $n \geq m$, then $10^n h_m$ is an integer, so

$$\{10^n(a + h_m)\} - \{10^n a\} = 0.$$

On the other hand, for $n < m$ we can write

$$\begin{aligned} 10^n a &= \text{integer} + 0.a_{n+1} a_{n+2} a_{n+3} \dots a_m \dots \\ 10^n(a + h_m) &= \text{integer} + 0.a_{n+1} a_{n+2} a_{n+3} \dots (a_m \pm 1) \dots \end{aligned}$$

(in order for the second equation to be true it is essential that we choose $h_m = -10^{-m}$ when $a_m = 9$). Now suppose that

$$0.a_{n+1}a_{n+2}a_{n+3}\dots a_m \dots \leq \frac{1}{2}.$$

Then we also have

$$0.a_{n+1}a_{n+2}a_{n+3}\dots (a_m \pm 1)\dots \leq \frac{1}{2}$$

(in the special case $m = n + 1$ the second equation is true because we chose $h_m = -10^{-m}$ when $a_m = 4$). This means that

$$\{10^n(a + h_m)\} - \{10^n a\} = \pm 10^{n-m},$$

and exactly the same equation can be derived when $0.a_{n+1}a_{n+2}a_{n+3}\dots > \frac{1}{2}$. Thus, for $n < m$ we have

$$10^{m-n}[\{10^n(a + h_m)\} - \{10^n a\}] = \pm 1.$$

In other words,

$$\frac{f(a + h_m) - f(a)}{h_m}$$

is the sum of $m - 1$ numbers, each of which is ± 1 . Now adding $+1$ or -1 to a number changes it from odd to even, and vice versa. The sum of $m - 1$ numbers each ± 1 is therefore an *even integer* if m is odd, and an *odd integer* if m is even. Consequently the sequence of ratios

$$\frac{f(a + h_m) - f(a)}{h_m}$$

cannot possibly converge, since it is a sequence of integers which are alternately odd and even. ■

In addition to its role in the previous theorem, the Weierstrass M -test is an ideal tool for analyzing functions which are very well behaved. We will give special attention to functions of the form

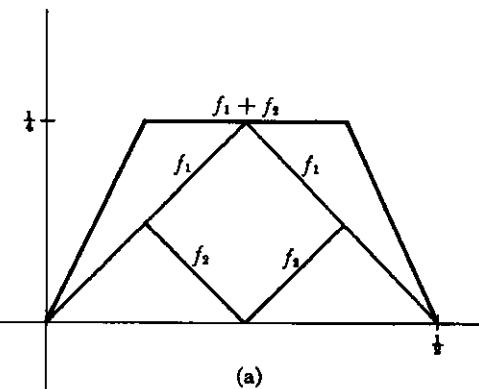
$$f(x) = \sum_{n=0}^{\infty} a_n(x - a)^n,$$

which can also be described by the equation

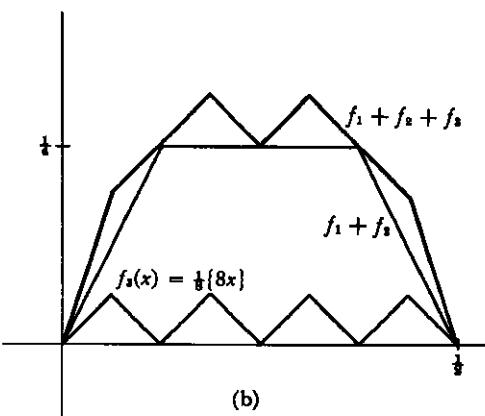
$$f(x) = \sum_{n=0}^{\infty} f_n(x),$$

for $f_n(x) = a_n(x - a)^n$. Such an infinite sum, of functions which depend only on powers of $(x - a)$, is called a **power series centered at a** . For the sake of simplicity, we will usually concentrate on power series centered at 0,

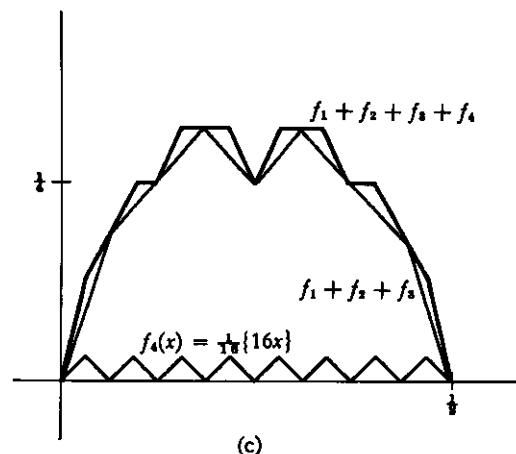
$$f(x) = \sum_{n=0}^{\infty} a_n x^n.$$



(a)



(b)



(c)

FIGURE 12

One especially important group of power series are those of the form

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^n,$$

where f is some function which has derivatives of all orders at a ; this series is called the **Taylor series for f at a** . Of course, it is not necessarily true that

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^n;$$

this equation holds only when the remainder terms satisfy $\lim_{n \rightarrow \infty} R_{n,a}(x) = 0$.

We already know that a power series $\sum_{n=0}^{\infty} a_n x^n$ does not necessarily converge for all x . For example, the power series

$$x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots$$

converges only for $|x| \leq 1$, while the power series

$$x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} + \dots$$

converges only for $-1 < x \leq 1$. It is even possible to produce a power series which converges only for $x = 0$. For example, the power series

$$\sum_{n=0}^{\infty} n! x^n$$

does not converge for $x \neq 0$; indeed, the ratios

$$\frac{(n+1)! (x^{n+1})}{n! x^n} = (n+1)x$$

are unbounded for any $x \neq 0$. If a power series $\sum_{n=0}^{\infty} a_n x^n$ does converge for some $x_0 \neq 0$ however, then a great deal can be said about the series $\sum_{n=0}^{\infty} a_n x^n$ for $|x| < |x_0|$.

THEOREM 6 Suppose that the series

$$f(x_0) = \sum_{n=0}^{\infty} a_n x_0^n$$

converges, and let a be any number with $0 < a < |x_0|$. Then on $[-a, a]$ the series

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

converges uniformly (and absolutely). Moreover, the same is true for the series

$$g(x) = \sum_{n=1}^{\infty} n a_n x^{n-1}.$$

Finally, f is differentiable and

$$f'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1}$$

for all x with $|x| < |x_0|$.

PROOF Since $\sum_{n=0}^{\infty} a_n x_0^n$ converges, the terms $a_n x_0^n$ approach 0. Hence they are surely bounded: there is some number M such that

$$|a_n x_0^n| = |a_n| \cdot |x_0^n| \leq M \quad \text{for all } n.$$

Now if x is in $[-a, a]$, then $|x| \leq |a|$, so

$$\begin{aligned} |a_n x^n| &= |a_n| \cdot |x^n| \\ &\leq |a_n| \cdot |a^n| \\ &= |a_n| \cdot |x_0|^n \cdot \left| \frac{a}{x_0} \right|^n \quad (\text{this is the clever step}) \\ &\leq M \left| \frac{a}{x_0} \right|^n. \end{aligned}$$

But $|a/x_0| < 1$, so the (geometric) series

$$\sum_{n=0}^{\infty} M \left| \frac{a}{x_0} \right|^n = M \sum_{n=0}^{\infty} \left| \frac{a}{x_0} \right|^n$$

converges. Choosing $M \cdot |a/x_0|^n$ as the number M_n in the Weierstrass M -test, it follows that $\sum_{n=0}^{\infty} a_n x^n$ converges uniformly on $[-a, a]$.

To prove the same assertion for $g(x) = \sum_{n=1}^{\infty} n a_n x^{n-1}$ notice that

$$\begin{aligned} |n a_n x^{n-1}| &= n |a_n| \cdot |x^{n-1}| \\ &\leq n |a_n| \cdot |a^{n-1}| \\ &= \frac{|a_n|}{|a|} \cdot |x_0|^n n \left| \frac{a}{x_0} \right|^n \\ &\leq \frac{M}{|a|} n \left| \frac{a}{x_0} \right|^n. \end{aligned}$$

Since $|a/x_0| < 1$, the series

$$\sum_{n=1}^{\infty} \frac{M}{|a|} n \left| \frac{a}{x_0} \right|^n = \frac{M}{|a|} \sum_{n=1}^{\infty} n \left| \frac{a}{x_0} \right|^n$$

converges (this fact was proved in Chapter 23 as an application of the ratio test).

Another appeal to the Weierstrass M -test proves that $\sum_{n=1}^{\infty} na_n x^{n-1}$ converges uniformly on $[-a, a]$.

Finally, our corollary proves, first that g is continuous, and then that

$$f'(x) = g(x) = \sum_{n=1}^{\infty} na_n x^{n-1} \quad \text{for } x \text{ in } [-a, a].$$

Since we could have chosen any number a with $0 < a < |x_0|$, this result holds for all x with $|x| < |x_0|$. ■

We are now in a position to manipulate power series with ease. Most algebraic manipulations are fairly straightforward consequences of general theorems about infinite series. For example, suppose that $f(x) = \sum_{n=0}^{\infty} a_n x^n$ and $g(x) = \sum_{n=0}^{\infty} b_n x^n$, where the two power series both converge for some x_0 . Then for $|x| < |x_0|$ we have

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

So the series $h(x) = \sum_{n=0}^{\infty} (a_n + b_n) x^n$ also converges for $|x| < |x_0|$, and $h = f + g$ for these x .

The treatment of products is just a little more involved. If $|x| < |x_0|$, then we know that the series $\sum_{n=0}^{\infty} a_n x^n$ and $\sum_{n=0}^{\infty} b_n x^n$ converge *absolutely*. So it follows from

Theorem 23-9 that the product $\sum_{n=0}^{\infty} a_n x^n \cdot \sum_{n=0}^{\infty} b_n x^n$ is given by

$$\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_i x^i b_j x^j,$$

where the elements $a_i x^i b_j x^j$ are arranged in any order. In particular, we can choose the arrangement

$$a_0 b_0 + (a_0 b_1 + a_1 b_0) x + (a_0 b_2 + a_1 b_1 + a_2 b_0) x^2 + \dots$$

which can be written as

$$\sum_{n=0}^{\infty} c_n x^n \quad \text{for } c_n = \sum_{k=0}^n a_k b_{n-k}.$$

This is the “Cauchy product” that was introduced in Problem 23-8. Thus, the Cauchy product $h(x) = \sum_{n=0}^{\infty} c_n x^n$ also converges for $|x| < |x_0|$ and $h = fg$ for these x .

Finally, suppose that $f(x) = \sum_{n=0}^{\infty} a_n x^n$, where $a_0 \neq 0$, so that $f(0) = a_0 \neq 0$.

Then we can try to find a power series $\sum_{n=0}^{\infty} b_n x^n$ which represents $1/f$. This means that we want to have

$$\sum_{n=0}^{\infty} a_n x^n \cdot \sum_{n=0}^{\infty} b_n x^n = 1 = 1 + 0 \cdot x + 0 \cdot x^2 + \dots$$

Since the left side of this equation will be given by the Cauchy product, we want to have

$$\begin{aligned} a_0 b_0 &= 1 \\ a_0 b_1 + a_1 b_0 &= 0 \\ a_0 b_2 + a_1 b_1 + a_2 b_0 &= 0 \\ &\dots \end{aligned}$$

Since $a_0 \neq 0$, we can solve the first of these equations for b_0 . Then we can solve the second for b_1 , etc. Of course, we still have to prove that the new series $\sum_{n=0}^{\infty} b_n x^n$ does converge for some $x \neq 0$. This is left as an exercise (Problem 17).

For derivatives, Theorem 6 gives us all the information we need. In particular, when we apply Theorem 6 to the infinite series

$$\begin{aligned} \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \dots, \\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \dots, \\ e^x &= 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots, \end{aligned}$$

we get precisely the results which are expected. Each of these converges for any x_0 , hence the conclusions of Theorem 6 apply for any x :

$$\begin{aligned} \sin'(x) &= 1 - \frac{3x^2}{3!} + \frac{5x^4}{5!} - \dots = \cos x, \\ \cos'(x) &= -\frac{2x}{2!} + \frac{4x^3}{4!} - \frac{6x^5}{6!} + \dots = -\sin x, \\ \exp'(x) &= 1 + \frac{2x}{2!} + \frac{3x^2}{3!} + \dots = \exp(x). \end{aligned}$$

For the functions \arctan and $f(x) = \log(1+x)$ the situation is only slightly more complicated. Since the series

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots$$

converges for $x_0 = 1$, it also converges for $|x| < 1$, and

$$\arctan'(x) = 1 - x^2 + x^4 - x^6 + \dots = \frac{1}{1+x^2} \quad \text{for } |x| < 1.$$

In this case, the series happens to converge for $x = -1$ also. However, the formula for the derivative is not correct for $x = 1$ or $x = -1$; indeed the series

$$1 - x^2 + x^4 - x^6 + \dots$$

diverges for $x = 1$ and $x = -1$. Notice that this does not contradict Theorem 6, which proves that the derivative is given by the expected formula only for $|x| < |x_0|$.

Since the series

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \dots$$

converges for $x_0 = 1$, it also converges for $|x| < 1$, and

$$\frac{1}{1+x} = \log'(1+x) = 1 - x + x^2 - x^3 + \dots \quad \text{for } |x| < 1.$$

In this case, the original series does not converge for $x = -1$; moreover, the differentiated series does not converge for $x = 1$.

All the considerations which apply to a power series will automatically apply to its derivative, at the points where the derivative is represented by a power series. If

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

converges for all x in some interval $(-R, R)$, then Theorem 6 implies that

$$f'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1}$$

for all x in $(-R, R)$. Applying Theorem 6 once again we find that

$$f''(x) = \sum_{n=2}^{\infty} n(n-1) a_n x^{n-2},$$

and proceeding by induction we find that

$$f^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1) \cdot \dots \cdot (n-k+1) a_n x^{n-k}.$$

Thus, a function defined by a power series which converges in some interval $(-R, R)$ is automatically infinitely differentiable in that interval. Moreover, the previous equation implies that

$$f^{(k)}(0) = k! a_k,$$

so that

$$a_k = \frac{f^{(k)}(0)}{k!}.$$

In other words, *a convergent power series centered at 0 is always the Taylor series at 0 of the function which it defines.*

On this happy note we could easily end our study of power series and Taylor series. A careful assessment of our situation will reveal some unexplained facts, however.

The Taylor series of \sin , \cos , and \exp are as satisfactory as we could desire; they converge for all x , and can be differentiated term-by-term for all x . The Taylor series of the function $f(x) = \log(1 + x)$ is slightly less pleasing, because it converges only for $-1 < x \leq 1$, but this deficiency is a necessary consequence of the basic nature of power series. If the Taylor series for f converged for any x_0 with $|x_0| > 1$, then it would converge on the interval $(-|x_0|, |x_0|)$, and on this interval the function which it defines would be differentiable, and thus continuous. But this is impossible, since it is unbounded on the interval $(-1, 1)$, where it equals $\log(1 + x)$.

The Taylor series for \arctan is more difficult to comprehend—there seems to be no possible excuse for the refusal of this series to converge when $|x| > 1$. This mysterious behavior is exemplified even more strikingly by the function $f(x) = 1/(1 + x^2)$, an infinitely differentiable function which is the next best thing to a polynomial function. The Taylor series of f is given by

$$f(x) = \frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + x^8 - \dots$$

If $|x| \geq 1$ the Taylor series does not converge at all. Why? What unseen obstacle prevents the Taylor series from extending past 1 and -1 ? Asking this sort of question is always dangerous, since we may have to settle for an unsympathetic answer: it happens because it happens—that's the way things are! In this case there does happen to be an explanation, but this explanation is impossible to give at the present time; although the question is about real numbers, it can be answered intelligently only when placed in a broader context. It will therefore be necessary to devote two chapters to quite new material before completing our discussion of Taylor series in Chapter 27.

PROBLEMS

1. For each of the following sequences $\{f_n\}$, determine the pointwise limit of $\{f_n\}$ (if it exists) on the indicated interval, and decide whether $\{f_n\}$ converges uniformly to this function.

(i) $f_n(x) = \sqrt[n]{x}$, on $[0, 1]$.

(ii) $f_n(x) = \begin{cases} 0, & x \leq n \\ x - n, & x \geq n, \end{cases}$ on $[a, b]$, and on \mathbf{R} .

(iii) $f_n(x) = \frac{e^x}{x^n}$, on $(1, \infty)$.

(iv) $f_n(x) = e^{-nx^2}$, on $[-1, 1]$.

(v) $f_n(x) = \frac{e^{-x^2}}{n}$, on \mathbf{R} .

2. This problem asks for the same information as in Problem 1, but the functions are not so easy to analyze. Some hints are given at the end.

(i) $f_n(x) = x^n - x^{2n}$ on $[0, 1]$.

(ii) $f_n(x) = \frac{nx}{1+n+x}$ on $[0, \infty)$.

(iii) $f_n(x) = \sqrt{x^2 + \frac{1}{n^2}}$ on $[a, \infty)$, $a > 0$.

(iv) $f_n(x) = \sqrt{x^2 + \frac{1}{n^2}}$ on \mathbf{R} .

(v) $f_n(x) = \sqrt{x + \frac{1}{n}} - \sqrt{x}$ on $[a, \infty)$, $a > 0$.

(vi) $f_n(x) = \sqrt{x + \frac{1}{n}} - \sqrt{x}$ on \mathbf{R} .

(vii) $f_n(x) = n \left(\sqrt{x + \frac{1}{n}} - \sqrt{x} \right)$ on $[a, \infty)$, $a > 0$.

(viii) $f_n(x) = n \left(\sqrt{x + \frac{1}{n}} - \sqrt{x} \right)$ on $[0, \infty)$.

Hints: (i) For each n , find the maximum of $|f - f_n|$ on $[0, 1]$. (ii) For each n , consider $|f(x) - f_n(x)|$ for x large. (iii) Express $f(x) - f_n(x)$ as a fraction and estimate $|f(x) - f_n(x)|$ for $x \geq a$. (iv) Give a separate estimate of $|f(x) - f_n(x)|$ for small $|x|$. (vii) Use (v).

3. Find the Taylor series at 0 for each of the following functions.

(i) $f(x) = \frac{1}{x-a}$, $a \neq 0$.

(ii) $f(x) = \log(x-a)$, $a \neq 0$.

(iii) $f(x) = \frac{1}{\sqrt{1-x}} = (1-x)^{-1/2}$. (Use Problem 20-7.)

(iv) $f(x) = \frac{1}{\sqrt{1-x^2}}$.

(v) $f(x) = \arcsin x$.

4. Find each of the following infinite sums.

(i) $1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} - \dots$

(ii) $1 - x^3 + x^6 - x^9 + \dots$. Hint: What is $1 - x + x^2 - x^3 + \dots$?

(iii) $\frac{x^2}{2} - \frac{x^3}{3 \cdot 2} + \frac{x^4}{4 \cdot 3} - \frac{x^5}{5 \cdot 4} + \dots$ for $|x| < 1$. Hint: Differentiate.

5. Evaluate the following infinite sums. (In most cases they are $f(a)$ where a is some obvious number and $f(x)$ is given by some power series. To evaluate the various power series, manipulate them until some well-known power series emerge.)

$$(i) \quad \sum_{n=0}^{\infty} \frac{(-1)^n 2^{2n} \pi^{2n}}{(2n)!}.$$

$$(ii) \quad \sum_{n=0}^{\infty} \frac{1}{(2n)!}.$$

$$(iii) \quad \sum_{n=0}^{\infty} \frac{1}{2n+1} \left(\frac{1}{2}\right)^{2n+1}$$

$$(iv) \quad \sum_{n=0}^{\infty} \frac{n}{2^n}.$$

$$(v) \quad \sum_{n=0}^{\infty} \frac{1}{3^n(n+1)}.$$

$$(vi) \quad \sum_{n=0}^{\infty} \frac{2n+1}{2^n n!}.$$

6. If $f(x) = (\sin x)/x$ for $x \neq 0$ and $f(0) = 1$, find $f^{(k)}(0)$. Hint: Find the power series for f .

7. In this problem we deduce the binomial series $(1+x)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n$, $|x| < 1$ without all the work of Problem 23-18, although we will use a fact established in part (a) of that problem—the series $f(x) = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n$ does converge for $|x| < 1$.

(a) Prove that $(1+x)f'(x) = \alpha f(x)$ for $|x| < 1$.

(b) Now show that any function f satisfying part (a) is of the form $f(x) = c(1+x)^\alpha$ for some constant c , and use this fact to establish the binomial series. Hint: Consider $g(x) = f(x)/(1+x)^\alpha$.

8. Prove that the series

$$\sum_{n=1}^{\infty} \frac{x}{n(1+nx^2)}$$

converges uniformly on \mathbf{R} .

9. (a) Prove that the series

$$\sum_{n=0}^{\infty} 2^n \sin \frac{1}{3^n x}$$

converges uniformly on $[a, \infty)$ for $a > 0$. Hint: $\lim_{h \rightarrow 0} (\sin h)/h = 1$.

- (b) By considering the sum from N to ∞ for $x = 2/(\pi 3^N)$, show that the series does not converge uniformly on $(0, \infty)$.
10. (a) Prove that the series
- $$f(x) = \sum_{n=0}^{\infty} \frac{nx}{1+n^4x^2}$$
- converges uniformly on $[a, \infty)$ for $a > 0$. Hint: First find the maximum of $nx/(1+n^4x^2)$ on $[0, \infty)$.
- (b) Show that
- $$f\left(\frac{1}{N}\right) \geq \frac{N}{2} \sum_{n \geq \sqrt{N}} \frac{1}{n^3},$$
- and by using an integral to estimate the sum, show that $f\left(\frac{1}{N}\right) \geq 1/4$. Conclude that the series does not converge uniformly on \mathbf{R} .
- (c) What about the series
- $$\sum_{n=0}^{\infty} \frac{nx}{1+n^5x^2}?$$
11. (a) Use Problem 15-33 and the method of proof used for Dirichlet's test (Problem 23-19) to obtain a uniform Cauchy condition for the series
- $$\sum_{n=1}^{\infty} \frac{\sin nx}{n}$$
- uniformly on $[\varepsilon, 2\pi - \varepsilon]$, $\varepsilon > 0$, and conclude that the series converges uniformly there.
- (b) For $x = \pi/N$, with N large, show that
- $$\left| \sum_{k=N}^{2N} \sin kx \right| = \left| \sum_{k=0}^N \sin kx \right| \geq \frac{N}{\pi}.$$
- Conclude that
- $$\left| \sum_{k=N}^{2N} \frac{\sin kx}{k} \right| \geq \frac{1}{2\pi},$$
- and that the series does not converge uniformly on $[0, 2\pi]$.
12. (a) Suppose that $f(x) = \sum_{n=0}^{\infty} a_n x^n$ converges for all x in some interval $(-R, R)$ and that $f(x) = 0$ for all x in $(-R, R)$. Prove that each $a_n = 0$. (If you remember the formula for a_n this is easy.)
- (b) Suppose we know only that $f(x_n) = 0$ for some sequence $\{x_n\}$ with $\lim_{n \rightarrow \infty} x_n = 0$. Prove again that each $a_n = 0$. Hint: First show that $f(0) = a_0 = 0$; then that $f'(0) = a_1 = 0$, etc.

This result shows that if $f(x) = e^{-1/x^2} \sin 1/x$ for $x \neq 0$, then f cannot possibly be written as a power series. It also shows that a function defined by a power series cannot be 0 for $x \leq 0$ but nonzero for $x > 0$ —thus a power series cannot describe the motion of a particle which has remained at rest until time 0, and then begins to move!

- (c) Suppose that $f(x) = \sum_{n=0}^{\infty} a_n x^n$ and $g(x) = \sum_{n=0}^{\infty} b_n x^n$ converge for all x in some interval containing 0 and that $f(t_m) = g(t_m)$ for some sequence $\{t_m\}$ converging to 0. Show that $a_n = b_n$ for each n . In particular, *a function can have only one representation as a power series centered at 0.*

13. Prove that if $f(x) = \sum_{n=0}^{\infty} a_n x^n$ is an even function, then $a_n = 0$ for n odd, and if f is an odd function, then $a_n = 0$ for n even.
14. Show that the power series for $f(x) = \log(1 - x)$ converges only for $-1 \leq x < 1$, and that the power series for $g(x) = \log[(1 + x)/(1 - x)]$ converges only for x in $(-1, 1)$.
- *15. Recall that the Fibonacci sequence $\{a_n\}$ is defined by $a_1 = a_2 = 1$, $a_{n+1} = a_n + a_{n-1}$.
- Show that $a_{n+1}/a_n \leq 2$.
 - Let

$$f(x) = \sum_{n=1}^{\infty} a_n x^{n-1} = 1 + x + 2x^2 + 3x^3 + \dots$$

Use the ratio test to prove that $f(x)$ converges if $|x| < 1/2$.

- (c) Prove that if $|x| < 1/2$, then

$$f(x) = \frac{-1}{x^2 + x - 1}.$$

Hint: This equation can be written $f(x) - xf(x) - x^2 f(x) = 1$.

- (d) Use the partial fraction decomposition for $1/(x^2 + x - 1)$, and the power series for $1/(x - a)$, to obtain another power series for f .
- (e) It follows from Problem 12 that the two power series obtained for f must be the same. Use this fact to show that

$$a_n = \frac{\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n}{\sqrt{5}}.$$

16. Let $f(x) = \sum_{n=0}^{\infty} a_n x^n$ and $g(x) = \sum_{n=0}^{\infty} b_n x^n$. Suppose we merely knew that $f(x)g(x) = \sum_{n=0}^{\infty} c_n x^n$ for some c_n , but we didn't know how to multiply series

in general. Use Leibniz's formula (Problem 10-18) to show directly that this series for fg must indeed be the Cauchy product of the series for f and g .

17. Suppose that $f(x) = \sum_{n=0}^{\infty} a_n x^n$ converges for some x_0 , and that $a_0 \neq 0$; for simplicity, we'll assume that $a_0 = 1$. Let $\{b_n\}$ be the sequence defined recursively by

$$\begin{aligned} b_0 &= 1 \\ b_n &= - \sum_{k=0}^{n-1} b_k a_{n-k}. \end{aligned}$$

The aim of this problem is to show that $\sum_{n=0}^{\infty} b_n x^n$ also converges for some $x \neq 0$, so that it represents $1/f$ for small enough $|x|$.

- (a) If all $|a_n x_0^n| \leq M$, show that

$$|b_n x^n| \leq M \sum_{k=0}^{n-1} |b_k x^k|.$$

- (b) Choose $M \geq \sqrt{2}$ with all $|a_n x_0^n| \leq M$. Show that

$$|b_n x_0^n| \leq M^{2n}.$$

- (c) Conclude that $\sum_{n=0}^{\infty} b_n x^n$ converges for $|x|$ sufficiently small.

18. Show that the series

$$\sum_{n=0}^{\infty} \frac{x^{2n+1}}{2n+1} - \frac{x^{n+1}}{2n+2}$$

converges uniformly to $\frac{1}{2} \log(x+1)$ on $[-a, a]$ for $0 < a < 1$, but that at 1 it converges to $\log 2!$

- *19. Suppose that $\sum_{n=0}^{\infty} a_n$ converges. We know that the series $f(x) = \sum_{n=0}^{\infty} a_n x^n$ must converge uniformly on $[-a, a]$ for $0 < a < 1$, but it may not converge uniformly on $[-1, 1]$; in fact, it may not even converge at the point -1 (for example, if $f(x) = \log(1+x)$). However, a beautiful theorem of Abel shows that the series *does* converge uniformly on $[0, 1]$. Consequently, f is continuous on $[0, 1]$ and, in particular, $\sum_{n=0}^{\infty} a_n = \lim_{x \rightarrow 1^-} \sum_{n=0}^{\infty} a_n x^n$. Prove Abel's Theorem by noticing that if $|a_m + \dots + a_n| < \varepsilon$, then $|a_m x^m + \dots + a_n x^n| < \varepsilon$, by Abel's Lemma (Problem 19-35).

20. A sequence $\{a_n\}$ is called **Abel summable** if $\lim_{x \rightarrow 1^-} \sum_{n=0}^{\infty} a_n x^n$ exists; Problem 19 shows that a summable sequence is necessarily Abel summable. Find a sequence which is Abel summable, but which is not summable. Hint: Look over the list of Taylor series until you find one which does not converge at 1, even though the function it represents is continuous at 1.
21. (a) Using Problem 19, find the following infinite sums.
- (i) $\frac{1}{2 \cdot 1} - \frac{1}{3 \cdot 2} + \frac{1}{4 \cdot 3} - \frac{1}{5 \cdot 4} + \dots$
 - (ii) $1 - \frac{1}{4} + \frac{1}{7} - \frac{1}{10} + \dots$
- (b) Let $\sum_{n=0}^{\infty} c_n$ be the Cauchy product of two convergent power series $\sum_{n=0}^{\infty} a_n$ and $\sum_{n=0}^{\infty} b_n$, and suppose merely that $\sum_{n=0}^{\infty} c_n$ converges. Prove that, in fact, it converges to the product $\sum_{n=0}^{\infty} a_n \cdot \sum_{n=0}^{\infty} b_n$.
22. (a) Suppose that $\{f_n\}$ is a sequence of bounded (not necessarily continuous) functions on $[a, b]$ which converge uniformly to f on $[a, b]$. Prove that f is bounded on $[a, b]$.
- (b) Find a sequence of continuous functions on $[a, b]$ which converge pointwise to an unbounded function on $[a, b]$.
- *23. Suppose that f is differentiable. Prove that the function f' is the pointwise limit of a sequence of continuous functions. (Since we already know examples of discontinuous derivatives, this provides another example where the pointwise limit of continuous functions is not continuous.)
24. Find a sequence of integrable functions $\{f_n\}$ which converges to the (nonintegrable) function f that is 1 on the rationals and 0 on the irrationals. Hint: Each f_n will be 0 except at a few points.
25. (a) Prove that if f is the uniform limit of $\{f_n\}$ on $[a, b]$ and each f_n is integrable on $[a, b]$, then so is f . (So one of the hypotheses in Theorem 1 was unnecessary.)
- (b) In Theorem 3 we assumed only that $\{f_n\}$ converges pointwise to f . Show that the remaining hypotheses ensure that $\{f_n\}$ actually converges uniformly to f .
- (c) Suppose that in Theorem 3 we do not assume $\{f_n\}$ converges to a function f , but instead assume only that $f_n(x_0)$ converges for some x_0 in $[a, b]$. Show that f_n does converge (uniformly) to some f (with $f' = g$).
- (d) Prove that the series

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{x+n}$$

converges uniformly on $[0, \infty)$.

26. Suppose that f_n are continuous functions on $[0, 1]$ that converge uniformly to f . Prove that

$$\lim_{n \rightarrow \infty} \int_0^{1-1/n} f_n = \int_0^1 f.$$

Is this true if the convergence isn't uniform?

27. (a) Suppose that $\{f_n\}$ is a sequence of continuous functions on $[a, b]$ which approach 0 pointwise. Suppose moreover that we have $f_n(x) \geq f_{n+1}(x) \geq 0$ for all n and all x in $[a, b]$. Prove that $\{f_n\}$ actually approaches 0 uniformly on $[a, b]$. Hint: Suppose not, choose an appropriate sequence of points x_n in $[a, b]$, and apply the Bolzano-Weierstrass theorem.
 (b) Prove Dini's Theorem: If $\{f_n\}$ is a nonincreasing sequence of continuous functions on $[a, b]$ which approach the continuous function f pointwise, then $\{f_n\}$ also approaches f uniformly on $[a, b]$. (The same result holds if $\{f_n\}$ is a nondecreasing sequence.)
 (c) Does Dini's Theorem hold if f isn't continuous? How about if $[a, b]$ is replaced by the open interval (a, b) ?
28. (a) Suppose that $\{f_n\}$ is a sequence of continuous functions on $[a, b]$ that converges uniformly to f . Prove that if x_n approaches x , then $f_n(x_n)$ approaches $f(x)$.
 (b) Is this statement true without assuming that the f_n are continuous?
 (c) Prove the converse of part (a): If f is continuous on $[a, b]$ and $\{f_n\}$ is a sequence with the property that $f_n(x_n)$ approaches $f(x)$ whenever x_n approaches x , then f_n converges uniformly to f on $[a, b]$. Hint: If not, there is an $\varepsilon > 0$ and a sequence x_n with $|f_n(x_n) - f(x_n)| > \varepsilon$. Then use the Bolzano-Weierstrass theorem.
29. This problem outlines a completely different approach to the integral; consequently, it is unfair to use any facts about integrals learned previously.
- (a) Let s be a step function on $[a, b]$, so that s is constant on (t_{i-1}, t_i) for some partition $\{t_0, \dots, t_n\}$ of $[a, b]$. Define $\int_a^b s$ as $\sum_{i=1}^n s_i \cdot (t_i - t_{i-1})$ where s_i is the (constant) value of s on (t_{i-1}, t_i) . Show that this definition does not depend on the partition $\{t_0, \dots, t_n\}$.
 (b) A function f is called a **regulated** function on $[a, b]$ if it is the uniform limit of a sequence of step functions $\{s_n\}$ on $[a, b]$. Show that in this case there is, for every $\varepsilon > 0$, some N such that for $m, n > N$ we have $|s_n(x) - s_m(x)| < \varepsilon$ for all x in $[a, b]$.
 (c) Show that the sequence of numbers $\left\{ \int_a^b s_n \right\}$ will be a Cauchy sequence.
 (d) Suppose that $\{t_n\}$ is another sequence of step functions on $[a, b]$ which converges uniformly to f . Show that for every $\varepsilon > 0$ there is an N such that for $n > N$ we have $|s_n(x) - t_n(x)| < \varepsilon$ for x in $[a, b]$.

516 Infinite Sequences and Infinite Series

- (e) Conclude that $\lim_{n \rightarrow \infty} \int_a^b s_n = \lim_{n \rightarrow \infty} \int_a^b t_n$. This means that we can *define* $\int_a^b f$ to be $\lim_{n \rightarrow \infty} s_n$ for any sequence of step functions $\{s_n\}$ converging uniformly to f . The only remaining question is: Which functions are regulated? Here is a partial answer.
- *(f) Prove that a continuous function is regulated. Hint: To find a step function s on $[a, b]$ with $|f(x) - s(x)| < \varepsilon$ for all x in $[a, b]$, consider all y for which there is such a step function on $[a, y]$.
- *30. Find a sequence $\{f_n\}$ approaching f uniformly on $[0, 1]$ for which $\lim_{n \rightarrow \infty} (\text{length of } f_n \text{ on } [0, 1]) \neq \text{length of } f \text{ on } [0, 1]$. (Length is defined in Problem 13-25, but the simplest example will involve functions the length of whose graphs will be obvious.)

CHAPTER 25

COMPLEX NUMBERS

With the exception of the last few paragraphs of the previous chapter, this book has presented unremitting propaganda for the real numbers. Nevertheless, the real numbers do have a great deficiency—not every polynomial function has a root. The simplest and most notable example is the fact that no number x can satisfy $x^2 + 1 = 0$. This deficiency is so severe that long ago mathematicians felt the need to “invent” a number i with the property that $i^2 + 1 = 0$. For a long time the status of the “number” i was quite mysterious: since there is no number x satisfying $x^2 + 1 = 0$, it is nonsensical to say “let i be the number satisfying $i^2 + 1 = 0$.” Nevertheless, admission of the “imaginary” number i to the family of numbers seemed to simplify greatly many algebraic computations, especially when “complex numbers” $a + bi$ (for a and b in \mathbb{R}) were allowed, and all the laws of arithmetical computation enumerated in Chapter 1 were assumed to be valid. For example, every quadratic equation

$$ax^2 + bx + c = 0 \quad (a \neq 0)$$

can be solved formally to give

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{or} \quad x = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

If $b^2 - 4ac \geq 0$, these formulas give correct solutions; when complex numbers are allowed the formulas seem to make sense in all cases. For example, the equation

$$x^2 + x + 1 = 0$$

has no real root, since

$$x^2 + x + 1 = (x + \frac{1}{2})^2 + \frac{3}{4} > 0, \quad \text{for all } x.$$

But the formula for the roots of a quadratic equation suggest the “solutions”

$$x = \frac{-1 + \sqrt{-3}}{2} \quad \text{and} \quad x = \frac{-1 - \sqrt{-3}}{2};$$

if we understand $\sqrt{-3}$ to mean $\sqrt{3 \cdot (-1)} = \sqrt{3} \cdot \sqrt{-1} = \sqrt{3}i$, then these numbers would be

$$-\frac{1}{2} + \frac{\sqrt{3}}{2}i \quad \text{and} \quad -\frac{1}{2} - \frac{\sqrt{3}}{2}i.$$

It is not hard to check that these, as yet purely formal, numbers do indeed satisfy the equation

$$x^2 + x + 1 = 0.$$

It is even possible to “solve” quadratic equations whose coefficients are themselves complex numbers. For example, the equation

$$x^2 + x + 1 + i = 0$$

ought to have the solutions

$$x = \frac{-1 \pm \sqrt{1 - 4(1+i)}}{2} = \frac{-1 \pm \sqrt{-3 - 4i}}{2},$$

where the symbol $\sqrt{-3 - 4i}$ means a complex number $\alpha + \beta i$ whose square is $-3 - 4i$. In order to have

$$(\alpha + \beta i)^2 = \alpha^2 - \beta^2 + 2\alpha\beta i = -3 - 4i$$

we need

$$\begin{aligned}\alpha^2 - \beta^2 &= -3, \\ 2\alpha\beta &= -4.\end{aligned}$$

These two equations can easily be solved for real α and β ; in fact, there are two possible solutions:

$$\begin{array}{lll}\alpha = 1 & \text{and} & \alpha = -1 \\ \beta = -2 & & \beta = 2.\end{array}$$

Thus the two “square roots” of $-3 - 4i$ are $1 - 2i$ and $-1 + 2i$. There is no reasonable way to decide which one of these should be called $\sqrt{-3 - 4i}$, and which $-\sqrt{-3 - 4i}$; the conventional usage of \sqrt{x} makes sense only for real $x \geq 0$, in which case \sqrt{x} denotes the (real) nonnegative root. For this reason, the solution

$$x = \frac{-1 \pm \sqrt{-3 - 4i}}{2}$$

must be understood as an abbreviation for:

$$x = \frac{-1 + r}{2}, \quad \text{where } r \text{ is one of the square roots of } -3 - 4i.$$

With this understanding we arrive at the solutions

$$\begin{aligned}x &= \frac{-1 + 1 - 2i}{2} = -i, \\ x &= \frac{-1 - 1 + 2i}{2} = -1 + i;\end{aligned}$$

as you can easily check, these numbers do provide formal solutions for the equation

$$x^2 + x + 1 + i = 0.$$

For cubic equations complex numbers are equally useful. Every cubic equation

$$ax^3 + bx^2 + cx + d = 0 \quad (a \neq 0)$$

with real coefficients a, b, c , and d , has, as we know, a real root α , and if we divide $ax^3 + bx^2 + cx + d$ by $x - \alpha$ we obtain a second-degree polynomial whose roots are the other roots of $ax^3 + bx^2 + cx + d = 0$; the roots of this second-degree polynomial

may be complex numbers. Thus a cubic equation will have either three real roots or one real root and 2 complex roots. The existence of the real root is guaranteed by our theorem that every odd degree equation has a real root, but it is not really necessary to appeal to this theorem (which is of no use at all if the coefficients are complex); in the case of a cubic equation we can, with sufficient cleverness, actually find a formula for all the roots. The following derivation is presented not only as an interesting illustration of the ingenuity of early mathematicians, but as further evidence for the importance of complex numbers (whatever they may be).

To solve the most general cubic equation, it obviously suffices to consider only equations of the form

$$x^3 + bx^2 + cx + d = 0.$$

It is even possible to eliminate the term involving x^2 , by a fairly straight-forward manipulation. If we let

$$x = y - \frac{b}{3},$$

then

$$\begin{aligned} x^3 &= y^3 - by^2 + \frac{b^2y}{3} - \frac{b^3}{27}, \\ x^2 &= y^2 - \frac{2by}{3} + \frac{b^2}{9}, \end{aligned}$$

so

$$\begin{aligned} 0 &= x^3 + bx^2 + cx + d \\ &= \left(y^3 - by^2 + \frac{b^2y}{3} - \frac{b^3}{27} \right) + \left(by^2 - \frac{2b^2y}{3} + \frac{b^3}{9} \right) + \left(cy - \frac{bc}{3} \right) + d \\ &= y^3 + \left(\frac{b^2}{3} - \frac{2b^2}{3} + c \right) y + \left(\frac{b^3}{9} - \frac{b^3}{27} - \frac{bc}{3} + d \right). \end{aligned}$$

The right-hand side now contains no term with y^2 . If we can solve the equation for y we can find x ; this shows that it suffices to consider in the first place only equations of the form

$$x^3 + px + q = 0.$$

In the special case $p = 0$ we obtain the equation $x^3 = -q$. We shall see later on that every complex number does have a cube root, in fact it has three, so that this equation has three solutions. The case $p \neq 0$, on the other hand, requires quite an ingenious step. Let

$$(*) \quad x = w - \frac{p}{3w}$$

Then

$$\begin{aligned} 0 &= x^3 + px + q = \left(w - \frac{p}{3w}\right)^3 + p\left(w - \frac{p}{3w}\right) + q \\ &= w^3 - \frac{3w^2p}{3w} + \frac{3wp^2}{9w^2} - \frac{p^3}{27w^3} + pw - \frac{p^2}{3w} + q \\ &= w^3 - \frac{p^3}{27w^3} + q. \end{aligned}$$

This equation can be written

$$27(w^3)^2 + 27q(w^3) - p^3 = 0,$$

which is a quadratic equation in w^3 (!!).

Thus

$$\begin{aligned} w^3 &= \frac{-27q \pm \sqrt{(27)^2q^2 + 4 \cdot 27p^3}}{2 \cdot 27} \\ &= -\frac{q}{2} \pm \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}. \end{aligned}$$

Remember that this really means:

$$w^3 = -\frac{q}{2} + r, \quad \text{where } r \text{ is a square root of } \frac{q^2}{4} + \frac{p^3}{27}.$$

We can therefore write

$$w = \sqrt[3]{-\frac{q}{2} \pm \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}};$$

this equation means that w is some cube root of $-q/2+r$, where r is some square root of $q^2/4 + p^3/27$. This allows six possibilities for w , but when these are substituted into (*), yielding

$$x = \sqrt[3]{-\frac{q}{2} \pm \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} - \frac{p}{3 \cdot \sqrt[3]{-\frac{q}{2} \pm \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}},$$

it turns out that only 3 different values for x will be obtained! An even more surprising feature of this solution arises when we consider a cubic equation all of whose roots are real; the formula derived above may still involve complex numbers in an essential way. For example, the roots of

$$x^3 - 15x - 4 = 0$$

are 4 , $-2 + \sqrt{3}$, and $-2 - \sqrt{3}$. On the other hand, the formula derived above (with $p = -15$, $q = -4$) gives as one solution

$$\begin{aligned}x &= \sqrt[3]{2 + \sqrt{4 - 125}} - \frac{-15}{3 \cdot \sqrt[3]{2 + \sqrt{4 - 125}}} \\&= \sqrt[3]{2 + 11i} + \frac{15}{3 \cdot \sqrt[3]{2 + 11i}}.\end{aligned}$$

Now,

$$\begin{aligned}(2+i)^3 &= 2^3 + 3 \cdot 2^2 i + 3 \cdot 2 \cdot i^2 + i^3 \\&= 8 + 12i - 6 - i \\&= 2 + 11i,\end{aligned}$$

so one of the cube roots of $2 + 11i$ is $2 + i$. Thus, for one solution of the equation we obtain

$$\begin{aligned}x &= 2 + i + \frac{15}{6 + 3i} \\&= 2 + i + \frac{15}{6 + 3i} \cdot \frac{6 - 3i}{6 - 3i} \\&= 2 + i + \frac{90 - 45i}{36 + 9} \\&= 4 (!).\end{aligned}$$

The other roots can also be found if the other cube roots of $2 + 11i$ are known. The fact that even one of these real roots is obtained from an expression which depends on complex numbers is impressive enough to suggest that the use of complex numbers cannot be entirely nonsense. As a matter of fact, the formulas for the solutions of the quadratic and cubic equations can be interpreted entirely in terms of real numbers.

Suppose we agree, for the moment, to write all complex numbers as $a + bi$, writing the real number a as $a + 0i$ and the number i as $0 + 1i$. The laws of ordinary arithmetic and the relation $i^2 = -1$ show that

$$\begin{aligned}(a + bi) + (c + di) &= (a + c) + (b + d)i \\(a + bi) \cdot (c + di) &= (ac - bd) + (ad + bc)i.\end{aligned}$$

Thus, an equation like

$$(1 + 2i) \cdot (3 + 1i) = 1 + 7i$$

may be regarded simply as an abbreviation for the *two* equations

$$\begin{aligned}1 \cdot 3 - 2 \cdot 1 &= 1, \\1 \cdot 1 + 2 \cdot 3 &= 7.\end{aligned}$$

The solution of the quadratic equation $ax^2 + bx + c = 0$ with real coefficients could be paraphrased as follows:

$$\text{If } \begin{cases} u^2 - v^2 = b^2 - 4ac, \\ uv = 0, \end{cases} \text{ (i.e., if } (u + vi)^2 = b^2 - 4ac),$$

$$\text{then } \begin{cases} a \left[\left(\frac{-b+u}{2a} \right)^2 - \left(\frac{v}{2a} \right)^2 \right] + b \left[\frac{-b+u}{2a} \right] + c = 0, \\ a \left[2 \left(\frac{-b+u}{2a} \right) \left(\frac{v}{2a} \right) \right] + b \left[\frac{v}{a} \right] = 0, \end{cases} \text{ (i.e., then } a \left(\frac{-b+u+vi}{2a} \right)^2 + b \left(\frac{-b+u+vi}{2a} \right) + c = 0 \text{).}$$

It is not very hard to check this assertion about real numbers without writing down a single “ i ,” but the complications of the statement itself should convince you that equations about complex numbers are worthwhile as abbreviations for pairs of equations about real numbers. (If you are still not convinced, try paraphrasing the solution of the cubic equation.) If we really intend to use complex numbers consistently, however, it is going to be necessary to present some reasonable definition.

One possibility has been implicit in this whole discussion. All mathematical properties of a complex number $a + bi$ are determined completely by the real numbers a and b ; any mathematical object with this same property may reasonably be used to define a complex number. The obvious candidate is the ordered pair (a, b) of real numbers; we shall accordingly *define* a complex number to be a pair of real numbers, and likewise *define* what addition and multiplication of complex numbers is to mean.

DEFINITION

A **complex number** is an ordered pair of real numbers; if $z = (a, b)$ is a complex number, then a is called the **real part** of z , and b is called the **imaginary part** of z . The set of all complex numbers is denoted by \mathbf{C} . If (a, b) and (c, d) are two complex numbers we define

$$(a, b) + (c, d) = (a + c, b + d) \\ (a, b) \cdot (c, d) = (a \cdot c - b \cdot d, a \cdot d + b \cdot c).$$

(The $+$ and \cdot appearing on the left side are new symbols being defined, while the $+$ and \cdot appearing on the right side are the familiar addition and multiplication for real numbers.)

When complex numbers were first introduced, it was understood that real numbers were, in particular, complex numbers; if our definition is taken seriously this is not true—a real number is not a pair of real numbers, after all. This difficulty

is only a minor annoyance, however. Notice that

$$(a, 0) + (b, 0) = (a + b, 0 + 0) = (a + b, 0), \\ (a, 0) \cdot (b, 0) = (a \cdot b - 0 \cdot 0, a \cdot 0 + 0 \cdot b) = (a \cdot b, 0);$$

this shows that the complex numbers of the form $(a, 0)$ behave precisely the same with respect to addition and multiplication of complex numbers as real numbers do with their own addition and multiplication. For this reason we will adopt the convention that $(a, 0)$ will be denoted simply by a . The familiar $a + bi$ notation for complex numbers can now be recovered if one more definition is made.

DEFINITION

$$i = (0, 1).$$

Notice that $i^2 = (0, 1) \cdot (0, 1) = (-1, 0) = -1$ (the last equality sign depends on our convention). Moreover

$$\begin{aligned} (a, b) &= (a, 0) + (0, b) \\ &= (a, 0) + (b, 0) \cdot (0, 1) \\ &= a + bi. \end{aligned}$$

You may feel that our definition was merely an elaborate device for defining complex numbers as “expressions of the form $a + bi$.” That is essentially correct; it is a firmly established prejudice of modern mathematics that new objects must be defined as something specific, not as “expressions.” Nevertheless, it is interesting to note that mathematicians were sincerely worried about using complex numbers until the modern definition was proposed. Moreover, the precise definition emphasizes one important point. Our aim in introducing complex numbers was to avoid the necessity of paraphrasing statements about complex numbers in terms of their real and imaginary parts. This means that we wish to work with complex numbers in the same way that we worked with rational or real numbers. For example, the solution of the cubic equation required writing $x = w - p/3w$, so we want to know that $1/w$ makes sense. Moreover, w^2 was found by solving a quadratic equation, which requires numerous other algebraic manipulations. In short, we are likely to use, at some time or other, any manipulations performed on real numbers. We certainly do not want to stop each time and justify every step. Fortunately this is not necessary. Since all algebraic manipulations performed on real numbers can be justified by the properties listed in Chapter 1, it is only necessary to check that these properties are also true for complex numbers. In most cases this is quite easy, and these facts will not be listed as formal theorems. For example, the proof of P1,

$$[(a, b) + (c, d)] + (e, f) = (a, b) + [(c, d) + (e, f)]$$

requires only the application of the definition of addition for complex numbers. The left side becomes

$$([a + c] + e, [b + d] + f),$$

and the right side becomes

$$(a + [c + e], b + [d + f]);$$

these two are equal because P1 is true for real numbers. It is a good idea to check P2–P6 and P8 and P9. Notice that the complex numbers playing the role of 0 and 1 in P2 and P6 are $(0, 0)$ and $(1, 0)$, respectively. It is not hard to figure out what $-(a, b)$ is, but the multiplicative inverse for (a, b) required in P7 is a little trickier: if $(a, b) \neq (0, 0)$, then $a^2 + b^2 \neq 0$ and

$$(a, b) \cdot \left(\frac{a}{a^2 + b^2}, \frac{-b}{a^2 + b^2} \right) = (1, 0).$$

This fact could have been guessed in two ways. To find (x, y) with

$$(a, b) \cdot (x, y) = (1, 0)$$

it is only necessary to solve the equations

$$\begin{aligned} ax - by &= 1, \\ bx + ay &= 0. \end{aligned}$$

The solutions are $x = a/(a^2 + b^2)$, $y = -b/(a^2 + b^2)$. It is also possible to reason that if $1/(a + bi)$ means anything, then it should be true that

$$\frac{1}{a + bi} = \frac{1}{a + bi} \cdot \frac{a - bi}{a - bi} = \frac{a - bi}{a^2 + b^2}.$$

Once the existence of inverses has actually been proved (after guessing the inverse by some method), it follows that this manipulation is really valid; it is the easiest one to remember when the inverse of a complex number is actually being sought—it was precisely this trick which we used to evaluate

$$\begin{aligned} \frac{15}{6 + 3i} &= \frac{15}{6 + 3i} \cdot \frac{6 - 3i}{6 - 3i} \\ &= \frac{90 - 45i}{36 + 9}. \end{aligned}$$

Unlike P1–P9, the rules P10–P12 do not have analogues: it is easy to prove that there is no set P of *complex* numbers such that P10–P12 are satisfied for all *complex* numbers. In fact, if there were, then P would have to contain 1 (since $1 = 1^2$) and also -1 (since $-1 = i^2$), and this would contradict P10. The absence of P10–P12 will not have disastrous consequences, but it does mean that we cannot define $z < w$ for complex z and w . Also, you may remember that for the real numbers, P10–P12 were used to prove that $1 + 1 \neq 0$. Fortunately, the corresponding fact for complex numbers can be reduced to this one: clearly $(1, 0) + (1, 0) \neq (0, 0)$.

Although we will usually write complex numbers in the form $a + bi$, it is worth remembering that the set of all complex numbers \mathbf{C} is just the collection of all pairs of real numbers. Long ago this collection was identified with the plane, and for this reason the plane is often called the “complex plane.” The horizontal axis, which consists of all points $(a, 0)$ for a in \mathbf{R} , is often called the *real axis*, and the

vertical axis is called the *imaginary axis*. Two important definitions are also related to this geometric picture.

DEFINITION

If $z = x + iy$ is a complex number (with x and y real), then the **conjugate** \bar{z} of z is defined as

$$\bar{z} = x - iy,$$

and the **absolute value or modulus** $|z|$ of z is defined as

$$|z| = \sqrt{x^2 + y^2}.$$

(Notice that $x^2 + y^2 \geq 0$, so that $\sqrt{x^2 + y^2}$ is defined unambiguously; it denotes the nonnegative real square root of $x^2 + y^2$.)

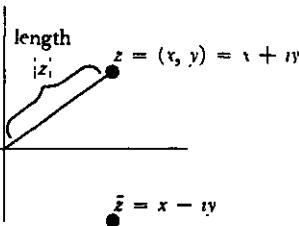


FIGURE 1

Geometrically, \bar{z} is simply the reflection of z in the real axis, while $|z|$ is the distance from z to $(0, 0)$ (Figure 1). Notice that the absolute value notation for complex numbers is consistent with that for real numbers. The **distance** between two complex numbers z and w can be defined quite easily as $|z - w|$. The following theorem lists all the important properties of conjugates and absolute values.

THEOREM 1

Let z and w be complex numbers. Then

- (1) $\bar{\bar{z}} = z$.
- (2) $\bar{z} = z$ if and only if z is real (i.e., is of the form $a + 0i$, for some real number a).
- (3) $\overline{z + w} = \bar{z} + \bar{w}$.
- (4) $\overline{-z} = -\bar{z}$.
- (5) $\overline{z \cdot w} = \bar{z} \cdot \bar{w}$.
- (6) $\overline{z^{-1}} = (\bar{z})^{-1}$, if $z \neq 0$.
- (7) $|z|^2 = z \cdot \bar{z}$.
- (8) $|z \cdot w| = |z| \cdot |w|$.
- (9) $|z + w| \leq |z| + |w|$.

PROOF

Assertions (1) and (2) are obvious. Equations (3) and (5) may be checked by straightforward calculations and (4) and (6) may then be proved by a trick:

$$0 = \bar{0} = \overline{z + (-z)} = \bar{z} + \overline{-z}, \quad \text{so } \overline{-z} = -\bar{z},$$

$$1 = \bar{1} = \overline{z \cdot (z^{-1})} = \bar{z} \cdot \overline{z^{-1}}, \quad \text{so } \overline{z^{-1}} = (\bar{z})^{-1}.$$

Equations (7) and (8) may also be proved by a straightforward calculation. The only difficult part of the theorem is (9). This inequality has, in fact, already occurred (Problem 4-9), but the proof will be repeated here, using slightly different terminology.

It is clear that equality holds in (9) if $z = 0$ or $w = 0$. It is also easy to see that (9) is true if $z = \lambda w$ for any real number λ (consider separately the cases $\lambda > 0$ and $\lambda < 0$). Suppose, on the other hand, that $z \neq \lambda w$ for any real number λ , and that

$w \neq 0$. Then, for all real numbers λ ,

$$\begin{aligned} (*) \quad 0 < |z - \lambda w|^2 &= (z - \lambda w) \cdot \overline{(z - \lambda w)} \\ &= (z - \lambda w) \cdot (\bar{z} - \lambda \bar{w}) \\ &= z\bar{z} + \lambda^2 w\bar{w} - \lambda(w\bar{z} + z\bar{w}) \\ &= \lambda^2|w|^2 + |z|^2 - \lambda(w\bar{z} + z\bar{w}). \end{aligned}$$

Notice that $w\bar{z} + z\bar{w}$ is real, since

$$\overline{w\bar{z} + z\bar{w}} = \bar{w}\bar{\bar{z}} + \bar{z}\bar{\bar{w}} = \bar{w}z + \bar{z}w = w\bar{z} + z\bar{w}.$$

Thus the right side of $(*)$ is a quadratic equation in λ with real coefficients and no real solutions; its discriminant must therefore be negative. Thus

$$(w\bar{z} + z\bar{w})^2 - 4|w|^2 \cdot |z|^2 < 0;$$

it follows, since $w\bar{z} + z\bar{w}$ and $|w| \cdot |z|$ are real numbers, and $|w| \cdot |z| \geq 0$, that

$$(w\bar{z} + z\bar{w}) < 2|w| \cdot |z|.$$

From this inequality it follows that

$$\begin{aligned} |z + w|^2 &= (z + w) \cdot (\bar{z} + \bar{w}) \\ &= |z|^2 + |w|^2 + (w\bar{z} + z\bar{w}) \\ &< |z|^2 + |w|^2 + 2|w| \cdot |z| \\ &= (|z| + |w|)^2, \end{aligned}$$

which implies that

$$|z + w| < |z| + |w|. \blacksquare$$

The operations of addition and multiplication of complex numbers both have important geometric interpretations. The picture for addition is very simple (Figure 2). Two complex numbers $z = (a, b)$ and $w = (c, d)$ determine a parallelogram having for two of its sides the line segment from $(0, 0)$ to z , and the line segment from $(0, 0)$ to w ; the vertex opposite $(0, 0)$ is $z + w$ (a proof of this geometric fact is left to you [compare Appendix 1 to Chapter 4]).

The interpretation of multiplication is more involved. If $z = 0$ or $w = 0$, then $z \cdot w = 0$ (a one-line computational proof can be given, but even this is unnecessary—the assertion has already been shown to follow from P1–P9), so we may restrict our attention to nonzero complex numbers. We begin by putting every nonzero complex number into a special form (compare Appendix 3 to Chapter 4).

For any complex number $z \neq 0$ we can write

$$z = |z| \frac{z}{|z|};$$

in this expression, $|z|$ is a positive real number, while

$$\left| \frac{z}{|z|} \right| = \frac{|z|}{|z|} = 1,$$

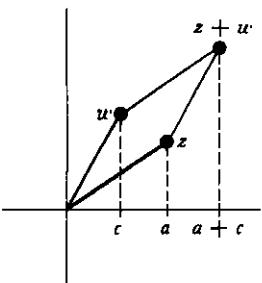


FIGURE 2