Chapter I: The Real Numbers

Don't play poker at someone's house unless you know the rules of the house. And don't play mathematics unless you know the rules of the subject – the axioms.

Accordingly, it is only fair that we set down the rules we will play by: the axioms for the real numbers. This is simple enough to do. However, some basic consequences of the axioms should also be presented so that you know how some "rules" you have been taught, which are not axioms, follow from the axioms. For example, "a minus times a minus is a plus", "zero times any number is zero","$0 < 1$", "$\frac{x}{y} + \frac{z}{w} = \frac{xw+yz}{yw}$" are not "rules" and formulas to be committed to memory for future use; they all follow from the axioms. Once you understand *how* they follow from the axioms, you will understand them better; put another way, the axioms will focus your understanding.

We can not verify *all* the familiar consequences of the axioms. We verify some of the more prominent consequences of the axioms; we hope that what we do, and what you are asked to do in exercises, is enough to make you feel that all the arithmetic you use could, indeed, be verified from the axioms.

All this having been said, I have to admit to being slightly disingenuous. I am referring to certain facts about the natural numbers $1, 2, 3, ...$ stated in section 4 (1.18). These facts are not consequences of the axioms for the real numbers listed in section 1; instead, they come from a way the natural numbers can be constructed. We do not do the construction, but the facts are easy for you to accept based on your past experience with the natural numbers. In effect, we will accept certain facts about the natural numbers as though they are axioms, but we postpone mentioning them until section 4.

1. The Axioms

We denote the set of real numbers by R^1.

We state the axioms for the real numbers. You are familiar with most of the axioms as the "rules of arithmetic"; the exception may be the Completeness Axiom. The Completeness Axiom is necessary since the rational numbers satisfy all the other axioms – thus, without the Completeness Axiom, we are not guaranteed that $\sqrt{2}$, π, etc. are real numbers (we prove that positive numbers have square roots in section 5).

(A) Addition Axioms. There is a function, $+$, defined on the Cartesian product $\mathsf{R}^1 \times \mathsf{R}^1$ satisfying A1 - A5 below (we write $a + b$ to stand for $+(a, b)$):

A1: For any $a, b \in \mathsf{R}^1$, $a + b \in \mathsf{R}^1$. (Closure)

A2: For any $a, b \in \mathsf{R}^1$, $a + b = b + a$. (Commutativity)

A3: For any $a, b, c \in \mathsf{R}^1$, $(a + b) + c = a + (b + c)$. (Associativity)

A4: There is a real number, denoted by 0, such that $a+0 = a$ for all $a \in \mathsf{R}^1$. (Identity)

A5: For each $a \in \mathsf{R}^1$, there is a real number, denoted by $-a$, such that $a + (-a) = 0$. (Inverse)

(M) Multiplication Axioms. There is a function, $\cdot$, defined on the Cartesian product $\mathsf{R}^1 \times \mathsf{R}^1$ satisfying M1 - M5 below (we write $a \cdot b$ to stand for $\cdot(a, b)$):

M1: For any $a, b \in \mathsf{R}^1$, $a \cdot b \in \mathsf{R}^1$. (Closure)

M2: For any $a, b \in \mathsf{R}^1$, $a \cdot b = b \cdot a$. (Commutativity)

M3: For any $a, b, c \in \mathsf{R}^1$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$. (Associativity)

M4: There is a real number, denoted by 1, such that $1 \neq 0$ and $a \cdot 1 = a$ for all $a \in \mathsf{R}^1$. (Identity)

M5: For each $a \in \mathsf{R}^1$ such that $a \neq 0$, there is a real number, denoted by a^{-1} or by $\frac{1}{a}$, such that $a \cdot a^{-1} = 1$, equivalently, $a \cdot \frac{1}{a} = 1$. (Inverse)

(D) Distributive Axiom. For all $a, b, c \in \mathsf{R}^1$, $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$.

(O) Order Axioms. There is a relation, $<$, on R^1 satisfying O1 - O4 below ($x < y$ is read x is less than y):

O1: For any $x, y \in \mathsf{R}^1$, one and only one of the following holds: $x < y$, $y < x$, or $x = y$. (Trichotomy)

O2: If $x, y, z \in \mathsf{R}^1$ and if $x < y$ and $y < z$, then $x < z$. (Transitivity)

O3: If $x, y, z \in \mathsf{R}^1$ and if $x < y$, then $x + z < y + z$.

O4: If $x, y, z \in \mathsf{R}^1$ and if $x < y$ and $0 < z$, then $x \cdot z < y \cdot z$.

We write $x \leq y$ to mean $x < y$ or $x = y$. We sometimes write $y > x$ or $y \geq x$ to mean $x < y$ or $x \leq y$, respectively. A number x is said to be *positive* if $x > 0$, *negative* if $x < 0$ and *nonnegative* if $x \geq 0$.

(C) Completeness Axiom.[1] If A is a nonempty subset of R^1 such that A has an upper bound u (i.e., $a \leq u$ for all $a \in A$), then A has a least upper bound ℓ (i.e., ℓ is an upper bound for A and $\ell \leq u$ for all upper bounds u for A). (See Remarks about the Completeness Axiom below.)

We summarize the axioms: Axioms (A), (M) and (D) say that R^1 is a field; axioms (A), (M), (D) and (O) say that R^1 is an ordered field; axioms (A), (M), (D), (O) and (C) say that R^1 is a complete ordered field. It can be shown that there is a complete ordered field and that there is only one complete ordered field (up to isomorphism). Thus, the reals are the unique complete ordered field.

Remarks about the Completeness Axiom. We make three clarifying observations about the Completeness Axiom.

First, the word *has* in the axiom is not intended to be possessive: the open interval $(0, 1)$ has 1 as an upper bound, but $1 \notin (0, 1)$.

Second, a nonempty set A that has an upper bound has *only one* least upper bound: If ℓ_1 and ℓ_2 were two different least upper bounds of A, then $\ell_1 < \ell_2$ (since ℓ_2 is an upper bound) and $\ell_2 < \ell_1$, in contradiction to O1.

Third, the requirement in the axiom that A be nonempty is necessary. This is because *every* real number is an upper bound of the empty set $\emptyset$, which we see as follows: If a real number x is not an upper bound of $\emptyset$, then $a \not\leq x$ for some $a \in \emptyset$, which is impossible since there is no point a in $\emptyset$.

[1] The Completeness Axiom is often called the Least Upper Bound Axiom. We will say more about this in section 8.

2. Some Notation, Intervals

Eventually, we adopt all the notation used in arithmetic and algebra. For now, we minimize our notation to the notation in section 1 together with the following convenient extensions of that notation:

- $\frac{a}{c}$ stands for $a \cdot \frac{1}{c}$ (assuming $c \neq 0$ – recall M5); in particular, $\frac{a+b}{c}$ stands for $(a+b) \cdot \frac{1}{c}$ and $\frac{a \cdot b}{c}$ stands for $(a \cdot b) \cdot \frac{1}{c}$.
- $a^n = \underbrace{a \cdot a \cdot \cdots \cdot a}_{n \text{ terms}}$ for $a \in \mathsf{R}^1$ and $n = 1, 2, \ldots$. (We discuss $a^{\frac{1}{n}}$ in section 5, where we show $a^{\frac{1}{n}}$ is a number for all $a \geq 0$ and $n = 1, 2, \ldots$.)
- We frequently juxtapose order relations: for example, $a < b \leq c$ means $a < b$ and $b \leq c$.

We use the usual notation for intervals. We divide intervals into three kinds:

1. Open intervals: (a, b), (a, ∞), $(-\infty, a)$ and $(-\infty, \infty) = \mathsf{R}^1$, where
$$(a, b) = \{x \in \mathsf{R}^1 : a < x < b\}, \quad a < b;$$
$$(a, \infty) = \{x \in \mathsf{R}^1 : x > a\};$$
$$(-\infty, a) = \{x \in \mathsf{R}^1 : x < a\}.$$
2. Closed intervals: $[a, b]$, $[a, \infty)$ and $(-\infty, a]$, where
$$[a, b] = \{x \in \mathsf{R}^1 : a \leq x \leq b\}, \quad a \leq b;$$
$$[a, \infty) = \{x \in \mathsf{R}^1 : x \geq a\};$$
$$(-\infty, a] = \{x \in \mathsf{R}^1 : x \leq a\}.$$
3. Half-open (or half-closed) bounded intervals: $[a, b)$ and $(a, b]$, where
$$[a, b) = \{x \in \mathsf{R}^1 : a \leq x < b\}, \quad a < b;$$
$$(a, b] = \{x \in \mathsf{R}^1 : a < x \leq b\}, \quad a < b.$$

The notation for an open interval and an ordered pair is the same; nevertheless, the context will prevent confusion. In the notation for intervals, we used ∞ and $-\infty$ only as abstract symbols; in particular, the symbols ∞ and $-\infty$ never denote real numbers.

3. Algebra and Arithmetic

After stating the axioms in section 1, we remarked that the axioms say that the reals are the *unique* complete ordered field. Thus, if your previous experience with real numbers leads you to believe that real numbers satisfy the axioms, then you should believe that the axioms yield all the "facts" you have used about real numbers all your life (except for what we have said about the natural numbers in the introduction to the chapter). We show how some of these facts are consequences of the axioms; many other facts are left as exercises for you to do.

Our first theorem verifies cancellation for addition.

Theorem 1.1: Let $x, y, z \in \mathsf{R}^1$. If $x + y = x + z$, then $y = z$.

Proof: Since $x + y = x + z$ and $-x$ is a real number (by A5), clearly $(-x) + (x + y) = (-x) + (x + z)$. Hence, by A3,

$$((-x) + x) + y = ((-x) + x) + z).$$

Thus, by A2, $(x + (-x)) + y = (x + (-x)) + z$. Hence, by A5, $0 + y = 0 + z$. Hence, by A2, $y + 0 = z + 0$. Therefore, by A4, $y = z$. ¥

Theorem 1.2: If $x, y \in \mathsf{R}^1$ and $x + y = 0$, then $y = -x$. In other words, the additive inverse $-a$ of a in A5 is unique.

Proof: Since $x + y = 0$ by assumption and $x + (-x) = 0$ by A5, we have that

$$x + y = x + (-x).$$

Therefore, by Theorem 1.1, $y = -x$. ¥

The following corollary shows that the familiar adage "a minus times a minus is a plus" is true:

Corollary 1.3: For any $x \in \mathsf{R}^1$, $-(-x) = x$.

Proof: By A5, $x + (-x) = 0$. Hence, by A2, $(-x) + x = 0$. Therefore, by Theorem 1.2, $x = -(-x)$. ¥

Theorem 1.4: For any $x \in \mathsf{R}^1$, $x \cdot 0 = 0$.

Proof: We have

$$x + (x \cdot 0) \overset{\text{M4}}{=} (x \cdot 1) + (x \cdot 0) \overset{\text{D}}{=} x \cdot (1 + 0) \overset{\text{A4}}{=} x \cdot 1 \overset{\text{M4}}{=} x \overset{\text{A4}}{=} x + 0.$$

Therefore, by Theorem 1.1, $x \cdot 0 = 0$. ¥

Theorem 1.5: For any $x \in \mathsf{R}^1$, $-x = (-1) \cdot x$.

Proof: First, note that

$$x + (-1) \cdot x \overset{\text{M4}}{=} (x \cdot 1) + ((-1) \cdot x) \overset{\text{M2}}{=} (x \cdot 1) + (x \cdot (-1))$$

$$\overset{\text{D}}{=} x \cdot (1 + (-1)) \overset{\text{A5}}{=} x \cdot 0 \overset{1.4}{=} 0.$$

Therefore, by Theorem 1.2, $(-1) \cdot x = -x$. ¥

Before we give our next theorem, we comment about the statement and the proof of the theorem.

The conclusion of our next theorem is a compound statement, where the two parts are connected with the word *or*. When two statements are connected by *or*, we include the possibility that both statements may be true. This is not always the case in common usage: "At 7:00 P.M., I will be in New Orleans or I will be in Boston" obviously excludes both statements from being true. We use "either ... or" when we mean one or the other but not both. If P and Q are statements, then our meaning for the statement "P or Q" is called the *inclusive*

disjunction of P *and* Q; the statement "either P or Q" is called the *exclusive disjunction of* P *and* Q.

The proof of our next theorem illustrates a logical principle: To prove that the disjunction of two statements P and Q is true, it is sufficient to assume one of the statements is false and prove the other statement is true.

As you know from experience, the following theorem is useful in finding solutions to equations and inequalities.

Theorem 1.6: If $x, y \in \mathsf{R}^1$ and $x \cdot y = 0$, then $x = 0$ or $y = 0$.

Proof: Assume that $x \neq 0$. Then $\frac{1}{x}$ is a real number by M5. Thus, since $x \cdot y = 0$ by assumption and since $\frac{1}{x} \cdot 0 = 0$ by Theorem 1.4, we have that

$$\tfrac{1}{x} \cdot (x \cdot y) = 0.$$

Hence, by M3, $(\frac{1}{x} \cdot x) \cdot y = 0$. Thus, by M2, $(x \cdot \frac{1}{x}) \cdot y = 0$. Therefore, by M5, $1 \cdot y = 0$. Hence, by M2, $y \cdot 1 = 0$. Therefore, by M4, $y = 0$. ¥

The following theorem verifies cancellation for multiplication.

Theorem 1.7: If $x, y \in \mathsf{R}^1$ and $y \neq 0$, then $y \cdot \frac{x}{y} = x$.

Proof: Recall from the notation in section 2 that $\frac{x}{y} = x \cdot \frac{1}{y}$. Therefore,

$$y \cdot \tfrac{x}{y} = y \cdot (x \cdot \tfrac{1}{y}) \overset{\text{M2}}{=} y \cdot (\tfrac{1}{y} \cdot x) \overset{\text{M3}}{=} (y \cdot \tfrac{1}{y}) \cdot x \overset{\text{M5}}{=} 1 \cdot x$$

$$\overset{\text{M2}}{=} x \cdot 1 \overset{\text{M4}}{=} x. \quad ¥$$

We now come to the familiar formula for adding fractions.

Theorem 1.8: If $x, y, z, w \in \mathsf{R}^1$ such that $y \neq 0$ and $w \neq 0$, then

$$\tfrac{x}{y} + \tfrac{z}{w} = \tfrac{x \cdot w + y \cdot z}{y \cdot w}.$$

Proof: Recall from section 2 that the right-hand side of the equation is shorthand for $(x \cdot w + y \cdot z) \cdot \frac{1}{y \cdot w}$. Thus, we must first know that $\frac{1}{y \cdot w}$ is a real number: Since $y \neq 0$ and $w \neq 0$, $y \cdot w \neq 0$ by Theorem 1.6; therefore, $\frac{1}{y \cdot w}$ is a real number by M5.

Now,

$$(y \cdot w) \cdot (\tfrac{x}{y} + \tfrac{z}{w}) \overset{\text{D}}{=} [(y \cdot w) \cdot \tfrac{x}{y}] + [(y \cdot w) \cdot \tfrac{z}{w}]$$

$$\overset{\text{M2}}{=} [(w \cdot y) \cdot \tfrac{x}{y}] + [(y \cdot w) \cdot \tfrac{z}{w}] \overset{\text{M3}}{=} [w \cdot (y \cdot \tfrac{x}{y})] + [y \cdot (w \cdot \tfrac{z}{w})]$$

$$\overset{1.7}{=} w \cdot x + y \cdot z \overset{\text{M2}}{=} x \cdot w + y \cdot z.$$

Hence,

$$(*)\ \tfrac{1}{y \cdot w} \cdot [(y \cdot w) \cdot (\tfrac{x}{y} + \tfrac{z}{w})] = \tfrac{1}{y \cdot w} \cdot [x \cdot w + y \cdot z]$$

Our theorem follows from (*) since the left-hand side of (*) is

$$\tfrac{1}{y\cdot w}\cdot[(y\cdot w)\cdot(\tfrac{x}{y}+\tfrac{z}{w})] \overset{\text{M3}}{=} [\tfrac{1}{y\cdot w}\cdot(y\cdot w)]\cdot(\tfrac{x}{y}+\tfrac{z}{w})]$$

$$\overset{\text{M2}}{=} [(y\cdot w)\cdot\tfrac{1}{y\cdot w}]\cdot(\tfrac{x}{y}+\tfrac{z}{w})] \overset{\text{M5}}{=} 1\cdot(\tfrac{x}{y}+\tfrac{z}{w}) \overset{\text{M2}}{=} (\tfrac{x}{y}+\tfrac{z}{w})\cdot 1$$

$$\overset{\text{M4}}{=} \tfrac{x}{y}+\tfrac{z}{w}$$

and the right-hand side of (*) is

$$\tfrac{1}{y\cdot w}\cdot[x\cdot w+y\cdot z] \overset{\text{M2}}{=} [x\cdot w+y\cdot z]\cdot\tfrac{1}{y\cdot w} = \tfrac{x\cdot w+y\cdot z}{y\cdot w}.$$ ¥

Theorem 1.9: $0 < 1$.

Proof: By M4, $0 \neq 1$. Hence, by O1, either $1 < 0$ or $0 < 1$ (not both). Assume by way of contradiction that $1 < 0$. Then, by 03,

$$1 + (-1) < 0 + (-1).$$

Thus, since $1 + (-1) = 0$ by A5 and since $0 + (-1) \overset{\text{A2}}{=} (-1) + 0 \overset{\text{A4}}{=} -1$, we have that $0 < -1$. Hence, by O4, $0 \cdot (-1) < (-1) \cdot (-1)$. Therefore, by M2, $(-1) \cdot 0 < (-1) \cdot (-1)$. Hence, by Theorem 1.4, $0 < (-1) \cdot (-1)$. Thus, since $(-1) \cdot (-1) = -(-1)$ by Theorem 1.5, $0 < -(-1)$. Hence, by Corollary 1.3, $0 < 1$. Therefore, we have a contradiction to our assumption that $1 < 0$ (since O1 says $1 < 0$ and $0 < 1$ can not *both* occur). ¥

Corollary 1.10: For any $x \in \mathsf{R}^1$, $x < x + 1$.

Proof: By Theorem 1.9, $0 < 1$. Hence, by O3, $0 + x < 1 + x$. Thus, by A2, $x + 0 < x + 1$. Therefore, since $x + 0 = x$ by A4, $x < x + 1$. ¥

Exercise 1.11: For any $x \in \mathsf{R}^1$, $x + (-1) < x$.

Exercise 1.12: Let $x, y, z \in \mathsf{R}^1$ such that $x \neq 0$. If $x \cdot y = x \cdot z$, then $y = z$.

Exercise 1.13: If $x \in \mathsf{R}^1$ such that $x \neq 0$, then $(x^{-1})^{-1} = x$.

Exercise 1.14: If $x > 0$, then $-x < 0$ and $\frac{1}{x} > 0$.

Exercise 1.15: If $x, y, z, w \in \mathsf{R}^1$ such that $y \neq 0$ and $w \neq 0$, then

$$\tfrac{x}{y} \cdot \tfrac{z}{w} = \tfrac{x\cdot z}{y\cdot w}.$$

Exercise 1.16: If $x < y$ and $z < 0$, then $x \cdot z > y \cdot z$.

Exercise 1.17: For any $x \in \mathsf{R}^1$ such that $x \neq 0$, $x \cdot x > 0$.

4. The Natural Numbers

As mentioned in section 1, the reals are the unique complete ordered field. The existence of a complete ordered field is proved by constructing one. The process of constructing a complete ordered field often begins with constructing what will become the natural numbers. These are the numbers you have always

seen denoted by $1, 2, 3, \ldots$. We are not going to construct the natural numbers; we merely assume the natural numbers are the numbers $1, 2, 3, \ldots$ and denote the set of all natural numbers by N.

To attempt a little rigor, note that 1 is a real number by M4; then, by A1, $1 + 1$ is a real number which we denote by 2; then, by A1, $(1 + 1) + 1$ is a real number which we denote by 3; then, by A1, $((1 + 1) + 1) + 1$ is a real number which we denote by 4; and so on. However, what do we mean by "and so on"? This is troublesome since we are indicating an infinite process that is, heretofore, not well defined. How are we assured we have defined a set of objects? We return to this later (after the proof of Theorem 1.20).

Even though we will not construct the natural numbers, we need to assume facts about the natural numbers that are by-products of the construction. The facts are easy to accept since they seem obvious from our experience with the natural numbers. However, we emphsize that the facts are not merely intuitive – they come from the construction of the natural numbers and they hold for the "numbers" that are (properly) designated as the natural numbers in any construction of a complete ordered field.

1.18 Facts Assumed about N:

- $1 \in \mathsf{N}$ and $1 \leq n$ for all $n \in \mathsf{N}$.
- If $n, m \in \mathsf{N}$, then $n + m \in \mathsf{N}$ and $n \cdot m \in \mathsf{N}$. (Closure)
- If $n \in \mathsf{N}$, then $n > 0$ and $n + (-1) < n$.
- If $n \in \mathsf{N}$ and $n > 1$, then $n + (-1) \in \mathsf{N}$.
- **Well Ordering Principle:** Every nonempty subset S of N has a least member ℓ (i.e., there exists $\ell \in S$ such that $\ell \leq s$ for all $s \in S$).

We illustrate the usefulness of the Well Ordering Principle by proving the following seemingly obvious result (think about how you might prove the result without using the Well Ordering Principle):

Theorem 1.19: There is no natural number between 0 and 1.

Proof: Let $S = \{x \in \mathsf{N} : 0 < x < 1\}$, and assume by way of contradiction that $S \neq \emptyset$. Then, by the Well Ordering Principle, there is a least member ℓ of S. Since $\ell \in S$, $0 < \ell < 1$. Hence, by O4,

$$0 \cdot \ell < \ell \cdot \ell < 1 \cdot \ell;$$

furthermore, $0 \cdot \ell = 0$ (by M2 and Theorem 1.4), and $1 \cdot \ell = \ell$ (by M2 and M4). Thus, $0 < \ell \cdot \ell < \ell$. Combining this with the fact that $\ell < 1$, we have

$$(*)\ 0 < \ell \cdot \ell < \ell < 1.$$

Since $\ell \in S$, $\ell \in \mathsf{N}$. Hence, by our assumption that N is closed under multiplication (1.18), $\ell \cdot \ell \in \mathsf{N}$. Thus, since $0 < \ell \cdot \ell < 1$ (by (*)), $\ell \cdot \ell \in S$. Therefore, since $\ell \cdot \ell < \ell$ (by (*)), ℓ is not the least member of S. This is a contradiction to our choice of ℓ. ¥

We now prove an important consequence of the Well Ordering Principle.

Theorem 1.20 (Induction Principle): For each $n \in \mathsf{N}$, let P_n be a statement. If P_{1} is true and if $P_{n+\mathsf{1}}$ is true whenever P_n is true, then P_n is true for all $n \in \mathsf{N}$.

Proof: Let $S = \{n \in \mathsf{N} : P_n \text{ is false}\}$, and assume by way of contradiction that $S \neq \emptyset$. Then, by the Well Ordering Principle, there is a least member ℓ of S. Since P_{1} is true, $1 \notin S$; thus, $\ell \neq 1$. Also, since $\ell > 0$ (1.18), we see from Theorem 1.19 that $\ell \not< 1$. Hence, by O1, $\ell > 1$. Thus, $\ell + (-1) \in \mathsf{N}$ (1.18); also, $\ell + (-1) < \ell$ (1.18). Thus, since ℓ is the least member of S, $P_{\ell+(-1)}$ is true. Hence, by assumption in our theorem, $P_{[\ell+(-1)]+\mathsf{1}}$ is true (note: $P_{[\ell+(-1)]+\mathsf{1}}$ is indeed one of the statements in our theorem, for since $\ell + (-1) \in \mathsf{N}$ and since $1 \in \mathsf{N}$ (1.18), we know that $[\ell + (-1)] + 1 \in \mathsf{N}$ (1.18)). This says that P_ℓ is true since

$$[\ell + (-1)] + 1 \overset{\text{A3}}{=} \ell + [(-1) + 1] \overset{\text{A2}}{=} \ell + [1 + (-1)] \overset{\text{A5}}{=} \ell + 0 \overset{\text{A4}}{=} \ell.$$

Therefore, having proved that P_ℓ is true, we have that $\ell \notin S$. This establishes a contradiction. ¥

Recall our attempt at rigor in the second paragraph of the section. You can now answer the questions we asked there: I invite you to use the Induction Principle to define the set of natural numbers in the manner indicated.

We prove one more theorem about the natural numbers. First, we motivate the importance of the theorem by showing that it is needed in elementary situations.

We recall the proof that the sequence $\{\frac{\mathsf{1}}{n}\}_{n=\mathsf{1}}^{\infty}$ converges to 0 as presented in most calculus books (the proof that follows is taken verbatim from Edwards and Penney, *Calculus*, Prentice Hall, fifth edition, p. 627) :

"Suppose that we want to establish rigorously the intuitively evident fact that the sequence $\{\frac{\mathsf{1}}{n}\}_{n=\mathsf{1}}^{\infty}$ coverges to zero,

$$\lim_{n\to\infty} \tfrac{\mathsf{1}}{n} = 0.$$

Because $L = 0$ here, we only need to convince ourselves that to each positive number ϵ there corresponds an integer N such that

$$\left|\tfrac{\mathsf{1}}{n}\right| = \tfrac{\mathsf{1}}{n} < \epsilon \quad \text{if } n \geq N.$$

But evidently it suffices to choose any fixed integer $N > \frac{\mathsf{1}}{\epsilon}$. Then $n \geq N$ implies immediately that

$$\tfrac{\mathsf{1}}{n} \leq \tfrac{\mathsf{1}}{N} < \epsilon,$$

as desired (Fig. 11.2.3)."

The proof is "correct" but incomplete. In fact, the essential point is missed: How do you know there is an integer N such that $N > \frac{1}{\epsilon}$? We correct the deficiency with our next theorem, which gives an important property of the natural numbers. First, we prove a lemma; the lemma is obvious based on our experience with the natural numbers.

Lemma 1.21: N has no upper bound.

Proof: Suppose by way of contradiction that N has an upper bound. Then, since $\mathsf{N} \neq \emptyset$ (because $1 \in \mathsf{N}$ (1.18)), the Completeness Axiom says that N has a least upper bound ℓ. By Exercise 1.11, $\ell + (-1) < \ell$. Hence, $\ell + (-1)$ can not be an upper bound for N (since ℓ is the *least* upper bound for N). Thus, there exists $k \in \mathsf{N}$ such that $k \not\leq \ell + (-1)$; hence, by O1, $\ell + (-1) < k$. Hence, $\ell < k + (-(-1))$ (use O3, and A3 - A5). Thus, since $-(-1) = 1$ by Corollary 1.3, $\ell < k + 1$. Hence, by O1,

$$\ell \not\geq k + 1;$$

furthermore, $k + 1 \in \mathsf{N}$ (by 1.18 since $k \in \mathsf{N}$ and $1 \in \mathsf{N}$). Therefore, ℓ is not an upper bound of N. This is a contradition (since ℓ *is* an upper bound of N). ¥

Theorem 1.22 (Archimedean Property): If $x, y \in \mathsf{R^1}$ and $x > 0$, then there exists $n \in \mathsf{N}$ such that $y < n \cdot x$.

Proof: Since $x \in \mathsf{R^1}$ and $x \neq 0$ (by O1), $\frac{1}{x} \in \mathsf{R^1}$ (by M5). Thus, $y \cdot \frac{1}{x} \in \mathsf{R^1}$ (by M1). Hence, by Lemma 1.21, there exists $n \in \mathsf{N}$ such that $y \cdot \frac{1}{x} < n$ (we are also using O1 here). Thus, since $x > 0$, we have by O4 that

$$(y \cdot \tfrac{1}{x}) \cdot x < n \cdot x.$$

Therefore, since

$$(y \cdot \tfrac{1}{x}) \cdot x \overset{\text{M3}}{=} y \cdot (\tfrac{1}{x} \cdot x) \overset{\text{M2}}{=} y \cdot (x \cdot \tfrac{1}{x}) \overset{\text{M5}}{=} y \cdot 1 \overset{\text{M4}}{=} y,$$

we have that $y < n \cdot x$. ¥

Exercise 1.23: In line with the discussion preceding Lemma 1.21, prove that for any given $\epsilon > 0$, there exists $N \in \mathsf{N}$ such that $N > \frac{1}{\epsilon}$.

Let $\epsilon > 0$, and let $n_i \in \mathsf{N}$ for each $i \in \mathsf{N}$ such that $n_i < n_{i+1}$ for each i. Prove that there exists N such that $\frac{1}{n_i} < \epsilon$ for all $i \geq N$.

Exercise 1.24: 1 is the least upper bound of the open interval $(0, 1)$.

5. Proof That Nonnegative Numbers Have nth Roots

We presented the axioms for the real numbers in section 1 and made certain assumptions, which we take as axioms, about the natural numbers in section 4 (1.18). We claimed at the beginning of the chapter that all properties of real numbers that you are familiar with can be proved on the basis of the axioms in section 1 and the facts assumed about N in section 4 (1.18). We supported the

claim in section 3. We support the claim further here by proving an important theorem that everyone believes, but, perhaps, only from experience.

Let $a \geq 0$ and let n be a natural number. We are all familiar with the symbols $\sqrt[n]{a}$ and $a^{\frac{1}{n}}$ as representing the n^{th} root of a. From past experience, we are use to thinking of $\sqrt[n]{a}$ as a number. We show that $\sqrt[n]{a}$ is a number and that this result follows from the axioms we have given. The proof of the theorem is very long – about six pages. The reason for the length of the proof is not the difficulty of the proof, but rather the fact that we carry out the details of many computations in the proof. Our purpose for including detailed computations is to give you further assurance that manipulations with arithmetic and inequalities that you already know can be verified rigorously. The main parts of the proof are broken down into steps that will help you understand how the proof is organized.

For a real number $a \geq 0$ and a natural number n, we define an n^{th} *root of* a to be a nonnegative real number b such that $b^n = a$. We will see that there is only one n^{th} root of each number $a \geq 0$; thus, we can say *the* n^{th} root of a rather than *an* n^{th} root of a. We denote the n^{th} of a by $a^{\frac{1}{n}}$ or $\sqrt[n]{a}$. (A further discussion of roots of numbers is in section 4 of Chapter VIII, where we include odd roots of negative numbers.)

The proof of our theorem is based on the Completeness Axiom. At the beginning of section 1 we said that without the Completeness Axiom we are not guaranteed that $\sqrt{2}$ is a real number. The proof of our theorem shows that *with* the Completeness Axiom we can prove that $\sqrt{2}$ *is* a real number.

Theorem 1.25: For any real number $a \geq 0$, there is a unique real number $b \geq 0$ such that $b^n = a$. In other words, every nonnegative real number has a unique (nonnegative) n^{th} root.

Proof: We prove the theorem only for square roots. The proof for n^{th} roots uses similar ideas (with some computations aided by the Binomial Theorem (Theorem 21.41), whose proof you can read now). Another proof for n^{th} roots is in section 4 of Chapter VIII, but that proof depends on previous material.

We first dispense with the case when $a = 0$: By Theorem 1.4, 0 is a square root of 0 and, by Exercise 1.17 and O1, 0 is the only square root of 0.

Now, fix $a > 0$. The proof centers around considering the following set:

$$S = \{x \geq 0 : x^2 \leq a\}.$$

We show that S has a least upper bound b; then we show that b is the square root of a.

Step 1: Proof that S has a least upper bound b

First, note that $S \neq \emptyset$ since $0 \in S$ by Theorem 1.4.

Next, we show that $a + 1$ is an upper bound for S. Assume by way of contradiction that there is an $x \in S$ such that $x \not\leq a + 1$. Then, by O1, $a + 1 < x$. Hence,

$$0 < a \overset{1.10}{<} a + 1 < x.$$

Thus, $a+1>0$ and $x>0$ (by O2); therefore, since $a+1<x$, we see that

$$(a+1)^2 \overset{\text{O4}}{<} x\cdot(a+1) \overset{\text{M2}}{=} (a+1)\cdot x \overset{\text{O4}}{<} x^2;$$

also, using the Distributive Axiom (D) and various axioms in (A) and (M), we see that

$$(a+1)^2 = (a^2+1)+(a+a).$$

Hence, we have

(1) $(a^2+1)+(a+a)<x^2$.

We obtain a contradiction to our assumption that $x\in S$ by proving

(2) $a<x^2$.

Proof of (2)*:* Since $a^2>0$ by Exercise 1.17, $a^2+1>0$ by Corollary 1.10 and O2. Hence,

$$(a^2+1)+(a+a) \overset{\text{O3}}{>} 0+(a+a) \overset{\text{A2}}{=} (a+a)+0 \overset{\text{A4}}{=} (a+a);$$

furthermore, since $a>0$,

$$a+a \overset{\text{O3}}{>} 0+a \overset{\text{A2}}{=} a+0 \overset{\text{A4}}{=} a.$$

Combining the last two inequalities using O2, we obtain that

$$a<(a^2+1)+(a+a).$$

Therefore, by (1) and O2, $a<x^2$. This proves (2).

By (2) and O1, $x^2\not\leq a$. Hence, $x\notin S$. This is a contradiction. Therefore, $a+1$ is an upper bound for S.

We have proved that $S\neq\emptyset$ and that S has an upper bound. Therefore, by the Completeness Axiom, S has a least upper bound b. This completes Step 1.

We note the following fact for use several times: Since $n>0$ for all $n\in\mathsf{N}$ (1.18), we have from Exercise 1.14 that

(3) $\frac{1}{n}>0$ for all $n\in\mathsf{N}$.

Step 2: Proof that $b>0$

We prove that $b>0$ by finding a positive number in S that is smaller than b.

Since $a>0$, Theorem 1.22 says there exists $n\in\mathsf{N}$ such that $1<n\cdot a$. Thus, by (3) and O4, $1\cdot\frac{1}{n}<(n\cdot a)\cdot\frac{1}{n}$. Hence, $\frac{1}{n}<a$ (use M2 - M5). Thus, by (3) and O4, we have that

(4) $\left(\frac{1}{n}\right)^2<a\cdot\frac{1}{n}$.

We prove that the right - hand side of (4) is $\leq a$ as follows: Since $1\leq n$ (1.18) and $a>0$, we have by O4 that $1\cdot a\leq n\cdot a$. Hence, $a\leq a\cdot n$ (by M2 and M4). Thus, since $\frac{1}{n}>0$ (by (3)),

$$a \cdot \tfrac{1}{n} \overset{\text{O4}}{\leq} (a \cdot n) \cdot \tfrac{1}{n} \overset{\text{M3}}{=} a \cdot (n \cdot \tfrac{1}{n}) \overset{\text{M5}}{=} a \cdot 1 \overset{\text{M4}}{=} a,$$

which proves that $a \cdot (\frac{1}{n}) \leq a$.

Now, having proved that $a \cdot (\frac{1}{n}) \leq a$, we have by (4) and O2 that

$$\left(\tfrac{1}{n}\right)^2 < a.$$

Therefore, since $\frac{1}{n} > 0$ (by (3)) we have proved that $\frac{1}{n} \in S$. Thus, since b is an upper bound for S, $b \geq \frac{1}{n}$. Therefore, since $\frac{1}{n} > 0$, $b > 0$ (by O2). This completes Step 2.

We show that $b^2 = a$ by showing that $b^2 \not> a$ and that $b^2 \not< a$ (and then applying O1).

Step 3: Proof that $b^2 \not> a$

Since $b > 0$ by Step 2, $b + b \overset{\text{O3}}{>} 0 + b \overset{\text{A2}}{=} b + 0 \overset{\text{A4}}{=} b > 0$; therefore, by O2, we have that

(5) $b + b > 0$.

Now, suppose by way of contradiction that $b^2 > a$. Note from (5) and O1 that $b + b \neq 0$ and, hence, that $\frac{1}{b+b}$ is a number by M5. We prove that

(6) $\frac{b^2 + (-a)}{b+b} > 0$.

Proof of (6)*:* By our assumption that $b^2 > a$ (by assumption), we see that

$$b^2 + (-a) \overset{\text{O3}}{>} a + (-a) \overset{\text{A5}}{=} 0;$$

thus, since $\frac{1}{b+b} > 0$ (by (5) and Exercise 1.14), we have by O4 that

$$[b^2 + (-a)] \cdot \tfrac{1}{b+b} > 0 \cdot \tfrac{1}{b+b}.$$

Therefore, since $0 \cdot \frac{1}{b+b} = 0$ (by M2 and Theorem 1.4), we have proved (6).

By (6) and Theorem 1.22, there exists $k \in \mathsf{N}$ such that

$$1 < k \cdot \tfrac{b^2 + (-a)}{b+b}.$$

Thus, since $\frac{1}{k} > 0$ (by (3)), we see from O4 (using M2 - M5) that

(7) $\frac{1}{k} < \frac{b^2 + (-a)}{b+b}$.

We show that

(8) $b + (-\frac{1}{k}) > 0$.

Proof of (8)*:* Since $a > 0$,

$$a + b^2 \overset{\text{O3}}{>} 0 + b^2 \overset{\text{A2}}{=} b^2 + 0 \overset{\text{A4}}{=} b^2;$$

hence, by O3 and A2 - A5, $b^2 > b^2 + (-a)$. Thus, since $\frac{1}{b+b} > 0$ (by (5) and Exercise 1.14), we have by O4 that

(8i) $\frac{b^2+(-a)}{b+b} < \frac{b^2}{b+b}$;

Since $b > 0$ (by Step 2), $b + b \overset{\text{O3}}{>} 0 + b \overset{\text{A2}}{=} b + 0 \overset{\text{A4}}{=} b$; thus, again since $b > 0$, $(b+b) \cdot b \overset{\text{O4}}{>} b^2$. Therefore, since $\frac{1}{b+b} > 0$ (by (5) and Exercise 1.14), we have

$$[(b+b) \cdot b] \cdot \tfrac{1}{b+b} \overset{\text{O4}}{>} b^2 \cdot \tfrac{1}{b+b} = \tfrac{b^2}{b+b};$$

Thus, since $[(b+b) \cdot b] \cdot \frac{1}{b+b} = b$ (by M2 - M5),

$$b > \tfrac{b^2}{b+b}.$$

Hence, by (8i) and O2, $\frac{b^2+(-a)}{b+b} < b$. Thus, by (7) and O2, $\frac{1}{k} < b$. Hence,

$$\tfrac{1}{k} + (-\tfrac{1}{k}) \overset{\text{O3}}{<} b + (-\tfrac{1}{k});$$

Therefore, by A5, $0 < b + (-\frac{1}{k})$. This proves (8).

Next, we show that

(9) $\left(b + (-\frac{1}{k})\right)^2 > a$.

Proof of (9)*:* Since $b + b > 0$ (by (5)), we see from (7) and O4 that

$$\tfrac{1}{k} \cdot (b+b) < \tfrac{b^2+(-a)}{b+b} \cdot (b+b);$$

furthermore, using M2 - M5, we see that

$$\tfrac{b^2+(-a)}{b+b} \cdot (b+b) = b^2 + (-a).$$

Hence,

$$\tfrac{1}{k} \cdot (b+b) < b^2 + (-a).$$

Thus, since $\frac{1}{k} \cdot (b+b) \overset{\text{M2}}{=} (b+b) \cdot \frac{1}{k} = \frac{b+b}{k}$, we have

$$\tfrac{b+b}{k} + [a + (-\tfrac{b+b}{k})] \overset{\text{O3}}{<} [b^2 + (-a)] + [a + (-\tfrac{b+b}{k})].$$

Hence, using A2 - A5, we see that

(9i) $a < b^2 + (-\frac{b+b}{k})$.

Since $\frac{1}{k} > 0$ (by (3)), $\frac{1}{k} \neq 0$ (by O1); hence, $\left(\frac{1}{k}\right)^2 > 0$ (by Exercise 1.17). Thus,

$$\left(\tfrac{1}{k}\right)^2 + [b^2 + (-\tfrac{b+b}{k})] \overset{\text{O3}}{>} 0 + [b^2 + (-\tfrac{b+b}{k})];$$

hence, by A2 and A4,

$$\left(\tfrac{1}{k}\right)^2 + [b^2 + (-\tfrac{b+b}{k})] > b^2 + (-\tfrac{b+b}{k});$$

thus, by (9i) and O2, we have that

(9ii) $\left(\tfrac{1}{k}\right)^2 + [b^2 + (-\tfrac{b+b}{k})] > a.$

We show that the left-hand side of (9ii) is $\left(b + (-\tfrac{1}{k})\right)^2$, which completes the proof of (9). The computations that follow are tedious, but are done so that you can see that what we prove depends only on the axioms (and some previous theorems). We note that when we change from $\frac{x}{y}$ to $x \cdot \frac{1}{y}$ or vice versa, this change is justified by the notation in section 2; in particular, the change does not use an axiom or a theorem. Now,

$$\left(b + (-\tfrac{1}{k})\right)^2 \overset{\text{D, M2}}{=} \left[b^2 + (b \cdot \left(-\tfrac{1}{k}\right))\right] + \left[(b \cdot \left(-\tfrac{1}{k}\right)) + \left(-\tfrac{1}{k}\right)^2\right]$$

$$\overset{1.5}{=} \left[b^2 + \{b \cdot \left((-1) \cdot \tfrac{1}{k}\right)\}\right] + \left[\{b \cdot \left((-1) \cdot \tfrac{1}{k}\right)\} + \left((-1) \cdot \tfrac{1}{k}\right)^2\right]$$

$$\overset{\text{A3, M2, M3}}{=} b^2 + \left[\left(((-1) \cdot \tfrac{b}{k}) + ((-1) \cdot \tfrac{b}{k})\right) + \left(\{(-1) \cdot (-1)\} \cdot \{\tfrac{1}{k}\}^2\right)\right]$$

$$\overset{1.5,\ 1.3}{=} b^2 + \left[\left(((-1) \cdot \tfrac{b}{k}) + ((-1) \cdot \tfrac{b}{k})\right) + \left(1 \cdot \{\tfrac{1}{k}\}^2\right)\right]$$

$$\overset{\text{D, M2}}{=} b^2 + \left[\{(-1) \cdot (\tfrac{b}{k} + \tfrac{b}{k})\} + (\{\tfrac{1}{k}\}^2 \cdot 1)\right]$$

$$\overset{1.8,\ \text{M4}}{=} b^2 + \left[\{(-1) \cdot \tfrac{(k \cdot b) + (k \cdot b)}{k^2}\} + \{\tfrac{1}{k}\}^2\right]$$

$$\overset{\text{D},\ 1.5}{=} b^2 + [(-\tfrac{k \cdot (b+b)}{k^2}) + \left(\tfrac{1}{k}\right)^2] \overset{1.15}{=} b^2 + [-(\tfrac{k}{k} \cdot \tfrac{b+b}{k}) + \left(\tfrac{1}{k}\right)^2]$$

$$\overset{\text{M5, M4}}{=} b^2 + [-\tfrac{b+b}{k} + \left(\tfrac{1}{k}\right)^2] \overset{\text{A2, A3}}{=} \left(\tfrac{1}{k}\right)^2 + [b^2 + (-\tfrac{b+b}{k})].$$

Therefore, as remaked after (9ii), we have proved (9).

We now show that

(10) $b + (-\frac{1}{k})$ is an upper bound for S.

Proof of (10)*:* Let $y \in \mathsf{R}^1$ such that $y > b + (-\frac{1}{k})$. We show that $y \notin S$ (which proves (10) contrapositively by O1).

Since $y > b + (-\frac{1}{k})$ and $b + (-\frac{1}{k}) > 0$ (by (8)), we see from O2 that $y > 0$. Thus, since $y > b + (-\frac{1}{k})$, we have that

$$y^2 \overset{04}{>} (b + (-\tfrac{1}{k})) \cdot y \overset{\text{M2}}{=} y \cdot (b + (-\tfrac{1}{k}));$$

also, since $b + (-\frac{1}{k}) > 0$ (by (8)) and $y > b + (-\frac{1}{k})$, we have by O4 that

$$y \cdot (b + (-\tfrac{1}{k})) > \left((b + (-\tfrac{1}{k})\right)^2.$$

Therefore, applying O2 to the two inequalities above, we have

$$y^2 > \left((b + (-\tfrac{1}{k})\right)^2.$$

Hence, by (9) and O2, $y^2 > a$. Thus, $y^2 \not\leq a$ (by O1). Hence, $y \notin S$. Therefore, by the comment at the beginning of the proof of (10), we have proved (10).

Finally, we complete the proof for Step 3. Specifically, we obtain a contradiction to b being the *least* upper bound for S by showing that $b + (-\frac{1}{k}) < b$ and applying (10).

Since $\frac{1}{k} > 0$ by (3), $-\frac{1}{k} < 0$ by Exercise 1.14. Hence, by O3, $(-\frac{1}{k})+b < 0+b$; thus, by A2 and A4, $b + (-\frac{1}{k}) < b$. Hence, by O1, $b \not\leq b + (-\frac{1}{k})$. Therefore, by (10), b is not the *least* upper bound for S. This is a contradiction. The contradiction is the result of our supposition near the beginning of the proof of Step 3 that $b^2 > a$. Therefore, $b^2 \not> a$. This completes Step 3.

Step 4: Proof that $b^2 \not< a$

We omit references to the use of axioms in section 1, some theorems and exercises in section 3, and the assumptions in 1.18. In other words, we use the familiar "rules" of arithmetic without reference; we invite the reader to fill in the details (which are similar to the details in Step 3).

Suppose by way of contradiction that $b^2 < a$. Then, since $\frac{a-b^2}{(b+b)+1} > 0$, we see from Theorem 1.22 that there exists $m \in \mathsf{N}$ such that

$$1 < m \cdot \tfrac{a-b^2}{(b+b)+1}.$$

Hence, it follows that

(11) $b^2 + \frac{(b+b)+1}{m} < a.$

Now, note that

$$(b + \tfrac{1}{m})^2 = b^2 + \tfrac{b+b}{m} + (\tfrac{1}{m})^2 \leq b^2 + \tfrac{b+b}{m} + \tfrac{1}{m} = b^2 + \tfrac{(b+b)+1}{m}.$$

Hence, by (11),

$$(b + \tfrac{1}{m})^2 < a;$$

furthermore, $b + \frac{1}{m} > 0$ (since b is an upper bound of S and $0 \in S$). Therefore, $b + \frac{1}{m} \in S$. However, since $b < b + \frac{1}{m}$, this contradicts that b is an upper bound for S. The contradiction comes from our assumption that $b^2 < a$. Therefore, $b^2 \not< a$. This completes Step 4.

Step 5: Completing the proof

We know from Steps 3 and 4 that $b^2 \not> a$, and that $b^2 \not< a$. Therefore, by O1, $b^2 = a$. By Step 2, $b > 0$. Therefore, it only remains to prove the uniqueness part of our theorem, that is, that b is the only nonnegative number such that $b^2 = a$. As in the proof of Step 4, we omit references to the use of the axioms in section 1, etc.

Assume that $c \geq 0$ and that $c^2 = a$. We show that $b = c$. Since $b^2 = c^2$, $b^2 + (-[c^2]) = 0$. Thus, since $b^2 + (-[c^2]) = (b + c)(b + [-c])$, we have that

$$(b + c)(b + [-c]) = 0.$$

Thus, since $b + c \neq 0$ (recall from Step 2 that $b > 0$), we have by Theorem 1.6 that $b + [-c] = 0$. Therefore, $b = c$. ¥

6. The Betweenness Property for the Rational Numbers

The *rational numbers* are the numbers that can be written in the form $\frac{m}{n}$ or $-\frac{m}{n}$, where $m, n \in \mathsf{N}$, together with the number 0. We denote the set of all rational numbers by Q.

We prove a fundamental result about the rational numbers: There is a rational number between any two (different) real numbers.

Theorem 1.26: If $a, b \in \mathsf{R}^1$ such that $a < b$, then there is a rational number r such that $a < r < b$.

Proof: As we did in the latter part of the proof of Theorem 1.25, we omit references to the use of axioms in section 1, some theorems and exercises in section 3, and assumptions in 1.18 (except when we use the Well Ordering Principle).

The theorem is obvious when $a < 0 < b$ (take $r = 0$). Thus, there are only two cases to consider: $a \geq 0$ and $b \leq 0$.

Assume first that $a \geq 0$. Since $a < b$, $b + (-a) > 0$. Hence,by Theorem 1.22, there exists $n \in \mathsf{N}$ such that $1 < n \cdot (b + (-a))$. (To envision the proof that follows, rewrite the inequality as $\frac{1}{n} < b + (-a)$.)

Let

$$S = \{k \in \mathsf{N} : a < \tfrac{k}{n}\}.$$

By Lemma 1.21, $n \cdot a < k$ for some $k \in \mathsf{N}$; thus, $S \neq \emptyset$. Hence, by the Well Ordering Principle for N (1.18), S has a least member ℓ. Since $\ell, n \in \mathsf{N}$, $\frac{\ell}{n} \in \mathsf{Q}$.

Since $\ell \in S$, $a < \frac{\ell}{n}$. We complete the proof for the case when $a \geq 0$ by proving that

$$(*)\ \tfrac{\ell}{n} < b.$$

Proof of (*)*:* We first prove that

(1) $\frac{\ell + (-1)}{n} \leq a$.

Proof of (1)*:* Suppose by way of contradiction that $\frac{\ell + (-1)}{n} > a$. Then, since $n > 0$, $\ell + (-1) > n \cdot a$. Thus, since $n \cdot a \geq 0$ (here is where we use that $a \geq 0$), $\ell + (-1) > 0$. Hence, $\ell > 1$. Thus, since $\ell \in \mathsf{N}$, $\ell + (-1) \in \mathsf{N}$. Therefore, since $\frac{\ell + (-1)}{n} > a$, we have that $\ell + (-1) \in S$; however, since $\ell + (-1) < \ell$ (by Exercise 1.11), this contradicts the fact that ℓ is the least member of S. Therefore, we have proved (1).

Now, to prove (*), note from (1) that $\frac{\ell}{n} + (-\frac{1}{n}) \leq a$. Hence, $\frac{\ell}{n} \leq a + \frac{1}{n}$. Also, $\frac{1}{n} < b + (-a)$ by our choice of n. Therefore,

$$\frac{\ell}{n} \leq a + \frac{1}{n} < a + (b + (-a)) = b.$$

This proves (*) and, therefore, completes the proof of our theorem for the case when $a \geq 0$.

Finally, consider the case when $b \leq 0$. Then $0 \leq -b < -a$; hence, having already proved the theorem for this type of situation, there exists $q \in \mathsf{Q}$ such that $-b < q < -a$. Therefore, $a < -q < b$ and $-q \in \mathsf{Q}$. ¥

Exercise 1.27: If $a, b \in \mathsf{R}^1$ such that $a < b$, then there are infinitely many rational numbers in the open interval (a, b).

7. Absolute value

For any number a, the *absolute value of* a is denoted by $|a|$ and is defined by

$$|a| = \begin{cases} a & \text{, if } a \geq 0 \\ -a & \text{, if } a < 0. \end{cases}$$

Intuitively, $|a|$ is how far a is from the origin 0 (without regard to whether a is positive or negative). Thus, $|a + (-b)|$ can be thought of as the distance between a and b. Therefore, we call $|a + (-b)|$ the *distance between a and b* (or the *distance from a to b*).

We note four basic properties of absolute value (the reader should supply proofs that the properties hold). The properties can be used to show that the function that assigns to an ordered pair (a, b) of real numbers the number $|a + (-b)|$ satisfies the general definition of a distance function (see Exercise 1.30).

1. $|a| \geq 0$ for all $a \in \mathsf{R}^1$.
2. $|a| = 0$ if and only if $a = 0$.
3. $|a \cdot b| = |a| \cdot |b|$ for all $a, b \in \mathsf{R}^1$.
4. $|a + b| \leq |a| + |b|$ for all $a, b \in \mathsf{R}^1$ (Triangle Inequality).

Exercise 1.28: If $a \geq 0$, then $|x| \leq a$ if and only if $-a \leq x \leq a$.

Exercise 1.29: $\big||a| - |b|\big| \leq |a - b|$ for all $a, b \in \mathsf{R}^1$.

Exercise 1.30: Let X be a set. A *distance function* (or *metric*) *for* X is a real-valued function d defined on the Cartesian product $X \times X$ that satisfies the following four conditions:

(1) $d(x, y) \geq 0$ for all $x, y \in X$;

(2) $d(x, y) = 0$ if and only if $x = y$;

(3) $d(x, y) = d(y, x)$ for all $x, y \in X$ (Symmetry);

(4) $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$ (Triangle Inequality).

Define $d(a, b) = |a + (-b)|$ for all $a, b \in \mathsf{R}^1$. Prove that d is a distance function (in the general sense just defined).

8. Concluding Comments

The reader should now be convinced that familiar manipulations and properties of the real numbers can be verified on the basis of the axioms for the real numbers and various theorems in previous sections. Thus, we will no longer refer to many of the axioms and theorems when we use them. There are some exceptions: We will refer to the Completeness Axiom, the Well Ordering Principle, the Induction Principle and the Archimedean Principle when we use them.

From now on we use all the usual notation associated with arithmetic. For example, we write $a - b$ to mean $a + (-b)$ and ab to mean $a \cdot b$. In an expression involving combinations of addition and multiplication, we assume that the juxtaposed multiplications are carried out first; for example, $ab + c$ means $(ab) + c$.

We use the usual symbols for specific numbers. In particular, we assume the natural numbers are the numbers $1, 2, 3, ...$ with their familiar properties.

We assume standard arithmetic – the multiplication tables, long division, etc. – without proof.

To summarize, having taken a bath, we have now drained away all the dirty water!

We conclude by stating a dual to the Completeness Axiom.

The Completeness Axiom stated in section 1 is frequently called the Least Upper Bound Axiom. We will often have occasion to use an equivalent formulation of the axiom called the Greatest Lower Bound Axiom, which we state after giving relevant terminology.

Let $A \subset \mathsf{R}^1$. A *lower bound for* A is a number x such that $x \leq a$ for all $a \in A$. A *greatest lower bound for* A is a lower bound g for A such that $g \geq x$ for all lower bounds x for A.

Greatest Lower Bound Axiom. If A is a nonempty subset of R^1 such that A has an lower bound, then A has a greatest lower bound.

It is easy to prove the Greatest Lower Bound Axiom is equivalent to the Completeness Axiom in section 1. Thus, from the point of view of our development, the Greatest Lower Bound Axiom is actually a theorem.

The analogues of remarks we made about the Completeness Axiom at the end of section 1 apply to the Greatest Lower Bound Axiom. In particular, a nonempty set with a lower bound has *only one* greatest lower bound.

Notation:

- lub, **sup:** We denote the least upper bound of a set A by $lub\, A$ or by $\sup A$ (sup stands for supremum).
- glb, **inf:** We denote the greatest lower bound of a set A by $glb\, A$ or by $\inf A$ (inf stands for infimum).
- **max**, **min:** We sometimes use $\max A$ and $\min A$ to stand for $\sup A$ and $glb\, A$, respectively. (We usually use max and min only when we know the set A is finite.)

Chapter II: The Notion of Arbitrary Closeness

In the first two sections we study the notion of arbitrary closeness. We then show that the notion leads in a natural intuitive way to the idea of continuous functions.

1. Introduction to Arbitrary Closeness

We define what it means for a real number to be arbitrarily close to a nonempty set of real numbers. Then we present some examples and two basic theorems.

If p is a real number and A is a nonempty set of real numbers, then the *infimum distance of* p *to* A, denoted by $dist(p, A)$, is defined by

$$dist(p, A) = glb\,\{|p - a| : a \in A\}.$$

Definition. Let $p \in \mathsf{R}^1$ and let $A \subset \mathsf{R}^1$ such that $A \neq \emptyset$. We say that p *is arbitrarily close to* A, denoted by writing $p \sim A$, provided that $A \neq \emptyset$ and $dist(p, A) = 0$.

We write $p \not\sim A$ to mean that the number p is not arbitrarily close to the set A.

Example 2.1: $0 \sim (0, 1)$, $p \sim (0, 1)$ if $p \in (0, 1)$, and $2 \not\sim (0, 1)$.

Example 2.2: Every real number is arbitrarily close to the set Q of rational numbers.

Theorem 2.3: Let $p \in \mathsf{R}^1$ and let $A \subset \mathsf{R}^1$ such that $A \neq \emptyset$. Then $p \sim A$ if and only if for each open interval I such that $p \in I$, $I \cap A \neq \emptyset$.

Proof: Assume that there is an open interval $I = (a, b)$ such that $p \in I$ and $I \cap A = \emptyset$. Then

$$A \subset (-\infty, a] \cup [b, \infty);$$

thus, since $a < p < b$,

$$dist(p, A) \geq \min\{p - a, b - p\} > 0.$$

Therefore, $p \not\sim A$.

Conversely, assume that $p \not\sim A$. Then $dist(p, A) > 0$. Let J be the open interval given by

$$J = (p - dist(p, A), p + dist(p, A)).$$

Then $p \in J$ since $dist(p, A) > 0$. In addition, $J \cap A = \emptyset$ since if $x \in J$, then $|p - x| < dist(p, A)$ and, therefore, $x \notin A$. ¥

Lemma 2.4: Let $p \in \mathsf{R}^1$ and let (a, b) be an open interval such that $p \in (a, b)$. Then there exists $\epsilon > 0$ such that $(p - \epsilon, p + \epsilon) \subset (a, b)$.

Proof: Let

$$\epsilon = \min\{p - a, b - p\}.$$

Since $p \in (a, b)$, $\epsilon > 0$. Since $\epsilon \leq p - a$, $a \leq p - \epsilon$ and, since $\epsilon \leq b - p$, $p + \epsilon \leq b$; therefore, $(p - \epsilon, p + \epsilon) \subset (a, b)$. ¥

Theorem 2.5: Let $p \in \mathsf{R}^1$ and let $A \subset \mathsf{R}^1$ such that $A \neq \emptyset$. Then $p \sim A$ if and only if for each $\epsilon > 0$, $(p - \epsilon, p + \epsilon) \cap A \neq \emptyset$.

Proof: It follows easily from Lemma 2.4 that the condition involving ϵ here is equivalent to the condition involving I in Theorem 2.3. Therefore, Theorem 2.5 follows from (and is actually a reformulation of) Theorem 2.3. ¥

2. The Set of Points Arbitrarily Close to a Set

If A is a nonempty set of real numbers, we let $A^{\sim}$ denote the set of all real numbers that are arbitrarily close to A; in other words,

$$A^{\sim} = \{x \in \mathsf{R}^1 : x \sim A\}.$$

It is convenient to extend the notation to the empty set $\emptyset$ by making the intuitively reasonable assumption that no real number is arbitrarily close to the empty set; in symbols, $\emptyset^{\sim} = \emptyset$.

Exercise 2.6: For any open interval (a, b), what is $(a, b)^{\sim}$? (As for all exercises, prove that your answer is correct).

Theorem 2.7: For any $A \subset \mathsf{R}^1$, $A \subset A^{\sim}$.

Proof: Since $\emptyset^{\sim} = \emptyset$ (by definition), the theorem is true when $A = \emptyset$. So, assume that $A \neq \emptyset$. Let $p \in A$. If I is an open interval such that $p \in I$, then $p \in I \cap A$ and, hence, $I \cap A \neq \emptyset$. Therefore, by Theorem 2.3, $p \in A^{\sim}$. ¥

Example 2.8: If A is a finite subset of R^1, then $A^{\sim} = A$.

Example 2.9: $\mathsf{N}^{\sim} = \mathsf{N}$; if $A = \{\frac{1}{n} : n \in \mathsf{N}\}$, then $A^{\sim} = A \cup \{0\}$.

Exercise 2.10: If $A \subset B$, then $A^{\sim} \subset B^{\sim}$.

Theorem 2.11: For any $A, B \subset \mathsf{R}^1$, $(A \cup B)^{\sim} = A^{\sim} \cup B^{\sim}$.

Proof: Since $A \subset A \cup B$, $A^{\sim} \subset (A \cup B)^{\sim}$ by Exercise 2.10; similarly, $B^{\sim} \subset (A \cup B)^{\sim}$. Therefore, $A^{\sim} \cup B^{\sim} \subset (A \cup B)^{\sim}$.

We prove the reverse containment, namely, $(A \cup B)^{\sim} \subset A^{\sim} \cup B^{\sim}$.

Assume first that $A = \emptyset$. Then $A \cup B = B$ and, hence,

$$(A \cup B)^{\sim} = B^{\sim} = \emptyset^{\sim} \cup B^{\sim} = A^{\sim} \cup B^{\sim}.$$

Similarly, if $B = \emptyset$, then $(A \cup B)^{\sim} = A^{\sim} \cup B^{\sim}$. This proves that if $A = \emptyset$ or $B = \emptyset$, then $(A \cup B)^{\sim} = A^{\sim} \cup B^{\sim}$.

So, we assume from now on that $A \neq \emptyset$ and $B \neq \emptyset$. We prove that $(A \cup B)^{\sim} \subset A^{\sim} \cup B^{\sim}$ with a contrapositive argument (a direct argument can not be done with the present methods: see Exercise 2.12).

Assume that $p \in \mathsf{R}^1$ such that $p \notin A^{\sim} \cup B^{\sim}$. Then, by Theorem 2.5, there exist $\epsilon_1, \epsilon_2 > 0$ such that

$$(p-\epsilon_1, p+\epsilon_1) \cap A = \emptyset \quad \text{and} \quad (p-\epsilon_2, p+\epsilon_2) \cap B = \emptyset.$$

Hence, letting $\epsilon = \min\{\epsilon_1, \epsilon_2\}$, we see that $\epsilon > 0$ and that

$$(p-\epsilon, p+\epsilon) \cap (A \cup B) = \emptyset.$$

Therefore, by Theorem 2.5, $p \notin (A \cup B)^{\sim}$. ¥

Exercise 2.12: Concerning a comment in the proof of Theorem 2.11, find the flaw in the following direct argument for $(A \cup B)^{\sim} \subset A^{\sim} \cup B^{\sim}$:

As in the proof of Theorem 2.11, we can assume that $A \neq \emptyset$ and $B \neq \emptyset$. Now, let $p \in (A \cup B)^{\sim}$. Then, by Theorem 2.5, $(p-\epsilon, p+\epsilon) \cap (A \cup B) \neq \emptyset$ for each $\epsilon > 0$. Hence,

$$[(p-\epsilon, p+\epsilon) \cap A] \cup [(p-\epsilon, p+\epsilon) \cap B] \neq \emptyset \text{ for each } \epsilon > 0.$$

Thus, $(p-\epsilon, p+\epsilon) \cap A \neq \emptyset$ or $(p-\epsilon, p+\epsilon) \cap B \neq \emptyset$ for each $\epsilon > 0$. Hence, by Theorem 2.5, $p \in A^{\sim}$ or $p \in B^{\sim}$. Therefore, $p \in A^{\sim} \cup B^{\sim}$.

Exercise 2.13: If $A_1, A_2, ..., A_n$ are finitely many subsets of R^1, then

$$(\cup_{i=1}^{n} A_i)^{\sim} = \cup_{i=1}^{n} A_i^{\sim}.$$

Exercise 2.14: Would the result in Exercise 2.13 remain true for infinitely many subsets of R^1? In other words, if $\{A_i : i \in \mathcal{I}\}$ is an infinite collection of subsets of R^1, then is it true that

$$(\cup\{A_i : i \in \mathcal{I}\})^{\sim} = \cup\{A_i^{\sim} : i \in \mathcal{I}\})\,?$$

Theorem 2.15: For any $A \subset \mathsf{R}^1$, $(A^{\sim})^{\sim} = A^{\sim}$.

Proof: By Theorem 2.7, $A \subset A^{\sim}$. Therefore, by Exercise 2.10, $A^{\sim} \subset (A^{\sim})^{\sim}$.

To prove the reverse containment, first note that $(A^{\sim})^{\sim} \subset A^{\sim}$ if $A = \emptyset$ (since $\emptyset^{\sim} = \emptyset$); hence, we assume for the proof that $A \neq \emptyset$. Thus, $A^{\sim} \neq \emptyset$ by Theorem 2.7. Now, let $p \in (A^{\sim})^{\sim}$. Let I be an open interval such that $p \in I$. Then, since $A^{\sim} \neq \emptyset$ and $p \in (A^{\sim})^{\sim}$, $I \cap A^{\sim} \neq \emptyset$ by Theorem 2.3. Hence, there exists a point $q \in I \cap A^{\sim}$. Thus, since $A \neq \emptyset$, $I \cap A \neq \emptyset$ by Theorem 2.3. We have proved that $I \cap A \neq \emptyset$ for any open interval I such that $p \in I$. Therefore, again by Theorem 2.3, $p \in A^{\sim}$. ¥

Exercise 2.16: If $A_1, A_2, ..., A_n$ are finitely many subsets of R^1, then

$$(\cap_{i=1}^{n} A_i)^{\sim} \subset \cap_{i=1}^{n} A_i^{\sim}.$$

Exercise 2.17: Give an example of two subsets A and B of R^1 such that $(A \cap B)^{\sim} \neq A^{\sim} \cap B^{\sim}$.

Exercise 2.18: Would the result in Exercise 2.16 remain true for infinitely many subsets of R^1?

3. The Definition of Continuity

You have surely had some experience with the idea of a continuous function; based on your experience, you know the intuitive meaning of continuity – a continuous function is a function that does not jump. Did you ever stop and try to figure out what it really means to say that a function does not jump? Let us examine this idea.

In general, a question asked in a negative way is harder to deal with than the corresponding question posed in the positive way. So, we ask *What does it mean for a function to jump*? If a function jumps, it seems reasonable that it must jump at some point in its domain. Thus, we ask *What does it mean for a function to jump at a point p in the domain of the function*? Certainly, everyone has an instinctive feeling – some mental picture – for what this means. Let us consider an example that everyone will agree is stereotypical of the (intuitive) idea of a function jumping at p:

Example 2.19: Define $f : \mathsf{R}^1 \to \mathsf{R}^1$ by

$$f(x) = \begin{cases} 0 & \text{, if } x \leq 0 \\ 1 & \text{, if } x > 0. \end{cases}$$

The function f jumps at $p = 0$. Surely you agree. But what is the underlying reason you agree? The reason is that if you look at positive numbers that are as close as you like to 0, but not equal to 0, their values under f are one unit away from $f(0)$.

Let us look at another example, one that is more complicated than the previous one.

Example 2.20: Define $f : \mathsf{R}^1 \to \mathsf{R}^1$ by

$$f(x) = \begin{cases} 0 & \text{, if } x \text{ is rational} \\ 1 & \text{, if } x \text{ is irrational.} \end{cases}$$

The function seems to jump at every point p. Why? If p is irrational, then you know from Theorem 1.26 that there rationals as close to p as you like, and the value of f at each rational is one unit away from $f(p)$. If p is rational, then (by the natural analogue of Theorem 1.26 for irrationals) there irrationals as close to p as you like, and the value of f at each irrational is one unit away from $f(p)$.

The two examples shed light on what it means for a function to jump at p. One need only observe the common thread in the two examples: The function f in each example jumps at p because there is a set A such that p is arbitrarily close to A but $f(p)$ is *not* arbitrarily close to $f(A)$. Notice that we say the condition holds for *some* set A, not for every set. Indeed, there are some sets A in the examples such that p is arbitrarily close to A *and* $f(p)$ *is arbitrarily close to* $f(A)$. You can see this by taking A to be any set containing p in both examples or, as is more illustrative, by taking $A = (-\infty, 0)$ in Example 2.19 and by taking A to be the set of all rationals except p when p is rational in Example 2.20.

Our discussion suggests the following definition:

Definition. Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$, and let $p \in X$. We say that f *jumps at* p provided that there is a subset, A, of X such that $p \sim A$ but $f(p) \not\sim f(A)$.

Exercise 2.21: Define $f : \mathsf{R}^1 \to \mathsf{R}^1$ by

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

Then f jumps at 0.

Exercise 2.22: Define $f : \mathsf{R}^1 \to \mathsf{R}^1$ by

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

Then f jumps at 0.

We do not claim that our definition for a function to jump at a point is "correct" – that can only be ascertained by checking numerous examples to see if the definition fits our intuition, and by seeing if the definition leads to appropriate theoretical developments. At this point, we accept the definition and use it to define continuity which, after all, is why we wanted the definition in the first place.

Definition. Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$, and let $p \in X$. We say that f is *continuous at* p provided that f does not jump at p. In other words, f is *continuous at* p provided that whenever $A \subset X$ and $p \sim A$, then $f(p) \sim f(A)$.

We say that f is *continuous on* X (or just *continuous* when the domain X is clear) provided that f does not jump at any point of X.

A simple kind of function that we know from past experience is continuous is a function whose graph is a straight line. We show this kind of function is continuous in the sense of the definition above. Thus, the example lends credibility to our definition of continuity.

Example 2.23: Fix $m, b \in \mathsf{R}^1$, and let $f : \mathsf{R}^1 \to \mathsf{R}^1$ be given by

$$f(x) = mx + b, \quad \text{all } x \in \mathsf{R}^1.$$

The function f is continuous.

To prove this, let $p \in \mathsf{R}^1$ and let $A \subset \mathsf{R}^1$ such that $p \sim A$.

If $m = 0$, then $f(p) = b$ and $f(A) = \{b\}$; thus, since $b \sim \{b\}$ by Theorem 2.7, we have that $f(p) \sim f(A)$. This proves that f is continuous at p when $m = 0$.

Next, assume that $m > 0$. We show that $f(p) \sim f(A)$ by using Theorem 2.3. For this purpose, let $I = (a, c)$ be an open interval such that $f(p) \in I$. Let J be the open interval defined by

$$J = (\tfrac{a-b}{m}, \tfrac{c-b}{m}).$$

We see that $p \in J$ as follows: Since $f(p) \in I$, $a < mp+b < c$; thus, since $m > 0$, $\frac{a-b}{m} < p < \frac{c-b}{m}$, so $p \in J$. Therefore, since $p \sim A$, we have by Theorem 2.3 that there is a point $x \in J \cap A$. Since $x \in J$, $\frac{a-b}{m} < x < \frac{c-b}{m}$; thus, since $m > 0$,

$$a < mx + b < c;$$

hence, $f(x) \in I$. Also, since $x \in A$, $f(x) \in f(A)$. Hence, $f(x) \in I \cap f(A)$. This proves that any open interval containing $f(p)$ has a nonempty intersection with $f(A)$. Thus, by Theorem 2.3, $f(p) \sim f(A)$. Therefore, we have proved that f is continuous at p when $m > 0$.

Finally, assume that $m < 0$. Then the proof that f is continuous at p is similar to the proof when $m > 0$: Let I be as before, redefine J to be the interval $(\frac{c-b}{m}, \frac{a-b}{m})$, and make the obvious changes necessitated by the assumption that $m < 0$.

We prove in the next chapter that our definition of continuity is correct in the sense that it is equivalent to the definition of continuity that you have (probably) seen in your study of calculus. So, why did we define continuity as we did? The answer is purely philosophical: *We adhere to the principle that a definition should convey as much as possible the fundamental idea behind the notion being defined.*

Exercise 2.24: Define $f : \mathsf{R}^1 \to \mathsf{R}^1$ by letting $f(x) = x^2$. Then f is continuous.

Exercise 2.25: Define $f : \mathsf{R}^1 - \{0\} \to \mathsf{R}^1$ by letting $f(x) = \frac{1}{x}$. Then f is continuous.

Exercise 2.26: Define $f : [0,1] \to \mathsf{R}^1$ by

$$f(x) = \begin{cases} x & \text{, if } x = \frac{1}{n} \text{ for some } n \in \mathsf{N} \\ 0 & \text{, otherwise.} \end{cases}$$

At what points p is f continuous?

Exercise 2.27: Any function $f : \mathsf{N} \to \mathsf{R}^1$ is continuous.

4. Limit Points and Isolated Points

Limit points and isolated points of sets will be important in our discussion of limits in the next chapter.

Definition. Let $X \subset \mathsf{R}^1$. A point $p \in \mathsf{R}^1$ is called a *limit point of* X provided that $p \sim X - \{p\}$ [2]. A point of X that is not a limit point of X is called an *isolated point of* X.

We let X^ℓ denote the set of all limit points of X.

[2] $\mathsf{X} - \{\mathsf{p}\}$ denotes all the points of X except p (if $\mathsf{p} \notin \mathsf{X}$, obviously $\mathsf{X} - \{\mathsf{p}\} = \mathsf{X}$). More generally, for any two sets A and B, $\mathsf{A} - \mathsf{B} = \{\mathsf{x} \in \mathsf{A} : \mathsf{x} \notin \mathsf{B}\}$; the set $\mathsf{A} - \mathsf{B}$ is called the *complement of* B *in* A.

Exercise 2.28: What are the limit points of $\{15\}$? What are the limit points of the interval $(0,1)$? What are the limit points of Q? What are the limit points of $X = \{\frac{1}{n} : n \in \mathsf{N}\}$?

Exercise 2.29: For any $A \subset \mathsf{R}^1$, $A^{\sim} = A \cup A^{\ell}$.

Exercise 2.30: If $A, B \subset \mathsf{R}^1$ such that $A \subset B$, then $A^{\ell} \subset B^{\ell}$.

Exercise 2.31: For any $A, B \subset \mathsf{R}^1$, $(A \cup B)^{\ell} = A^{\ell} \cup B^{\ell}$.

Exercise 2.32: For any $A \subset \mathsf{R}^1$, $(A^{\ell})^{\ell} \subset A^{\ell}$. Must $(A^{\ell})^{\ell} = A^{\ell}$?

Exercise 2.33: Let $A \subset \mathsf{R}^1$, and let $p \in \mathsf{R}^1$. Then p is a limit point of A if and only if for each $\epsilon > 0$, the open interval $(p - \epsilon, p + \epsilon)$ contains infinitely many points of A.

Exercise 2.34: Let $X \subset \mathsf{R}^1$ and let $A \subset X$. If p is an isolated point of X, then $p \sim A$ if and only if $p \in A$.

We conclude with a simple theorem that shows that functions are always continuous at any isolated point of their domain. In other words, continuity is only in question at points of the domain that are limit points of the domain.

Theorem 2.35: Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let p be an isolated point of X. Then f is continuous at p.

Proof: We prove that f is continuous at p by showing that f satisfies the definition of continuity at p (which is below Exercise 2.22).

Let $A \subset X$ such that $p \sim A$. Then, by Exercise 2.34, $p \in A$. Hence, $f(p) \in f(A)$. Thus, by Theorem 2.7, $f(p) \sim f(A)$. Therefore, we have proved that f is continuous at p. ¥

Chapter III: The Notion of Limit

We define and discuss the notion of limit of a function, commonly denoted in calculus by $\lim_{x\to p} f(x)$. In section 2, we reformulate the notion of limit completely in terms of arbitrary closeness. In section 3, we use the result in section 2 to show that our definition of continuity in the preceding chapter is equivalent to the definition of continuity as presented in calculus. In section 4, we present our rationale for introducing continuity before limits (which is contrary to common practice). In the final section, we discuss one - sided limits.

1. The Definition of Limit

You encountered limits in calculus. We state the definition for $\lim_{x\to p} f(x)$ as it is presented in calculus but in a slightly more general way – we replace the assumption in the calculus definition that f is defined at all points $x \neq p$ in an open interval about p with the less restrictive assumption that p is a limit point of the domain of f. (See the comments at the end of section 5.)

Definition. Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in \mathsf{R}^1$ such that p is a limit point of X. We say that L *is the limit of* f *as* x *approaches* p, written $\lim_{x\to p} f(x) = L$, provided that for any given number $\epsilon > 0$, there is a number $\delta > 0$ such that for all $x \in X - \{p\}$ such that $|x - p| < \delta$, we have

$$|f(x) - L| < \epsilon.$$

The definition is complicated. Let us interpret the definition informally as a game: You give me any error $\epsilon > 0$, meaning that you will allow the values of f to deviate from L but only by less than ϵ; I win the game if for any such allowed error, I can find a δ - neighborhood of p such that the values of the function f on the neighborhood with p removed are within the prescribed error ϵ from L.

It is important in the definition of limit that we did not require p to be a point of X. Indeed, many important limits are considered when p is *not* a point of X. For example, the derivative of a function f at a point p is $\lim_{h\to 0} \frac{f(p+h)-f(p)}{h}$; the expression $\frac{f(p+h)-f(p)}{h}$ defines a function of h for which 0 is not in its domain.

We also note that the requirement that p be a limit point of X is important in the definition. For if p is not a limit point of X, then any number whatsoever is a limit of f as x approaches p, even when $p \in X$; this is seen by taking $\delta = dist(p, X - \{p\})$ (try this for any function $f : \mathsf{N} \to \mathsf{R}^1$ and any choice of L). What we are suggesting here is that the requirement that p be a limit point of X makes the limit unique (if the limit exists); we now prove that this is the case.

Theorem 3.1: Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in \mathsf{R}^1$ such that p is a limit point of X. If $\lim_{x\to p} f(x) = L_1$ and $\lim_{x\to p} f(x) = L_2$, then $L_1 = L_2$.

Proof: Suppose by way of contradiction that $L_1 \neq L_2$. Let $\epsilon = \frac{|L_1-L_2|}{2}$, and note that $\epsilon > 0$. Since $\lim_{x\to p} f(x) = L_i$ for each $i = 1$ and 2, there exist $\delta_1, \delta_2 > 0$ satisfying the following for each $i = 1$ and 2 :

(*) For all $x \in X - \{p\}$ such that $|x - p| < \delta_i$, $|f(x) - L_i| < \epsilon$.

Now, let $\delta = \min\{\delta_1, \delta_2\}$, and note that $\delta > 0$. Thus, since $p \sim X - \{p\}$, we have by Theorem 2.5 that

$$(p - \delta, p + \delta) \cap (X - \{p\}) \neq \emptyset.$$

Hence, there is a point $x_0 \in (p - \delta, p + \delta) \cap (X - \{p\})$. Thus, since $\delta \leq \delta_i$ for each $i = 1$ and 2, we see from (*) that

$$|f(x_0) - L_i| < \epsilon \text{ for each } i.$$

Therefore,

$$|L_1 - L_2| = |L_1 - f(x_0) + f(x_0) - L_2| \leq |L_1 - f(x_0)| + |f(x_0) - L_2| < 2\epsilon.$$

Thus, since $\epsilon = \frac{|L_1 - L_2|}{2}$, $|L_1 - L_2| < |L_1 - L_2|$; however, this is impossible. ¥

It is convenient to have the following general agreement: *When we consider an algebraic expression as being a function, we assume, often without saying so, that the domain of the function is the largest set of real numbers for which the expression makes sense (unless we say otherwise).*

In the example below, we illustrate the thought process for computing limits of specific functions. The thought process is important even though we establish general theorems for evaluating limits in the next chapter.

Example 3.2: $\lim_{x \to 7} \frac{1}{x-4} = \frac{1}{3}$. To prove this, let $\epsilon > 0$. We want to find a $\delta > 0$ such that for all $x \in \mathsf{R}^1 - \{4\}$ (which is the understood domain of the function $f(x) = \frac{1}{x-4}$),

$$(*) \left| \frac{1}{x-4} - \frac{1}{3} \right| < \epsilon \text{ when } x \neq 7 \text{ and } |x - 7| < \delta.$$

We start our search for δ by writing $\left| \frac{1}{x-4} - \frac{1}{3} \right|$ in a way that tells us how its value depends on $|x - 7|$:

$$(1) \left| \frac{1}{x-4} - \frac{1}{3} \right| = \left| \frac{3-(x-4)}{3(x-4)} \right| = \left| \frac{-x+7}{3(x-4)} \right| = \frac{|x-7|}{3|x-4|}.$$

Next, we make an initial restriction on δ so that we can bound the size of the last expression in (1) when $|x - 7| < \delta$. This means we want δ small enough so that if $|x - 7| < \delta$, then x is bounded away from 4. This happens for any fixed $\delta < 3$. So, we assume temporarily that $\delta \leq 1$ and, of course, that $\delta > 0$. (We will see when we make our final choice for δ why we do not simply take $\delta = 1$ here).

Now, we examine what our assumption $|x - 7| < \delta \leq 1$ says about the size of $\left| \frac{1}{x-4} - \frac{1}{3} \right|$. Since $|x - 7| < \delta \leq 1$, we see that $x > 6$ and, thus, $2 < |x - 4|$. Hence,

$$\frac{1}{|x-4|} < \frac{1}{2}.$$

Thus, $\frac{|x-7|}{3|x-4|} < \frac{|x-7|}{6}$. Therefore, by (1), we have that

(2) $\left|\frac{1}{x-4} - \frac{1}{3}\right| < \frac{|x-7|}{6}$ when $0 < \delta \leq 1$.

We now make our final choice for δ and verify that our choice works. Note that $\frac{|x-7|}{6} < \epsilon$ if $|x-7| < 6\epsilon$, and let

$$\delta = \min\{1, 6\epsilon\}.$$

Then, for all $x \in \mathsf{R}^1 - \{4\}$ such that $x \neq 7$ and $|x-7| < \delta$, we have

$$\left|\frac{1}{x-4} - \frac{1}{3}\right| \overset{(2)}{<} \frac{|x-7|}{6} < \frac{6\epsilon}{6} = \epsilon.$$

This proves (*).

Exercise 3.3: Prove that $\lim_{x\to 1} \frac{x}{x-3} = \frac{-1}{2}$.

Exercise 3.4: Prove that $\lim_{x\to 2} 4x + 5 = 13$.

Exercise 3.5: Prove that $\lim_{x\to 4} \frac{x-4}{x^2-2x-8} = \frac{1}{6}$.

Exercise 3.6: Prove that $\lim_{x\to p} |x| = |p|$.

Exercise 3.7: Prove that $\lim_{x\to p} \sqrt{x} = \sqrt{p}$ for all $p \geq 0$. (Note: $f(x) = \sqrt{x}$ is a function on $[0, \infty)$ by Theorem 1.25.)

Exercise 3.8: Prove that $\lim_{x\to -3} \frac{|x+3|}{x+3}$ does not exist.

Exercise 3.9: Assume that $\lim_{x\to p} f(x) = \sqrt{82} - 9$, where p is a limit point of the domain X of f. Prove that there is a $\delta > 0$ such that $f(x) > 0$ for all $x \in X - \{p\}$ such that $|x-p| < \delta$. If $p \in X$, must $f(p) > 0$?

Exercise 3.10: Give an example of functions $f, g : \mathsf{R}^1 \to \mathsf{R}^1$ such that $\lim_{x\to 0} f(x) = 0$, $\lim_{x\to 0} g(x) = 0$, and $\lim_{x\to 0} \frac{f(x)}{g(x)} = 23$.

2. Limits in Terms of Arbitrary Closeness

We reformulate the definition of limit entirely in terms of the notion arbitrary closeness. We use the reformulation in the next section.

Theorem 3.11: Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in \mathsf{R}^1$ such that p is a limit point of X. Then $\lim_{x\to p} f(x) = L$ if and only if whenever $A \subset X$ such that $p \sim A - \{p\}$, then $L \sim f(A - \{p\})$.

Proof: Assume that $\lim_{x\to p} f(x) = L$. Let $A \subset X$ such that $p \sim A - \{p\}$. We show that $L \sim f(A - \{p\})$ by using Theorem 2.5. Let $\epsilon > 0$. Then, since $\lim_{x\to p} f(x) = L$, there exists $\delta > 0$ such that for all $x \in X - \{p\}$ such that $|x-p| < \delta$, we have

$$|f(x) - L| < \epsilon.$$

Since $p \sim A - \{p\}$, there is a point $x_0 \in (p-\delta, p+\delta) \cap (A - \{p\})$ by Theorem 2.5. Hence, $|x_0 - p| < \delta$ and $x_0 \in X - \{p\}$. Thus, $|f(x_0) - L| < \epsilon$; also, since $x_0 \in A - \{p\}$, $f(x_0) \in f(A - \{p\})$. Hence,

$$f(x_0) \in (L-\epsilon, L+\epsilon) \cap f(A - \{p\}).$$

We have shown that for any $\epsilon > 0$, $(L-\epsilon, L+\epsilon) \cap f(A - \{p\}) \neq \emptyset$. Therefore, by Theorem 2.5, $L \sim f(A - \{p\})$.

Conversely, assume that $\lim_{x \to p} f(x) \neq L$. Then there exists $\epsilon > 0$ such that for every $\delta > 0$, there is a point $x_\delta \in X - \{p\}$ such that $|x_\delta - p| < \delta$ and $|f(x_\delta) - L| \geq \epsilon$. In other words, the following set is nonempty for each $\delta > 0$:

$$A_\delta = \{x \in X - \{p\} : |x - p| < \delta \text{ and } |f(x) - L| \geq \epsilon\}.$$

Now, let $A = \cup_{\delta > 0} A_\delta$. Since $A_\delta \neq \emptyset$ for each $\delta > 0$, we see that

$$(p-\delta, p+\delta) \cap A \neq \emptyset \text{ for each } \delta > 0.$$

Hence, by Theorem 2.5, $p \sim A$. Thus, since $p \notin A$, we have that

(1) $p \sim A - \{p\}$.

Since $|f(x) - L| \geq \epsilon$ for all $x \in A$,

$$(L-\epsilon, L+\epsilon) \cap f(A) = \emptyset,$$

which gives $(L-\epsilon, L+\epsilon) \cap f(A - \{p\}) = \emptyset$. Thus, by Theorem 2.5, we have that

(2) $L \not\sim f(A - \{p\})$.

Finally, we see from (1) and (2) that the condition in the second part of our theorem is false for the set A we have defined. ¥

3. The Limit Characterization of Continuity

We show that our definition of continuity in the preceding chapter is equivalent to the definition of continuity as presented in calculus. In other words, the standard definition of continuity (in terms of limits) is, for us, a theorem. The reason for this seemingly strange development is discussed in section 4.

Theorem 3.12: Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in X$ such that p is a limit point of X. Then f is continuous at p if and only if $\lim_{x \to p} f(x) = f(p)$.

Proof: Assume that f is continuous at p. Then, for any $A \subset X$ such that $p \sim A - \{p\}$, we see from our definition of continuity that $f(p) \sim f(A - \{p\})$. Therefore, by Theorem 3.11, $\lim_{x \to p} f(x) = f(p)$.

Conversely, assume that f is not continuous at p. Then, by our definition of continuity, there exists $A \subset X$ such that $p \sim A$ but $f(p) \not\sim f(A)$.

Since $f(p) \not\sim f(A)$, $f(p) \notin f(A)$ (by Theorem 2.7); hence, $p \notin A$, which shows that $A = A - \{p\}$. Thus, since $p \sim A$ and $f(p) \not\sim f(A)$, we have that

$p \sim A - \{p\}$ and $f(p) \not\sim f(A - \{p\})$. Therefore, $\lim_{x \to p} f(x) \neq f(p)$ by Theorem 3.11. ¥

The theorem we just proved characterizes continuity only at limit points of X. The following corollary completes the characterization.

Corollary 3.13: Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in X$. Then f is continuous at p if and only if p is an isolated point of X or $\lim_{x \to p} f(x) = f(p)$ when p is a limit point of X.

Proof: Assume that f is continuous at p and that p is not an isolated point of X. Then p is a limit point of X and, hence, $\lim_{x \to p} f(x) = f(p)$ by Theorem 3.12. This proves that continuity at p implies the second conditions in the corollary.

Conversely, if p is an isolated point of X, then f is continuous at p by Theorem 2.35. If p is a limit point of X and if $\lim_{x \to p} f(x) = f(p)$, then f is continuous at p by Theorem 3.12. ¥

Exercise 3.14: Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in X$. Then f is continuous at p if and only if for any open interval I such that $f(p) \in I$, there is an open interval J such that $p \in J$ and $f(J) \subset I$.

4. Limits in Terms of Continuity

In all calculus books, limits are defined before continuity and continuity is then defined in terms of limits. In our presentation, we have reversed the order for introducing these ideas. The reason we have done this is our realization that in trying to understand limits, you are really trying to understand continuity; the theorem below explains this. It is my opinion that continuity is simpler and easier to understand than limits. Thus, why *not* introduce continuity first and use it as a vehicle for building up intuition for the more subtle idea of limits.

In general terms, the following theorem says that $\lim_{x \to p} f(x)$ exists if and only if the function f can be defined or redefined at p so that the resulting function is continuous at p.

Theorem 3.15: Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in \mathsf{R}^1$ such that p is a limit point of X. Then $\lim_{x \to p} f(x) = L$ if and only if the function $g : X \cup \{p\} \to \mathsf{R}^1$ given by

$$g(x) = \begin{cases} f(x) & \text{, if } x \in X \\ L & \text{, if } x = p \end{cases}$$

is continuous at p.

Proof: Note that $g(x) = f(x)$ for all $x \in X - \{p\}$. Thus, we see easily from the definition of limit (section 1) that $\lim_{x \to p} f(x) = L$ if and only if $\lim_{x \to p} g(x) = L$. Thus, since $L = g(p)$, $\lim_{x \to p} f(x) = L$ if and only if $\lim_{x \to p} g(x) = g(p)$. Therefore, by Theorem 3.12, $\lim_{x \to p} f(x) = L$ if and only if g is continuous at p. ¥

5. One - sided Limits

A point in the real line can be "approached" from the left and from the right. This simple observation leads us to a way to break limits down into two cases – limits from the left and limits from the right. Considering the two cases separately is sometimes helpful in computing limits or in showing limits do not exist. This is especially true when a function is defined by a formula that changes at a point (the change can happen explicitly or implicitly – compare Exercises 3.17 and 3.18). We prove a theorem that can be applied in such situations.

We note the definition of the restriction of a function. Let X and Y be sets, and let $f : X \to Y$ be a function. For any set $X' \subset X$, the *restriction of f to X'*, denoted by $f|X'$, is the function from X' to Y defined in the following simple way:

$$(f|X')(x') = f(x'), \text{ all } x' \in X.$$

We define one - sided limits:

Definition. Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in \mathsf{R}^1$ such that p is a limit point of $X \cap (-\infty, p]$. We say L *is the limit of f as x approaches p from the left*, or the *left - hand limit of f as x approaches p*, written $\lim_{x \to p^-} f(x) = L$, provided that

$$\lim_{x \to p}(f|X \cap (-\infty, p])(x) = L.$$

Similarly, assuming that p is a limit point of $X \cap (p, \infty]$, we say L *is the limit of f as x approaches p from the right*, or the *right - hand limit of f as x approaches p*, written $\lim_{x \to p^+} f(x) = L$, provided that

$$\lim_{x \to p}(f|X \cap [p, \infty))(x) = L.$$

The following terminology is descriptive and will help make statements succinct: Let $X \subset \mathsf{R}^1$ and let $p \in \mathsf{R}^1$; we call p a *two - sided limit point of X* provided that p is a limit point of $X \cap (-\infty, p]$ and p is a limit point of $X \cap [p, \infty)$.

Theorem 3.16: Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in \mathsf{R}^1$ such that p is a two - sided limit point of X. Then $\lim_{x \to p} f(x) = L$ if and only if

$$\lim_{x \to p^-} f(x) = L = \lim_{x \to p^+} f(x).$$

Proof: Assume that $\lim_{x \to p} f(x) = L$. Let $\epsilon > 0$. Then, by the definition of limit, there exists $\delta > 0$ such that for all $x \in X - \{p\}$ such that $|x - p| < \delta$,

$$|f(x) - L| < \epsilon.$$

Therefore, it is clear that $|f(x) - L| < \epsilon$ for all $x \in X \cap (-\infty, p)$, as well as for all $x \in X \cap (p, \infty)$, such that $|x - p| < \delta$. This proves that

$$\lim_{x \to p^-} f(x) = L = \lim_{x \to p^+} f(x).$$

Conversely, assume that $\lim_{x\to p^-} f(x) = L = \lim_{x\to p^+} f(x)$. Let $\epsilon > 0$. Since $\lim_{x\to p^-} f(x) = L$ and since (by definition)

$$\lim_{x\to p^-} f(x) = \lim_{x\to p}(f|X \cap (-\infty, p])(x),$$

there exists $\delta_1 > 0$ such that for all $x \in X \cap (-\infty, p)$ such that $|x - p| < \delta_1$,

$$|(f|X \cap [p, \infty])(x) - L| < \epsilon;$$

similarly, since $\lim_{x\to p^+} f(x) = L$, there exists $\delta_2 > 0$ such that for all $x \in X \cap (p, \infty)$ such that $|x - p| < \delta_2$,

$$|(f|X \cap [p, \infty])(x) - L| < \epsilon.$$

Therefore, letting $\delta = \min\{\delta_1, \delta_2\}$, we see that for all $x \in X - \{p\}$ such that $|x - p| < \delta$,

$$|f(x) - L| < \epsilon.$$

This proves that $\lim_{x\to p} f(x) = L$ (note: for us to conclude that $\lim_{x\to p} f(x) = L$, the definition of limit in section 1 requires us to know that p is a limit point of X; this follows from Exercise 2.30 since p is a limit point of $X \cap (-\infty, p]$). ¥

We conclude with comments about limits and one-sided limits. When we defined $\lim_{x\to p} f(x)$ in section 1, we did not make the common assumption that the point p lies in an open interval contained in the domain of f. Thus, for example, we can properly write $\lim_{x\to p} \sqrt{x}$ even when $p = 0$, whereas common practice forces authors to write $\lim_{x\to 0^+} \sqrt{x}$. In general, when the domain of f is an interval, we write $\lim_{x\to p} f(x)$ whether p is an end point of the interval or not, whereas other authors *are forced* to make the distinction. In this situation, we consider the distinction between limits and one-sided limits a distraction – a nuisance – rather than substantive. On the other hand, there are situations in which it is important to consider one-sided limits. By defining limits as we did, all our general theorems about limits in the next chapter *automatically* hold for their one-sided analogues.

Exercise 3.17: Find $\lim_{x\to 3} f(x)$ (if the limit exists) when

$$f(x) = \begin{cases} x + 1 & \text{, if } x \leq 3 \\ -4x + 16 & \text{, if } x > 3. \end{cases}$$

Exercise 3.18: Find $\lim_{x\to 4} \frac{|x-4|}{x-4}$ (if the limit exists).

Exercise 3.19: Find $\lim_{x\to 0} \frac{x^2}{|x|}$ (if the limit exists).

Exercise 3.20: Find $\lim_{x\to 1} \frac{x-1}{|x^2+x-2|}$ (if the limit exists).

Chapter IV: Limit Theorems

We prove theorems about limits of sums, differences, products and quotients of functions whose limits separately exist. We obtain general results about continuity as corollaries; as consequences, we show that all polynomials are continuous and that all rational functions are continuous (on their domains). We then prove theorems about limits of compositions of functions, including the Substitution Theorem. Next, we prove the simple but useful Squeeze Theorem. Finally, we briefly discuss limits of sequences.

All our theorems concerning limits hold for one - sided limits (see the comments at the end of the last section of Chapter III). We keep this in mind rather than stating the one - sided versions of the theorems.

1. Limits for Sums and Differences

We prove theorems about limits and continuity of sums and differences of two functions. We then extend the sum theorems to finitely many functions.

Definition. Let $X \subset \mathsf{R}^1$, and let $f, g : X \to \mathsf{R}^1$ be functions. The *sum of f and g* is the function $f + g : X \to \mathsf{R}^1$ defined by

$$(f+g)(x) = f(x) + g(x) \quad \text{for all } x \in X.$$

Similarly, the *difference of f and g* is the function $f - g : X \to \mathsf{R}^1$ defined by

$$(f-g)(x) = f(x) - g(x) \quad \text{for all } x \in X.$$

We first prove that the limit of the sum of two functions whose limits separately exist is the sum of the limits of the two functions. Note that this shows, in particular, that the limit of the sum exists (provided that the separate limits exist).

Theorem 4.1: Let $X \subset \mathsf{R}^1$, let $f, g : X \to \mathsf{R}^1$ be functions, and let $p \in \mathsf{R}^1$ such that p is a limit point of X. If

$$\lim_{x\to p} f(x) = L \quad \text{and} \quad \lim_{x\to p} g(x) = M,$$

then $\lim_{x\to p}(f+g)(x) = L + M$.

Proof: Let $\epsilon > 0$. We want to find a $\delta > 0$ such that for all $x \in X - \{p\}$ such that $|x - p| < \delta$, $|(f+g)(x) - (L+M)| < \epsilon$.

The clue to how to find δ comes from rewriting $|(f+g)(x) - (L+M)|$ so that expressions related to different assumptions in the theorem are grouped together:

$$\begin{aligned} |(f+g)(x) - (L+M)| &= |(f(x) - L) + (g(x) - M)| \\ &\leq |f(x) - L| + |g(x) - M|. \end{aligned}$$

Thus, we want to find a $\delta > 0$ such that $|f(x) - L| < \frac{\epsilon}{2}$ and $|g(x) - M| < \frac{\epsilon}{2}$ for all $x \in X - \{p\}$ such that $|x - p| < \delta$. It is fairly easy to find such a δ; we now prove the theorem using what we have just observed as a guide (a cheat sheet!).

Since $\lim_{x\to p} f(x) = L$, there is a $\delta_1 > 0$ such that for all $x \in X - \{p\}$ such that $|x - p| < \delta_1$,

$$|f(x) - L| < \tfrac{\epsilon}{2}.$$

Since $\lim_{x\to p} g(x) = M$, there is a $\delta_2 > 0$ such that for all $x \in X - \{p\}$ such that $|x - p| < \delta_2$,

$$|g(x) - M| < \tfrac{\epsilon}{2}.$$

Let $\delta = \min\{\delta_1, \delta_2\}$. Then $\delta > 0$ and for all $x \in X - \{p\}$ such that $|x - p| < \delta$,

$$|(f + g)(x) - (L + M)| \leq |f(x) - L| + |g(x) - M| < \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = \epsilon. \quad ¥$$

Our next theorem is the analogue of Theorem 4.1 for the difference of two functions.

Theorem 4.2: Let $X \subset \mathsf{R}^1$, let $f, g : X \to \mathsf{R}^1$ be functions, and let $p \in \mathsf{R}^1$ such that p is a limit point of X. If

$$\lim_{x\to p} f(x) = L \quad \text{and} \quad \lim_{x\to p} g(x) = M,$$

then $\lim_{x\to p}(f - g)(x) = L - M$.

Exercise 4.3: Prove Theorem 4.2.

Corollary 4.4: Let $X \subset \mathsf{R}^1$, let $f, g : X \to \mathsf{R}^1$ be functions, and let $p \in X$. If f and g are continuous at p, then $f + g$ and $f - g$ are continuous at p.

Proof: The corollary follows immediately from Theorem 4.1 and Theorem 4.2 using Corollary 3.13. ¥

We extend Theorem 4.1 to the sum of finitely many functions. The *sum of finitely many functions* is defined inductively: Having already defined the sum of two functions, assume inductively that we have defined the sum of n functions (with the same domain) for some natural number $n \geq 2$; then, for any $n + 1$ functions with the same domain, define $f_1 + \cdots + f_n + f_{n+1}$ to be the function $(f_1 + \cdots + f_n) + f_{n+1}$ (see Theorem 1.20).

Theorem 4.5: Let $X \subset \mathsf{R}^1$, let $f_i : X \to \mathsf{R}^1$ be a function for each $i = 1, 2, ..., n$, and let $p \in \mathsf{R}^1$ such that p is a limit point of X. If

$$\lim_{x\to p} f_i(x) = L_i \text{ for each } i = 1, 2, ..., n,$$

then $\lim_{x\to p}(f_1 + f_2 + \cdots + f_n)(x) = L_1 + L_2 + \cdots + L_n$.

Proof: We prove the theorem by induction on the number n of functions. The Induction Principle is Theorem 1.20.

The theorem is obviously true when $n = 1$.

Assume inductively that for some natural number k, the theorem is true for *any* k functions.

Let $f_1, f_2, ..., f_{k+1}$ be any $k + 1$ functions satisfying the assumptions in the theorem; that is, for each $i = 1, 2, ..., k + 1$, f_i is a function from X to R^1 such that $\lim_{x\to p} f_i(x) = L_i$. Then, by our inductive assumption,

$$\lim_{x\to p}(f_1 + f_2 + \cdots + f_k)(x) = L_1 + L_2 + \cdots + L_k.$$

Thus, since $\lim_{x\to p} f_{k+1}(x) = L_{k+1}$, Theorem 4.1 gives us that

$$\lim_{x\to p}((f_1 + f_2 + \cdots + f_k) + f_{k+1})(x) = (L_1 + L_2 + \cdots + L_k) + L_{k+1}.$$

Therefore, by our definition of finite sums of functions,

$$\lim_{x\to p}(f_1 + f_2 + \cdots + f_k + f_{k+1})(x) = L_1 + L_2 + \cdots + L_k + L_{k+1}.$$

This proves the theorem is true for $k+1$ functions under the assumption that it is true for k functions.

The theorem now follows from the Induction Principle. ¥

Corollary 4.6: Let $X \subset \mathsf{R}^1$, and let $p \in X$. If each of finitely many functions is continuous at p, then the sum function is continuous at p.

Proof: Apply Theorem 4.5 and Corollary 3.13. ¥

Exercise 4.7: Give an example of two functions $f, g : \mathsf{R}^1 \to \mathsf{R}^1$ such that for some point $p \in \mathsf{R}^1$, $\lim_{x\to p}(f+g)(x)$ exists but $\lim_{x\to p} f(x)$ and $\lim_{x\to p} g(x)$ do not exist.

Exercise 4.8: Are there two functions $f, g : \mathsf{R}^1 \to \mathsf{R}^1$ such that for some point $p \in \mathsf{R}^1$, $\lim_{x\to p}(f+g)(x)$ and $\lim_{x\to p} f(x)$ both exist but $\lim_{x\to p} g(x)$ does not exist ?

2. Limits for Products

We prove theorems about limits and continuity of products of finitely many functions.

Definition. Let $X \subset \mathsf{R}^1$, and let $f, g : X \to \mathsf{R}^1$ be functions. The *product of f and g* is the function $f \cdot g : X \to \mathsf{R}^1$ defined by

$$(f \cdot g)(x) = f(x)g(x) \quad \text{for all } x \in X.$$

We first prove that the limit of the product of two functions whose limits separately exist is the product of the limits of the two functions.

Theorem 4.9: Let $X \subset \mathsf{R}^1$, let $f, g : X \to \mathsf{R}^1$ be functions, and let $p \in \mathsf{R}^1$ such that p is a limit point of X. If

$$\lim_{x\to p} f(x) = L \quad \text{and} \quad \lim_{x\to p} g(x) = M,$$

then $\lim_{x\to p}(f \cdot g)(x) = LM$.

Proof: Let $\epsilon > 0$. We want to find a $\delta > 0$ such that for all $x \in X - \{p\}$ such that $|x - p| < \delta$, $|(f \cdot g)(x) - LM| < \epsilon$.

As in the proof of Theorem 4.1, the clue for finding δ comes from rewriting $|(f \cdot g)(x) - LM|$ so that expressions related to different assumptions in the

theorem are grouped together. To group the proper expressions, we employ the trick of subtracting and adding an expression, namely, $Lg(x)$:

$$\begin{aligned}|(f\cdot g)(x)-LM| &= |f(x)g(x)-Lg(x)+Lg(x)-LM|\\ &\le |g(x)(f(x)-L)|+|L(g(x)-M)|\\ &= |g(x)|\,|f(x)-L|+|L|\,|g(x)-M|.\end{aligned}$$

Thus, we want to find a $\delta>0$ such that for all $x\in X-\{p\}$ such that $|x-p|<\delta$,

$$(*)\ |g(x)|\,|f(x)-L|<\tfrac{\epsilon}{2}\ \text{ and }\ (**)\ |L|\,|g(x)-M|<\tfrac{\epsilon}{2}.$$

We show how to find such a δ as follows.

We first bound $|g(x)|$: Since $\lim_{x\to p} g(x)=M$, there is a $\delta_1>0$ such that for all $x\in X-\{p\}$ such that $|x-p|<\delta_1$, $|g(x)-M|<1$; hence, by Exercise 1.29, $\big||g(x)|-|M|\big|<1$. Thus, we have that

(1) $|g(x)|<1+|M|$ for all $x\in X-\{p\}$ such that $|x-p|<\delta_1$.

Next, with (1) and (*) in mind, we note that since $\lim_{x\to p} f(x)=L$, there is a $\delta_2>0$ such that

(2) $|f(x)-L|<\frac{\epsilon}{2(1+|M|)}$ for all $x\in X-\{p\}$ such that $|x-p|<\delta_2$.

Then we see from (1) and (2) that $\min\{\delta_1,\delta_2\}$ is a δ that makes (*) hold for all $x\in X-\{p\}$ such that $|x-p|<\delta$.

Next, we find a $\delta_3>0$ that makes (**) hold for all $x\in X-\{p\}$ such that $|x-p|<\delta_3$. Our immediate inclination is to use that $\lim_{x\to p} g(x)=M$ to choose $\delta_3>0$ such that $|g(x)-M|<\frac{\epsilon}{2|L|}$ for the relevant points x, hence (**) holds. However, this obviously does not work when $L=0$; nevertheless, if $L=0$, then any $\delta_3>0$ makes (**) hold for the relevant points x. Thus, we can take two cases in defining δ_3 – the case when $L\neq 0$ and the case when $L=0$ – or we can use the trick of considering the positive number $\frac{\epsilon}{2(1+|L|)}$. We choose the latter: Since $\lim_{x\to p} g(x)=M$, there is a $\delta_3>0$ such that

(3) $|g(x)-M|<\frac{\epsilon}{2(1+|L|)}$ for all $x\in X-\{p\}$ such that $|x-p|<\delta_3$.

Finally, let $\delta=\min\{\delta_1,\delta_2,\delta_3\}$. Then $\delta>0$ and for all $x\in X-\{p\}$ such that $|x-p|<\delta$, we see using (1), (2) and (3) that

$$\begin{aligned}|(f\cdot g)(x)-LM| &\le |g(x)|\,|f(x)-L|+|L|\,|g(x)-M|\\ &< (1+|M|)\tfrac{\epsilon}{2(1+|M|)}+|L|\,\tfrac{\epsilon}{2(1+|L|)}<\tfrac{\epsilon}{2}+\tfrac{\epsilon}{2}=\epsilon.\ \text{¥}\end{aligned}$$

Corollary 4.10: Let $X\subset \mathsf{R}^1$, let $f,g: X\to \mathsf{R}^1$ be functions, and let $p\in X$. If f and g are continuous at p, then $f\cdot g$ is continuous at p.

Proof: Simply apply Theorem 4.9 and Corollary 3.13. ¥

We extend Theorem 4.9 to the product of finitely many functions. The *product of finitely many functions* is defined inductively in the same way that we defined the sum of finitely many functions in the preceding section.

Theorem 4.11: Let $X\subset \mathsf{R}^1$, let $f_i: X\to \mathsf{R}^1$ be a function for each $i=1,2,...,n$, and let $p\in \mathsf{R}^1$ such that p is a limit point of X. If

$$\lim_{x\to p} f_i(x) = L_i \text{ for each } i = 1, 2, ..., n\,,$$

then $\lim_{x\to p}(f_1 \cdot f_2 \cdot \cdots \cdot f_n)(x) = L_1 L_2 \cdots L_n$.

Exercise 4.12: Prove Theorem 4.11.

Corollary 4.13: Let $X \subset \mathbb{R}^1$, and let $p \in X$. If each of finitely many functions is continuous at p, then the product function is continuous at p.

Proof: Apply Theorem 4.11 and Corollary 3.13. ¥

Exercise 4.14: Give an example of two functions $f, g : \mathbb{R}^1 \to \mathbb{R}^1$ such that for some point $p \in \mathbb{R}^1$, $\lim_{x\to p}(f \cdot g)(x)$ exists but $\lim_{x\to p} f(x)$ and $\lim_{x\to p} g(x)$ do not exist.

Exercise 4.15: Are there two functions $f, g : \mathbb{R}^1 \to \mathbb{R}^1$ such that for some point $p \in \mathbb{R}^1$, $\lim_{x\to p}(f \cdot g)(x)$ and $\lim_{x\to p} f(x)$ both exist but $\lim_{x\to p} g(x)$ does not exist ?

3. Continuity of Polynomials

We are now in a position to easily prove the important fact that all polynomials are continuous.

Definition. A *polynomial* is a function f that can be written in the form

$$f(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n, \text{ all } x \in \mathbb{R}^1,$$

where $c_0, c_1, ..., c_n$ are constants.

The constants $c_0, c_1, ..., c_n$ are called the *coefficients of the polynomial* f; c_i is called the i^{th} *coefficient of* f. If $c_n \neq 0$, we say that f is a *polynomial of degree* n.

Note that we say f is a polynomial if it *can be written* in the form indicated. Thus, for example, the function f defined by $f(x) = 3(x-4)(x^6 + 5x^2)^3$ is a polynomial.

We use the following functions in the proof that polynomials are continuous: A *constant function* is a function all of whose values are the same (i.e., a polynomial of degree 0); the *identity function* is the function f given by $f(x) = x$ for all $x \in \mathbb{R}^1$.

Theorem 4.16: All polynomials are continuous on $\mathbb{R}^1$.

Proof: Any constant function and the identity function are continuous, as we showed in Example 2.23. Thus, for any fixed real number c and for any fixed natural number k, the function $f(x) = cx^k$ (all $x \in \mathbb{R}^1$) is continuous by Corollary 4.13. Our theorem now follows from Corollary 4.6. ¥

Theorems really make life easy: Can you imagine proving with epsilons and deltas, without theorems about limits, that the function f given by $f(x) = 6x^{89} + \frac{168}{31}x^{25} - \sqrt{17}x^{13} + 49$ is continuous?

Exercise 4.17: At which real numbers p is the function f given by $f(x) = \frac{8x^3-64}{2(x-2)}$ continuous?

Exercise 4.18: Is the function f given by $f(x) = \frac{x^2-x}{x}$ a polynomial?

4. Limits for Quotients

We prove theorems about limits and continuity of quotients of two functions.

Definition. Let $X \subset \mathsf{R}^1$, and let $f, g : X \to \mathsf{R}^1$ be functions such that $g(x) \neq 0$ for any $x \in X$. The *quotient of f and g* is the function $\frac{f}{g} : X \to \mathsf{R}^1$ defined by

$$\tfrac{f}{g}(x) = \tfrac{f(x)}{g(x)} \quad \text{for all } x \in X.$$

We prove that the limit of the quotient of two functions whose limits separately exist is the quotient of the limits of the two functions provided, of course, that the limit of the function in the denominator is not zero. When the limit of the denominator *is* zero, the limit of the quotient may or may not exist: $\lim_{x\to 0} \frac{1}{x}$ does not exist and $\lim_{x\to 0} \frac{x}{x} = 1$.

We prove a lemma about reciprocals; then our theorem about limits of quotients follows easily using the theorem about limits of products (Theorem 4.9).

Lemma 4.19: Let $X \subset \mathsf{R}^1$, let $g : X \to \mathsf{R}^1$ be a function such that $g(x) \neq 0$ for any $x \in X$, and let $p \in \mathsf{R}^1$ such that p is a limit point of X. If

$$\lim_{x\to p} g(x) = M \neq 0,$$

then $\lim_{x\to p} \frac{1}{g}(x) = \frac{1}{M}$.

Proof: Let $\epsilon > 0$. We want to find a $\delta > 0$ such that for all $x \in X - \{p\}$ such that $|x - p| < \delta$, $\left|\frac{1}{g}(x) - \frac{1}{M}\right| < \epsilon$.

As we did in proofs of previous theorems of this type, let us first examine what is involved in finding δ. We rewrite $\left|\frac{1}{g}(x) - \frac{1}{M}\right|$ so that the expression that we know can be made small, namely $|g(x) - M|$, is by itself (and hope that we can take care of the rest):

$$\left|\tfrac{1}{g}(x) - \tfrac{1}{M}\right| = \left|\tfrac{1}{g(x)} - \tfrac{1}{M}\right| = \left|\tfrac{M-g(x)}{Mg(x)}\right| = \tfrac{1}{|M|}\tfrac{1}{|g(x)|}\,|g(x) - M|.$$

Hence, we want to find a $\delta > 0$ such that for all $x \in X - \{p\}$ such that $|x - p| < \delta$,

$$\tfrac{1}{|M|}\tfrac{1}{|g(x)|}\,|g(x) - M| < \epsilon.$$

We now proceed with the proof, using what we have written as a guide.

Since $\lim_{x\to p} g(x) = M$, we see easily using Exercise 1.29 that

$$\lim_{x\to p} |g(x)| = |M|.$$

Thus, since $M \neq 0$, there is a $\delta_1 > 0$ such that

(1) $|g(x)| > \frac{|M|}{2}$ for all $x \in X - \{p\}$ such that $|x - p| < \delta_1$.

Since $M \neq 0$, $\frac{M^2\epsilon}{2} > 0$ (use of the quantity $\frac{M^2\epsilon}{2}$ comes from (1) and the observations we referred to as a guide). Thus, since $\lim_{x\to p} g(x) = M$, there is a $\delta_2 > 0$ such that

(2) $|g(x) - M| < \frac{M^2\epsilon}{2}$ for all $x \in X - \{p\}$ such that $|x - p| < \delta_2$.

Now, let $\delta = \min\{\delta_1, \delta_2\}$. Then $\delta > 0$ and for all $x \in X - \{p\}$ such that $|x - p| < \delta$, we see from (1) and (2) that

$$\left|\tfrac{1}{g}(x) - \tfrac{1}{M}\right| = \tfrac{1}{|M|}\tfrac{1}{|g(x)|}\,|g(x) - M| < \tfrac{1}{|M|}\tfrac{2}{|M|}\tfrac{M^2\epsilon}{2} = \epsilon. \quad ¥$$

Theorem 4.20: Let $X \subset \mathsf{R}^1$, let $f, g : X \to \mathsf{R}^1$ be functions such that $g(x) \neq 0$ for any $x \in X$, and let $p \in \mathsf{R}^1$ such that p is a limit point of X. If

$$\lim_{x\to p} f(x) = L \quad \text{and} \quad \lim_{x\to p} g(x) = M \neq 0,$$

then $\lim_{x\to p} \frac{f}{g}(x) = \frac{L}{M}$.

Proof: Observe that $\frac{f}{g} = f \cdot \frac{1}{g}$; then use Lemma 4.19 to apply Theorem 4.9 to the product $f \cdot \frac{1}{g}$. ¥

Corollary 4.21: Let $X \subset \mathsf{R}^1$, let $f, g : X \to \mathsf{R}^1$ be functions such that $g(x) \neq 0$ for any $x \in X$, and let $p \in X$. If f and g are continuous at p, then $\frac{f}{g}$ is continuous at p.

Proof: Use Theorem 4.20 and Corollary 3.13. ¥

Exercise 4.22: Give an example of two functions $f, g : \mathsf{R}^1 \to \mathsf{R}^1$, $g(x) \neq 0$ for all $x \in \mathsf{R}^1$, such that for some point $p \in \mathsf{R}^1$, $\lim_{x\to p} \frac{f}{g}(x)$ exists but $\lim_{x\to p} f(x)$ and $\lim_{x\to p} g(x)$ do not exist.

Exercise 4.23: Are there two functions $f, g : \mathsf{R}^1 \to \mathsf{R}^1$, $g(x) \neq 0$ for all $x \in \mathsf{R}^1$, such that for some point $p \in \mathsf{R}^1$, $\lim_{x\to p} \frac{f}{g}(x)$ and $\lim_{x\to p} f(x)$ both exist but $\lim_{x\to p} g(x)$ does not exist ?

Exercise 4.24: Are there two functions $f, g : \mathsf{R}^1 \to \mathsf{R}^1$, $g(x) \neq 0$ for all $x \in \mathsf{R}^1$, such that for some point $p \in \mathsf{R}^1$, $\lim_{x\to p} \frac{f}{g}(x)$ and $\lim_{x\to p} g(x)$ both exist but $\lim_{x\to p} f(x)$ does not exist ?

5. Continuity of Rational Functions

Definition. A *rational function* is a function that can be written as a quotient of two polynomials.

The following theorem is trivial to prove in view of what we have already done.

Theorem 4.25: Every rational function is continuous on its domain.

Proof: By Theorem 4.16, polynomials are continuous on R^1. Therefore, our theorem follows from Corollary 4.21. ¥

Exercise 4.26: Is the function f given by $f(x) = \frac{1}{(\frac{1}{x})}$ (all $x \neq 0$) a rational function ?

6. Compositions of Functions and Limits

We prove a theorem about the continuity of compositions of functions and a generalization concerning limits of compositions.

Definition. Let X, Y, and Z be sets, and let $f : X \to Y$ and $g : Y \to Z$ be functions. The *composition f followed by g* is the function from X to Z denoted by $g \circ f$ and defined by letting

$$(g \circ f)(x) = g(f(x)), \text{ all } x \in X.$$

We often use the phrase *the composition of f and g* when the context makes it clear (or unimportant) which function is first.[3]

Perhaps you have never drawn the graph of a composition of two specific functions. If not, try the following exercise:

Exercise 4.27: Let $f, g : [0,1] \to [0,1]$ be defined by

$$f(x) = \begin{cases} x + \frac{1}{2} & \text{, if } 0 \leq x \leq \frac{1}{2} \\ -2x + 2 & \text{, if } \frac{1}{2} \leq x \leq 1 \end{cases}, \quad g(x) = \begin{cases} -x + \frac{1}{2} & \text{, if } 0 \leq x \leq \frac{1}{2} \\ 2x - 1 & \text{, if } \frac{1}{2} \leq x \leq 1. \end{cases}$$

Draw the graphs of $f \circ f$, $g \circ f$ and $f \circ g$.

Our first theorem concerns the continuity of the composition of two functions. The theorem is simple to prove using only the definition of continuity (above Example 2.23).

Theorem 4.28: Let $X, Y, Z \subset \mathsf{R}^1$, and let $f : X \to Y$ and $g : Y \to Z$ be functions. If f is continuous at p and g is continuous at $f(p)$, then $g \circ f$ is continuous at p.

Proof: Let $A \subset X$ such that $p \sim A$. Then, by the definition of continuity, $f(p) \sim f(A)$. Thus, since g is continuous at $f(p)$, $g(f(p)) \sim g(f(A))$. Hence, we have proved that for any $A \subset X$ such that $p \sim A$,

$$(g \circ f)(p) \sim (g \circ f)(A).$$

Therefore, $g \circ f$ is continuous at p. ¥

Our next theorem is called the Substitution Theorem because it says that under certain conditions, $\lim_{x \to p}(g \circ f)(x)$ can be found by substituting $\lim_{x \to p} f(x)$

[3] In the definition of composition, the order of the functions is important: $\mathsf{f} \circ \mathsf{g}$ is not defined on all of Y when $\mathsf{g}(\mathsf{Y}) \not\subset \mathsf{X}$; furthermore, even if $\mathsf{X} = \mathsf{Y} = \mathsf{Z}$, $\mathsf{g} \circ \mathsf{f}$ is almost always different from $\mathsf{f} \circ \mathsf{g}$.

into the function g. After we prove the theorem, we discuss the assumptions in the theorem.

Theorem 4.29 (Substitution Theorem): Let $X, Y, Z \subset \mathsf{R}^1$, and let $f : X \to Y$ and $g : Y \to Z$ be functions. Let $p \in \mathsf{R}^1$ such that p is a limit point of X. If $\lim_{x\to p} f(x) = L$ and if g is continuous at L, then

$$\lim_{x\to p}(g \circ f)(x) = g(L).$$

Proof: We use Theorem 3.11. Let $A \subset X$ such that $p \sim A - \{p\}$. Then, since $\lim_{x\to p} f(x) = L$, we have by Theorem 3.11 that

$$L \sim f(A - \{p\}).$$

Thus, since g is continuous at L, the definition of continuity (above Example 2.23) gives us that

$$g(L) \sim g[f(A - \{p\})].$$

Hence, we have proved that for any $A \subset X$ such that $p \sim A - \{p\}$,

$$g(L) \sim (g \circ f)(A - \{p\}).$$

Therefore, by Theorem 3.11, $\lim_{x\to p}(g \circ f)(x) = g(L)$. ¥

Theorem 4.28 follows immediately from Theorem 4.29 using the characterization of continuity in Corollary 3.13. Nevertheless, we presented Theorem 4.28 first since it is less technical than Theorem 4.29 and since it is obviously the origin for Theorem 4.29.

The analogue of Theorem 4.29 for limits as x approaches infinity is in Theorem 18.6. It may enhance your understanding of Theorem 4.29 to read the proof of Theorem 18.6 now and adapt the proof to give an "epsilon - delta proof" of Theorem 4.29.

There is a natural question to ask about Theorem 4.29. It is the question of whether the analogous theorem is true when we interchange the assumptions about f and g; that is, assume f is continuous at p and $\lim_{y\to f(p)} g(y) = L$, and then conclude that $\lim_{x\to p}(g \circ f)(x) = L$. Of course, the assumption that $\lim_{y\to f(p)} g(y) = L$, makes no sense unless $f(p)$ is a limit point of Y (recall the definition of limit at the beginning of Chapter III). So, let's add the assumption that $f(p)$ is a limit point of Y to our other assumptions here. Now, what can go wrong? We see the problem when we try to prove the result:

Let $A \subset X$ such that $p \sim A - \{p\}$. Then, by our assumption that f is continuous at p,

$$f(p) \sim f(A - \{p\}).$$

Now, according to Theorem 3.11, we must prove $L \sim g(f(A - \{p\}))$ in order to know that $\lim_{x\to p}(g \circ f)(x) = L$. The definition of arbitrary closeness (section 1 of Chapter II) says that $L \sim g(f(A - \{p\}))$ means

$$dist(L, g(f(A - \{p\})) = 0.$$

Could $dist(L, g(f(A - \{p\})) > 0$? In other words, could $g(f(A - \{p\}))$ be bounded away from L? The simplest thing that could happen that would make $g(f(A - \{p\}))$ be bounded away from L is the following: $f(A - \{p\})$ is a single point q at which g jumps and for which $g(q) \neq L$. This suggests considering f to be the constant function $f(x) = q$ and letting g be a function that jumps at q but for which $g(q) \neq L = \lim_{y \to q} g(y)$. You should now be prepared to write down a counterexample to the analogue of Theorem 4.29 that we have tried to prove:

Exercise 4.30: Using the preceding discussion as a guide, find functions $f, g : \mathsf{R}^1 \to \mathsf{R}^1$ such that f is continuous at $p = 0$, $\lim_{y \to f(p)} g(y) = L$, but $\lim_{x \to p}(g \circ f)(x) \neq L$.

Exercise 4.31: If $f : \mathsf{R}^1 \to \mathsf{R}^1$ is a function such that $\lim_{x \to p} f(x)$ exists for some point $p \in \mathsf{R}^1$, then $\lim_{x \to p} |f(x)| = |\lim_{x \to p} f(x)|$.

Exercise 4.32: If $f : \mathsf{R}^1 \to \mathsf{R}^1$ is a function such that $\lim_{x \to p} f(x)$ exists for some point $p \geq 0$, then $\lim_{x \to p} \sqrt{f(x)} = \sqrt{\lim_{x \to p} f(x)}$.

Exercise 4.33: For any two functions $f, g : \mathsf{R}^1 \to \mathsf{R}^1$, define the *maximum function of f and g*, written $f \bigvee g$, and *the minimum function of f and g*, written $f \bigwedge g$, as follows: For each $x \in \mathsf{R}^1$,

$$(f \bigvee g)(x) = \max\{f(x), g(x)\}, \quad (f \bigwedge g)(x) = \min\{f(x), g(x)\}.$$

Prove that if f and g are continuous at p, then $f \bigvee g$ and $f \bigwedge g$ are continuous at p.

(*Hint:* What is $\frac{x+y}{2} + \frac{|x-y|}{2}$ for real numbers x and y?)

7. The Squeeze Theorem

The name Squeeze Theorem is very descriptive of what the theorem says: Consider three functions f, g and h defined on X into R^1 satisfying the conditions (where p is a limit point of X)

$$g(x) \leq f(x) \leq h(x), \text{ all } x \in X - \{p\}, \quad \lim_{x \to p} g(x) = \lim_{x \to p} h(x);$$

the conditions suggest that the function f is squeezed by the functions g and h as x approaches p; the Squeeze Theorem says that f is then forced to have the same limit as x approaches p that g and h have.

The Squeeze Theorem is one of those theorems that is easy to prove but that states an important and useful point of view. The idea of squeezing to obtain a limit will come up in other contexts (e.g., the definition of the integral in Chapter XII).

Theorem 4.34 (Squeeze Theorem): Let $X \subset \mathsf{R}^1$, and let $p \in \mathsf{R}^1$ such that p is a limit point of X. Assume that $f, g, h : X \to \mathsf{R}^1$ are functions such that

$$g(x) \leq f(x) \leq h(x), \text{ all } x \in X - \{p\}, \quad \lim_{x \to p} g(x) = \lim_{x \to p} h(x) = L.$$

Then $\lim_{x \to p} f(x) = L$.

Proof: We prove the theorem using only the definition of limit.

Let $\epsilon > 0$. Since $\lim_{x\to p} g(x) = L$ and $\lim_{x\to p} h(x) = L$, there exist $\delta_1, \delta_2 > 0$ such that

$$|g(x) - L| < \epsilon \text{ for all } x \in X - \{p\} \text{ such that } |x - p| < \delta_1$$

and

$$|h(x) - L| < \epsilon \text{ for all } x \in X - \{p\} \text{ such that } |x - p| < \delta_2.$$

Let $\delta = \min\{\delta_1, \delta_2\}$. Then, for all $x \in X - \{p\}$ such that $|x - p| < \delta$,

$$L - \epsilon < g(x), h(x) < L + \epsilon$$

Thus, since $g(x) \leq f(x) \leq h(x)$, we must have that $L - \epsilon < f(x) < L + \epsilon$ for all $x \in X - \{p\}$ such that $|x - p| < \delta$. This proves that $\lim_{x\to p} f(x) = L$. ¥

Exercise 4.35: Find $\lim_{x\to 0} x \sin(\frac{1}{x})$.

Exercise 4.36: Find $\lim_{x\to 0} \sqrt{x^3 + x + 1} \sin(\frac{1}{x})$.

8. Limits of Sequences

We briefly introduce sequences and limits of sequences for use later. We will present an in-depth study of the general theory of sequences beginning with Chapter XIX. In the meantime, we will not use sequences often, and the material we present here is all we need.

In this section, we will see that the standard definition of limits for sequences can be considered to be a special case of the notion of limits for functions as defined in section 1 of Chapter III. As a consequence, almost all theorems about limits of functions that we have proved are automatically true for limits of sequences (Theorem 4.38).

A *sequence* is simply a function defined on the set $\mathsf{N} = \{1, 2, ...\}$. The range of a sequence can consist of any types of objects (for example, a sequence of statements, a sequence of vectors, a sequence of numbers and statements together, and so on). If s is a sequence, we denote the value of s at n by s_n (i.e., $s_n = s(n)$). We often denote a sequence s by writing $\{s_n\}_{n=1}^{\infty}$.

We want to focus on sequences whose values are real numbers. Such sequences are sometimes called *numerical sequences*. We prefer to just use the term sequence rather than numerical sequence; thus, unless we say otherwise or it is evident from context, we assume that the values of a sequence are real numbers.

We say that a *sequence* $\{s_n\}_{n=1}^{\infty}$ *coverges to a point* p provided that for each $\epsilon > 0$, there exists N such that $|s_n - p| < \epsilon$ for all $n \geq N$; we call p the *limit of the sequence* $\{s_n\}_{n=1}^{\infty}$. We write $\lim_{n\to\infty} s_n = p$ or $\{s_n\}_{n=1}^{\infty} \to p$ to denote that a sequence $\{s_n\}_{n=1}^{\infty}$ converges to p.

A sequence that converges is called a *convergent sequence*; a sequence that does not converge is called a *divergent sequence*.

For example, the sequence $\{\frac{1}{n}\}_{n=1}^{\infty}$ converges to 0 by the second part of Exercise 1.23 with $n_i = n$ for each i. However, we remind the reader that the fact that $\{\frac{1}{n}\}_{n=1}^{\infty}$ converges to 0 (as well as Exercise 1.23) depends essentially on the Archimedean Property – see the discussion leading to Lemma 1.21.

On the other hand, the sequence $\{(-1)^n\}_{n=1}^{\infty}$ diverges (can you see why?).

We reformulate the definition of limits of sequences in terms of limits of functions as defined at the beginning of Chapter III. At first glance, the two notions of limit seem incompatible: limits of sequences are concerned with unbounded domain values, whereas limits of functions in Chapter III are involved only with bounded domain values. Nevertheless, the definition of $\lim_{x \to p} f(x)$ in Chapter III does *not* require p to be a point of the domain X of f; this is the underlying reason that we are able to formulate limits of sequences in terms of limits of functions as follows:

Exercise 4.37: Let $\{s_n\}_{n=1}^{\infty}$ be a sequence, and let $f : \{\frac{1}{n} : n \in \mathsf{N}\} \to R^1$ be the function defined by $f(\frac{1}{n}) = s_n$ for each $n \in \mathsf{N}$. Then $\{s_n\}_{n=1}^{\infty} \to p$ if and only if $\lim_{\frac{1}{n} \to 0} f(\frac{1}{n}) = p$.

Theorem 4.38: Results about limits in Chapter III and in this chapter apply to limits of sequences as well (except for Theorem 3.16).

Proof: The theorem is immediate from Exercise 4.37. ¥

We remark the fundamental notions of arbitrary closeness and continuity, which we introduced in Chapter II, can each be reformulated in terms of sequences. The reformulations are postponed until we begin a systematic study of sequences in Chapter XIX. The relevant results are Theorem 19.38 and Theorem 19.39; you are well prepared to read the results and their proofs now: the only background you need, aside from what we have already covered, is the analogue of Theorem 4.29 for limits as x approaches infinity (Theorem 18.6).

Chapter V: Two Properties of Continuous Functions

Of all the properties of continuous functions that I know, two properties stand out as being the most important. Both properties concern continuous functions on intervals: Let I be an interval, and let $f : I \to \mathsf{R}^1$ be a continuous function; then any number between two values of f is a value of f; if I is a closed and bounded interval, then f has a (unique) largest value, and a (unique) smallest value. We verify the two properties and give applications (in the exercises). We will see many more applications in subsequent chapters.

1. The Intermediate Value Theorem

A continuous function defined on a subset of R^1 may have two values such that no number between those two values is a value of the function. For example, let $X = \{0, 1\}$ and define f by letting $f(0) = 0$ and $f(1) = 1$. In fact, this kind of behavior can always happen when X is not an interval:

Exercise 5.1: Let X be any nonempty subset of R^1 such that X is not an interval. Then there is a continuous function $f : X \to \mathsf{R}^1$ such that for some two values, $f(a)$ and $f(b)$, no number between $f(a)$ and $f(b)$ is a value of f.

However, the behavior illustrated above can not occur when the domain of a continuous function is an interval:

Theorem 5.2 (Intermediate Value Theorem): Let I be an interval, and let $f : I \to \mathsf{R}^1$ be continuous. Then every number that lies between two values of f is a value of f. In other words, $f(I)$ is an interval.

Proof: Assume that $x_0, x_1 \in I$, with $x_0 < x_1$, and that $c \in \mathsf{R}^1$ such that

(1) $f(x_0) < c < f(x_1)$.

We prove that c is a value of f. (We will also need to prove that c is a value of f when $f(x_1) < c < f(x_0)$; we do this after we prove the theorem under the assumption in (1).)

We begin with a simple observation: Since $x_0, x_1 \in I$ and I is an interval, it is clear that

(2) $[x_0, x_1] \subset I$.

The following set A is central to the proof:

$$A = \{x \in [x_0, x_1] : f(x) < c\}.$$

By (1), $x_0 \in A$; hence, $A \neq \emptyset$. Also, x_1 is an upper bound for A by the way A is defined. Therefore, there is a least upper bound ℓ for A by the Completeness Axiom (section 1 of Chapter I). Since $x_0 \in A$ and x_1 is an upper bound for A, it is clear that

(3) $\ell \in [x_0, x_1]$.

We prove that $f(\ell) = c$ (note that f is defined at ℓ by (2) and (3)). We prove that $f(\ell) = c$ by proving (4) and (5) below.

(4) $f(\ell) \geq c$.

Proof of (4)*:* Suppose by way of contradiction that $f(\ell) < c$. Then, since f is continuous at ℓ, there is an open interval J such that $\ell \in J$ and $f(J) \subset (-\infty, c)$ (by Exercise 3.14). Since $f(\ell) < c$, we see from (1) that $\ell \neq x_1$; hence, by (3), $\ell < x_1$. Thus, there exists $p \in J$ such that $\ell < p < x_1$. Hence, by (3), $p \in [x_0, x_1]$. Therefore, by (2), $p \in I$, so f is defined at p. Thus, since $p \in J$, $f(p) \in (-\infty, c)$. Hence, since $p \in [x_0, x_1]$, we have that $p \in A$. Therefore, since $\ell < p$, we have a contradiction to ℓ being an upper bound for A. This proves (4).

(5) $f(\ell) \leq c$.

Proof of (5)*:* Suppose by way of contradiction that $f(\ell) > c$. Then, since f is continuous at ℓ, there is an open interval J' such that $\ell \in J'$ and $f(J') \subset (c, \infty)$ (by Exercise 3.14). Since $f(\ell) > c$, we see from (1) that $\ell \neq x_0$; hence, by (3), $\ell > x_0$. Thus, since ℓ is the *least* upper bound of A, there is a point $a \in A \cap J'$. Since $a \in A$, $f(a) < c$; on the other hand, since $a \in J'$, $f(a) > c$. This establishes a contradiction. Therefore, we have proved (5).

By (4) and (5), $f(\ell) = c$. This proves the theorem under the assumption in (1).

To complete the proof of the theorem, we must consider the case when (1) is changed to $f(x_1) < c < f(x_0)$ (and, as before, $x_0 < x_1$). We can prove the theorem for this case by going through the proof we have done and making the necessary adjustments; however, we prefer to sieze this opportunity to introduce a standard technique – a trick.

Observe that we have proved the theorem for *any* continuous function $g : I \to \mathsf{R}^1$ and for *any* point $b \in \mathsf{R}^1$ such that $g(x_0) < b < g(x_1)$, where $x_0, x_1 \in I$ and $x_0 < x_1$, We can use this to prove the theorem for the given function f when $f(x_1) < c < f(x_0)$, as follows:

Assume that $x_0, x_1 \in I$, with $x_0 < x_1$, and that $c \in \mathsf{R}^1$ such that the reverse inequalities in (1) hold, namely,

$$f(x_1) < c < f(x_0).$$

Consider the function $g = -f$. Then g is continuous by Corollary 4.10 and

$$g(x_0) < -c < g(x_1).$$

Hence, by the observation in the preceding paragraph, there is a point $p \in I$ such that $g(p) = -c$. Clearly, then, $f(p) = c$. ¥

Remember the "trick" employed in the last part of the proof of the Intermediate Value Theorem and use the trick to your advantage in the future.

The Intermediate Value Theorem is an existence theorem – it tells us that certain values of the function exist, but it does not tell us *where* in the domain a particular value of the function is attained. In addition, the proof is not constructive – the proof does not locate where a particular value is attained. For example, see Exercise 5.5.

We will see many applications of the Intermediate Value Theorem. We mention one application that is particularly important: We use the Intermediate Value Theorem to characterize one-to-one continuous functions on intervals as being either strictly increasing or strictly decreasing (Theorem 8.4); this leads to a proof of the Inverse Function Theorem (Theorem 8.7).

If X is any subset of R^1 and $f : X \to \mathsf{R}^1$ is continuous, then the Intermediate Value Theorem can be applied to intervals contained in X. We see why this is so from the first exercise below.

Exercise 5.3: The restriction of a continuous function is continuous (the restriction of a function is defined in the second paragraph of section 5 of Chapter III); in fact, if $f : X \to \mathsf{R}^1$ is continuous at a point $p \in X$ and if $X' \subset X$ such that $p \in X'$, then $f|X'$ is continuous at p.

Exercise 5.4: Use Theorem 5.2 to give a very short (and elegant) proof that every positive real number has a positive square root (which we proved earlier in Theorem 1.25).

Exercise 5.5: Let $f(x) = \frac{35}{\sqrt{2}x^{14} + 5x^{10} + 9x^8 + 3x^4 + 7}$. Show that $f(x) = \frac{10}{7}$ for some $x \in [0, 1]$.

Exercise 5.6: Prove that if f is a polynomial of odd degree, then f has a root (i.e., $f(x) = 0$ for some $x \in \mathsf{R}^1$).

Give an example of a polynomial of even degree that does not have a root.

Exercise 5.7: If $f : [0, 1] \to [0, 1]$ is continuous, then $f(p) = p$ for some point p. (For any function f, a point p such that $f(p) = p$ is called a *fixed point of the function* f.)

Exercise 5.8: Assume that $f : [0, 1] \to \mathsf{R}^1$ is continuous, $f(0) \leq 0$ and $f(1) \geq 1$. Then the equation $f(x) = x^2$ has a solution.

Exercise 5.9: You leave your home at 8 P.M. and walk to a friend's home, arriving at 8 : 30 P.M. You stay overnight, and the next evening you leave your friend's home at 8 P.M. and arrive home at 8 : 30 P.M., retracing exactly the same route as the evening before. At some time between 8 P.M. and 8 : 30 P.M., you are at exactly the same place on the route both evenings. Why?

Exercise 5.10: No interval is the union of two or more (including infinitely many) mutually disjoint open intervals.

(*Hint:* Assume to the contrary, and find a continuous function that contradicts the Intermediate Value Theorem (Theorem 5.2).)

2. The Maximum - Minimum Theorem

A continuous function defined on an interval may not have a largest or smallest value (e.g., $f(x) = x$ for $0 < x < 1$). On the other hand, when the interval is closed and bounded, any continuous function on the interval has *both* a largest value and a smallest value; this result is called the Maximum - Minimum Theorem. The theorem fails for bounded intervals that are not closed (by the example above) and for closed intervals that are not bounded (by restricting $f(x) = x$ to the closed interval $[1, \infty)$).

We devote this section to proving the Maximum - Minimum Theorem, which is Theorem 5.13.

We introduce terminology that we use throughout the section (and later).

Definition: We define bounded set, bounded function, maximum value and minimum value (extreme values).

- A subset X of R^1 is a *bounded set* provided that there exists $M > 0$ such that $X \subset (-M, M)$.
- A function $f : X \to \mathsf{R}^1$ is a *bounded function*, or is *bounded on* X, provided that $f(X)$ is a bounded set.
- The *maximum value*, or *largest value*, of a function $f : X \to \mathsf{R}^1$ is the value $f(p)$, if it exists, such that $f(p) \geq f(x)$ for all $x \in X$.
- The *minimum value*, or *smallest value*, of a function $f : X \to \mathsf{R}^1$ is the value $f(p)$, if it exists, such that $f(p) \leq f(x)$ for all $x \in X$.
- Taken together, the maximum value and the minimum value of a function (if they exist) are called the *extreme values* of the function.

We first prove a general theorem commonly called the Nested Interval Property. The Nested Interval Property seems to have nothing to do with the subject at hand; however, the property is the basis for the proof the Maximum - Minimum Theorem. The proof of the Nested Interval Property only uses the Completeness Axiom; thus, we could have presented the result and its proof back in Chapter I. We waited until now in order to give an immediate application of the result to a situation that is far removed from the result itself. Thus, the use of the Nested Interval Property to prove the Maximum - Minimum Theorem illustrates the many diverse and unexpected applications of the Nested Interval Property.

Theorem 5.11 (Nested Interval Property): Let $I_1, I_2, ..., I_n, ...$ be closed and bounded intervals such that $I_n \supset I_{n+1}$ for each n. Then

$$\cap_{n=1}^{\infty} I_n \neq \emptyset.$$

Proof: For each $n = 1, 2, ...$, let $I_n = [a_n, b_n]$. Since $I_n \supset I_{n+1}$ for each n, we see easily that

(1) $a_i \leq b_j$ for all $i, j = 1, 2, \ldots$.

Let $A = \{a_n : n = 1, 2, ...\}$. By (1), b_1 is an upper bound for A. Therefore, since $A \neq \emptyset$, A has a least upper bound ℓ (by the Completeness Axiom).

We show that $\ell \in \cap_{n=1}^{\infty} I_n$, which proves the theorem.

Since ℓ is an upper bound for A, $a_n \leq \ell$ for each n. Hence, to show that $\ell \in \cap_{n=1}^{\infty} I_n$, we are left to show that $\ell \leq b_n$ for each n. But this is easy to show: If it were true that $\ell > b_j$ for some j, then, since b_j is an upper bound for A by (1), ℓ would not be the *least* upper bound for A. ¥

The Nested Interval Property is simply another way of stating the Completeness Axiom; you will be asked to prove this (Exercise 5.17).

Next, we prove a preliminary lemma. The lemma will be subsumed by our main theorem; nevertheless, the lemma is a convenient way to break the proof of our main theorem into two parts.

Lemma 5.12: If $f : [a, b] \to \mathsf{R}^1$ is continuous, then f is bounded.

Proof: Let

$$A = \{x \in [a, b] : f([a, x]) \text{ is bounded}\}.$$

We prove the lemma by proving that $b \in A$.

Since $a \in A$, $A \neq \emptyset$; also, b is an upper bound for A. Therefore, by the Completeness Axiom, there is a least upper bound ℓ for A.

We prove that $b \in A$ by proving that $\ell \in A$ and then proving that $\ell = b$.

Note that $\ell \in [a, b]$ (since $a \in A$ and b is an upper bound for A). Hence, f is defined at ℓ and, therefore, f is continuous at ℓ. Thus, by Exercise 3.14, there is an open interval $J = (s, t)$ in R^1 such that

(1) $\ell \in J = (s, t)$ and $f(J) \subset (f(\ell) - 1, f(\ell) + 1)$.

We now prove that $\ell \in A$ (draw a picture as the proof progresses). Since $a \in A$, we assume for the proof that $\ell \neq a$; hence, $\ell \in (a, b]$. Thus, since $\ell \in J = (s, t)$ (by (1)) and since ℓ is the least upper bound for A, there is a point $x \in A$ such that $s < x \leq \ell$. Since $x \in A$, $f([a, x])$ is bounded. Also, since $[x, \ell] \subset J$ (by (1)), $f([x, \ell])$ is bounded (by (1)). Thus, since the union of two bounded sets is bounded and since $f([a, x]) \cup f([x, \ell]) = f([a, \ell])$, we have that $f([a, \ell])$ is bounded. Therefore, since $\ell \in [a, b]$, we have proved that $\ell \in A$.

Finally, we show that $\ell = b$. Suppose by way of contradiction that $\ell \neq b$. Then, since $\ell \in [a, b]$, $\ell < b$. Thus, since $\ell \in J$ (by (1)), there is a point $z \in [a, b] \cap J$ such that $z > \ell$. Since $\ell, z \in J$, we see that $[\ell, z] \subset J$; hence, by (1), $f([\ell, z])$ is bounded. Also, since $\ell \in A$ (proved above), $f([a, \ell])$ is bounded. Hence, $f([a, \ell]) \cup f([\ell, z])$ is bounded. Thus, $f([a, z])$ is bounded. Therefore, since $z \in [a, b]$, we have proved that $z \in A$. Thus, since $z > \ell$, we have a contradition to the fact that ℓ is an upper bound for A. Therefore, $\ell = b$.

We have proved that $\ell \in A$ and that $\ell = b$. Hence, $b \in A$. Therefore, from the definition of A, we see that f is bounded. ¥

We are ready to prove our main theorem. The proof illustrates an important technique - the bisection procedure - that has numerous applications in calculus as well as in other areas such as continuum theory, dynamics and chaos.

Theorem 5.13 (Maximum - Minimum Theorem): If $f : [a, b] \to \mathsf{R}^1$ is continuous, then f has a (unique) maximum value and a (unique) minimum value. Thus, $f([a, b])$ is a closed and bounded interval.

Proof: Since the theorem is trivially true when $a = b$, we assume for the proof that $a < b$. We prove that f has a maximum value; the proof that f has a minimum value is left for the reader (Exercise 5.14).

By Lemma 5.12, $f([a, b])$ is bounded. Therefore (since $f([a, b])$ is nonempty), we have by the Completeness Axiom that $f([a, b])$ has a least upper bound ℓ.

We prove that ℓ is a value of f; obviously, then, ℓ is the maximum value of f.

We inductively define closed and bounded intervals $I_1 \supset I_2 \supset \cdots \supset I_n \supset \cdots$ by bisecting, as follows: Let $I_1 = [a, b]$, and note that $\ell = lub\, f(I_1)$. Assume inductively that we have defined a closed and bounded interval $I_n = [a_n, b_n]$ for some $n \geq 1$ such that $\ell = lub\, f(I_n)$. Let m denote the midpoint of I_n (i.e., $m = \frac{a_n + b_n}{2}$). Then, since $\ell = lub\, f(I_n)$, we see easily that

$$\text{(i) } \ell = lub\, f([a_n, m]) \quad \text{or} \quad \text{(ii) } \ell = lub\, f([m, b_n]).$$

Define I_{n+1} to be $[a_n, m]$ if (i) holds and define I_{n+1} to be $[m, b_n]$ if (ii) holds and (i) does not hold. Then, by the Induction Principle (Theorem 1.20), we have defined I_n for each $n = 1, 2, \ldots$.

The intervals I_n have the following three important properties, each of which follows easily from the way we defined the intervals:

(1) For each n, $I_n \supset I_{n+1}$ and I_n is a closed and bounded interval;

(2) the length of I_n is $\frac{1}{2^{n-1}}(b - a)$ for each n;

(3) $\ell = lub\, f(I_n)$ for each n.

By (1) and the Nested Interval Property (Theorem 5.11), there is a point $p \in \cap_{n=1}^{\infty} I_n$.

We prove that $f(p) = \ell$. Suppose, as will lead to a contradiction, that $f(p) \neq \ell$. Then, since ℓ is an upper bound for $f([a, b])$, $f(p) < \ell$. Hence,

$$f(p) \in (-\infty, \tfrac{f(p)+\ell}{2}).$$

Thus, since f is continuous at p, we see from Exercise 3.14 that there is an open interval J such that $p \in J$ and

$$f(J) \subset (-\infty, \tfrac{f(p)+\ell}{2}).$$

Since J is an open interval containing p and since $p \in I_n$ for all n, it follows from (2) that $I_k \subset J$ for some k. Therefore,

$$f(I_k) \subset (-\infty, \tfrac{f(p)+\ell}{2}).$$

Hence, $\frac{f(p)+\ell}{2}$ is an upper bound for $f(I_k)$. Thus, since $\frac{f(p)+\ell}{2} < \ell$, we have a contradiction to (3). Therefore, $f(p) = \ell$.

We have proved that ℓ is the maximum value of f. Finally, assuming we have proved that f has a minimum value m (Exercise 5.14), we see that $f([a,b])$ is a closed and bounded interval: For by Theorem 5.2, $f([a,b]) = [m,\ell]$. ¥

The Maximum - Minimum Theorem and its proof have the same inherent characteristics as the Intermediate Value Theorem and its proof: The Maximum - Minimum Theorem is an existence theorem, and its proof does not tell us what the extreme values of f are or where in the domain $[a,b]$ they are attained. Let us illustrate. Consider the function f defined on the interval $[0,6]$ by

$$f(x) = 4x^3 - 36x^2 + 77x;$$

we know from the Maximum - Minimum Theorem that f has extreme values on $[0,6]$; however, we do not know (at this time) what the extreme values are or at which points in $[0,6]$ they are attained. See if you can find them, even with a hand calculator (but, *no* calculus allowed!).

Finding extreme values and where they are attained is a very important problem; differential calculus is designed to provide solutions to the problem. We will return to the problem of finding extreme values in our study of derivatives. We focus on the problem in Chapters IX and X (you will be asked to analyze the example above in Exercises 10.21 and 10.38).

Exercise 5.14: Finish the proof of Theorem 5.13 by proving that f has a minimum value.

Exercise 5.15: There are two positive real numbers such that the sum of their squares is $\sqrt{3}$ and such that their product is as large as possible.

Exercise 5.16: Use the bisection procedure in the proof of Theorem 5.13 to prove that any bounded infinite subset of R^1 has a limit point in R^1.

Exercise 5.17: Prove that the Nested Interval Property (Theorem 5.11) is equivalent to the Completeness Axiom. (The proof of Theorem 5.11 shows that the Completeness Axiom implies the Nested Interval Property).

Exercise 5.18: In addition to the assumptions for the Nested Interval Property (Theorem 5.11), assume that the length of the interval I_n is less than $\frac{1}{n}$. Then $\cap_{n=1}^{\infty} I_n$ contains exactly one point.

Exercise 5.19: True or false: If $f : (a,b] \to \mathsf{R}^1$ is continuous, then f has a maximum value or f has a minimum value.

Exercise 5.20: If X is an unbounded subset of R^1, then there is a continuous function $f : X \to \mathsf{R}^1$ such that f is unbounded.

Exercise 5.21: Is there a continuous function on R^1 that attains each of its values exactly twice?

Exercise 5.22: Is there a continuous function on R^1 that attains each of its values exactly three times?

Chapter VI: Introduction to the Derivative

The derivative of a function at a point is a general notion with numerous interpretations. The two most prominent interpretations are the instantaneous rate of change of a function at a point and the slope of the tangent line to the graph of a function at a point. In section 1 we first discuss the physical and geometric ideas that lead to the definition of the derivative; then we present the formal definition of the derivative and illustrate the definition in connection with tangent lines. In section 2 we relate differentiability to continuity. In the last section, we discuss linear approximation.

1. Definition of the Derivative

The definition of the derivative of a function took almost two millennia to be developed and rigorously understood. The notion comes from classical geometry.

For centuries, geometers were concerned with finding tangents to surfaces. Apollonius (262 - 190 B.C.) constructed tangents to conic sections. R. Descartes (1596 - 1650) used tangents to circles to find tangents to curves by "fitting" a circle to a point on the curve and declaring the tangent to the circle at the point to be the tangent to the curve at the point. P. Fermat (1601 - 1665) found tangents to curves using the so - called difference quotient we use today.[4]

But tangents to curves are not just a curiosity for geometers: Tangents to curves can describe physical action. We mention two examples. The first example concerns the tangent line itself, and the second example concerns the slope of the tangent line.

A tangent line to a curve can be interpreted as the path along which an object would naturally move were the object not constrained. You can see this experimentally: Attach a small weighty object to one end of a piece of string, twirl the object while holding the other end of the string, and then let go – the object goes in the direction tangent to its originally circular path at the point where it was released.

The slope of the tangent line to a curve at a point can be thought of as representing the velocity of a particle at the point. But wait! We all know what velocity is – $\frac{\text{distance}}{\text{time}}$ – but what is *velocity at a point* – $\frac{0}{0}$? Not hardly! And this is where the notion of limit steps in: Let $d(t)$ denote the distance a particle has moved from its initial position at time $t = 0$ to its position at time $t > 0$; assume that the particle is at a point p at time $t = t_0$. Then the velocity $v(p)$ of the particle at the point p should be the limit of the velocities over times $t \neq 0$ as $t \to t_0$ (if the limit exists):

$$v(p) = \lim_{t \to 0} \frac{d(t_0 + t) - d(t_0)}{t} .$$

[4] Fermat used difference quotients without the notion of limit, which came later. The modern day definition of limit is due to K. Weierstrass (1815 - 1897), but the essence of the notion is traceable back to I. Newton (1642 - 1727), who thought in terms of "ultimate ratios" of infinitesimal increments (without a definition).

We call $v(p)$ the *instantaneous velocity of the particle at* $x = p$, and we call $\frac{d(t_0+t)-d(t_0)}{t}$ the *average velocity of the particle over time* t ($t \neq 0$). Thus, the instantaneous velocity at p is the limit of the average velocities.

Now, note that for each time $t \neq 0$, the expression $\frac{d(t_0+t)-d(t_0)}{t}$ is the slope of the secant line joining the points $(t_0, d(t_0))$ and $(t_0 + t, d(t_0 + t))$ on the graph of the function d. The limit of these slopes, $\lim_{t\to 0} \frac{d(t_0+t)-d(t_0)}{t}$, should be considered to be the slope of the tangent line to the graph of the function d at $(t_0, d(t_0))$.

We conclude that we have two interpretations for $\lim_{t\to 0} \frac{d(t_0+t)-d(t_0)}{t}$: the instantaneous velocity of a particle at time t_0, and the slope of the tangent line to the graph of the function d at $(t_0, d(t_0))$.

The derivative of a function (which we define below) is merely a general formulation of what we have described. Indeed, there is no reason whatsoever to restrict what we have done to the setting of moving particles or to slopes of tangent lines. Nevertheless, the preceding discussion gives us an intuitive understanding, a point of reference if you like, for the definition of derivative.

Definition: Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in X$ such that p is a limit point of X. We say that f *is differentiable at* p provided that the limit

$$\lim_{h\to 0} \frac{f(p+h)-f(p)}{h}$$

exists, in which case we call the limit the *derivative of* f *at* p, denoted by $f'(p)$.

We say that f is *differentiable on* X (or just *differentiable* when the domain X is evident) provided that f is differentiable at each point of X.

If f is differentiable on X, then the *derivative of* f (*on* I) is the function f' that assigns the value $f'(x)$ to each point $x \in X$.

With the discussion above in mind, we sometimes use descriptive terminology: We call $f'(p)$ the *slope of the tangent line to the graph of* f *at* $(p, f(p))$; when we consider f to be the distance an object has traveled with respect to a variable x (usually time), we call $f'(p)$ the *instantaneous velocity of the object at* $x = p$. The descriptive phrases are no longer just intuitive ideas – as of now, they are defined to be the derivative of f at p.

We see that, in general, $f'(p)$ can be thought of as the *instantaneous rate of change of* f *at* p. Rate of change can refer to a number of physical quantities that change, for example, with time: The size of a population, the financial return on an investment, the amount of rainfall, the amount of a product produced in a chemical reaction or in a business, etc. The study of derivatives is, therefore, the study of many ideas at the same time. This illustrates a major aspect of the beauty of mathematics – the ability of mathematics to unify a number of seemingly different ideas.

The definition of the derivative at a point presupposes that the point is in the domain of the function and that the point is a limit point of the domain. Thus, when we assume that a function is differentiable at a point, we do not explicitly mention the conditions about the point.

Before proceeding, we make two comments about the defintion of derivative.

Our first comment concerns terminology. Assume that $f : [a, b] \to \mathsf{R}^1$ is a differentiable function. Then, according to our terminology, $f'(a)$ and $f'(b)$ are derivatives. This terminology contrasts with the usual terminology: in other books, $f'(a)$ and $f'(b)$ are called one-sided derivatives (i.e., derivatives from the right or from the left, respectively). The reason we prefer our terminology goes back to our comments about limits in the last section of Chapter III (above Exercise 3.17). We will consider one-sided derivatives when there is a good reason to do so; for example, one situation in which one-sided derivatives come up naturally is in the next section.

Our second comment is a clarification concerning the limit in the definition of derivative. In order that $\lim_{h\to 0} \frac{f(p+h)-f(p)}{h}$ make sense, 0 must be a limit point of $\{h \in \mathsf{R}^1 : p + h \in X\}$ (as required in the definition of limit in section 1, Chapter III); the reader should check that this is so:

Exercise 6.1: If $X \subset \mathsf{R}^1$ and p is a limit point of X, then 0 is a limit point of $\{h \in \mathsf{R}^1 : p + h \in X\}$.

We conclude the section with three examples. The examples illustrate various aspects of the definition of the derivative in relation to tangent lines.

We have defined the slope of the tangent line to the graph of a function f at $(p, f(p))$ to be $f'(p)$; thus, we had better make sure that when the graph of f itself is a line, then $f'(p)$ is the slope of the line (in the sense of precalculus):

Example 6.2: Let $f(x) = mx + b$, the slope-intercept form of a line with slope m. We show that $f'(p) = m$ for any point $p \in \mathsf{R}^1$:

$$f'(p) = \lim_{h\to 0} \frac{f(p+h)-f(p)}{h} = \lim_{h\to 0} \frac{m(p+h)+b-(mp+b)}{h} = \lim_{h\to 0} \frac{mh}{h}$$
$$= \lim_{h\to 0} m = m.$$

Our next example illustrates how to find the equation of the tangent line to the graph of a function at a point of the graph.

Example 6.3: Let $f(x) = x^2$ (all $x \in \mathsf{R}^1$). We find the equation of the tangent line to the graph of f at the point $(3, 9)$. We first compute the derivative of f: For any given point x,

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} \frac{(x+h)^2-x^2}{h} = \lim_{h\to 0} \frac{2xh+h^2}{h}$$
$$= \lim_{h\to 0} 2x + h \overset{4.16}{=} 2x + 0 = 2x.$$

Thus, $f'(3) = 6$. Therefore, the equation of the tangent line to the graph of f at the point $(3, 9)$ is $y - 9 = 6(x - 3)$, or $y = 6x - 9$.

In geometry we are accustomed to a tangent line being on "one side" of the curve and only touching the curve at the point of tangency. The following example shows that tangent lines as we have defined them do not always behave that way:

Example 6.4: Let $f(x) = x^3$ (all $x \in \mathsf{R}^1$). Then, for any given point x,

$$f'(x) = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = \lim_{h\to 0} \frac{x^3+3x^2h+3xh^2+h^3-x^3}{h}$$
$$= \lim_{h\to 0} \frac{h(3x^2+3xh+h^2)}{h} = \lim_{h\to 0} 3x^2+3xh+h^2 \overset{4.16}{=} 3x^2.$$

Hence, $f'(0) = 0$, so the equation of the tangent line to the graph of f at $(0,0)$ is $y = 0$; therefore, since $x^3 < 0$ when $x < 0$ and $x^3 > 0$ when $x > 0$, the tangent line crosses the graph of f – the tangent line is not even locally on one side of the graph of f. Next, note that since $f'(1) = 3$, the equation of the tangent line to the graph of f at $(1,1)$ is $y = 3x - 2$; therefore, the tangent line intersects the graph of f at the two points $(1,1)$ and $(-2,-8)$.

Exercise 6.5: Find the points at which the function $f(x) = \frac{x}{x+1}$ is differentiable and find its derivative at those points.

Exercise 6.6: Find the points at which the function $f(x) = \sqrt{x}$ is differentiable and find its derivative at those points.

Exercise 6.7: Find the points at which the function $f(x) = \sqrt{x^2+1}$ is differentiable and find its derivative at those points.

Exercise 6.8: Let $f(x) = \frac{1}{\sqrt{x}}$. Find the equation of the tangent line to the graph of f at the point $(4, \frac{1}{2})$.

Exercise 6.9: Assume that f is a function defined on an open interval I and that f is differentiable at some point $p \in I$ with $f'(p) \neq 0$. Then there exists $\delta > 0$ such that for all $x \in I$ with $x \neq p$ and $|x-p| < \delta$, $f(x) \neq f(p)$.

Exercise 6.10: Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in X$ such that p is a limit point of X. Then f is differentiable at p if and only if $\lim_{x\to p} \frac{f(x)-f(p)}{x-p}$ exists, in which case $\lim_{x\to p} \frac{f(x)-f(p)}{x-p} = f'(p)$.

Exercise 6.11: Assume that f is a function defined on an open interval I and that f is differentiable at some point $p \in I$. Find

$$\lim_{h\to 0} \frac{f(p+h)-f(p-h)}{h}.$$

Exercise 6.12: Assume that f is a function defined on an open interval I and that f is differentiable at some point $p \in I$. Find

$$\lim_{h\to 0} \frac{f(p+2h)-f(p)}{h}.$$

Exercise 6.13: Let f be a function defined on an open interval I such that f is differentiable at some point $p \in I$. Let φ be a function defined on an open interval J about 0 such that φ is continuous at 0, $\varphi(0) = 0$, and $\varphi(x) \neq 0$ when $x \neq 0$. Then

$$\lim_{h\to 0} \frac{f(p+\varphi(h))-f(p)}{\varphi(h)} = f'(p).$$

(*Hint:* The formula

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

defines a function g on some open interval about 0. Make use of g.)

2. Differentiability and Continuity

In the forthcoming discussion, we assume that the functions are defined on intervals (or at least on sets that have no isolated points).

Continuity and differentiability can be thought of as smoothness conditions on the graph of a function: Continuity says the graph of a continuous function is smooth enough so that the function does not jump; differentiability says the graph of a differentiable function is smooth enough to have a (unique) tangent line at each point. The natural question is Are the two smoothness conditions related? In other words, does differentiability imply continuity, does continuity imply differentiability, or does neither imply the other?

It seems reasonable to expect that differentiability implies continuity: if the graph of a function is smooth enough to have a (unique) tangent line at each point of its graph, then the graph of the function should be smooth enough to keep the function from jumping. However, our experience with Example 6.4 shows that tangent lines do not always work the way we expect, so we had better proceed with caution.

We want to try to show that if $\lim_{h\to 0} \frac{f(p+h)-f(p)}{h}$ exists, then $\lim_{x\to p} f(x) = f(p)$ (recall Theorem 3.12). To have a chance to prove this, we need to write x in terms of h (or h in terms of x); we have essentially already done this in Exercise 6.10, which says that

$$\lim_{h\to 0} \frac{f(p+h)-f(p)}{h} = \lim_{x\to p} \frac{f(x)-f(p)}{x-p}.$$

Now, can you see what to do next? Think about it before reading the proof of the theorem below.

Theorem 6.14: Let $X \subset \mathsf{R}^1$, and let $f : X \to \mathsf{R}^1$ be a function. If f is differentiable at p, then f is continuous at p.

Proof: We want to prove $\lim_{x\to p} f(x) = f(p)$ (recall Theorem 3.12). The key to answering the question in the preceding discussion is the observation that writing $\lim_{x\to p} f(x) = f(p)$ is the same as writing $\lim_{x\to p}(f(x) - f(p)) = 0$.

To see why what we just said works, recall that $\lim_{x\to p} \frac{f(x)-f(p)}{x-p} = f'(p)$ (by Exercise 6.10) and that $\lim_{x\to p}(x - p) = 0$ (by Theorem 4.16); then

$$\lim_{x\to p}(f(x) - f(p)) = \lim_{x\to p} \frac{f(x)-f(p)}{x-p}(x - p) \overset{4.9}{=} f'(p) \cdot 0 = 0.$$

Therefore, $\lim_{x\to p} f(x) = f(p)$. ¥

Now, having proved that differentiability implies continuity, we address the question of whether the converse is true: Does continuity imply differentiability? If you worked Exercise 6.6, you already know the answer is *no*: The function $f(x) = \sqrt{x}$ is continuous at every point of $[0, \infty)$, but the function is not differentiable at $p = 0$. Perhaps you think this example is not very satisfactory because 0 is an end point of the interval on which f is defined – indeed, unusual things can happen at end points. But we can extend the function f so that the resulting function is continuous on the entire real line and not differentiable at 0 : Simply let

$$g(x) = \begin{cases} \sqrt{x} & \text{, if } x \geq 0 \\ 0 & \text{, if } x < 0. \end{cases}$$

"OK," you say, "the example with g is fine, but maybe we should extend the notion of tangent line to include $x = 0$ as a tangent line to the graph of $f(x) = \sqrt{x}$ at the origin."

"Yes, we can do that," I reply, "but that's a subject for another time."

An example showing something is false is one thing; a general principle showing *why* it is false is quite another. What is a general underlying principle that would lead easily to many continuous functions that are not differentiable?

The key to answering the question is to carefully examine why g in the discussion above is not differentiable; if you do so, you will arrive naturally at the notion of one-sided derivatives, defined below, and the simple theorem that follows.

Definition. Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in X$ such that p is a limit point of $X \cap (-\infty, p]$. We say f is *differentiable from the left at* p, written $f'_-(p)$, provided that the function $f|X \cap (-\infty, p]$ is differentiable at p, in which case $f'_-(p)$ is the derivative of $f|X \cap (-\infty, p]$ at p. We call $f'_-(p)$ *the left-hand derivative of* f *at* p (when it exists).

Similarly, assuming that p is a limit point of $X \cap [p, \infty)$, we say f is *differentiable from the right at* p, written $f'_+(p)$, provided that the function $f|X \cap [p, \infty)$ is differentiable at p, in which case $f'_+(p)$ is the derivative of $f|X \cap [p, \infty)$ at p. We call $f'_+(p)$ *the right-hand derivative of* f *at* p (when it exists).

Theorem 6.15: Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in X$ such that p is a two-sided limit point of X. Then f is differentiable at p if and only if $f'_-(p) = f'_+(p)$, in which case

$$f'_-(p) = f'(p) = f'_+(p).$$

Proof: The theorem follows immediately from the theorem on one-sided limits (Theorem 3.16). ¥

We can now easily construct many functions that are continuous on R^1 but that are not differentiable at certain points. All we need to do is cut and paste: Start with two functions, f and g, that are continuous on R^1, that agree at a point p, but that have different derivatives at p (e.g., two functions whose graphs are straight lines with different slopes); then cut the domains at p and paste the restricted functions together, thereby forming the new function h given by

$$h(x) = \begin{cases} f(x) & \text{, if } x \leq p \\ g(x) & \text{, if } x > p. \end{cases}$$

The effect is that h has a "corner" in its graph at p which keeps h from being differentiable at p (but the "corner" does not keep h from being continuous). We give a specific example.

Example 6.16: Let

$$h(x) = \begin{cases} x & \text{, if } x \leq 0 \\ 3x & \text{, if } x > 0. \end{cases}$$

Then h is continuous (by Theorem 4.16 and Exercise 5.3 when $x \neq 0$, and by Theorem 3.12 and Theorem 3.16 when $x = 0$), and h is not differentiable at $x = 0$ by Theorem 6.15.

One-sided derivatives can be used to show that lines that really look like tangent lines are not tangent lines in the sense that we have defined them. Consider the function $f : \mathsf{R}^1 \to \mathsf{R}^1$ given by

$$f(x) = \begin{cases} x^2 & \text{, if } x \leq 0 \\ \sqrt{x} & \text{, if } x > 0. \end{cases}$$

Draw a picture of the graph of f and it will look like the x-axis is tangent to the graph of f at the origin. However, using one-sided derivatives (Theorem 6.15), we see that f is not differentiable at $x = 0$; thus, the x-axis is not tangent to the graph of f at the origin. Note that the x-axis *is* tangent to the graph of $f|(-\infty, 0]$ at the origin but that the x-axis is *not* tangent to the graph of $f|[0, \infty)$ at the origin.

We summarize the section in terms of our discussion at the beginning of the section: The smoothness that differentiability imposes on the graph of a function is stronger than the smoothness that continuity imposes on the graph.

Exercise 6.17: The function f given by $f(x) = |x|$ is continuous at every point of R^1, but f is not differentiable at 0.

Exercise 6.18: Let

$$f(x) = \begin{cases} x & \text{, if } x \leq 0 \\ x^2 + x & \text{, if } x > 0. \end{cases}$$

Is there a tangent line to the graph of f at the point $(0, 0)$?

Exercise 6.19: Are there constants a and b such that the function f given by

$$f(x) = \begin{cases} x^2 + 5 & \text{, if } x \leq 1 \\ ax + b & \text{, if } x > 1 \end{cases}$$

is differentiable?

Exercise 6.20: Are there constants a and b such that the function f given by

$$f(x) = \begin{cases} ax + 2 & \text{, if } x \leq b \\ x^3 + 3 & \text{, if } x > b \end{cases}$$

is differentiable?

Exercise 6.21: Give an example of a function that is continuous on R^1 and that is not differentiable at any integer.

Exercise 6.22: Give an example of a function that is continuous on R^1 and that is not differentiable at any of the points 1, $\frac{1}{2}$, $\frac{1}{3}$, ...

Exercise 6.23: Give an example of a function that is continuous on $[0, \infty)$ and that is not differentiable at 0, 1, $\frac{1}{2}$, $\frac{1}{3}$, ...

Exercise 6.24: Let

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

Determine all points x at which f is differentiable.

Exercise 6.25: True or false: If $f : \mathsf{R}^1 \to \mathsf{R}^1$ is a function such that $f'_-(0)$ and $f'_+(0)$ both exist, then f is continuous at 0.

Exercise 6.26: Find all differentiable functions $f : \mathsf{R}^1 \to \mathsf{R}^1$ such that $f \circ f = f$.

3. Linear Approximation

We refer to a function whose graph is a straight line as a *linear function* (not to be confused with linear functions in the setting of linear algebra).

We can immediately determine the exact value of a given linear function at any particular point; for example, if $f(x) = 2x + 7$, then $f(65) = 137$. On the other hand, it is more difficult, sometimes impossible, to determine exact values of other types of functions – for example, $\sqrt{65}$ and $\sin(65)$.

Consider the tangent line to graph of a differentiable function f at a point $(p, f(p))$; the tangent line is a fairly good approximation to the graph of f near the point $(p, f(p))$. Thus, tangent lines should provide a way to find fairly close approximate values for functions such as $\sqrt{x}$ and $\sin(x)$.

The general procedure of using tangent lines to find approximate values is called *linear approximation*. We describe the procedure as follows: We are given (perhaps implicitly) a differentiable function f and a point x_0 at which we want to approximate $f(x_0)$. We first find a point p close to x_0 for which we know the value $f(p)$. Next, we determine the equation of the tangent line to the graph of f at $(p, f(p))$. Finally, we use the formula for the tangent line to obtain the desired approximation.

We illustrate:

Example 6.27: We approximate $\sqrt{65}$ using linear approximation. We are implicitly given the function $f(x) = \sqrt{x}$. The function f is differentiable at every $x > 0$ and the derivative is $f'(x) = \frac{1}{2\sqrt{x}}$ (which you know if you worked Exercise 6.6). We know $\sqrt{64} = 8$ and $\sqrt{81} = 9$, so we choose $p = 64$ since 64 is closer to 65 than 81 is (we could choose 65.61, which we find by computing $(8.1)^2$, but our point here is to avoid such tedious computations). The equation of the tangent line to the graph of f at $(64, 8)$ is

$$y = \frac{1}{2\sqrt{64}}x + 4 = \frac{1}{16}x + 4.$$

We write the equation of the tangent line using function notation, $y(x) = \frac{1}{16}x+4$, to show explicitly that we consider y to be a function of x. Then, finally,

$$y(65) = \frac{65}{16} + 4 = \frac{129}{16},$$

which is our linear approximation to $\sqrt{65}$.

We could expand the approximation in Example 6.27 into its decimal equivalent, 8.0625. If we wanted accuracy only to three decimal places to the right of the decimal point, would we round up or round down? We can numerically determine the answer:

$$(8.062)^2 = 64.996 \quad \text{and} \quad (8.063)^2 = 65.012,$$

thus we would round down to 8.062. However, if you know what the graph of $f(x) = \sqrt{x}$ looks like, then you know that all tangent lines to the graph of f lie above the graph of f (except where they touch the graph); therefore, rounding down is a good bet when we use linear approximation to estimate the square root of any positive number.

Exercise 6.28: Approximate $(5.137)^3$ using linear approximation.

Assume you only want accuracy to two decimal places to the right of the decimal point; would you round your answer up or would you round your answer down? Explain why without finding the decimal representing $(5.137)^3$,

Exercise 6.29: Approximate $\sin(31°)$ using linear approximation. (Use that the derivative of $\sin(x)$ is $\cos(x)$ when x is radian measure, which we will prove in Theorem 8.20; $1° = \frac{\pi}{180}$ radians.)

Chapter VII: Derivatives of Combinations

We show that various combinations of differentiable functions (including compositions) are differentiable; we derive formulas for the derivatives of the combinations in terms of the derivatives of the functions separately. We apply our results to show that polynomials are differentiable and that rational functions are differentiable where they are defined.

1. Sums, Differences, Products and Quotients

We show that sums, differences, products and quotients of two differentiable functions are differentiable; in the process, we derive formulas for the derivatives of the combined functions in terms of the derivatives of the functions separately. We apply our results in the next section to show polynomials and rational functions are differentiable.

Theorem 7.1: Let $X \subset \mathsf{R}^1$, and let $f, g : X \to \mathsf{R}^1$ be functions. If f and g are each differentiable at p, then $f + g$ is differentiable at p and

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

Proof: Using Theorem 4.1 for the third equality below, we have

$$\begin{aligned}
&\lim_{h\to 0} \tfrac{(f+g)(p+h)-(f+g)(p)}{h} = \lim_{h\to 0} \tfrac{f(p+h)-f(p)+g(p+h)-g(p)}{h} \\
&= \lim_{h\to 0}[\tfrac{f(p+h)-f(p)}{h} + \tfrac{g(p+h)-g(p)}{h}] \\
&\overset{4.1}{=} \lim_{h\to 0} \tfrac{f(p+h)-f(p)}{h} + \lim_{h\to 0} \tfrac{g(p+h)-g(p)}{h} = f'(p) + g'(p). \quad ¥
\end{aligned}$$

Corollary 7.2: Let $X \subset \mathsf{R}^1$, and let $f_1, f_2, ..., f_n : X \to \mathsf{R}^1$ be finitely many functions. If each of the functions $f_1, f_2, ..., f_n$ is differentiable at p, then the sum function $f_1 + f_2 + \cdots + f_n$ is differentiable at p and

$$(f_1 + f_2 + \cdots + f_n)'(p) = f_1'(p) + f_2'(p) + \cdots + f_n'(p).$$

Proof: The corollary follows from Theorem 7.1 by a simple induction (much like the proof of Theorem 4.5). ¥

Theorem 7.3: Let $X \subset \mathsf{R}^1$, let $f, g : X \to \mathsf{R}^1$ be functions. If f and g are each differentiable at p, then $f - g$ is differentiable at p and

$$(f - g)'(p) = f'(p) - g'(p).$$

Proof: The proof is similar to the proof of Theorem 7.1 using Theorem 4.2 (instead of Theorem 4.1). ¥

We know from the previous two theorems that derivatives "distribute over" sums and differences. Furthermore, the proofs of the two theorems use nothing more than the corresponding results about limits. Therefore, since limits "distribute over" products (Theorem 4.9), it is natural to expect that derivatives would do the same; in other words, we should expect that if f and g are differentiable at p, then

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

So, let's try to verify the formula and see what happens:

$$(f \cdot g)'(p) = \lim_{h\to 0} \frac{(f\cdot g)(p+h)-(f\cdot g)(p)}{h} = \lim_{h\to 0} \frac{f(p+h)g(p+h)-f(p)g(p)}{h};$$

this might not look promising, but remember the trick we used in proving the limit theorem for products (Theorem 4.9)? We subtracted and added an expression that enabled us to isolate expressions that related directly to the assumptions in the theorem. Let's try that here. Since we want to isolate the difference quotients for f and g, let's subtract and add $f(p)g(p+h)$ to the numerator of the second difference quotient above. The limit then becomes

$$\lim_{h\to 0} \frac{f(p+h)g(p+h)-f(p)g(p+h)+f(p)g(p+h)-f(p)g(p)}{h}$$
$$= \lim_{h\to 0}[g(p+h)\frac{f(p+h)-f(p)}{h} + f(p)\frac{g(p+h)-g(p)}{h}]$$

This doesn't look at all like $f'(p)g'(p)$! In fact, since $\lim_{h\to 0} g(p+h) = g(p)$ by Theorem 6.14 (and Theorem 3.12), we have uncovered a completely unexpected formula:

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

(for this step, we are using the sum and product theorems for limits (Theorems 4.1 and 4.9)).

Thus, even though our initial guess about a formula for the derivative of a product was wrong, we have discovered the following theorem:

Theorem 7.4: Let $X \subset \mathsf{R}^1$, and let $f, g : X \to \mathsf{R}^1$ be functions. If f and g are each differentiable at p, then $f \cdot g$ is differentiable at p and

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

Proof: The proof is in the discussion above. ¥

It is an understatement to say that the formula in Theorem 7.3 is not intuitive. But, at the very least, could we have known that our original "formula" – $(f \cdot g)'(p) = f'(p)g'(p)$ – could not be true before we tried to prove it? Yes, if we had tried to apply our "formula" in any one of several simple cases, such as to the product $x \cdot x$ or even to the function x written as $1x$ (we already computed the relevent derivatives in Examples 6.2 and 6.3).

Do we now discard our false formula so no one will know we made such a silly mistake? No! We turn our mistake into a question: For what differentiable functions f and g on R^1 is it true that $(f \cdot g)'(x) = f'(x)g'(x)$ for all $x \in \mathsf{R}^1$? We return to this question later.

Next, we show that the quotient of two differentiable functions is differentiable and, at the same time, we derive a formula for the derivative of the quotient.

Note that a quotient $\frac{f}{g}$ can be viewed as the product $f \cdot \frac{1}{g}$. Therefore, to simplify the proof of our theorem about quotients, we first prove a lemma

concerning reciprocals. We did the same thing when we proved the theorem on limits of quotients in section 4 of Chapter IV (Lemma 4.19 and Theorem 4.20).

Lemma 7.5: Let $X \subset \mathsf{R}^1$, and let $g : X \to \mathsf{R}^1$ be a function. If g is differentiable at p and $g(p) \neq 0$, then $\frac{1}{g}$ is differentiable at p and

$$(\tfrac{1}{g})'(p) = \tfrac{-g'(p)}{[g(p)]^2}.$$

Proof: We begin by trying to get a feeling for what is going on:

$$(\tfrac{1}{g})'(p) = \lim_{h\to 0} \frac{\frac{1}{g(p+h)} - \frac{1}{g(p)}}{h} = \lim_{h\to 0} \tfrac{1}{h} \tfrac{g(p)-g(p+h)}{g(p+h)g(p)};$$

hence, isolating the difference quotient for g from the rest, we have

(1) $(\tfrac{1}{g})'(p) = \lim_{h\to 0}[\tfrac{g(p+h)-g(p)}{h} \tfrac{-1}{g(p+h)g(p)}]$.

Now, we see what to do: We evaluate the limits of the two quotients on the right-hand side of (1) separately.

Since g is differentiable at p, we have that

(2) $\lim_{h\to 0} \tfrac{g(p+h)-g(p)}{h} = g'(p)$.

Since g is continuous at p by Theorem 6.14, $\lim_{h\to 0} g(p+h) = g(p)$ by Theorem 4.29; thus, since $g(p) \neq 0$, $\lim_{h\to 0} \frac{1}{g(p+h)} = \frac{1}{g(p)}$ by Lemma 4.19. Therefore, by the limit theorem on products (Theorem 4.9), we have

(3) $\lim_{h\to 0} \tfrac{-1}{g(p+h)g(p)} = \tfrac{-1}{[g(p)]^2}$.

By (1), (2) and (3), we can apply the limit theorem on products again to obtain that

$$(\tfrac{1}{g})'(p) = g'(p)\tfrac{-1}{[g(p)]^2} = \tfrac{-g'(p)}{[g(p)]^2}.$$

Have we proved the lemma? Yes, except for a technical detail: Even though $g(p) \neq 0$, there may be values h for which $g(p+h) = 0$, in which case the expression $\frac{1}{g(p+h)}$, which we used throughout the proof, does not make sense. However, this is easy to take care of: As already observed above (3),

$$\lim_{h\to 0} g(p+h) = g(p) \neq 0;$$

thus, there is a $\delta > 0$ such that

$$|g(p+h)| > \tfrac{|g(p)|}{2} \quad \text{when } p+h \in X \text{ and } |h| < \delta;$$

hence, $g(p+h) \neq 0$ for such h. Therefore, by stipulating at the beginning of the proof that all values h in the proof are restricted to those for which $|h| < \delta$, we take care of the matter. ¥

Theorem 7.6: Let $X \subset \mathsf{R}^1$, and let $f, g : X \to \mathsf{R}^1$ be functions. If f and g are each differentiable at p and $g(p) \neq 0$, then $\frac{f}{g}$ is differentiable at p and

$$(\tfrac{f}{g})'(p) = \tfrac{g(p)f'(p) - f(p)g'(p)}{[g(p)]^2}.$$

Proof: Since $\frac{f}{g} = f \cdot \frac{1}{g}$,

$$(\tfrac{f}{g})'(p) = (f \cdot \tfrac{1}{g})'(p) \overset{7.4}{=} f(p)(\tfrac{1}{g})'(p) + \tfrac{1}{g(p)} f'(p)$$

$$\overset{7.5}{=} f(p)\tfrac{-g'(p)}{[g(p)]^2} + \tfrac{1}{g(p)} f'(p) = \tfrac{-f(p)g'(p) + g(p)f'(p)}{[g(p)]^2}. \quad ¥$$

Exercise 7.7: Assume that $(f + g)(x) = x^3 + 5x - 3$, where f and g are differentiable functions and $f'(4) = 2$. Find $g'(4)$.

Exercise 7.8: Assume that $(f \cdot g)(x) = \frac{3x}{x^2+8}$, where f and g are differentiable functions such that $f(2) = 4$ and $f'(2) = 5$. Find $g'(2)$.

Exercise 7.9: Assume that $\frac{f}{g}(x) = x^2 + 2x$, where f and g are differentiable functions such that $f(2) = 2$ and $f'(2) = 3$. Find $g'(2)$.

Exercise 7.10: Let f and g be differentiable functions with $g(x) \neq 0$ for all x. Assume that the equation of the tangent line to the graph of f at $(2, f(2))$ is $3x - y - 5 = 0$ and that the equation of the tangent line to the graph of $\frac{f}{g}$ at $(2, \frac{f}{g}(2))$ is $2x + y + 4 = 0$. Find the equation of the tangent line to the graph of g at $(2, g(2))$.

2. Differentiating Polynomials and Rational Functions

In Chapter IV we showed that polynomials and rational functions are continuous. We now prove these types of functions are differentiable.

Our results will follow immediately from theorems in the preceding section once we prove a lemma.

Lemma 7.11: The function $f(x) = x^n$ is differentiable for each $n = 1, 2, \ldots$. In fact, for each $n = 1, 2, \ldots$,

$$(x^n)' = nx^{n-1}.$$

Proof: We prove the lemma by induction (Theorem 1.20).

We already know that $f(x) = x$ is differentiable and that $x' = 1 = 1x^0$ (Example 6.2); in other words, the lemma is true when $n = 1$.

Assume inductively that the lemma is true for a given natural number k.

We show using our inductive assumption that $(x^{k+1})' = (k + 1)x^k$.

Note that $x^{k+1} = xx^k$ and that, by our inductive assumption, $(x^k)' = kx^{k-1}$. Thus, since $x' = 1$, we can apply Theorem 7.3 on products to obtain

$$(x^{k+1})' = (xx^k)' = x(x^k)' + (x^k)x' = x(kx^{k-1}) + x^k$$

$$= kx^k + x^k = (k + 1)x^k.$$

The lemma now follows from the Induction Principle (Theorem 1.20). ¥

Theorem 7.12: Every polynomial is differentiable. Furthermore, if

$$f(x) = c_0 + c_1x + c_2x^2 + c_3x^3 + \cdots + c_nx^n,$$

then

$$f'(x) = c_1 + 2c_2x + 3c_3x^2 + \cdots + nc_nx^{n-1}.$$

Proof: By Theorem 7.4 on products and Lemma 7.11, $(cx^m)' = cmx^{m-1}$ for any constant c and any $m = 1, 2, \ldots$. Therefore, the theorem follows from Corollary 7.2. ¥

Theorem 7.13: Every rational function is differentiable on its domain.

Proof: Every point of the domain of a rational function is a limit point of its domain (you are asked to prove this in Exercise 7.14). Therefore, our theorem follows from Theorem 7.12 and Theorem 7.6. ¥

We close with a word of caution about computing derivatives. We know from Lemma 7.11 that $(x^4)' = 4x^3$. However, this does not say that $((2x)^4)' = 4(2x)^3$; in fact, since $(2x)^4 = 16x^4$, we see from Lemma 7.11 and Theorem 7.4 on products that $((2x)^4)' = (16x^4)' = 64x^3$. In other words, in general, if f is differentiable, Lemma 7.11 does not tell us how to differentiate $(f(x))^n$ or even whether $(f(x))^n$ is differentiable. We will learn about this in the next section.

Exercise 7.14: Prove the statement *every point of the domain of a rational function is a limit point of its domain*, which we used in the proof of Theorem 7.13. (No fair using that polynomials have only finitely many roots).

Exercise 7.15: $(x^n)' = nx^{n-1}$ for each $n = -1, -2, \ldots$.

Exercise 7.16: Find $f'(2)$ for each of the following functions f:

$$f(x) = -4x^5 + \frac{2}{x^3} - 7;\ f(x) = \frac{3x^2-2x+1}{(2x-1)^2};\ f(x) = \frac{x}{(4x-6)^3}.$$

Exercise 7.17: Let $f(x) = \frac{x}{(1+\frac{1}{x})^2}$. Find the equation of the tangent line to the graph of f at $(1, f(1))$.

Exercise 7.18: Find a function whose derivative is $3x^5 - 2x^2 + 1$.

Exercise 7.19: Find a function whose derivative is $\frac{1}{x^3} - (4x^2 + 1)^3$.

Exercise 7.20: Is there a polynomial of degree 3 that has horizontal tangent lines to its graph at three different points?

Exercise 7.21: Recall our discussion of the bogus formula $(f \cdot g)'(x) = f'(x)g'(x)$ following Theorem 7.4. When do polynomials f and g satisfy the formula?

3. The Chain Rule

We have proved that the composition of two continuous functions is continuous (Theorem 4.28). We now prove that the composition of two differentiable functions is differentiable and derive a formula for the derivative of the composition. The formula is called the Chain Rule (Theorem 7.23). The Chain Rule is useful in computing derivatives and has far-reaching theoretical consequences.

We will see applications of the Chain Rule in the next chapter (e.g., proof of Theorem 8.16) and in other chapters as well.

Assume that f and g are differentiable functions. Let us try to find out what the derivative of the composition $g \circ f$ should be. First, using the form in Exercise 6.10 for appearance sake only,

$$(g \circ f)'(p) = \lim_{x \to p} \frac{g(f(x)) - g(f(p))}{x - p} \text{ (if the limit exists).}$$

Next, as we have done on numerous occasions, we manipulate algebraically to obtain expressions that relate to our assumptions. Since we are assuming that f is differentiable, we want the difference quotient $\frac{f(x)-f(p)}{x-p}$ to appear as part of what we take the limit of to get $(g \circ f)'(p)$. We force this to happen by multiplying and dividing the expression $\frac{g(f(x))-g(f(p))}{x-p}$ by $f(x) - f(p)$, thereby obtaining

$$(g \circ f)'(p) = \lim_{x \to p} \frac{g(f(x)) - g(f(p))}{f(x) - f(p)} \frac{f(x) - f(p)}{x - p}.$$

Like the proverbial ostrich, we bury our head in the sand in order to believe that we have not divided by 0. Since $\lim_{x \to p}[f(x) - f(p)] = 0$ by Theorems 6.14 and 3.12, $\lim_{x \to p} \frac{g(f(x))-g(f(p))}{f(x)-f(p)}$ looks a lot like $g'(f(p))$. If the limit is $g'(f(p)$, then we can apply our theorem on limits of products (Theorem 4.9) to arrive at

$$(g \circ f)'(p) = \lim_{x \to p} \frac{g(f(x)) - g(f(p))}{f(x) - f(p)} \lim_{x \to p} \frac{f(x) - f(p)}{x - p} = g'(f(p))f'(p).$$

We have found a possible formula for $(g \circ f)'(p)$; we have not verified the formula (or even proved that $g \circ f$ is differentiable) since we may have divided by 0 in our computations. The following lemma overcomes this obstacle: the lemma will allow us to avoid limits of quotients with $f(x) - f(p)$ in the denominator, thereby verifying that the formula is indeed correct (Theorem 7.23).

Lemma 7.22: Let $X, Y, Z \subset \mathsf{R}^1$, and let $f : X \to Y$ and $g : Y \to Z$ be functions. Assume that f is continuous at p and that g is differentiable at $f(p) = q$. Define $G : Y \to \mathsf{R}^1$ by

$$G(y) = \begin{cases} \frac{g(y)-g(q)}{y-q} & \text{, if } y \neq q \\ g'(q) & \text{, if } y = q. \end{cases}$$

Then $\lim_{x \to p} G(f(x)) = g'(q)$ and

$$G(f(x))\frac{f(x)-f(p)}{x-p} = \frac{g(f(x))-g(f(p))}{x-p}, \text{ all } x \in X - \{p\}.$$

Proof: Since g is differentiable at q, we see from Exercise 6.10 that

$$\lim_{y \to q} \frac{g(y)-g(q)}{y-q} = g'(q) = G(q);$$

hence, G is continuous at q by Theorem 3.12. Thus, since f is continuous at p and $f(p) = q$, $G \circ f$ is continuous at p by Theorem 4.28. Therefore, by Theorem 3.12,

$$\lim_{x\to p}(G\circ f)(x) = (G\circ f)(p) = G(f(p)) = G(q) = g'(q).$$

This proves the first part of the lemma.

To verify the equation in the second part of the lemma, let $x \in X - \{p\}$. Assume first that $f(x) \neq q$. Then, by the definition of G,

$$G(f(x)) = \tfrac{g(f(x))-g(q)}{f(x)-q};$$

thus, since $q = f(p)$,

$$G(f(x))\tfrac{f(x)-f(p)}{x-p} = \tfrac{g(f(x))-g(f(p))}{f(x)-f(p)}\tfrac{f(x)-f(p)}{x-p} = \tfrac{g(f(x))-g(f(p))}{x-p}.$$

This verifies the equation in the second part of the lemma when $f(x) \neq q$. Finally, when $f(x) = q$, we have $f(x) = f(p)$, so both sides of the equation are equal to 0. ¥

Theorem 7.23 (Chain Rule): Let $X, Y, Z \subset \mathsf{R}^1$, and let $f : X \to Y$ and $g : Y \to Z$ be functions. Assume that f is differentiable at p and that g is differentiable at $f(p) = q$. Then $g \circ f$ is differentiable at p and

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

Proof: All the work is done (we can apply Lemma 7.22 below since f is continuous at p by Theorem 6.14):

$$(g\circ f)'(p) \overset{6.10}{=} \lim_{x\to p}\tfrac{g(f(x))-g(f(p))}{x-p} \overset{7.22}{=} \lim_{x\to p} G(f(x))\tfrac{f(x)-f(p)}{x-p};$$

also, $\lim_{x\to p} G(f(x)) = g'(q)$ (by Lemma 7.22) and $\lim_{x\to p}\frac{f(x)-f(p)}{x-p} = f'(p)$ (by Exercise 6.10). Therefore, we can apply Theorem 4.9 on limits of products to obtain

$$(g\circ f)'(p) = g'(q)f'(p) = g'(f(p))f'(p).$$ ¥

We conclude by illustrating how to use the Chain Rule in finding derivatives.

Example 7.24: Let $f(x) = (4x+5)^{12}$. Note that $f = h \circ g$, where

$$g(x) = 4x + 5, \quad h(y) = y^{12}.$$

Hence, by the Chain Rule,

$$f'(x) = h'(g(x))g'(x) = 12(g(x))^{11}(4) = 48(4x+5)^{11}.$$

Exercise 7.25: Find $f'(3)$ for each of the following functions f:

$$f(x) = \tfrac{1}{(1-x)^5};\ \ f(x) = \sqrt{x^6+3x^2+1};\ \ f(x) = [x + (x-x^3)^6]^7.$$

Exercise 7.26: Assume that $(g\circ f)(x) = \frac{x}{x+1}$, where f and g are differentiable functions such that $f(1) = 4$ and $g'(4) = 5$. Find $f'(1)$.

Exercise 7.27: Assume that $(g\circ f)(x) = x^4 + 3x$, where f and g are differentiable functions such that $f(2) = 3$ and $f'(2) = 5$. Find $g'(3)$.

Chapter VIII: The Inverse Function Theorem

The Inverse Function Theorem is concerned with one - to - one differentiable functions defined on an interval. The theorem tells us when the inverse of such a function is differentiable and provides a formula for the derivative.

After necessary preliminary results, we prove the Inverse Function Theorem in section 3. We apply the theorem in section 4 to show that rational powers of x are differentiable (where defined and for $x \neq 0$). We study the trigonometric functions in section 5: We show that the trigonometric functions are differentiable, and then we apply the Inverse Function Theorem to show that the inverse trigonometric functions are differentiable. We obtain formulas for the derivatives of rational powers, trigonometric functions and inverse trigonometric functions.

1. One - to - one Functions and Inverses

We recall some notions and notation from precalculus.

Let X and Y be sets. A function $f : X \to Y$ is said to be *one - to - one* provided that whenever $x_1, x_2 \in X$ and $x_1 \neq x_2$, then $f(x_1) \neq f(x_2)$.

Assume that $f : X \to Y$ is one - to - one. Then we can define a function $g : f(X) \to X$ as follows: For each $y \in f(X)$, $g(y)$ is the unique point in X that maps to y under f. In other words, $f(g(y)) = y$ for all $y \in f(X)$; in addition, $g(f(x)) = x$ for all $x \in X$. The function g is called the *inverse of* f, which we denote from now on by f^{-1}.

Do not confuse the notation f^{-1} with $\frac{1}{f}$; f^{-1} is the (unique) function such that $f \circ f^{-1}$ is the identity function on $f(X)$ and $f^{-1} \circ f$ is the identity function on X.

Let $X \subset \mathsf{R}^1$, and let $f : X \to \mathsf{R}^1$ be one - to - one. Then the graph of f^{-1} is obtained by reflecting the graph of f about the line $y = x$ in the plane. The reason is quite simple: The reflection about $y = x$ changes a point $(x, f(x))$ to the point $(f(x), x)$, and $f^{-1}(f(x)) = x$.

The simple relation between the graphs of f and f^{-1} just mentioned can provide geometric intuition for the Inverse Function Theorem and for some results preceding it. In particular, examining the graphs of f and f^{-1} in the same picture can serve to motivate the results and provide insight. I leave it to the reader to draw pictures of continuous one - to - one functions on intervals, together with their inverses, and differentiable one - to - one functions on intervals, together with their inverses, before reading further – try to predict (from the pictures) a geometric characterization of one - to - one continuous functions on intervals, and try to determine what the formula should be for the derivative of f^{-1} in terms of the derivative of f.

2. Continuity of the Inverse Function

We prove that the inverse of a one - to - one continuous function on an interval is continuous. We will use this result in the next section to prove that the

inverse of a differentiable function (on an interval) is differentiable. The pattern should be familiar from the preceding chapter: There we used the continuity of compositions (in the proof of Lemma 7.22) in proving the Chain Rule.

Our result about continuity of the inverse function is Theorem 8.6. We prove the result by first characterizing one-to-one continuous functions defined on intervals in a geometric way. The terminology for the characterization is as follows:

Definition: Let $X \subset \mathsf{R}^1$ and let $f : X \to \mathsf{R}^1$ be a function. We say that f is *increasing on* X provided that whenever $x_1, x_2 \in X$ such that $x_1 < x_2$, then $f(x_1) \leq f(x_2)$; f is *strictly increasing on* X provided that whenever $x_1, x_2 \in X$ such that $x_1 < x_2$, then $f(x_1) < f(x_2)$.

Similarly, f is *decreasing on* X (or *strictly decreasing on* X) provided that whenever $x_1, x_2 \in X$ such that $x_1 < x_2$, then $f(x_1) \geq f(x_2)$ (or $f(x_1) > f(x_2)$, respectively).

If $Y \subset X$, we say f is *increasing* (*strictly increasing*, etc.) *on* Y to mean $f|Y$ is increasing (strictly increasing, etc.) on Y.

Exercise 8.1: Let $X \subset \mathsf{R}^1$ and let $f : X \to \mathsf{R}^1$ be a function. If f is strictly increasing on X, then f^{-1} is strictly increasing on $f(X)$; if f is strictly decreasing on X, then f^{-1} is strictly decreasing on $f(X)$.

It is obvious that if a function f is either strictly increasing or strictly decreasing, then f is one-to-one. Our characterization theorem says that the converse is also true when f is continuous on an interval (Theorem 8.4). We first prove a lemma; we use the lemma in the proof of the characterization theorem and in the proof of the subsequent theorem about the continuity of the inverse function. The proof of the lemma uses the Intermediate Value Theorem and the Maximum-Minimum Theorem.

Lemma 8.2: Let $f : [a, b] \to \mathsf{R}^1$ be a one-to-one continuous function.

(1) If $f(a) < f(b)$, then f is strictly increasing on $[a, b]$, $f([a, b]) = [f(a), f(b)]$, and f^{-1} is strictly increasing on $[f(a), f(b)]$.

(2) If $f(a) > f(b)$, then f is strictly decreasing on $[a, b]$, $f([a, b]) = [f(b), f(a)]$, and f^{-1} is strictly decreasing on $[f(b), f(a)]$.

Proof: We prove part (1); part (2) follows easily from part (1) (Exercise 8.3).

Assume that $f(a) < f(b)$. Then, by the last part of Theorem 5.13,

$$f([a, b]) = [c, d] \text{ for some } c < d.$$

We show that $f(a) = c$ and $f(b) = d$. Since $f([a, b]) = [c, d]$, there exist $s, t \in [a, b]$ such that $f(s) = c$ and $f(t) = d$. Let J denote the closed interval with end points s and t (i.e., $J = [s, t]$ if $s < t$ and $J = [t, s]$ if $t < s$). Since $f(s) = c$ and $f(t) = d$ and since $f(J) \subset [c, d]$, we see by the Intermediate Value Theorem (Theorem 5.2) that $f(J) = [c, d]$. Thus, since $f(a), f(b) \in [c, d]$, there exist $p, q \in J$ such that $f(p) = f(a)$ and $f(q) = f(b)$. Now, since f is one-to-one on $[a, b]$, $p = a$ and $q = b$. Hence, $a, b \in J$. Thus, $J = [a, b]$. Therefore, $s = a$ or b, and $t = a$ or b; furthermore, if $s = b$, then $t = a$, and we have

$$f(b) = f(s) = c < d = f(t) = f(a),$$

which contradicts our assumption that $f(a) < f(b)$. Hence, $s = a$ and, consequently, $t = b$. Therefore, $f(a) = c$ and $f(b) = d$.

We have proved the following:

$$(*)\ f([a,b]) = [f(a), f(b)].$$

We use (*) to prove that f is strictly increasing on $[a,b]$. Suppose that f is not strictly increasing on $[a,b]$. Then, since f is one-to-one, there are points x_1 and x_2 such that

$$a \leq x_1 < x_2 \leq b \quad \text{and} \quad f(x_1) > f(x_2).$$

Furthermore, $x_1 > a$ (otherwise, $x_1 = a$ and, hence, $f(a) > f(x_2)$, which contradicts (*)); also, since f is one-to-one and $a \neq x_2$, we see from (*) that $f(a) < f(x_2)$. To summarize, we have that

$$a < x_1 \quad \text{and} \quad f(a) < f(x_2) < f(x_1).$$

Thus, by the Intermediate Value Theorem (Theorem 5.2), there exists a point $c \in (a, x_1)$ such that $f(c) = f(x_2)$. However, this contradicts that f is one-to-one (since $c < x_1 < x_2$). Therefore, we have proved that f is strictly increasing on $[a,b]$.

Finally, we have shown that f is strictly increasing on $[a,b]$ and that $f([a,b]) = [f(a), f(b)]$ (by (*)); therefore, by Exercise 8.1, f^{-1} is strictly increasing on $[f(a), f(b)]$.

This completes the proof of part (1) of the lemma; part (2) is left as Exercise 8.3. ¥

Exercise 8.3: Finish the proof of Lemma 8.2 by showing how part (2) follows quickly from part (1).

We are now ready to prove the characterization theorem.

Theorem 8.4: Let I be an interval, and let $f : I \to \mathsf{R}^1$ be a continuous function. Then f is one-to-one if and only if f is either strictly increasing on I or strictly decreasing on I.

Proof: If f is either strictly increasing on I or strictly decreasing on I, then it is clear that f is one-to-one. Therefore, we need only prove the converse.

Any interval can be written as a countable union of closed and bounded intervals $[a_n, b_n]$, $n = 1, 2, ...$, where $[a_n, b_n] \subset [a_{n+1}, b_{n+1}]$ for all n. For example, $(a,b) = \cup_{n=1}^{\infty}[a + \frac{b-a}{2n}, b - \frac{b-a}{2n}]$, $[a,b) = \cup_{n=1}^{\infty}[a, b - \frac{b-a}{2n}]$, $(a,\infty) = \cup_{n=1}^{\infty}[a + \frac{1}{n}, a+n]$, and so on. Thus, whatever kind of interval the interval I in our theorem is (excluding the trivial case when $I = [a,a]$), we have

$$I = \cup_{n=1}^{\infty}[a_n, b_n], \quad [a_n, b_n] \subset [a_{n+1}, b_{n+1}] \text{ for all } n, \quad a_1 < b_1.$$

Now, assume that $f : I \to \mathsf{R}^1$ is one-to-one. Then either $f(a_1) < f(b_1)$ or $f(b_1) < f(a_1)$.

Assume first that $f(a_1) < f(b_1)$. Then, by part (1) of Lemma 8.2, f is strictly increasing on $[a_1, b_1]$. Assume inductively that f is strictly increasing on $[a_k, b_k]$ for some given k. Since f is one-to-one, either $f(a_{k+1}) < f(b_{k+1})$ or $f(b_{k+1}) < f(a_{k+1})$. If $f(b_{k+1}) < f(a_{k+1})$, then we see from part (2) of Lemma 8.2 that f is strictly decreasing on $[a_{k+1}, b_{k+1}]$, hence on $[a_k, b_k]$; this contradicts our inductive assumption that f is strictly increasing on $[a_k, b_k]$. Hence, $f(a_{k+1}) < f(b_{k+1})$. Therefore, by part (1) of Lemma 8.2, f is strictly increasing on $[a_{k+1}, b_{k+1}]$. Hence, by the Induction Principle (Theorem 1.20), we have proved that f is strictly increasing on $[a_n, b_n]$ for all n. Therefore, since $I = \cup_{n=1}^{\infty}[a_n, b_n]$, it follows easily that f is strictly increasing on I.

We leave the case when $f(a_1) > f(b_1)$ as an exercise (Exercise 8.5). ¥

Exercise 8.5: Finish the proof of Theorem 8.4 (by taking care of the case when $f(a_1) > f(b_1)$).

Finally, we prove our main theorem.

Theorem 8.6: Let I be an interval. If $f : I \to \mathsf{R}^1$ is a one-to-one continuous function, then f^{-1} is continuous on $f(I)$.

Proof: By Theorem 8.4, f is either strictly increasing on I or strictly decreasing on I. We assume that

(1) f is strictly increasing on I.

By (1) and Exercise 8.1, we have that

(2) f^{-1} is strictly increasing on $f(I)$.

Now, to prove that f^{-1} is continuous on $f(I)$, let $p \in f(I)$. We prove that $\lim_{y \to p} f^{-1}(y) = f^{-1}(p)$.

Let $\epsilon > 0$. Let $q = f^{-1}(p)$.

By the Intermediate Value Theorem, $f(I)$ is an interval. We take two cases:

Case 1: p is not an end point of $f(I)$. Then it follows from (2) that q is not an end point of I. Hence, we can assume that ϵ is small enough so that

$$[q - \epsilon, q + \epsilon] \subset I.$$

Thus, since f is strictly increasing on $[q - \epsilon, q + \epsilon]$ (by (1)), we have

(3) $f(q - \epsilon) < f(q) = p < f(q + \epsilon)$.

By (1), f is strictly increasing on $[q-\epsilon, q+\epsilon]$; hence, by Lemma 8.2, we have that

(4) $f([q - \epsilon, q + \epsilon]) = [f(q - \epsilon), f(q + \epsilon)]$.

Now, let

$$\delta = \min\{p - f(q - \epsilon), f(q + \epsilon) - p\}.$$

By (3), $\delta > 0$. Assume that $|y - p| < \delta$. Then $y \in (f(q-\epsilon), f(q+\epsilon))$ since

$$f(q-\epsilon) = p - [p - f(q-\epsilon)] \leq p - \delta < y < p + \delta$$
$$\leq p + [f(q+\epsilon) - p] = f(q+\epsilon).$$

Hence, by (4), $f^{-1}(y) \in (q-\epsilon, q+\epsilon)$; in other words, $\left|f^{-1}(y) - q\right| < \epsilon$. Therefore, since $q = f^{-1}(p)$, $\left|f^{-1}(y) - f^{-1}(p)\right| < \epsilon$. Thus, we have proved that

$$\lim_{y \to p} f^{-1}(y) = f^{-1}(p).$$

Therefore, f^{-1} is continuous at p by Theorem 3.12.

Case 2: p is an end point of $f(I)$. Then it follows from (2) that q is an end point of I, and it is easy to modify the argument for Case 1 to prove that f^{-1} is continuous at p (replace $[q-\epsilon, q+\epsilon]$ with either $[q, q+\epsilon]$ or $[q-\epsilon, q]$, and make the obvious adjustments in the rest of the proof for Case 1). ¥

3. The Inverse Function Theorem

We prove the main theorem of the chapter. The assumption in the theorem that $f'(p) \neq 0$ is unconditionally necessary (see Exercise 8.8).

Theorem 8.7 (Inverse Function Theorem): Let I be an interval, let $f : I \to \mathsf{R}^1$ be a one-to-one continuous function, and let $p \in I$. If f is differentiable at p and $f'(p) \neq 0$, then f^{-1} is differentiable at $f(p) = q$ and

$$(f^{-1})'(q) = \tfrac{1}{f'(p)} = \tfrac{1}{f'(f^{-1}(q))}.$$

Proof: We will use Lemma 7.22 with f in the lemma replaced by f^{-1} here and g in the lemma replaced by f here (thus, the roles of p and q in the lemma are switched here). The function F defined below is the function G in Lemma 7.22 with the replacements just mentioned:

$$F(x) = \begin{cases} \frac{f(x)-f(p)}{x-p} & \text{, if } x \neq p \\ f'(p) & \text{, if } x = p. \end{cases}$$

The assumptions of Lemma 7.22 are satisfied since f^{-1} is continuous at q (by Theorem 8.6) and f is differentiable at $f^{-1}(q) = p$ (by assumption in our theorem). Hence, by Lemma 7.22 (as adjusted here),

$$\lim_{y \to q} F(f^{-1}(y)) = f'(p).$$

Thus, since $f'(p) \neq 0$ (by assumption), $\lim_{y \to q} \frac{1}{F(f^{-1}(y))} = \frac{1}{f'(p)}$ by Lemma 4.19. Therefore, using the formula for F for the first equality below,

$$\tfrac{1}{f'(p)} = \lim_{y \to q} \frac{1}{\frac{f(f^{-1}(y))-f(p)}{f^{-1}(y)-p}} = \lim_{y \to q} \frac{f^{-1}(y)-p}{f(f^{-1}(y))-f(p)}$$

$$= \lim_{y \to q} \frac{f^{-1}(y)-f^{-1}(q)}{y-q} \overset{6.10}{=} (f^{-1})'(q). \text{ ¥}$$

Exercise 8.8: The assumption that $f'(p) \neq 0$ in Theorem 8.7 is absolutely necessary: *If I is an interval and $f : I \to R^1$ is a one-to-one differentiable function such that $f'(p) = 0$ for some point p, then f^{-1} is not differentiable at $f(p)$.* Prove this result, and explain why the result is to be expected from a picture of the graphs of f and f^{-1}.

Exercise 8.9: In the proof of Theorem 8.7, we proved that f^{-1} is differentiable at q by deriving the formula for $(f^{-1})'(q)$. If we had known beforehand that f^{-1} is differentiable at q, then we could have derived the formula for $(f^{-1})'(q)$ using the Chain Rule (Theorem 7.23). Show how to derive the formula for $(f^{-1})'(q)$ using the Chain Rule under the assumption that f^{-1} is differentiable at q (and the assumptions about f in Theorem 8.7).

Exercise 8.10: Find $(f^{-1})'(1)$ when $f(x) = x^7 + x^3 + x + 1$.

Exercise 8.11: Find $(f^{-1})'(6)$ when $f(x) = \sqrt{x^3 + 2x + 3}$.

Exercise 8.12: Let $f : \mathsf{R}^1 \to \mathsf{R}^1$ be a one-to-one differentiable function such that $f(3) = 4$ and $f'(3) = \frac{1}{4}$. Let $h = \frac{1}{f^{-1}}$. Find $h'(4)$.

4. Differentiating Rational Powers

We know that for all integers n, x^n is differentiable and $(x^n)' = nx^{n-1}$ (by Lemma 7.11 and Exercise 7.15). We extend the result to expressions of the form $x^{\frac{m}{n}}$, where m and n are integers ($n \neq 0$) and $x \neq 0$. The proof is an application of the Inverse Function Theorem (Theorem 8.7) and the Chain Rule (Theorem 7.23).

We begin by examining the function $f(x) = x^n$, where n is a natural number. We need to distinguish between the case when n is even and the case when n is odd; the reason will be apparent when we use the following lemma as a guide for defining the n^{th} root function.

Lemma 8.13: Let n be a natural number, and let $f(x) = x^n$ for all $x \in \mathsf{R}^1$.

(1) If n is even, then f is strictly increasing, hence one-to-one, on $[0, \infty)$ and $f([0, \infty)) = [0, \infty)$.

(2) If n is odd, then f is strictly increasing, hence one-to-one, on R^1 and $f(\mathsf{R}^1) = \mathsf{R}^1$.

Proof: All the numbers $0, 1, 2^n, 3^n, ..., k^n, ...$ are values of f; in addition, if n is odd, $n = 2m+1$, all the numbers $-k(-k)^{2m}$ for $k = 0, 1, 2, ...$ are values of f. Also, f is continuous by Theorem 4.16. Hence, it follows from the Intermediate Value Theorem (Theorem 5.2) and Lemma 1.21 that $f([0, \infty)) = [0, \infty)$ and, if n is odd, $f(\mathsf{R}^1) = \mathsf{R}^1$.

The fact that f is strictly increasing can be proved by induction; we leave this to the reader (Exercise 8.14).

Finally, f is one-to-one since a strictly increasing function is obviously one-to-one. ¥

Exercise 8.14: Finish the proof of Lemma 8.13 as indicated.

Lemma 8.13 provides a completely different proof of Theorem 1.25 and extends Theorem 1.25 to negative real numbers when n is odd. Which proof of Theorem 1.25 do you like better – the original proof or this proof?

Definition: Let n be a natural number, and let $f(x) = x^n$ for all $x \in \mathsf{R}^1$.

- With Lemma 8.13 in mind, we define the n^{th} *root function* to be the inverse of $f|[0, \infty)$ if n is even and to be the inverse of f if n is odd. Hence, the n^{th} root function has domain and range $[0, \infty)$ when n is even, and the n^{th} root function has domain and range R^1 when n is odd.
- The value of the n^{th} root function at x is denoted by $x^{\frac{1}{n}}$. Thus, we have *defined* $x^{\frac{1}{n}}$ for all $x \geq 0$ when n is even and for all real numbers x when n is odd; in other words, $x^{\frac{1}{n}}$ is defined to be the unique number such that

$$(x^{\frac{1}{n}})^n = x = (x^n)^{\frac{1}{n}}.$$

- For any integer m, $x^{\frac{m}{n}}$ is defined to be $(x^{\frac{1}{n}})^m$ for all x such that $x^{\frac{1}{n}}$ is defined; in addition, $x = 0$ is excluded if $m < 0$. Thus, except for $x = 0$ if $m < 0$, the function $h(x) = x^{\frac{m}{n}}$ is defined (only) on $[0, \infty)$ if n is even and on all of R^1 if n is odd.

The following theorem, which we use to prove our main theorem, is a consequence of the Inverse Function Theorem.

Theorem 8.15: Let g denote the n^{th} root function for some natural number n. Then g is differentiable at every point x in its domain except $x = 0$ and

$$g'(x) = (x^{\frac{1}{n}})' = \tfrac{1}{n} x^{\frac{1}{n} - 1}.$$

Proof: Let $f(x) = x^n$, where f is restricted to $[0, \infty)$ if n is even. Note that $g = f^{-1}$. We will apply the Inverse Function Theorem (Theorem 8.7) to g. To know that we can do so, note the following: f is one-to-one (by Lemma 8.13), f is continuous (by Theorem 4.16), and $f'(x) = nx^{n-1}$ (by Lemma 7.11), hence $f'(x) \neq 0$ if $x \neq 0$. Therefore, by the Inverse Function Theorem, if $x \neq 0$,

$$g'(x) = \frac{1}{f'(g(x))} = \frac{1}{f'(x^{\frac{1}{n}})} = \frac{1}{n(x^{\frac{1}{n}})^{n-1}} = \frac{1}{nx^{1-\frac{1}{n}}} = \tfrac{1}{n} x^{\frac{1}{n} - 1}.$$

Finally, since $f'(0) = 0$, we know that g is not differentiable at $x = 0$ by Exercise 8.8. ¥

We now prove our main theorem using Theorem 8.15 and the Chain Rule.

Theorem 8.16: Let n be a natural number, and let $m \neq 0$ be an integer. The function $h(x) = x^{\frac{m}{n}}$ is differentiable at every point x in its domain except $x = 0$ and

$$h'(x) = (x^{\frac{m}{n}})' = \tfrac{m}{n} x^{\frac{m}{n} - 1}.$$

Proof: By the definition of $x^{\frac{m}{n}}$ (above Theorem 8.15), $h(x) = (x^{\frac{1}{n}})^m$. Hence, $h = f \circ g$, where $g(x) = x^{\frac{1}{n}}$ and $f(x) = x^m$. Thus, by the Chain Rule (Theorem 7.23), $h'(x) = f'(g(x))g'(x)$. Therefore, since $g'(x) = \frac{1}{n}x^{\frac{1}{n}-1}$ (by Theorem 8.15) and $f'(x) = mx^{m-1}$ (by Lemma 7.11 and Exercise 7.15), we have

$$h'(x) = f'(x^{\frac{1}{n}})g'(x) = m(x^{\frac{1}{n}})^{m-1}(\tfrac{1}{n}x^{\frac{1}{n}-1}) = \tfrac{m}{n}x^{\frac{m}{n}-1}. \text{ ¥}$$

It is natural to wonder if Theorem 8.16 holds for *all* powers of x rather than just for rational powers (when considering an irrational power of x, we assume that $x > 0$). However – we must first wonder what x^p means for a given irrational number p: What do we mean by $2^{\sqrt{2}}$, 3^{π}, etc.? Once we give an appropriate definition of x^p for any given irrational number p (and $x > 0$), we will see that $(x^p)' = px^{p-1}$ for any given real number p and all $x > 0$. The definition of x^p for irrational powers p awaits further developments, namely, the natural logarithm, which we define in Chapter XVI using the integral. The definition for x^p is above Exercise 16.20, and the result about the derivative of x^p is Theorem 16.31.

5. Differentiating Trigonometric Functions and Their Inverses

We first show that the trigonometric functions are differentiable. The fact that the inverse trigonometric functions are differentiable is then a consequence of the Inverse Function Theorem. As we have done in the past, we obtain formulas for all derivatives.

We assume that the reader is familiar with the definitions of the trigonometric functions and basic trigonometric identities. The independent variable, x, for a trigonometric function is a real number that is to be understood as the angle whose radian measure is x. Thus, when we write $\sin(x)$, $\cos(x)$ and so on, we assume x is radian measure; when we use degree measure, we will specifically write x° to mean x measured in degrees. We note the relationship between radian measure and degree measure: $1^\circ = \frac{\pi}{180}$ radians.

We denote a trigonometric function raised to a power with a superscript directly after the function; for example $\sin^2(x)$ denotes $(\sin(x))^2$. As is consistent with our notation for inverse functions in general (section 1), we denote inverse trigonometric functions with a superscript of -1 directly after the function; for example, $\sin^{-1}(x)$ denotes the inverse sine of x, not $\frac{1}{\sin(x)}$ (which we denote by $(\sin(x))^{-1}$). The reader should *not*, for example, confuse $\sin^{-p}(x)$ with $(\sin^{-1}(x))^p$ when $p \neq 1$; $\sin^{-p}(x)$ for $p \neq 1$ always means $(\sin(x))^{-p} = \frac{1}{\sin^p(x)}$.

We use notation for the derivative of a trigonometric function and the derivative of its inverse that is consistent with our notation for derivatives in general: $\sin'$ or $\sin'(x)$ denotes the derivative of the sine function, $(\sin^{-1})'$ or $(\sin^{-1})'(x)$ denotes the derivative of the inverse sine function, and so forth.

We let S^1 denote the unit circle in the plane R^2 (i.e., S^1 is all points (x, y) in R^2 such that $\sqrt{x^2 + y^2} = 1$).

We note that the sine and cosine functions are continuous for use later:

Lemma 8.17: The sine function and the cosine function are continuous.

Proof: We do not belabor the proof that the sine and cosine functions are continuous. Their continuity is geometrically clear: Simply recognize that $(\cos(x), \sin(x))$ is the point on the unit circle S^1 corresponding to the angle whose radian measure is x, and observe that small changes in x result in small changes in the points $(\cos(x), \sin(x))$. ¥

To show that all trigonometric functions are differentiable, we focus on the sine function. Once we prove the sine function is differentiable, the differentiability of *all* trigonometric functions follows using elementary facts from trigonometry.

Let us see what is involved in showing that the sine function is differentiable: For a given x and for any $h \neq 0$,

$$\frac{\sin(x+h)-\sin(x)}{h} = \frac{\sin(x)\cos(h)+\sin(h)\cos(x)-\sin(x)}{h};$$

since $\lim_{h\to 0} \frac{\sin(x)}{h}$ and $\lim_{h\to 0} \frac{\cos(x)}{h}$ do not exist, we have no hope of proving that $\sin(x)$ is differentiable unless we put the expressions involving h together, obtaining

$$\frac{\sin(x+h)-\sin(x)}{h} = \sin(x)\frac{\cos(h)-1}{h} + \cos(x)\frac{\sin(h)}{h}.$$

Thus, we need to find two limits, $\lim_{h\to 0} \frac{1-\cos(h)}{h}$ and $\lim_{h\to 0} \frac{\sin(h)}{h}$, if the limits do, indeed, exist. The problem is not trivial, but can be solved with the aid of some elementary geometry:

Lemma 8.18: $\lim_{x\to 0} \frac{\sin(x)}{x} = 1$ and $\lim_{x\to 0} \frac{1-\cos(x)}{x} = 0$.

Proof: We first prove that

(1) $\lim_{x\to 0^+} \frac{\sin(x)}{x} = 1$.

Proof of (1)*:* Assume that $0 < x < \frac{\pi}{2}$. Referring to Figure 8.18 below, we see that

$$area(\Delta OAB) < area(\text{sector}\, OAC) < area(\Delta ODC).$$

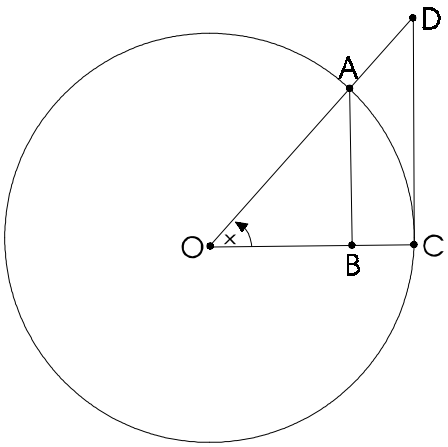


Figure 8.18

We write the inequalities above Figure 8.18 in terms of x (note: since a semi-circle has area $\frac{\pi}{2}$ and is a sector with angle π, sector OAC with angle x has proportional area $\frac{x}{\pi}\frac{\pi}{2}$, which is $\frac{x}{2}$):

$$\tfrac{1}{2}\cos(x)\sin(x) < \tfrac{x}{2} < \tfrac{1}{2}\tan(x) = \tfrac{\sin(x)}{2\cos(x)}.$$

Since $0 < x < \frac{\pi}{2}$, $\sin(x) > 0$; hence, the inequalities remain in the same direction when we multiply through by $\frac{2}{\sin(x)}$, obtaining

$$\cos(x) < \tfrac{x}{\sin(x)} < \tfrac{1}{\cos(x)}.$$

Hence, taking reciprocals (thereby reversing inequalities), we have

$$\tfrac{1}{\cos(x)} > \tfrac{\sin(x)}{x} > \cos(x).$$

Moreover, by Lemma 8.17, Theorem 3.12, and Lemma 4.19,

$$\lim_{x\to 0^+}\cos(x) = 1 = \lim_{x\to 0^+}\tfrac{1}{\cos(x)}$$

Therefore, by the Squeeze Theorem (Theorem 4.34), $\lim_{x\to 0^+}\frac{\sin(x)}{x} = 1$. This proves (1).

Next, we prove that

(2) $\lim_{x\to 0^-}\frac{\sin(x)}{x} = 1$.

Proof of (2)*:* Define $g : [0, \frac{\pi}{2}) \to \mathbf{R}^1$ by

$$g(x) = \begin{cases} \frac{\sin(x)}{x} & \text{, if } x > 0 \\ 1 & \text{, if } x = 0. \end{cases}$$

By (1) and Theorem 3.15, g is continuous. Define $f : (-\frac{\pi}{2}, 0] \to [0, \frac{\pi}{2})$ by $f(x) = -x$; obviously, f is continuous. Hence, by Theorem 4.28, $g \circ f$ is continuous. Thus, by Theorem 3.12,

$$\lim_{x\to 0^-}(g\circ f)(x) = (g\circ f)(0) = g(0) = 1;$$

furthermore, since $\sin(-x) = -\sin(x)$, we have that

$$(g\circ f)(x) = g(-x) = \tfrac{\sin(-x)}{-x} = \tfrac{-\sin(x)}{-x} = \tfrac{\sin(x)}{x}, \quad -\tfrac{\pi}{2} < x < 0.$$

Therefore,

$$\lim_{x\to 0^-}\tfrac{\sin(x)}{x} = \lim_{x\to 0^-}(g\circ f)(x) = 1.$$

This proves (2).

By (1), (2), and Theorem 3.16, $\lim_{x\to 0}\frac{\sin(x)}{x} = 1$. This proves that the first part of the lemma.

To prove the second part of the lemma, first observe that when $-\frac{\pi}{2} < x < \frac{\pi}{2}$ (to assure that $\cos(x) \neq -1$),

$$\frac{1-\cos(x)}{x} = \frac{1-\cos(x)}{x}\frac{1+\cos(x)}{1+\cos(x)} = \frac{\sin^2(x)}{x[1+\cos(x)]} = \frac{\sin(x)}{x}\frac{\sin(x)}{1+\cos(x)}.$$

Next, note that $\lim_{x\to 0}\frac{\sin(x)}{x} = 1$ (by the first part of the lemma) and that $\lim_{x\to 0}\frac{\sin(x)}{1+\cos(x)} = 0$ (by Lemma 8.17, Corollary 4.21 and Theorem 3.12). Therefore, by Theorem 4.9 on limits of products,

$$\lim_{x\to 0}\frac{1-\cos(x)}{x} = (\lim_{x\to 0}\frac{\sin(x)}{x})(\lim_{x\to 0}\frac{\sin(x)}{1+\cos(x)}) = (1)(0) = 0. \text{ ¥}$$

Exercise 8.19: Fix nonzero real numbers a and b. Find $\lim_{x\to 0}\frac{\sin(ax)}{bx}$ by making use of Theorem 3.15. Show all work carefully.

It is now easy to prove our result for the sine function:

Theorem 8.20: $\sin'(x) = \cos(x)$.

Proof: Fix $x \in \mathsf{R}^1$. Continuing from where we left off above Lemma 8.18,

$$\sin'(x) = \lim_{h\to 0}[\sin(x)\frac{\cos(h)-1}{h} + \cos(x)\frac{\sin(h)}{h}].$$

Therefore, by Lemma 8.18 and Theorem 4.1 on limits of sums,

$$\sin'(x) = \lim_{h\to 0}\sin(x)\frac{\cos(h)-1}{h} + \lim_{h\to 0}\cos(x)\frac{\sin(h)}{h} = \cos(x). \text{ ¥}$$

Corollary 8.21: $\cos'(x) = -\sin(x)$.

Proof: Since $\cos(x) = \sin(\frac{\pi}{2} - x)$ for all x, we see from Theorem 8.20 and the Chain Rule (Theorem 7.23) that

$$\cos'(x) = [\cos(\tfrac{\pi}{2} - x)][-1] = -\cos(\tfrac{\pi}{2} - x).$$

Therefore, since $\cos(\frac{\pi}{2} - x) = \sin(x)$, we have that $\cos'(x) = -\sin(x)$. ¥

Exercise 8.22: Using that all trigonometric functions can be expressed in terms of the sine and/or cosine functions, prove that the following formulas hold (for x in the domain of each function): $\tan'(x) = \sec^2(x)$, $\cot'(x) = -\csc^2(x)$, $\sec'(x) = \sec(x)\tan(x)$, and $\csc'(x) = -\csc(x)\cot(x)$.

Exercise 8.23: Would you expect the rate of change of a trigonometric function with respect to radian measure to be greater, smaller, or the same as the rate of change of the trigonometric function with respect to degree measure? Explain your answer intuitively, and prove your answer is correct.

Exercise 8.24: Direct computations using the Chain Rule (Theorem 7.23) and Theorem 8.20 give that

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

Thus, $(\sin^2(x))' = \sin(2x)$. Is this a coincidence, or can you explain why the result is to be expected from, say, a geometric point of view?

We turn our attention to derivatives of the inverse trigonometric functions.

The inverse sine function, $\sin^{-1}$, has domain $[-1,1]$ and range $[-\frac{\pi}{2},\frac{\pi}{2}]$. Differentiating the inverse sine function is simply a matter of applying the Inverse Function Theorem (Theorem 8.7) in conjunction with Theorem 8.20:

Theorem 8.25: $(\sin^{-1})'(x) = \frac{1}{\sqrt{1-x^2}}$.

Proof: Fix $x \in [-1,1]$. Since $\sin' = \cos$ (Theorem 8.20), we see from the Inverse Function Theorem (Theorem 8.7) that

$$(\sin^{-1})'(x) = \frac{1}{\sin'(\sin^{-1}(x))} = \frac{1}{\cos(\sin^{-1}(x))} = \frac{1}{\sqrt{1-x^2}}. \quad ¥$$

Exercise 8.26: The inverse cosine function has domain $[-1,1]$ and range $[0,\pi]$. Prove that $(\cos^{-1})'(x) = \frac{-1}{\sqrt{1-x^2}}$.

Exercise 8.27: The inverse tangent has domain R^1 and range $(-\frac{\pi}{2},\frac{\pi}{2})$. Prove that $(\tan^{-1})'(x) = \frac{1}{1+x^2}$.

Exercise 8.28: The inverse cotangent has domain R^1 and range $(0,\pi)$. Prove that $(\cot^{-1})'(x) = \frac{-1}{1+x^2}$.

Exercise 8.29: The inverse secant has domain $(-\infty,-1)\cup(1,\infty)$ and range $[0,\frac{\pi}{2})\cup(\frac{\pi}{2},\pi]$. Prove that $(\sec^{-1})'(x) = \frac{1}{|x|\sqrt{x^2-1}}$.

Exercise 8.30: The inverse cosecant has domain $(-\infty,-1)\cup(1,\infty)$ and range $[-\frac{\pi}{2},0)\cup(0,\frac{\pi}{2}]$. Prove that $(\csc^{-1})'(x) = \frac{-1}{|x|\sqrt{x^2-1}}$.

Chapter IX: Maxima, Minima and Derivatives

As a student in plane geometry, you may have seen the following problem: *If p and q are points on the same side of a line ℓ, find a point r on ℓ such that the sum of the distances pr and rq is a minimum.* The problem is solved easily by reflecting q across the line ℓ to the point q', and then observing that r must be the point on ℓ where the line from p to q' meets ℓ. What you may not have observed is that the minimum path from p to ℓ to q is the path for which the angles formed by pr and ℓ and by qr and ℓ are equal. Is this symmetry only a coincidence?

Is it merely a coincidence that the largest area enclosed by all curves in the plane of a given length is the area enclosed by the most symmetric of those curves (the circle)? And is it a coincidence that of all the rectangles having a given perimeter, the one with the largest area is the one that is most symmetric (the square)?

Surely, beauty in nature is intimately connected with symmetry, and it would appear that symmetry is connected with maxima and minima. Perhaps this is why maximum and minimum problems have been a constant theme throughout history. Leonhard Euler (1707 - 1783) articulated the importance of maxima and minima by saying that all interesting phenomena in this world can be explained in terms of maxima and minima.

We began our study of maxima and minima in Chapter V in the setting of continuous functions; there we proved that every continuous function on a closed and bounded interval has a maximum value and a minimum value (Theorem 5.13). We now localize the notions of maxima and minima and relate the local notions to derivatives. Our main result is Theorem 9.7, which lays the foundation for further study of maxima and minima. Theorem 9.7 sets the stage for the proof of the Mean Value Theorem (which we prove in the next chapter); the Mean Value Theorem is the essential ingredient for proving theorems that are used to classify local maxima and minima.

1. Neighborhoods

The following descriptive terminology will help us formulate statements concisely.

Definition: Let $X \subset \mathsf{R}^1$, and let $p \in X$. A *neighborhood of p in X* is the intersection of X with any open interval in R^1 containing p; in other words, if (a, b) is an open interval in R^1 such that $p \in (a, b)$, then $X \cap (a, b)$ is a neighborhood of p in X.

If $\epsilon > 0$, then $X \cap (p - \epsilon, p + \epsilon)$ is called the *ϵ - neighborhood of p in X*; thus, the ϵ - neighborhood of p in X is $\{x \in X : |p - x| < \epsilon\}$.

Example 9.1: $(-1, 1)$ is a neighborhood of 0 in $[-1, 1]$ (the 1 - neighborhood of 0 in $[-1, 1]$); $[0, \frac{1}{2})$ is a neighborhood of 0 in $[0, 1]$ (the $\frac{1}{2}$ - neighborhood of 0 in $[0, 1)$), but $[0, \frac{1}{2})$ is not a neighborhood of 0 in $[-1, 1]$; $(-\frac{1}{4}, \frac{1}{2})$ is a

neighborhood of 0 in $[-\frac{1}{4}, 1)$, but $(-\frac{1}{4}, \frac{1}{2})$ is not an ϵ-neighborhood of 0 in $[-\frac{1}{4}, 1)$; for $X = \{0, 1, \frac{1}{2}, \frac{1}{3}, ..., \frac{1}{n}, ...\}$, $\{0, \frac{1}{9}, \frac{1}{10}, \frac{1}{11}, ..., \frac{1}{n}, ...\}$ is a neighborhood of 0 in X (the $\frac{1}{9}$-neighborhood of 0 in X), but $\{0, \frac{1}{9}, \frac{1}{11}, \frac{1}{13}, ..., \frac{1}{9+2n}, ...\}$ is not a neighborhood of 0 in X.

Exercise 9.2: Let $X \subset \mathsf{R}^1$, and let $p \in X$. The intersection of finitely many neighborhoods of p in X is a neighborhood of p in X.

Exercise 9.3: If I is an open interval and $p \in I$, then any neighborhood U of p in I contains an open interval J such that $p \in J$; hence, J is an open neighborhood of p in I and both J and U are open neighborhoods of p in R^1.

Exercise 9.4: Let $X \subset \mathsf{R}^1$, and let $p \in X$. When is $\{p\}$ a neighborhood of p in X?

2. Local and Global Maxima and Minima

We localize the notions of maximum value and minimum value of a function (defined in section 2 of Chapter V). In order to avoid ambiguity, from now on we call the maximum value and the minimum value of a function the global maximum value and the global minimum value of the function.

Definition: Let $X \subset \mathsf{R}^1$, let $f : X \to \mathsf{R}^1$ be a function, and let $p \in X$.

- f has a *local maximum at* p provided that there is a neighborhood U of p in X such that $f(p) \geq f(x)$ for all $x \in U$.
- f has a *local minimum at* p provided that there is a neighborhood U of p in X such that $f(p) \leq f(x)$ for all $x \in U$.
- f has a *global* (or *absolute*) *maximum at* p provided that $f(p) \geq f(x)$ for all $x \in X$, in which case we call $f(p)$ the *global maximum value of* f.
- f has a *global* (or *absolute*) *minimum at* p provided that $f(p) \leq f(x)$ for all $x \in X$, in which case we call $f(p)$ the *global minimum value of* f.
- Local maxima and local minima are called *local extrema*; global maxima and global minima are called *global* (or *absolute*) *extrema*.

We give an example to illustrate the concepts we just introduced.

Example 9.5: Define f on $[0, 3]$ as follows:

$$f(x) = \begin{cases} 3x & \text{, if } 0 \leq x \leq 1 \\ -x + 4 & \text{, if } 1 \leq x \leq 2 \\ 2x - 2 & \text{, if } 2 \leq x \leq 3. \end{cases}$$

Then f has local minima at $x = 0$ and 2, local maxima at $x = 1$ and 3, and global extrema at $x = 0$ and 3.

Next, we give an example for which we have more questions than answers. Our purpose is to motivate the value of the theorem we are about to prove; we return to the example after we prove the theorem.

Example 9.6: Define $f : [0, 4] \to \mathsf{R}^1$ by $f(x) = x(x-2)(x-4)$. Note that $f(x) = 0$ when $x = 0, 2$ and 4; also, from the signs of the terms, we see that $f(x) > 0$ when $0 < x < 2$ and that $f(x) < 0$ when $2 < x < 4$. It now follows from the Maximum - Minimum Theorem (Theorem 5.13) that f has a global maximum value at some point of $[0, 2]$ and a global minimum value at some point of $[2, 4]$. Also, f has a local minimum at $x = 0$ and a local maximum at $x = 4$; obviously, f does not have global extrema at $x = 0, 4$. The questions are: At what points are the global extrema attained? What are the values of the global extrema? Are there any local extrema occurring at points in the open interval $(0, 4)$ that are not global extrema and, if so, at what points do they occur? We answer the questions in Example 9.10.

The theorem below gives an important relation between local extrema and derivatives. The relation is only true in the direction stated; for example, $f(x) = x^3$ (all $x \in \mathsf{R}^1$) has derivative zero at $p = 0$ and yet has no local extrema.

Theorem 9.7: Let I be an open interval, and let $f : I \to \mathsf{R}^1$ be a function that is differentiable at a point $p \in I$. If f has a local extremum at p, then $f'(p) = 0$.

Proof: Assume that f has a local maximum at p. Then there is a neighborhood U of p in I such that $f(p) \geq f(x)$ for all $x \in U$. By Exercise 9.3, there is an open interval $(s, t) \subset U$ such that $p \in (s, t)$. Note that $(s, t) \subset I$ (since $U \subset I$); hence, we have that

(1) $f(p) \geq f(x)$ for all $x \in (s, t)$.

The proof now proceeds by analyzing the sign of $\frac{f(x)-f(p)}{x-p}$ when $s < x < p$ and when $p < x < t$: By (1), $f(x) - f(p) \leq 0$ for all $x \in (s, t)$; hence,

(2) $\frac{f(x)-f(p)}{x-p} \geq 0$ if $s < x < p$ and $\frac{f(x)-f(p)}{x-p} \leq 0$ if $p < x < t$.

Now, since f is differentiable at p, we know from Theorem 6.15 that

$$f'_-(p) = f'(p) = f'_+(p).$$

Furthermore, by (2), $f'_-(p) \geq 0$ and $f'_+(p) \leq 0$. Therefore, $f'(p) = 0$.

This proves the theorem when f has a local maximum at p. We leave the case when f has a local minimum at p as an exercise (below). ¥

Exercise 9.8: Prove Theorem 9.7 for the case when f has a local minimum at p.

Exercise 9.9: Give an example to show that the analogue of Theorem 9.7 for closed intervals is false.

Theorem 9.7 gives us a way to determine where a differentiable function on an interval may have local or global extrema. Sometimes, we can even determine the types of extrema the function has. We illustrate with two examples. The first example is a continuation of Example 9.6.

Example 9.10: Define $f : [0,4] \to \mathsf{R}^1$ by $f(x) = x(x-2)(x-4)$, the function in Example 9.6. We apply Theorem 9.7 to answer the questions we asked in Example 9.6.

To find the derivative of f, it is convenient to write f in unfactored form (to avoid using the product theorem for derivatives twice): $f(x) = x^3 - 6x^2 + 8x$ and thus, by Theorem 7.12,

$$f'(x) = 3x^2 - 12x + 8.$$

Hence, $f'(x) = 0$ when $x = 2 \pm \frac{2}{3}\sqrt{3}$. Moreover, we knew in Example 9.6 that f has its global maximum value at some point of $(0,2)$ and its global minimum value at some point of $(2,4)$. Therefore, by Theorem 9.7, we can now conclude that f must have its global maximum at $x = 2 - \frac{2}{3}\sqrt{3}$, its global minimum at $x = 2 + \frac{2}{3}\sqrt{3}$, and the global extrema do not occur at any other point; the global maximum value is $f(2 - \frac{2}{3}\sqrt{3}) = \frac{16}{9}\sqrt{3}$ and the global minimum value is $f(2 - \frac{2}{3}\sqrt{3}) = -\frac{16}{9}\sqrt{3}$. Finally, by Theorem 9.7, f has no local extrema at points in the open interval $(0,4)$ that are not global extrema. Thus, taking into account the end points, the local extrema occur at $x = 0, 2 \pm \frac{2}{3}\sqrt{3}, 4$ and the global extrema occur at $x = 2 \pm \frac{2}{3}\sqrt{3}$.

Example 9.11: Define $f : [-2,3] \to \mathsf{R}^1$ by $f(x) = x^3 - 3x^2 + 2$. Then, by Theorem 7.12,

$$f'(x) = 3x^2 - 6x.$$

Hence, $f'(x) = 0$ when $x = 0$ or 2. Thus, by Theorem 9.7, the only possible points at which f could have local extrema are $0, 2$ and the end points -2 and 3 (end point extrema are not taken care of by Theorem 9.7). Now, we see whether extrema occur at these points and, if so, what types of extrema they are. We list the values of f at the four points $-2, 0, 2$ and 3:

$$f(-2) = -18,\ f(0) = 2,\ f(2) = -2,\ f(3) = 2.$$

Therefore, $f(-2) = -18$ is the global minimum of f and $f(0) = f(3) = 2$ is the global maximum of f. What about $f(2) = -2$? This appears be a local minimum for f since the function f seems to go down to -2 on $[0,2]$ and then up to 2 on $[2,3]$; but, can we be sure that f has a local minimum at 2? Yes – we can be sure by using Theorem 9.7 together with the Maximum - Minimum Theorem (Theorem 5.13). We argue as follows: By the Maximum - Minimum Theorem, f has a minimum value m on $[0,3]$; since $f(2) = -2$, m does not occur at the end points of $[0,3]$; thus, by Theorem 9.7 applied to the open interval $(0,3)$, m occurs when $x \in (0,3)$ and $f'(x) = 0$; therefore, $x = 2$ is the only possibility and, hence, $f(2) = m$. This proves that $f(2) = -2$ is a local minimum for f.

The argument in Example 9.11 to show f has a local minimum at $x = 2$ is somewhat tedious. Later, we will have a simple test at our disposal which will enable us to avoid such arguments (Theorem 10.19).

We clarify one point so as not to be misled by the examples above: A differentiable function on a closed interval *need not* have local extrema at end points of the interval even if the derivative of the function is zero at an end point. You are asked to find an example:

Exercise 9.12: Give an example of a differentiable function f on $[0, 1]$ such that $f'(0) = 0$ and, yet, 0 is not a local extremum of f. A picture of the function (rather than a formula) is sufficient, even preferred!

Exercise 9.13: Let $f(x) = x^3 + x^2 - 6x$. Find all points where f has local maxima and local minima; determine what kind of extremum occurs at each such point. Are there any global extrema?

3. Critical Points

In this section we bring into sharper focus the main ideas in the theorem and examples in the preceding section. We conclude with general comments.

We have seen that three types of points play the crucial role in finding and classifying extrema of a function on an interval: Points at which the derivative of the function is zero, end points of the interval (if there are any), and points at which the function is not differentiable (Example 9.5). We give a name to the types of these points that involve derivatives:

Definition: Let I be an interval, and let $f : I \to \mathsf{R}^1$ be a function. A point $p \in I$ that is not an end point of I is called a *critical point of* f provided that $f'(p) = 0$ or f is not differentiable at p.

We can now summarize what we have shown in the examples and the theorem in section 2 in a concise way:

Corollary 9.14: Let I be an interval, and let $f : I \to \mathsf{R}^1$ be a function. Then the local and global extrema (if they exist) must be attained at critical points of f or at an end point of I.

Proof: Assume that f has a local extremum at a point $p \in I$. Assume further that f is differentiable at p and that p is not an end point of I (remember: functions can be differentiable at end points according to our definition of derivative). Then, by Theorem 9.7 (applied to I without its end points), $f'(p) = 0$; therefore, p is a critical point of f. ¥

We comment in general about the ideas and, especially, the direction initiated in this chapter.

We have shifted our emphasis from finding global extrema to finding local extrema. At the same time, we have stressed the importance of finding global extrema. Why don't we just narrow down on finding global extrema and leave the problem of finding local extrema for later or omit it completely? The answer is simple: Finding local extrema *is* narrowing down on finding global extrema, as we have illustrated in examples, and it is easier to find local extrema first than it is to find global extrema directly (by virtue of Theorem 9.7).

Using local extrema to find global extrema is a special case of a general mathematical procedure – approximation. You have seen approximation at work when rounding off decimals, finding areas (if you had some contact with integral calculus or the work of the ancient Greeks), and in the section on linear approximation (section 3 of Chapter VI); in fact, the very definitions of limits and derivatives are based on approximation. We are now approximating global extrema by finding local extrema; as we have seen, this leads to finding the global extrema. "Necessity is the mother of invention," and, in this case, local extrema were born out of the desire to find global extrema.

We note that local extrema are important in connection with many aspects of mathematics and science. To mention only a few, local extrema are used in the physical sciences, in optimization, in dynamical systems, in economics, and in analyzing statistical data. That being said, we must add that local extrema are themselves interesting and that is enough reason to study them.

Exercise 9.15: Define $f : [-1, 1] \to \mathsf{R}^1$ by $f(x) = x^{\frac{4}{5}} + 3$. Find all points where f has local maxima and local minima; determine what kind of extremum occurs at each such point.

Exercise 9.16: Let f be defined on R^1 by $f(x) = 5x^{\frac{2}{3}} + x^{\frac{5}{3}} + 1$. Find all points where f has local maxima and local minima; determine what kind of extremum occurs at each such point.

Exercise 9.17: Prove the assertion in the introduction to the chapter that of all the rectangles having a given perimeter, the one with the largest area is the one that is most symmetric (the square).

Exercise 9.18: Find the point on the circle $x^2 + y^2 = 1$ that is closest to $(2, 0)$. (You know the answer, but use the methods in this chapter.)

Exercise 9.19: Assume that $f : \mathsf{R}^1 \to \mathsf{R}^1$ is differentiable and that $f'(x) \neq 0$ for all x. Then f is one-to-one.

Exercise 9.20: Give examples of polynomials of degree 3 that have no critical point, only one critical point, and two critical points.

Exercise 9.21: A polynomial of degree $n > 0$ has at most n roots. (A *root of a function* is a point at which the function has value 0.)

Exercise 9.22: Give an example of a nonconstant function $f : \mathsf{R}^1 \to \mathsf{R}^1$ such that every real number is a critical point of f and such that $f'_+(x)$ exists for every $x \in \mathsf{R}^1$.

Chapter X: The Mean Value Theorem and Consequences

We prove the Mean Value Theorem in section 1. Then, section by section, we derive different types of important results from the theorem. We emphasize curve sketching in conjunction with the results in sections 3 and 4.

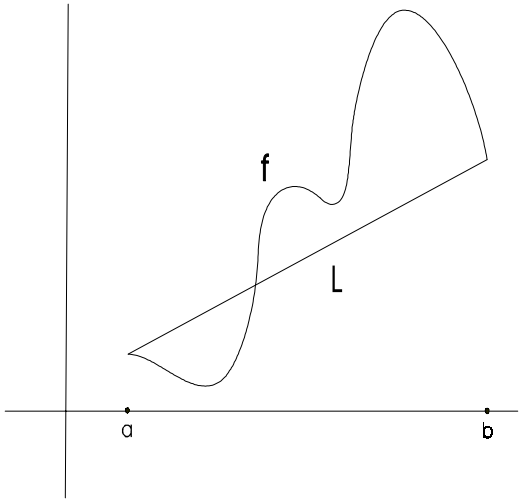
1. The Mean Value Theorem

If you travel 100 miles in 2 hours, it is obvious that at some point during the trip your velocity must be 50 miles per hour (your average velocity). In general terms, let f be a differentiable function that gives the distance $f(t)$ an object has traveled as a function of time t; then it is intuitively evident that the average velocity of the object over a time interval $[a, b]$ must be its instantaneous velocity at some time t_0 between a and b:

$$f'(t_0) = \frac{f(b)-f(a)}{b-a}.$$

This is the substance of the Mean Value Theorem, but certainly not the proof! Let us indicate the geometrical idea behind the proof.

In the figure below, the slope of the line segment L joining $(a, f(a))$ and $(b, f(b))$ is the average velocity of the object. Imagine that we continuously move L up (or down) parallel to itself. We eventually arrive at the last time the moving line segments touch the graph of f; at that moment, the line segment is tangent to the graph of f at a point $(t_0, f(t_0))$, which says $f'(t_0) = \frac{f(b)-f(a)}{b-a}$.



The discussion we just presented is not a proof; for example, how do we know there is a last time the moving line segments touch the graph of f ? Nevertheless,

the discussion is an intuitively plausible argument that provides insight into why the Mean Value Theorem is true.

We proceed to the precise statement and proof of the Mean Value Theorem. We first prove a special case of the theorem from which the theorem follows. The special case is due to Michel Rolle (1652 - 1719), who eventually became a vocal opponent of calculus, calling it a "collection of ingenious fallacies."

Lemma 10.1 (Rolle's Theorem): Assume that f is continuous on $[a, b]$, differentiable on (a, b), and that $f(a) = f(b) = 0$. Then there is a point $p \in (a, b)$ such that $f'(p) = 0$.

Proof: If f is a constant function, then $f'(x) = 0$ for all x and, thus, the lemma is true. Hence, we assume for the purpose of proof that f is not a constant function. Then there is a point $x_0 \in (a, b)$ such that $f(x_0) \neq 0$. Hence, either $f(x_0) < 0$ or $f(x_0) > 0$.

Assume first that $f(x_0) < 0$. By the Maximum - Minimum Theorem (Theorem 5.13), f attains its global minimum value at a point p. Since $f(x_0) < 0$, clearly $f(p) < 0$; hence, $p \in (a, b)$. In particular, then, f is differentiable at p. Therefore, by Theorem 9.7, $f'(p) = 0$.

The case when $f(x_0) > 0$ is handled similarly by taking p to be a point at which f attains its global maximum value (or, perhaps you have a simpler proof based on past experience?). ¥

Theorem 10.2 (Mean Value Theorem): Assume that f is continuous on $[a, b]$ and differentiable on (a, b). Then there is a point $p \in (a, b)$ such that

$$f'(p) = \tfrac{f(b)-f(a)}{b-a}.$$

Proof: In functional notation, the equation of the line going through the two points $(a, f(a))$ and $(b, f(b))$ is

$$g(x) = \tfrac{f(b)-f(a)}{b-a}(x - a) + f(a).$$

Define $h : [a, b] \to \mathsf{R}^1$ by letting $h = f - g$. (For geometric insight into what we do next, locate the local extrema of h in the figure on the preceding page.)

We see that h satisfies the assumptions of Lemma 10.1: h is continuous on $[a, b]$ by Corollary 4.4, h is differentiable on (a, b) by Theorem 7.3, and $h(a) = h(b) = 0$ by the formulas for g and h. Hence, by Lemma 10.1, there is a point $p \in (a, b)$ such that $h'(p) = 0$. Therefore,

$$0 = h'(p) \overset{7.3}{=} f'(p) - g'(p) \overset{6.2}{=} f'(p) - \tfrac{f(b)-f(a)}{b-a},$$

which gives that $f'(p) = \frac{f(b)-f(a)}{b-a}$. ¥

Exercise 10.3: Define $f : [-2, 2] \to \mathsf{R}^1$ by $f(x) = x^3 - 3x + 3$. Find all numbers p in $[-2, 2]$ that satisfy the conclusion of the Mean Value Theorem.

Exercise 10.4: If $f : \mathsf{R}^1 \to \mathsf{R}^1$ is differentiable and $f'(x) \neq 1$ for all $x \in \mathsf{R}^1$, then there is at most one point $p \in \mathsf{R}^1$ such that $f(p) = p$.

Exercise 10.5: Let I be an open interval, and let $p \in I$. Assume that f is continuous on I and differentiable on $I - \{p\}$ and that $\lim_{x \to p} f'(x)$ exists. Then f is differentiable at p.

Exercise 10.6: Assume that f and g are continuous on $[a, b]$ and differentiable on (a, b). Then there is a point $p \in (a, b)$ such that

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

2. Functions with Equal Derivatives

All constant functions (on an interval) have derivative zero. We prove that there are no other functions with derivative zero. Perhaps you think this is obvious. But then you may also think it is obvious that the only function whose derivative is itself is the function $f(x) = 0$; however, this is false! Furthermore, the prime example showing it is false is not just a curiosity – it is the exponential function $f(x) = e^x$, which has numerous applications in probabilty theory, economics and the physical sciences. See Corollary 16.24; in Exercise 16.25 we determine all functions f such that $f' = f$.

Once we prove that constant functions are the only functions whose derivative is zero, it follows easily that any two functions on an interval that have the same derivative must differ by a constant; stated more insightfully, the graphs of the functions are vertical translations of one another. This result is so important that it is often referred to as the fundamental theorem of differential calculus. When we study the integral, we will see that the fundamental theorem of differential calculus is crucial to evaluating integrals – it is the important ingredient in proving the second part of the Fundamental Theorem of Calculus (Theorem 14.2).

Theorem 10.7: If f is continuous on $[a, b]$ and $f'(x) = 0$ for all $x \in (a, b)$, then f is a constant function.

Proof: Let $x \in [a, b]$ such that $x \neq a$. Note that f is continuous on the interval $[a, x]$ (by Exercise 5.3). Hence, we can apply the Mean Value Theorem (Theorem 10.2) to f on the interval $[a, x]$, thereby obtaining a point $p \in (a, x)$ such that

$$f'(p) = \tfrac{f(x) - f(a)}{x - a}.$$

Thus, since $f'(p) = 0$ (by assumption in the theorem), we see that $f(x) = f(a)$. This proves that $f(x) = f(a)$ for all $x \in [a, b]$. ¥

Theorem 10.8: If f and g are continuous on $[a, b]$ and $f'(x) = g'(x)$ for all $x \in (a, b)$, then f and g differ by a constant; in other words, there is a constant C such that $f(x) - g(x) = C$ for all $x \in [a, b]$.

Proof: Define $h : [a, b] \to \mathsf{R}^1$ by letting $h = f - g$. Then h is continuous on $[a, b]$ (by Corollary 4.4) and

$$h'(x) \overset{7.3}{=} f'(x) - g'(x) = 0, \quad \text{all } x \in (a, b).$$

Therefore, by Theorem 10.7, h is a constant function. ¥

We note that Theorem 10.7 and Theorem 10.8 are really the same theorem: Theorem 10.7 follows immediately from Theorem 10.8 by taking g in Theorem 10.8 to be the constant function $g(x) = 0$.

We close by noting that Theorem 10.8 holds when the functions are defined on any interval:

Theorem 10.9: Let I be any interval, and let E denote the set of end points of I (E may be empty). If $f, g : I \to \mathsf{R}^1$ are continuous on I and if $f'(x) = g'(x)$ for all $x \in I - E$, then f and g differ by a constant.

Proof: Recall from the proof of Theorem 8.4 that any interval is the countable union of an "increasing sequence" of closed and bounded intervals. Using this fact and Theorem 10.8, our theorem follows (we leave the details for the first exercise below). ¥

Exercise 10.10: Do the details for the proof of Theorem 10.9.

Exercise 10.11: Let $f(x) = x^5 - 3x^2 + 2$. Find all functions whose derivatives are f.

Exercise 10.12: Let $f(x) = (2x+4)^8$. Find all functions whose derivatives are f.

Exercise 10.13: Let $f(x) = x\sqrt{x^2+7}$. Find all functions whose derivatives are f.

Exercise 10.14: Let $f(x) = \frac{1}{x^2}$. Find all functions whose derivatives are f. (Be careful – there may be more than you think!)

Exercise 10.15: Let $f(x) = |x - 1|$. Find all functions whose derivatives are f.

Exercise 10.16: Let f be the function given by

$$f(x) = \begin{cases} x+2 & \text{, if } x < 0 \\ x & \text{, if } x \geq 0. \end{cases}$$

Is there a function $g : \mathsf{R}^1 \to \mathsf{R}^1$ such that $g' = f$?

3. Derivative Test for Local Extrema

Recall how hard we had to work in Example 9.11 to determine whether the function f had a local maximum or a local minimum at $x = 2$. We now provide a simple general test that will enable us to classify local extrema easily.

The test for classifying local extrema is based on the sign of the derivative. The following theorem shows what the sign of the derivative of a function says about the function. After we prove the theorem and discuss it, we give the test for classifying local extrema (Theorem 10.19).

Theorem 10.17: Assume that f is continuous on $[a, b]$ and differentiable on (a, b).

(1) If $f'(x) > 0$ for all $x \in (a, b)$, then f is strictly increasing on $[a, b]$.

(2) If $f'(x) < 0$ for all $x \in (a, b)$, then f is strictly decreasing on $[a, b]$.

Proof: Let $x_1, x_2 \in [a, b]$ such that $x_1 < x_2$. By the Mean Value Theorem (Theorem 10.2), there is a point $p \in (x_1, x_2)$ such that

$$f'(p) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

Thus, under the assumption in part (1), $f(x_2) > f(x_1)$ and, under the assumption in part (2), $f(x_2) < f(x_1)$. ¥

Exercise 10.18: Give examples to show that the converses of parts (1) and (2) are false. However, prove the following partial converses to parts (1) and (2): If f is increasing (decreasing) on $[a, b]$, then $f'(x) \geq 0$ ($f'(x) \leq 0$, respectively) for all $x \in [a, b]$.

Theorem 10.17 is intuitively obvious: If all the tangent lines to the graph of a differentiable function have positive slopes, hence are strictly increasing, then surely the function is strictly increasing. However persuasive this argument may seem, it is still not a proof; it is no more a proof than saying, as a "proof" for Theorem 10.7, that if every tangent line to the graph of f is horizontal, then surely the function f is constant. The proof we gave for Theorem 10.17 is certainly short, deceptively short because the proof rests on so many previous results: The proof of Theorem 10.17 used the Mean Value Theorem, whose proof used Rolle's Theorem, whose proof depended essentially on the Maximum - Minimum Theorem; the proof of the Maximum - Minimum Theorem was by no means trivial and depended indispensably on the Completeness Axiom. Thus, in the final analysis, the underlying reason Theorem 10.17 is true is the Completeness Axiom. We conclude that Theorem 10.17 is not as obvious as it would seem to be or as trivial as its brief proof would suggest.

It is worthwhile to consider Theorem 10.17 and Theorem 10.7 together: The theorems show that the sign of a derivative on an interval has a lot to say about the nature of a function.

One final comment about Theorem 10.17: A differentiable function on an open interval can have a positive derivative at a particular point but not be strictly increasing in any neighborhood of the point. See Exercise 10.53.

We are ready to prove the derivative test for classifying local extrema.

Theorem 10.19 (First Derivative Test for Local Extrema): Let I be an open interval, and let $p \in I$. Assume that f is continuous on I and differentiable at each point of I except possibly at p. Let $[s, t] \subset I$ such that $p \in (s, t)$.

(1) If $f'(x) > 0$ for all $x \in (s, p)$ and $f'(x) < 0$ for all $x \in (p, t)$, then f has a local maximum at p.

(2) If $f'(x) < 0$ for all $x \in (s, p)$ and $f'(x) > 0$ for all $x \in (p, t)$, then f has a local minimum at p.

(3) If $f'(x) > 0$ for all $x \in (s,p) \cup (p,t)$, or if $f'(x) < 0$ for all $x \in (s,p) \cup (p,t)$, then f does not have a local extremum at p.

Proof: Assume the conditions in part (1). Then, by Theorem 10.17, f is strictly increasing on $(s,p]$ and f is strictly decreasing on $[p,t)$. Hence, $f(x) < f(p)$ when $s < x < p$ and $f(p) > f(x)$ when $p < x < t$. Thus, $f(p) \geq f(x)$ for all $x \in (s,t)$. Therefore, f has a local maximum at p. This proves part (1).

The proof of part (2) is similar.

We prove part (3) for the case when $f'(x) > 0$ for all $x \in (s,p) \cup (p,t)$. In this case, we have by Theorem 10.17 that f is strictly increasing on $(s,p]$ and on $[p,t)$. It follows easily that f is strictly increasing on (s,t). Thus,

$$f(y) < f(p) < f(z) \text{ whenever } s < y < p < z < t.$$

Therefore, we see that f does not have a local extremum at p (we leave the details to the reader).

The proof of part (3) for the case when $f'(x) < 0$ for all $x \in (s,p) \cup (p,t)$ is similar. ¥

The converse of part (1) of Theorem 10.19 is false (Exercise 10.26).

We illustrate how well the First Derivative Test for Local Extrema works:

Example 10.20: Let $f(x) = 2x^5 - 5x^4 - 10x^3$ for all $x \in \mathsf{R}^1$. We find all points at which f has local and global extrema and determine which extrema are local (or global) minima and which are local (or global) maxima. We also determine the maximal intervals on which f is strictly increasing or strictly decreasing. Finally, we sketch the graph of f using the information we have obtained (however, the sketch is incomplete, as we will see).

By the formula for differentiating polynomials (Theorem 7.12),

$$f'(x) = 10x^4 - 20x^3 - 30x^2.$$

To find where $f'(x) = 0$ (in order to apply Theorem 9.7), we factor $f'(x)$:

$$f'(x) = 10x^2(x^2 - 2x - 3) = 10x^2(x-3)(x+1).$$

Hence, by Theorem 9.7, the only possible points at which f has local extrema are $x = -1, 0, 3$.

The critical step for using Theorem 10.19 is to find the sign of f' on small intervals about the points $x = -1, 0, 3$. How small do we need the intervals to be? The answer comes from noting that f' is continuous: Hence, we can apply the Intermediate Value Theorem (Theorem 5.2) to f' to know that f' can not have opposite signs at two points without being 0 somewhere between the two points; thus, we only need to check the signs of f' at one point of each of the open intervals determined by the points $x = -1, 0, 3$. We can do this readily by inspecting the factored form of f'; we obtain the table below:

interval $\rightarrow$	$(-\infty, -1)$	$(-1, 0)$	$(0, 3)$	$(3, \infty)$
$sign f'(x) \rightarrow$	$+$	$-$	$-$	$+$

From the table and from Theorem 10.19, f has a local maximum at $x = -1$, a local minimum at $x = 3$, and no local extremum at $x = 0$. Furthermore, from the table and Theorem 10.17, the maximal intervals on which f is strictly increasing are $(-\infty, -1]$ and $[3, \infty)$, and the maximal interval on which f is strictly decreasing is $[-1, 3]$.

Next, we see that f has no global extrema: f has no global maximum since its only local maximum is $f(-1) = 3$ and $f(4) = 128$; f has no global minimum since its only local minimum is $f(3) = -189$ and $f(-3) = -621$. Actually, we can see that f has no global extrema without these types of numerical computations: Simply note that

$$f(x) = x^5(2 - \tfrac{5}{x} - \tfrac{10}{x^2}) \text{ for } x \neq 0,$$

which easily shows that f is neither bounded above nor bounded below.

Finally, using the information available, we obtain a picture of the graph of f (Figure 10.20 below). However, something is wrong: $f'(0) = 0$, so the x-axis should be tangent to the graph of f at the origin. In correcting this flaw, we must change the shape of the graph at some point to the left of the origin; we must also change the shape of the graph at some point to the right of the origin in order to avoid having a cusp at $(3, f(3))$. At the present time, it is not at all obvious where these changes should be made; moreover, for all we know, there may be many such changes, perhaps even at points $x < -1$ or at points $x > 3$. If this makes you wonder whether you really know how to graph $y = x^2$, then that is good!

We return to the problem of what is wrong with the graph of f in the next section. There we develop general ideas that solve the problem and that can be applied to other graphs. We arrive at a correct graph of the function f in Example 10.34.

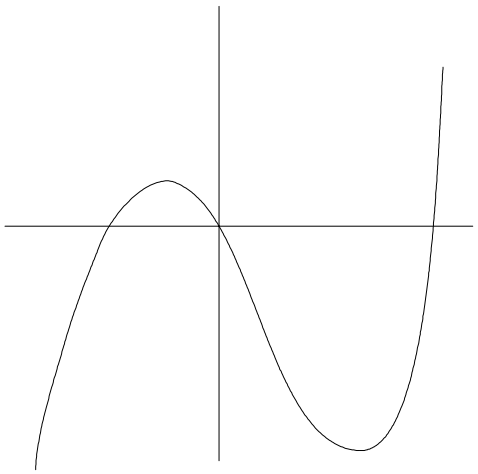


Figure 10.20

Exercise 10.21: Define $f : [0, 6] \to \mathsf{R}^1$ by $f(x) = 4x^3 - 36x^2 + 77x$. Find all points at which f has local and global extrema, determine which extrema are local (or global) minima and which are local (or global) maxima, and determine the maximal intervals on which f is strictly increasing or strictly decreasing. Sketch the graph of f and discuss possible flaws in your graph as per the discussion of the graph we gave for Example 10.19. (We briefly discussed the function f after the proof of the Maximum - Minimum Theorem (Theorem 5.13)).

Exercise 10.22: Define $f : [-2, 2] \to \mathsf{R}^1$ by $f(x) = x^4 - 2x^2 + 1$. Repeat Exercise 10.21 for this function.

Exercise 10.23: Define $f : [0, 2] \to \mathsf{R}^1$ by $f(x) = \frac{x}{x^2+1}$. Repeat Exercise 10.21 for this function.

Exercise 10.24: In Figure 10.24 below, we have drawn a picture of the graph of the derivative of a function f. Determine all points at which f has local and global extrema, determine which extrema are local (or global) minima and which are local (or global) maxima, and determine the maximal intervals on which f is strictly increasing or strictly decreasing. Sketch the graph of f assuming that $f(0) = 0$.

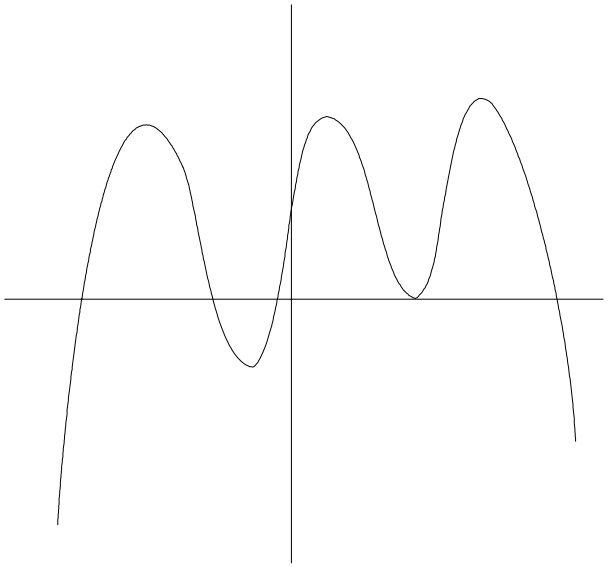


Figure 10.24

Exercise 10.25: Let $f, g : \mathsf{R}^1 \to \mathsf{R}^1$ be differentiable functions such that $f'(x) < g'(x)$ for all $x \in \mathsf{R}^1$. Then there is at most one point p such that $f(p) = g(p)$.

Exercise 10.26: Draw a picture of the graph of a differentiable function on an open interval such that the function has a unique global maximum at a point p for which part (1) Theorem 10.19 does not apply.

4. Concavity

This section follows up on the discussion above Figure 10.20: We introduce concepts that describe the flaws in the preliminary graph in Figure 10.20 and that we can use to refine our graphing techniques in general. Specifically, we define the notions of concavity and inflection point, and we obtain results that connect the notions to derivatives. At the end of the section, we sketch the graph for Example 10.20 (this time correctly!).

Let I be an interval, let $a, b \in I$ such that $a \neq b$, and let $f : I \to \mathsf{R}^1$ be a function. The *chord joining* $(a, f(a))$ *and* $(b, f(b))$ is the line segment in the plane with end points $(a, f(a))$ and $(b, f(b))$.

Definition: Let I be an interval, and let $f : I \to \mathsf{R}^1$ be a function.

- We say that f is *concave up on* I provided that for any two different points $a, b \in I$, the chord joining $(a, f(a))$ and $(b, f(b))$ lies above the graph of f on (a, b); in other words,

$$f(x) < \tfrac{f(b)-f(a)}{b-a}(x-a) + f(a) \quad \text{when } a < x < b.$$

- We say that f is *concave down on* I provided that for any two different points $a, b \in I$, the chord joining $(a, f(a))$ and $(b, f(b))$ lies below the graph of f on (a, b); in other words,

$$f(x) > \tfrac{f(b)-f(a)}{b-a}(x-a) + f(a) \quad \text{when } a < x < b.$$

For example, $f(x) = x^3$ is concave up on $[0, \infty)$ and concave down on $(-\infty, 0]$. On the other hand, a linear function $f(x) = mx + b$ is not concave up or down on any interval.

In Theorem 10.29, we characterize the two types of concavity for a differentiable function on an interval in terms of the derivative of the function.

Lemma 10.27: Let I be an interval, let $f : I \to \mathsf{R}^1$ be a function, and let $x_1, x_2, x_3 \in I$ such that $x_1 < x_2 < x_3$. For each $i \neq j$, let $C_{i,j}$ denote the the chord joining $(x_i, f(x_i))$ and $(x_j, f(x_j))$.

(1) If f is concave up on I, then

$$\text{slope of } C_{1,2} < \text{slope of } C_{1,3} < \text{slope of } C_{2,3}.$$

(2) If f is concave down on I, then

$$\text{slope of } C_{1,2} > \text{slope of } C_{1,3} > \text{slope of } C_{2,3}.$$

Proof: We prove part (1); we leave the proof of part (2) to the reader (Exercise 10.28).

Assume that f is concave up on I. Let y_2 denote the second coordinate of the point on $C_{1,3}$ with first coordinate x_2. Since f is concave up on I, $f(x_2) < y_2$; hence, $f(x_2) - f(x_1) < y_2 - f(x_1)$. Thus,

$$\frac{f(x_2)-f(x_1)}{x_2-x_1} < \frac{y_2-f(x_1)}{x_2-x_1}.$$

Therefore, since the slope of $C_{1,2} = \frac{f(x_2)-f(x_1)}{x_2-x_1}$ and the slope of $C_{1,3} = \frac{y_2-f(x_1)}{x_2-x_1}$, we have proved that the slope of $C_{1,2}$ < slope of $C_{1,3}$.

The proof of the second inequality in part (1) is similar to what we have done: We rewrite $f(x_2) < y_2$ as $-y_2 < -f(x_2)$; then $f(x_3)-y_2 < f(x_3)-f(x_2)$. Thus,

$$\frac{f(x_3)-y_2}{x_3-x_2} < \frac{f(x_3)-f(x_2)}{x_3-x_2}.$$

Therefore, since the slope of $C_{1,3} = \frac{f(x_3)-y_2}{x_3-x_2}$ and the slope of $C_{2,3} = \frac{f(x_3)-f(x_2)}{x_3-x_2}$, we have proved that the slope of $C_{1,3}$ < slope of $C_{2,3}$.

This completes the proof of part (1) of the lemma. ¥

Exercise 10.28: Formulate and prove a simple theorem that can be applied to prove part (2) of Lemma 10.27 directly from part (1).

Theorem 10.29: Assume that f is differentiable on an open interval I.

(1) f is concave up on I if and only if f' is strictly increasing on I.

(2) f is concave down on I if and only if f' is strictly decreasing on I.

Proof: We prove part (1), leaving the proof of part (2) to the reader (Exercise 10.30).

Assume that f is concave up on I. Fix points $a, b \in I$ such that $a < b$. We show that $f'(a) < f'(b)$.

Fix points c and d such that $a < c < d < b$. Then, using part (1) of Lemma 10.27 twice, we see that

(i) $\frac{f(c)-f(a)}{c-a} < \frac{f(d)-f(a)}{d-a} < \frac{f(c)-f(d)}{c-d} < \frac{f(c)-f(b)}{c-b} < \frac{f(d)-f(b)}{d-b}$.

Then, using (1) of Lemma 10.27 for the first and third inequalities below,

$$\frac{f(x)-f(a)}{x-a} < \frac{f(c)-f(a)}{c-a} \overset{\text{(i)}}{<} \frac{f(d)-f(b)}{d-b} < \frac{f(y)-f(b)}{y-b}, \text{ if } a < x < c \text{ and } d < y < b.$$

Thus, since $f'(a) = \lim_{x\to a} \frac{f(x)-f(a)}{x-a}$ and $f'(b) = \lim_{y\to b} \frac{f(y)-f(b)}{y-b}$ (by Exercise 6.10), we have that $f'(a) < f'(b)$. This proves that f' is strictly increasing on I.

Conversely, assume that f' is strictly increasing on I. Fix points $s, t \in I$ such that $s < t$. Fix a point x such that $s < x < t$. Note that f is continuous on $[s, t]$ by Theorem 6.14 (and Exercise 5.3). Therefore, we can apply the Mean Value Theorem (Theorem 10.2) to obtain points $p \in (s, x)$ and $q \in (x, t)$ such that

$$f'(p) = \frac{f(x)-f(s)}{x-s}, \quad f'(q) = \frac{f(t)-f(x)}{t-x}.$$

Thus, since f' is strictly increasing on (s, t) and $p < q$, we have that

(ii) $\frac{f(x)-f(s)}{x-s} < \frac{f(t)-f(x)}{t-x}$.

We now show that $f(x) < \frac{f(t)-f(s)}{t-s}(x-s) + f(s)$, which proves that f is concave up. By (ii),

$$[f(x) - f(s)](t - x) < [f(t) - f(x)](x - s);$$

hence,

$$f(x)(t - s) < f(t)(x - s) + f(s)(t - x);$$

thus,

(iii) $f(x) < f(t)\frac{x-s}{t-s} + f(s)\frac{t-x}{t-s}$.

Finally, subtracting and adding $f(s)\frac{x-s}{t-s}$ to the right - hand side of (iii), we have

$$\begin{aligned} f(x) &< \tfrac{f(t)-f(s)}{t-s}(x - s) + f(s)\tfrac{x-s}{t-s} + f(s)\tfrac{t-x}{t-s} \\ &= \tfrac{f(t)-f(s)}{t-s}(x - s) + f(s)\tfrac{t-s}{t-s} = \tfrac{f(t)-f(s)}{t-s}(x - s) + f(s). \end{aligned}$$ ¥

Exercise 10.30: Show that part (2) of Theorem 10.29 follows easily from part (1) using the theorem you discovered in Exercise 10.28.

In order to easily apply Theorem 10.29 to determine concavity, we need a simple test to determine whether a derivative f' is strictly increasing or strictly decreasing. Theorem 10.17 provides such a test when f' is differentiable; the corollary below states the test precisely.

We denote the derivative of f' by f''; f'' is called the *second derivative of* f. A function f that has a second derivative is said to be *twice differentiable*.

Corollary 10.31: Assume that f is twice differentiable on an open interval I.

(1) If $f''(x) > 0$ for all $x \in I$, then f is concave up on I.

(2) If $f''(x) < 0$ for all $x \in I$, then f is concave down on I.

Proof: The corollary follows directly from Theorems 10.17 and 10.29. ¥

We make two observations about Corollary 10.31. First, Corollary 10.31 does not apply to all differentiable functions since a function can be differentiable and yet not be twice differentiable. Second, the converses of parts (1) and (2) of Corollary 10.31 are false; for example, $f(x) = x^4$ shows the converse of part (1) is false.

Exercise 10.32: Verify the statements in the preceding paragraph (include an example to show that the converse of part (2) of Corollary 10.31 is false).

We will be concerned with points at which the concavity of a function changes. For example, we say that the function $f(x) = x^3$ changes concavity at $x = 0$ because f is concave down on $(-\infty, 0)$ and concave up on $(0, \infty)$. We give the following precise, general definition for change in concavity:

Definition. Let f be a function defined on an open interval I, and let $p \in I$. We say that f *changes concavity at the point* p provided that for some interval $(s, t) \subset I$, $f|(s, p)$ is concave one way (up or down) and $f|(p, t)$ is concave the other way (down or up, respectively).

When a function f changes concavity at a point p, we say that f *has an inflection point at* p, in which case we call $(p, f(p))$ an *inflection point of* f.[5]

The following theorem, in conjunction with Corollary 10.31, can enable us to determine the inflection points of a twice differentiable function; we will illustrate this in Example 10.34. We note that the theorem is analogous to Theorem 9.7 for local extrema.

Theorem 10.33: Assume that f is twice differentiable on an open interval I. If f has an inflection point at p, then $f''(p) = 0$.

Proof: Assume that f has an inflection point at p. Then there is an interval $[s, t] \subset I$ such that $f|(s, p)$ is concave one way and $f|(p, t)$ is concave the other way. Assume that $f|(s, p)$ is concave up and $f|(p, t)$ is concave down. Then, by Theorem 10.29, we have that

(*) f' is strictly increasing on (s, p) and f' is strictly decreasing on (p, t).

Since f' is differentiable and $[s, t] \subset I$, f' is continuous on $[s, t]$ by Theorem 6.14 (and Exercise 5.3). Hence, by the Maximum - Minimum Theorem (Theorem 5.13), the restricted function $f'|[s, t]$ has attains its maximum value at some point q of $[s, t]$. We see from (*) that $q = p$. Thus, f' has a local maximum at p. Therefore, by Theorem 9.7, $f''(p) = 0$.

The proof when $f|(s, p)$ is concave down and $f|(p, t)$ is concave up is similar and is omitted. ¥

Let's see how all this works. We return to Example 10.20:

Example 10.34: Let $f(x) = 2x^5 - 5x^4 - 10x^3$ for all $x \in \mathsf{R}^1$. In Example 10.20 we showed that f has a local maximum at $x = -1$, a local minimum at $x = 3$, and no global extrema. We noted some problems with the graph of f as depicted in Figure 10.20. We are now prepared to address the problems.

We use Theorem 10.33 to find all the *possible* inflection points of f:

$$f''(x) = 40x^3 - 60x^2 - 60x = 20x(2x^2 - 3x - 3);$$

thus, the points x at which $f''(x) = 0$ are $x = 0, \frac{3}{4} \pm \frac{\sqrt{33}}{4}$; hence, by Theorem 10.33, the only possible points at which f could have inflection points are $x = 0, \frac{3}{4} \pm \frac{\sqrt{33}}{4}$. Next, we use Corollary 10.31 to see which of the points $0, \frac{3}{4} \pm \frac{\sqrt{33}}{4}$ *is* a point at which f has an inflection point.

Note that f'' is continuous; thus, to apply Corollary 10.31, we only need to check the signs of f'' at one point of each of the open intervals determined by the points $x = 0, \frac{3}{4} \pm \frac{\sqrt{33}}{4}$ (we are using the Intermediate Value Theorem (Theorem 5.2)). Without using specific values for x, but merely inspecting the expression $f''(x) = 20x(2x^2 - 3x - 3)$ for any x very negative, for any $x < 0$ and very near 0, for any $x > 0$ and very near 0, and for any x very large (positive), we arrive at the following table:

[5] Notice in the definition of inflection point that an inflection point of f is a point $(\mathsf{p}, \mathsf{f}(\mathsf{p}))$ on the graph of f, not the point p; the distinction emphasizes the fact that the graph of f is where the geometry inherent in the notion of inflection point is visible.

interval $\rightarrow$	$(-\infty, \frac{3}{4} - \frac{\sqrt{33}}{4})$	$(\frac{3}{4} - \frac{\sqrt{33}}{4}, 0)$	$(0, \frac{3}{4} + \frac{\sqrt{33}}{4}$	$(\frac{3}{4} + \frac{\sqrt{33}}{4}, \infty)$
$sign f''(x) \rightarrow$	$-$	$+$	$-$	$+$

Hence, by Corollary 10.31, f changes concavity at each of the points $x = 0, \frac{3}{4} \pm \frac{\sqrt{33}}{4}$. Therefore, f has an inflection point at each of these points and at no other point (by Theorem 10.33).

We also know from the table and Corollary 10.31 that f is concave up on $(\frac{3}{4} - \frac{\sqrt{33}}{4}, 0) \cup (\frac{3}{4} + \frac{\sqrt{33}}{4}, \infty)$ and that f is concave down on $(-\infty, \frac{3}{4} - \frac{\sqrt{33}}{4}) \cup (0, \frac{3}{4} + \frac{\sqrt{33}}{4})$.

Taking into account inflection points and concavity, we correct the graph of f that we drew in Figure 10.20:

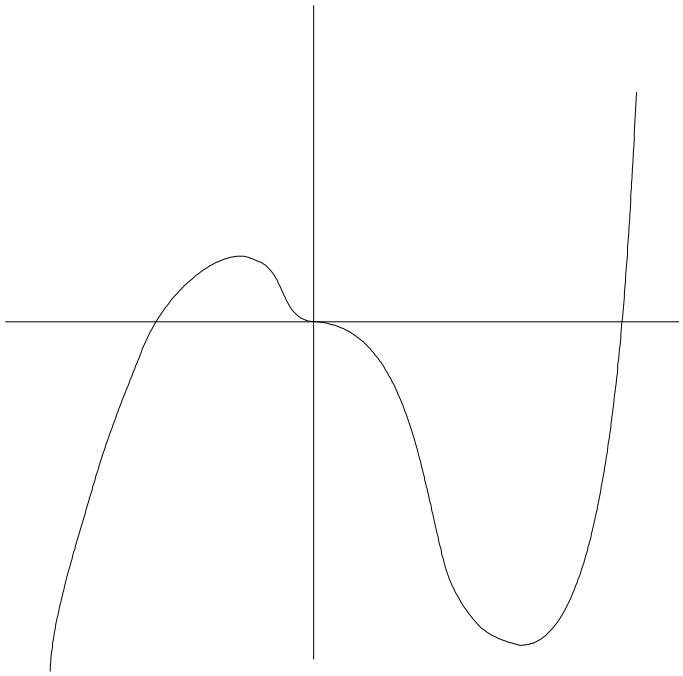


Figure 10.34

We conclude with an interesting theorem about polynomials. What really makes the theorem interesting is that the theorem is not true for differentiable functions in general, as you will be asked to show in Exercise 10.45. The proof of the theorem uses several previous results and must be done carefully (you will gain appreciation for the proof if you keep your solution to Exercise 10.26 in mind as you read the proof).

Theorem 10.35: A nonconstant polynomial has an inflection point at some point between any two points at which the polynomial has local extrema.

Proof: Assume the f is a polynomial with local extrema at p and q with $p < q$. Note that f has degree ≥ 3 (since nonconstant polynomials of degree ≤ 2 do not have two local extrema).

By Theorem 9.7, $f'(p) = 0$ and $f'(q) = 0$. Note that f' is a polynomial of degree ≥ 2 (by Theorem 7.12); hence, by Exercise 9.21, we can assume p and q were chosen so that $f'(x) \neq 0$ for all $x \in (p, q)$. Therefore, by the Intermediate Value Theorem (Theorem 5.2), $f'(x) > 0$ for all $x \in (p, q)$ or $f'(x) < 0$ for all $x \in (p, q)$. We assume for the proof that $f'(x) > 0$ for all $x \in (p, q)$ (the proof for the other case is similar).

By the Maximum - Minimum Theorem (Theorem 5.13), f' attains a maximum value on $[p, q]$ at a point r. Since $f'(p) = 0 = f'(q)$ and $f'(x) > 0$ for all $x \in (p, q)$, it is clear that $r \in (p, q)$. Hence, by Theorem 9.7, $f''(r) = 0$.

Now, since f'' is a polynomial of degree ≥ 1 (by Theorem 7.12), we see from Exercise 9.21 that there is a subinterval $[s, t]$ of (p, q) such that $r \in (s, t)$ and

$$f''(x) \neq 0 \text{ for all } x \in (s, r) \cup (r, t).$$

Thus, since f'' is continuous (because f'' is a polynomial), we have by the Intermediate Value Theorem that f'' does not change sign on (s, r) and f'' does not change sign on (r, t). Therefore, since f' has a local maximum at r, we see from part (3) of Theorem 10.19 that the sign of f'' on (s, r) is opposite to the sign of f'' on (r, t). Therefore, by Corollary 10.31, f changes concavity at r; in other words, f has an inflection point at r. ¥

In some exercises, we will ask you to find maximal intervals on which a function is concave up or down. The following theorem should be kept in mind when finding such maximal intervals:

Theorem 10.36: If f is continuous on $[a, b]$ and concave up (down) on (a, b), then f is concave up (down, respectively) on $[a, b]$.

Proof: Left as the first exercise below. ¥

Exercise 10.37: Prove Theorem 10.36.

Exercise 10.38: Define $f : [0, 6] \to \mathsf{R}^1$ by $f(x) = 4x^3 - 36x^2 + 77x$. Continue the analysis of f begun in Exercise 10.21 by finding all points at which f has an inflection point and determining the maximal intervals on which f is concave up or down. Sketch the graph of f eliminating flaws that may have occurred in your sketch for Exercise 10.21

Exercise 10.39: Define $f : [-2, 2] \to \mathsf{R}^1$ by $f(x) = x^4 - 2x^2 + 1$. This is the function in Exercise 10.22. Repeat Exercise 10.38 for this function.

Exercise 10.40: Define $f : [0, 2] \to \mathsf{R}^1$ by $f(x) = \frac{x}{x^2+1}$. This is the function in Exercise 10.23. Repeat Exercise 10.38 for this function.

Exercise 10.41: Sketch the graph of $f(x) = 8x^5 - 5x^4 - 20x^3 + 1$ identifying all local extrema, inflection points, and concavity.

Exercise 10.42: Repeat Exercise 10.41 for $f(x) = x^{\frac{2}{3}}(6 - x)^{\frac{1}{3}}$.

Exercise 10.43: Repeat Exercise 10.41 for $f(x) = |x|(x+1)$.

Exercise 10.44: In Figure 10.44 below, we have drawn a picture of the graph of the second derivative of a function f. Assuming that $f(0) = 0$ and that $f'(0) = 0$, sketch the graph of f indicating the points at which all local extrema occur, the points at which f has inflection points, and the maximal intervals on which f is concave up or down.

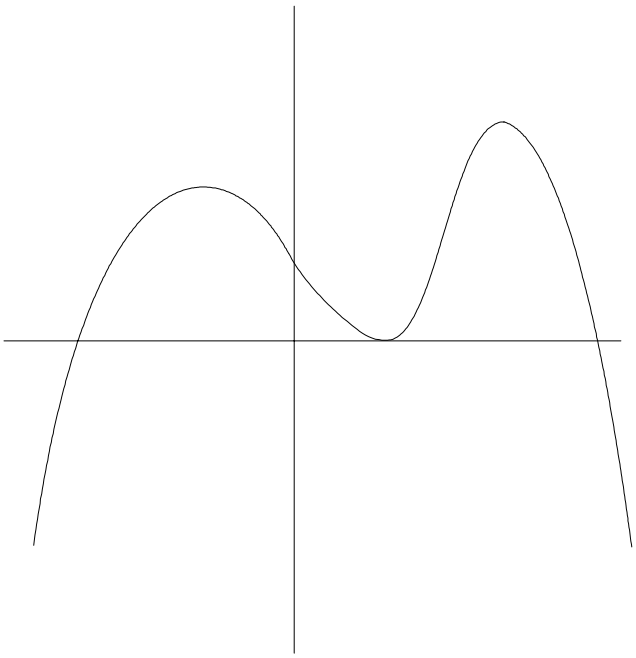


Figure 10.44

Exercise 10.45: Theorem 10.35 is not necessarily true for differentiable functions that are not polynomials. Show this by giving an example of a differentiable function on R^1 that has local extrema at different points but that has no inflection point.

Exercise 10.46: Any polynomial f of degree 3 has exactly one inflection point. Furthermore, if f crosses the x-axis at three distinct points a, b, c (i.e., has three distinct roots), then the inflection point of f occurs at the average $x = \frac{a+b+c}{3}$ of the roots.

Exercise 10.47: Let I be an open interval, and let $f : I \to \mathsf{R}^1$ be differentiable on I. Then f is concave up (down) on I if and only if for each $x \in I$, the graph of f lies above (below, respectively) the tangent line to the graph of f at $(x, f(x))$ except for the point $(x, f(x))$ itself.

5. Intermediate Value Property for Derivatives

When we sketched graphs of specific functions, we determined the sign of a derivative or a second derivative on an interval (complementary to the critical points) using the following procedure: We checked the sign at one point in the interval and then appealed to the Intermediate Value Theorem (Theorem 5.2) to conclude that the sign was the same throughout the interval. This works fine as long as the derivatives are continuous. Can a derivative fail to be continuous? If so, is there a systematic way to check signs for such a derivative in order to apply various tests easily? (I am referring to the tests in Theorem 10.19 and Corollary 10.31.)

The answer to the first question is yes, a derivative can fail to be continuous. The answer to the second question is that the answer to the first question is irrelevant: We can determine the sign of a derivative on an interval the way as we always did – by checking the sign at only one point of the interval – whether the derivative is continuous or not! In other words, derivatives do not change sign on an interval on which they are defined without having value zero at some point of the interval.

We give an example that verifies our answer to the first question, and we give a theorem that explains our answer to the second question.

Example 10.48: We give an example of a differentiable function $f : \mathsf{R}^1 \to \mathsf{R}^1$ such that its derivative is not continuous. Define f by

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

Using various results in Chapter VII and Theorem 8.20, we see that f is differentiable at every point $x \neq 0$ and that

$$f'(x) = x^2[\cos(\tfrac{1}{x})](\tfrac{-1}{x^2}) + 2x\sin(\tfrac{1}{x}) = 2x\sin(\tfrac{1}{x}) - \cos(\tfrac{1}{x}), \quad x \neq 0.$$

furthermore, we see that f is differentiable at $x = 0$ as follows: For $x \neq 0$,

$$0 \leq \left|\tfrac{f(x)-f(0)}{x-0}\right| = \left|x\sin(\tfrac{1}{x})\right| \leq |x|;$$

thus, since $\lim_{x\to 0} |x| = 0$, the Squeeze Theorem (Theorem 4.34) applies to give us that

$$\lim_{x\to 0} \left|\tfrac{f(x)-f(0)}{x-0}\right| = 0.$$

This proves that $f'(0) = 0$ (recall Exercise 6.10).

Finally, we show that f' is not continuous at 0 by showing that $\lim_{x\to 0} f'(x)$ does not exist. Recall that

$$f'(x) = x^2[\cos(\tfrac{1}{x})](\tfrac{-1}{x^2}) + 2x\sin(\tfrac{1}{x}) = 2x\sin(\tfrac{1}{x}) - \cos(\tfrac{1}{x}), \quad x \neq 0.$$

Note that

$$0 \leq \left|2x \sin(\tfrac{1}{x})\right| \leq |2x|, \quad x \neq 0;$$

thus, since $\lim_{x \to 0} |2x| = 0$, we have by the Squeeze Theorem (Theorem 4.34) that

$$\lim_{x \to 0} 2x \sin(\tfrac{1}{x}) = 0.$$

Hence, if $\lim_{x \to 0} f'(x)$ existed, then we would have

$$\lim_{x \to 0} \cos(\tfrac{1}{x}) \stackrel{4.2}{=} \lim_{x \to 0} 2x \sin(\tfrac{1}{x}) - \lim_{x \to 0} f'(x),$$

which is impossible (since, as is clear, $\lim_{x \to 0} \cos(\frac{1}{x})$ does not exist).

Next, we show why, even though derivatives may not be continuous, we can determine the sign of a derivative on an interval complementary to the critical points by checking the sign at only one point of the interval. The reason is simple enough – derivatives, continuous or not, satisfy the conclusion to the Intermediate Value Theorem (Theorem 5.2). We prove this in Theorem 10.50. First, we introduce relevant terminology and discuss the notion we define. (The terminology carries the name of the French mathematician G. Darboux (1842 - 1917) who proved the theorem we will prove.)

Definition: Let I be an interval, and let $f : I \to \mathsf{R}^1$ be a function. We say that f is a *Darboux function* provided that for any two points $p, q \in I$ and any point y between $f(p)$ and $f(q)$, there is a point x between p and q such that $f(x) = y$ (i.e., for any subinterval J of I, $f(J)$ is an interval).

There are fairly simple functions that are Darboux but not continuous: For example, let

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

Actually, the derivative f' of the function in Example 10.48 is another example of a discontinuous Darboux function. This fact about the function in Example 10.48 illustrates the content of the theorem we will prove: Any derivative on an interval is a Darboux function.

We use the following lemma in the proof of our theorem.

Lemma 10.49: Let f be a continuous function on an interval I, and let C denote the set of all slopes of chords joining any two points on the graph of f; that is,

$$C = \{\tfrac{f(s)-f(r)}{s-r} : s, r \in I \text{ and } s \neq r\}.$$

Then C is an interval.

Proof: Fix $p \in C$, say

$$p = \tfrac{f(a)-f(b)}{a-b}, \quad a < b.$$

We show that there is an interval in C joining p to any other point of C. To this end, let $z \in C$, say

$$z = \tfrac{f(u)-f(v)}{u-v}, \quad u < v.$$

Note that since $a - b < 0$ and $u - v < 0$, $[1-t](a-b) + t(u-v) \neq 0$ for all $t \in [0,1]$; in anticipation of what comes next, we write this as follows:

$$[1-t]a + tu - [1-t]b - tv \neq 0 \quad \text{for all } t \in [0,1].$$

Hence, the following formula defines a function $\sigma : [0,1] \to C$ such that $\sigma(0) = p$ and $\sigma(1) = z$ as follows:

$$\sigma(t) = \tfrac{f([1-t]a+tu)-f([1-t]b+tv)}{[1-t]a+tu-[1-t]b-tv}, \quad \text{all } t \in [0,1].$$

By the continuity of f and by various theorems about continuity in Chapter IV (notably, 4.4, 4.21 and 4.28), we see that σ is continuous. Thus, by the Intermediate Value Theorem (Theorem 5.2), $\sigma([0,1])$ is an interval. Therefore, since $\sigma(0) = p$ and $\sigma(1) = z$, we have proved that p and any other point z of C lie in an interval in C. It now follows easily that C is an interval. ¥

Theorem 10.50: If f is a differentiable function on an interval I, then f' is a Darboux function.

Proof: Let D be the set of all values of the first derivative of f on I,

$$D = \{f'(x) : x \in I\}.$$

We prove that D is an interval, which is simply another way of stating the theorem we are proving.

Let C be as in Lemma 10.49. Since f is continuous by Theorem 6.14, C is an interval by Lemma 10.49. Let E denote the set of end points of C (E may be empty).

The Mean Value Theorem (Theorem 10.2) says that $C \subset D$. The definition of the derivative says that every value of the first derivative of f is a limit of slopes of chords; hence, $D \subset C \cup E$ (since $C \cup E = C^{\sim}$, where $C^{\sim}$ is the set of all points arbitrarily close to C, as defined in sections 1 and 2 of Chapter II).

We have proved that

$$C \text{ is an interval and } C \subset D \subset C \cup E.$$

Therefore, it follows at once that D is an interval. ¥

In Exercise 10.16 you were asked if a certain function with a simple discontinuity was a derivative of some function. You probably worked the exercise in a fairly computational way. Theorem 10.50 yields the solution to Exercise 10.16 immediately and furnishes a completely different perspective on the exercise. We briefly discuss the situation in general.

Let f be a function defined on an open interval I. Then f is said to have a *simple discontinuity at a point* $p \in I$, sometimes called a *discontinuity of the first kind*, provided that f is not continuous at p and $\lim_{x\to p^-} f(x)$ and $\lim_{x\to p^+} f(x)$ exist. The function f is said to have a *discontinuity of the second kind at* p provided that f is not continuous at p and f does not have a simple discontinuity at p.

There are exactly two ways a function can have a simple discontinuity at p: Either $\lim_{x\to p^-} f(x) \neq \lim_{x\to p^+} f(x)$ or $\lim_{x\to p^-} f(x) = \lim_{x\to p^+} f(x) \neq f(p)$.

Corollary 10.51: If f is a differentiable function on an open interval I, then f' has no simple discontinuities.

Proof: Left as the first exercise below. ¥

Exercise 10.52: Prove Corollary 10.51. In fact, prove that the corollary extends to all Darboux functions; that is, any discontinuity of a Darboux function on an open interval is a discontinuity of the second kind.

Exercise 10.53: A differentiable function on R^1 can have derivative equal to zero at a point and yet not have a local extremum at the point (e.g., $f(x) = x^3$). Similarly, a differentiable function on R^1 can have a positive derivative at a point without being strictly increasing in any neighborhood of the point (compare with Theorem 10.17): Modify the function in Example 10.48 to give an example.

(*Hint:* Think geometrically: modify the graph of the function in Example 10.48.)

Exercise 10.54: Define $f : [\frac{9}{10}, \frac{21}{10}] \to \mathsf{R}^1$ by $f(x) = x^4 - 6x^3 + 12x^2$. Find

$$D = \{f'(x) : x \in [\tfrac{9}{10}, \tfrac{21}{10}]\}, \quad C = \{\tfrac{f(s)-f(r)}{s-r} : s, r \in [\tfrac{9}{10}, \tfrac{21}{10}] \text{ and } s \neq r\};$$

D and C are the sets in the proof of Theorem 10.50.

Exercise 10.55: Let $f : \mathsf{R}^1 \to \mathsf{R}^1$ be a polynomial of odd degree. Theorem 10.50 implies that the set D of all values of the first derivative of f is an interval. What types of intervals can D be? What types of intervals can the set C in Lemma 10.49 be?

Exercise 10.56: Repeat Exercise 10.55 for the case when f is a polynomial of even degree.

Exercise 10.57: Prove that at most one of the functions f and g below can be a derivative of a function:

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

Chapter XI: Area

The chapter is a bridge between previous chapters and the topic of subsequent chapters (the integral). We simply present an informal, nonrigorous discussion of an aspect of area for the purpose of motivating the integral. Our discussion connects derivatives with area!

Consider the continuous function f whose graph we have drawn in Figure 1 below. We want to find the area between the graph of f and the interval $[a, b]$ on x - axis.

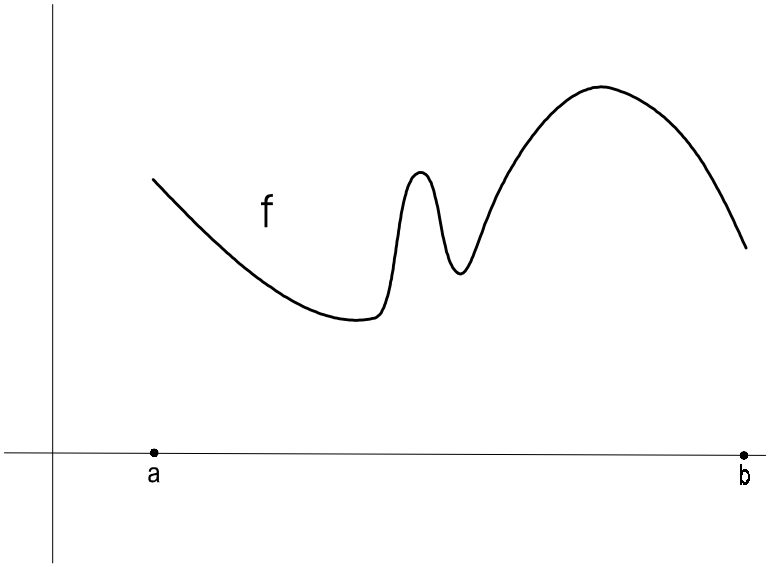


Figure 1

There is an obvious question here: What do we mean by area (referring to the area between the graph of f and the interval $[a, b]$)? We will answer the question in a precise way in Chapter XIV. Here we answer the question somewhat intuitively, and then we describe how to compute the area.

We start by dividing the interval $[a, b]$ into n intervals whose end points are

$$x_0 = a < x_1 < x_2 < \cdots < x_n = b.$$

We think of each of the intervals $[x_{i-1}, x_i]$ as being small, and we consider the rectangles R_i of height $f(x_i)$ and width $x_i - x_{i-1}$, as in Figure 2 (we use $f(x_i)$ as a matter of convenience; we could use $f(t_i)$ for any $t_i \in [x_{i-1}, x_i]$).

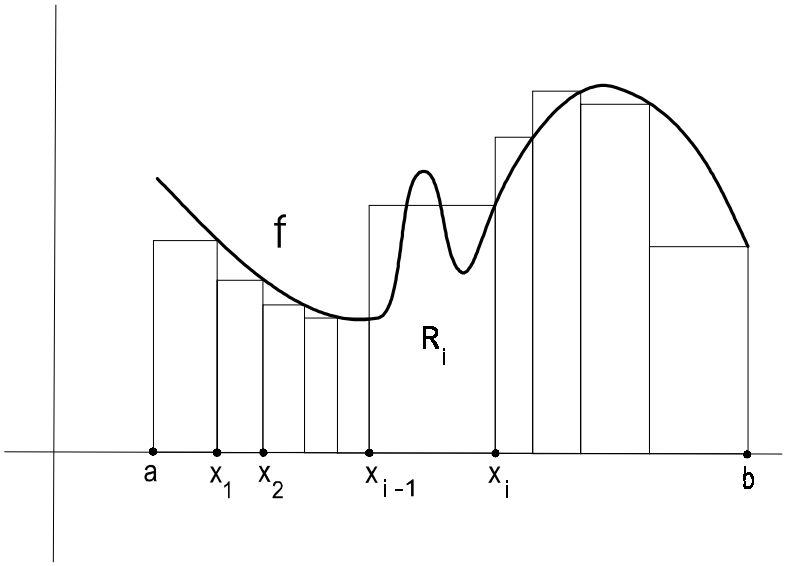


Figure 2

We know from elementary geometry that the area of each rectangle R_i is $f(x_i)(x_i - x_{i-1})$. Thus, the sum $S = \Sigma_{i=1}^{n} f(x_i)(x_i - x_{i-1})$ represents the area of the region covered by all the rectangles. Observe that if x_{i-1} and x_i are very close to one another for each i, then the sum S is very close to what we would call the area between the graph of f and the interval $[a, b]$. Consider dividing the interval $[a, b]$ into more and more subintervals in such a way that the end points x_{i-1} and x_i of the intervals get closer and closer together: If we can compute the "limit" of the sums S associated with the subdivisions, then we will have computed what we would call the area between the graph of f and the interval $[a, b]$.[6]

Now, having indicated what we mean by the area between the graph of f and the interval $[a, b]$, we give a procedure for computing the area. The method is so ingenious that it stands as a monument to human thought.

We make use of the *area function* $A : [a, b] \to \mathsf{R}^1$, defined as follows: For each $x \in [a, b]$, $A(x)$ is the area between the graph of $f|[a, x]$ and the interval $[a, x]$. (We will see in section 2 of Chapter XIV that $A(x)$ is the integral of f over the interval $[a, x]$.)

If we knew a formula for A, computing the area between the graph of f and the interval $[a, b]$ would be easy – we would simply plug b into the formula. Thus, we want to find a formula for A, or at least enough information about A to find $A(b)$.

[6] Note that the "limit" mentioned here is not a limit as we defined the term in Chapter III since each sum S depends on many points $\mathsf{x_i}$. In other words, S is not a function of a single real variable. We have used the term "limit" in an intuitive way – to conjure up a picture in the reader's mind. We give a rigorous definition in section 2 of Chapter XIV.

We "show" that the area function A is differentiable by "computing" its derivative (the quotes mean we show and compute as best as we can without a mathematically precise definition of area). Then we discover what the derivative of A has to do with finding the area we want.

Fix $x \in [a, b]$. In order to find

$$A'(x) = \lim_{h \to 0} \frac{A(x+h) - A(x)}{h},$$

it is clear that we must write the numerator with a factor of h.

We first examine the numerator $A(x + h) - A(x)$ for some given $h > 0$; we assume h to be near enough to 0 so that $x + h < b$ (if $x = b$, we only consider the case when $h < 0$, which we will consider later for any x).

We see from Figure 3 that $A(x+h) - A(x)$ is the area between the graph of $f|[x, x + h]$ and the interval $[x, x + h]$.

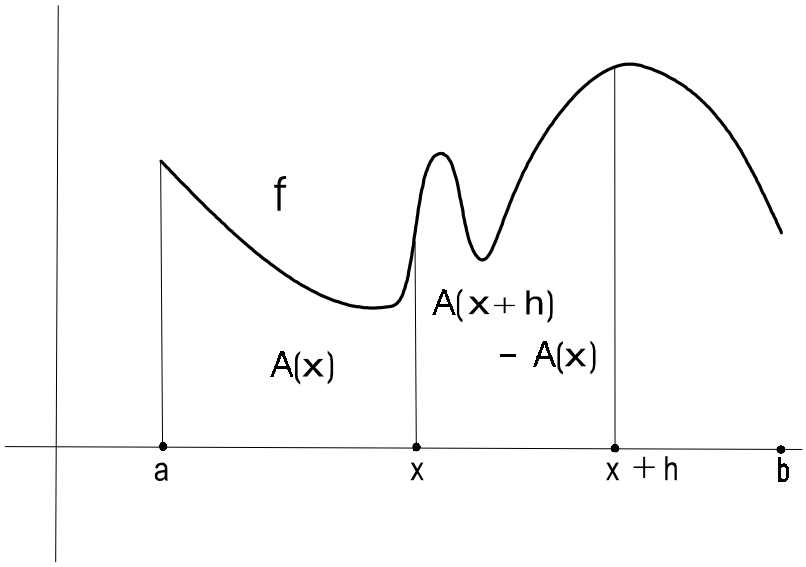


Figure 3

The continuous function f has a maximum value M and a minimum value m on $[x, x+h]$ (by Theorem 5.13). Consider the function $\varphi : [m, M] \to \mathsf{R}^1$ that assigns to a point $t \in [m, M]$ the area of the rectangle $[x, x + h] \times [0, t]$ (see Figure 4); since the height of the rectangle is t and its width is h,

$$\varphi(t) = th \text{ for each } t \in [m, M].$$

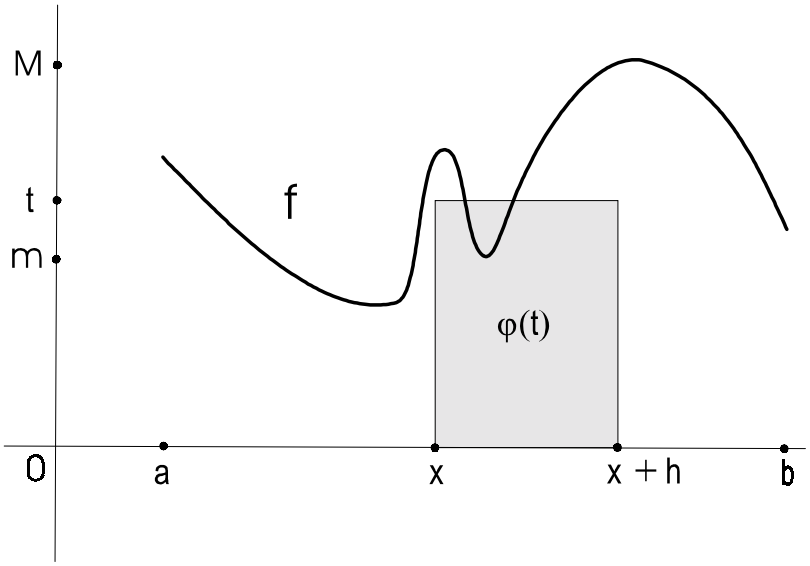


Figure 4

We know that the function φ is continuous (see Example 2.23); furthermore, since $A(x+h) - A(x)$ is the area between the graph of $f|[x, x+h]$ and the interval $[x, x+h]$, we know that

$$\varphi(m) \leq A(x+h) - A(x) \leq \varphi(M).$$

Hence, there is a point $t_h \in [m, M]$ such that $\varphi(t_h) = A(x+h) - A(x)$; in other words,

$$t_h h = A(x+h) - A(x).$$

Now, note that f is continuous on $[x, x+h]$ (by Exercise 5.3); thus, since $t_h \in [m, M]$ and since m and M are values of f on $[x, x+h]$, there is a point $x_h \in [x, x+h]$ such that $f(x_h) = t_h$ (by Theorem 5.2). Therefore, by the previous displayed item, we have

$$(*)\ f(x_h)h = A(x+h) - A(x).$$

The equality in (*) also holds when $h < 0$ (and near enough to 0 so that $x + h > a$): For then the area between the graph of $f|[x+h, x]$ and the interval $[x+h, x]$ is $A(x) - A(x+h)$, and the rectangle $[x+h, x] \times [0, t]$ has width $-h$ for any $t \in [m, M]$; hence, by the analogue of the argument above (in this case, $\varphi(t) = t(-h)$), there is a point $t_h \in [m, M]$ such that

$$t_h(-h) = A(x) - A(x+h),$$

and there is a point $x_h \in [x+h, h]$ such that $f(x_h) = t_h$, thus

$$f(x_h)(-h) = A(x) - A(x+h),$$

which is the same as (*).

We are ready to compute the derivative of A at x: Using that (*) holds whether h is positive or negative, we have that

$$\frac{A(x+h)-A(x)}{h} = \frac{f(x_h)h}{h} = f(x_h), \text{ where } x_h \text{ lies between } x \text{ and } x+h.$$

Hence,

$$A'(x) = \lim_{h\to 0} \frac{A(x+h)-A(x)}{h} = \lim_{h\to 0} f(x_h);$$

furthermore, since $\lim_{h\to 0} x_h = x$ by the Squeeze Theorem (Theorem 4.34) and since f is continuous at x, we see that $\lim_{h\to 0} f(x_h) = f(x)$ (by Theorem 4.29 by considering the function $h \mapsto x_h$). Therefore,

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

So, the derivative of the area function is f; but what does that have to do with computing the area between the graph of f and the interval $[a,b]$? Think about it before reading further. Here is a hint: The area we want to compute is $A(b)$, and $A(b) = A(b) - A(a)$.

We show the way to compute $A(b)$. The method is theoretical, but after we discuss the method we will illustrate that it works quite well in practice.

Let g be *any* function whose derivative on $[a,b]$ is f. Then, since $g' = A'$, g and A differ by a constant (by Theorem 10.8), say $A - g = C$. Thus,

$$A(b) - A(a) = \Big(g(b) + C\Big) - \Big(g(a) + C\Big) = g(b) - g(a).$$

Therefore, since $A(b) = A(b) - A(a)$, we can now conclude the following:

(#) *To find the area between the graph of f and the interval $[a,b]$, we need only find a function g whose derivative on $[a,b]$ is f; then the area between the graph of f and $[a,b]$ is $g(b) - g(a)$.*

We give two examples to illustrate how easy it is to apply the procedure we have found.

Example 11.1: We find the area between the graph of $f(x) = x^2$ and the interval $[1,3]$. The function $g(x) = \frac{x^3}{3}$ has derivative f (by Lemma 7.11); therefore, by (#), the area between the graph of f and the interval $[1,3]$ is

$$g(3) - g(1) = 9 - \tfrac{1}{3} = \tfrac{26}{3}.$$

Example 11.2: We find the area between the graph of $f(x) = x^{\frac{2}{5}} + 3x^3$ and the interval $[1,3]$. The function $g(x) = \frac{5}{7}x^{\frac{7}{5}} + \frac{3}{4}x^4$ has derivative f (by Theorem 7.1 and Theorem 8.16); hence, by (#), the area between the graph of f and the interval $[1,3]$ is

$$g(3) - g(1) = \frac{5}{7}3^{\frac{7}{5}} + \frac{243}{4} - \frac{41}{28}.$$

How do we know that the procedure in (#) really does give the area? The most reasonable way to check this is to see if the procedure gives various areas that are known from geometry. We offer the following exercise as a start:

Exercise 11.3: Show that the procedure in (#) gives the formulas from geometry for the areas of rectangles, triangles and circles.

(*Hint:* In the case of a circle of radius r about the origin, consider the function $g(x) = \frac{x}{2}\sqrt{r^2 - x^2} + \frac{r^2}{2}\sin^{-1}(\frac{x}{2})$.)

When more complicated figures (than those in Exercise 11.3) whose areas are known from geometry are analyzed using the procedure in (#), the answer is always the same: Applying (#) results in arriving at the known areas. In the end, therefore, we will be jusified in *defining* area in terms of the integral and using the procedure in (#) to find the area – see section 2 of Chapter XIV.

We conclude with a few exercises.

Exercise 11.4: Find the area between the graph of $f(x) = \sin(x)$ and the interval $[0, \pi]$.

Exercise 11.5: Find the area between the graph of $f(x) = \frac{1}{\sqrt{1-x^2}}$ and the interval $[0, \frac{1}{2}]$.

Exercise 11.6: Find formulas for the area functions for Examples 11.1 and 11.2.

Exercise 11.7: Using the intuitive observation that the area of two nonoverlapping regions is the sum of the areas of the two regions, find the area above the interval $[0, 1]$ between the graphs of the two functions $f_1(x) = x^4$ and $f_2(x) = x^5$.

Chapter XII: The Integral

In the first part of preceding chapter, we intuitively discussed a way of defining area in order to provide a tangible picture to keep in mind when studying the integral. In this chapter, we begin a rigorous treatment of the integral. This is the first of four chapters concerned directly with the theory of the integral. (There are many types of integrals; we will only study one type – the Riemann integral – which we simply refer to as the integral.)

After presenting preliminary notions and results, we define the integral in section 3. In section 4, we prove an existence theorem that gives a necessary and sufficient condition for a function to be integrable (Theorem 12.15); we also prove a theorem that provides a way (albeit limited) to evaluate the integral (Theorem 12.17). In section 5, we use the existence theorem in section 4 to prove that all continuous functions are integrable.

1. Partitions

In this section (and the next) we present a rigorous and systematic treatment of some of the ideas that we introduced informally in the preceding chapter. Thus, we consider the preceding chapter as motivation for what follows.

Definition. A *partition of* $[a, b]$ when $a < b$ is a finite subset P of $[a, b]$ that can be indexed so that $P = \{x_0, x_1, ..., x_n\}$, where

$$x_0 = a < x_1 < x_2 < \cdots < x_n = b, \text{ some } n \geq 1.$$

It is also to be understood that the interval $[a, a]$ has a (unique) partition, namely, $\{a\}$.

For example, $\{0, 1\}$ and $\{0, \frac{1}{3}, \frac{1}{2}, 1\}$ are partitions of $[0, 1]$. Obviously, every interval $[a, b]$ has a partition.

Whenever P is a partition and we write $P = \{x_0, x_1, ..., x_n\}$, we assume (without explicitly saying so) that the points x_i satisfy the condition in the definition above. We prove all results that involve partitions, directly or indirectly (as in the case of integrals), assuming that $a < b$. It will be evident that the results hold when $a = b$.

Definition. Let P_1 and P_2 be partitions of $[a, b]$. We say that P_2 *is a refinement of* P_1, written $P_2 \preceq P_1$, provided that $P_2 \supset P_1$.

We can think of a refinement of a partition P as being obtained from P by adding points to P (although, of course, a partition is a refinement of itself). Obviously, every partition of $[a, b]$ is a refinement of $\{a, b\}$.

Exercise 12.1: Give an example of two partitions of $[a, b]$ such that neither one is a refinement of the other.

A relation $\ll$ between elements of a set S is a *partial order on* S provided that the relation is reflexive ($s \ll s$ for all $s \in S$), antisymmetric (if $s_1 \ll s_2$ and $s_2 \ll s_1$, then $s_1 = s_2$), and transitive (if $s_1 \ll s_2$ and $s_2 \ll s_3$, then $s_1 \ll s_3$).

For example, $\leq$ is a partial order on R^1 by axioms O1 and O2 in section 1 of Chapter I.

Note the following simple fact:

Exercise 12.2: The relation $\preceq$ of refinement on the collection $\mathcal{P}$ of all partitions of a given interval $[a, b]$ is a partial order.

Definition. Let P_1 and P_2 be refinements of $[a, b]$. A *common refinement of P_1 and P_2* is a partition P of $[a, b]$ such that $P \preceq P_1$ and $P \preceq P_2$.

Exercise 12.3: For any two partitions P_1 and P_2 of $[a, b]$, there is a smallest common refinement of P_1 and P_2; that is, there is a common refinement, P, of P_1 and P_2 such that every common refinement of P_1 and P_2 contains P.

2. Upper and Lower Sums

We continue with our presentation of the background necessary for defining the integral and understanding the definition.

We adopt the following notation: Let $f : [a, b] \to \mathsf{R}^1$ be a bounded function, and let $P = \{x_0, x_1, ..., x_n\}$ be a partition of $[a, b]$. For each $i = 1, 2, ..., n$,

$$\Delta x_i = x_i - x_{i-1}, \quad M_i(f) = lub\, f([x_{i-1}, x_i]), \quad m_i(f) = glb\, f([x_{i-1}, x_i]).$$

Definition. Let $f : [a, b] \to \mathsf{R}^1$ be a bounded function, and let $P = \{x_0, x_1, ..., x_n\}$ be a partition of $[a, b]$.

- The *upper sum of f with respect to P*, denoted by $U_P(f)$, is defined by

$$U_P(f) = \Sigma_{i=1}^{n} M_i(f)\Delta x_i.$$

- The *lower sum of f with respect to P*, denoted by $L_P(f)$, is defined by

$$L_P(f) = \Sigma_{i=1}^{n} m_i(f)\Delta x_i.$$

Exercise 12.4: Define $f : [-4, 4] \to \mathsf{R}^1$ by $f(x) = x^3 - 12x$. Evaluate $U_P(f)$ and $L_P(f)$ for the partition $P = \{-4, 1, 4\}$.

Exercise 12.5: Define $f : [0, 4] \to \mathsf{R}^1$ by $f(x) = x^3 - 9x^2 + 26x - 24$. Evaluate $U_P(f)$ and $L_P(f)$ for the partition $P = \{0, 1, 3, 4\}$.

Lemma 12.6: Let $f : [a, b] \to \mathsf{R}^1$ be a bounded function. For any partition $P = \{x_0, x_1, ..., x_n\}$ of $[a, b]$, $L_P(f) \leq U_P(f)$.

Proof: For each i, $m_i(f) \leq M_i(f)$ and $\Delta x_i > 0$, hence $m_i(f)\Delta x_i \leq M_i(f)\Delta x_i$. Therefore, the lemma follows immediately by summing over i. ¥

Lemma 12.7: Let $f : [a, b] \to \mathsf{R}^1$ be a bounded function. Let P be a partition of $[a, b]$, and let q be a point of $[a, b]$ such that $q \notin P$. Let $Q = P \cup \{q\}$ (considered as a partition of $[a, b]$). Then

$$U_Q(f) \leq U_P(f) \quad \text{and} \quad L_Q(f) \geq L_P(f).$$

Proof: Assume that $P = \{x_0, x_1, ..., x_n\}$. Let k be such that $x_k < q < x_{k+1}$. Then, letting

$$\alpha = \Big(lub\, f([x_k, q])\Big)\,(q - x_k) + \Big(lub\, f([q, x_{k+1}])\Big)\,(x_{k+1} - q),$$

we have that $U_Q(f) = \Sigma_{i \neq k+1} M_i(f)\Delta x_i + \alpha$. Also, since $lub(A) \leq lub(B)$ when $A \subset B$,

$$\begin{aligned}
\alpha &= \Big(lub\, f([x_k, q])\Big)\,(q - x_k) + \Big(lub\, f([q, x_{k+1}])\Big)\,(x_{k+1} - q) \\
&\leq \Big(lub\, f([x_k, x_{k+1}])\Big)\,(q - x_k) + \Big(lub\, f([x_k, x_{k+1}])\Big)\,(x_{k+1} - q) \\
&= \Big(lub\, f([x_k, x_{k+1}])\Big)\,(x_{k+1} - x_k).
\end{aligned}$$

Therefore,

$$U_Q(f) = \Sigma_{i \neq k+1} M_i(f)\Delta x_i + \alpha \leq \Sigma_{i=1}^{n} M_i(f)\Delta x_i = U_P(f).$$

Similarly, $L_Q(f) \geq L_P(f)$. ¥

Lemma 12.8: Let $f : [a, b] \to \mathsf{R}^1$ be a bounded function, and let P_1 and P_2 be partitions of $[a, b]$ such that $P_2 \preceq P_1$. Then

$$U_{P_2}(f) \leq U_{P_1}(f) \quad \text{and} \quad L_{P_2}(f) \geq L_{P_1}(f).$$

Proof: Let $y_1, y_2, ..., y_m$ be the points in $P_2 - P_1$ (we assume that $P_1 \neq P_2$ since, otherwise, the lemma is obvious). We successively define partitions Q_j, $j = 1, ..., m$, of $[a, b]$ as follows:

$$Q_1 = P_1, \;\; Q_2 = Q_1 \cup \{y_1\}, \;\; Q_3 = Q_2 \cup \{y_2\}, \; ... \; , \;\; Q_m = P_2.$$

Since Q_{j+1} has exactly one more point than Q_j for each j, each successive inequality below follows at once from Lemma 12.7:

$$U_{P_2}(f) = U_{Q_m}(f) \leq U_{Q_{m-1}}(f) \leq \cdots \leq U_{Q_2}(f) \leq U_{Q_1}(f) = U_{P_1}(f)$$

and

$$L_{P_2}(f) = L_{Q_m}(f) \geq L_{Q_{m-1}}(f) \geq \cdots \geq L_{Q_2}(f) \geq L_{Q_1}(f) = L_{P_1}(f). \text{ ¥}$$

Lemma 12.9: Let $f : [a, b] \to \mathsf{R}^1$ be a bounded function, and let P_1 and P_2 be partitions of $[a, b]$. Then

$$L_{P_1}(f) \leq U_{P_2}(f).$$

Proof: Let P be a common refinement of P_1 and P_2 (see Exercise 12.3). Then

$$L_{P_1}(f) \overset{12.8}{\leq} L_P(f) \overset{12.6}{\leq} U_P(f) \overset{12.8}{\leq} U_{P_2}(f). \text{ ¥}$$

The numbers $lub_{P\in\mathcal{P}}L_P(f)$ and $glb_{P\in\mathcal{P}}U_P(f)$ in the next lemma are the basis for our definition of the integral in the next section.

Lemma 12.10: Let $f:[a,b]\to\mathsf{R}^1$ be a bounded function, and let $\mathcal{P}$ denote the collection of all partitions of $[a,b]$. Then $lub_{P\in\mathcal{P}}L_P(f)$ and $glb_{P\in\mathcal{P}}U_P(f)$ exist and

$$lub_{P\in\mathcal{P}}L_P(f)\le glb_{P\in\mathcal{P}}U_P(f).$$

Proof: There is a partition P_1 of $[a,b]$. By Lemma 12.9, $L_{P_1}(f)$ is a lower bound for the set of all upper sums of f with respect to all partitions of $[a,b]$. Hence, by the Greatest Lower Bound Axiom (section 8 of Chapter I), $glb_{P\in\mathcal{P}}U_P(f)$ exists, and

$$(*)\ L_{P_1}(f)\le glb_{P\in\mathcal{P}}U_P(f).$$

Note that we have proved (*) for any partition P_1 of $[a,b]$. Hence, $glb_{P\in\mathcal{P}}U_P(f)$ is an upper bound for the set of all lower sums of f with respect to all partitions of $[a,b]$. Therefore, by the Least Upper Bound Axiom (Completeness Axiom), $lub_{P\in\mathcal{P}}L_P(f)$ exists, and it is clear that

$$lub_{P\in\mathcal{P}}L_P(f)\le glb_{P\in\mathcal{P}}U_P(f).\ \text{¥}$$

Except for very simple functions, it is difficult to directly compute the numbers $lub_{P\in\mathcal{P}}L_P(f)$ and $glb_{P\in\mathcal{P}}U_P(f)$ in Lemma 12.10. For example, the reader might try to compute the numbers in Lemma 12.10 for the case when f is the function on $[0,1]$ defined by $f(x)=x$. In fact, computing the numbers $lub_{P\in\mathcal{P}}L_P(f)$ and $glb_{P\in\mathcal{P}}U_P(f)$ is actually evaluating integrals or showing integrals do not exist, as we will see from the definition of the integral (in the next section). Nevertheless, we can at this time compute the numbers in Lemma 12.10 for a few functions. We illustrate how to do this in the two examples below. In the first example, $lub_{P\in\mathcal{P}}L_P(f)=glb_{P\in\mathcal{P}}U_P(f)$; in the second example, $lub_{P\in\mathcal{P}}L_P(f)\ne glb_{P\in\mathcal{P}}U_P(f)$.

Example 12.11: Define $f:[0,2]\to\mathsf{R}^1$ by

$$f(x)=\begin{cases}1 & \text{, if } x\ne 1\\ 2 & \text{, if } x=1.\end{cases}$$

Let $\mathcal{P}$ denote the collection of all partitions of $[0,2]$. We show that

$$lub_{P\in\mathcal{P}}L_P(f)=glb_{P\in\mathcal{P}}U_P(f)=2.$$

Let $P=\{x_0,x_1,...,x_n\}$ be a partition of $[0,2]$. Note that each of the intervals $[x_{i-1},x_i]$ contains a point different from 1; hence, $m_i(f)=1$ for each i. Thus,

$$L_P(f)=\Sigma_{i=1}^{n}\Delta x_i=x_n-x_0=2-0=2.$$

Therefore, $lub_{P\in\mathcal{P}}L_P(f)=2$.

We now show that $glb_{P\in\mathcal{P}}U_P(f)=2$. Let $\epsilon>0$ such that $\epsilon<1$. Consider the following very simple partition Q of $[0,2]$:

$$Q = \{0, 1-\epsilon, 1+\epsilon, 2\}.$$

We compute $U_Q(f)$:

$$U_Q(f) = 1([1-\epsilon]-0) + 2([1+\epsilon]-[1-\epsilon]) + 1(2-[1+\epsilon]) = 2+2\epsilon.$$

Thus, since ϵ can be as close to zero as we like, we have proved that

$$glb_{P\in\mathcal{P}} U_P(f) \leq 2.$$

Also, having proved above that $lub_{P\in\mathcal{P}} L_P(f) = 2$, we know from Lemma 12.10 that $2 \leq glb_{P\in\mathcal{P}} U_P(f)$. Therefore,

$$glb_{P\in\mathcal{P}} U_P(f) = 2 = lub_{P\in\mathcal{P}} L_P(f).$$

Example 12.12: Define $f : [0,1] \to \mathsf{R}^1$ by

$$f(x) = \begin{cases} 0 & \text{, if } x \text{ is rational} \\ 1 & \text{, if } x \text{ is irrational.} \end{cases}$$

Let $\mathcal{P}$ denote the collection of all partitions of $[0,1]$. We show that

$$lub_{P\in\mathcal{P}} L_P(f) = 0 \quad \text{and} \quad glb_{P\in\mathcal{P}} U_P(f) = 1.$$

Let $P = \{x_0, x_1, ..., x_n\}$ be a partition of $[0,1]$. By Theorem 1.26 (and its analogue for irrational numbers), there is a rational number and an irrational number in each of the intervals $[x_{i-1}, x_i]$. Hence,

$$L_P(f) = \Sigma_{i=1}^n (0)\Delta x_i = 0$$

and

$$\begin{aligned} U_P(f) &= \Sigma_{i=1}^n (1)\Delta x_i = (x_1 - x_0) + (x_2 - x_1) + \cdots + (x_n - x_{n-1}) \\ &= x_n - x_0 = 1 - 0 = 1. \end{aligned}$$

Therefore, $lub_{P\in\mathcal{P}} L_P(f) = 0$ and $glb_{P\in\mathcal{P}} U_P(f) = 1$.

The cancellation that gave $\Sigma_{i=1}^n \Delta x_i = x_n - x_0$ in Example 12.12 is trivial but has far-reaching generalizations in multi-dimensional calculus (for example, in the proof of Green's Theorem).

Exercise 12.13: Let f be a constant function on an interval $[a,b]$, say $f(x) = c$ for all $x \in [a,b]$. Compute $lub_{P\in\mathcal{P}} L_P(f)$ and $glb_{P\in\mathcal{P}} U_P(f)$.

Exercise 12.14: Define $f : [0,2] \to \mathsf{R}^1$ by

$$f(x) = \begin{cases} 1 & \text{, if } 0 \leq x < 1 \\ 3 & \text{, if } 1 \leq x \leq 2. \end{cases}$$

Compute $lub_{P\in\mathcal{P}} L_P(f)$ and $glb_{P\in\mathcal{P}} U_P(f)$.

3. Definition of the Integral

We are ready to define the integral.

Definition. Let $f : [a, b] \to \mathsf{R}^1$ be a bounded function, and let $\mathcal{P}$ denote the collection of all partitions of $[a, b]$. Recall that we showed in Lemma 12.10 that the numbers $glb_{P\in\mathcal{P}}U_P(f)$ and $lub_{P\in\mathcal{P}}L_P(f)$ exist.

- The *upper integral of f over $[a, b]$* is $glb_{P\in\mathcal{P}}U_P(f)$, which we denote from now on by $\overline{\int}_a^b f$.
- The *lower integral of f over $[a, b]$* is $lub_{P\in\mathcal{P}}L_P(f)$, which we denote from now on by $\underline{\int}_a^b f$.
- We say that f *is integrable over* $[a, b]$ provided that $\overline{\int}_a^b f = \underline{\int}_a^b f$, in which case we call the common value $\overline{\int}_a^b f = \underline{\int}_a^b f$ *the integral of f over $[a, b]$* (or *the integral of f from a to b*). We denote the integral of f over $[a, b]$ by $\int_a^b f$ or by $\int_a^b f(x)dx$. The notation $\int_a^b f(x)dx$ is read *integral of f over $[a, b]$ with respect to the variable x.*[7]

In the expressions $\overline{\int}_a^b f$, $\underline{\int}_a^b f$ and $\int_a^b f$, the numbers a and b are referred to as *the limits of integration* (a being *the lower limit of integration* and b being *the upper limit of integration*) The function f is called the *integrand.*

From what we showed in Example 12.11, we can now say that the function f in the example is integrable and $\int_0^2 f = 2$. On the other hand, from what we showed in Example 12.12, the function f in Example 12.12 is not integrable.

We prove results about integrals over $[a, b]$ as though $a < b$ without saying so. The reader can easily check that the results are true when $a = b$ ($\int_a^a f = 0$ since $\{a\}$ is the only partition of the interval $[a, a]$).

4. Two Theorems about Integrability

We prove two theorems about integrability and show how the theorems can be applied.

Our first theorem is useful for proving that a function is integrable; we illustrate this for a specific function after we prove the theorem. We use the theorem in the next section to prove that all continuous functions are integrable, and we use the theorem in many other places as well.

[7] Regarding the notation $\int_a^b f(x)dx$, the symbol dx has absolutely no mathematical content other than to indicate the variable with respect to which the integration is being performed. Thus, the symbol dx can be used to clarify situations when the expression being integrated contains two or more letters as symbols; for example, simply writing $\int_a^b t^2x^3$ puts in doubt whether we are integrating with respect to t or with respect to x, whereas writing $\int_a^b t^2x^3dx$ and $\int_a^b t^2x^3dt$ makes it clear what the variable of integration is in each case.

Theorem 12.15: Let $f : [a,b] \to \mathsf{R}^1$ be a bounded function. Then f is integrable over $[a,b]$ if and only if for each $\epsilon > 0$, there is a partition P of $[a,b]$ such that

$$U_P(f) - L_P(f) < \epsilon.$$

Proof: Assume that f is integrable over $[a,b]$. Let $\epsilon > 0$. Since

$$\int_a^b f = \overline{\int}_a^b f = glb_{P\in\mathcal{P}} U_P(f) \quad \text{and} \quad \int_a^b f = \underline{\int}_a^b f = lub_{P\in\mathcal{P}} L_P(f),$$

there are a partitions P_1 and P_2 of $[a,b]$ such that

(1) $U_{P_1}(f) < \int_a^b f + \frac{\epsilon}{2}$ and $L_{P_2}(f) > \int_a^b f - \frac{\epsilon}{2}$.

Let P be a common refinement of P_1 and P_2 (see Exercise 12.3). Then, by Lemma 12.6 and Lemma 12.8 , we have

(2) $L_{P_2}(f) \le L_P(f) \le U_P(f) \le U_{P_1}(f)$.

Now,

$$U_P(f) - L_P(f) \overset{(2)}{\le} U_{P_1}(f) - L_{P_2}(f) \overset{(1)}{<} \int_a^b f + \tfrac{\epsilon}{2} - (\int_a^b f - \tfrac{\epsilon}{2}) = \epsilon.$$

This proves that P is as required in the theorem.

Conversely, assume that for each $\epsilon > 0$, there is a partition P_ϵ of $[a,b]$ such that

$$U_{P_\epsilon}(f) - L_{P_\epsilon}(f) < \epsilon.$$

Then, since $\overline{\int}_a^b f = glb_{P\in\mathcal{P}} U_P(f)$ and $\underline{\int}_a^b f = lub_{P\in\mathcal{P}} L_P(f)$,

$$0 \overset{12.10}{\le} \overline{\int}_a^b f - \underline{\int}_a^b f \le U_{P_\epsilon}(f) - L_{P_\epsilon}(f) < \epsilon \text{ for all } \epsilon > 0.$$

Hence, $\overline{\int}_a^b f - \underline{\int}_a^b f = 0$ (it follows from the axioms in section 1 of Chapter I that if $0 \le x < \epsilon$ for all $\epsilon > 0$, then $x = 0$). Therefore, $\overline{\int}_a^b f = \underline{\int}_a^b f$, which proves that f is integrable. ¥

Lest it escape us without notice, we point out that Theorem 12.15 says that we need only find *one* appropriate partition for each $\epsilon > 0$ in order to show a function is integrable. This feature of Theorem 12.15 makes it significantly easier to show a function is integrable than it would be to show the function is integrable using the definition of integrability directly. We illustrate this with the following example:

Example 12.16: Define $f : [0,2] \to \mathsf{R}^1$ by $f(x) = x^2$. We show that f is integrable over $[0,2]$ by applying Theorem 12.15.

Let $\epsilon > 0$. Let n be a natural number such that $\frac{4}{n} < \epsilon$ (the number n exists by the Archimedean Property (Theorem 1.22)). Let P be the partition of $[0,2]$ given by

$$P = \{x_0 = 0, x_1 = \tfrac{1}{n}, ..., x_i = \tfrac{i}{n}, ..., x_{2n} = 2\}.$$

Note that f is strictly increasing (by Theorem 10.17 since $f'(x) = 2x > 0$ for all $x \in [0, 2]$). Hence,

$$M_i(f) = x_i^2, \;\; m_i(f) = x_{i-1}^2, \;\; \text{each } i = 1, 2, ..., 2n.$$

Thus, since $\Delta x_i = \frac{1}{n}$ for each i,

$$U_P(f) - L_P(f) = \Sigma_{i=1}^{2n} x_i^2 \tfrac{1}{n} - \Sigma_{i=1}^{2n} x_{i-1}^2 \tfrac{1}{n} = \tfrac{1}{n}(\Sigma_{i=1}^{2n} x_i^2 - \Sigma_{i=1}^{2n} x_{i-1}^2)$$
$$= \tfrac{1}{n}(x_{2n}^2 - x_0^2) = \tfrac{1}{n}(4 - 0) < \epsilon.$$

Therefore, by Theorem 12.15, f is integrable over $[0, 2]$.

Note that we did not evaluate the integral in Example 12.16 – Theorem 12.15 is not set up to evaluate integrals. Our next theorem gives a condition that can be used to evaluate integrals (in practice, however, the theorem has very limited use for this purpose). After we prove the theorem, we apply the theorem to evaluate the integral in the example above.

We note that the limits in the following theorem are limits of sequences, which we discussed in section 8 of Chapter IV.

Theorem 12.17: Let $f : [a, b] \to \mathsf{R}^1$ be a bounded function. Assume that $P_1, P_2, ..., P_n, ...$ are partitions of $[a, b]$ such that

$$\lim_{n\to\infty} U_{P_n}(f) = \lim_{n\to\infty} L_{P_n}(f) = c.$$

Then f is integrable over $[a, b]$ and $\int_a^b f = c$.

Proof: By definition, $\underline{\int}_a^b f = lub_{P\in\mathcal{P}} L_P(f)$ and $\overline{\int}_a^b f = glb_{P\in\mathcal{P}} U_P(f)$; hence,

$$L_{P_n}(f) \le \underline{\int}_a^b f \overset{12.10}{\le} \overline{\int}_a^b f \le U_{P_n}(f), \;\; \text{all } n = 1, 2, \ldots .$$

Thus, by the Squeeze Theorem (Theorem 4.34), which holds for sequences by Theorem 4.38, we have that

$$\underline{\int}_a^b f = c \;\; \text{and} \;\; \overline{\int}_a^b f = c.$$

Therefore, f is integrable and $\int_a^b f = c$. ¥

Example 12.18: We use Theorem 12.17 to evaluate the integral of the function in Example 12.16; we show that $\int_0^2 x^2 = \frac{8}{3}$.

We use following formula; the formula can be verified by induction (we leave the verification for the reader in Exercise 12.19):

$$(*)\; \Sigma_{i=1}^n i^2 = \tfrac{n(n+1)(2n+1)}{6} \text{ for each } n = 1, 2, \ldots .$$

For each $n = 1, 2, ...$, let P_n be the partition of $[0, 2]$ given by

$$P_n = \{x_0 = 0, x_1 = \tfrac{1}{n}, ..., x_i = \tfrac{i}{n}, ..., x_{2n} = 2\}.$$

Then, since $M_i(f) = x_i^2$ and $m_i(f) = x_{i-1}^2$ for each i (as in Example 12.16),

$$U_{P_n}(f) = \Sigma_{i=1}^{2n} x_i^2 \tfrac{1}{n} \quad \text{and} \quad L_{P_n}(f) = \Sigma_{i=1}^{2n} x_{i-1}^2 \tfrac{1}{n} \quad \text{for each } n.$$

Hence, for each n,

$$\begin{aligned} U_{P_n}(f) &= \tfrac{1}{n}\Sigma_{i=1}^{2n}(\tfrac{i}{n})^2 = \tfrac{1}{n^3}\Sigma_{i=1}^{2n} i^2 \overset{(*)}{=} \tfrac{1}{n^3}\tfrac{2n(2n+1)(4n+1)}{6} \\ &= \tfrac{(2n+1)(4n+1)}{3n^2} = \tfrac{8}{3} + \tfrac{2}{n} + \tfrac{1}{3n^2} \end{aligned}$$

and

$$\begin{aligned} L_{P_n}(f) &= \Sigma_{i=1}^{2n}(\tfrac{i-1}{n})^2 \tfrac{1}{n} = \tfrac{1}{n^3}\Sigma_{i=1}^{2n}(i-1)^2 = \tfrac{1}{n^3}\Sigma_{i=1}^{2n-1} i^2 \\ &\overset{(*)}{=} \tfrac{1}{n^3}\tfrac{(2n-1)(2n)(4n-1)}{6} = \tfrac{(2n-1)(4n-1)}{3n^2} = \tfrac{8}{3} - \tfrac{2}{n} + \tfrac{1}{3n^2}\,. \end{aligned}$$

Thus, $\lim_{n\to\infty} U_{P_n}(f) = \frac{8}{3}$ and $\lim_{n\to\infty} L_{P_n}(f) = \frac{8}{3}$. Therefore, by Theorem 12.17, $\int_0^2 x^2 = \frac{8}{3}$.

Exercise 12.19: Verify that $\Sigma_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$ for each $n = 1, 2, ...$ by using induction (Theorem 1.20). (We used the formula in Example 12.18.)

In Examples 12.16 and 12.18, we used partitions that divide the interval of integration into intervals of equal length. These types of partitions are useful because we can factor Δx_i out of summations when computing upper and lower sums. We call a partition of an interval $[a, b]$ that divides $[a, b]$ into intervals of equal length Δx_i a *regular partition*.

Exercise 12.20: Evaluate $\int_a^b x$ for any $a \le b$.
(*Hint:* First prove that $\Sigma_{i=1}^{n} i = \frac{n(n+1)}{2}$ for each $n = 1, 2, ...$.)

Exercise 12.21: Determine if f is integrable, where $f : [0, 1] \to \mathsf{R}^1$ is defined as follows (Q denotes the set of all rational numbers; for integers m and n, $\frac{m}{n}$ *in lowest terms* means m and n have no common divisor other than ± 1):

$$f(x) = \begin{cases} 0 & \text{, if } x \text{ is irrational} \\ 1 & \text{, if } x = 0 \\ \frac{1}{n} & \text{, if } x \in \mathsf{Q} - \{0\} \text{ and } x = \frac{m}{n} \text{ in lowest terms.} \end{cases}$$

Exercise 12.22: Assume that $f(x) \le g(x) \le h(x)$ for all $x \in [a, b]$ and that f and h are integrable over $[a, b]$. If $\int_a^b f = \int_a^b h$, then g is integrable and $\int_a^b g$ is equal to $\int_a^b f = \int_a^b h$.

Exercise 12.23: If f is increasing on $[a, b]$ or decreasing on $[a, b]$, then f is integrable over $[a, b]$.

Exercise 12.24: In connection with Exercise 12.23, is every one-to-one bounded function on an interval $[a, b]$ integrable over $[a, b]$?

Exercise 12.25: If $f : [a, b] \to \mathsf{R}^1$ is a nonnegative function that is integrable over $[a, b]$, then $\int_a^b f \geq 0$.

Exercise 12.26: Let $f : [a, b] \to \mathsf{R}^1$ be a nonnegative function that is integrable over $[a, b]$. Then $\int_a^b f = 0$ if and only if $glbf(I) = 0$ for each open interval I in $[a, b]$.

Exercise 12.27: Let $f : [a, b] \to \mathsf{R}^1$ be a function that is integrable over $[a, b]$, and let $g : [a, b] \to \mathsf{R}^1$ be a function that agrees with f except at finitely many points. Is g integrable over $[a, b]$?

5. Continuous Functions Are Integrable

We prove that any continuous function defined on a closed and bounded interval is integrable. This is an existence theorem – it does not show how to evaluate the integral. We will be able to evaluate integrals of many simple continuous functions using the Fundamental Theorem of Calculus, which we prove in Chapter XIV. However, evaluating integrals of most continuous functions is difficult, usually impossible; ad hoc methods can sometimes be employed, but most often one has to settle for approximate evaluations by numerical methods.

The following notion is of general importance and is the key idea that we use to prove our theorem:

Definition: Let $X \subset \mathsf{R}^1$, and let $f : X \to \mathsf{R}^1$ be a function. We say that f is *uniformly continuous on* X provided that for any $\epsilon > 0$, there is a $\delta > 0$ such that if $x_1, x_2 \in X$ and $|x_1 - x_2| < \delta$, then $|f(x_1) - f(x_2)| < \epsilon$.

Exercise 12.28: Let $X \subset \mathsf{R}^1$. If $f : X \to \mathsf{R}^1$ is uniformly continuous, then f is continuous.

Exercise 12.29: The converse of the result in Exercise 12.28 is false: The function $f : \mathsf{R}^1 \to \mathsf{R}^1$ given by $f(x) = x^2$ is continuous but not uniformly continuous.

Exercise 12.30: Any linear function f (i.e., $f(x) = mx + b$) is uniformly continuous on R^1. More generally, if f is differentiable on an interval I and the derivative f' is bounded on I, then f is uniformly continuous on I.

The following theorem is not concerned with integrals, but it is the basis of our proof that continuous functions are integrable. The theorem is so important in all of mathematics that even though it plays the role of a lemma here, we can not bring ourselves to call the theorem a lemma. The theorem shows that the converse of the result in Exercise 12.28 is true when X is a closed and bounded interval.

Theorem 12.31: If $f : [a, b] \to \mathsf{R}^1$ is continuous, then f is uniformly continuous.

Proof: Suppose by way of contradiction that f is not uniformly continuous. Then, for some $\epsilon > 0$, there are points $x_n, y_n \in [a, b]$ for each $n \in \mathsf{N}$ such that

(1) $|x_n - y_n| < \frac{1}{n}$ for each $n \in \mathsf{N}$

and

(2) $|f(x_n) - f(y_n)| \geq \epsilon$ for each $n \in \mathsf{N}$.

Let

$$X = \{x_n : n \in \mathsf{N}\}.$$

We prove that X is an infinite set. Suppose that the set X is finite. Then there is a point $q \in X$ such that $q = x_n$ for infinitely many n. Hence, we can assume that

(3) $q = x_{n_i}$, where $n_i < n_{i+1}$ for each $i \in \mathsf{N}$.

Then, by (3) and (1), we have

(4) $|q - y_{n_i}| < \frac{1}{n_i}$ for each $i \in \mathsf{N}$

and, by (3) and (2), we have

(5) $|f(q) - f(y_{n_i})| \geq \epsilon$ for each $i \in \mathsf{N}$.

Let

$$Y = \{y_{n_i} : i \in \mathsf{N}\}.$$

Recall from (3) that $n_i < n_{i+1}$ for each $i \in \mathsf{N}$; hence, by (4) and the second part of Exercise 1.23, the sequence $\{y_{n_i}\}_{i=1}^{\infty}$ converges to q. Thus, $q \sim Y$ (definition in section 1 of Chapter II). However, by (5), $f(q) \not\sim f(Y)$. Hence, f is not continuous at q by the definition of continuity (section 3 of Chapter II). This contradicts the assumption in our theorem. Therefore, we have proved that X is an infinite set.

Now, since X is a bounded infinite set, X has a limit point p in R^1 (by Exercise 5.16); furthermore, since $X \subset [a, b]$ and $p \sim X$, it follows easily that $p \in [a, b]$ (if $p \notin [a, b]$, then $p \not\sim [a, b]$ by Theorem 2.5; hence, $p \not\sim X$ by Exercise 2.10, a contradiction).

Since $p \in [a, b]$, f is continuous at p. Hence, by Theorem 3.12, there is a $\delta > 0$ such that

(6) $|f(x) - f(p)| < \frac{\epsilon}{2}$ whenever $x \in [a, b]$ and $|x - p| < \delta$.

Since p is a limit point of X, we have by Exercise 2.33 that

$$|x_n - p| < \tfrac{\delta}{2} \text{ for infinitely many } n.$$

Hence, by the Archimedean Property (Theorem 1.22), there is a natural number k such that $\frac{1}{k} < \frac{\delta}{2}$ and $|x_k - p| < \frac{\delta}{2}$. Thus,

$$|y_k - p| \leq |y_k - x_k| + |x_k - p| \overset{(1)}{<} \tfrac{1}{k} + \tfrac{\delta}{2} < \delta.$$

Since $|x_k - p| < \frac{\delta}{2} < \delta$ and $|y_k - p| < \delta$,

$$|f(x_k) - f(y_k)| \leq |f(x_k) - f(p)| + |f(p) - f(y_k)| \overset{(6)}{<} \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = \epsilon.$$

This contradicts (2). ¥

We note the following terminology, which calls attention to an important idea in connection with integrals.

Definition: The *norm of a partition* $P = \{x_0, x_1, ..., x_n\}$ of $[a, b]$, which we denote by $\|P\|$, is defined by

$$\|P\| = \max\{\Delta x_i : i = 1, 2, ..., n\}.$$

Exercise 12.32: For any $\eta > 0$, there is a partition P of $[a, b]$ such that $\|P\| < \eta$.

We now prove our theorem.

Theorem 12.33: If $f : [a, b] \to \mathsf{R}^1$ is a continuous function, then f is integrable over $[a, b]$.

Proof: We will apply Theorem 12.15. Let $\epsilon > 0$. By Theorem 12.31, f is uniformly continuous. Hence, there is a $\delta > 0$ such that

(1) $|f(y) - f(z)| < \frac{\epsilon}{b-a}$ whenever $y, z \in [a, b]$ and $|y - z| < \delta$.

Let $P = \{x_0, x_1, ..., x_n\}$ be a partition of $[a, b]$ such that $\|P\| < \delta$ (P exists by Exercise 12.32). For each $i = 1, 2, ..., n$, $f|[x_{i-1}, x_i]$ is continuous (by Exercise 5.3); hence, by the Maximum - Minimum Theorem (Theorem 5.13), $f|[x_{i-1}, x_i]$ attains its maximum value at a point $y_i \in [x_{i-1}, x_i]$ and its minimum value at a point $z_i \in [x_{i-1}, x_i]$; in other words, we have that

(2) $M_i(f) = f(y_i)$ and $m_i(f) = f(z_i)$ for each $i = 1, 2, ..., n$.

Now, we show that P satisfies the condtion in Theorem 12.15:

$$\begin{aligned} U_P(f) - L_P(f) &= \Sigma_{i=1}^n M_i(f)\Delta x_i - \Sigma_{i=1}^n m_i(f)\Delta x_i \\ &= \Sigma_{i=1}^n [M_i(f) - m_i(f)]\Delta x_i \overset{(2)}{=} \Sigma_{i=1}^n [f(y_i) - f(z_i)]\Delta x_i \\ &\overset{(1)}{<} \Sigma_{i=1}^n \tfrac{\epsilon}{b-a}\Delta x_i = \tfrac{\epsilon}{b-a}\Sigma_{i=1}^n \Delta x_i = \tfrac{\epsilon}{b-a}\Sigma_{i=1}^n (x_i - x_{i-1}) \\ &= \tfrac{\epsilon}{b-a}(x_n - x_0) = \tfrac{\epsilon}{b-a}(b - a) = \epsilon. \end{aligned}$$

Therefore, since $\epsilon > 0$ was arbitrary, we have by Theorem 12.15 that f is integrable over $[a, b]$. ¥

Exercise 12.34: If $f : [a, b] \to \mathsf{R}^1$ is a bounded function that is continuous at all but finitely many points, then f is integrable over $[a, b]$.

Chapter XIII: The Algebra of the Integral

We show that sums, differences, products, quotients (with a condition), and absolute values of integrable functions are integrable. Finally, we examine integrals over subintervals (which we discuss again in the first section of Chapter XVI).

We remark that most results in this chapter follow immediately from a characterization of integrability in Chapter XV. In addition, the best general theorem about quotients follows easily from the characterization in Chapter XV (we were unable to find a proof using the methods we use in the present chapter; compare Theorem 13.36 with the result in Exercise 15.34). Nevertheless, this chapter is important for two reasons: First, it is always a good idea to understand why theorems are true from the most basic point of view; second, several results we prove here are necessary for proving the Fundamental Theorem of Calculus, which we want to prove as soon as possible, and some of those results are not consequences of the characterization in Chapter XV (e.g., the inequality in Theorem 13.17 and Theorem 13.40).

1. Integrability of Sums

We prove that the sum of two integrable functions is integrable and that the integral of the sum is the sum of the integrals (Theorem 13.3).

Lemma 13.1: Let f and g be bounded functions defined on a nonempty set X. Then

(1) $lub_{x\in X}(f(x)+g(x)) \le lub_{x\in X}f(x) + lub_{x\in X}g(x)$

and

(2) $glb_{x\in X}(f(x)+g(x)) \ge glb_{x\in X}f(x) + glb_{x\in X}g(x)$.

Proof: For any $y \in X$,

$$f(y)+g(y) \le lub_{x\in X}f(x) + lub_{x\in X}g(x);$$

hence, $lub_{x\in X}f(x) + lub_{x\in X}g(x)$ is an upper bound for $\{f(x)+g(x) : x \in X\}$. Therefore, by the Completeness Axiom, $lub_{x\in X}(f(x)+g(x))$ exists and, clearly,

$$lub_{x\in X}(f(x)+g(x)) \le lub_{x\in X}f(x) + lub_{x\in X}g(x),$$

which proves (1). The proof of (2) is similar. ¥

The inequalities in Lemma 13.1 are in general strict, as can be seen, for example, by taking $f(x) = x$ and $g(x) = -\frac{1}{2}x + \frac{1}{2}$ for all $x \in [0,1]$.

Lemma 13.2: Let $f, g : [a,b] \to \mathsf{R}^1$ be bounded functions, and let P be a partition of $[a,b]$. Then

(1) $U_P(f+g) \le U_P(f) + U_P(g)$

and

(2) $L_P(f+g) \ge L_P(f) + L_P(g)$.

Proof: Assume that $P = \{x_0, x_1, ..., x_n\}$. By Lemma 13.1,

(*) $M_i(f+g) \le M_i(f) + M_i(g)$ and $m_i(f+g) \ge m_i(f) + m_i(g)$, all i.

Therefore,

$$\begin{aligned} U_P(f+g) &= \Sigma_{i=1}^n M_i(f+g)\Delta x_i \overset{(*)}{\le} \Sigma_{i=1}^n [M_i(f) + M_i(g)]\Delta x_i \\ &= \Sigma_{i=1}^n M_i(f)\Delta x_i + \Sigma_{i=1}^n M_i(g)\Delta x_i = U_P(f) + U_P(g) \end{aligned}$$

and, similarly, $L_P(f+g) \ge L_P(f) + L_P(g)$. ¥

We now prove our theorem.

Theorem 13.3: If f and g are integrable over $[a,b]$, then $f+g$ is integrable over $[a,b]$ and

$$\int_a^b (f+g) = \int_a^b f + \int_a^b g.$$

Proof: Let $\epsilon > 0$. Then, by the definition of the integrals of f and g (section 3 of Chapter XII), there are partitions P_1 and P_2 of $[a,b]$ such that

$$U_{P_1}(f) < \int_a^b f + \tfrac{\epsilon}{2} \quad \text{and} \quad U_{P_2}(g) < \int_a^b g + \tfrac{\epsilon}{2}.$$

Let P be a common refinement of P_1 and P_2 (see Exercise 12.3). Then, by Lemma 12.8, $U_P(f) \le U_{P_1}(f)$ and $U_P(g) \le U_{P_2}(g)$. Hence,

$$U_P(f) < \int_a^b f + \tfrac{\epsilon}{2} \quad \text{and} \quad U_P(g) < \int_a^b g + \tfrac{\epsilon}{2}.$$

Thus, since $\overline{\int}_a^b (f+g) \le U_P(f+g) \overset{13.2}{\le} U_P(f) + U_P(g)$, we have proved that

(1) $\overline{\int}_a^b (f+g) < \left(\int_a^b f + \int_a^b g\right) + \epsilon$.

Similarly, there are partitions Q_1 and Q_2 of $[a,b]$ such that

$$L_{Q_1}(f) > \int_a^b f - \tfrac{\epsilon}{2} \quad \text{and} \quad L_{Q_2}(g) > \int_a^b g - \tfrac{\epsilon}{2},$$

and, for a common refinement, Q, of Q_1 and Q_2, Lemma 12.8 shows that $L_Q(f) \ge L_{Q_1}(f)$ and $L_Q(g) \ge L_{Q_2}(g)$; hence,

$$L_Q(f) > \int_a^b f - \tfrac{\epsilon}{2} \quad \text{and} \quad L_Q(g) > \int_a^b g - \tfrac{\epsilon}{2}.$$

Thus, since $\underline{\int}_a^b (f+g) \ge L_Q(f+g) \overset{13.2}{\ge} L_Q(f) + L_Q(g)$, we have proved that

(2) $\underline{\int}_a^b (f+g) > \left(\int_a^b f + \int_a^b g\right) - \epsilon$.

We now have that

$$\left(\int_a^b f + \int_a^b g\right) - \epsilon \overset{(2)}{<} \underline{\int}_a^b (f+g) \overset{12.10}{\le} \overline{\int}_a^b (f+g) \overset{(1)}{<} \left(\int_a^b f + \int_a^b g\right) + \epsilon.$$

Therefore, since ϵ was an arbitrary positive number, we see that

$$\underline{\int}_a^b (f+g) = \overline{\int}_a^b (f+g) = \int_a^b f + \int_a^b g.$$

Thus, $f+g$ is integrable (by the first equality) and $\int_a^b (f+g) = \int_a^b f + \int_a^b g$. ¥

Corollary 13.4: If each of finitely many functions $f_1, f_2, ..., f_n$ is integrable over $[a,b]$, then their sum is integrable over $[a,b]$ and

$$\int_a^b (f_1 + f_2 + ... + f_n) = \int_a^b f_1 + \int_a^b f_2 + \cdots + \int_a^b f_n.$$

Proof: Left as an exercise. ¥

Exercise 13.5: Prove Corollary 13.4.

In analogy with Theorem 13.3, the difference of two integrable functions is integrable and the integral of the difference is the difference of the integrals. We prove this result in the next section (Corollary 13.12).

Exercise 13.6: Define $f : [0,2] \to \mathsf{R}^1$ by

$$f(x) = \begin{cases} 2 & \text{, if } 0 \le x < 1 \\ 5 & \text{, if } x = 1 \\ 4 & \text{, if } 1 < x \le 2. \end{cases}$$

Using Example 12.11 and Exercise 12.14, evaluate $\int_0^2 f$.

2. Integrability of Scalar Products

A *scalar product of a function* $f : X \to \mathsf{R}^1$ is the function $\lambda f : X \to \mathsf{R}^1$ obtained by multiplying each value of f by a fixed real number λ; that is,

$$(\lambda f)(x) = \lambda f(x) \text{ for all } x \in X.$$

The term *scalar product* is from vector spaces, where it refers to the product of a vector by a field element. The terminology is, therefore, appropriate here since the set of all real-valued functions defined on a nonempty set X forms a vector space under pointwise addition of functions (the vectors) and scalar product as defined above.

We prove that a scalar product λf of an integrable function f on $[a,b]$ is integrable and that the expected formula holds:

$$\int_a^b \lambda f = \lambda \int_a^b f.$$

Combining this result with the result about sums in the preceding section (Theorem 13.3), we have that the set V of all integrable functions defined on $[a,b]$ is a vector space and that $\int_a^b$ is a linear transformation from V to the vector space R^1; in other words,

$$\int_a^b (\lambda_1 f + \lambda_2 g) = \lambda_1 \int_a^b f + \lambda_2 \int_a^b g, \text{ all } f, g \in V \text{ and } \lambda_1, \lambda_2 \in \mathsf{R}^1.$$

In connection with $\int_a^b$ being a linear transformation, note that Exercise 12.26 characterizes all the nonnegative integrable functions in the null space of $\int_a^b$.

We remark that our theorem about the integrability of scalar products is a special case of the theorem about products that we will prove in section 4.

For any subset X of R^1 and any real number λ, we let λX denote the set defined by

$$\lambda X = \{\lambda x : x \in X\}.$$

Lemma 13.7: Let X be a nonempty bounded subset of R^1.

(1) If $\lambda \geq 0$, then $lub\lambda X = \lambda lubX$ and $glb\lambda X = \lambda glbX$.

(2) If $\lambda < 0$, then $lub\lambda X = \lambda glbX$ and $glb\lambda X = \lambda lubX$.

Proof: We prove part (1). Since (1) is trivial when $\lambda = 0$, we assume that $\lambda > 0$.

Since $x \leq lubX$ for all $x \in X$ and since $\lambda > 0$, it is clear that $\lambda x \leq \lambda lubX$ for all $x \in X$. Hence, $\lambda lubX$ is an upper bound for λX; thus, by the Completeness Axiom, $lub\lambda X$ exists and, obviously,

(a) $lub\lambda X \leq \lambda lubX$.

Since $\lambda x \leq lub\lambda X$ for all $x \in X$ and since $\lambda > 0$, we have that $x \leq \frac{1}{\lambda} lub\lambda X$ for all $x \in X$. Hence, $lubX \leq \frac{1}{\lambda} lub\lambda X$; thus, since $\lambda > 0$, we have

(b) $\lambda lubX \leq lub\lambda X$.

By (a) and (b), $lub\lambda X = \lambda lubX$. Similarly, by replacing lub with glb and reversing inequalities in the argument above, we obtain that $glb\lambda X = \lambda glbX$. This proves part (1).

We leave the proof of part (2) as an exercise. ¥

Exercise 13.8: Prove part (2) of Lemma 13.7.

Lemma 13.9: Let $f : [a, b] \to \mathsf{R}^1$ be a bounded function, and let P be a partition of $[a, b]$.

(1) For any $\lambda \geq 0$, $U_P(\lambda f) = \lambda U_P(f)$ and $L_P(\lambda f) = \lambda L_P(f)$.

(2) For any $\lambda < 0$, $U_P(\lambda f) = \lambda L_P(f)$ and $L_P(\lambda f) = \lambda U_P(f)$.

Proof: Assume that $P = \{x_0, x_1, ..., x_n\}$.

We prove part (1). Let $\lambda \geq 0$. Then, by the first part of Lemma 13.7,

$$(*)\ M_i(\lambda f) = \lambda M_i(f) \quad \text{and} \quad m_i(\lambda f) = \lambda m_i(f), \quad \text{all } i.$$

Therefore,

$$U_P(\lambda f) = \Sigma_{i=1}^n M_i(\lambda f)\Delta x_i \overset{(*)}{=} \Sigma_{i=1}^n \lambda M_i(f)\Delta x_i = \lambda U_P(f)$$

and

$$L_P(\lambda f) = \Sigma_{i=1}^n m_i(\lambda f)\Delta x_i \overset{(*)}{=} \Sigma_{i=1}^n \lambda m_i(f)\Delta x_i = \lambda L_P(f).$$

This proves part (1).

We leave the proof of part (2) as an exercise. ¥

Exercise 13.10: Prove part (2) of Lemma 13.9.

Theorem 13.11: Let $\lambda \in \mathsf{R}^1$. If f is integrable over $[a,b]$, then λf is integrable over $[a,b]$ and

$$\int_a^b \lambda f = \lambda \int_a^b f.$$

Proof: Let $\mathcal{P}$ denote the collection of all partitions of $[a,b]$. By the definitions of the upper and lower integrals (section 3 of Chapter XII),

$$(1)\ \overline{\int}_a^b \lambda f = glb_{P\in\mathcal{P}} U_P(\lambda f),\ \ \lambda\overline{\int}_a^b f = \lambda glb_{P\in\mathcal{P}} U_P(f)$$

and

$$(2)\ \underline{\int}_a^b \lambda f = lub_{P\in\mathcal{P}} L_P(\lambda f),\ \ \lambda\underline{\int}_a^b f = \lambda lub_{P\in\mathcal{P}} L_P(f)$$

We use the first part of Lemma 13.9 and then the first part of Lemma 13.7 in each of the following:

$$(3)\ glb_{P\in\mathcal{P}} U_P(\lambda f) = glb_{P\in\mathcal{P}} \lambda U_P(f) = \lambda glb_{P\in\mathcal{P}} U_P(f),\ \text{ if } \lambda \geq 0;$$

$$(4)\ lub_{P\in\mathcal{P}} L_P(\lambda f) = lub_{P\in\mathcal{P}} \lambda L_P(f) = \lambda lub_{P\in\mathcal{P}} L_P(f),\ \text{ if } \lambda \geq 0.$$

We use the second part of Lemma 13.9 and then the second part of Lemma 13.7 in each of the following:

$$(5)\ glb_{P\in\mathcal{P}} U_P(\lambda f) = glb_{P\in\mathcal{P}} \lambda L_P(f) = \lambda lub_{P\in\mathcal{P}} L_P(f),\ \text{ if } \lambda < 0;$$

$$(6)\ lub_{P\in\mathcal{P}} L_P(\lambda f) = lub_{P\in\mathcal{P}} \lambda U_P(f) = \lambda glb_{P\in\mathcal{P}} U_P(f),\ \text{ if } \lambda < 0.$$

Now, assume that $\lambda \geq 0$. Then, using that f is integrable over $[a,b]$ for the last equalities below,

$$\overline{\int}_a^b \lambda f \overset{(1)}{=} glb_{P\in\mathcal{P}} U_P(\lambda f) \overset{(3)}{=} \lambda glb_{P\in\mathcal{P}} U_P(f) \overset{(1)}{=} \lambda\overline{\int}_a^b f = \lambda \int_a^b f$$

and

$$\underline{\int}_a^b \lambda f \overset{(2)}{=} lub_{P\in\mathcal{P}} L_P(\lambda f) \overset{(4)}{=} \lambda lub_{P\in\mathcal{P}} L_P(f) \overset{(2)}{=} \lambda\underline{\int}_a^b f = \lambda \int_a^b f;$$

therefore, $\overline{\int}_a^b \lambda f = \lambda \int_a^b f = \underline{\int}_a^b \lambda f$, which proves the lemma when $\lambda \geq 0$.

Finally, assume that $\lambda < 0$. Then, using that f is integrable over $[a,b]$ for the last equalities below,

$$\overline{\int}_a^b \lambda f \overset{(1)}{=} glb_{P\in\mathcal{P}} U_P(\lambda f) \overset{(5)}{=} \lambda lub_{P\in\mathcal{P}} L_P(f) \overset{(2)}{=} \lambda\underline{\int}_a^b f = \lambda \int_a^b f$$

and

$$\underline{\int}_a^b \lambda f \overset{(2)}{=} lub_{P\in\mathcal{P}} L_P(\lambda f) \overset{(6)}{=} \lambda glb_{P\in\mathcal{P}} U_P(f) \overset{(1)}{=} \lambda\overline{\int}_a^b f = \lambda \int_a^b f;$$

therefore, $\overline{\int}_a^b \lambda f = \lambda \int_a^b f = \underline{\int}_a^b \lambda f$, which proves the lemma when $\lambda < 0$. ¥

We can now easily obtain the theorem for the difference of two integrable functions that is analogous to the theorem for the sum of two integrable functions in the preceding section (Theorem 13.3):

Corollary 13.12: If f and g are integrable over $[a, b]$, then $f-g$ is integrable over $[a, b]$ and

$$\int_a^b (f - g) = \int_a^b f - \int_a^b g.$$

Proof: By Theorem 13.11, $-g$ is integrable over $[a, b]$ and $\int_a^b -g = -\int_a^b g$. Therefore, since $f - g = f + (-g)$, we have by Theorem 13.3 that $f - g$ is integrable over $[a, b]$ and

$$\int_a^b (f - g) = \int_a^b (f + (-g)) \overset{13.3}{=} \int_a^b f + \int_a^b -g \overset{13.11}{=} \int_a^b f - \int_a^b g. \quad ¥$$

Exercise 13.13: Using Exercise 12.13, Example 12.18 and Exercise 12.20, evaluate $\int_0^2 (5x^2 - 3x + 4)$.

3. Integrability of Absolute Values

The *absolute value of a function* $f : X \to \mathsf{R}^1$ is the function $|f| : X \to \mathsf{R}^1$ defined by

$$|f|(x) = |f(x)| \text{ for all } x \in X.$$

We prove that the absolute value $|f|$ of an integrable function f on $[a, b]$ is integrable and that

$$\left| \int_a^b f \right| \leq \int_a^b |f|.$$

The inequality is what we would expect in view of the definition of upper and lower sums and the Triangle Inequality for absolute values (above Exercise 1.28); however, our proof of the inequality is along different lines.

Lemma 13.14: Assume that f and g are integrable over $[a, b]$ and that $f(x) \leq g(x)$ for all $x \in [a, b]$. Then

$$\int_a^b f \leq \int_a^b g.$$

Proof: Let $h = g - f$. Then, by Corollary 13.12, h is integrable and

$$\int_a^b h = \int_a^b g - \int_a^b f.$$

Also, since $h(x) \geq 0$ for all $x \in [a, b]$, $\int_a^b h \geq 0$ (by Exercise 12.25). Therefore,

$$\int_a^b g - \int_a^b f \geq 0,$$

which proves the lemma. ¥

Exercise 13.15: If f is integrable over $[a, b]$ and $\alpha \leq f(x) \leq \beta$ for all $x \in [a, b]$, then

$$\alpha(b - a) \leq \int_a^b f \leq \beta(b - a).$$

Lemma 13.16: Let X be a nonempty set, and let $f : X \to \mathsf{R}^1$ be a bounded function. Then

$$lub_{x\in X}\,|f(x)| - glb_{x\in X}\,|f(x)| \le lub_{x\in X} f(x) - glb_{x\in X} f(x).$$

Proof: Let $p, q \in X$. Then

$$\begin{aligned}|f(p)| - |f(q)| &\overset{1.29}{\le} |f(p) - f(q)| = \max\{f(p), f(q)\} - \min\{f(p), f(q)\} \\ &\le lub_{x\in X} f(x) - glb_{x\in X} f(x),\end{aligned}$$

which says

$$|f(p)| \le lub_{x\in X} f(x) - glb_{x\in X} f(x) + |f(q)|.$$

Since this inequality holds for all points $p \in X$, we have that

$$lub_{x\in X}\,|f(x)| \le lub_{x\in X} f(x) - glb_{x\in X} f(x) + |f(q)|.$$

Hence,

$$lub_{x\in X}\,|f(x)| - lub_{x\in X} f(x) + glb_{x\in X} f(x) \le |f(q)|.$$

Since this inequality holds for all points $q \in X$, we now have that

$$lub_{x\in X}\,|f(x)| - lub_{x\in X} f(x) + glb_{x\in X} f(x) \le glb_{x\in X}\,|f(x)|,$$

which proves the lemma. ¥

We now prove our theorem.

Theorem 13.17: If f is integrable over $[a, b]$, then $|f|$ is integrable over $[a, b]$ and

$$\left|\int_a^b f\right| \le \int_a^b |f|.$$

Proof: Let $\epsilon > 0$. Since f is integrable over $[a, b]$, we know from Theorem 12.15 that there is a partition $P = \{x_0, x_1, ..., x_n\}$ of $[a, b]$ such that

$$U_P(f) - L_P(f) < \epsilon.$$

Hence, using Lemma 13.16 on each of the intervals $[x_{i-1}, x_i]$ for the inequality is the third step below, we have

$$\begin{aligned}U_P(|f|) - L_P(|f|) &= \Sigma_{i=1}^n [M_i(|f|) - m_i(|f|)]\Delta x_i \\ &\le \Sigma_{i=1}^n [M_i(f) - m_i(f)]\Delta x_i = U_P(f) - L_P(f) < \epsilon.\end{aligned}$$

Therefore, since $\epsilon > 0$ was arbitrary, we have by Theorem 12.15 that $|f|$ is integrable over $[a, b]$.

Finally, we prove the inequality in the theorem. Having just proved that $|f|$ is integrable over $[a, b]$, we know from Theorem 13.11 that $-|f|$ is integrable over $[a, b]$. Therefore,

$$-\int_a^b |f| \overset{13.11}{=} \int_a^b -|f| \overset{13.14}{\leq} \int_a^b f \overset{13.14}{\leq} \int_a^b |f|,$$

which shows that $\left|\int_a^b f\right| \leq \int_a^b |f|$. ¥

Exercise 13.18: If $|f|$ is integrable over $[a, b]$, then is f integrable over $[a, b]$?

Exercise 13.19: If f and g are integrable over $[a, b]$, then the maximum function $f \bigvee g$ and the minimum function $f \bigwedge g$ are integrable over $[a, b]$. (We defined $f \bigvee g$ and $f \bigwedge g$ in Exercise 4.33.)

(*Hint:* The same as the hint for Exercise 4.33.)

Exercise 13.20: We defined the notion of a distance function for a set in Exercise 1.30. Let $\mathcal{I}([a, b])$ denote the set of all functions that are integrable over $[a, b]$. For any $f, g \in \mathcal{I}([a, b])$, let

$$d(f, g) = \int_a^b |f - g|.$$

Determine whether d is a metric for $\mathcal{I}([a, b])$.

4. Integrability of Products

We prove that the product of two integrable functions is integrable. It is not possible to give a formula for the integral of the product; in particular, the integral of the product of two integrable functions is not necessarily the product of the integrals (e.g., $\int_0^2 x^2 \neq (\int_0^2 x)(\int_0^2 x)$ by Example 12.18 and Exercise 12.20).

Our theorems about products (along with other results) can be used to show that all polynomials are integrable over any closed and bounded interval $[a, b]$, $a < b$ (Exercise 13.28).

We first prove that the product of two integrable functions is integrable for the case when the functions are nonnegative; we then apply Exercise 13.19 to obtain the general result (Theorem 13.26).

Note the following exercise for use in Lemma 13.22.

Exercise 13.21: Using axioms in section 1 of Chapter I, prove that if $0 \leq a \leq b$ and $0 \leq c \leq d$, then $ac \leq bd$.

Lemma 13.22: Let f and g be bounded nonnegative functions defined on a nonempty set X. Then

(1) $[glb_{x\in X} f(x)][glb_{x\in X} g(x)] \leq glb_{x\in X}(f \cdot g)(x)$

and

(2) $lub_{x\in X}(f \cdot g)(x) \leq [lub_{x\in X} f(x)][lub_{x\in X} g(x)]$.

Proof: By Exercise 13.21, we have

$$[glb_{x\in X} f(x)][glb_{x\in X} g(x)] \leq f(y)g(y), \text{ all } y \in X.$$

Hence, $[glb_{x\in X} f(x)][glb_{x\in X} g(x)]$ is a lower bound for $\{(f \cdot g)(x) : x \in X\}$. Therefore, $glb_{x\in X}(f \cdot g)(x)$ exists by the Greatest Lower Bound Axiom (section 8 of Chapter I) and

$$[glb_{x\in X} f(x)][glb_{x\in X} g(x)] \leq glb_{x\in X}(f\cdot g)(x).$$

This proves (1). The proof of (2) is similar. ¥

Lemma 13.23: Let $f, g : [a,b] \to \mathsf{R}^1$ be bounded nonnegative functions, and let s be an upper bound for both $f([a,b])$ and $g([a,b])$. If P is a partition of $[a,b]$, then

$$U_P(f\cdot g) - L_P(f\cdot g) \leq s[U_P(f) - L_P(f)] + s[U_P(g) - L_P(g)].$$

Proof: Let $P = \{x_0, x_1, ..., x_n\}$. We see from Lemma 13.22 that

$$(*)\ m_i(f)m_i(g) \leq m_i(f\cdot g) \leq M_i(f\cdot g) \leq M_i(f)M_i(g),\ \text{ all } i.$$

Hence,

$$\begin{aligned}
U_P(f\cdot g) - L_P(f\cdot g) &= \Sigma_{i=1}^n [M_i(f\cdot g) - m_i(f\cdot g)]\Delta x_i \\
&\overset{(*)}{\leq} \Sigma_{i=1}^n [M_i(f)M_i(g) - m_i(f)m_i(g)]\Delta x_i \\
&= \Sigma_{i=1}^n [M_i(f)M_i(g) - m_i(f)M_i(g) + m_i(f)M_i(g) - m_i(f)m_i(g)]\Delta x_i \\
&= \Sigma_{i=1}^n M_i(g)[M_i(f) - m_i(f)]\Delta x_i + \Sigma_{i=1}^n m_i(f)[M_i(g) - m_i(g)]\Delta x_i \\
&\leq s\Sigma_{i=1}^n [M_i(f) - m_i(f)]\Delta x_i + s\Sigma_{i=1}^n [M_i(g) - m_i(g)]\Delta x_i \\
&= s[U_P(f) - L_P(f)] + s[U_P(g) - L_P(g)].\ ¥
\end{aligned}$$

Our next lemma is the product theorem for nonnegative functions.

Lemma 13.24: If f and g are nonnegative functions that are integrable over $[a,b]$, then $f\cdot g$ is integrable over $[a,b]$.

Proof: Let $\epsilon > 0$ (we will use Theorem 12.15). Since f and g are integrable over $[a,b]$, f and g are bounded; thus, we have an upper bound $s > 0$ for both $f([a,b])$ and $g([a,b])$. Now, by Theorem 12.15, there are partitions P_1 and P_2 of $[a,b]$ such that

(1) $U_{P_1}(f) - L_{P_1}(f) < \frac{\epsilon}{2s}$ and $U_{P_2}(g) - L_{P_2}(g) < \frac{\epsilon}{2s}$.

Let P be a common refinement of P_1 and P_2 (see Exercise 12.3). Then, by Lemma 12.6 and Lemma 12.8, we have

(2) $L_{P_1}(f) \leq L_P(f) \leq U_P(f) \leq U_{P_1}(f)$

and

(3) $L_{P_2}(g) \leq L_P(g) \leq U_P(g) \leq U_{P_2}(g)$.

Now,

$$U_P(f) - L_P(f) \overset{(2)}{\leq} U_{P_1}(f) - L_{P_1}(f) \overset{(1)}{<} \frac{\epsilon}{2s}$$

and, similarly,

$$U_P(g) - L_P(g) \overset{(3)}{\leq} U_{P_2}(g) - L_{P_2}(g) \overset{(1)}{<} \tfrac{\epsilon}{2s}.$$

Hence, by Lemma 13.23 (and since $s > 0$),

$$U_P(f \cdot g) - L_P(f \cdot g) < s\tfrac{\epsilon}{2s} + s\tfrac{\epsilon}{2s} = \epsilon.$$

Therefore, since $\epsilon > 0$ was arbitrary, we have by Theorem 12.15 that $f \cdot g$ is integrable over $[a, b]$. ¥

Exercise 13.25: If f and g are integrable over $[a, b]$ and neither f nor g changes sign on $[a, b]$, then $f \cdot g$ is integrable over $[a, b]$.

We are ready to prove the general theorem about products.

Theorem 13.26: If f and g are integrable over $[a, b]$, then $f \cdot g$ is integrable over $[a, b]$.

Proof: Let $\overline{0}$ denote the zero function on $[a, b]$ (i.e., $\overline{0}(x) = 0$ for all $x \in [a, b]$). Consider the maximum and minimum functions of f and $\overline{0}$ and of g and $\overline{0}$ (as defined in Exercise 4.33): $f \bigvee \overline{0}$, $f \bigwedge \overline{0}$, $g \bigvee \overline{0}$, and $g \bigwedge \overline{0}$. By taking the four cases involving the possible signs of $f(x)$ and $g(x)$ for a fixed (but arbitrary) point x, we see that

$$(1)\ f \cdot g = (f \bigvee \overline{0}) \cdot (g \bigvee \overline{0}) + (f \bigvee \overline{0}) \cdot (g \bigwedge \overline{0}) + (f \bigwedge \overline{0}) \cdot (g \bigvee \overline{0}) + (f \bigwedge \overline{0}) \cdot (g \bigwedge \overline{0}).$$

Each of the functions $f \bigvee \overline{0}$, $f \bigwedge \overline{0}$, $g \bigvee \overline{0}$, and $g \bigwedge \overline{0}$ is integrable over $[a, b]$ by Exercise 13.19; furthermore, none of these functions changes sign on $[a, b]$. Hence, by Exercise 13.25, each of the four product functions on the right-hand side of (1) is integrable on $[a, b]$. Therefore, by (1) and Corollary 13.4, $f \cdot g$ is integrable on $[a, b]$. ¥

Corollary 13.27: If each of finitely many functions is integrable over $[a, b]$, then their product is integrable over $[a, b]$.

Proof: The corollary follows from Theorem 13.26 by a straightforward induction (Theorem 1.20). ¥

Exercise 13.28: Use results in this chapter to prove that polynomials are integrable over any closed and bounded interval. (Note: The result also follows from Theorem 12.33 since polynomials are continuous by Theorem 4.16.)

Exercise 13.29: If the product of two bounded functions is integrable over $[a, b]$, then must each of the functions be integrable over $[a, b]$? In other words, is the converse of Theorem 13.26 true?

Exercise 13.30: True or false: If f and g are integrable over $[a, b]$, then

$$\left|\int_a^b f \cdot g\right| \leq \left|\int_a^b f\right| \left|\int_a^b g\right|.$$

5. Integrability of Quotients

Let f and g be integrable functions defined on an interval $[a, b]$ such that g is never zero. The quotient $\frac{f}{g}$ is not necessarily integrable since $\frac{f}{g}$ may not be bounded. However, $\frac{f}{g}$ is integrable when $\frac{f}{g}$ is bounded. We do not know how to prove this theorem without using a characterization theorem in Chapter XV. At this time, we prove that $\frac{f}{g}$ is integrable when g is bounded away from zero (Theorem 13.36). You will be asked to prove the general theorem later (in Exercise 15.34).

If X is a nonempty set and $g : X \to \mathsf{R}^1$ is a function, then we say that g is *bounded away from zero* provided that there is an $\alpha > 0$ such that $|g(x)| > \alpha$ for all $x \in X$.

We first prove the theorem about integrability of quotients for the case of reciprocals (Theorem 13.35); then the theorem about quotients follows easily using our previous theorem about the integrability of products (Theorem 13.26). This pattern is analogous to what we did to obtain theorems about limits of quotients and derivatives of quotients in previous chapters.

We prove three lemmas. We use the first two lemmas to prove the third lemma, which we use to obtain the result about reciprocals.

Lemma 13.31: Let g be bounded function defined on a nonempty set X such that $glb_{x\in X} g(x) > 0$. Then

(1) $glb_{x\in X} \frac{1}{g(x)} = \frac{1}{lub_{x\in X} g(x)}$

and

(2) $lub_{x\in X} \frac{1}{g(x)} = \frac{1}{glb_{x\in X} g(x)}$.

Proof: Since $0 < g(y) \leq lub_{x\in X} g(x)$ for all $y \in X$, we have

$$\frac{1}{lub_{x\in X} g(x)} \leq \frac{1}{g(y)} \text{ for all } y \in X.$$

Hence, $\frac{1}{lub_{x\in X} g(x)}$ is a lower bound for $\{\frac{1}{g(x)} : x \in X\}$. Therefore, $glb_{x\in X} \frac{1}{g(x)}$ exists by the Greatest Lower Bound Axiom (section 8 of Chapter I) and

$$\frac{1}{lub_{x\in X} g(x)} \leq glb_{x\in X} \frac{1}{g(x)}.$$

Therefore, $\frac{1}{lub_{x\in X} g(x)} = glb_{x\in X} \frac{1}{g(x)}$ since if $\frac{1}{lub_{x\in X} g(x)} < glb_{x\in X} \frac{1}{g(x)}$, then there exists $z \in X$ such that $\frac{1}{g(z)} < glb_{x\in X} \frac{1}{g(x)}$, a contradiction. This proves (1). The proof of (2) is similar (and is left as an exercise). ¥

Exercise 13.32: Prove part (2) of Lemma 13.31.

Lemma 13.33: Let g be bounded function defined on a nonempty set X such that $lub_{x\in X} g(x) < 0$. Then

(1) $glb_{x\in X} \frac{1}{g(x)} = \frac{1}{lub_{x\in X} g(x)}$

and

(2) $lub_{x\in X} \frac{1}{g(x)} = \frac{1}{glb_{x\in X} g(x)}$.

Proof: Let $h = -g$. Then, since $glb_{x\in X} h(x) \overset{13.7}{=} -lub_{x\in X} g(x) > 0$, we can apply 13.31 to h to obtain that

(a) $glb_{x\in X} \frac{1}{h(x)} = \frac{1}{lub_{x\in X} h(x)}$

and

(b) $lub_{x\in X} \frac{1}{h(x)} = \frac{1}{glb_{x\in X} h(x)}$.

Therefore,

$$glb_{x\in X} \frac{1}{g(x)} \overset{13.7}{=} -lub \frac{1}{h(x)} \overset{(b)}{=} \frac{-1}{glb_{x\in X} h(x)} \overset{13.7}{=} \frac{1}{lub_{x\in X} g(x)}$$

and

$$lub_{x\in X} \frac{1}{g(x)} \overset{13.7}{=} -glb_{x\in X} \frac{1}{h(x)} \overset{(a)}{=} \frac{-1}{lub_{x\in X} h(x)} \overset{13.7}{=} \frac{1}{glb_{x\in X} g(x)}. \quad ¥$$

Lemma 13.34: Let g be bounded function defined on a nonempty set X such that g is bounded away from zero, say $|g(x)| > \alpha$ for all $x \in X$. Let

$$M = lub_{x\in X} \frac{1}{g(x)}, \quad m = glb_{x\in X} \frac{1}{g(x)}$$

Then $M - m < \frac{lub_{x\in X} g(x) - glb_{x\in X} g(x)}{\alpha^2}$.

Proof: Let

$$X^+ = \{x \in X : g(x) > 0\}, \quad X^- = \{x \in X : g(x) < 0\}.$$

We prove the lemma by considering three cases.

Case 1: $X = X^+$. Then

$$M - m \overset{13.31}{=} \frac{1}{glb_{x\in X} g(x)} - \frac{1}{lub_{x\in X} g(x)}$$
$$= \frac{lub_{x\in X} g(x) - glb_{x\in X} g(x)}{[glb_{x\in X} g(x)][lub_{x\in X} g(x)]} < \frac{lub_{x\in X} g(x) - glb_{x\in X} g(x)}{\alpha^2}.$$

Case 2: $X = X^-$. Then

$$M - m \overset{13.33}{=} \frac{1}{glb_{x\in X} g(x)} - \frac{1}{lub_{x\in X} g(x)}$$
$$= \frac{lub_{x\in X} g(x) - glb_{x\in X} g(x)}{[glb_{x\in X} g(x)][lub_{x\in X} g(x)]} < \frac{lub_{x\in X} g(x) - glb_{x\in X} g(x)}{\alpha^2}.$$

Case 3: $X^+ \neq \emptyset$ and $X^- \neq \emptyset$. Then we can let

$$\gamma^+ = glb\, g(X^+), \quad \gamma^- = lub\, g(X^-).$$

We see that

$$M - m \overset{13.31}{=} \frac{1}{\gamma^+} - m \overset{13.33}{=} \frac{1}{\gamma^+} - \frac{1}{\gamma^-} = \frac{\gamma^- - \gamma^+}{\gamma^+\gamma^-} = \frac{\gamma^+ - \gamma^-}{|\gamma^+\gamma^-|}$$
$$\leq \frac{lub_{x\in X} g(x) - glb_{x\in X} g(x)}{|\gamma^+\gamma^-|} < \frac{lub_{x\in X} g(x) - glb_{x\in X} g(x)}{\alpha^2}. \quad ¥$$

Theorem 13.35: If g is integrable over $[a, b]$ and g is bounded away from zero, then $\frac{1}{g}$ is integrable over $[a, b]$.

Proof: Since g is bounded away from zero, there exists $\alpha > 0$ such that $|g(x)| > \alpha$ for all $x \in [a, b]$.

Let $\epsilon > 0$. Since g is integrable over $[a, b]$, we have by Theorem 12.15 that there is a partition $P = \{x_0, x_1, ..., x_n\}$ of $[a, b]$ such that

(1) $U_P(g) - L_P(g) < \alpha^2\epsilon$.

Now,

$$U_P(\tfrac{1}{g}) - L_P(\tfrac{1}{g}) = \Sigma_{i=1}^{n}[M_i(\tfrac{1}{g}) - m_i(\tfrac{1}{g})]\Delta x_i$$

$$\overset{13.34}{<} \Sigma_{i=1}^{n}\tfrac{M_i(g)-m_i(g)}{\alpha^2}\Delta x_i = \tfrac{1}{\alpha^2}\Sigma_{i=1}^{n}[M_i(g) - m_i(g)]\Delta x_i$$

$$= \tfrac{1}{\alpha^2}[U_P(g) - L_P(g)] \overset{(1)}{<} \epsilon.$$

Therefore, since $\epsilon > 0$ was arbitrary, we have by Theorem 12.15 that $\frac{1}{g}$ is integrable over $[a, b]$. ¥

We are ready to prove our main result.

Theorem 13.36: If f and g are are integrable over $[a, b]$ and g is bounded away from zero, then $\frac{f}{g}$ is integrable over $[a, b]$.

Proof: By Theorem 13.35, $\frac{1}{g}$ is integrable over $[a, b]$. Therefore, since $\frac{f}{g} = f \cdot \frac{1}{g}$, $\frac{f}{g}$ is integrable over $[a, b]$ by Theorem 13.26. ¥

Exercise 13.37: Rational functions are integrable over any closed and bounded interval contained in their domain. Prove this without using Theorem 12.33.

Exercise 13.38: Give an example of integrable functions $f, g : [0, 1] \to \mathsf{R}^1$ such that $\frac{f}{g}$ is integrable over $[0, 1]$ but g is not bounded away from zero.

6. Integrability Over Subintervals

We prove that if f is integrable over $[a, b]$ and if $[c, d]$ is a subinterval of $[a, b]$, then the restriction of f to $[c, d]$ is integrable over $[c, d]$.[8] As a consequence, we obtain the following sum formula (where c is a point of $[a, b]$):

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

Theorem 13.39: If f is integrable over $[a, b]$ and $[c, d]$ is a subinterval of $[a, b]$, then the restricted function $f|[c, d]$ is integrable over $[c, d]$.

Proof: Let $\epsilon > 0$. Since f is integrable over $[a, b]$, Theorem 12.15 gives us a partition P of $[a, b]$ such that

(1) $U_P(f) - L_P(f) < \epsilon$.

Let $Q = P \cup \{c, d\}$, a partition of $[a, b]$. Then, since Q is a refinement of P,

[8] To avoid cumbersome notation, we write $\int_{\mathsf{c}}^{\mathsf{d}} \mathsf{f}$ instead of $\int_{\mathsf{c}}^{\mathsf{d}} \mathsf{f}|[\mathsf{c}, \mathsf{d}]$ for the integral of f over $[\mathsf{c}.\mathsf{d}]$.

$$L_P(f) \overset{12.8}{\leq} L_Q(f) \overset{12.6}{\leq} U_Q(f) \overset{12.8}{\leq} U_P(f).$$

Hence, by (1), we have that

(2) $U_Q(f) - L_Q(f) < \epsilon$.

Let $R = Q \cap [c,d]$. Since $c, d \in Q$ and Q is a partition of $[a,b]$, R is a partition of $[c,d]$; furthermore, each term in the sum for $U_R(f|[c,d]) - L_R(f|[c,d])$ is a term in the sum for $U_Q(f) - L_Q(f)$, and all terms in the sums are positive (since each term is of the form $[M_i(f) - m_i(f)]\Delta x_i$). Thus,

$$U_R(f|[c,d]) - L_R(f|[c,d]) \leq U_Q(f) - L_Q(f).$$

Hence, by (2),

$$U_R(f|[c,d]) - L_R(f|[c,d]) < \epsilon.$$

Therefore, since $\epsilon > 0$ was arbitrary, we have by Theorem 12.15 that $f|[c,d]$ is integrable over $[c,d]$. ¥

Theorem 13.40: If f is integrable over $[a,b]$ and c is any point of $[a,b]$, then

$$\int_a^b f = \int_a^c f + \int_c^b f.$$

Proof: Let $\epsilon > 0$. By Theorem 13.39, $\int_a^c f$ and $\int_c^b f$ exist. Thus, since the integral of a function is, by definition, the common value of the upper and the lower integrals of the function, there are partitions P_1 of $[a,c]$ and P_2 of $[c,b]$ such that

(1) $U_{P_1}(f|[a,c]) < \int_a^c f + \frac{\epsilon}{2}$, $L_{P_1}(f|[a,c]) > \int_a^c f - \frac{\epsilon}{2}$,

$U_{P_2}(f|[c,b]) < \int_c^b f + \frac{\epsilon}{2}$, $L_{P_2}(f|[c,b]) > \int_c^b f - \frac{\epsilon}{2}$.

Let $P = P_1 \cup P_2$, a partition of $[a,b]$. Then

$$\left(\int_a^c f - \tfrac{\epsilon}{2}\right) + \left(\int_c^b f - \tfrac{\epsilon}{2}\right) \overset{(1)}{<} L_{P_1}(f|[a,c]) + L_{P_2}(f|[c,b]) = L_P(f) \leq \int_a^b f$$

and

$$\int_a^b f \leq U_P(f) = U_{P_1}(f|[a,c]) + U_{P_2}(f|[c,b]) \overset{(1)}{<} \left(\int_a^c f + \tfrac{\epsilon}{2}\right) + \left(\int_c^b f + \tfrac{\epsilon}{2}\right).$$

The first and last parts of the expressions give us that

$$\int_a^c f + \int_c^b f - \epsilon < \int_a^b f < \int_a^c f + \int_c^b f + \epsilon.$$

Therefore, since $\epsilon > 0$ was arbitrary, we have that

$$\int_a^b f = \int_a^c f + \int_c^b f. \text{ ¥}$$

A useful result in the reverse direction to Theorem 13.39 is in Exercise 13.43.

Exercise 13.41: Evaluate $\int_{-2}^{2} f$ when

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

Be sure to explain why f is integrable over $[-2, 2]$.

Exercise 13.42: Evaluate $\int_{-2}^{2} f$ when $f(x) = |x + 1|$. Be sure to explain why f is integrable over $[-2, 2]$.

Exercise 13.43: If f is integrable over $[a, c]$ and f is integrable over $[c, b]$, then f is integrable over $[a, b]$ and $\int_a^b f = \int_a^c f + \int_c^b f$.

Exercise 13.44: Evaluate $\int_{1}^{3} f$ when

$$f(x) = \begin{cases} x & \text{, if } 1 \leq x \leq 2 \\ x + 1 & \text{, if } 2 < x \leq 3. \end{cases}$$

Be sure to explain why f is integrable over $[1, 3]$.

Chapter XIV: The Fundamental Theorem of Calculus

We prove the Fundamental Theorem of Calculus. This beautiful theorem shows a surprising connection between derivatives and integrals – in short, the theorem unifies the subject of calculus. Thus, it is only appropriate that the theorem stand alone, in a chapter all by itself. Nevertheless, we include an application that illustrates a geometric aspect of the theorem; namely (in section 2), we discuss the use of the theorem in connection with computing area, which puts our informal discussion in Chapter XI on a firm (rigorous) foundation.

1. The Fundamental Theorem

We prove the Fundamental Theorem of Calculus after we prove the following technical lemma.

Lemma 14.1: Assume that f is integrable over $[a, b]$, where $a < b$. Let $p \in [a, b]$, and let $h \neq 0$ be a real number such that $p + h \in [a, b]$. For each $x \in [a, b]$, let $F(x) = \int_a^x f$ (which exists by Theorem 13.39).

(1) If $h > 0$, then $\left|\frac{F(p+h)-F(p)}{h} - f(p)\right| \leq \frac{1}{h}\int_p^{p+h} |f - f(p)|$.

(2) If $h < 0$, then $\left|\frac{F(p+h)-F(p)}{h} - f(p)\right| \leq \frac{-1}{h}\int_{p+h}^{p} |f - f(p)|$.

Proof: To prove (1), assume that $h > 0$. Then, since $a \leq p < p + h$,

$$\int_a^{p+h} f \overset{13.40}{=} \int_a^p f + \int_p^{p+h} f.$$

Thus,

$$\frac{F(p+h)-F(p)}{h} = \frac{1}{h}\left(\int_a^{p+h} f - \int_a^p f\right) = \frac{1}{h}\int_p^{p+h} f;$$

also, since $\int_p^{p+h} f(p) = hf(p)$ (by Exercise 12.13),

$$f(p) = \frac{1}{h}\int_p^{p+h} f(p).$$

Hence,

$$\left|\frac{F(p+h)-F(p)}{h} - f(p)\right| = \left|\frac{1}{h}\int_p^{p+h} f - \frac{1}{h}\int_p^{p+h} f(p)\right|$$

$$\overset{13.12}{=} \left|\frac{1}{h}\int_p^{p+h}(f - f(p))\right| = \frac{1}{h}\left|\int_p^{p+h}(f - f(p))\right| \overset{13.17}{\leq} \frac{1}{h}\int_p^{p+h} |f - f(p)|.$$

This proves (1).

To prove (2), assume that $h < 0$. Then, since $a \leq p + h < p$,

$$\int_a^p f \overset{13.40}{=} \int_a^{p+h} f + \int_{p+h}^p f.$$

Thus,

$$\tfrac{F(p+h)-F(p)}{h} = \tfrac{1}{h}\left(\int_a^{p+h} f - \int_a^p f\right) = \tfrac{-1}{h}\int_{p+h}^p f;$$

also, since $\int_{p+h}^p f(p) = -hf(p)$ (by Exercise 12.13),

$$f(p) = \tfrac{-1}{h}\int_{p+h}^p f(p).$$

Hence (note for equality in third row below that $\frac{-1}{h} > 0$),

$$\begin{aligned}\left|\tfrac{F(p+h)-F(p)}{h} - f(p)\right| &= \left|\tfrac{-1}{h}\int_{p+h}^p f + \tfrac{1}{h}\int_{p+h}^p f(p)\right| \\ &= \left|\tfrac{-1}{h}\left(\int_{p+h}^p f - \int_{p+h}^p f(p)\right)\right| \overset{13.12}{=} \left|\tfrac{-1}{h}\int_{p+h}^p (f - f(p))\right| \\ &= \tfrac{-1}{h}\left|\int_{p+h}^p (f - f(p))\right| \overset{13.17}{\leq} \tfrac{-1}{h}\int_{p+h}^p |f - f(p)|.\end{aligned}$$

This proves (2). ¥

Theorem 14.2 (The Fundamental Theorem of Calculus): Assume that $a < b$ and that $f : [a,b] \to \mathsf{R}^1$ is a continuous function.

(1) The function F given by $F(x) = \int_a^x f$ for each $x \in [a,b]$ is differentiable on $[a,b]$ and $F' = f$.

(2) If g is any differentiable function on $[a,b]$ such that $g' = f$, then

$$\int_a^b f = g(b) - g(a).$$

Proof: To prove part (1), first note that the function F in (1) is, indeed, defined for each $x \in [a,b]$ by Theorem 12.33 (since $f|[a,x]$ is continuous by Exercise 5.3).

Now, fix a point $p \in [a,b]$. We want to show that $F'(p) = f(p)$. We show this using the definition of the derivative (in section 1 of Chapter VI),

$$F'(p) = \lim_{h\to 0} \tfrac{F(p+h)-F(p)}{h} \quad \text{(if the limit exists)}.$$

Specifically (recalling the definition of limit in section 1 of Chapter III), we show that for any $\epsilon > 0$, there is a $\delta > 0$ such that

$$(*)\ \left|\tfrac{F(p+h)-F(p)}{h} - f(p)\right| < \epsilon \text{ when } h \neq 0,\ p+h \in [a,b], \text{ and } |h| < \delta.$$

Proof of (*)*:* Let $\epsilon > 0$. Then, since f is continuous at p, Theorem 3.12 gives us a $\delta > 0$ such that

(i) $|f(x) - f(p)| < \frac{\epsilon}{2}$ for all $x \in [a,b]$ such that $|x - p| < \delta$.

We prove that this choice of δ satisfies (*).

Fix $h \neq 0$ such that $p + h \in [a,b]$ and $|h| < \delta$.

Assume first that $h > 0$. Then, by (i), $|f(x) - f(p)| < \frac{\epsilon}{2}$ for all $x \in [p, p+h]$; hence,

$$\int_p^{p+h} |f - f(p)| \overset{13.14}{\leq} \int_p^{p+h} \tfrac{\epsilon}{2} \overset{12.13}{=} h\tfrac{\epsilon}{2}.$$

Thus,

$$\left|\tfrac{F(p+h)-F(p)}{h} - f(p)\right| \overset{14.1}{\leq} \tfrac{1}{h}\int_p^{p+h} |f - f(p)| \leq \tfrac{1}{h}(h\tfrac{\epsilon}{2}) = \tfrac{\epsilon}{2} < \epsilon.$$

This proves (*) when $h > 0$.

Assume next that $h < 0$ (the proof is the same as when $h > 0$, as we will see). Then, by (i), $|f(x) - f(p)| < \frac{\epsilon}{2}$ for all $x \in [p + h, p]$; hence,

$$\int_{p+h}^{p} |f - f(p)| \overset{13.14}{\leq} \int_{p+h}^{p} \tfrac{\epsilon}{2} \overset{12.13}{=} -h\tfrac{\epsilon}{2}.$$

Thus,

$$\left|\tfrac{F(p+h)-F(p)}{h} - f(p)\right| \overset{14.1 \text{ (part (2))}}{\leq} \tfrac{-1}{h}\int_{p+h}^{p} |f - f(p)| \leq \tfrac{-1}{h}(-h\tfrac{\epsilon}{2}) = \tfrac{\epsilon}{2} < \epsilon.$$

This proves (*) when $h < 0$.

Therefore, we have proved that for any $\epsilon > 0$, there is a $\delta > 0$ such that (*) holds.

Hence, $F'(p) = f(p)$. This completes the proof of part (1) of our theorem.

To prove part (2), let g be any differentiable function on $[a, b]$ such that $g' = f$. Then, by part (1) of our theorem, $g' = F'$. Hence, by Theorem 10.8, g and F differ by a constant, say

$$F(x) - g(x) = C \text{ for all } x \in [a, b].$$

Let's evaluate C: Since $F(a) = g(a) + C$ and $F(a) = \int_a^a f = 0$ (by the definition of F), we have that $C = -g(a)$. Therefore,

$$\int_a^b f = F(b) = g(b) + C = g(b) - g(a).$$

This proves part (2) of our theorem. ¥

It is easy to apply the Fundamental Theorem of Calculus to evaluate integrals of many continuous functions whose integrals we could not have evaluated up until now. For example,

$$\int_1^3 x^4 = \tfrac{3^5}{5} - \tfrac{1^5}{5} = \tfrac{242}{5}, \quad \int_1^4 \sqrt{x} = \tfrac{2}{3}4^{\frac{3}{2}} - \tfrac{2}{3}1^{\frac{3}{2}} = \tfrac{14}{3},$$
$$\int_0^{\frac{\pi}{3}} \sin(x) = -\cos(\tfrac{\pi}{3}) + \cos(0) = \tfrac{1}{2},$$

and so on. However, there are numerous continuous functions whose integrals we still can not evaluate: For example, $\int_1^2 \frac{x^2+1}{\sqrt{x^5+3x+2}}$ or even one as simple as $\int_1^2 \frac{1}{x}$. We will never be able to evaluate the first integral; however, we will see in

Chapter XVI that the second integral is ln(2) and, thus, that logarithm tables can be used to approximate the value of the second integral.

Exercise 14.3: The assumption that f is continuous in part (1) of the Fundamental Theorem of Calculus is necessary: Give an example to show that part (1) of the theorem would be false if we had only assumed that f is integrable over $[a, b]$.

(Note: Part (2) of the Fundamental Theorem of Calculus generalizes to functions f that are only assumed to be integrable over $[a, b]$. We will not prove this.)

Exercise 14.4: Let $f : [0, 6] \to \mathsf{R}^1$ be the continuous function whose graph is drawn in Figure 14.4 below. Let F be the function in part (1) of Theorem 14.2, $F(x) = \int_0^x f$ for all $x \in [0, 6]$. At which points (on the x-axis) does F have local or global extrema? What type of extremum occurs at each such point? Sketch a rough graph of F (showing where F has inflection points).

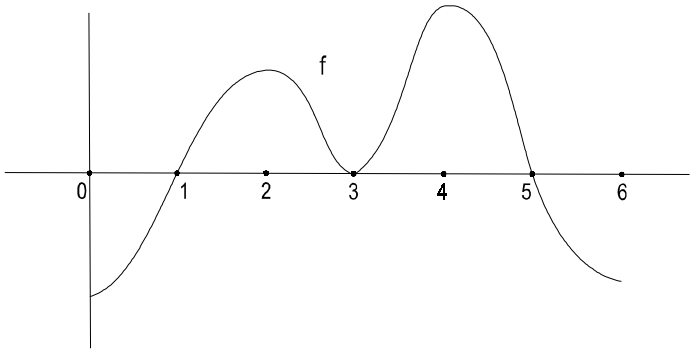


Figure 14.4

Exercise 14.5: If f is a continuous on $[a, b]$ and $a < b$, then there is a point $p \in (a, b)$ such that

$$f(p) = \frac{\int_a^b f}{b-a}.$$

This result is called the Mean Value Theorem for Integrals, and $f(p) = \frac{\int_a^b f}{b-a}$ is often referred to as the *average value of f over* $[a, b]$. (The next three exercises are follow ups to this one.)

Exercise 14.6: Find the average value of $f(x) = \sin(2x)$ over $[0, \frac{\pi}{2}]$. (Average value is as defined in Exercise 14.5.)

The mean daily temperature in Morgantown t months after May 1 ($t \leq 6$) is given by the formula $f(t) = 63 + 30\sin(\frac{\pi t}{12})$. Determine the average value of the temperature between June 1 and September 1.

Exercise 14.7: Give an example to show that the result in Exercise 14.5 would fail if we had only assumed that f is integrable on $[a, b]$.

Exercise 14.8: Assume that f is continuous on $[a, b]$ and that $a < b$. For any $x \in [a, b]$ with $x > a$, the average value of f over $[a, x]$ is $\frac{\int_a^x f}{x-a}$ (Exercise 14.5); on the other hand, the average of the values $f(a)$ and $f(x)$ is $\frac{f(a)+f(x)}{2}$.

Determine all the continuous functions f on $[a, b]$ such that for all $x \in [a, b]$ with $x > a$, the average value of f over $[a, x]$ is the average of the values $f(a)$ and $f(x)$.

2. Area Again

Let f be a continuous nonnegative function on an interval $[a, b]$. In Chapter XI, we intuitively discussed the idea of the area between the graph of f and the interval $[a, b]$, and we indicated how to compute the area. It is almost evident that the Fundamental Theorem of Calculus gives us a rigorous definition for the area function A that we used in Chapter XI and is the theorem behind the procedure we arrived at for computing area in Chapter XI. It is only *almost evident* because our approach to area in Chapter XI was slightly different than our approach to the integral in Chapter XII. We show in this section that the two approaches are actually equivalent.

We temporarily disregard the approach to area in Chapter XI. In its place, we define the area between the graph of any integrable function f and the interval $[a, b]$ on which f is defined in terms of the integral. This general definition does not require f to be continuous or to be nonnegative (as was required in Chapter XI).

Definition: Let f be an integrable function on the interval $[a, b]$. We define the *area between the graph of f and the interval* $[a, b]$ to be $\int_a^b |f|$. (Recall that $|f|$ is integrable over $[a, b]$ by Theorem 13.17.)

In the definition we assume f is integrable, not just that $|f|$ is integrable even though the area is the integral of $|f|$. By doing so, Theorem 13.3 assures us that the *existence* of area is invariant under vertical translation; this is obviously a property that any notion called area should have. In fact, this property would fail if we had only assumed in the definition that $|f|$ is integrable: For example, if f is defined on $[0, 1]$ by

$$f(x) = \begin{cases} 1 & \text{, if } x \text{ is rational} \\ -1 & \text{, if } x \text{ is irrational} \end{cases}$$

then $\int_0^1 |f| = 1$ but $\int_0^1 |f + 1|$ does not exist (just like in Example 12.12).

Next, we bring the definition of area above in sync with the approach to area in Chapter XI. We first provide terminology for the types of sums we used in Chapter XI.

Definition: Let $f : [a, b] \to \mathsf{R}^1$ be a function, and let $P = \{x_0, x_1, ..., x_n\}$ be a partition of $[a, b]$. A *Riemann sum for f with respect to P* is a sum of the

form $\Sigma_{i=1}^{n} f(t_i)\Delta x_i$ for any choice of points $t_i \in [x_{i-1}, x_i]$ for each i. We denote any such Riemann sum by $R_P(f)$ (without reference to the points t_i).

We now define the notion of limit for Riemann sums. The definition gives rigorous meaning to the intuitive idea for limits of sums that we worked with in Chapter XI (see the footnote on the second page of Chapter XI). Recall that the norm, $\|P\|$, of a partition P is defined above Exercise 12.32.

Definition: Let $f : [a, b] \to \mathsf{R}^1$ be a function. We say that L *is the limit of the Riemann sums for* f *as the norms of the partitions of* $[a, b]$ *go to* 0, written

$$\lim_{\|P\| \to 0} R_P(f) = L,$$

provided that for each $\epsilon > 0$, there is a $\delta > 0$ such that if $P = \{x_0, x_1, ..., x_n\}$ is any partition of $[a, b]$ and $\|P\| < \delta$, then $|R_P(f) - L| < \epsilon$, meaning that

$$|\Sigma_{i=1}^{n} f(t_i)\Delta x_i - L| < \epsilon \quad \text{for all choices of points } t_i \in [x_{i-1}, x_i].$$

Finally, the following theorem will show that our approach to area in Chapter XI is the same as area defined in terms of the integral at the beginning of this section (see comments following the proof):

Theorem 14.9: If f is a continuous function on $[a, b]$, then

$$\int_a^b f = \lim_{\|P\| \to 0} R_P(f).$$

Proof: Let $\epsilon > 0$. Then, since f is uniformly continuous (by Theorem 12.31), there is a $\delta > 0$ such that

$$|f(y) - f(z)| < \tfrac{\epsilon}{b-a} \text{ whenever } y, z \in [a, b] \text{ and } |y - z| < \delta.$$

Let $P = \{x_0, x_1, ..., x_n\}$ be any partition of $[a, b]$ such that $\|P\| < \delta$. Then, since δ satisfies the condition for δ in the proof of Theorem 12.33, the calculations in the proof of Theorem 12.33 show that

$$(*)\ U_P(f) - L_P(f) < \epsilon.$$

Now, choose any points $t_i \in [x_{i-1}, x_i]$ for each i. Since $m_i(f) \leq f(t_i) \leq M_i(f)$ for each i, it is clear from the definitions of $L_P(f)$ and $U_P(f)$ (section 2 of Chapter XII) that

$$L_P(f) \leq \Sigma_{i=1}^{n} f(t_i)\Delta x_i \leq U_P(f);$$

also, by the definition of the integral (section 3 of Chapter XII),

$$L_P(f) \leq \int_a^b f \leq U_P(f).$$

Thus, by (*),

$$\left|\Sigma_{i=1}^{n} f(t_i)\Delta x_i - \int_a^b f\right| < \epsilon.$$

Therefore, since the points t_i are any points in the intervals $[x_{i-1}, x_i]$,

$$\left|R_P(f) - \int_a^b f\right| < \epsilon. \text{ ¥}$$

As in Chapter XI, let f be a continuous nonnegative function on an interval $[a, b]$. By Theorem 14.9, we can now conclude that the area function A in Chapter XI is the function in part (1) of the Fundamental Theorem of Calculus; that is,

$$A(x) = \int_a^x f \text{ for each } x \in [a, b].$$

We also see that the procedure for computing area in Chapter XI, which is summarized in (#) above Example 11.1, is justified by part (2) of the Fundamental Theorem of Calculus.

Exercise 14.10: Let $f(x) = x^2 + x - 2$. Find the area between the graph of f and the interval $[-2, 3]$.

Exercise 14.11: Let $f(x) = \frac{1}{96+x^3}$. Find $c > 0$ such that the area between the graph of f and the interval $[c, 3c]$ is largest.

Exercise 14.12: Using only Theorem 14.9 and the Mean Value Theorem (Theorem 10.2), give a short, elegant proof of part (2) of the Fundamental Theorem of Calculus.

Exercise 14.13: If $f : [a, b] \to \mathsf{R}^1$ is a function such that $\lim_{\|P\|\to 0} R_P(f)$ exists, then f is bounded on $[a, b]$.

Exercise 14.14: If $f : [a, b] \to \mathsf{R}^1$ is a function such that $\lim_{\|P\|\to 0} R_P(f)$ exists, then f is integrable over $[a, b]$ and

$$\int_a^b f = \lim_{\|P\|\to 0} R_P(f).$$

(*Hint:* Let $L = \lim_{\|P\|\to 0} R_P(f)$. Let $\epsilon > 0$. Give reasons for each of the following statements: There is a partition $P = \{x_0, x_1, ..., x_n\}$ of $[a, b]$ such that $L - \frac{\epsilon}{2} < \Sigma_{i=1}^{n} f(t_i)\Delta x_i < L + \frac{\epsilon}{2}$ for all $t_i \in [x_{i-1}, x_i]$. For each i, there exist $p_i, q_i \in [x_{i-1}, x_i]$ such that $M_i(f) - \frac{\epsilon}{2(b-a)} < f(p_i)$ and $f(q_i) < m_i(f) + \frac{\epsilon}{2(b-a)}$ (note that $M_i(f)$ and $m_i(f)$ exist by Exercise 14.13). Then $U_P(f) - \frac{\epsilon}{2} = \Sigma_{i=1}^{n}[M_i(f) - \frac{\epsilon}{2(b-a)}]\Delta x_i < L + \frac{\epsilon}{2}$, hence $U_P(f) < L + \epsilon$; similarly, $L_P(f) > L - \epsilon$. Thus, $\overline{\int}_a^b f < L + \epsilon$ and $L - \epsilon < \underline{\int}_a^b f$. The result now follows.)

The converse of the first part of Exercise 14.14 is true: If f is integrable over $[a, b]$, then $\lim_{\|P\|\to 0} R_P(f)$ exists. Thus, a function f is integrable over $[a, b]$ if and only if $\lim_{\|P\|\to 0} R_P(f)$ exists, in which case $\int_a^b f = \lim_{\|P\|\to 0} R_P(f)$. This equivalence justifies the somewhat common practice of *defining* the integral in terms of Riemann sums.

Riemann sums are useful for envisioning how to set up an integral to solve a mathematical or physical problem. One illustration of this is in Chapter XI – it was only natural to use Riemann sums to arrive at the notion of area. We give another illustration in the following exercise:

Exercise 14.15: Let f be a continuous nonnegative function on $[a, b]$. Using Riemann sums, find a reasonable formula for the volume of the solid obtained by revolving the graph of f about the x-axis.

Indicate that your formula is reasonable by showing that it gives the known value ($\frac{4}{3}\pi r^3$) for the volume of the sphere of radius r centered at the origin in 3-space.