Mathematical Analysis

Mohammad Safdari

Contents

Chapter 1

Real Numbers

1.1 Ordered Fields

Definition 1.1. A **field** is a nonempty set F equipped with two binary operations

$$\begin{array}{c} F \times F \longrightarrow F \\ (a,b) \mapsto a+b \end{array}, \qquad \begin{array}{c} F \times F \longrightarrow F \\ (a,b) \mapsto ab \end{array},$$

called respectively **addition** and **multiplication**, such that

(i) The operations are *associative* and *commutative*, i.e. for every $a, b, c \in F$

$$\begin{aligned} a + (b + c) = (a + b) + c, &\qquad a(bc) = (ab)c, \\ a + b = b + a, &\qquad ab = ba. \end{aligned}$$

(ii) There exist elements $0, 1 \in F$, called respectively *zero* and *identity* of F, such that $1 \neq 0$, and for every $a \in F$

$$a + 0 = a, \qquad a1 = a.$$

(iii) For every $a \in F$ there exists an element $-a \in F$, called the **opposite** of a, such that

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

(iv) For every $a \in F - \{0\}$ there exists an element $a^{-1} \in F$, called the **inverse** of a, such that

$$aa^{-1} = 1.$$

(v) Multiplication is *distributive* over addition, i.e. for every $a, b, c \in F$

$$a(b + c) = ab + ac.$$

Let a, b be two elements in a field F. Then $a+b$ is called the *sum* of a, b. And a, b are called the *summands* of $a+b$. Also, ab is called the *product* of a, b; and a, b are called the *factors* of ab. We sometimes denote the product of two elements a, b by $a \cdot b$, or $a \times b$. In addition, the *square* and the *cube* of an element a are respectively defined as

$$a^2 := aa, \qquad a^3 := a^2 a = aaa.$$

The inverse of a is also called the *reciprocal* of a.

Furthermore, the **subtraction** and the **division** of two elements $a, b \in F$ are respectively defined as follows

$$a - b := a + (-b), \qquad a/b = \frac{a}{b} := ab^{-1} \text{ when } b \neq 0.$$

The element $a - b$ is called the *difference* of a, b. The element $\frac{a}{b}$ is called the *quotient* or the *ratio* of a, b. We also call $\frac{a}{b}$ a *fraction*. In a fraction $\frac{a}{b}$, a is called the *numerator*, and b is called the *denominator*. The *reciprocal* of the fraction $\frac{a}{b}$ is the fraction $\frac{b}{a}$, provided that a is also nonzero. We will see that $\frac{b}{a}$ is the inverse of $\frac{a}{b}$; so the two uses of the term "reciprocal" are compatible.

Remark. Note that due to the commutativity we have

$$0 + a = a, \qquad 1a = a, \qquad (-a) + a = 0.$$

We also have $a^{-1}a = 1$, if $a \neq 0$.

Remark. Informally, a field is a structure in which we can perform the four basic arithmetic operations, i.e. addition, subtraction, multiplication, and division.

Notation. We will assume that in a field, multiplication binds stronger than addition; thus, for example, $ab + c$ means $(ab) + c$, not $a(b + c)$. In addition, we assume that multiplication binds stronger than taking the opposite. So for example, $-ab$ means $-(ab)$.

***Example* 1.2.** $\mathbb{Q}$ is a field.

Theorem 1.3. *Let F be a field. Then for every $a, b, c, d \in F$ we have*

(i) **(Cancellation Laws)**

$$a + c = b + c \implies a = b,$$
$$ac = bc, \ c \neq 0 \implies a = b.$$

(ii) *The zero and identity of F are unique.*

(iii) *The opposite of every element of F is unique; and the inverse of every nonzero element of F is unique.*

(iv) $-0 = 0$, *and* $1^{-1} = 1$.

(v) $0a = 0 = a0$. *And* $ab = 0 \implies a = 0$ *or* $b = 0$.

(vi) $-(-a) = a$, *and* $-(a+b) = (-a) + (-b) = -a - b$.

(vii) *If* $a \neq 0$ *then* $(a^{-1})^{-1} = a$. *And if* $a, b \neq 0$ *then* $(ab)^{-1} = a^{-1}b^{-1}$.

(viii) $(-a)b = -ab = a(-b)$, *and* $(-a)(-b) = ab$.

(ix) $-a = (-1)a$, $-(b-c) = c - b$, *and* $a(b-c) = ab - ac$.

(x) *If* $a \neq 0$ *then* $a^{-1} = \frac{1}{a}$, *and* $(-a)^{-1} = -a^{-1}$.

(xi) *If* $b, d \neq 0$ *then*

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}, \qquad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}.$$

(xii) *If* $b, c, d \neq 0$ *we have*

$$-\frac{a}{b} = \frac{-a}{b} = \frac{a}{-b}, \qquad \left(\frac{c}{d}\right)^{-1} = \frac{d}{c}, \qquad \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}.$$

Proof. **(i)** By adding $-c$ to both sides of $a + c = b + c$ we obtain $(a+c)+(-c) = (b+c)+(-c)$. Now by associativity of addition we have $a+(c+(-c)) = b+(c+(-c))$. Since $c + (-c) = 0$, we get $a + 0 = b + 0$; and hence $a = b$. The multiplicative case can be proved similarly using c^{-1}.

(ii) Suppose $0, \tilde{0}$ are both zeros of F. Then we have $\tilde{0} = \tilde{0} + 0$, since 0 is a zero. We also have $0 + \tilde{0} = 0$, since $\tilde{0}$ is a zero. However we know that $\tilde{0} + 0 = 0 + \tilde{0}$, because addition is commutative. Therefore we must have $\tilde{0} = 0$, as desired. Similarly, for two identities $1, \tilde{1}$ we have $\tilde{1} = \tilde{1}1 = 1$.

(iii) Suppose $\tilde{a}$ is also an opposite of a. Then we have

$$\tilde{a} + a = 0 = (-a) + a.$$

Thus by cancellation law we get $\tilde{a} = -a$, as desired. The case of inverse is similar.

(iv) We have $0 + 0 = 0$, and $1 \cdot 1 = 1$. Now the result follows from the uniqueness of opposite and inverse.

(v) We have

$$0 + 0a = 0a = (0+0)a = 0a + 0a,$$

Thus by cancellation law we get $0 = 0a$. Now we have $0a = a0 = 0$.

Next, if $ab = 0$ and $a \neq 0$ then we have

$$b = 1b = (a^{-1}a)b = a^{-1}(ab) = a^{-1}0 = 0.$$

(vi) The proof is similar to (vii). Note that the last equality in (vi) holds by definition.

(vii) First note that if $a^{-1} = 0$ then we get $1 = aa^{-1} = a0 = 0$; which contradicts the fact that $1 \neq 0$. Therefore we must have $a^{-1} \neq 0$. Hence a^{-1} is invertible. Now note that $(a^{-1})^{-1}a^{-1} = 1 = aa^{-1}$. Thus by cancellation law we get $(a^{-1})^{-1} = a$, as desired.

Next, note that when a, b are nonzero, ab is also nonzero by (v); hence ab is invertible. Furthermore

$$\begin{aligned}(a^{-1}b^{-1})(ab) &= (b^{-1}a^{-1})(ab) = b^{-1}(a^{-1}(ab))\\ &= b^{-1}((a^{-1}a)b) = b^{-1}(1b) = b^{-1}b = 1.\end{aligned}$$

Hence $(ab)(a^{-1}b^{-1}) = 1$. Thus by the uniqueness of inverse we obtain $(ab)^{-1} = a^{-1}b^{-1}$.

(viii) Note that

$$ab + (-a)b = (a + (-a))b = 0b = 0.$$

Thus uniqueness of opposite implies $(-a)b = -ab$. The other equalities can be proved similarly.

(ix) We have $(-1)a + a = ((-1) + 1)a = 0a = 0$. Therefore $(-1)a = -a$, since the opposite is unique. We also have

$$\begin{aligned}-(b - c) &= -(b + (-c)) = -b + (-(-c)) = -b + c = c - b,\\ a(b - c) &= a(b + (-c)) = ab + a(-c) = ab + (-ac) = ab - ac.\end{aligned}$$

(x) By the definition of division, for $a \neq 0$ we have $\frac{1}{a} = 1a^{-1} = a^{-1}$. Also, $(-a^{-1})(-a) = a^{-1}a = 1$. Hence we get the desired by uniqueness of inverse.

(xi) By the definition of division we have

$$\begin{aligned}\frac{ad + bc}{bd} &= (ad + bc)(bd)^{-1} = (ad + bc)b^{-1}d^{-1}\\ &= adb^{-1}d^{-1} + bcb^{-1}d^{-1} = ab^{-1} + cd^{-1} = \frac{a}{b} + \frac{c}{d},\\ \frac{a}{b} \cdot \frac{c}{d} &= (ab^{-1})(cd^{-1}) = acb^{-1}d^{-1} = ac(bd)^{-1} = \frac{ac}{bd}.\end{aligned}$$

(xii) We have

$$\begin{aligned}\frac{-a}{b} &= (-a)b^{-1} = -ab^{-1} = -\frac{a}{b},\\ \frac{a}{-b} &= a(-b)^{-1} = a(-b^{-1}) = -ab^{-1} = -\frac{a}{b}.\end{aligned}$$

We also have

$$\begin{aligned}\Big(\frac{c}{d}\Big)^{-1} &= (cd^{-1})^{-1} = c^{-1}(d^{-1})^{-1} = c^{-1}d = \frac{d}{c},\\ \frac{\frac{a}{b}}{\frac{c}{d}} &= \Big(\frac{a}{b}\Big)\Big(\frac{c}{d}\Big)^{-1} = \frac{a}{b} \cdot \frac{d}{c} = \frac{ad}{bc}.\end{aligned}$$

■

Definition 1.4. An **ordered field** is a field F equipped with a binary relation $<$, such that

(i) The relation $<$ is a linear order, which means $<$ is transitive and satisfies the trichotomy, i.e. for every $a, b, c \in F$

$$a < b,\ b < c \implies a < c,$$
$$\text{Exactly one of the } a < b,\ a = b,\ b < a, \text{ is true.}$$

(ii) For every $a, b, c \in F$ we have

$$a < b \implies a + c < b + c,$$
$$0 < a,\, 0 < b \implies 0 < ab.$$

Notation. As usual, we define $a \leq b$ to mean $a < b$ or $a = b$. Similarly we define $a > b$ to mean $b < a$, and $a \geq b$ to mean $b \leq a$. We say a is **positive** if $a > 0$, and a is **negative** if $a < 0$. We also say a is *nonnegative* or *nonpositive*, if $a \geq 0$ or $a \leq 0$ respectively.

***Example* 1.5.** $\mathbb{Q}$ is an ordered field.

Theorem 1.6. *Suppose F is an ordered field. Then for every $a, b, c, d \in F$ we have*

(i) *$a < b$ if and only if $b - a > 0$.*
(ii) *$a > 0$ if and only if $-a < 0$.*
(iii) *If $a \neq 0$ then $a^2 > 0$. As a result $1 = 1^2 > 0$.*
(iv) *$a < b$ if and only if $a + c < b + c$.*
(v) *If $a < c$, $b \leq d$ then $a + b < c + d$.*
(vi) *If $c > 0$, then $a < b$ if and only if $ac < bc$.*
(vii) *If $c < 0$, then $a < b$ if and only if $ac > bc$.*
(viii) *If $a, b < 0$ then $ab > 0$.*
(ix) *If $0 \leq a < c$, $0 < b \leq d$ then $ab < cd$.*
(x) *If $a > 0$ then $a^{-1} > 0$.*
(xi) *If $0 < a < b$ then $a^{-1} > b^{-1} > 0$. In particular, $a > 1$ if and only if $0 < a^{-1} < 1$.*
(xii) *If $a \leq b$ and $a \geq b$ then $a = b$.*
(xiii) *If $a_1, \ldots, a_n \geq 0$ and $a_1 + \cdots + a_n = 0$ then $a_i = 0$ for all i.*

Proof. **(i)** We have

$$a < b \implies a + (-a) < b + (-a) \implies 0 < b - a.$$

Similarly, for the converse we can add a to both sides of $0 < b - a$.

(ii) By (i) we have $-a < 0 \iff 0 < 0 - (-a) = a$.

(iii) If $a \neq 0$ then either $0 < a$ or $a < 0$. Hence by (ii) we either have $0 < a$ or $0 < -a$. Thus either

$$0 < aa = a^2, \qquad \text{or} \qquad 0 < (-a)(-a) = a^2.$$

(iv) $a < b$ implies $a+c < b+c$ by definition. On the other hand, if $a+c < b+c$ then we cannot have $a \geq b$. Since in that case we would obtain $a+c \geq b+c$, which is a contradiction. So we must have $a < b$. Alternatively, we can add $-c$ to both sides of $a+c < b+c$ to get $a < b$.

(v) When $b = d$ then the claim holds by the definition of ordered fields. So suppose $b < d$. Then we add b to both sides of $a < c$, and add c to both sides of $b < d$, to obtain

$$a+b < c+b < c+d.$$

(vi) By (i) we have $0 < b-a$. Thus if $c > 0$ then

$$0 < (b-a)c = bc - ac \implies ac < bc.$$

Conversely, suppose $ac < bc$. Then we cannot have $a \geq b$, since in this case we would obtain $ac \geq bc$, which is a contradiction. So we must have $a < b$.

(vii) When $c < 0$ we have $-c > 0$. Hence

$$0 < (b-a)(-c) = -(bc-ac) = ac - bc \implies bc < ac.$$

Conversely, suppose $ac > bc$. Then we cannot have $a \geq b$, since in this case we would obtain $ac \leq bc$, which is a contradiction. So we must have $a < b$.

(viii) We multiply both sides of $a < 0$ by b to obtain $ab > 0b = 0$.

(ix) When $b = d$ then the claim holds by (v). So suppose $b < d$. Then we multiply both sides of $a < c$ by b, and both sides of $b < d$ by c, to obtain

$$ab < cb < cd.$$

(x) Note that $a^{-1} \neq 0$, since otherwise we would have $1 = aa^{-1} = a0 = 0$. Therefore if the claim does not hold we must have $a^{-1} < 0$. But this implies that $1 = aa^{-1} < a0 = 0$, which is a contradiction. So $a^{-1} > 0$.

(xi) Note that $a^{-1} \neq b^{-1}$, since $a \neq b$ and the inverse is unique. Also note that by (x) we have $a^{-1}, b^{-1} > 0$. Therefore if the claim does not hold we must have $0 < a^{-1} < b^{-1}$. But then by (ix) we get $1 = aa^{-1} < bb^{-1} = 1$, which is a contradiction. The last statement follows easily since $1^{-1} = 1$.

(xii) We either have $a < b$, $a = b$, or $a > b$. But we cannot have $a < b$, since we know that $a \geq b$. Similarly we cannot have $a > b$. Therefore we must have $a = b$.

(xiii) We have $a_i = -\sum_{j \neq i} a_j \leq 0$, hence $a_i = 0$. ■

Remark. The following version of the above theorem can be proved easily from it. We only need to consider some trivial cases in which the inequalities become equalities.

(i) If $a \le b$, $b \le c$ then $a \le c$.
(ii) $a \le b$ if and only if $b - a \ge 0$.
(iii) $a \ge 0$ if and only if $-a \le 0$.
(iv) $a^2 \ge 0$.
(v) $a \le b$ if and only if $a + c \le b + c$.
(vi) If $a \le c$, $b \le d$ then $a + b \le c + d$.
(vii) If $a \le b$ then

$$\begin{aligned} c \ge 0 &\implies ac \le bc, \\ c \le 0 &\implies ac \ge bc. \end{aligned}$$

(viii) If $a, b \ge 0$, or $a, b \le 0$, then $ab \ge 0$.
(ix) If $0 \le a \le c$, $0 \le b \le d$ then $ab \le cd$.

Theorem 1.7. *Suppose F is an ordered field, and $a, b, c \in F$. If $a > 0$, then the quadratic expression*

$$f(x) = ax^2 + bx + c$$

*is nonnegative for every $x \in F$ if and only if its **discriminant** $\Delta := b^2 - 4ac$ is nonnegative.*

Proof. First note that in a field F, the element n is defined to be $\overbrace{1 + 1 + \cdots + 1}^{n \text{ times}}$, where 1 is the identity of F. By the last proposition, in an ordered field we have $n > 0$, since $1 > 0$. Thus in particular, $n \neq 0$ in an ordered field.

Now for every $x \in F$ we have

$$\begin{aligned} ax^2 + bx + c &= a\big(x^2 + 2\frac{b}{2a}x + \frac{c}{a}\big) \\ &= a\big(x^2 + 2\frac{b}{2a}x + \frac{b^2}{4a^2} - \frac{b^2}{4a^2} + \frac{c}{a}\big) \\ &= a\Big(\big(x + \frac{b}{2a}\big)^2 - \frac{b^2 - 4ac}{4a^2}\Big) = a\Big(\big(x + \frac{b}{2a}\big)^2 + \frac{-\Delta}{4a^2}\Big). \end{aligned}$$

Note that the square of any element in an ordered field is nonnegative. Thus if $\Delta \le 0$ we have

$$\big(x + \frac{b}{2a}\big)^2 + \frac{-\Delta}{4a^2} = \big(x + \frac{b}{2a}\big)^2 + (-\Delta)(4a^2)^{-1} \ge 0.$$

Hence $f(x) \ge 0$. On the other hand, if $\Delta > 0$ then we have $f(-\frac{b}{2a}) = -\frac{\Delta}{4a} < 0$. Therefore if $f(x) \ge 0$ for every x, then we must have $\Delta \le 0$, as desired. ■

Theorem 1.8. *Let F be an ordered field. Then there exists a one-to-one function $\varphi : \mathbb{Q} \to F$ such that for every $p, q \in \mathbb{Q}$ we have*

(i) $\varphi(p+q) = \varphi(p) + \varphi(q)$*,*
(ii) $\varphi(pq) = \varphi(p)\varphi(q)$*,*
(iii) $p < q$ *if and only if* $\varphi(p) < \varphi(q)$*.*

Remark. Therefore F has a subset $\varphi(\mathbb{Q})$, which looks like $\mathbb{Q}$. Informally, we can say that F contains a "copy" of $\mathbb{Q}$.

Proof. For every $n \in \mathbb{N}$ let $\varphi(n) := \sum_{j \leq n} 1 \in F$, where 1 is the identity of F. Note that $\varphi(n) > 0$, since $1 > 0$ in F. Also, let $\varphi(0) := 0$, and $\varphi(-n) := -\varphi(n)$. Then it is easy to show by induction that for every $n, m \in \mathbb{Z}$ we have $\varphi(n+m) = \varphi(n) + \varphi(m)$, and $\varphi(nm) = \varphi(n)\varphi(m)$. Now suppose $\varphi(n) = \varphi(m)$. If $n > m$ then we have $\varphi(n-m) = \varphi(n) - \varphi(m) = 0$, which contradicts the fact that $\varphi(n-m) > 0$. Similarly, we cannot have $n < m$; so $n = m$. Thus φ is one-to-one on $\mathbb{Z}$. Next let us define $\varphi(p)$ for $p \in \mathbb{Q}$. We know that $p = \frac{m}{n}$ for some $n, m \in \mathbb{Z}$ with $n \neq 0$. We define

$$\varphi(p) := \frac{\varphi(m)}{\varphi(n)} \in F.$$

Note that $\varphi(n) \neq 0$, since $n \neq 0$ and φ is one-to-one on $\mathbb{Z}$. Now note that the value of $\varphi(p)$ does not depend on the representing fraction $\frac{m}{n}$. Because if we also have $p = \frac{m'}{n'}$ then $mn' = m'n$. Hence

$$\varphi(m)\varphi(n') = \varphi(mn') = \varphi(m'n) = \varphi(m')\varphi(n).$$

Therefore we get $\frac{\varphi(m)}{\varphi(n)} = \frac{\varphi(m')}{\varphi(n')}$. Note that $\varphi(n), \varphi(n') \neq 0$, since $n, n' \neq 0$. So, φ is a well-defined function from $\mathbb{Q}$ to F.

Next let us show that φ is one-to-one. Suppose $\varphi(\frac{m}{n}) = \varphi(\frac{k}{l})$. Then similarly to the above we get

$$\varphi(ml) = \varphi(m)\varphi(l) = \varphi(k)\varphi(n) = \varphi(kn).$$

Hence $ml = kn$, since φ is one-to-one on $\mathbb{Z}$. Thus $\frac{m}{n} = \frac{k}{l}$, as wanted. Now note that φ preserves the addition and multiplication of $\mathbb{Z}$. Therefore we get

$$\begin{aligned}
\varphi(\frac{m}{n} + \frac{k}{l}) &= \varphi(\frac{ml+kn}{nl}) = \frac{\varphi(ml+kn)}{\varphi(nl)} = \frac{\varphi(ml) + \varphi(kn)}{\varphi(n)\varphi(l)} \\
&= \frac{\varphi(m)\varphi(l) + \varphi(k)\varphi(n)}{\varphi(n)\varphi(l)} = \frac{\varphi(m)}{\varphi(n)} + \frac{\varphi(k)}{\varphi(l)} = \varphi(\frac{m}{n}) + \varphi(\frac{k}{l}), \\
\varphi(\frac{m}{n} \cdot \frac{k}{l}) &= \varphi(\frac{mk}{nl}) = \frac{\varphi(mk)}{\varphi(nl)} = \frac{\varphi(m)\varphi(k)}{\varphi(n)\varphi(l)} = \frac{\varphi(m)}{\varphi(n)} \cdot \frac{\varphi(k)}{\varphi(l)} = \varphi(\frac{m}{n}) \cdot \varphi(\frac{k}{l}).
\end{aligned}$$

Hence φ also preserves the addition and multiplication of $\mathbb{Q}$. Finally, suppose $\frac{m}{n} < \frac{k}{l}$, and $n, l > 0$. Then we have $ml < kn$; so $kn - ml > 0$. Therefore

$$0 < \varphi(kn - ml) = \varphi(k)\varphi(n) - \varphi(m)\varphi(l).$$

Also, $\varphi(n), \varphi(l) > 0$. Thus we get

$$\varphi(\frac{m}{n}) = \frac{\varphi(m)}{\varphi(n)} < \frac{\varphi(k)}{\varphi(l)} = \varphi(\frac{k}{l}),$$

as desired. Conversely, we can similarly show that $\frac{m}{n} \geq \frac{k}{l}$ implies $\varphi\big(\frac{m}{n}\big) \geq \varphi\big(\frac{k}{l}\big)$. Hence $\varphi\big(\frac{m}{n}\big) < \varphi\big(\frac{k}{l}\big)$ also implies that $\frac{m}{n} < \frac{k}{l}$. ■

Definition 1.9. Suppose F is an ordered field, and $S \subset F$. A function $f : S \to F$ is called **increasing** if for every $a, b \in S$ we have

$$a \leq b \implies f(a) \leq f(b).$$

Similarly, f is called **decreasing** if for every $a, b \in S$ we have

$$a \leq b \implies f(a) \geq f(b).$$

We also say f is **strictly increasing** or **strictly decreasing**, if $a < b$ implies $f(a) < f(b)$, or $f(a) > f(b)$ respectively. Finally, we say f is **monotone** if it is increasing or decreasing. And we say f is **strictly monotone**, if it is strictly increasing or strictly decreasing.

Definition 1.10. Suppose F is an ordered field and $S \subset F$. The set S is **bounded above** if there exists $M \in F$ such that for all $a \in S$ we have $a \leq M$. In this case we say M is an **upper bound** for S. Similarly, the set S is *bounded below* if there exists $m \in F$ such that for all $a \in S$ we have $a \geq m$. In this case we say M is a *lower bound* for S. We say S is **bounded** if it is bounded above and bounded below. Also, we say S is *unbounded* if it is not bounded, i.e. either it is not bounded above, or it is not bounded below.

Definition 1.11. If S is nonempty and bounded above, we say $c \in F$ is the **least upper bound** of S, if c is an upper bound for S, and if c is less than or equal to any upper bound for S, i.e.

$$\begin{aligned} &\forall a \in S \qquad a \leq c, \\ &\forall b \in F \qquad \text{If } \forall a \in S \ a \leq b \implies c \leq b. \end{aligned}$$

Similarly, we say $d \in F$ is the *greatest lower bound* of a nonempty bounded below set S, if d is a lower bound for S, and if d is greater than or equal to any lower bound for S, i.e.

$$\begin{aligned} &\forall a \in S \qquad a \geq d, \\ &\forall b \in F \qquad \text{If } \forall a \in S \ a \geq b \implies d \geq b. \end{aligned}$$

Remark. It is easy to see that the least upper bound and the greatest lower bound are unique, when they exist.

Definition 1.12. An ordered field F is **complete**, if every nonempty and bounded above subset $S \subset F$ has a least upper bound in F.

Theorem 1.13. *Let F be a complete ordered field. Suppose $S \subset F$ is nonempty and bounded below. Then S has a greatest lower bound in F.*

Proof. Let

$$T := \{-x : x \in S\}.$$

Then T is nonempty. Also, if z is a lower bound for S then $-z$ is an upper bound for T. So T is bounded above. Hence it has a least upper bound, which we call y. We claim that $-y$ is the greatest lower bound of S. First note that for every $x \in S$ we have $-x \in T$; so $-x \leq y$. Thus $x \geq -y$. So $-y$ is a lower bound for S. Now let w be a lower bound for S. Then $-w$ is an upper bound for T. Hence $-w \geq y$. Therefore $w \leq -y$. Thus $-y$ is the greatest lower bound of S, as desired. ■

***Example* 1.14.** $\mathbb{Q}$ is an ordered field which is not complete. For example

$$\{r \in \mathbb{Q} : r \geq 0,\ r^2 < 2\}$$

is a nonempty bounded above subset of $\mathbb{Q}$ which has no least upper bound in $\mathbb{Q}$. Because, as we will show in the proof of Theorem 1.46, if the above set has a supremum, its supremum must be a square root of 2. However, in Theorem 1.50 we will show that no rational number can be a square root of 2.

1.2 Real Numbers

In the field of rational numbers, we can perform the four basic arithmetic operations. Therefore rational numbers are adequate for many purposes. However, the field $\mathbb{Q}$ has some deficiencies. For example, the length of the diagonal of a square whose side length is one cannot be expressed by a rational number, i.e. by the ratio of two integers. In other words, there is no rational number p that satisfies $p^2 = 2$ (for the proof, see Theorem 1.50). Hence, in order to fill the gaps of the set of rational numbers, we need to extend it to a larger set, which is known as the set of real numbers.

But, what is a real number? For example, what is "square root of two"? It turns out that from a mathematical point of view, the inherent nature of "square root of two" is not very important. The important thing about "square root of two" is that it is a positive number whose square is two. Hence, we just need to find a set whose elements represent the real numbers, and have the properties that we expect from the real numbers; and we are not concerned with the inherent nature of real numbers.

The basic idea, due to Dedekind, is to consider a real number as the set of all rational numbers smaller than it. Because, intuitively we know that if two real numbers are different, then there is a rational number between them; so the set of rational numbers less than the larger real number is distinct from the set of rational numbers less than the smaller real number. Thus, a real number is uniquely determined by the set of rational numbers smaller than it.

Definition 1.15. A **Dedekind cut** is a pair (A, B), where $A, B \subset \mathbb{Q}$ satisfy the following conditions:

(i) $A \neq \emptyset$, $B \neq \emptyset$, $A \cap B = \emptyset$, and $A \cup B = \mathbb{Q}$.
(ii) If $a \in A$ and $b \in B$, then $a < b$.
(iii) A does not contain a largest element.

Remark. We denote a Dedekind cut (A, B) by $A|B$. Note that in a Dedekind cut $A|B$ we have $B = \mathbb{Q} - A$; so B is uniquely determined from A, and vice versa.

Remark. Also note that if $a \in A$ and $c \leq a$, then we must have $c \in A$. Since if $c \in B$ then we would have $a < c$, contrary to our assumption. Similarly, if $b \in B$ and $d \geq b$, then we must have $d \in B$.

Remark. Note that B may or may not have a smallest element.

Definition 1.16. A **real number** is a Dedekind cut. The set of all real numbers is

$$\mathbb{R} := \{x \in \mathcal{P}(\mathbb{Q}) \times \mathcal{P}(\mathbb{Q}) : x \text{ is a Dedekind cut}\},$$

where $\mathcal{P}(\mathbb{Q})$ is the power set of $\mathbb{Q}$.

Remark. Intuitively, we think of $\mathbb{R}$ as a line. Thus, a real number $x = A|B$ represents the cut in the line $\mathbb{R}$ at the point x, and A, B are the set of rational numbers in the two pieces of the line that remain after the cut.

Proposition 1.17. *Let $p \in \mathbb{Q}$. Also let*

$$A = \{r \in \mathbb{Q} : r < p\}, \qquad B = \{r \in \mathbb{Q} : r \geq p\}.$$

Then $A|B$ is a real number, i.e. a Dedekind cut.

Proof. First note that A, B are nonempty; because $\mathbb{Q}$ does not have a largest or smallest element. Also $A \cap B = \emptyset$, since we cannot have $r < p$ and $r \geq p$ for any r. In addition, note that $A \cup B = \mathbb{Q}$, since by trichotomy law we have $r < p$ or $r \geq p$ for every r. Furthermore, for $a \in A$ and $b \in B$ we have $a < p \leq b$; so $a < b$. Finally, note that A does not have a largest element. Because for every $a \in A$ we have $a < p$. Hence there is a rational number c such that $a < c < p$. Thus $c \in A$, and $c > a$. Therefore a cannot be the largest element of A. So $A|B$ is a Dedekind cut. ■

Now let us define the order of real numbers. If $x = A|B$ and $y = C|D$ are two real numbers, then intuitively we know that A consists of rational numbers less than x, and similarly C consists of rational numbers less than y. Hence if $x < y$ we must have $A \subset C$. In addition, we intuitively know that there are rational numbers between x, y; hence we also must have $A \neq C$. Thus we arrive at the following definition of order for real numbers.

Definition 1.18. Let $x = A|B$ and $y = C|D$ be two real numbers. Then we say $x < y$ if $A \subset C$ and $A \neq C$.

Remark. We know that by definition, $x \leq y$ means that $x < y$ or $x = y$. But $x = y$ is equivalent to $A = C$, since B, D are uniquely determined from A, C respectively. Therefore we have $x \leq y$ if and only if $A \subset C$.

Theorem 1.19. *The relation $<$ on $\mathbb{R}$ is a linear order. In addition, $\mathbb{R}$ does not have a smallest element, nor a largest element.*

Proof. Let $x = A|B$, $y = C|D$, and $z = E|F$. First note that $x \not< x$, since $A = A$. So $<$ is irreflexive. Now suppose $x < y$ and $y < z$. Then we have

$$A \subset C,\ A \neq C, \qquad C \subset E,\ C \neq E.$$

Hence we get $A \subset E$. We also have $A \neq E$, since otherwise we would get $C \subset A$, which contradicts the fact that $A \neq C$ and $A \subset C$. Thus $<$ is also transitive.

Next suppose $x \neq y$. Then $A \neq C$. We need to show that either $A \subset C$, or $C \subset A$; in order to conclude that either $x < y$, or $x > y$. Suppose $A \not\subset C$. Then there is $a \in A$ such that $a \notin C$. Hence we have $a \in D$. Thus for every $c \in C$ we must have $c < a$. Now note that if $c \in B$ then we cannot have $a \in A$, since $c < a$. Therefore $c \notin B$. Hence $c \in A$. Thus we have $C \subset A$, as desired. So $<$ is a linear order on $\mathbb{R}$.

Finally, let us show that for every $x = A|B$ there are $y, z \in \mathbb{R}$ such that $y < x < z$. Let $p \in A$ and $q \in B$. Let $y := C|D$ and $z := E|F$, where

$$\begin{aligned} C &= \{r \in \mathbb{Q} : r < p\}, & D &= \{r \in \mathbb{Q} : r \geq p\}, \\ E &= \{r \in \mathbb{Q} : r < q + 1\}, & F &= \{r \in \mathbb{Q} : r \geq q + 1\}. \end{aligned}$$

Then we have $A \subset E$, since for every $a \in A$ we have $a < q < q + 1$. Also, $A \neq E$; because $q \in E - A$. In addition, we have $C \subset A$, since if $r \in B$ then $r > p$; so for every $r < p$ we must have $r \in A$. Furthermore, $C \neq A$; because A must contain an element larger than p, since A does not have a largest element. So we have shown that $y < x < z$. Therefore no real number like x can be the smallest element, nor the largest element of $\mathbb{R}$. ■

Theorem 1.20. *Suppose $S \subset \mathbb{R}$ is nonempty and bounded above. Then S has a least upper bound in $\mathbb{R}$.*

Proof. Let

$$A := \{r \in \mathbb{Q} : r \in C \text{ for some } C|D \in S\}, \qquad B := \mathbb{Q} - A.$$

First note that $A|B$ is a cut. It is obvious that $A \cap B = \emptyset$, and $A \cup B = \mathbb{Q}$. It is also obvious that $A \neq \emptyset$, since S is nonempty. We also have $B \neq \emptyset$. To see this let $E|F$ be an upper bound for S. Then for every $C|D \in S$ we have $C \subset E$; so $A \subset E$. Thus $F \subset B$. Hence $B \neq \emptyset$. Now let $a \in A$ and $b \in B$. Then $a \in C$ for some $C|D \in S$. If $b \in C$ then $b \in A$, which contradicts our assumption. So $b \in D$. Thus we must have $a < b$. Finally, let $a \in A$. Then $a \in C$ for some $C|D \in S$. But we know that there is $c \in C$ such that $a < c$, since C does not have a largest element. However, we also have $c \in A$. Thus no element $a \in A$ can be the largest element of A. Therefore we have shown that $A|B$ is a cut.

Next let us show that $A|B$ is an upper bound for S. Let $C|D \in S$, and let $r \in C$. Then by definition we have $r \in A$. Hence $C \subset A$. Thus $C|D \leq A|B$, as desired. Now let $E|F$ be an upper bound for S. Let $r \in A$. Then there is $C|D \in S$ such that $r \in C$. However, we know that $C \subset E$. Thus we have $r \in E$. Hence $A \subset E$. Therefore $A|B \leq E|F$. So $A|B$ is the least upper bound of S. ■

Next, we have to define the addition and multiplication on $\mathbb{R}$, and show that they have the expected properties. Let $x = A|B, y = C|D \in \mathbb{R}$. Informally, these cuts represent the real numbers which are the supremum of the set of rational numbers in A, C respectively. Thus, at first glance, if we want to add and multiply x, y, the results should be cuts whose first components are respectively

$$\{a + c : a \in A, c \in C\},$$
$$\{ac : a \in A, c \in C\}.$$

Because if $a < x$ and $c < y$, then we have $a + c < x + y$. We also have $ac < xy$, provided that $a, c > 0$. However, if a, c are negative, then ac can be a large positive number greater than xy. Thus the definition of multiplication of real numbers needs more attention, which will be discussed later.

Furthermore, given a real number like $x = A|B$, we need to find its opposite and its inverse (provided that x is nonzero). Intuitively, we know that if the real number $B'|A'$ is the opposite of x, then B' consists of rational numbers s less than $-x$. However $s < -x$ if and only if $-s > x$. Therefore we must have $-s \in B$. In other words, the cut whose first component, i.e. B', consists of the opposite of the elements of B is our candidate for $-x$. But we must note that B can have a least element, and B' cannot have a largest element. Therefore when we construct B' from B we have to exclude the least element of B, if it exists. Hence we arrive at the following definition. The idea for finding the inverse of a nonzero real number is similar, and will be discussed later. First let us check that these proposed definitions for the sum and opposite are actually Dedekind cuts.

Proposition 1.21. *Let $x, y \in \mathbb{R}$, and suppose $x = A|B$, $y = C|D$. Let*

$$E = \{a + c : a \in A, c \in C\}, \qquad F = \mathbb{Q} - E.$$

Also let

$$B' = \{-b : b \in B,\ b \textit{ is not the smallest element of } B\}, \qquad A' = \mathbb{Q} - B'.$$

Then $E|F$ and $B'|A'$ are real numbers.

Proof. First note that E is nonempty, since A, C are nonempty. It is also obvious that $E \cap F = \emptyset$, and $E \cup F = \mathbb{Q}$. Now let $b \in B$ and $d \in D$. Then for every $a \in A$ and $c \in C$ we have $a < b$ and $c < d$. Thus $a + c < b + d$. So $b + d \notin E$. Hence $F \neq \emptyset$. Next let $a + c \in E$ and $r \in F$. If $r \leq a + c$ then $r - a \leq c$. Thus $r - a \in C$. Hence we have $r = a + (r - a) \in E$, which is a contradiction. Therefore we must have $a + c < r$. Finally, let $a + c \in E$. Then there are $p \in A$ and $q \in C$ such that $a < p$ and $c < q$; because A, C do not have largest elements. Now we have $p + q \in E$, and $a + c < p + q$. Thus no element $a + c$ can be the largest element of E.

Next lest us consider B', A'. Note that B' is nonempty. Because B is nonempty, and for every $b \in B$ we also have $b + 1 \in B$; so B has elements which are not its smallest element. It is also obvious that $B' \cap A' = \emptyset$, and $B' \cup A' = \mathbb{Q}$. In addition, for $a \in A$ we must have $-a \in A'$. Thus $A' \neq \emptyset$. Now let $b' \in B'$ and $a' \in A'$. Then $-b' \in B$. Also, $-a'$ either belongs to A, or is the smallest element of B (otherwise a' belongs to B'). In either case we have $-a' < -b'$. Hence $b' < a'$.

Finally, note that if $b' \in B'$ is its largest element, then $-b'$ must be less than or equal to every element of B except its smallest element (if B has a smallest element). Now if B does not have a smallest element, then $-b'$ is less than or equal to every element of B. Hence $-b'$ is the smallest element of B, contrary to our assumption. So suppose $p \in B$ is its smallest element. Then we have $p < -b'$. Let q be a rational number such that $p < q < -b'$. Then we have $q \in B$, since $p < q$. However, q is not the smallest element of B. So it cannot be smaller than $-b'$. Thus we have a contradiction, and therefore B' cannot have a largest element. ■

Definition 1.22. Let $x, y \in \mathbb{R}$, and suppose $x = A|B$, $y = C|D$. Then we define their **addition** to be $x + y := E|F$, where

$$E = \{a + c : a \in A, c \in C\}, \qquad F = \mathbb{Q} - E.$$

The **zero** and **identity** of $\mathbb{R}$ are

$$0 := \{r \in \mathbb{Q} : r < 0\} \,|\, \{r \in \mathbb{Q} : r \geq 0\},$$
$$1 := \{r \in \mathbb{Q} : r < 1\} \,|\, \{r \in \mathbb{Q} : r \geq 1\},$$

respectively. The **opposite** of x is $-x := B'|A'$, where

$$B' = \{-b : b \in B,\ b \text{ is not the smallest element of } B\}, \qquad A' = \mathbb{Q} - B'.$$

Remark. It is easy to see that in $\mathbb{R}$ we have $0 < 1$; because $\{r < 0\} \subset \{r < 1\}$, and

$$\frac{1}{2} \in \{r \in \mathbb{Q} : r < 1\} - \{r \in \mathbb{Q} : r < 0\}.$$

So in particular we have $0 \neq 1$.

Remark. Note that if $x = A|B$ and $x > 0$, then $\{r < 0\} \subsetneqq A$. Therefore there must be a nonnegative rational number r such that $r \in A$. However, A does not have a largest element. So 0 cannot be the only nonnegative rational number in A. Hence there must be a positive rational number r such that $r \in A$. In addition, note that every element of B is positive.

Theorem 1.23. *The addition of real numbers has the following properties: for every $x, y, z \in \mathbb{R}$ we have*

(i) *Associativity* :

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

(ii) *Commutativity* :

$$x + y = y + x.$$

(iii) *Identity element* :

$$x + 0 = x.$$

(iv) *Additive inverse* :

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

(v) *If $x < y$ then $x + z < y + z$.*

Proof. Let $x = A|B$, $y = C|D$, and $z = E|F$.

(i) Suppose $x + (y + z) = G|H$ and $(x + y) + z = I|J$. Then we have

$$\begin{aligned} r \in G &\iff \exists a \in A\ \exists c \in C\ \exists e \in E\ r = a + (c + e) \\ &\iff \exists a \in A\ \exists c \in C\ \exists e \in E\ r = (a + c) + e \iff r \in I. \end{aligned}$$

Hence $G = I$. Thus $x + (y + z) = (x + y) + z$.

(ii) Suppose $x + y = G|H$ and $y + x = I|J$. Then we have

$$\begin{aligned} r \in G &\iff \exists a \in A\ \exists c \in C\ r = a + c \\ &\iff \exists c \in C\ \exists a \in A\ r = c + a \iff r \in I. \end{aligned}$$

Hence $G = I$. Thus $x + y = y + x$.

(iii) Suppose $x + 0 = G|H$. Let $r \in G$. Then there are $a \in A$ and $c < 0$ such that $r = a + c$. Thus $r < a$. Hence $r \in A$. So $G \subset A$. Conversely, let $p \in A$. Then there is $q > p$ such that $q \in A$. Now we have $p = q + (p - q)$, and $p - q < 0$. Therefore $p \in G$. Hence $A \subset G$. So $A = G$. Thus we have $x + 0 = x$.

(iv) Suppose $-x = B'|A'$, and $x + (-x) = G|H$. Let $r \in G$. Then there are $a \in A$ and $b' \in B'$ such that $r = a + b'$. We know that $-b' \in B$. Therefore we have $a < -b'$. Hence $r = a + b' < 0$. Thus $G \subset \{r < 0\}$.

Conversely, suppose $r < 0$. We claim that there is $a \in A$ such that $a - r \in B$. Because otherwise for every $a \in A$ we would have $a - r \in A$. Hence we would also have $a - 2r = (a - r) - r \in A$. In fact, by induction we can show that $a - nr \in A$ for every $n \in \mathbb{N}$. But this leads to a contradiction, since we can make $a - nr$ larger than any rational number by taking n large enough (this is known as the Archimedean property of $\mathbb{Q}$). To see this let $b \in B$. Then $b > a$. We know that $-r, b - a$ are positive rational numbers. So we have $-r = m/k$ and $b - a = l/j$, for some $m, k, l, j \in \mathbb{N}$. Now for $n = kl$ we have $-nr = n(-r) = kl \times m/k = lm \geq l/j = b - a$, since $mj \geq 1$. Hence we would get $b \leq a - nr$, which implies that $b \in A$; and this is a contradiction.

Therefore there is $a \in A$ such that $a - r \in B$. Now we need $r - a = -(a - r)$ to be in B'. The only obstruction is that $a - r$ might be the smallest element of B. To solve this problem let $c \in A$ be such that $c > a$. Then we also have $c - r \in B$, since $a - r < c - r$. Let $b := c - r$. Then b is not the smallest element of B. Hence $-b \in B'$. Now we have

$$r = c + r - c = c + (-b) \in G.$$

Thus we have shown that $\{r < 0\} \subset G$. Therefore $G = \{r < 0\}$, and we have $x + (-x) = 0$, as desired.

(v) Suppose $x < y$, $x + z = G|H$, and $y + z = I|J$. Then we know that $A \subset C$. Let $r \in G$. Then we have $r = a + e$, for some $a \in A$ and $e \in E$. However, we also have $a \in C$. Hence $r \in I$ too. Thus $G \subset I$. So we have $x + z \leq y + z$. But $x + z = y + z$ implies that

$$\begin{aligned} x = x + 0 = x + (z + (-z)) &= (x + z) + (-z) \\ &= (y + z) + (-z) = y + (z + (-z)) = y + 0 = y; \end{aligned}$$

which contradicts our assumption. So we must have $x + z \neq y + z$. Therefore $x + z < y + z$, as desired. ■

Remark. As a consequence of the above theorem, for every $x, y \in \mathbb{R}$ we have

$$-(-x) = x, \qquad -(x + y) = (-x) + (-y).$$

These are proved in Theorem 1.3 for arbitrary fields; however, note that we did not use the multiplication of field in their proof. So, that proof works here too. Thus, as a trivial consequence we have $y = -x$ if and only if $-y = x$. In addition, we have

$$x < 0 \qquad \Longleftrightarrow \qquad -x > 0.$$

Because $x < 0 \implies x + (-x) < 0 + (-x) \implies 0 < -x$, and $0 < -x \implies 0 + x < (-x) + x \implies x < 0$.

The next step is to define multiplication and inverse, and to conclude their properties. Let $x = A|B, y = C|D \in \mathbb{R}$. We have seen that a suitable candidate for the product of x, y should be a cut whose first component is

$$\{ac : a \in A, c \in C\}.$$

Because if $a < x$ and $c < y$, then we have $ac < xy$, provided that $a, c > 0$. However as we noted before, a, c can be negative, and consequently ac can be a large positive number greater than xy. In order to overcome this difficulty, we first assume that $x, y > 0$. Then we can represent the positive rational numbers less than xy by ac, where $a, c > 0$. And to construct the first component of the Dedekind cut of xy we also include all nonpositive rational numbers, since they are all less than xy. Finally we can extend the definition of multiplication to all real numbers by taking the opposites, and reducing the general case to the case of positive real numbers, as explained below.

In addition, to find the inverse of a given nonzero real number like $x = A|B$, we first assume $x > 0$. Intuitively, we know that if the real number $\tilde{B}|\tilde{A}$ is the inverse of x, then $\tilde{B}$ consists of rational numbers s less than x^{-1}. However for $s > 0$ we have $s < x^{-1}$ if and only if $s^{-1} > x$. Therefore we must have $s^{-1} \in B$. In other words, the cut whose first component, i.e. $\tilde{B}$, consists of nonpositive rational numbers together with the inverse of the elements of B (which are all positive, since they are not less than x) is our candidate for x^{-1}. But we must note that B can have a least element, and $\tilde{B}$ cannot have a largest element. Therefore when we construct $\tilde{B}$ from B we have to exclude the least element of B, if it exists. Hence we arrive at the following definition. For the inverse of negative real numbers we can take their opposites and use the definition of inverse of positive numbers, as explained below. But first let us check that these proposed definitions for the product and inverse of positive numbers are actually Dedekind cuts.

Proposition 1.24. *Let $x, y \in \mathbb{R}$, and suppose $x = A|B$, $y = C|D$. Suppose $x, y > 0$. Let*

$$G = \{r \in \mathbb{Q} : r \le 0\} \cup \{ac : a \in A, c \in C,\ a, c > 0\}, \qquad H = \mathbb{Q} - G.$$

Also let

$$\tilde{B} = \{r \in \mathbb{Q} : r \le 0\} \cup \{b^{-1} : b \in B,\ b \textit{ is not the smallest element of } B\},$$
$$\tilde{A} = \mathbb{Q} - \tilde{B}.$$

Then $G|H$ and $\tilde{B}|\tilde{A}$ are real numbers.

Remark. Note that when $x > 0$, A contains a positive rational number; so every element of B is positive, since every element of B is greater than every element of A. Thus the definition of $\tilde{B}$ makes sense.

Proof. It is obvious that G is nonempty, $G \cap H = \emptyset$, and $G \cup H = \mathbb{Q}$. Now let $b \in B$ and $d \in D$. Let $a \in A$ and $c \in C$ be positive. We know that $a < b$ and $c < d$. Hence we have $ac < bd$. So $bd \notin G$. Thus $H \neq \emptyset$. Next let $g \in G$ and $h \in H$. Note that by definition every element of H is positive. Thus if $g \leq 0$ then $g < h$. So suppose $g = ac > 0$. If $h \leq ac$ then $h/a \leq c$. Thus $h/a \in C$. Hence we have $h = a \times h/a \in G$, which is a contradiction. Therefore we must have $ac < h$. Finally, let ac be a positive element of G. Then there are $p \in A$ and $q \in C$ such that $a < p$ and $c < q$; because A, C do not have largest elements. Now we have $pq \in G$, and $ac < pq$. Thus no element ac can be the largest element of G. It is also obvious that a nonpositive element of G cannot be its largest element.

Next lest us consider $\tilde{B}, \tilde{A}$. Note that all the elements of B are positive, since $x > 0$. It is obvious that $\tilde{B}$ is nonempty, $\tilde{B} \cap \tilde{A} = \emptyset$, and $\tilde{B} \cup \tilde{A} = \mathbb{Q}$. In addition, for a positive $a \in A$ we have $a^{-1} > 0$; therefore $a^{-1} \in \tilde{A}$. Thus $\tilde{A} \neq \emptyset$. Now let $b' \in \tilde{B}$ and $a' \in \tilde{A}$. Note that by definition every element of $\tilde{A}$ is positive. Thus if $b' \leq 0$ then $b' < a'$. So suppose $b' > 0$. Then $b'^{-1} \in B$. Also, a'^{-1} either belongs to A, or is the smallest element of B (otherwise a' belongs to $\tilde{B}$). In either case we have $a'^{-1} < b'^{-1}$. Hence $b' < a'$, since $a', b' > 0$.

Finally, note that if $b' \in \tilde{B}$ is its largest element, then b' must be positive. So $b'^{-1} \in B$, and it must be less than or equal to every element of B except its smallest element (if B has a smallest element). Now if B does not have a smallest element, then b'^{-1} is less than or equal to every element of B. Hence b'^{-1} is the smallest element of B, contrary to our assumption. So suppose $p \in B$ is its smallest element. Then we have $p < b'^{-1}$. Let q be a rational number such that $p < q < b'^{-1}$. Then we have $q \in B$, since $p < q$. However, q is not the smallest element of B. So it cannot be smaller than b'^{-1}. Thus we have a contradiction, and therefore $\tilde{B}$ cannot have a largest element. ■

Definition 1.25. Let $x, y \in \mathbb{R}$, and suppose $x = A|B$, $y = C|D$. When $x, y > 0$, we define the **multiplication** of x, y to be $xy := G|H$, where

$$G = \{r \in \mathbb{Q} : r \leq 0\} \cup \{ac : a \in A, c \in C,\ a, c > 0\}, \qquad H = \mathbb{Q} - G.$$

In other cases, we define

(i) When $x < 0$ and $y > 0$, $xy := -((-x)y)$.
(ii) When $x > 0$ and $y < 0$, $xy := -(x(-y))$.
(iii) When $x < 0$ and $y < 0$, $xy := (-x)(-y)$.
(iv) For every $x \in \mathbb{R}$, $x0 := 0$ and $0x := 0$.

Suppose $x \neq 0$. When $x > 0$ the **inverse** of x is $x^{-1} := \tilde{B}|\tilde{A}$, where

$$\begin{aligned}&\tilde{B} = \{r \in \mathbb{Q} : r \leq 0\} \cup \{b^{-1} : b \in B,\ b \text{ is not the smallest element of } B\},\\ &\tilde{A} = \mathbb{Q} - \tilde{B}.\end{aligned}$$

And when $x < 0$ we define $x^{-1} := -(-x)^{-1}$.

Remark. Note that when $x, y > 0$ we have $xy > 0$. Hence if one of the x, y is positive, and the other one is negative, we have $xy < 0$; and if both x, y are negative we have $xy > 0$. Also note that when $x > 0$ we have $x^{-1} > 0$. Thus when $x < 0$ we have $x^{-1} < 0$.

Theorem 1.26. *The multiplication of real numbers has the following properties: for every $x, y, z \in \mathbb{R}$ we have*

(i) *Associativity* :
$$x(yz) = (xy)z.$$

(ii) *Commutativity* :
$$xy = yx.$$

(iii) *Identity element* :
$$x1 = x.$$

(iv) *Multiplicative inverse* :
$$x \neq 0 \implies xx^{-1} = 1.$$

(v) *Distributivity* :
$$x(y + z) = xy + xz.$$

(vi) *If $x > 0$ and $y > 0$, then $xy > 0$.*

Remark. This theorem and the previous theorem show that $\mathbb{R}$ is an ordered field.

Proof. Let $x = A|B$, $y = C|D$, and $z = E|F$.

(i) Suppose $x(yz) = G|H$ and $(xy)z = I|J$. If one of x, y, z is zero, then both $x(yz), (xy)z$ are zero; so $x(yz) = (xy)z$. So we can assume that x, y, z are all nonzero. First suppose $x, y, z > 0$. Then we have

$$\begin{aligned}r \in G &\iff r \leq 0 \ \text{ or } \ \exists a \in A\ \exists c \in C\ \exists e \in E\ a, c, e > 0 \text{ and } r = a(ce)\\ &\iff r \leq 0 \ \text{ or } \ \exists a \in A\ \exists c \in C\ \exists e \in E\ a, c, e > 0 \text{ and } r = (ac)e\\ &\iff r \in I.\end{aligned}$$

Hence $G = I$. Thus $x(yz) = (xy)z$. Now we have (Note that the sign of xy and yz, which are needed in the following computations, can be determined from the sign of x, y, z.)

$$\begin{aligned}
x<0, y>0, z>0 \implies x(yz) &= -[(-x)(yz)] = -[((-x)y)z] \\
&= -[(-(xy))z] = (xy)z, \\
x>0, y<0, z>0 \implies x(yz) &= -[x(-(yz))] = -[x((-y)z)] \\
&= -[(x(-y))z] = -[(-(xy))z] = (xy)z, \\
x<0, y<0, z>0 \implies x(yz) &= (-x)(-(yz)) = (-x)((-y)z)) \\
&= ((-x)(-y))z = (xy)z, \\
x>0, y>0, z<0 \implies x(yz) &= -[x(-(yz))] = -[x(y(-z))] \\
&= -[(xy)(-z)] = (xy)z, \\
x<0, y>0, z<0 \implies x(yz) &= (-x)(-(yz)) = (-x)(y(-z)) \\
&= ((-x)y)(-z) = (-(xy))(-z) = (xy)z, \\
x>0, y<0, z<0 \implies x(yz) &= x((-y)(-z)) = (x(-y))(-z) \\
&= (-(xy))(-z) = (xy)z, \\
x<0, y<0, z<0 \implies x(yz) &= -[(-x)(yz)] = -[(-x)((-y)(-z))] \\
&= -[((-x)(-y))(-z)] = -[(xy)(-z)] = (xy)z.
\end{aligned}$$

(ii) Suppose $xy = G|H$ and $yx = I|J$. If one of x, y is zero, then both xy, yx are zero; so $xy = yx$. So we can assume that both x, y are nonzero. First suppose $x, y > 0$. Then we have

$$\begin{aligned}
r \in G &\iff r \le 0 \text{ or } \exists a \in A\ \exists c \in C\ a, c > 0 \text{ and } r = ac \\
&\iff r \le 0 \text{ or } \exists c \in C\ \exists a \in A\ c, a > 0 \text{ and } r = ca \iff r \in I.
\end{aligned}$$

Hence $G = I$. Thus $xy = yx$. Now we have

$$\begin{aligned}
x<0, y>0 &\implies xy = -((-x)y) = -(y(-x)) = yx, \\
x>0, y<0 &\implies xy = -(x(-y)) = -((-y)x) = yx, \\
x<0, y>0 &\implies xy = (-x)(-y) = (-y)(-x) = yx.
\end{aligned}$$

(iii) Suppose $x1 = G|H$. If $x = 0$ we have $x1 = 0 \times 1 = 0 = x$. Let us assume $x > 0$. Let $r \in G$. If $r \le 0$ then we have $r \in A$, since $x > 0$. So suppose $r > 0$. Then there are $a \in A$ and $c < 1$ such that $a, c > 0$, and $r = ac$. Hence we have $r = ac < a1 = a$. Thus $r \in A$. So $G \subset A$. Conversely, let $p \in A$. If $p \le 0$ then $p \in G$ by definition. So suppose $p > 0$. Then there is $q > p$ such that $q \in A$. Now we have $p = q \times p/q$, and $0 < p/q < 1$. Therefore $p \in G$. Hence $A \subset G$. So $A = G$. Thus we have $x1 = x$. Finally, let us assume $x < 0$. Then we have

$$x1 = -((-x)1) = -(-x) = x,$$

since $-x, 1 > 0$.

(iv) Suppose $x \neq 0$, $x^{-1} = \tilde{B}|\tilde{A}$, and $xx^{-1} = G|H$. First let us assume that $x > 0$. Note that in this case all the elements of B are positive. Let $r \in G$. If $r \leq 0$ then $r < 1$. So suppose $r > 0$. Then there are $a \in A$ and $b' \in \tilde{B}$ such that $a, b' > 0$, and $r = ab'$. We know that $b'^{-1} \in B$. Therefore we have $a < b'^{-1}$. Hence $r = ab' < 1$, since $b' > 0$. Thus $G \subset \{r < 1\}$.

Conversely, suppose $r < 1$. If $r \leq 0$ then $r \in G$ by definition. So suppose $r > 0$. Note that A contains some positive rational numbers, since $x > 0$. We claim that there is $a \in A$ such that $a > 0$, and $a/r \in B$. Because otherwise for every positive $a \in A$ we would have $a/r \in A$. Hence we would also have $a/r^2 = (a/r)/r \in A$, since a/r is positive too. In fact, by induction we can show that $a/r^n \in A$ for every $n \in \mathbb{N}$. But this leads to a contradiction, since we can make a/r^n larger than any rational number by taking n large enough (this also follows from the Archimedean property of $\mathbb{Q}$, but we have to convert the multiplicative form a/r^n to the additive form $a + nd$ for some positive rational number d). To see this note that $r < 1$, so $s := 1 - r > 0$. It is easy to show by induction that

$$\frac{1}{(1-s)^n} \geq 1 + ns.$$

Because for $n = 0$ both sides are 1. And if the inequality holds for some n, then for $n + 1$ we have

$$\frac{1}{(1-s)^{n+1}} = \frac{1}{1-s}\frac{1}{(1-s)^n} \geq \frac{1+ns}{1-s} \geq 1 + (n+1)s,$$

because we have

$$(1-s)(1+(n+1)s) = 1 - s + (n+1)s - (n+1)s^2 \leq 1 + ns,$$

since $1 - s, (n+1)s^2$ are positive. Therefore we get

$$\frac{a}{r^n} = \frac{a}{(1-s)^n} \geq a(1+ns) = a + nas.$$

Let $b \in B$. Then $b > a$. We know that $as, b - a$ are positive rational numbers. So we have $as = m/k$ and $b - a = l/j$, for some $m, k, l, j \in \mathbb{N}$. Now for $n = kl$ we have $nas = kl \times m/k = lm \geq l/j = b - a$, since $mj \geq 1$. Thus we get

$$b \leq a + nas \leq \frac{a}{r^n}.$$

However, this implies that $b \in A$, which is a contradiction.

Therefore there is a positive $a \in A$ such that $a/r \in B$. Now we need $r/a = (a/r)^{-1}$ to be in $\tilde{B}$. The only obstruction is that a/r might be the smallest element of B. To solve this problem let $c \in A$ be such that $c > a$. Then we also have $c/r \in B$, since $a/r < c/r$. Let $b := c/r$. Then b is not the smallest element of B. Hence $b^{-1} \in \tilde{B}$. Now we have

$$r = c \times \frac{r}{c} = c \times \big(\frac{c}{r}\big)^{-1} = cb^{-1} \in G.$$

Thus we have shown that $\{r < 1\} \subset G$. Therefore $G = \{r < 1\}$, and we have $xx^{-1} = 1$, as desired. Finally, suppose $x < 0$. Then $-x > 0$, and by definition we have $x^{-1} = -(-x)^{-1}$. Hence $x^{-1} < 0$ too, and we have $-x^{-1} = -(-(-x)^{-1}) = (-x)^{-1}$. Thus we get

$$xx^{-1} = (-x)(-x^{-1}) = (-x)(-x)^{-1} = 1,$$

as desired.

(v) Suppose $x(y+z) = G|H$, $xy + xz = I|J$, $y + z = K|K'$, $xy = L|L'$, and $xz = M|M'$. If $x = 0$ then

$$x(y+z) = 0 = 0 + 0 = xy + xz.$$

And if one of y, z is zero, say y is zero, then

$$x(y+z) = x(0+z) = xz = 0 + xz = xy + xz.$$

So we can assume that x, y, z are all nonzero. First suppose $x, y, z > 0$. Note that $y + z > 0 + z = z > 0$. In addition, note that $xy, xz > 0$. Hence we can similarly show that $xy + xz > 0$. Let $r \in G$. If $r \le 0$ then $r \in I$, since $xy + xz > 0$. So suppose $r > 0$. Then there are $a \in A$ and $k \in K$ such that $a, k > 0$, and $r = ak$. On the other hand, there are $c \in C$ and $e \in E$ such that $k = c + e$. Thus we have $r = ac + ae$. If $c \le 0$ then $ac \le 0$; so $ac \in L$, since $xy > 0$. And if $c > 0$ then $ac \in L$ by definition of xy. Similarly, we have $ae \in M$. Therefore $r = ac + ae \in I$. Hence $G \subset I$.

Conversely, let $r \in I$. Then we know that $r = l + m$, for some $l \in L$ and $m \in M$. Suppose $l, m > 0$. Then there are positive numbers $a, a' \in A$, $c \in C$, and $e \in E$ such that $l = ac$, and $m = a'e$. Without loss of generality we can assume $a' \le a$. Then we have

$$r = l + m = ac + a'e \le ac + ae = a(c+e).$$

However, $c + e \in K$, and $c + a > 0$; so $a(c+e) \in G$. Therefore we also have $r \in G$. Next suppose one of l, m is positive and the other one is nonpositive, say $l > 0, m \le 0$. Then there are positive numbers $a \in A$ and $c \in C$ such that $l = ac$. Let $e \in E$ be positive. Then we have

$$r = l + m \le l = ac \le ac + ae = a(c+e);$$

which implies $r \in G$. Finally, if $l, m \leq 0$ then $r \leq 0$. Thus by definition of G we have $r \in G$. Therefore we have $I \subset G$. Hence $I = G$, and we get $x(y+z) = xy+xz$, as desired.

Next suppose $x > 0$, $y < 0$, and $z > 0$. Then $-y > 0$. We have to consider two cases. If $y + z \geq 0$ we have

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

Hence we get $x(y+z) = xy+xz$ by adding xy to both sides. And if $y+z < 0$ then $-(y+z) > 0$. Thus we have

$$\begin{aligned} -x(y+z) + xz &= x(-(y+z)) + xz \\ &= x(-(y+z)+z) = x(-y-z+z) = x(-y) = -xy. \end{aligned}$$

Hence we get $x(y+z) = xy+xz$ by adding $xy, x(y+z)$ to both sides. Now suppose $x > 0$, $y > 0$, and $z < 0$. Then we can switch y, z and use the previous case to obtain

$$x(y+z) = x(z+y) = xz + xy = xy + xz.$$

Next suppose $x > 0$ and $y, z < 0$. Then we have $y + z < 0$. Hence

$$\begin{aligned} x(y+z) = -[x(-(y+z))] &= -[x((-y)+(-z))] = -[x(-y)+x(-z)] \\ &= -[(-xy)+(-xz)] = -[-(xy+xz)] = xy+xz. \end{aligned}$$

Finally, suppose $x < 0$. Let us first show that for every $y \in \mathbb{R}$ we have

$$(-x)y = -xy. \tag{$*$}$$

If $y = 0$ then both sides of the above equation are 0. If $y > 0$ then by definition we have $xy = -((-x)y)$. Thus $-xy = (-x)y$. And if $y < 0$ then by definition we have $xy = (-x)(-y)$. We also have $(-x)y = -((-x)(-y))$, since $-x > 0$. Hence we get $(-x)y = -xy$, as desired. Now by the previous paragraph, for every $y, z \in \mathbb{R}$ we have

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

since $-x > 0$. Hence by $(*)$ we get

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

Thus we obtain $x(y+z) = xy+xz$; because the additive inverse of every number is unique.

(vi) As we have noted before, this is a trivial consequence of the definition of multiplication. ■

Although $\mathbb{R}$ does not contain $\mathbb{Q}$ as a subset, it has a subset which looks like $\mathbb{Q}$. Informally, we can say that $\mathbb{R}$ contains a "copy" of $\mathbb{Q}$. The next theorem shows that this subset of $\mathbb{R}$ behaves similarly to $\mathbb{Q}$.

Theorem 1.27. *Let $E : \mathbb{Q} \to \mathbb{R}$ be defined as*

$$E(p) = \{r \in \mathbb{Q} : r < p\} \,|\, \{r \in \mathbb{Q} : r \geq p\},$$

for every $p \in \mathbb{Q}$. Then E is a one-to-one function, and for every $p, q \in \mathbb{Q}$ we have
(i) $E(p+q) = E(p) + E(q)$.
(ii) $E(pq) = E(p)E(q)$.
(iii) $p < q$ *if and only if* $E(p) < E(q)$.

Proof. First note that E is one-to-one. Because if $E(p) = E(q)$ then we have $\{r \geq p\} = \{r \geq q\}$. But both p, q are the minimum of this set. So we must have $p = q$.

(i) Suppose $E(p) + E(q) = A|B$. Let $r \in A$. Then there are $a < p$ and $b < q$ such that $r = a + b$. Hence we have $r = a + b < p + q$. On the other hand, suppose $r < p + q$. Then $r - p < q$. So there is b such that $r - p < b < q$. Hence we have $r - b < p$. Thus

$$r = r - b + b \in A.$$

Therefore we have $A = \{r < p + q\}$, as desired.

(ii) Note that by definition we have $E(0) = 0$. Thus for every p we have

$$-E(p) = E(-p),$$

since $E(p) + E(-p) = E(p + (-p)) = E(0) = 0$. Now let $E(p)E(q) = C|D$. If one of p, q is zero, then both $E(pq)$ and $E(p)E(q)$ are zero. So suppose that both p, q are nonzero. First let us assume that $p, q > 0$. Then by the next part we have $E(p), E(q) > E(0) = 0$. Let $r \in C$. If $r \leq 0$ then we obviously have $r < pq$. So suppose $r > 0$. Then there are $a < p$ and $b < q$ such that $a, b > 0$, and $r = ab$. Thus we have $r = ab < pq$. On the other hand, suppose $r < pq$. If $r \leq 0$ then by definition we have $r \in C$. So suppose $r > 0$. Then we have $r/p < q$. So there is b such that $r/p < b < q$. Hence we have $r/b < p$, since $b > 0$. Thus

$$r = \frac{r}{b} \times b \in C.$$

Therefore we have $C = \{r < pq\}$, as desired.

Next suppose $p > 0$ and $q < 0$. Then $E(q) < 0$. Hence we have

$$\begin{aligned} E(p)E(q) &= -[E(p)(-E(q))] = -[E(p)E(-q)] \\ &= -E(p(-q)) = -E(-pq) = -(-E(pq)) = E(pq). \end{aligned}$$

Now suppose $p < 0$ and $q > 0$. Then we can switch p, q and use the previous case to obtain

$$E(p)E(q) = E(q)E(p) = E(qp) = E(pq).$$

Finally, suppose $p, q < 0$. Then $E(p), E(q) < 0$. Hence we have

$$E(p)E(q) = (-E(p))(-E(q)) = E(-p)E(-q) = E((-p)(-q)) = E(pq).$$

(iii) If $p < q$ then we clearly have $\{r < p\} \subset \{r < q\}$. In addition, we have $p \in \{r < q\} - \{r < p\}$. Hence $\{r < p\} \subsetneqq \{r < q\}$. Thus $E(p) < E(q)$. Conversely, if $E(p) < E(q)$ then we must have $p < q$. Because if $p \ge q$ then we have shown that $E(p) \ge E(q)$; and this contradicts our assumption. ■

Definition 1.28. A real number x is called **rational** if there is $p \in \mathbb{Q}$ such that $x = E(p)$. In other words, if $x = A|B$ and

$$A = \{r \in \mathbb{Q} : r < p\}, \qquad B = \{r \in \mathbb{Q} : r \ge p\}.$$

A real number which is not rational is called **irrational**. A real number x is an **integer**, if $x = E(p)$ for some $p \in \mathbb{Q}$, where p is an integer.

Remark. If x is a rational real number as above, we identify it with the rational number p. Note that this identification is mostly harmless, since by Theorem 1.27, the addition, multiplication, and order relation will be preserved under this identification.

***Example* 1.29.** An example of an irrational number is $x = A|B$ where

$$A = \{r \in \mathbb{Q} : r < 0 \text{ or } r^2 < 2\}, \qquad B = \{r \in \mathbb{Q} : r > 0 \text{ and } r^2 \ge 2\}.$$

The number x is actually $\sqrt{2}$, and we will see later that it is indeed irrational.

1.3 More about Real Numbers

In the last section, we have shown that $\mathbb{R}$ is a complete ordered field. In fact, up to isomorphism, $\mathbb{R}$ is the only complete ordered field, i.e. every complete ordered field is essentially the same as $\mathbb{R}$. Intuitively, this means that every complete ordered field can be obtained from $\mathbb{R}$ by renaming its elements. The next theorem states this fact precisely.

Theorem 1.30. *Let F be a complete ordered field. Then F is **isomorphic** to $\mathbb{R}$, i.e. there exists a one-to-one and onto function $\varphi : \mathbb{R} \to F$ such that for every $x, y \in \mathbb{R}$ we have*

(i) $\varphi(x + y) = \varphi(x) + \varphi(y)$,

(ii) $\varphi(xy) = \varphi(x)\varphi(y)$,

(iii) $x < y$ *if and only if* $\varphi(x) < \varphi(y)$.

Proof. In Theorem 1.8 we have shown that there exists a one-to-one function $\varphi : \mathbb{Q} \to F$ that satisfies the above three conditions. Also remember that the image of φ is the set of "rationals" $\frac{m}{n}$ in F, where m, n are the "integers" in F, i.e. they are constructed by adding the identity of F to itself. Furthermore, since F is complete, we can show that between any two of its elements there is a "rational" element. The proof is similar to the proof of the same property for $\mathbb{R}$, as will be demonstrated later in this section. In this proof, r, s, p denote rational numbers.

Now we extend φ to all of $\mathbb{R}$. For $x \in \mathbb{R}$ let

$$A_x := \{\varphi(r) : r \in \mathbb{Q} \text{ and } r < x\}.$$

Then A_x is nonempty and bounded above, since for some rational number $s > x$ we have $\varphi(s) > \varphi(r)$, for every rational number $r < x$. Now let

$$\hat{\varphi}(x) := \sup A_x.$$

Note that $\sup A_x$ exists, because F is complete. First let us show that $\hat{\varphi}(r) = \varphi(r)$ for every $r \in \mathbb{Q}$. Note that for every rational number $p < r$ we have $\varphi(p) < \varphi(r)$; so $\varphi(r)$ is an upper bound for A_r. Hence $\hat{\varphi}(r) \le \varphi(r)$. Suppose to the contrary that $\hat{\varphi}(r) < \varphi(r)$. Then there is a rational element $\varphi(s) \in F$ such that $\hat{\varphi}(r) < \varphi(s) < \varphi(r)$. But then we must have $s < r$; so $s \in A_r$. However, then we get $\varphi(s) \le \hat{\varphi}(r)$, which is a contradiction. Therefore $\hat{\varphi}(r) = \varphi(r)$. Hence $\hat{\varphi}$ is an extension of φ. So in the rest of this proof, we will simply denote $\hat{\varphi}$ by φ.

Next let us show that φ is onto. Let $a \in F$, and let

$$\tilde{A}_a := \{r : r \in \mathbb{Q} \text{ and } \varphi(r) < a\}.$$

Then $\tilde{A}_a$ is a nonempty and bounded above subset of $\mathbb{R}$. Because there is a "rational" element $a - 1 < \varphi(p) < a$. And for $\varphi(s) > a$ we have $r < s$ for every $r \in \tilde{A}_a$. Let $x := \sup \tilde{A}_a$. We claim that $a = \varphi(x)$. Note that if $\varphi(s) \ge a$ then $\varphi(s)$ is an upper bound for $\varphi(\tilde{A}_a)$; so s is an upper bound for $\tilde{A}_a$. Hence we get $s \ge x$, since x is the least upper bound of $\tilde{A}_a$. Therefore if $r < x$ then $\varphi(r) < a$. Thus a is an upper bound for A_x. So $\varphi(x) = \sup A_x \le a$. Suppose to the contrary that $\varphi(x) < a$. Then there is $\varphi(x) < \varphi(r) < a$. Hence $r \in \tilde{A}_a$. So we must have $r \le x$. But $r \ne x$, since $\varphi(x) \ne \varphi(r)$. On the other hand, $r < x$ implies that $\varphi(r) \in A_x$. Hence we get $\varphi(r) \le \varphi(x)$, which is a contradiction. Thus $\varphi(x) = a$. So φ is onto.

Now suppose $x < y$. Then we have $A_x \subset A_y$. Hence $\varphi(x) \le \varphi(y)$. Also, we know that there are rational numbers $x < r < s < y$. So $\varphi(s) \in A_y$, and $\varphi(r)$ is an upper bound for A_x. Therefore $\varphi(x) \le \varphi(r) < \varphi(s) \le \varphi(y)$. Thus $\varphi(x) < \varphi(y)$.

Hence φ is one-to-one. In addition, $\varphi(x) < \varphi(y)$ also implies $x < y$; because otherwise we would have $x \geq y$, which implies $\varphi(x) \geq \varphi(y)$. Next let us show that

$$\varphi(x+y) = \varphi(x) + \varphi(y).$$

If this does not happen, then we either have $\varphi(x+y) < \varphi(x)+\varphi(y)$, or $\varphi(x+y) > \varphi(x) + \varphi(y)$. Suppose the latter inequality holds; the other case can be treated similarly. Then there is a rational number r such that

$$\varphi(x+y) > \varphi(r) > \varphi(x) + \varphi(y).$$

Hence we get $r < x+y$; so $r - y < x$. Thus there is also a rational number s such that $r - y < s < x$. Then we have $r - s < y$. Hence we have $\varphi(s) < \varphi(x)$, and $\varphi(r-s) < \varphi(y)$. Therefore we obtain

$$\varphi(x) + \varphi(y) > \varphi(s) + \varphi(r-s) = \varphi(s+r-s) = \varphi(r),$$

which is a contradiction.

Finally, let us show that $\varphi(xy) = \varphi(x)\varphi(y)$. Note that since φ preserves addition, we have $\varphi(-x) = -\varphi(x)$. We also know that $\varphi(0) = 0$. Hence it suffices to check the equality $\varphi(xy) = \varphi(x)\varphi(y)$ for $x, y > 0$. If the equality does not hold, then either $\varphi(xy) < \varphi(x)\varphi(y)$, or $\varphi(xy) > \varphi(x)\varphi(y)$. Suppose the former inequality holds; the other case can be treated similarly. Then there is a rational number r such that

$$\varphi(xy) < \varphi(r) < \varphi(x)\varphi(y).$$

Hence we get $0 < xy < r$; so $x < \frac{r}{y}$. Thus there is also a rational number s such that $0 < x < s < \frac{r}{y}$. Then we have $y < \frac{r}{s}$. Hence we have $\varphi(x) < \varphi(s)$, and $\varphi(y) < \varphi(\frac{r}{s})$. Therefore we obtain

$$\varphi(x)\varphi(y) < \varphi(s)\varphi(\tfrac{r}{s}) = \varphi(s\tfrac{r}{s}) = \varphi(r),$$

which is a contradiction. ■

Although we have constructed real numbers by using Dedekind cuts, in practice we do not think of a real number as a pair of sets of rational numbers. Rather, as it is common, we think of real numbers as the points of a line. And when we want to prove something about real numbers, we will use the fact that $\mathbb{R}$ is a complete ordered field.

The Extended Real Number System. There does not exist a largest or smallest real number, but we consider the so-called "infinite real numbers" $+\infty$ and $-\infty$. They are called **(positive) infinity** and **negative infinity** respectively. We also denote $+\infty$ simply by ∞. Technically, $\pm\infty$ are two distinct objects that are different from all real numbers. In contrast to infinities, the elements of $\mathbb{R}$ are called "finite real numbers". Also, the set

$$\mathbb{R} \cup \{+\infty, -\infty\}$$

is called "the extended real number system". We extend the order of $\mathbb{R}$ to $\mathbb{R} \cup \{+\infty, -\infty\}$, so that for all $x \in \mathbb{R}$ we have

$$-\infty < x < +\infty.$$

We also define

$$\begin{aligned}
&\pm \infty + x = x + (\pm\infty) = \pm\infty, && \pm \infty + (\pm\infty) = \pm\infty,\\
&\pm \infty \cdot x = x \cdot (\pm\infty) = \pm\infty \text{ if } x > 0, && \pm \infty \cdot (+\infty) = \pm\infty,\\
&\pm \infty \cdot x = x \cdot (\pm\infty) = \mp\infty \text{ if } x < 0, && \pm \infty \cdot (-\infty) = \mp\infty,\\
&-(\pm\infty) = \mp\infty, && \frac{x}{\pm\infty} = 0.
\end{aligned}$$

Note that the following expressions are not defined

$$\pm\infty - (\pm\infty), \qquad \pm\infty + (\mp\infty), \qquad \frac{\pm\infty}{\pm\infty},$$

where in the last expression the infinities in the numerator and denominator can have the same or the opposite signs. Also, $0 \cdot (\pm\infty)$ is sometimes defined to be zero. But we postpone using this convention until needed, and for now consider $0 \cdot (\pm\infty)$ to be undefined. ■

Definition 1.31. Let $A \subset \mathbb{R}$. When A is nonempty and bounded above, we denote its least upper bound by $\sup A$, and we call it the **supremum** of A. When A is nonempty and has no upper bound we define

$$\sup A := +\infty.$$

When A is nonempty and bounded below, we denote its greatest lower bound by $\inf A$, and we call it the **infimum** of A. When A is nonempty and has no lower bound we define

$$\inf A := -\infty.$$

Remark. It is obvious that for a nonempty set $A \subset \mathbb{R}$ we always have $\inf A \leq \sup A$.

Definition 1.32. Suppose $a, b \in \mathbb{R}$, and $a < b$. The **closed interval** and the **open interval** with **endpoints** a, b are respectively

$$\begin{aligned} [a, b] &:= \{x \in \mathbb{R} : a \le x \le b\}, \\ (a, b) &:= \{x \in \mathbb{R} : a < x < b\}. \end{aligned}$$

We can similarly define the *half-open* (or *half-closed*) intervals

$$\begin{aligned} [a, b) &:= \{x \in \mathbb{R} : a \le x < b\}, \\ (a, b] &:= \{x \in \mathbb{R} : a < x \le b\}. \end{aligned}$$

All these intervals are called **bounded**. The **length** of these bounded intervals is the positive real number $b - a$. We also define the **unbounded intervals**, whose endpoints can be $\pm\infty$, as follows

$$\begin{aligned} (-\infty, a) &:= \{x \in \mathbb{R} : x < a\}, \\ (-\infty, a] &:= \{x \in \mathbb{R} : x \le a\}, \\ (b, +\infty) &:= \{x \in \mathbb{R} : x > b\}, \\ [b, +\infty) &:= \{x \in \mathbb{R} : x \ge b\}, \\ (-\infty, +\infty) &:= \mathbb{R}. \end{aligned}$$

An unbounded interval that contains its finite endpoint is called closed, and an unbounded interval that does not contain its finite endpoint is called open. We consider $\mathbb{R}$ to be both an open interval and a closed interval.

Archimedean Property. *For every $x \in \mathbb{R}$ there exists $n \in \mathbb{N}$ such that $x < n$. Also, for every $x \in (0, \infty)$ there exists $n \in \mathbb{N}$ such that $\frac{1}{n} < x$.*

Proof. Suppose to the contrary that there is $x_0 \in \mathbb{R}$ such that $x_0 \ge n$ for all $n \in \mathbb{N}$. Let

$$A := \{x \in \mathbb{R} : x \ge n \; \forall n \in \mathbb{N}\}.$$

Then A is nonempty by our assumption. A is also bounded below, since for example 1 is a lower bound for it. Therefore $s := \inf A$ is a finite number. Now $s + \frac{1}{2}$ is not a lower bound for A, since s is the greatest lower bound for A. Hence there is $x_1 \in A$ such that $x_1 < s + \frac{1}{2}$. On the other hand $s - \frac{1}{2}$ does not belong to A, since s is a lower bound for A. Thus there is $n_0 \in \mathbb{N}$ such that $s - \frac{1}{2} < n_0$. But this implies that

$$x_1 < s + \frac{1}{2} < n_0 + 1 \in \mathbb{N},$$

which is a contradiction.

For the second statement, note that there is $n \in \mathbb{N}$ such that $\frac{1}{x} < n$. Thus as $\frac{1}{x} > 0$ we can deduce that $x > \frac{1}{n}$. ■

The meaning of the Archimedean property is that $\mathbb{R}$ does not contain infinitely large or infinitely small elements, since intuitively we consider $n = 1 + \cdots + 1$ to be finite, no matter how large n is. Similarly, we consider $\frac{1}{n}$ to be finite, even though it might be very small.

There is an alternative way to formulate the Archimedean property, which also states the intuition that $\mathbb{R}$ does not have infinitely large or small elements. Let $x, y > 0$ be real numbers. Then there is $n \in \mathbb{N}$ such that $nx > y$. In other words, no matter how small x is, and how large y is, in finitely many steps of length x we can surpass y. For the proof note that there is n such that $\frac{y}{x} < n$; so we have $y < nx$, since x is positive. Conversely, if this formulation of the Archimedean property holds, then by setting $x = 1$ we obtain the previous formulation.

Remark. The Archimedean property of $\mathbb{R}$ is a consequence of its least upper bound property, but the two properties are not equivalent. For example $\mathbb{Q}$ is not complete but it is Archimedean, since for every $\frac{p}{q} \in \mathbb{Q}$ we have $\frac{p}{q} < |p| + 1$. Finally we mention that there are ordered fields that are not Archimedean.

It is easy to show that the well-ordering of natural numbers implies the following property for sets of integers in $\mathbb{R}$. However, we will show that this property can be deduced from the fact that $\mathbb{R}$ is a complete ordered field.

Theorem 1.33. *Let A be a set of integers in $\mathbb{R}$, and suppose A is nonempty.*
(i) *If A is bounded below then it has a least element.*
(ii) *If A is bounded above then it has a largest element.*

Proof. **(i)** Since A is nonempty and bounded below, it has a greatest lower bound, which we call m. Then $m + \frac{1}{2}$ cannot be a lower bound for A. Hence there is $n \in A$ such that $n < m + \frac{1}{2}$. Then we have $n - 1 < m - \frac{1}{2} < m$. We also have $m \leq n$. Thus

$$n - 1 < m \leq n.$$

Now if $m \neq n$ then n cannot be a lower bound for A. Therefore there must be $k \in A$ such that $k < n$. However $k \geq m$ too. So we get $n - 1 < k < n$, which contradicts the assumption of k, n being integers. Hence we must have $m = n$; which implies $m \in A$. Therefore m is the least element of A, since it belongs to A, and it is a lower bound for A.

(ii) This part can be proved similarly to the previous part. We can also use the set $B := \{-a : a \in A\}$, and use the previous part, similarly to the proof of Theorem 1.13. ■

Integer Part. Let $x \in \mathbb{R}$. Then by the Archimedean property there is $n_1 \in \mathbb{N}$ such that $x < n_1$. Thus the set

$$A := \{n \in \mathbb{Z} : x < n\}$$

is nonempty and bounded below. Hence A has a smallest element, which we call $n_0 + 1$. So $n_0 \notin A$. Therefore we have

$$n_0 \leq x < n_0 + 1.$$

We set $\lfloor x \rfloor := n_0$, and call $\lfloor x \rfloor$ the **integer part** of x. Note that $\lfloor x \rfloor$ is the greatest integer less than or equal to x, and we have

$$\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1, \qquad x - 1 < \lfloor x \rfloor \leq x.$$

The function $\lfloor\ \rfloor : \mathbb{R} \to \mathbb{Z}$ is called the greatest integer function or the **floor** function. ■

Theorem 1.34. *Between any two real numbers there is a rational and an irrational number.*

Proof. Suppose $a < b$. Let $c = \frac{a+b}{2}$. It is easy to see that $a < c < b$. Then $c - a > 0$, and there is $n \in \mathbb{N}$ such that $c - a > \frac{1}{n}$. Now we have $nc - 1 < \lfloor nc \rfloor \leq nc$. Thus $a < c - \frac{1}{n} < \frac{\lfloor nc \rfloor}{n} \leq c < b$, and $\frac{\lfloor nc \rfloor}{n} \in \mathbb{Q}$.

Next, let r be a rational number between $a + \sqrt{2}$ and $b + \sqrt{2}$. Then $r - \sqrt{2}$ is an irrational number between a, b. ■

Theorem 1.35. *A subset $I \subset \mathbb{R}$ is an interval, if and only if it has more than one element; and for every $a, b \in I$ with $a < b$, and every c where $a < c < b$, we have $c \in I$.*

Proof. First suppose I is an interval with endpoints $\alpha < \beta$, where the endpoints can be $\pm\infty$. Then by Theorem 1.34, I has at least two elements. Let $a, b \in I$ with $a < b$, and let $a < c < b$. Then we have $\alpha \leq a$, and $b \leq \beta$. Hence $\alpha < c < \beta$, and therefore $c \in I$ by the definition of intervals.

Now suppose conversely that I has the specified property. Then in particular I is nonempty. Let

$$\alpha := \inf I, \qquad \beta := \sup I.$$

Then $\alpha < \beta$, since otherwise I cannot have more than one element. We claim that $(\alpha, \beta) \subset I$. Let $\alpha < c < \beta$. Then there is $c < x < \beta$. But x cannot be an upper bound for I. Hence there is $x < b \leq \beta$ such that $b \in I$. Similarly there is $\alpha \leq a < c$ such that $a \in I$. Therefore $c \in I$. Thus $(\alpha, \beta) \subset I$.

On the other hand, if $c > \beta$ then $c \notin I$, since β is an upper bound for I. Similarly I cannot contain any number less than α. Hence $I \subset [\alpha, \beta]$. Therefore I equals one of the sets

$$(\alpha, \beta), \quad [\alpha, \beta), \quad (\alpha, \beta], \quad [\alpha, \beta].$$

Thus I is an interval. Note that if one of the α, β is infinite, then we have to eliminate the sets containing it. ■

Definition 1.36. Let $x \in \mathbb{R}$. Then the **absolute value** of x is

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

Theorem 1.37. *For all $x, y \in \mathbb{R}$ we have*

(i) $|x| \geq 0$, *and for $x \neq 0$ we have* $|x| > 0$.

(ii) $|-x| = |x|$.

(iii) $|xy| = |x||y|$.

(iv) $|\frac{x}{y}| = \frac{|x|}{|y|}$ *when* $y \neq 0$.

(v) $|x| < r$ *if and only if* $-r < x < r$; *and* $|x| \leq r$ *if and only if* $-r \leq x \leq r$.

(vi) **(Triangle Inequality)** $|x + y| \leq |x| + |y|$.

Proof. **(i)** For $x > 0$ we have $|x| = x > 0$, and for $x < 0$ we have $|x| = -x > 0$. Also, $|0| = 0$.

(ii) We have

$$|-x| = \begin{cases} -x & \text{if } -x \geq 0 \\ -(-x) & \text{if } -x < 0 \end{cases} = \begin{cases} -x & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ x & \text{if } x > 0 \end{cases} = |x|.$$

(iii) When $x, y \geq 0$ we have $xy \geq 0$, hence $|xy| = xy = |x||y|$. When one of the x or y is negative, we consider its opposite and apply the last argument using (ii). For example when $x < 0$, and $y \geq 0$ we have

$$|xy| = |-(-x)y| = |(-x)y| = |-x||y| = |x||y|.$$

The other two cases are similar.

(iv) We have $y\frac{1}{y} = 1$. Thus $|y||\frac{1}{y}| = |1| = 1$, since $1 > 0$. Therefore $|\frac{1}{y}| = \frac{1}{|y|}$. Then we have

$$|\frac{x}{y}| = |x||\frac{1}{y}| = |x|\frac{1}{|y|} = \frac{|x|}{|y|}.$$

(v) We have

$$|x| \leq r \iff \begin{cases} x \leq r & \text{if } x \geq 0 \\ -x \leq r & \text{if } x < 0 \end{cases} \iff 0 \leq x \leq r, \quad \text{or} \quad -r \leq x < 0.$$

The other one is similar.

(vi) Since $|x| \leq |x|$, by (v) we have

$$-|x| \leq x \leq |x|, \qquad -|y| \leq y \leq |y|.$$

If we add these two inequalities we get

$$-|x| - |y| \leq x + y \leq |x| + |y|.$$

Therefore again by (v) we obtain $|x + y| \leq |x| + |y|$. ■

Theorem 1.38. *If for all $\epsilon > 0$ we have $|x - y| \le \epsilon$ then $x = y$.*

$\boxed{\textbf{Proof.}}$ Suppose to the contrary that $x \neq y$. Then $|x - y| > 0$. Let $\epsilon = \frac{1}{2}|x - y|$. Then we have $|x-y| \le \frac{1}{2}|x-y|$, so $2|x-y| \le |x-y|$. Therefore we get $|x-y| \le 0$, which is a contradiction. ■

Remark. Similarly if $x \le y + \epsilon$ for all $\epsilon > 0$ then $x \le y$.

1.4 Decimal Expansion

Theorem 1.39. *Consider the set of sequences of the form*

$$a_0.a_1a_2a_3\ldots,$$

where $a_0 \in \mathbb{N} \cup \{0\}$, $a_j \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ for $j > 0$, and for every j there is $k > j$ such that $a_k \neq 9$. Then there exists a one-to-one and onto map from this set of sequences to $[0, \infty) \subset \mathbb{R}$, that maps the sequence $a_0.a_1a_2a_3\ldots$ to the real number

$$x := \sup\{x_n : n \in \mathbb{N}\},$$

where $x_n := a_0 + \sum_{j=1}^{n} \dfrac{a_j}{10^j}$.

Definition 1.40. The unique sequence $a_0.a_1a_2a_3\ldots$ whose existence is proved in the above theorem is called the **decimal expansion** of x.

Remark. By using the concepts of *limit* and *series*, we can easily conclude from the following proof that

$$x = \lim_{n\to\infty} x_n = a_0 + \sum_{j=1}^{\infty} \frac{a_j}{10^j}.$$

Remark. In the following proof, we actually provide a concrete method for finding the decimal expansion of x.

$\boxed{\textbf{Proof.}}$ First note that $x_{n+1} = x_n + \frac{a_{n+1}}{10^{n+1}} \ge x_n$. So x_n forms an increasing sequence, i.e. it is an increasing function of n. Next note that the set $\{x_n\}$ is nonempty and bounded above, since

$$a_0 + \sum_{j\le n} \frac{a_j}{10^j} \le a_0 + \sum_{j \le n} \frac{9}{10^j} = a_0 + 9\frac{\frac{1}{10} - \frac{1}{10^{n+1}}}{1 - \frac{1}{10}} = a_0 + 1 - \frac{1}{10^n} < a_0 + 1.$$

Therefore $x = \sup\{x_n\}$ exists. In addition, for some n we have $x \ge x_n \ge a_0 \ge 0$; thus $x \in [0, \infty)$.

Now let us show that the map defined in the theorem is one-to-one. Suppose $a_0.a_1a_2a_3\dots$ and $b_0.b_1b_2b_3\dots$ are distinct sequences, which are mapped to x, y respectively. We want to show that $x \neq y$. Suppose l is the smallest index for which we have $a_l \neq b_l$. We will show that if $a_l < b_l$ then $x < y$. We know that there is $k > l$ such that $a_k \neq 9$. Hence $a_k \le 8 = 9 - 1$. We also have $a_l + 1 \le b_l$. Then for $n > k$ we have

$$\begin{aligned}
x_n = a_0 + \sum_{j\le n} \frac{a_j}{10^j} &= a_0 + \sum_{j<l} \frac{a_j}{10^j} + \frac{a_l}{10^l} + \sum_{l<j<k} \frac{a_j}{10^j} + \frac{a_k}{10^k} + \sum_{k<j\le n} \frac{a_j}{10^j} \\
&\le a_0 + \sum_{j<l} \frac{a_j}{10^j} + \frac{a_l}{10^l} + \sum_{l<j<k} \frac{9}{10^j} + \frac{9-1}{10^k} + \sum_{k<j\le n} \frac{9}{10^j} \\
&= a_0 + \sum_{j<l} \frac{a_j}{10^j} + \frac{a_l}{10^l} - \frac{1}{10^k} + 9\frac{\frac{1}{10^{l+1}} - \frac{1}{10^{n+1}}}{1 - \frac{1}{10}} \\
&= b_0 + \sum_{j<l} \frac{b_j}{10^j} + \frac{a_l}{10^l} - \frac{1}{10^k} + \frac{1}{10^l} - \frac{1}{10^n} \\
&\le b_0 + \sum_{j<l} \frac{b_j}{10^j} + \frac{b_l}{10^l} - \frac{1}{10^k} - \frac{1}{10^n} \\
&= y_l - \frac{1}{10^k} - \frac{1}{10^n} < y_l - \frac{1}{10^k} \le y_n - \frac{1}{10^k}.
\end{aligned}$$

Thus we get $x_n + \frac{1}{10^k} < y_n \le y$, since y is the supremum of $\{y_n\}$. Also note that for $n \le k$ we have $x_n + \frac{1}{10^k} \le x_{k+1} + \frac{1}{10^k} < y$, because x_n is an increasing sequence. Therefore we get $x \le y - \frac{1}{10^k}$, since x is the supremum of $\{x_n\}$. Hence we obtain $x < y$, as desired.

Next let us prove that the map defined in the theorem is onto. Let x be a nonnegative real number. Set a_0 to be the integer part of x, i.e.

$$a_0 := \lfloor x \rfloor.$$

We know that $a_0 \le x < a_0 + 1$. Hence $0 \le x - a_0 < 1$. Also note that since $x \ge 0$ we have $a_0 > -1$; so $a_0 \ge 0$, since it is an integer. Next let

$$a_1 := \lfloor 10(x - a_0) \rfloor.$$

Note that $0 \le 10(x - a_0) < 10$. Thus a_1 is a nonnegative integer less than 10, i.e. it belongs to $\{0, 1, \dots, 9\}$. In addition we have

$$0 \le 10(x - a_0) - a_1 < 1 \implies 0 \le x - \big(a_0 + \frac{a_1}{10}\big) < \frac{1}{10}.$$

We continue this process to inductively define a_n. More precisely, we define

(i) $x_0 := \lfloor x \rfloor$,

(ii) $x_{n+1} := x_n + \frac{\lfloor 10^{n+1}(x-x_n)\rfloor}{10^{n+1}}$.

Then for $n \geq 0$ we define

$$a_{n+1} := 10^{n+1}(x_{n+1} - x_n) = \lfloor 10^{n+1}(x - x_n)\rfloor.$$

It is easy to show that $x_n = a_0 + \sum_{j=1}^{n} \frac{a_j}{10^j}$. Because for $n = 0$ we have $x_0 = \lfloor x \rfloor = a_0$. And if the equality holds for some n, then for $n+1$ we get

$$x_{n+1} = x_n + \frac{a_{n+1}}{10^{n+1}} = a_0 + \sum_{j=1}^{n} \frac{a_j}{10^j} + \frac{a_{n+1}}{10^{n+1}} = a_0 + \sum_{j=1}^{n+1} \frac{a_j}{10^j},$$

as desired.

Now let us show that

$$0 \leq x - x_n < \frac{1}{10^n}. \tag{$*$}$$

Note that as a consequence we get $0 \leq 10^{n+1}(x - x_n) < 10$. Therefore $a_{n+1} = \lfloor 10^{n+1}(x-x_n)\rfloor$ is a nonnegative integer less than 10, i.e. it belongs to $\{0, 1, \dots, 9\}$. The proof of the inequality $(*)$ is by induction on n. For $n = 0$ we have $x_0 = \lfloor x \rfloor$, and the inequality holds due to the properties of the integer part, as we have seen above. Suppose the inequality holds for some n. We know that

$$0 \leq 10^{n+1}(x - x_n) - \lfloor 10^{n+1}(x - x_n)\rfloor < 1,$$

due to the properties of the integer part. Hence we get

$$0 \leq x - \big(x_n + \frac{\lfloor 10^{n+1}(x - x_n)\rfloor}{10^{n+1}}\big) < \frac{1}{10^{n+1}},$$

which is the desired inequality for $n+1$.

Next, note that we have

$$a_n = \lfloor 10^n(x - x_{n-1})\rfloor \leq 10^n(x - x_{n-1}) < a_n + 1.$$

Hence for small positive ϵ we have $10^n(x - x_{n-1}) < a_n + 1 - \epsilon$. But by Archimedean property there is k such that $k > \frac{1}{10\epsilon}$; so $10^k \geq 10k > \frac{1}{\epsilon}$. Thus we get

$$x - x_{n-1} < \frac{a_n}{10^n} + \frac{1}{10^n} - \frac{1}{10^{n+k}}.$$

So $x - x_n < \frac{1}{10^n} - \frac{1}{10^{n+k}}$. Now suppose to the contrary that $a_{n+1} = \cdots = a_{n+k} = 9$. Then we have

$$x_{n+k} - x_n = \sum_{j=n+1}^{n+k} \frac{a_j}{10^j} = \sum_{j=n+1}^{n+k} \frac{9}{10^j} = 9\frac{\frac{1}{10^{n+1}} - \frac{1}{10^{n+k+1}}}{1 - \frac{1}{10}} = \frac{1}{10^n} - \frac{1}{10^{n+k}}.$$

Hence we get

$$x - x_{n+k} = x - x_n + x_n - x_{n+k} < 0,$$

which contradicts the inequality $(*)$. Therefore one of the $a_{n+1}, \dots, a_{n+k}$ must be less than 9. So the sequence $a_0.a_1a_2a_3\dots$ satisfies all the properties required in the theorem.

Finally let us show that $a_0.a_1a_2a_3\dots$ is mapped to x, i.e. x is the supremum of $\{x_n\}$. First note that by inequality $(*)$ we have $x_n \le x$ for every n. Thus x is an upper bound for $\{x_n\}$. On the other hand, suppose y is an upper bound for $\{x_n\}$. Then $(*)$ implies that $x - \frac{1}{10^n} < x_n \le y$. So $x - y < \frac{1}{10^n}$ for every n. However, as we have shown above, for every $\epsilon > 0$ there is n such that $\epsilon > \frac{1}{10^n}$, due to the Archimedean property. Hence $x - y < \epsilon$ for every $\epsilon > 0$; so $x - y \le 0$. Thus $x = \sup\{x_n\}$, as desired. ■

1.5 Powers and Roots

Definition 1.41. We define the **powers** of $a \in \mathbb{R}$ as follows. For a positive integer n we inductively define

$$a^1 := a, \ \dots \ a^n := a^{n-1}a.$$

Here a is called the **base**, and n is called the **exponent**. If $a \neq 0$, we define

$$a^0 := 1, \qquad a^{-n} := (a^{-1})^n.$$

Remark. We also use the convention that $0^0 = 1$. This is useful in some algebraic manipulations, but we must be careful that 0^0 does not have a definite value when we deal with limits in later chapters.

Notation. We assume that exponentiation binds stronger than multiplication and addition; so, for example, $d + a^n c$ means $d + ((a^n)c)$. We also use the convention that a^{m^n} means $a^{(m^n)}$.

Definition 1.42. Let $n \in \mathbb{N}$. The ***n* factorial** is

$$n! := n \times (n-1) \times \cdots \times 2 \times 1.$$

We also set $0! := 1$. Suppose $n, k \in \mathbb{Z}$, and $0 \le k \le n$. The number

$$\binom{n}{k} := \frac{n!}{k!(n-k)!}$$

is called a **binomial coefficient**.

Remark. Note that for all $n \ge 1$ we have $n! = n(n-1)!$. It is also trivial to see that $\binom{n}{0} = 1 = \binom{n}{n}$ for all $n \ge 0$.

Proposition 1.43. *For all integers $1 \le k \le n$ we have*

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}.$$

As a result $\binom{n}{k}$ is always a positive integer for every $0 \le k \le n$.

Proof. We have

$$\begin{aligned}\binom{n}{k} + \binom{n}{k-1} &= \frac{n!}{k!(n-k)!} + \frac{n!}{(k-1)!(n-k+1)!} \\ &= \frac{n!}{(k-1)!(n-k)!}(\frac{1}{k} + \frac{1}{n-k+1}) \\ &= \frac{n!}{(k-1)!(n-k)!}\,\frac{n+1}{k(n-k+1)} \\ &= \frac{(n+1)!}{k!(n+1-k)!} = \binom{n+1}{k}.\end{aligned}$$

Next, we show by induction on n that $\binom{n}{k}$ is a positive integer for all $0 \le k \le n$. For $n = 1$ we have $\binom{1}{0} = \binom{1}{1} = 1 \in \mathbb{N}$. Suppose the claim holds for n. Then for $0 < k < n+1$ we have

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1} \in \mathbb{N}.$$

Also note that $\binom{n+1}{0} = \binom{n+1}{n+1} = 1 \in \mathbb{N}$. ■

Theorem 1.44. *Suppose $a, b \in \mathbb{R}$, and $n, m \in \mathbb{Z}$. In all of the following statements, when the base is zero the exponent must be nonnegative.*

(i) *If $a \neq 0$ then $(a^n)^{-1} = a^{-n} = (a^{-1})^n$.*

(ii) $a^n a^m = a^{n+m}$.

(iii) $(a^n)^m = a^{nm}$.

(iv) $a^n b^n = (ab)^n$.

(v) **(Binomial Theorem)** *For $n > 0$ we have*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k.$$

(vi) *For $n > 0$ we have*

$$a^n - b^n = (a-b)\Big(\sum_{k=0}^{n-1} a^{n-1-k} b^k\Big).$$

(vii) *For $n \geq m \geq 0$ and $a \neq 1$ we have*

$$\sum_{k=m}^{n} a^k = \frac{a^{n+1} - a^m}{a - 1}.$$

(viii) *Suppose $n \geq 0$. If $a > 1$ then $a^{n-1} < a^n$, and if $0 < a < 1$ then $a^{n-1} > a^n$.*

(ix) *If $a > 0$ then $a^n > 0$.*

(x) *If n is even then $(-a)^n = a^n$, and if n is odd then $(-a)^n = -a^n$.*

(xi) *If $n > 0$ and $0 \leq a < b$ then $a^n < b^n$.*

(xii) *If $n > 0$ is odd and $a < b$ then $a^n < b^n$.*

(xiii) $|a^n| = |a|^n$.

Proof. Almost all of the proofs are by induction. We will only write the induction steps below, since the base of inductions can be checked easily.

(i) When $n \geq 0$ we have

$$(a^{n+1})a^{-n-1} = a^n a(a^{-1})^{n+1} = aa^n(a^{-1})^n a^{-1} = aa^n a^{-n} a^{-1} = aa^{-1} = 1.$$

When $n = -m < 0$ we have $a^{-m} = (a^{-1})^m$. Hence by the previous part we get

$$(a^{-m})^{-1} = ((a^{-1})^m)^{-1} = (a^{-1})^{-m} = ((a^{-1})^{-1})^m = a^m.$$

The second equality holds by definition when $n > 0$. When $n = 0$ we have

$$(a^{-1})^0 = 1 = a^0 = a^{-0}.$$

And when $n = -m < 0$ we have

$$(a^{-1})^{-m} = ((a^{-1})^{-1})^m = a^m = a^{-n}.$$

(ii) When $n, m \geq 0$ we have

$$a^n a^{m+1} = a^n a^m a = a^{n+m} a = a^{n+m+1}.$$

Now suppose a is nonzero. Then we have

$$\begin{aligned} a^{-n}a^{m+1} &= a^{-n}a^m a \\ &= a^{-n+m}a = \begin{cases} (a^{-1})^{n-m}a = (a^{-1})^{n-m-1}a^{-1}a & \text{if } -n+m < 0, \\ \qquad = (a^{-1})^{n-m-1} = a^{-n+m+1} & \\ & \\ a^{-n+m+1} & \text{if } -n+m \geq 0. \end{cases} \end{aligned}$$

We also have

$$\begin{aligned} a^n a^{-m} &= (a^{-1})^{-n}(a^{-1})^m = (a^{-1})^{-n+m} = a^{n-m}, \\ a^{-n} a^{-m} &= (a^{-1})^{n}(a^{-1})^m = (a^{-1})^{n+m} = a^{-n-m}. \end{aligned}$$

(iii) For $n, m \geq 0$ we have

$$(a^n)^{m+1} = (a^n)^m a^n = a^{nm} a^n = a^{nm+n} = a^{n(m+1)}.$$

If a is nonzero we have

$$\begin{aligned} (a^{-n})^m &= ((a^{-1})^n)^m = (a^{-1})^{nm} = a^{-nm}, \\ (a^{\pm n})^{-m} &= ((a^{\pm n})^m)^{-1} = (a^{\pm nm})^{-1} = a^{\mp nm}. \end{aligned}$$

(iv) For $n \geq 0$ we have

$$a^{n+1}b^{n+1} = a^n a b^n b = a^n b^n ab = (ab)^n ab = (ab)^{n+1},$$

and if a, b are nonzero we have

$$a^{-n}b^{-n} = (a^{-1})^n (b^{-1})^n = (a^{-1}b^{-1})^n = ((ab)^{-1})^n = (ab)^{-n}.$$

(v) We have

$$\begin{aligned} (a+b)^{n+1} &= (a+b)^n(a+b) = \left(\sum_{k=0}^{n} \binom{n}{k} a^{n-k}b^k\right)(a+b) \\ &= \sum_{k=0}^{n} \binom{n}{k} a^{n-k}b^k(a+b) = \sum_{k=0}^{n} \binom{n}{k} (a^{n-k}b^k a + a^{n-k}b^k b) \\ &= \sum_{k=0}^{n} \binom{n}{k} a^{n-k+1}b^k + \sum_{k=0}^{n} \binom{n}{k} a^{n-k}b^{k+1} \\ &= \sum_{k=0}^{n} \binom{n}{k} a^{n+1-k}b^k + \sum_{j=1}^{n+1} \binom{n}{j-1} a^{n+1-j}b^j \\ &\qquad\qquad \text{(We replaced } k \text{ with } j-1 \text{ in the 2nd sum.)} \\ &= a^{n+1} + \left(\sum_{k=1}^{n} \binom{n}{k} a^{n+1-k}b^k + \sum_{k=1}^{n} \binom{n}{k-1} a^{n+1-k}b^k\right) + b^{n+1} \\ &\qquad\qquad \text{(We replaced } j \text{ with } k \text{ in the 2nd sum.)} \\ &= a^{n+1} + \left(\sum_{k=1}^{n} \left[\binom{n}{k} + \binom{n}{k-1}\right] a^{n+1-k}b^k\right) + b^{n+1} \\ &= \sum_{k=0}^{n+1} \binom{n+1}{k} a^{n+1-k}b^k. \end{aligned}$$

(vi) We have

$$\begin{aligned}
(a-b)\Big(\sum_{k=0}^{n-1} a^{n-1-k}b^k\Big) &= \sum_{k=0}^{n-1}(a-b)a^{n-1-k}b^k \\
&= \sum_{k=0}^{n-1}(aa^{n-1-k}b^k - ba^{n-1-k}b^k) \\
&= \sum_{k=0}^{n-1} a^{n-k}b^k - \sum_{k=0}^{n-1} a^{n-1-k}b^{k+1} = \sum_{k=0}^{n-1} a^{n-k}b^k - \sum_{j=1}^{n} a^{n-j}b^j \\
&\qquad \text{(We replaced } k \text{ with } j-1 \text{ in the 2nd sum.)} \\
&= a^n + \left(\sum_{k=1}^{n-1} a^{n-k}b^k - \sum_{k=1}^{n-1} a^{n-k}b^k\right) - b^n \\
&\qquad \text{(We replaced } j \text{ with } k \text{ in the 2nd sum.)} \\
&= a^n - b^n.
\end{aligned}$$

(vii) We have

$$\begin{aligned}
(a-1)\Big(\sum_{k=m}^{n} a^k\Big) &= \sum_{k=m}^{n}(a^{k+1}-a^k) \\
&= a^{n+1} - a^n + a^n - a^{n-1} + \cdots + a^{m+1} - a^m = a^{n+1} - a^m.
\end{aligned}$$

(viii) For $a > 1$ we have $a^{n-1} < a^n$; hence

$$a^n = aa^{n-1} < aa^n = a^{n+1}.$$

The case of $0 < a < 1$ is similar.

(ix) For $n > 0$ we multiply both sides of $a^n > 0$ by a to get $a^{n+1} > 0$. When $n = 0$ we have $a^0 = 1 > 0$. And when $n = -m < 0$ we have $a^n = (a^{-1})^m > 0$, since $a^{-1} > 0$.

(x) Since $(-a)^n = ((-1)a)^n = (-1)^n a^n$, we only need to compute $(-1)^n$. Now if $(-1)^{2k} = 1$ then

$$(-1)^{2(k+1)} = (-1)^{2k}(-1)^2 = 1 \cdot 1 = 1.$$

For negative powers we have the same result since $(-1)^{-1} = -1$. Finally for odd powers we have $(-1)^{2k+1} = (-1)^{2k}(-1) = -1$.

(xi) First suppose $0 < a < b$. Then by induction hypothesis and (ix) we know that $0 < a^n < b^n$. Now we multiply these two inequalities to get

$$0 < a^{n+1} < b^{n+1}.$$

Since $0^n = 0$, we can allow $a = 0$ too by (ix).

(xii) If $0 \leq a < b$ then the claim holds by last part. If $a < b \leq 0$ then $0 \leq -b < -a$, hence

$$-b^n = (-b)^n < (-a)^n = -a^n.$$

Thus $a^n < b^n$. Finally if $a < 0 < b$, then $b^n > 0$. Also $-a^n = (-a)^n > 0$, since $-a > 0$. Hence $a^n < 0 < b^n$.

(xiii) For $n \geq 0$ we have

$$|a^{n+1}| = |a^n a| = |a^n||a| = |a|^n|a| = |a|^{n+1}.$$

When $n = -m < 0$ we have

$$|a^{-m}| = |(a^{-1})^m| = |a^{-1}|^m = (|a|^{-1})^m = |a|^{-m}. \qquad \blacksquare$$

***Exercise* 1.45.** Show that for $a_1, \dots, a_m \in \mathbb{R}$ we have

$$(a_1 + \cdots + a_m)^2 = \sum_{i \leq m} a_i^2 + 2 \sum_{j \leq m} \sum_{i<j} a_i a_j.$$

Theorem 1.46. *For every real number $x \geq 0$ and every $n \in \mathbb{N}$ there is a unique real number $y \geq 0$ such that $y^n = x$. We denote y by $\sqrt[n]{x}$ or $x^{\frac{1}{n}}$, and call it the **nth root** of x.*

Notation. We denote $\sqrt[2]{x}$ by $\sqrt{x}$, and we call it the **square root** of x.

Proof. The uniqueness of y is obvious, since if $0 \leq y_1 < y_2$ then $y_1^n < y_2^n$. For the existence, consider the set

$$A := \{z \geq 0 : z^n \leq x\}.$$

Note that $0 \in A$, so A is nonempty. Also if $z \geq 1 + x \geq 1$, then $z^n \geq (1+x)^n \geq 1 + x > x$. Thus $1 + x$ is an upper bound for A. Let $y := \sup A$. We need to show that $y^n = x$.

Suppose to the contrary that $y^n < x$. Then for $0 < \epsilon < 1$ and $k \in \mathbb{N}$ we have $\epsilon^k \leq \epsilon$. Thus by using the binomial theorem we get

$$(y+\epsilon)^n = \sum_{k=0}^{n} \binom{n}{k} y^{n-k} \epsilon^k \leq y^n + \sum_{k=1}^{n} \binom{n}{k} y^{n-k} \epsilon = y^n + M\epsilon,$$

where $M := \sum\limits_{k=1}^{n} \binom{n}{k} y^{n-k} > 0$. Now for $0 < \epsilon < \min\{\frac{x-y^n}{M}, 1\}$ we have

$$(y+\epsilon)^n \leq y^n + M\epsilon < x,$$

which contradicts the fact that y is an upper bound for A.

Next, suppose to the contrary that $y^n > x \geq 0$. Suppose $0 < \epsilon < \min\{1, y\}$. Let $z := y - \epsilon$. Then as $0 < z < y$ we have

$$y^n = (z+\epsilon)^n = \sum_{k=0}^{n} \binom{n}{k} z^{n-k}\epsilon^k < z^n + \sum_{k=1}^{n} \binom{n}{k} y^{n-k}\epsilon = z^n + M\epsilon,$$

where $M := \sum\limits_{k=1}^{n} \binom{n}{k} y^{n-k} > 0$. Now for $0 < \epsilon < \min\{\frac{y^n - x}{M}, 1, y\}$ we have

$$(y-\epsilon)^n = z^n > y^n - M\epsilon > x.$$

But y is the supremum of A, so $y - \epsilon$ is not an upper bound for A. Hence there is $a \in A$ such that $a > y - \epsilon$. Therefore $a^n > (y-\epsilon)^n > x$, which is a contradiction. ■

Remark. When $n \in \mathbb{N}$ is odd we have $(-1)^n = -1$. Thus for $x < 0$ we have

$$(-\sqrt[n]{-x})^n = (-1)^n(\sqrt[n]{-x})^n = (-1)(-x) = x.$$

Since for $y_1 < y_2$ we have $y_1^n < y_2^n$, there is no other real number whose nth power is x. So we define

$$\sqrt[n]{x} := -\sqrt[n]{-x}.$$

Theorem 1.47. *For any two nonnegative real numbers x, y, and any $n \in \mathbb{N}$ we have*

$$\sqrt[n]{xy} = \sqrt[n]{x}\,\sqrt[n]{y}.$$

When n is odd we can allow x and/or y to be negative too.

Proof. We have

$$(\sqrt[n]{x}\,\sqrt[n]{y})^n = (\sqrt[n]{x})^n(\sqrt[n]{y})^n = xy.$$

Thus when $x, y \geq 0$ we get the desired result since $\sqrt[n]{x}\,\sqrt[n]{y} \geq 0$. When n is odd, we do not need to assume anything about the sign of x, y, since there is only one real number whose nth power is xy. ■

Theorem 1.48. *For any two nonnegative real numbers x, y, and any $n \in \mathbb{N}$ we have*

$$x < y \implies \sqrt[n]{x} < \sqrt[n]{y}.$$

When n is odd we can allow x and/or y to be negative too.

Proof. Suppose to the contrary that $\sqrt[n]{x} \geq \sqrt[n]{y}$. Then as $\sqrt[n]{x}, \sqrt[n]{y} \geq 0$ we have

$$x = (\sqrt[n]{x})^n \geq (\sqrt[n]{y})^n = y,$$

which is a contradiction. When n is odd the same argument works, except that we do not need to assume $\sqrt[n]{x}, \sqrt[n]{y} \geq 0$, so we can allow x, y to be negative too. ■

Theorem 1.49. *For any real number x we have $\sqrt{x^2} = |x|$.*

Proof. We have $|x|^2 = |x^2| = x^2$, since $x^2 \geq 0$. Hence we get the desired result because $|x| \geq 0$. ■

Theorem 1.50. *$\sqrt{2}$ is irrational.*

Proof. Suppose to the contrary that $\sqrt{2} \in \mathbb{Q}$. Thus $\sqrt{2} = \frac{p}{q}$ where $p, q \in \mathbb{N}$, since $\sqrt{2} > 0$. Therefore we have $p^2 = 2q^2$. We can assume that p, q have no common factors. Thus both of them cannot be even. Suppose p is even; so $p = 2k$ for some $k \in \mathbb{N}$. Then q is odd; thus q^2 is odd too. But we have $4k^2 = p^2 = 2q^2$. Hence $q^2 = 2k^2$ is even, which is a contradiction. Next suppose p is odd. Then p^2 is odd too. But $p^2 = 2q^2$ must be even, which is again a contradiction. Hence $\sqrt{2}$ cannot be rational. ■

Rational and Real Exponents.

Suppose $p \in \mathbb{Q}$ and $x \in \mathbb{R}$ are positive. Then there are $n, k \in \mathbb{N}$ with no common factor such that $p = \frac{k}{n}$. Furthermore, $p = \frac{km}{nm}$ for every $m \in \mathbb{N}$, and these are all the representations of p as a fraction with positive denominator. Now we have $(\sqrt[n]{x})^k = \sqrt[n]{x^k}$. We also have

$$((\sqrt[nm]{x})^{km})^n = ((\sqrt[nm]{x})^{nm})^k = x^k \implies (\sqrt[nm]{x})^{km} = \sqrt[n]{x^k} = (\sqrt[n]{x})^k.$$

So if we set

$$x^p = x^{\frac{k}{n}} := (\sqrt[n]{x})^k,$$

then x^p is well defined. We also set

$$x^{-p} := \frac{1}{x^p} = (\sqrt[n]{x})^{-k}.$$

Also remember that $x^0 := 1$. Hence we have defined x^p for all $p \in \mathbb{Q}$ and all $x > 0$.

Remark. It is obvious from the definition that for all $p \in \mathbb{Q}$ we have $1^p = 1$, and $x^p > 0$ for all $x > 0$.

Proposition 1.51. *Suppose $p, q \in \mathbb{Q}$ and $p < q$. Then for $x > 1$ we have $x^p < x^q$, and for $0 < x < 1$ we have $x^p > x^q$.*

Proof. First suppose $x > 1$. Let $p = \frac{k}{n}$ and $q = \frac{l}{m}$, where $k, l \in \mathbb{Z}$, and $n, m \in \mathbb{N}$. Then we have $\frac{k}{n} < \frac{l}{m}$, so $mk < nl$. If $0 \leq p < q$ then $0 \leq mk < nl$. Thus $x^{mk} < x^{nl}$. Hence

$$x^p = \sqrt[n]{x^k} = \sqrt[nm]{x^{mk}} < \sqrt[nm]{x^{nl}} = \sqrt[m]{x^l} = x^q.$$

If $p < q \le 0$ then $0 \le -q < -p$. Therefore $0 < x^{-q} < x^{-p}$. Hence

$$x^p = \frac{1}{x^{-p}} < \frac{1}{x^{-q}} = x^q.$$

And if $p < 0 < q$ then $x^p < x^0 < x^q$. The case of $0 < x < 1$ is similar. ■

Now we can define x^r for all $x > 0$ and all $r \in \mathbb{R}$. We set

$$x^r := \begin{cases} \sup\{x^p : p \in \mathbb{Q}, p \le r\} & x \ge 1, \\ \inf\{x^p : p \in \mathbb{Q}, p \le r\} & 0 < x < 1. \end{cases}$$

First we have to check that the above supremum and infimum are finite. But this is easy since if $q \ge r$ is a rational number, then for all rational numbers $p \le r$ we have

$$\begin{cases} x^p \le x^q & x > 1, \\ x^p \ge x^q & 0 < x < 1. \end{cases} \tag{$*$}$$

Hence the set $\{x^p : p \le r\}$, which is obviously nonempty, has the appropriate bound to ensure the finiteness of the above supremum and infimum. The inequalities in $(*)$ also show that when r is rational, the new definition of x^r agrees with the old definition. Because in this case we have $x^r \in \{x^p : p \le r\}$, so x^r is the required supremum, or infimum, of this set.

Remark. Note that for all $r \in \mathbb{R}$ we have

$$1^r = \sup\{1^p : p \le r\} = \sup\{1\} = 1.$$

Also, for every $x > 0$ and $r \in \mathbb{R}$ we have $x^r > 0$. This is obvious when $x > 1$, since in this case x^r is the supremum of a set of positive numbers. When $0 < x < 1$, the second inequality in $(*)$ implies that

$$x^r = \inf\{x^p : p \le r\} \ge x^q > 0,$$

for some rational number $q \ge r$.

Remark. We can develop the properties of powers with real exponents using the tools that we already have, but the proofs are cumbersome and lengthy. So we will postpone this to Chapter 6, in which we define and study the logarithm.

Remark. There are other equivalent ways to define x^r. For example we can use the notion of limit of sequences, as defined in the next chapter, and set

$$x^r := \lim x^{p_i},$$

where p_i is a sequence of rational numbers converging to r. We have to check that the limit exists, and does not depend on the particular sequence p_i, i.e. if p_i, q_i are two sequences of rational numbers converging to r then $|x^{p_i} - x^{q_i}| \to 0$.

1.6 Euclidean Spaces

Definition 1.52. Let $\mathbb{R}^n := \{(x_1, \dots, x_n) : x_i \in \mathbb{R}\}$. The elements of $\mathbb{R}^n$ are called **(n-dimensional) vectors**. Let $x, y \in \mathbb{R}^n$ and $r \in \mathbb{R}$. The **vector addition**, **scalar multiplication**, and **inner product** are defined respectively as follows

$$\begin{aligned} x + y &:= (x_1 + y_1, \dots, x_n + y_n), \\ rx &:= (rx_1, \dots, rx_n), \\ \langle x, y\rangle = x \cdot y &:= x_1y_1 + \cdots + x_ny_n. \end{aligned}$$

Notation. Suppose $r \in \mathbb{R}$ is nonzero, and $x \in \mathbb{R}^n$. Sometimes we use the shorthand notation $\frac{x}{r}$ for $\frac{1}{r}x$.

Theorem. *For every $x, y, z \in \mathbb{R}^n$ and $r \in \mathbb{R}$ we have*
(i) $\langle x + ry, z\rangle = \langle x, z\rangle + r\langle y, z\rangle$.
(ii) $\langle x, y\rangle = \langle y, x\rangle$.
(iii) $\langle x, x\rangle \geq 0$, *and* $\langle x, x\rangle = 0 \iff x = 0$.

Proof. **(i)** We have

$$\begin{aligned} \langle x + ry, z\rangle &= (x_1 + ry_1)z_1 + \cdots + (x_n + ry_n)z_n \\ &= x_1z_1 + \cdots + x_nz_n + r(y_1z_1 + \cdots + y_nz_n) = \langle x, z\rangle + r\langle y, z\rangle. \end{aligned}$$

(ii) $\langle x, y\rangle = x_1y_1 + \cdots + x_ny_n = y_1x_1 + \cdots + y_nx_n = \langle y, x\rangle$.
(iii) We have

$$\langle x, x\rangle = x_1^2 + \cdots + x_n^2 \geq 0.$$

We also have $\langle 0, 0\rangle = 0^2 + \cdots + 0^2 = 0$. Now if $\langle x, x\rangle = 0$ then $x_1^2 + \cdots + x_n^2 = 0$. However $x_j^2 \geq 0$, hence $x_j^2 = 0$ for every j. Thus $x_j = 0$ for every j. So $x = 0$. ■

Definition 1.53. The **norm** or **length** of a vector $x \in \mathbb{R}^n$ is the nonnegative real number

$$|x| := \sqrt{\langle x, x\rangle} = \sqrt{x_1^2 + \cdots + x_n^2}.$$

Theorem 1.54. *For all $x, y \in \mathbb{R}^n$ and $r \in \mathbb{R}$ we have*
(i) $|x| \geq 0$, *and* $|x| = 0 \iff x = 0$.
(ii) $|rx| = |r||x|$.
(iii) **(Cauchy-Schwarz Inequality)** $|\langle x, y\rangle| \leq |x||y|$.
(iv) **(Triangle Inequality)** $|x + y| \leq |x| + |y|$.

Proof. **(i)** Obviously we have $|x| \geq 0$, and $|0| = 0$. Now if $|x| = 0$ then $\langle x, x\rangle = 0$. Therefore $x = 0$.

(ii) We have

$$|rx| = \sqrt{r^2x_1^2 + \cdots + r^2x_n^2} = \sqrt{r^2}\sqrt{x_1^2 + \cdots + x_n^2} = |r||x|.$$

(iii) If $y = 0$ then the inequality holds trivially. So suppose that $y \neq 0$. Let $t \in \mathbb{R}$. Then

$$0 \leq |x + ty|^2 = \langle x + ty, x + ty\rangle = |x|^2 + 2t\langle x, y\rangle + t^2|y|^2.$$

Since this inequality holds for all $t \in \mathbb{R}$, and $|y|^2 > 0$, the discriminant of the above quadratic function in t must be nonpositive, i.e.

$$(\langle x, y\rangle)^2 - |x|^2|y|^2 \leq 0.$$

Hence $(\langle x, y\rangle)^2 \leq |x|^2|y|^2$, and we get the desired by taking the square root of both sides of this inequality.

(iv) We have

$$\begin{aligned} |x + y|^2 &= |x|^2 + |y|^2 + 2\langle x, y\rangle \\ &\leq |x|^2 + |y|^2 + 2|x||y| = (|x| + |y|)^2. \end{aligned}$$

■

Theorem 1.55. *For all $x, y \in \mathbb{R}^n$ we have*

$$\big||x| - |y|\big| \leq |x - y|.$$

Proof. Note that by the triangle inequality we have $|x| \leq |x - y| + |y|$. Therefore $|x| - |y| \leq |x - y|$. By switching x, y we get

$$|y| - |x| \leq |y - x| = |x - y| \implies |x| - |y| \geq -|x - y|.$$

Therefore $\big||x| - |y|\big| \leq |x - y|$. Note that here we have used $|\ |$ to denote both the length of a vector, and the absolute value of a real number. ■

Definition 1.56. For two vectors $x, y \in \mathbb{R}^n$ we define their **Euclidean distance** to be

$$d(x, y) := |x - y| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

d is also called the **Euclidean metric**.

Theorem 1.57. *For all $x, y, z \in \mathbb{R}^n$ we have*

(i) $d(x, y) \geq 0$, *and* $d(x, y) = 0 \iff x = y$.

(ii) $d(x, y) = d(y, x)$.

(iii) $d(x, z) \leq d(x, y) + d(y, z)$.

Proof. (i) We have $d(x, y) = |x - y| \geq 0$, and $d(x, x) = |x - x| = |0| = 0$. Now if $d(x, y) = 0$ then $x - y = 0$, hence $x = y$.

(ii) We have

$$d(x, y) = |x - y| = |(-1)(y - x)| = |-1||y - x| = d(y, x).$$

(iii) We have

$$d(x, z) = |x - z| = |x - y + y - z| \leq |x - y| + |y - z| = d(x, y) + d(y, z). \qquad \blacksquare$$

1.7 Complex Numbers

Definition 1.58. The set $\mathbb{C}$ of **complex numbers** is the set $\mathbb{R}^2$ equipped with the following addition and multiplication

$$\begin{aligned} (a, b) + (c, d) &:= (a + c, b + d), \\ (a, b)(c, d) &:= (ac - bd, ad + bc). \end{aligned}$$

Theorem 1.59. *$\mathbb{C}$ is a field, whose zero and identity are respectively*

$$(0, 0), \text{ and } (1, 0).$$

Also the opposite of a complex number $z = (a, b)$ is

$$-z := (-a, -b),$$

and when z is nonzero its inverse is

$$z^{-1} := \Big(\frac{a}{a^2 + b^2}, \frac{-b}{a^2 + b^2}\Big).$$

Proof. Let $z = (a, b)$, $w = (c, d)$, and $u = (e, f)$ be complex numbers. It is easy to check that addition is associative and commutative:

$$\begin{aligned} z + (w + u) &= \big(a + (c + e), b + (d + f)\big) \\ &= \big((a + c) + e, (b + d) + f\big) = (z + w) + u, \\ z + w &= (a + c, b + d) = (c + a, d + b) = w + z. \end{aligned}$$

We can also easily check that

$$\begin{aligned} (a, b) + (0, 0) &= (a + 0, b + 0) = (a, b), \\ z + (-z) &= (a + (-a), b + (-b)) = (0, 0). \end{aligned}$$

It is obvious that $(1,0) \neq (0,0)$. In addition, we have

$$(a,b)(1,0) = (a1 - b0, a0 + b1) = (a,b).$$

Now let us check that multiplication is associative, commutative, and distributive over addition. We have

$$\begin{aligned} zw &= (ac - bd, ad + bc) = (ca - db, cb + da) = wz, \\ z(wu) &= (a,b)(ce - df, cf + de) \\ &= \big(ace - adf - bcf - bde, acf + ade + bce - bdf\big) \\ &= \big((ac - bd)e - (ad + bc)f, (ac - bd)f + (ad + bc)e\big) \\ &= (ac - bd, ad + bc)(e,f) = (zw)u, \\ z(w + u) &= (a,b)(c + e, d + f) \\ &= (ac + ae - bd - bf, ad + af + bc + be) \\ &= (ac - bd, ad + bc) + (ae - bf, af + be) = zw + zu. \end{aligned}$$

Finally, suppose $z \neq (0,0)$. Then $a \neq 0$ or $b \neq 0$. Hence we must have $a^2 + b^2 > 0$. To simplify the notation let $r := a^2 + b^2$. Then we have

$$zz^{-1} = (a,b)\big(\frac{a}{r}, \frac{-b}{r}\big) = \big(\frac{a^2}{r} - \frac{-b^2}{r}, \frac{-ab}{r} + \frac{ba}{r}\big) = \big(\frac{a^2 + b^2}{r}, 0\big) = (1,0),$$

as desired. ■

Remark. The map $a \mapsto (a,0)$ from $\mathbb{R}$ into $\mathbb{C}$ is a one-to-one map that preserves addition and multiplication, i.e.

$$(a,0) + (b,0) = (a + b, 0), \qquad (a,0)(b,0) = (ab, 0).$$

Thus $\mathbb{C}$ contains a copy of the field $\mathbb{R}$. We will abuse the notation and denote the element $(a,0)$ by a. We also define $i := (0,1)$. Then any complex number $z = (a,b)$ can be written as

$$z = (a,b) = (a,0) + (0,b) = (a,0) + (0,1)(b,0) = a + ib.$$

Note that we have

$$i^2 = (0,1)^2 = (-1,0) = -1,$$

i.e. i is a square root of -1.

Definition 1.60. Let $z = (a,b) = a + ib$ be a complex number. The real numbers a, b are called the **real part** and the **imaginary part** of z, respectively, and we will denote them by

$$a = \operatorname{Re} z, \qquad b = \operatorname{Im} z.$$

The **conjugate** of z is the complex number

$$\bar{z} := (a, -b) = a - ib.$$

The **modulus** or the **absolute value** of z is the nonnegative real number

$$|z| := \sqrt{a^2 + b^2}.$$

Remark. Note that $|z|$ is the length of the vector $(a, b) \in \mathbb{R}^2$. Thus for every $z, w \in \mathbb{C}$ we have

(i) $|z| \geq 0$, and $|z| = 0 \iff z = 0$.

(ii) $|z + w| \leq |z| + |w|$.

Also note that when $a \in \mathbb{R}$ we have $|(a, 0)| = \sqrt{a^2} = |a|$. Thus the absolute value of complex numbers is compatible with the absolute value of real numbers.

Remark. Note that $\mathbb{C}$ does not have a natural order. In fact, there is no order on $\mathbb{C}$ that makes it into an ordered field. The reason is that in any ordered field the square of any nonzero element is positive. In particular $1 = 1^2 > 0$. Thus $-1 < 0$. But in $\mathbb{C}$ we have $i^2 = -1$. Hence if $\mathbb{C}$ was an ordered field, then -1 must have been simultaneously positive and negative, which is impossible.

Remark. We can define the integer powers of complex numbers, similarly to the case of real numbers. Then all the basic properties of powers expressed in Theorem 1.44 also hold for powers of complex numbers, except obviously those properties that are related to the order structure.

Theorem 1.61. *For all $z, w \in \mathbb{C}$ and $n \in \mathbb{Z}$ we have*

(i) $\overline{z + w} = \bar{z} + \bar{w}$.

(ii) $\overline{zw} = \bar{z}\bar{w}$.

(iii) $\bar{\bar{z}} = z$.

(iv) $z\bar{z} = |z|^2$, *hence* $z^{-1} = |z|^{-2}\bar{z}$.

(v) $|\bar{z}| = |z|$.

(vi) $|zw| = |z||w|$.

(vii) $|\operatorname{Re} z| \leq |z|$, *and* $|\operatorname{Im} z| \leq |z|$.

(viii) $z = \bar{z}$ *if and only if* $z \in \mathbb{R}$.

(ix) $z + \bar{z} = 2\operatorname{Re} z$, *and* $z - \bar{z} = 2i\operatorname{Im} z$.

(x) $|z^n| = |z|^n$ *(when $z = 0$ we assume $n > 0$).*

(xi) $\overline{z^n} = \bar{z}^n$ *(when $z = 0$ we assume $n > 0$).*

$\boxed{\text{Proof.}}$ Let $z = a + ib$ and $w = c + id$, where $a, b, c, d \in \mathbb{R}$.

(i) We have

$$\begin{aligned} \overline{z + w} &= \overline{(a + c) + i(b + d)} \\ &= (a + c) - i(b + d) = a - ib + c - id = \bar{z} + \bar{w}. \end{aligned}$$

(ii) We have

$$\begin{aligned}\overline{zw} &= \overline{(ac-bd)+i(ad+bc)} = (ac-bd)-i(ad+bc)\\ &= (ac-(-b)(-d))+i(a(-d)+(-b)c) = (a-ib)(c-id) = \bar{z}\bar{w}.\end{aligned}$$

(iii) $\bar{\bar{z}} = \overline{a-ib} = a-(-ib) = a+ib = z$.

(iv) We have

$$\begin{aligned}z\bar{z} &= (a+ib)(a-ib)\\ &= (a^2-b(-b))+i(a(-b)+ba) = a^2+b^2 = |z|^2.\end{aligned}$$

Thus we have $(|z|^{-2}\bar{z})z = |z|^{-2}|z|^2 = 1$. Hence $|z|^{-2}\bar{z} = z^{-1}$, since the inverse is unique.

(v) $|\bar{z}| = \sqrt{a^2+(-b)^2} = \sqrt{a^2+b^2} = |z|$.

(vi) We have

$$\begin{aligned}|zw| &= \sqrt{(ac-bd)^2+(ad+bc)^2}\\ &= \sqrt{a^2c^2-2acbd+b^2d^2+a^2d^2+2adbc+b^2c^2}\\ &= \sqrt{(a^2+b^2)(c^2+d^2)} = \sqrt{a^2+b^2}\sqrt{c^2+d^2} = |z||w|.\end{aligned}$$

(vii) $|\operatorname{Re} z| = |a| = \sqrt{a^2} \leq \sqrt{a^2+b^2} \leq |z|$. Note that we have used the monotonicity of the square root function over nonnegative real numbers. The other inequality can be proved similarly.

(viii) $z = \bar{z} \iff b = -b \iff b = 0 \iff z = a \in \mathbb{R}$.

(ix) $z+\bar{z} = a+ib+a-ib = 2a = 2\operatorname{Re} z$. The other one is similar.

(x) The proof is by induction on n, when $n > 0$. The claim holds obviously for $n = 1$, so suppose it also holds for n. Then for $n+1$ we have

$$|z^{n+1}| = |z^n z| = |z^n||z| = |z|^n|z| = |z|^{n+1}.$$

Now suppose $z \neq 0$. When $n = 0$ both sides of the equation are one. For $n = -1$ we have

$$|z^{-1}| = ||z|^{-2}\bar{z}| = ||z|^{-2}||\bar{z}| = |z|^{-2}|z| = |z|^{-1}.$$

Note that $|z|^{-2}$ is a positive real number, therefore its modulus equals its absolute value as a real number, which is itself. Finally for $n = -m < 0$ we have

$$|z^n| = |(z^{-1})^m| = |z^{-1}|^m = (|z|^{-1})^m = |z|^n.$$

(xi) The proof is by induction on n, when $n > 0$. The claim holds obviously for $n = 1$, so suppose it also holds for n. Then for $n+1$ we have

$$\overline{z^{n+1}} = \overline{z^n z} = \overline{z^n}\,\bar{z} = \bar{z}^n\bar{z} = \bar{z}^{n+1}.$$

Now suppose $z \neq 0$. When $n = 0$ both sides of the equation are one. For $n = -1$ we have

$$\overline{z^{-1}} = \overline{|z|^{-2}\bar{z}} = \overline{|z|^{-2}}\,\overline{\bar{z}} = |z|^{-2}\overline{\bar{z}} = |\bar{z}|^{-2}\overline{\bar{z}} = (\bar{z})^{-1}.$$

Note that $|z|^{-2}$ is a real number, therefore its conjugate is itself. Finally for $n = -m < 0$ we have

$$\overline{z^n} = \overline{(z^{-1})^m} = \left(\overline{(z^{-1})}\right)^m = ((\bar{z})^{-1})^m = (\bar{z})^{-m} = \bar{z}^n. \qquad ■$$

Remark. Suppose $z \in \mathbb{C}$ is nonzero. Then $r := |z| > 0$. Now $\frac{z}{r}$ has modulus one, so it belongs to the unit circle in $\mathbb{C}$. Hence by Theorem 6.59, there is a unique $\theta \in [0, 2\pi)$ such that $\frac{z}{r} = e^{i\theta} = \cos\theta + i\sin\theta$. Therefore

$$z = re^{i\theta} = r(\cos\theta + i\sin\theta).$$

This is called the **polar representation** of z. The number θ is called the **argument** of z, and is denoted by $\arg z$. In fact θ is the signed angle between the segment connecting z and 0, and the half line of nonnegative real numbers.

Remark. Suppose $z = re^{i\theta}$ and $w = se^{i\phi}$. Then by Theorem 6.56 we have

$$zw = rse^{i(\theta+\phi)}.$$

The interpretation of this formula is that when you multiply a complex number w by a complex number z, you scale the modulus of w by the modulus of z, and you rotate w around the origin by the angle $\arg z$.

1.8 Polynomials

Definition 1.62. A **polynomial** is a function $p : \mathbb{R} \to \mathbb{R}$ for which there are $a_0, \dots, a_n \in \mathbb{R}$, called the **coefficients** of p, such that for all $x \in \mathbb{R}$ we have

$$p(x) = a_n x^n + \cdots + a_1 x + a_0.$$

If $a_n \neq 0$ then we define the **degree** of the polynomial to be n, and we denote it by $\deg p$. Similarly a function $p : \mathbb{C} \to \mathbb{C}$ is a polynomial if

$$p(z) = a_n z^n + \cdots + a_1 z + a_0,$$

for some $a_0, \dots, a_n \in \mathbb{C}$, and all $z \in \mathbb{C}$.

Remark. The zero polynomial is the function p such that $p(x) = 0$ for all x. We define the degree of the zero polynomial to be $-\infty$.

Remark. Note that we can regard a polynomial with real coefficients as either a function on $\mathbb{R}$, or a function on $\mathbb{C}$.

Theorem 1.63. *The degree and coefficients of a polynomial are uniquely determined.*

Proof. We prove the theorem for complex polynomials. The real case is similar. Suppose to the contrary that a polynomial p has two different representations

$$a_n z^n + \cdots + a_0 = p(z) = b_m z^m + \cdots + b_0.$$

We can move every term to one side of the equality, and subtract the coefficients of the same powers of z, to obtain

$$c_k z^k + \cdots + c_0 = 0,$$

for all $z \in \mathbb{C}$. Now if $n \neq m$, or $a_i \neq b_i$ for some i, then some $c_j \neq 0$. We show that this results in a contradiction. Let k be the largest integer for which $c_k \neq 0$. Then after some rearrangement of terms we have

$$\begin{aligned} z^k &= -c_k^{-1} c_{k-1} z^{k-1} - \cdots - c_k^{-1} c_1 z - c_k^{-1} c_0 \\ &=: \alpha_{k-1} z^{k-1} + \cdots + \alpha_1 z + \alpha_0. \end{aligned}$$

Now let $z_0 = 1 + |\alpha_0| + \cdots + |\alpha_{k-1}|$. Then for $i \leq k-1$ we have $|z_0|^i \leq |z_0|^{k-1}$, since $|z_0| = z_0 > 1$. Hence we have

$$\begin{aligned} |\alpha_{k-1} z_0^{k-1} + \cdots + \alpha_1 z_0 + \alpha_0| &\leq |\alpha_{k-1}||z_0|^{k-1} + \cdots + |\alpha_1||z_0| + |\alpha_0| \\ &\leq (|\alpha_{k-1}| + \cdots + |\alpha_1| + |\alpha_0|)|z_0|^{k-1} \\ &< |z_0||z_0|^{k-1} = |z_0|^k = |z_0^k|, \end{aligned}$$

which is a contradiction. ■

Definition 1.64. Let p be a polynomial. If a number a satisfies $p(a) = 0$, then we say a is a **root** of p.

Theorem 1.65. *Suppose p is a nonzero polynomial, and a is a number. Then $p(a) = 0$ if and only if there is a nonzero polynomial q such that*

$$p(x) = (x - a)q(x),$$

and $\deg q = \deg p - 1$. As a result, the number of distinct roots of a nonzero polynomial p is at most $\deg p$.

Proof. If $p = (x - a)q$ then $p(a) = 0$ obviously. Conversely, suppose $p(a) = 0$, where

$$p(x) = a_n x^n + \cdots + a_1 x + a_0,$$

with $a_n \neq 0$. Note that we must have $n \geq 1$, since otherwise p cannot have a root. Then by the binomial theorem we have

$$\begin{aligned} p(x) &= a_n(x-a+a)^n + \cdots + a_1(x-a+a) + a_0 \\ &= a_n(x-a)^n + b_{n-1}(x-a)^{n-1} + \cdots + b_1(x-a) + b_0, \end{aligned}$$

for some $b_0, \dots, b_{n-1}$. Note that the coefficients of $(x-a)^n$ is a_n, since $(x-a)^n$ appears with coefficient one in the expansion of $(x-a+a)^n$, and other terms $(x-a+a)^i$ with $i < n$ do not produce $(x-a)^n$. Now by evaluating at $x = a$ we get $b_0 = p(a) = 0$. Hence

$$p(x) = (x-a)[a_n(x-a)^{n-1} + \cdots + b_1] =: (x-a)q(x).$$

Note that q is also a polynomial, since again by the binomial theorem we have

$$q(x) = a_n(x-a)^{n-1} + \cdots + b_2(x-a) + b_1 = a_n x^{n-1} + c_{n-2}x^{n-2} + \cdots + c_0,$$

for some $c_0, \dots, c_{n-2}$. By the same argument as above, the coefficient of x^{n-1} is a_n. But $n-1 \geq 0$, and $a_n \neq 0$, hence q is a nonzero polynomial of degree $n-1$.

The last statement can be proved by induction on $\deg p$. Nonzero polynomials of degree zero are constant polynomials which have no root. Suppose the claim holds for all polynomials with degree less than $\deg p$. If p has no root, then there is nothing to prove. So let a be a root of p. Then $p = (x-a)q$. We know that $\deg q = \deg p - 1$. Now if b is another root of p we must have $q(b) = 0$. But q has at most $\deg q$ distinct roots, hence p has at most $\deg q + 1 = \deg p$ distinct roots. ■

Theorem 1.66. *Suppose p is a polynomial with real coefficients, and $\alpha \in \mathbb{C}$ is a root of p. Then $\bar{\alpha}$ is also a root of p.*

Proof. Suppose $p(z) = a_n z^n + \cdots + a_0$ where $a_i \in \mathbb{R}$. Then we have

$$\begin{aligned} p(\bar{\alpha}) &= a_n \bar{\alpha}^n + \cdots + a_1\bar{\alpha} + a_0 \\ &= \bar{a}_n \overline{\alpha^n} + \cdots + \bar{a}_1\bar{\alpha} + \bar{a}_0 \\ &= \overline{a_n \alpha^n + \cdots + a_1\alpha + a_0} = \overline{p(\alpha)} = \bar{0} = 0. \end{aligned}$$ ■

Definition 1.67. Let F denote $\mathbb{R}$ or $\mathbb{C}$. A **polynomial in *n* variables** is a function $p : F^n \to F$ that is the sum of finitely many functions of the form

$$(x_1, \dots, x_n) \mapsto c x_1^{m_1} \cdots x_n^{m_n},$$

where $c \in F$, and m_i's are nonnegative integers. Each term $c x_1^{m_1} \cdots x_n^{m_n}$ is called a **monomial**, and the numbers c are called the **coefficients** of p.

Remark. The coefficients of a multivariable polynomial are also uniquely determined by the polynomial. There is an easy proof of this fact that uses partial derivatives. See Exercise 7.38.

1.9 Countable Sets

Definition 1.68. A function is called **injective** or an injection if it is one-to-one. A function is called **surjective** or a surjection if it is onto. Finally, a function is called **bijective** or a bijection if it is one-to-one and onto.

Definition 1.69. A set A is **finite** if there exists a bijection from A onto the set $\{1, 2, \dots, n\}$ for some $n \in \mathbb{N}$. A set that is not finite is **infinite**. We consider the empty set as a finite set.

A set A is **countably infinite** or **denumerable** if there exists a bijection from A onto $\mathbb{N}$. A set is **countable** if it is finite or countably infinite. A set that is not countable is **uncountable**.

Theorem 1.70. *We have*

(i) *A set A is countable if and only if there exists a surjective map from $\mathbb{N}$ to A.*

(ii) *A set A is countable if and only if there exists an injective map from A to $\mathbb{N}$.*

(iii) *A subset of a countable set is countable.*

(iv) *The Cartesian product of finitely many countable sets is countable.*

(v) *The union of countably many countable sets is countable.*

Theorem 1.71. $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ *are countably infinite.*

Proof. $\mathbb{N}$ is countably infinite, since the identity map is a bijection from $\mathbb{N}$ onto $\mathbb{N}$. For $\mathbb{Z}$ we have the bijection

$$f(n) := \begin{cases} 2n & n > 0 \\ -2n+1 & n \leq 0. \end{cases}$$

Finally, $\mathbb{Q}$ is infinite and can be written as the union of countably many countable sets as follows

$$\mathbb{Q} = \bigcup_{q \in \mathbb{N}} \{\frac{p}{q} : p \in \mathbb{Z}\}.$$

■

Theorem 1.72. *The set of all sequences of* $0, 1$, *i.e. the set of all functions from* $\mathbb{N}$ *into* $\{0, 1\}$, *is uncountable.*

Proof. Let f be an injective map from $\mathbb{N}$ into the set of all sequences of $0, 1$. We show that f cannot be surjective. We use the diagonal method due to Cantor. Consider

$$\begin{aligned} f(1) &= a_{11}a_{12}a_{13}\cdots \\ f(2) &= a_{21}a_{22}a_{23}\cdots \\ &\vdots \end{aligned}$$

where $a_{ij} \in \{0, 1\}$. Now we define the sequence $b = b_1 b_2 b_3 \cdots$ as follows

$$b_i := \begin{cases} 1 & \text{if } a_{ii} = 0, \\ 0 & \text{if } a_{ii} = 1. \end{cases}$$

Then $b \neq f(n)$ for each $n \in \mathbb{N}$, since $b_n \neq a_{nn}$. Hence b is not in the image of f, and f is not onto. ■

Theorem 1.73. *$\mathbb{R}$ is uncountable.*

Proof. The proof is similar to the last theorem. We only need to use the decimal expansion of real numbers instead of the sequences. ■

Theorem 1.74. *Suppose $a, b \in \mathbb{R}$, and $a < b$. Then the intervals $(a, b), [a, b]$ are uncountable.*

Proof. We will show that there are bijections from the $\mathbb{R}$ onto (a, b). Then it follows that the (a, b) cannot be countable, since otherwise $\mathbb{R}$ would be countable too. It also follows that $[a, b]$ is uncountable, since $(a, b) \subset [a, b]$.

Now the function

$$x \mapsto \frac{x}{1 + |x|}$$

is a bijection from $\mathbb{R}$ onto $(-1, 1)$ (why?), and $x \mapsto a + \frac{x+1}{2}(b - a)$ is a bijection from $(-1, 1)$ onto (a, b). So their composition is the required bijection from $\mathbb{R}$ onto (a, b). ■

Chapter 2

Metric Spaces

2.1 Topology of Metric Spaces

Definition 2.1. A **metric space** (X, d), is a nonempty set X equipped with a map

$$d : X \times X \to \mathbb{R},$$

called the **metric** or the **distance function**, such that

(i) d is *positive definite*, i.e. for every $x, y \in X$ we have

$$d(x, y) \geq 0, \text{ and } d(x, y) = 0 \iff x = y.$$

(ii) d is *symmetric*, i.e. for every $x, y \in X$ we have

$$d(x, y) = d(y, x).$$

(iii) d satisfies the **triangle inequality**, i.e. for every $x, y, z \in X$ we have

$$d(x, z) \leq d(x, y) + d(y, z).$$

Remark. We refer to the elements of a metric space as *points* of the metric space.

Notation. We denote the metric of a metric space X by d_X. When it causes no confusion, we simply denote the metric by d.

***Example* 2.2.** $\mathbb{R}^n$ with its standard Euclidean metric is a metric space.

***Example* 2.3.** On any set we can define the distance between any two distinct elements to be 1. This is a metric called the **discrete metric**.

***Example* 2.4.** Suppose (X, d) is a metric space and $A \subset X$. Then $d|_{A\times A}$ is a metric on A; and A is called a **subspace** of X. For example the $\boldsymbol{n}$**-sphere**

$$S^n := \{x \in \mathbb{R}^{n+1} : |x| = 1\}$$

is a subspace of $\mathbb{R}^{n+1}$.

Definition 2.5. A **sequence** (a_n) in a set X is a function

$$\begin{aligned} \mathbb{N} &\to X \\ n &\mapsto a_n \end{aligned}.$$

We also denote this sequence by $(a_n)_{n\in\mathbb{N}}$. A sequence $(b_k)_{k\in\mathbb{N}}$ is called a **subsequence** of $(a_n)_{n\in\mathbb{N}}$ if $b_k = a_{n_k}$, for a strictly increasing sequence $n_1 < n_2 < \cdots$ of positive integers.

Definition 2.6. Suppose (a_n) is a sequence in the metric space X. We say the sequence (a_n) converges to the **limit** $a \in X$, and write $\lim a_n = a$ or $a_n \to a$, if

$$\forall \epsilon > 0\ \exists N \in \mathbb{N} \text{ such that } \forall n \geq N \text{ we have } d(a_n, a) < \epsilon.$$

Remark. It follows immediately from the definition that a sequence $a_n \to a$ if and only if $d(a_n, a) \to 0$ as a sequence in $\mathbb{R}$.

Remark. When we want to emphasize the index with respect to which we take the limit, we write $\lim_{n\to\infty} a_n = a$, or we say $a_n \to a$ as $n \to \infty$.

Theorem 2.7. *The limit of a convergent sequence is unique.*

$\boxed{\textbf{Proof.}}$ Suppose to the contrary that $a_n \to a$ and $a_n \to b$ with $a \neq b$. Let $0 < \epsilon < \frac{1}{2}d(a,b)$. Then for n large enough we have $d(a_n, a) < \epsilon$ and $d(a_n, b) < \epsilon$, which is a contradiction. ■

Theorem 2.8. *Every subsequence of a convergent sequence converges to the same limit as the original sequence.*

$\boxed{\textbf{Proof.}}$ Suppose $a_n \to a$ and $b_k = a_{n_k}$. Given $\epsilon > 0$ there is N so that $d(a_n, a) < \epsilon$ for $n \geq N$. Now since $n_k \geq k$, the same N works for (b_k). ■

Definition 2.9. Suppose X is a metric space, and r is a positive real number. Let $x \in X$. The set

$$B_r(x) := \{y \in X : d(y, x) < r\}$$

is called the **open ball** of radius r around x. A set $U \subset X$ is **open** if

$$\forall x \in U\ \exists r > 0 \text{ such that } B_r(x) \subset U.$$

A **neighborhood** of a point x is a set that contains an open set containing x.

Definition 2.10. A set $C \subset X$ is **closed** if the limit of every convergent sequence of points in C belongs to C, i.e. if $a_n \in C$ for every n, and $a_n \to a$, then $a \in C$.

Remark. In other words, a set is closed if it is closed under taking the limit of sequences.

Proposition 2.11. *Open balls are open.*

Proof. If $y \in B_r(x)$ then $s := r - d(y,x) > 0$. Now we have $B_s(y) \subset B_r(x)$, since by the triangle inequality for $z \in B_s(y)$ we have

$$d(z,x) \leq d(z,y) + d(y,x) < s + d(y,x) = r.$$ ■

Theorem 2.12. *A set is open if and only if its complement is closed.*

Proof. Suppose U is open. Take a sequence (a_n) in U^c, and assume that $a_n \to a$. We have to show that $a \in U^c$. If this does not happen we would have $a \in U$. Therefore $B_\epsilon(a) \subset U$ for some $\epsilon > 0$. But for large enough n we have $d(a_n, a) < \epsilon$ i.e. $a_n \in B_\epsilon(a)$, which is a contradiction.

On the other hand suppose that A is closed. Let $a \in A^c$. We need to show that $B_r(a) \subset A^c$ for some $r > 0$. If this does not hold then $B_r(a) \cap A$ is nonempty for all positive r. Let a_n be an element of $B_{\frac{1}{n}}(a) \cap A$. Now for any $\epsilon > 0$ there is $N \in \mathbb{N}$ such that $\frac{1}{N} < \epsilon$, due to the Archimedean property. Hence for $n \geq N$ we have

$$d(a_n, a) < \frac{1}{n} \leq \frac{1}{N} < \epsilon.$$

Thus $a_n \to a$, and we must have $a \in A$, which is again a contradiction. ■

Proposition 2.13. *Open intervals in $\mathbb{R}$ are open, and closed intervals are closed.*

Proof. The openness of open intervals is easy to show. Let $x \in (a,b)$, where a, b can be respectively $-\infty$ or $+\infty$ too. Then for $r = \min\{1, x-a, b-x\}$ we have

$$B_r(x) = (x-r, x+r) \subset (a,b).$$

The complement of a closed interval is either empty, an open interval, or the union of two disjoint open intervals, depending on whether the closed interval is $\mathbb{R}$, an unbounded interval other than $\mathbb{R}$, or a bounded interval. The empty set and an open interval are open. The union of two disjoint open intervals is also open as the following theorem shows (or we can prove it similarly to the above). ■

***Example* 2.14.** The subset $[0,1)$ of $\mathbb{R}$ is neither open nor closed. It is not open since it does not contain any open ball around 0; and it is not closed since it contains the convergent sequence $(1 - \frac{1}{n})_{n\in\mathbb{N}}$ but not its limit 1.

Definition 2.15. The family of all open subsets of X is called the **topology** of X.

Theorem 2.16. *The topology has the following properties*

(i) *$\emptyset, X$ are open sets.*

(ii) *The union of any collection of open sets is an open set.*

(iii) *The intersection of finitely many open sets is an open set.*

$\boxed{\text{Proof.}}$ **(i)** For every $a \in X$ and $r > 0$ we have $B_r(a) \subset X$, since by definition $B_r(a)$ is a subset of X. So X is open. Also, the sentence "for every $a \in \emptyset$ there is $r > 0$ such that $B_r(a) \subset \emptyset$" is vacuously true, since $\emptyset$ has no elements. Thus $\emptyset$ is open, as desired.

(ii) Let $\{U_\alpha\}_{\alpha\in I}$ be a collection of open sets. If $a \in \bigcup_{\alpha\in I} U_\alpha$ then $a \in U_{\alpha_0}$ for some α_0. Hence there is $r > 0$ such that $B_r(a) \subset U_{\alpha_0} \subset \bigcup_{\alpha\in I} U_\alpha$.

(iii) Let $a \in U_1 \cap \cdots \cap U_k$ where U_i's are open. Then a belongs to each U_i. Therefore there are positive numbers r_i so that $B_{r_i}(a) \subset U_i$. Now for $r = \min_{i\le k} r_i$ we have $B_r(a) \subset B_{r_i}(a) \subset U_i$, for every i. Hence we have

$$B_r(a) \subset U_1 \cap \cdots \cap U_k.$$

Note that the finiteness of the number of U_i's is needed to ensure that $r > 0$. ■

Theorem 2.17. *The family of closed sets has the following properties*

(i) $\emptyset, X$ *are closed sets.*

(ii) *The intersection of any collection of closed sets is a closed set.*

(iii) *The union of finitely many closed sets is a closed set.*

$\boxed{\text{Proof.}}$ Take the complement of the respective properties for open sets, and use De Morgan's laws. ■

***Exercise* 2.18.** Prove the above theorem using the definition of closed sets.

***Example* 2.19.** The intersection of infinitely many open sets is not open in general, and the union of infinitely many closed sets is not closed in general. For example

$$\bigcap_{n\ge1}(-\frac{1}{n},1) = [0,1), \qquad \bigcup_{n\ge1}[\frac{1}{n},1] = (0,1].$$

Theorem 2.20. *Every open set in $\mathbb{R}$ is the union of countably many disjoint open intervals.*

$\boxed{\text{Proof.}}$ Let $U \subset \mathbb{R}$ be an open set. Define an equivalence relation on U as follows

$$x \sim y \iff x, y \text{ belong to an open interval } I \subset U.$$

It is easy to show that $\sim$ is an equivalence relation. We know that U is the disjoint union of the equivalence classes; so it is enough to show that each equivalence class is an open interval, and there are at most countably many different equivalence classes.

Let I_x be the equivalence class of $x \in U$. Let α, β be the infimum and the supremum of I_x respectively. We want to show that $I_x = (\alpha, \beta)$. If any number

$r \geq \beta$ belongs to I_x, then there is an open interval I containing x, r. Thus in particular I_x must contain numbers greater than β, which is impossible. The case of $r \leq \alpha$ is similar; so we have $I_x \subset (\alpha, \beta)$. To prove the other inclusion, suppose there is $r \in (\alpha, \beta)$ that does not belong to I_x. Suppose for example that $r > x$. Then since r is not the supremum of I_x, there is $s \in (r, \beta) \cap I_x$. Therefore there is an open interval $I \subset U$ such that $x, s \in I$. But then we must have $r \in I$, and consequently $r \in I_x$. This contradiction gives the desired result.

Finally, countability of the number of intervals follows, since we can choose a distinct rational number from each interval. ■

Remark. The above characterization of open subsets of $\mathbb{R}$ is a consequence of its special order structure. There is no similar simple description of closed subsets of $\mathbb{R}$, nor of open subsets of $\mathbb{R}^n$ for $n > 1$.

Definition 2.21. Suppose X is a metric space, and $A \subset X$. The **closure** of A, denoted by $\bar{A}$, is the set of limits of all convergent sequences of points in A. In other words, $\bar{A}$ is the set of points $a \in X$ for which there is a sequence (a_n) in A such that $a_n \to a$.

Theorem 2.22. *Suppose X is a metric space, and $A \subset X$. Then $\bar{A}$ is the smallest closed set that contains A, i.e. it is closed, contains A, and is contained in any closed set containing A.*

Proof. First let us show that $\bar{A}$ is closed. Let (a_n) be a sequence in $\bar{A}$, and suppose $a_n \to a$. We must prove that $a \in \bar{A}$. Since $a_n \in \bar{A}$, there is a sequence $(b_{k,n})_{k \in \mathbb{N}}$ in A such that

$$b_{k,n} \to a_n \qquad \text{as } k \to \infty.$$

Thus there is $k(n) \in \mathbb{N}$ such that $d(b_{k(n),n}, a_n) < \frac{1}{n}$. Now we claim that

$$b_{k(n),n} \to a \qquad \text{as } n \to \infty.$$

The reason is

$$d(b_{k(n),n}, a) \leq d(b_{k(n),n}, a_n) + d(a_n, a) < \frac{1}{n} + d(a_n, a) \underset{n \to \infty}{\longrightarrow} 0.$$

Hence by the definition of $\bar{A}$ we have $a \in \bar{A}$.

It is obvious that $A \subset \bar{A}$, since every $a \in A$ is the limit of the constant sequence (a). Finally, suppose $C \supset A$ is closed. Then every convergent sequence in A is also in C, so the limits of those convergent sequences are in C. Therefore $C \supset \bar{A}$. ■

***Exercise* 2.23.** Suppose X is a metric space, and $A \subset X$. Show that $\bar{A} = A$ if and only if A is closed. As a result we have $\bar{\bar{A}} = \bar{A}$.

***Exercise* 2.24.** Give an example of a metric space in which the closure of some open ball is not the corresponding closed ball, i.e.

$$\overline{B_r(x)} \neq \{y \in X : d(y,x) \leq r\},$$

for some r, x. But show that we always have $\overline{B_r(x)} \subset \{y \in X : d(y,x) \leq r\}$.

Proposition 2.25. *Suppose A is a bounded above (below) nonempty subset of $\mathbb{R}$. Then the supremum (infimum) of A belongs to $\bar{A}$. In particular, a closed and bounded nonempty subset of $\mathbb{R}$ contains its supremum and infimum.*

Proof. Suppose s is the supremum of A. Then for every $n \in \mathbb{N}$ there is $a_n \in A$ such that $s - \frac{1}{n} < a_n \leq s$, since $s - \frac{1}{n}$ is not an upper bound of A. It is easy to see that $a_n \to s$. Thus we must have $s \in \bar{A}$. The case of infimum is similar. ■

Definition 2.26. Suppose X is a metric space, and $A \subset X$. A point $x \in X$ is called a **limit point** or an **accumulation point** of $A \subset X$, if every $B_r(x)$ intersects A in a point other than x. If a point $y \in A$ is not a limit point of A, we call it an **isolated point** of A.

***Example* 2.27.** Let $A = \{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\}$. Then 0 is a limit point of A. Also, every point of A is an isolated point.

Theorem 2.28. *Suppose X is a metric space, and $A \subset X$. A point a is a limit point of A if and only if there exists a sequence of distinct points of A converging to a.*

Proof. If there exists a sequence of distinct points (a_n) in A that converges to a, then every $B_r(a)$ contains a_n's for n large enough. Hence $B_r(a)$ contains at least one point of A other than a.

Now suppose a is a limit point of A. We want to build a sequence of distinct points of A that converges to a. Let a_1 be an arbitrary point of $(A - \{a\}) \cap B_1(a)$. Suppose we have chosen $a_1, \dots, a_n$, then let a_{n+1} be a point of $(A - \{a\}) \cap B_{r_{n+1}}(a)$, where

$$r_{n+1} = \min\{\frac{1}{n+1}, d(a_1, a), \dots, d(a_n, a)\}.$$

Now we have $a_n \to a$, since $d(a_n, a) < r_n \leq \frac{1}{n}$. It is also obvious that a_n's are distinct, because if $a_m = a_n$ for some $m > n$ then we have $d(a_m, a) < r_m \leq d(a_n, a)$ which is impossible. ■

Remark. A consequence of the above theorem is that if a is a limit point of A then any open ball around a contains infinitely many points of A.

Theorem 2.29. *The closure of a set is the union of the set and its limit points.*

Proof. Let A be the set. Then $A \subset \bar{A}$. Also by the above theorem, every limit point of A is the limit of a sequence of points in A, so it belongs to $\bar{A}$. Thus we only need to show that every point of $\bar{A}$ is either an element of A or a limit point of A. Let $a \in \bar{A}$ and suppose that $a \notin A$. It is enough to show that a is a limit point of A. We know that there is a sequence (a_n) in A that converges to a. Thus every open ball around a contains some a_n. But $a_n \neq a$ as a does not belong to A. Hence every open ball around a contains some point of A other than a. ■

***Example* 2.30.** Let $A = \{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\}$. Then $\bar{A} = \{0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots\}$ (why?).

Definition 2.31. A subset A of a metric space X is **dense** in X, if $\bar{A} = X$.

***Example* 2.32.** $\mathbb{Q}$ is dense in $\mathbb{R}$. Because $\frac{\lfloor nx \rfloor}{n} \to x$ for every $x \in \mathbb{R}$.

Definition 2.33. Suppose X is a metric space, and $A \subset X$. The **interior** of A is

$$A^\circ := \{x \in A : \exists r > 0 \text{ such that } B_r(x) \subset A\},$$

and the **boundary** of A is $\partial A := \bar{A} - A^\circ$.

Remark. By definition we have $A^\circ \cap \partial A = \emptyset$, and $\bar{A} = A^\circ \cup \partial A$. It is also easy to show that $\bar{A} = A \cup \partial A$.

***Exercise* 2.34.** Show that

(i) A° is the largest open set contained in A, i.e. it is open, is contained in A, and contains any open subset of A.
(ii) $A^\circ = A$ if and only if A is open. Hence $A^{\circ\circ} = A^\circ$.
(iii) $\partial A = \bar{A} \cap \overline{A^c}$; hence ∂A is closed.
(iv) A is closed if and only if $\partial A \subset A$.
(v) A is open if and only if $A \cap \partial A = \emptyset$.
(vi) $\partial A^c = \partial A$.

***Exercise* 2.35.** Show that the closure of a set A is the set of points that any open neighborhood of them intersects A. As a result, ∂A is the set of points that any open neighborhood of them intersects both A, A^c.

***Exercise* 2.36.** Show that

(i) If $A \subset B$ then $\bar{A} \subset \overline{B}$, and $A^\circ \subset B^\circ$.
(ii) $\overline{(A \cup B)} = \bar{A} \cup \overline{B}$.
(iii) $(A \cup B)^\circ \supset A^\circ \cup B^\circ$.
(iv) $\partial(A \cup B) \subset \partial A \cup \partial B$.
(v) $\overline{(A \cap B)} \subset \bar{A} \cap \overline{B}$.
(vi) $(A \cap B)^\circ = A^\circ \cap B^\circ$.
(vii) Show that the equality does not necessarily hold in (iii), (iv), and (v).

(viii) Is there any relation between $\partial(A \cap B)$ and $\partial A \cap \partial B$?

***Exercise* 2.37.** Let (a_n) be a sequence in a metric space X. Show that the set of all **subsequential limits** of (a_n), i.e. the limits of all subsequences of (a_n), is a closed set. Note that this set of subsequential limits is not the same as $\overline{\{a_n\}}$, since it does not necessarily contain the a_n's themselves. However, show that the limit points of the set $\{a_n\}$ are subsequential limits of the sequence (a_n).

Definition 2.38. Two metrics d_1 and d_2 on X are **equivalent** if there exist $c, C > 0$ such that

$$c\, d_1(x,y) \le d_2(x,y) \le C\, d_1(x,y)$$

for all $x, y \in X$.

Remark. Note that the equivalence of metrics is an equivalence relation.

Theorem 2.39. *Two equivalent metrics d_1, d_2 induce the same topology, i.e. their open sets and closed sets are the same. In addition, if $a_n \to a$ with respect to d_1, then $a_n \to a$ with respect to d_2 as well.*

Proof. We use $B_r(x, d_i)$ to denote the open balls with respect to d_i. Suppose U is an open set with respect to d_1, i.e. for every $x \in U$ there is $r > 0$ such that $B_r(x, d_1) \subset U$. To show that U is open with respect to d_2 it suffices to show that $B_s(x, d_2) \subset B_r(x, d_1) \subset U$ for some $s > 0$. But we have $B_{cr}(x, d_2) \subset B_r(x, d_1)$, because if $d_2(y, x) < cr$ then

$$d_1(y,x) \le \frac{1}{c} d_2(y,x) < \frac{1}{c} cr = r.$$

Thus the theorem is proved for open sets. The result for closed sets follows from the duality between open sets and closed sets.

Finally, suppose $a_n \to a$ with respect to d_1. Then for any $\epsilon > 0$ there is $N \in \mathbb{N}$ so that for $n \ge N$ we have $d_1(a_n, a) < \frac{\epsilon}{C}$. Hence for $n \ge N$ we have $d_2(a_n, a) \le C d_1(a_n, a) < C \frac{\epsilon}{C} = \epsilon$. Therefore $a_n \to a$ with respect to d_2 as well. ■

***Exercise* 2.40.** Give an example of two metrics on a space that induce the same topology, but are not equivalent.

2.2 Subspaces and Products

Definition 2.41. A subset A of a metric space X, is itself a metric space with the induced (or inherited) metric $d|_{A \times A}$. We call A a **subspace** of X.

Theorem 2.42. *Suppose Y is a subspace of X. Then $V \subset Y$ is open in Y if and only if there is an open subset $U \subset X$ such that $V = U \cap Y$. The same is true for closed sets.*

Proof. Let $z \in V$. We use the notations $B_r(z, X)$ and $B_r(z, Y)$ for the open balls around z in X, Y respectively. Note that we have

$$B_r(z, Y) = \{y \in Y : d(y, z) < r\} = B_r(z, X) \cap Y.$$

Now, since V is open in Y, there is $r_z > 0$ such that $B_{r_z}(z, Y) \subset V$. Set

$$U := \bigcup_{z \in V} B_{r_z}(z, X).$$

Then U is a union of open balls in X, hence it is open in X. In addition

$$U \cap Y = \bigcup_{z \in V} [B_{r_z}(z, X) \cap Y] = \bigcup_{z \in V} B_{r_z}(z, Y) = V.$$

Next suppose U is open in X. Then for any $z \in U \cap Y$ we have $B_r(z, X) \subset U$, for some $r > 0$. Hence

$$B_r(z, Y) = B_r(z, X) \cap Y \subset U \cap Y.$$

Thus $U \cap Y$ is open in Y. Finally, the result for closed sets follows by taking the complement of the results for open sets. For example if C is closed in X, then C^c is open in X. Therefore $C^c \cap Y$ is open in Y. Hence its complement in Y is closed in Y. But its complement in Y is

$$Y - (C^c \cap Y) = Y \cap (C^c \cap Y)^c = Y \cap (C \cup Y^c) = Y \cap C.$$

The other direction is similar. ■

***Example* 2.43.** The open and closed subsets of a subspace are not necessarily open or closed in the larger space. For example, $(0, 1)$ is an open subset of $\mathbb{R}$, but if we identify $\mathbb{R}$ with the x-axis of $\mathbb{R}^2$, $(0, 1)$ is not open in $\mathbb{R}^2$.

***Example* 2.44.** The closure of a set in a subspace is also not necessarily the same as its closure in the larger space. For example when we consider $(0, 1)$ as a subset of the metric space $X = (0, 1)$, it is closed, since any metric space is closed in itself. Thus the closure of $(0, 1)$ as a subset of X is $(0, 1)$. But the closure of $(0, 1)$ as a subset of $\mathbb{R}$ is $[0, 1]$.

Theorem 2.45. *Suppose A is a subspace of X. Then*

(i) *Open sets in A are open in X when A is open in X.*

(ii) *Closed sets in A are closed in X when A is closed in X.*

Proof. Open sets in A are of the form $U \cap A$ where U is open in X. Hence if A is open in X, $U \cap A$ is also open in X. The proof is the same for closed sets. ■

Theorem 2.46. *Suppose $(X_1, d_1), \dots, (X_n, d_n)$ are metric spaces. Then on the* ***product space***

$$\prod_{i=1}^{n} X_i := X_1 \times \cdots \times X_n$$

there are three equivalent metrics

$$d(x,y) := \left[\sum_{i=1}^{n} d_i(x_i, y_i)^2\right]^{\frac{1}{2}},$$
$$d_{\text{sum}}(x,y) := \sum_{i=1}^{n} d_i(x_i, y_i),$$
$$d_{\max}(x,y) := \max_{i \le n} \{d_i(x_i, y_i)\}.$$

Proof. All of these functions satisfy the first two conditions of a metric obviously. Also it is easy to see that d_{sum} satisfies the triangle inequality. For $d_{\max}$ we have

$$\begin{aligned} d_{\max}(x,z) &= \max_{i \le n}\{d_i(x_i, z_i)\} \le \max_{i \le n}\{d_i(x_i, y_i) + d_i(y_i, z_i)\} \\ &\le \max_{i \le n}\{d_i(x_i, y_i)\} + \max_{i \le n}\{d_i(y_i, z_i)\} = d_{\max}(x,y) + d_{\max}(y,z). \end{aligned}$$

Finally for d we have

$$\begin{aligned} d(x,z) &= \left[\sum_{i=1}^{n} d_i(x_i, z_i)^2\right]^{\frac{1}{2}} \le \left[\sum_{i=1}^{n} \big(d_i(x_i, y_i) + d_i(y_i, z_i)\big)^2\right]^{\frac{1}{2}} \\ &\le \left[\sum_{i=1}^{n} d_i(x_i, y_i)^2\right]^{\frac{1}{2}} + \left[\sum_{i=1}^{n} d_i(y_i, z_i)^2\right]^{\frac{1}{2}} = d(x,y) + d(y,z). \end{aligned}$$

Here we applied the triangle inequality for the standard norm on $\mathbb{R}^n$. Thus it only remains to show that these metrics are equivalent. Let $a_i := d_i(x_i, y_i)$. Then

$$\max\{a_i\} \le \big(\sum a_i^2\big)^{\frac{1}{2}} \le \sum a_i \le n \max\{a_i\} \le n\big(\sum a_i^2\big)^{\frac{1}{2}} \le n \sum a_i.$$

The second inequality above follows from squaring its both sides. ■

Theorem 2.47. *A sequence $(a_n)_{n \in \mathbb{N}} = ((a_{n,1}, \dots, a_{n,k}))_{n \in \mathbb{N}}$ in the product space $X_1 \times \cdots \times X_k$ converges to $a = (a_1, \dots, a_k)$ if and only if $a_{n,i} \to a_i$ for each i.*

Proof. We use the metric $d_{\max}$. We have

$$d_i(a_{n,i}, a_i) < \epsilon \text{ for all } i \iff d_{\max}(a_n, a) < \epsilon.$$

From this the result follows easily. ■

Theorem 2.48. *Suppose A_i is a subset of the metric space X_i for $i = 1, \dots k$. If each A_i is open then $\prod_{i=1}^k A_i$ is an open subset of $\prod_{i=1}^k X_i$, and if each A_i is closed then $\prod_{i=1}^k A_i$ is a closed subset of $\prod_{i=1}^k X_i$.*

Proof. Suppose A_i's are closed. Let $\big((a_{n,1}, \dots, a_{n,k})\big)_{n\in\mathbb{N}}$ be a sequence in $\prod_{i=1}^k A_i$ converging to $(a_1, \dots, a_k)$. Then we have $a_{n,i} \to a_i$, so we must have $a_i \in A_i$. Thus

$$(a_1, \dots, a_k) \in \prod_{i=1}^k A_i.$$

Now suppose A_i's are open. We have

$$\left(\prod_{i=1}^k A_i\right)^c = \bigcup_{j\le k} \big(X_1 \times \cdots \times X_{j-1} \times A_j^c \times X_{j+1} \times \cdots \times X_k\big).$$

Each of the sets $X_1 \times \cdots \times A_j^c \times \cdots \times X_k$ is a product of closed sets, hence it is closed. Thus their union is closed too. Therefore $\prod_{i=1}^k A_i$ is open. ■

***Example* 2.49.** Not every open or closed set in a product space is a product of open or closed sets. For example the open unit disk in $\mathbb{R}^2$ is not a product of two open subsets of $\mathbb{R}$. (Why?)

2.3 Continuous Functions

Definition 2.50. A function $f : X \to Y$ between two metric spaces is **continuous at a point** $a \in X$ if

$$\begin{aligned}&\forall \epsilon > 0 \; \exists \delta > 0 \text{ such that } \forall x \in X \\ &\qquad d_X(x, a) < \delta \implies d_Y(f(x), f(a)) < \epsilon.\end{aligned}$$

We say f is **continuous** if it is continuous at every point of its domain. A function that is not continuous is called **discontinuous**.

Proposition 2.51. *The constant functions are continuous. Also, the identity map of any space*

$$\begin{aligned}\mathrm{id}_X : X &\to X \\ x &\mapsto x\end{aligned}$$

is continuous.

Theorem 2.52. *A function $f : X \to Y$ is continuous at $a \in X$ if and only if for any sequence $a_n \to a$ we have $f(a_n) \to f(a)$.*

Proof. Suppose f is continuous at a, and $a_n \to a$. Then for n large enough we have $d_X(a_n, a) < \delta$, hence $d_Y(f(a_n), f(a)) < \epsilon$. Thus $f(a_n) \to f(a)$.

For the converse, we prove the contrapositive. Therefore suppose f is not continuous at a. Then there is $\epsilon > 0$ such that for all $n \in \mathbb{N}$ there is $a_n \in X$ such that

$$d_X(a_n, a) < \frac{1}{n}, \qquad \text{but } \ d_Y(f(a_n), f(a)) \geq \epsilon.$$

This means $a_n \to a$ but $f(a_n) \not\to f(a)$. ■

Theorem 2.53. *Suppose $f : X \to Y$ and $g : Y \to Z$ are continuous respectively at $x, f(x)$. Then $g \circ f : X \to Z$ is continuous at x. As a result, the composition of continuous functions is continuous.*

Proof. If $x_n \to x$ then $f(x_n) \to f(x)$, hence $g(f(x_n)) \to g(f(x))$. ■

Definition 2.54. Let $f : X \to Y$, and $A \subset Y$. Then the **preimage** or the **inverse image** of A is

$$f^{-1}(A) := \{x \in X : f(x) \in A\}.$$

Note that we do not require the map f to be invertible.

Theorem 2.55. *Let $f : X \to Y$ be a function between two metric spaces. Then the following assertions are equivalent.*

(i) *f is continuous.*
(ii) *For every convergent sequence $x_n \to x$ we have $f(x_n) \to f(x)$.*
(iii) *For each open subset U of Y, the set $f^{-1}(U)$ is an open subset of X.*
(iv) *For each closed subset C of Y, the set $f^{-1}(C)$ is a closed subset of X.*

Proof. We have seen that (i) and (ii) are equivalent. It is easy to see that (iii) and (iv) are equivalent too, since

$$f^{-1}(A^c) = \{x \in X : f(x) \in A^c\} = \{x \in X : f(x) \notin A\} = (f^{-1}(A))^c.$$

Now we only need to invoke the duality of open sets and closed sets.

Hence it suffices to prove that (ii) implies (iv), and (iii) implies (i). Suppose C is a closed subset of Y. Let x_n be a sequence in $f^{-1}(C)$ that converges to x. It is enough to show that $x \in f^{-1}(C)$. We know that $f(x_n) \to f(x)$. We also know that $f(x_n) \in C$. Hence $f(x) \in C$ and therefore $x \in f^{-1}(C)$.

Now suppose (iii) holds. We want to show that f is continuous. Let $x \in X$ be an arbitrary point. Then for any given $\epsilon > 0$, $B_\epsilon(f(x))$ is an open set in Y. Hence $f^{-1}(B_\epsilon(f(x)))$ is open in X. But this set obviously contains x. Therefore there is $\delta > 0$ such that

$$B_\delta(x) \subset f^{-1}(B_\epsilon(f(x))).$$

This means that $d_X(y, x) < \delta$ implies that $d_Y(f(y), f(x)) < \epsilon$, as desired. ■

Remark. The image of an open set under a continuous function is not necessarily open. For example $x \mapsto x^2$ takes $(-1,1)$ to $[0,1)$. The continuous image of a closed set is not necessarily closed either. For example the continuous function $x \mapsto \frac{1}{x}$ from $(0,\infty)$ to $\mathbb{R}$, takes $[1,\infty)$ to $(0,1]$.

Proposition 2.56. *Continuity of a map $f : X \to Y$ depends only on the topology of X, Y; so it is unaffected if we change the metrics of X, Y with equivalent metrics.*

Proof. By the last theorem, continuity can be expressed solely in terms of open sets. On the other hand, equivalent metrics induce the same topology. ■

Proposition 2.57. *The projections*

$$\begin{aligned} \pi_i : X_1 \times \cdots \times X_n &\to X_i \\ (x_1, \dots, x_n) &\mapsto x_i \end{aligned}$$

from a product space to any of its components are continuous.

Proof. We have

$$d_{\max}(x,y) < \epsilon \implies d_i(x_i, y_i) < \epsilon. \qquad ■$$

Theorem 2.58. *For functions into product spaces we have*

(i) *The function*

$$\begin{aligned} f = (f_1, \dots, f_n) : X &\longrightarrow Y_1 \times \cdots \times Y_n \\ x &\mapsto (f_1(x), \dots, f_n(x)) \end{aligned}$$

is continuous at $a \in X$ if and only if each $f_i : X \to Y_i$ is continuous at a.

(ii) *The function*

$$\begin{aligned} f = f_1 \times \cdots \times f_n : X_1 \times \cdots \times X_n &\longrightarrow Y_1 \times \cdots \times Y_n \\ (x_1, \dots, x_n) &\mapsto (f_1(x_1), \dots, f_n(x_n)) \end{aligned}$$

is continuous at $\mathbf{a} = (a_1, \dots, a_n) \in \prod X_i$ if and only if each $f_i : X_i \to Y_i$ is continuous at a_i.

Proof. **(i)** Let $d_{\max}$ be the max metric on $\prod Y_i$. The result follows from

$$d_{\max}(f(x), f(a)) < \epsilon \iff d_{Y_i}(f_i(x), f_i(a)) < \epsilon \text{ for all } i.$$

(ii) Again, the result follows from

$$d_{\max}(f(\mathbf{x}), f(\mathbf{a})) < \epsilon \iff d_{Y_i}(f_i(x_i), f_i(a_i)) < \epsilon \text{ for all } i.$$

Note that in this case we also need to use the fact that

$$\tilde{d}_{\max}(\mathbf{x}, \mathbf{a}) < \delta \iff d_{X_i}(x_i, a_i) < \delta \text{ for all } i,$$

where $\tilde{d}_{\max}$ is the max metric on $\prod X_i$. ■

Theorem 2.59. *The addition, subtraction and multiplication from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$ are continuous. Also, the inversion from $\mathbb{R} - \{0\}$ to $\mathbb{R}$, and the division from $\mathbb{R} \times (\mathbb{R} - \{0\})$ to $\mathbb{R}$ are continuous.*

Proof. First consider addition. We want to show that the map

$$\begin{aligned} \mathbb{R} \times \mathbb{R} &\longrightarrow \mathbb{R} \\ (x, y) &\mapsto x + y \end{aligned}$$

is continuous at an arbitrary point (a, b). Let $\epsilon > 0$ be given. We use the $d_{\max}$ on $\mathbb{R} \times \mathbb{R}$. So suppose $d_{\max}\big((x, y), (a, b)\big) < \frac{\epsilon}{2}$, or equivalently $|x - a|\,, |y - b| < \frac{\epsilon}{2}$. Then

$$|x + y - (a + b)| = |x - a + y - b| \leq |x - a| + |y - b| < \epsilon.$$

Next consider multiplication. Let $\epsilon > 0$ be given. Suppose $|x - a|\,, |y - b| < \delta$. Then we have

$$|xy - ab| = |xy - ay + ay - ab| \leq |y||x - a| + |a||y - b|.$$

But we have $|y| \leq |y - b| + |b| < \delta + |b|$. Hence

$$|xy - ab| \leq |y||x - a| + |a||y - b| < \delta(\delta + |b| + |a|).$$

Thus for $\delta \leq \min\{1, \frac{\epsilon}{1+|b|+|a|}\}$ we have $|xy - ab| < \delta(1 + |b| + |a|) < \epsilon$.

Subtraction is a composition of continuous functions as follows

$$(x, y) \mapsto (x, (-1)y) \mapsto x + ((-1)y) = x - y.$$

Note that $y \mapsto (-1)y$ is continuous, since it is the composition of $y \mapsto (-1, y) \mapsto (-1)y$.

Now consider inversion, which is the map from $\mathbb{R} - \{0\}$ to $\mathbb{R}$ that takes x to $\frac{1}{x}$. We want to show that it is continuous at an arbitrary point $a \neq 0$. Let $\epsilon > 0$ be given. Suppose $|x - a| < \delta$. Then for $\delta < \frac{|a|}{2}$ we have $|x| > \frac{|a|}{2}$. Hence

$$\left|\frac{1}{x} - \frac{1}{a}\right| = \left|\frac{a - x}{ax}\right| < \frac{2\delta}{|a|^2}.$$

Thus for $\delta \leq \min\{\frac{|a|}{2}, \frac{\epsilon|a|^2}{2}\}$ we have $|\frac{1}{x} - \frac{1}{a}| < \epsilon$ as desired.

Finally, note that division is a composition of continuous functions as follows

$$\begin{array}{ccccc} \mathbb{R} \times (\mathbb{R} - \{0\}) & \longrightarrow & \mathbb{R} \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ (x, y) & \longmapsto & (x, \dfrac{1}{y}) & \longmapsto & x\dfrac{1}{y} = \dfrac{x}{y}. \end{array}$$

■

Theorem 2.60. *Suppose $f, g : X \to \mathbb{R}$ are continuous. Then $f \pm g, fg$ are continuous. Also, $\frac{f}{g}$ is continuous if $g \neq 0$.*

Proof. All these functions can be written as a composition of continuous functions. For example fg can be written as

$$x \mapsto (f(x), g(x)) \mapsto f(x)g(x) = (fg)(x).$$ ■

Theorem 2.61. *The functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ whose components are polynomials in n-variables, are continuous. Also, the norm $|\ | : \mathbb{R}^n \to \mathbb{R}$ is continuous.*

Proof. It is enough to show the continuity of each component, so we assume $m = 1$. Each polynomial in n-variables is a sum of monomials, and each monomial is of the form $cx_1^{k_1} \cdots x_n^{k_n}$. Since the projections

$$(x_1, \dots, x_n) \mapsto x_i$$

are continuous, and the product of continuous functions are continuous, we can show by an easy induction that the maps

$$(x_1, \dots, x_n) \mapsto x_i^{k_i}$$

are continuous. Now as the constant functions are continuous, and the sum and product of continuous functions are continuous, we obtain the continuity of our polynomial.

For the norm, just note that for any two points x, y we have $||x| - |y|| \leq |x - y|$. This implies that if two points are close, their norms are close too. ■

Remark. The addition and multiplication of complex numbers are polynomial functions from $\mathbb{R}^2$ to $\mathbb{R}^2$, so they are continuous. Therefore, by imitating the above proof we can show that functions from $\mathbb{C}^n$ to $\mathbb{C}^m$ whose components are polynomials in n-variables, are continuous. Also, the modulus of complex numbers is the same as their norm as a vector in $\mathbb{R}^2$, thus $|\ | : \mathbb{C} \to \mathbb{R}$ is continuous too.

Remark. The addition, scalar multiplication, and inner product of vectors in Euclidean spaces are all continuous, since their components are polynomial functions. Hence we can conclude that the addition, scalar multiplication, and inner product of continuous functions into Euclidean spaces, are all continuous too.

Theorem 2.62. *The metric $d : X \times X \to \mathbb{R}$ is continuous.*

Proof. First note that for all $x, y, z \in X$ we have by the triangle inequality

$$d(x, y) - d(y, z) \leq d(x, z).$$

By switching x, z we get

$$|d(x,y) - d(y,z)| \leq d(x,z).$$

Now if we use d_{sum} on $X \times X$, we get

$$\begin{aligned} |d(x,y) - d(a,b)| &\leq |d(x,y) - d(y,a)| + |d(y,a) - d(a,b)| \\ &\leq d(x,a) + d(y,b) = d_{\text{sum}}\big((x,y),(a,b)\big). \end{aligned}$$

Thus if (x,y) is close to (a,b), then $d(x,y)$ is close to $d(a,b)$. ■

Definition 2.63. A function $f : X \to Y$ between two metric spaces is **uniformly continuous** if

$$\begin{aligned} &\forall \epsilon > 0 \; \exists \delta > 0 \text{ such that } \forall x, y \in X \\ &\qquad d_X(x,y) < \delta \implies d_Y(f(x), f(y)) < \epsilon. \end{aligned}$$

Remark. The difference of uniform continuity with continuity is that for a given ϵ we can choose a δ that works for *all points* of X.

Definition 2.64. A function $f : X \to Y$ between two metric spaces is **Lipschitz (continuous)** if there exists $L > 0$ such that

$$d_Y(f(x), f(y)) \leq L \, d_X(x,y)$$

for all $x, y \in X$.

Proposition 2.65. *A Lipschitz function is uniformly continuous.*

Proof. Take $\delta < \frac{\epsilon}{L}$. ■

Definition 2.66. A continuous bijective function whose inverse is also continuous is called a **homeomorphism**. Two spaces are said to be **homeomorphic** if there exists a homeomorphism between them.

***Example* 2.67.** The inverse of an invertible continuous function is not necessarily continuous. For example the function

$$f : \theta \mapsto (\cos\theta, \sin\theta),$$

from $[0, 2\pi)$ to the unit circle S^1, is a continuous bijection whose inverse is not continuous. To see this note that $a_n := \big(\cos(2\pi - \frac{1}{n}), \sin(2\pi - \frac{1}{n})\big) \to (1,0)$ but

$$f^{-1}(a_n) = 2\pi - \frac{1}{n} \to 2\pi \neq 0 = f^{-1}\big((1,0)\big).$$

Homeomorphic spaces are topologically equivalent, i.e. from the viewpoint of topology they are the same. In other words, homeomorphic spaces are topologically indistinguishable. They may well have other differences, but those differences cannot be caught by continuous functions. For example a circle and a square in the plane are homeomorphic, but they have many geometric differences. Nevertheless, ignoring some differences and paying attention to only a few properties, is a useful idea. It helps us to study many different objects at the same time. It also helps us to understand the implications of those few properties, and to not confuse these implications with the specific properties of each particular object.

Remark. It is easy to check that being homeomorphic is an equivalence relation. This is one of the reasons that we require the inverse of a homeomorphism to be continuous.

***Example* 2.68.** The interval $(0, 1)$ is homeomorphic to $\mathbb{R}$. For example

$$x \mapsto \frac{x}{1 + |x|}$$

is a homeomorphism from $\mathbb{R}$ onto $(-1, 1)$, and $x \mapsto \frac{x+1}{2}$ is a homeomorphism from $(-1, 1)$ onto $(0, 1)$ (why?). Can you give a different homeomorphism between them?

Proposition 2.69. *Suppose $f : X \to Y$ is a homeomorphism. Then f induces a bijection between the topology of X and the topology of Y.*

Proof. Since f^{-1} is continuous, $f(U)$ is open for every open set $U \subset X$. It is easy to see that the map $U \mapsto f(U)$ is a bijection between the topologies. ■

2.4 Complete Metric Spaces

Definition 2.70. A sequence (a_n) in a metric space (X, d) is a **Cauchy** sequence if

$$\forall \epsilon > 0 \, \exists N \in \mathbb{N} \text{ such that } \forall n, m \geq N \text{ we have } d(a_n, a_m) < \epsilon.$$

Theorem 2.71. *Convergent sequences are Cauchy.*

Proof. Let $a_n \to a$, then $d(a_n, a_m) \leq d(a_n, a) + d(a, a_m) \to 0$ as $m, n \to \infty$. ■

Remark. The converse is not true in general. For example the sequence $(\frac{\lfloor n\sqrt{2} \rfloor}{n})$ in $\mathbb{Q}$ is Cauchy, but it is not convergent. Here $\lfloor x \rfloor$ is the integer part of x.

In fact, a Cauchy sequence that does not converge, indicates that the space has a cavity. In other words, the points of a Cauchy sequence cluster but the space lacks their limit.

Definition 2.72. A **complete** metric space is a metric space in which all Cauchy sequences are convergent.

***Example* 2.73.** The above remark shows that $\mathbb{Q}$ is not a complete metric space. Also the interval $(0,1)$ is not a complete metric space, since the sequence $(\frac{1}{n})_{n=1}^{\infty}$ in $(0,1)$ is Cauchy but it is not convergent in $(0,1)$.

Theorem 2.74. *If a subsequence of a Cauchy sequence converges, then the whole sequence converges to the same limit.*

Proof. Let (a_n) be a Cauchy sequence such that $a_{n_k} \to a$. We want to show that $a_n \to a$. For a given positive ϵ, there is $N_1 \in \mathbb{N}$ such that for $k \geq N_1$ we have $d(a_{n_k}, a) < \frac{\epsilon}{2}$. Also, there is $N_2 \in \mathbb{N}$ such that for $n, m \geq N_2$ we have $d(a_n, a_m) < \frac{\epsilon}{2}$. Keep in mind that $n_k \geq k$. Then for $m \geq \max\{N_1, N_2\}$ we have

$$d(a_m, a) \leq d(a_m, a_{n_m}) + d(a_{n_m}, a) < \epsilon.$$ ■

Theorem 2.75. $\mathbb{R}$ *is complete.*

Proof. Let (a_n) be a Cauchy sequence in $\mathbb{R}$. First note that $\{a_n : n \in \mathbb{N}\}$ is a bounded subset of $\mathbb{R}$. The reason is that there is $N \in \mathbb{N}$ such that for $n \geq N$ we have $|a_n - a_N| < 1$. Then it is obvious that for all n we have $|a_n| \leq R$, where

$$R := |a_N| + \max\{1, |a_1 - a_N|, \dots, |a_{N-1} - a_N|\}.$$

Now, it is enough to show that (a_n) has a convergent subsequence (a_{n_k}). Let $n_0 = 0$. Then suppose we have chosen $a_{n_1}, \dots, a_{n_{m-1}}$. Choose $n_m > n_{m-1}$ such that

$$a_{n_m} > \sup\{a_n : n > n_{m-1}\} - \frac{1}{m}.$$

Let $b_m := \sup\{a_n : n > n_{m-1}\}$. It is obvious that $|b_m| \leq R$ for all m. Also note that (b_m) is a decreasing sequence, i.e. $b_{m+1} \leq b_m$. Then set

$$b := \inf\{b_m : m \in \mathbb{N}\}.$$

We claim that $a_{n_m} \longrightarrow b$ as $m \to \infty$. To see this note that for a given $\epsilon > 0$, $b + \epsilon$ is not a lower bound of $\{b_m\}$, hence there is N such that $b_N < b + \epsilon$. Thus for $m \geq \max\{N, \frac{1}{\epsilon}\}$ we have

$$b - \epsilon \leq b - \frac{1}{m} \leq b_m - \frac{1}{m} < a_{n_m} \leq b_m \leq b_N < b + \epsilon.$$

Therefore $|a_{n_m} - b| < \epsilon$ as desired. ■

Remark. The completeness of a metric space depends on its metric and not just its topology. For example $(0,1)$ and $\mathbb{R}$ are homeomorphic, but $(0,1)$ is not complete.

Theorem 2.76. *The product of finitely many complete metric spaces, equipped with any of the metrics in Theorem 2.46, is complete. In particular $\mathbb{R}^n$ is complete.*

Proof. Suppose $(X_1, d_1), \dots, (X_k, d_k)$ are complete metric spaces. We equip $\prod X_i$ with $d_{\max}$. The other cases can be proved similarly. Let $(a_n) = \big((a_{n,1}, \dots, a_{n,k})\big)_{n\in\mathbb{N}}$ be a Cauchy sequence in $\prod X_i$. Then we know that for any given $\epsilon > 0$ we can take m, n to be large enough so that $d_{\max}(a_m, a_n) < \epsilon$. This means that for m, n large enough we have $d_i(a_{m,i}, a_{n,i}) < \epsilon$. Therefore $(a_{n,i})$ is a Cauchy sequence in X_i. Hence $(a_{n,i})$ is convergent in X_i, and consequently (a_n) is convergent in $\prod X_i$. ■

Theorem 2.77. *Closed subsets of a complete metric space are complete.*

Proof. Suppose A is a closed subset of the complete metric space X. Let (a_n) be a Cauchy sequence in A. Then (a_n) is also a Cauchy sequence in X. Hence $a_n \to a$. But A is closed, so $a \in A$. Thus (a_n) is convergent in A. ■

2.5 Connectedness

Definition 2.78. A **separation** of a metric space X is a pair of nonempty open subsets A, B of X such that

$$X = A \cup B, \text{ and } A \cap B = \emptyset.$$

A metric space X is **disconnected** if there exists a separation of X.

A metric space is **connected** if it is not disconnected, i.e. there does not exist a separation of it. A nonempty subset of a space is connected if it is connected as a metric space with the induced metric.

Notation. We use the notation $A \sqcup B$ for $A \cup B$, when $A \cap B = \emptyset$.

***Example* 2.79.** $\mathbb{Q}$ is disconnected. For example

$$\{r \in \mathbb{Q} : r < \sqrt{2}\} \bigsqcup \{r \in \mathbb{Q} : r > \sqrt{2}\}$$

is a separation of $\mathbb{Q}$.

Theorem 2.80. *A space X is connected if and only if the only subsets of X that are both open and closed in X are $\emptyset, X$.*

Proof. Any other closed and open subset A, has a nonempty open complement A^c; and together they form a separation of X. ■

Theorem 2.81. *The continuous image of a connected set is connected.*

Proof. Suppose X is connected and $f : X \to Y$ is continuous. We want to show that $f(X)$ is connected. Let A be a nonempty open and closed subset of $f(X)$. It is enough to show that $A = f(X)$. Now $f^{-1}(A)$ is both open and closed. It is also nonempty, since A is nonempty and contains elements of the image of f. Thus $f^{-1}(A) = X$ since X is connected. Hence we have $A = f(f^{-1}(A)) = f(X)$ as desired. ■

Remark. An immediate consequence of the above theorem is that a space which is homeomorphic to a connected space, is connected.

Theorem 2.82. *A nonempty subset of $\mathbb{R}$ is connected if and only if it is an interval, or has only one element.*

Proof. Subsets with only one element are obviously connected. So we assume that the subset has more than one element. Now suppose $I \subset \mathbb{R}$ is an interval with a separation $U \sqcup V$. Take $a \in U$ and $b \in V$, and suppose $a < b$. Set

$$S = \{x \in I : [a, x] \subset U,\ x < b\},$$

and $c = \sup S$. As $a \leq c \leq b$ and I is an interval, we have $c \in I$ by Theorem 1.35. If $c \in U$ then we must have $c < b$. Thus I contains $[c, c + \epsilon]$ for some positive ϵ. Since U is open in I, U also contains $[c, c + \epsilon]$ for some ϵ. Therefore U contains $[a, c + \epsilon]$. But this contradicts the fact that c is an upper bound for S.

Thus we must have $c \in V$. Then $a < c$ and as before, V contains $[c - \epsilon, c]$ for some positive ϵ. But this implies that $c - \epsilon$ is an upper bound for S, which is in contradiction with c being the supremum of S. Therefore I cannot be disconnected.

For the other direction, let I be a connected subset of $\mathbb{R}$ with more than one element. Suppose to the contrary that I is not an interval. Hence by Theorem 1.35 there are distinct points $a, b \in I$, and $a < c < b$ such that $c \notin I$. But then $(I \cap (-\infty, c)) \sqcup (I \cap (c, +\infty))$ is a separation of I, which is a contradiction. ■

***Example* 2.83.** S^1 is connected, since it is the continuous image of $\mathbb{R}$ under the map $t \mapsto (\cos t, \sin t)$ (see Theorem 6.59).

Intermediate Value Theorem. *Suppose $f : X \to \mathbb{R}$ is continuous and X is connected. If $a, b \in f(X)$ and $a < c < b$ then $c \in f(X)$.*

Proof. If $c \notin f(X)$, then $f^{-1}((-\infty, c)) \sqcup f^{-1}((c, +\infty))$ is a separation of X. ■

***Exercise* 2.84.** Suppose I is an interval, and $f : I \to \mathbb{R}$ is a continuous one-to-one function. Show that f is either strictly increasing, or strictly decreasing.

Solution. Suppose $a, b, c \in I$, and $a < b < c$. Note that $f(a), f(b), f(c)$ are distinct real numbers, because f is one-to-one. First we show that $f(b)$ must be between $f(a), f(c)$. Suppose to the contrary that this is not the case. Then $f(b)$ is either greater than both $f(a), f(c)$, or is less than both of them. Suppose for example $f(a) < f(c) < f(b)$, the other cases are similar. But we know that $[a, b]$ is connected, so by the intermediate value theorem there is $t \in [a, b]$ such that $f(t) = f(c)$, which contradicts the fact that f is one-to-one. Hence we get the desired.

Now suppose we have $f(a) < f(b)$. We will show that f is strictly increasing. Let $x \in I - \{a, b\}$. If $x > b$ then we have $f(a) < f(b) < f(x)$, since $f(b)$ must be between $f(a), f(x)$. Similarly, if $x < a$ then $f(x) < f(a) < f(b)$, and if $a < x < b$ then $f(a) < f(x) < f(b)$. Now let $x, y \in I$, and suppose $x < y$. We have to show that $f(x) < f(y)$. If $x = a$ then we have already shown that the desired result holds. Similarly, we can deal with the cases $x = b$, $y = a$, and $y = b$.

So suppose $x, y \in I - \{a, b\}$. If x, y do not belong to the same interval among $I \cap (-\infty, a)$, $I \cap (a, b)$, and $I \cap (b, +\infty)$, then we can easily deduce that $f(x) < f(y)$, by comparing them to $f(a)$ or $f(b)$. Now suppose that x, y belong to the same interval among the above three intervals. Suppose for example $x, y \in (a, b)$, the other cases are similar. Then we must have $f(a) < f(x) < f(y)$, since $f(x)$ must be between $f(a), f(y)$. Thus we get the desired result. Similarly, the assumption $f(a) > f(b)$ implies that f is strictly decreasing. ■

Definition 2.85. Let X be a metric space, and $x, y \in X$. A **path** from x to y, is a continuous function f from an interval $[a, b]$ to X such that $f(a) = x$ and $f(b) = y$. A metric space X is called **path connected** if for any two points $x, y \in X$ there exists a path from x to y. A nonempty subset A of a space X is path connected if it is path connected as a metric space with the induced metric, in other words between any two points of A there is a path inside A.

Remark. Note that by a linear change of variable, we can assume that the domain of a path is any given closed interval.

Theorem 2.86. *A path connected space is connected.*

Proof. Suppose to the contrary that X is path connected, and there is a separation $A \sqcup B$ of X. Let $x \in A$ and $y \in B$. Then there is a path $f : [a, b] \to X$ from x to y. But it is easy to see that $f^{-1}(A) \sqcup f^{-1}(B)$ is a separation of $[a, b]$, which is a contradiction. ■

***Example* 2.87.** S^n is path connected for $n \geq 1$. To see this note that for any $x, y \in S^n$, if $y \neq -x$ then

$$t \mapsto \frac{(1-t)x + ty}{|(1-t)x + ty|}$$

is a path from x to y with domain $t \in [0,1]$. When $y = -x$ we have the path

$$t \mapsto \frac{(1-2t)x + t(1-t)z}{|(1-2t)x + t(1-t)z|}$$

from x to $-x$ with domain $t \in [0,1]$, where z is a nonzero vector orthogonal to x.

***Example* 2.88.** S^1 is not homeomorphic to an interval I in $\mathbb{R}$. To see this, suppose to the contrary that there is a homeomorphism $f : S^1 \to I$. Let $c \in I$ be an interior point, so that $I - \{c\}$ is disconnected. Then

$$f : S^1 - \{f^{-1}(c)\} \to I - \{c\}$$

is still continuous and onto. But $S^1 - \{f^{-1}(c)\}$ is obviously path connected, hence it is connected. Therefore its continuous image cannot be the disconnected set $I - \{c\}$.

***Exercise* 2.89.** Show that $\mathbb{R}$ is not homeomorphic to $[0,1)$. In general, find necessary and sufficient conditions for two intervals to be homeomorphic.

***Example* 2.90.** Similarly to the above example, we can see that S^1 is not homeomorphic to S^2, nor to a disk in $\mathbb{R}^2$. The reason is that removing any two points of S^1 disconnects it, while the same is not true for S^2 or the disk. In fact, S^2 or the disk minus finitely many points are path connected. Can you prove this?

Remark. We cannot apply the above argument directly to show that S^2 is not homeomorphic to S^3. Because removing finitely many points from S^2 or S^3 does not make them disconnected. Intuitively, we have to remove a closed curve from S^2 to make it disconnected. And removing a closed curve does not disconnect S^3 (To imagine this keep in mind that locally S^3 looks like $\mathbb{R}^3$). But the problem with this modified argument is that the continuous image of a curve is not necessarily a curve. There are continuous functions that map a curve onto a two-dimensional set, and removing a two-dimensional set can disconnect S^3. However by using the machinery of *algebraic topology*, we can make this idea precise and prove that S^2, S^3 are not homeomorphic.

Theorem 2.91. *The union of a family of connected sets that have a point in common is connected.*

Proof. Let $X = \bigcup X_\alpha$, and suppose $p \in \bigcap X_\alpha$. Suppose that X has a nonempty closed and open subset U. We either have $p \in U$ or $p \in U^c$. Suppose U contains p, otherwise we can work with the nonempty closed and open subset U^c. Then $U \cap X_\alpha$ is nonempty for all α. Also as $X_\alpha \subset X$, $U \cap X_\alpha$ is a closed and open subset of X_α. Hence by connectedness of X_α, we have $U \cap X_\alpha = X_\alpha$. Therefore U contains all X_α's and we have $U = X$. ■

***Exercise* 2.92.** Give an example of two connected sets whose intersection is not connected.

Theorem 2.93. *Suppose A is a connected subset of X, and $A \subset B \subset \bar{A}$. Then B is also connected. In particular, the closure of a connected set is connected.*

Proof. Suppose V is a nonempty closed and open subset of B. Then there is an open set $U \subset X$ such that $V = U \cap B$. Now

$$V \cap A = (U \cap B) \cap A = U \cap A$$

is a closed and open subset of A. Let us show that $U \cap A$ is nonempty. We know that there is $b \in U \cap B$. If $b \in A$ we are done. Otherwise we have $b \in \bar{A} - A$. Therefore b is a limit point of A. Thus as U is an open set containing b, U must intersect A. Hence by connectedness of A we get $U \cap A = A$.

Therefore V contains A. Since V is also closed in B, there is a closed subset C of X such that $V = C \cap B$. But then C contains A, and as C is closed we have $C \supset \bar{A}$. Consequently

$$V = C \cap B = B.$$

Thus B is connected. ■

***Example* 2.94.** Not every connected set is path connected. For example the *topologist's sine curve*

$$\{(x, y) : y = \sin\frac{1}{x},\ x \in (0, 1]\} \bigcup \{(0, y) : y \in [-1, 1]\}$$

is a connected subset of $\mathbb{R}^2$ which is not path connected. The proof can be found in Example 11.93.

Theorem 2.95. *The product of finitely many connected spaces is connected.*

Proof. It is sufficient to prove the theorem for the product of two spaces. The general result follows by induction. Suppose X, Y are connected. Suppose to the contrary that $X \times Y$ has a separation $A \sqcup B$. Let $f : X \times Y \to \mathbb{R}$ be the function that takes A to 0 and takes B to 1. It is easy to see that f is continuous. Now let $(x_0, y_0) \in A$ and $(x_1, y_1) \in B$. Consider (x_0, y_1). Suppose for example that $f((x_0, y_1)) = 0$. Then the function

$$\begin{aligned} g : X &\longrightarrow X \times Y \longrightarrow \quad \mathbb{R} \\ x &\longmapsto (x, y_1) \longmapsto f((x, y_1)) \end{aligned}$$

is continuos. Now $g(x_0) = 0$ and $g(x_1) = 1$. But g does not achieve any intermediate value, which is in contradiction with the connectedness of X. The case of $f((x_0, y_1)) = 1$ is similar. ■

2.6 Compactness

Definition 2.96. Suppose X is a set, and $f : X \to \mathbb{R}$ is a function. We say f has a **(global) maximum** at $y \in X$ if $f(y) \geq f(x)$ for all $x \in X$. The number $f(y)$ is also called the **maximum (value)** of f. Similarly, we say f has a **(global) minimum** at $y \in X$ if $f(y) \leq f(x)$ for all $x \in X$. The number $f(y)$ is also called the **minimum (value)** of f. Finally, an **extremum** of f, is either a maximum of f, or a minimum of f.

Suppose we want to minimize a function $F : X \to \mathbb{R}$, where X is a metric space. X can simply be a subset of a Euclidean space, or it can be an infinite dimensional space of states of some physical system. In order for F to have a minimum, we need to assume that F is bounded below. Let

$$m := \inf_{x \in X} F(x).$$

Then by the definition of infimum, for every n there is $x_n \in X$ such that

$$m \leq F(x_n) \leq m + \frac{1}{n}.$$

Now we hope that the sequence (x_n) converges to some $x^* \in X$. If F is continuous (we actually need less than continuity, and this is important in infinite dimensions), then we obviously have $F(x^*) = m$. However, the hard part of this approach is to show that (x_n) is convergent. Unfortunately, the sequence of "approximate minima" (x_n) is in general not convergent. For example, the function $\frac{1}{4}x^4 - \frac{1}{2}x^2$ on $\mathbb{R}$ has two minima at $x = \pm 1$. And its minimizing sequence

$$x_n = (-1)^n + \frac{1}{n}$$

is not convergent. But, as the above example suggests, we actually do not need the convergence of the whole sequence (x_n). For our purpose, it is enough to show that a subsequence (x_{n_k}) is convergent. Then we can argue as before and conclude that the limit of this subsequence is the minimizer. Thus we need to extract a convergent subsequence from an arbitrary sequence in X. This is the property of the space X that we want to study in this section.

Definition 2.97. A subset A of a metric space X is **(sequentially) compact** if every sequence in A has a subsequence that converges in A.

Remark. Note that we consider the empty set $\emptyset$ to be compact. In all of the following theorems, you should check that the claim holds for the empty compact set trivially.

***Example* 2.98.** A finite subset of a metric space is compact. Because any sequence in a finite set has a constant subsequence.

Theorem 2.99. *Closed subsets of a compact space are compact.*

Proof. Suppose X is compact, and $A \subset X$ is closed. Let (a_n) be a sequence in A. Then a subsequence $a_{n_k} \to a$, as X is compact. Since A is closed we must have $a \in A$. ■

***Exercise* 2.100.** Show that

(i) The union of finitely many compact sets is compact.

(ii) The intersection of an arbitrary family of compact sets is compact.

Definition 2.101. A subset A of a metric space X is called **bounded**, if there is $x \in X$ such that $A \subset B_r(x)$ for some $r > 0$. A function into a metric space is called a **bounded function** if its image is a bounded set. A subset or a function that is not bounded, is called *unbounded.*

Remark. Note that the boundedness of a subset depends on the metric.

Remark. Note that the closure of a bounded set A is also bounded. Because if $A \subset B_r(x)$ then $\bar{A} \subset \overline{B_r(x)} \subset B_{r+1}(x)$.

Theorem 2.102. *Compact subsets of a metric space are closed and bounded.*

Proof. Suppose the sequence (a_n) is in the compact subset A, and $a_n \to a$. Then a subsequence (a_{n_i}) must converge in A. But the limit of the subsequence must also be a. Hence $a \in A$ and A is closed.

Now suppose A is not bounded. Then we can choose a sequence in it that has no convergent subsequence, as follows. Let a_1 be an arbitrary point of A. Suppose we have chosen $a_1, \dots, a_k$. Let $r_i := d(a_1, a_i)$, and $r > r_i + 1$ for all i. Then by our assumption $B_r(a_1)$ does not contain A. Hence there is $a_{k+1} \in A - B_r(a_1)$. Note that we have

$$d(a_{k+1}, a_i) \geq d(a_{k+1}, a_1) - d(a_1, a_i) > r - r_i > 1.$$

Thus we have constructed a sequence (a_n) in A such that $d(a_n, a_m) > 1$ for $n \neq m$. It is easy to see that (a_n) cannot have a convergent subsequence, which is in contradiction with the compactness of A. ■

Remark. The converse of the above theorem is not true in general. For example $\mathbb{N}$ with the discrete metric is bounded and closed, but it has the sequence $1, 2, 3, \dots$ with no convergent subsequence.

Theorem 2.103. *The continuous image of a compact set is compact.*

$\boxed{\textbf{Proof.}}$ Let f be a continuous map on the compact set A. Take a sequence (b_n) in $f(A)$. Then for each n there is $a_n \in A$ such that $f(a_n) = b_n$. Now the sequence (a_n) has a convergent subsequence (a_{n_i}). Therefore $(b_{n_i}) = (f(a_{n_i}))$ is a convergent subsequence of (b_n). ■

Remark. An immediate consequence of the above theorem is that a space which is homeomorphic to a compact space, is compact.

Extreme Value Theorem. *A continuous function from a nonempty compact set into $\mathbb{R}$ is bounded, and achieves its maximum and minimum values.*

$\boxed{\textbf{Proof.}}$ Let $f : X \to \mathbb{R}$ be continuous, and suppose X is compact. Then $f(X)$ is compact, hence it is bounded and closed. Every nonempty, closed and bounded subset of $\mathbb{R}$ contains its finite supremum and infimum. Therefore the supremum and the infimum of $f(X)$ are achieved by f, i.e. there are $x_1, x_2 \in X$ such that

$$f(x_2) = \sup\{f(x) : x \in X\}, \qquad f(x_1) = \inf\{f(x) : x \in X\}.$$

These are the maximum and the minimum of f respectively. ■

Theorem 2.104. *Closed bounded intervals in $\mathbb{R}$ are compact.*

$\boxed{\textbf{Proof.}}$ The proof is similar to the proof of the completeness of $\mathbb{R}$. Let (a_n) be a sequence in $[a, b]$. We will construct a convergent subsequence (a_{n_k}). Let $n_0 = 0$. Then suppose we have chosen $a_{n_1}, \dots, a_{n_{m-1}}$. Choose $n_m > n_{m-1}$ such that

$$a_{n_m} > \sup\{a_n : n > n_{m-1}\} - \frac{1}{m}.$$

Let $b_m := \sup\{a_n : n > n_{m-1}\}$. It is obvious that $b_m \in [a, b]$ for all m. Also note that (b_m) is a decreasing sequence, i.e. $b_{m+1} \le b_m$. Then set

$$b := \inf\{b_m : m \in \mathbb{N}\}.$$

We claim that $a_{n_m} \longrightarrow b$ as $m \to \infty$. To see this note that for a given $\epsilon > 0$, $b + \epsilon$ is not a lower bound of $\{b_m\}$, hence there is N such that $b_N < b + \epsilon$. Thus for $m \ge \max\{N, \frac{1}{\epsilon}\}$ we have

$$b - \epsilon \le b - \frac{1}{m} \le b_m - \frac{1}{m} < a_{n_m} \le b_m \le b_N < b + \epsilon.$$

Therefore $|a_{n_m} - b| < \epsilon$ as desired. ■

***Example* 2.105.** $\mathbb{R}$ is not homeomorphic to $[0, 1]$, since $\mathbb{R}$ is not compact as it is unbounded.

Theorem 2.106. *The product of finitely many compact spaces is compact.*

Proof. It is enough to prove the theorem for the product of two spaces. The general result follows by induction. Let (x_n, y_n) be a sequence in the product of compact spaces $X \times Y$. Then (x_n) has a convergent subsequence $x_{n_k} \to x$. Now (y_{n_k}) has a convergent subsequence $y_{n_{k_i}} \to y$. Thus we have

$$(x_{n_{k_i}}, y_{n_{k_i}}) \to (x, y). \qquad ■$$

***Example* 2.107.** The product of closed bounded intervals

$$[a_1, b_1] \times \cdots \times [a_n, b_n]$$

is compact in $\mathbb{R}^n$.

Heine-Borel Theorem. *A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded in the Euclidean metric.*

Proof. Any bounded subset is contained in the product of some closed bounded intervals, which is a compact set. Hence if the subset is closed it is compact. The converse is proved in Theorem 2.102. ■

Bolzano-Weierstrass Theorem. *A bounded sequence in $\mathbb{R}^n$ has a convergent subsequence.*

Proof. Any bounded sequence is contained in the product of some closed bounded intervals, which is a compact set. ■

Theorem 2.108. *A continuous bijection from a compact space into another metric space is a homeomorphism.*

Proof. Let $f : X \to Y$ be a continuous bijection, and suppose X is compact. We have to show that f^{-1} is continuous. For any closed set $C \subset X$ we have $(f^{-1})^{-1}(C) = f(C)$. But C is compact, hence $f(C)$ is compact too. Thus $f(C)$ is closed, and f^{-1} is continuous. ■

Second Proof. Suppose to the contrary that f^{-1} is not continuous at $y \in Y$. Then by using the definition of continuity we find $\epsilon > 0$, so that for each $n \in \mathbb{N}$ we can choose $y_n \in Y$ such that

$$d_Y(y_n, y) < \frac{1}{n} \quad \text{and} \quad d_X(f^{-1}(y_n), f^{-1}(y)) \geq \epsilon. \qquad (*)$$

Thus $y_n \to y$. Let $x_n := f^{-1}(y_n)$. Then (x_n) has a convergent subsequence $x_{n_i} \to x$. Hence we have $y_{n_i} = f(x_{n_i}) \to f(x)$. But we know that $y_{n_i} \to y$, so $y = f(x)$ and therefore $x = f^{-1}(y)$. Thus we get

$$f^{-1}(y_{n_i}) = x_{n_i} \longrightarrow x = f^{-1}(y).$$

However this is in contradiction with $(*)$. ■

Remark. The above theorem is not true when the domain is not compact. For example the function

$$\theta \mapsto (\cos\theta, \sin\theta)$$

from $[0, 2\pi)$ to S^1, is a continuous bijection which is not a homeomorphism. In fact there does not exist a homeomorphism between $[0, 2\pi)$ and S^1, as S^1 is compact while $[0, 2\pi)$ is not.

***Exercise* 2.109.** Show that for $p \in \mathbb{Q}$ the function $x \mapsto x^p$ is continuous on its domain.

Theorem 2.110. *Compact metric spaces are complete.*

Proof. A Cauchy sequence with a convergent subsequence is convergent. ■

***Exercise* 2.111.** Suppose A, B are disjoint subsets of a metric space. Also suppose A is closed and B is compact. Show the distance between the points of A, B has a positive lower bound, i.e. there is $c > 0$ such that $d(a, b) \geq c$ for every $a \in A$ and $b \in B$.

Solution. To prove this note that otherwise we would have sequences of points $a_j \in A$ and $b_j \in B$ such that $|a_j - b_j| \to 0$. Then by compactness of B we can choose a subsequence a_{j_k} such that $a_{j_k} \to a \in B$. As a consequence we have $b_{j_k} \to a$, because

$$|b_{j_k} - a| \leq |b_{j_k} - a_{j_k}| + |a_{j_k} - a| \to 0.$$

Hence we get $a \in A$, since A is closed. But this means that A and B have a nonempty intersection, contrary to our assumption. ■

Definition 2.112. Suppose A is a bounded nonempty subset of a metric space X. Then the **diameter** of A is

$$\operatorname{diam}(A) := \sup\{d(x, y) : x, y \in A\}.$$

We also set $\operatorname{diam}(\emptyset) := 0$.

Remark. A sequence of sets $\{A_n\}_{n=1}^{\infty}$ is called decreasing if $A_{n+1} \subset A_n$ for all n.

Theorem 2.113. *Let $\{A_n\}_{n=1}^{\infty}$ be a decreasing sequence of nonempty compact subsets of a metric space X. Then their intersection $\bigcap_{n\geq 1} A_n$ is compact and nonempty.*

Furthermore, if in addition we have $\operatorname{diam}(A_n) \to 0$, *then $\bigcap_{n\geq 1} A_n$ consists of a single point.*

$\boxed{\textbf{Proof.}}$ The intersection of A_n's is closed since A_n's are closed. Now, $\bigcap_{n\geq 1} A_n$ is a closed subset of the compact set A_1, hence it is compact. Thus we only need to show that the intersection is nonempty. Let a_n be an element of A_n. Then (a_n) is a sequence in the compact set A_1. Hence it has a subsequence $a_{n_k} \to a$. But $a_{n_k} \in A_m$ if $n_k \geq m$. So a must belong to A_m as A_m is closed. Therefore $a \in \bigcap A_m$.

For the second part, suppose $a, b \in \bigcap A_n$. Then $a, b \in A_n$ for all n. Hence

$$d(a,b) \leq \operatorname{diam}(A_n) \to 0.$$

Therefore we must have $a = b$. ■

Remark. The compactness assumption is essential in the above theorem. For example, the intervals $\{(0, \frac{1}{n})\}_{n=1}^{\infty}$ form a decreasing sequence of subsets of $\mathbb{R}$, but their intersection is empty. Even if we assume that the sets are closed, but not compact, we can not deduce that their intersection is nonempty. For example, the intervals $\{[n, \infty)\}_{n=1}^{\infty}$ form a decreasing sequence of closed subsets of $\mathbb{R}$, but their intersection is empty.

There is another formulation of compactness that is very useful and important. Although at first, the importance of this formulation is not as evident as the sequential formulation of compactness. We will show that the two formulations of compactness are equivalent for metric spaces.

Definition 2.114. An **open covering** of a subset A of a metric space X, is a family $\mathcal{U}$ of open subsets of X such that

$$\bigcup \mathcal{U} = \bigcup_{U \in \mathcal{U}} U \supset A.$$

A **subcovering** $\mathcal{V}$ of $\mathcal{U}$ is a subfamily of $\mathcal{U}$ which is itself an open covering of A.

Definition 2.115. A subset A of a metric space X is **compact** if every open covering of A has a finite subcovering.

Remark. Note that we do not say that A has a finite open covering. Rather, we say that from any open covering of A we can choose finitely many open sets whose union covers A.

***Example* 2.116.** $(0, 1)$ is not compact, since the open covering $\{(\frac{1}{n}, 1)\}_{n=1}^{\infty}$ has no finite subcovering.

Theorem 2.117. *A continuous function from a compact metric space into another metric space is uniformly continuous.*

Proof. Let $f : X \to Y$ be continuous, and suppose X is compact. Suppose to the contrary that f is not uniformly continuous. Then there is $\epsilon > 0$ such that for every $\delta > 0$ there are $x, y \in X$ so that $d_X(x, y) < \delta$ but $d_Y(f(x), f(y)) \geq \epsilon$. Set $\delta = \frac{1}{n}$, and let x_n, y_n be the corresponding points. Then (x_n) has a convergent subsequence $x_{n_i} \to a$, since X is compact. Now (y_{n_i}) is also a sequence in the compact space X. So it has a convergent subsequence $y_{n_{i_j}} \to b$. Then $x_{n_{i_j}} \to a$. Also as n_{i_j}'s are distinct positive integers and cannot remain bounded we have

$$0 \leq d_X(x_{n_{i_j}}, y_{n_{i_j}}) < \frac{1}{n_{i_j}} \underset{j\to\infty}{\longrightarrow} 0.$$

Thus due to the continuity of the metric we have $d_X(a, b) = \lim d_X(x_{n_{i_j}}, y_{n_{i_j}}) = 0$. Hence $a = b$. On the other hand we must have $f(x_{n_{i_j}}) \to f(a)$, and $f(y_{n_{i_j}}) \to f(a)$, since f is continuous. Therefore

$$d_Y(f(x_{n_{i_j}}), f(y_{n_{i_j}})) \to d_Y(f(a), f(a)) = 0.$$

But this contradicts the fact that $d_Y(f(x_{n_{i_j}}), f(y_{n_{i_j}})) \geq \epsilon$. ■

Second Proof. Let $f : X \to Y$ be continuous, and suppose X is compact. We denote the open balls of X by $B_r(x)$, and the open balls of Y by $B_r(y, Y)$. Now for a given $\epsilon > 0$ and $x \in X$, there is $\delta_x(\epsilon) > 0$ such that

$$d_X(y, x) < \delta_x(\epsilon) \implies d_Y(f(y), f(x)) < \epsilon.$$

This is equivalent to

$$f(B_{\delta_x(\epsilon)}(x)) \subset B_\epsilon(f(x), Y).$$

We need to show that we can choose δ independently of x. Consider the open covering

$$\{B_{\frac{1}{2}\delta_x(\frac{\epsilon}{2})}(x) : x \in X\}$$

of X. It has a finite subcovering

$$\{B_{\frac{1}{2}\delta_{x_1}(\frac{\epsilon}{2})}(x_1), \dots, B_{\frac{1}{2}\delta_{x_k}(\frac{\epsilon}{2})}(x_k)\},$$

for some $x_1, \dots, x_k \in X$. Set $\delta := \frac{1}{2}\min_{i\leq k} \delta_{x_i}(\frac{\epsilon}{2})$. Then for any x there is an x_i such that $x \in B_{\frac{1}{2}\delta_{x_i}(\frac{\epsilon}{2})}(x_i)$. Hence $B_\delta(x) \subset B_{\delta_{x_i}(\frac{\epsilon}{2})}(x_i)$, and $f(x) \in B_{\frac{\epsilon}{2}}(f(x_i), Y)$. Therefore we have

$$f(B_\delta(x)) \subset f(B_{\delta_{x_i}(\frac{\epsilon}{2})}(x_i)) \subset B_{\frac{\epsilon}{2}}(f(x_i), Y) \subset B_\epsilon(f(x), Y). \quad ■$$

Theorem 2.118. *A compact subset of a metric space is sequentially compact.*

Proof. Let (a_n) be a sequence in the compact set A. Suppose to the contrary that no subsequence of (a_n) converges to a point of A. Then for any $x \in A$ there is $r_x > 0$ such that the set

$$\{n : a_n \in B_{r_x}(x)\}$$

is finite. Since otherwise there is $a \in A$ such that for all $m \in \mathbb{N}$ the set $\{n : a_n \in B_{\frac{1}{m}}(a)\}$ is infinite. Then for each m we can choose $a_{n_m} \in B_{\frac{1}{m}}(a)$ such that $n_m > n_{m-1}$. Now it is easy to see that the subsequence (a_{n_m}) converges to a, contrary to our assumption.

Thus for every $x \in A$ there is $r_x > 0$ so that $\{n : a_n \in B_{r_x}(x)\}$ is a finite set. Now the family

$$\{B_{r_x}(x) : x \in A\}$$

is an open covering of A. Hence it has a finite subcovering, namely

$$A \subset B_{r_{x_1}}(x_1) \cup \cdots \cup B_{r_{x_k}}(x_k),$$

for some $x_1, \ldots, x_k \in A$. Therefore

$$\mathbb{N} = \{n : a_n \in A\} \subset \{n : a_n \in B_{r_{x_i}}(x_i) \quad i = 1, \ldots, k\}.$$

But the right hand side set is a finite set while $\mathbb{N}$ is infinite, which is a contradiction. Consequently (a_n) must have a convergent subsequence in A, and therefore A is sequentially compact. ■

Definition 2.119. Let $\mathcal{U}$ be an open covering of a subset A of a metric space X. A positive real number λ is called a **Lebesgue number** for $\mathcal{U}$ if for every $a \in A$ there exists $U \in \mathcal{U}$ such that $B_\lambda(a) \subset U$.

Remark. The point of this definition is that λ does not depend on a.

***Example* 2.120.** If a set is noncompact, then an open covering need not have a Lebesgue number even if it is finite. For example $\{(0,1)\}$ is an open covering of $(0,1)$ that has no positive Lebesgue number.

Lebesgue Number Lemma. *Every open covering of a sequentially compact subset of a metric space has a Lebesgue number.*

Proof. Suppose to the contrary that a sequentially compact set A has an open covering $\mathcal{U}$ that has no Lebesgue number. Then for each positive integer n we can find a point $a_n \in A$ such that no element of $\mathcal{U}$ contains $B_{\frac{1}{n}}(a_n)$. Then a subsequence (a_{n_k}) converges to some $a \in A$. Now as $\mathcal{U}$ covers A, there is $U \in \mathcal{U}$ such that $a \in U$. Also as U is open, there is $r > 0$ such that $B_r(a) \subset U$. But for large enough k we have $a_{n_k} \in B_{\frac{r}{2}}(a)$, and $\frac{1}{n_k} < \frac{r}{2}$. Thus we have $B_{\frac{1}{n_k}}(a_{n_k}) \subset U$, which is a contradiction. ■

Theorem 2.121. *A sequentially compact subset of a metric space is compact.*

Proof. Let A be a sequentially compact subset of the metric space X. Let $\mathcal{U}$ be an open covering of A. Suppose to the contrary that $\mathcal{U}$ has no finite subcovering. Let λ be a Lebesgue number for $\mathcal{U}$. Let $a_1 \in A$ be an arbitrary point. Then there is $U_1 \in \mathcal{U}$ such that $B_\lambda(a_1) \subset U_1$. Since $A \not\subset U_1$ by our assumption, there is $a_2 \in A - U_1$. Choose $U_2 \in \mathcal{U}$ such that $B_\lambda(a_2) \subset U_2$. Again note that $A \not\subset U_1 \cup U_2$. We continue this way and we choose

$$a_n \in A - (U_1 \cup \cdots \cup U_{n-1}),$$

and $U_n \in \mathcal{U}$ such that $B_\lambda(a_n) \subset U_n$.

Thus we have a sequence (a_n). Let (a_{n_k}) be a subsequence converging to $a \in A$. Then there is a large enough k such that $a_{n_k} \in B_\lambda(a)$. Therefore $a \in B_\lambda(a_{n_k}) \subset U_{n_k}$. Hence there is $r > 0$ such that $B_r(a) \subset U_{n_k}$. But then for large enough l we must have

$$a_{n_l} \in B_r(a) \subset U_{n_k}.$$

This is a contradiction since for $l > k$ we have $a_{n_l} \notin U_{n_k}$. ■

Remark. Note our crucial use of the Lebesgue number in the above proof. If λ was dependent on a_n, then from $a_{n_k} \in B_\lambda(a)$ we could not deduce $a \in U_{n_k}$.

2.7 The Cantor Set

Definition 2.122. Let $C_0 = [0, 1]$. We construct the sets C_n inductively as follows. To obtain C_{n+1} we remove the open middle third of each interval of C_n. Thus we have

$$\begin{aligned}
&C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1] = C_0 - (\frac{1}{3}, \frac{2}{3})\\
&C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1] = C_1 - \{(\frac{1}{9}, \frac{2}{9}) \cup (\frac{7}{9}, \frac{8}{9})\}\\
&\vdots
\end{aligned}$$

Each C_n is the disjoint union of 2^n closed intervals of length $\frac{1}{3^n}$. Let

$$C = \bigcap_{n \geq 0} C_n.$$

Then C is called the standard **Cantor set**.

Definition 2.123. A subset A of a metric space is called **totally disconnected** if every connected subset of A has at most one element.

***Example* 2.124.** $\mathbb{Q}$ is totally disconnected (why?), while $[0,1]\cup[2,3]$ is not.

Theorem 2.125. *The Cantor set is nonempty, compact, totally disconnected, and has no isolated point.*

Proof. Each C_n is compact and nonempty, and they form a decreasing sequence of sets. Thus their intersection, i.e. the Cantor set C, is also nonempty and compact. In particular, note that the endpoints of every interval of each C_n is in C, since those endpoints were not removed in any step.

Now let $x \in C$. Suppose $r > 0$ is given. We want to find a point in $B_r(x) \cap C$ other than x. Let n be large enough so that $\frac{1}{3^n} < r$. Then as $x \in C_n$, there is an interval $[a,b] \subset C_n$ such that $x \in [a,b]$. Suppose $b \neq x$. Then $|b-x| \leq |b-a| = \frac{1}{3^n} < r$. Hence $b \in B_r(x)$. But b is an endpoint of an interval of C_n, therefore $b \in C$. Thus x is a limit point of C.

Finally, suppose $A \subset C$ is connected and contains two distinct points x, y. Let n be large enough so that $\frac{1}{3^n} < |y-x|$. Then as $x, y \in C_n$, there are intervals $[a,b],[c,d] \subset C_n$ such that $x \in [a,b]$ and $y \in [c,d]$. Note that x, y cannot belong to the same interval of C_n, since the length of each interval in C_n is less than $|y-x|$. Now suppose $b < c$. Let $r \in (b,c) - C$. Then $A \cap [0,r)$ and $A \cap (r,1]$ form a separation of A, which is a contradiction. ■

Remark. Any metric space that has the four properties mentioned in the above theorem, is actually homeomorphic to the Cantor set.

Theorem 2.126. *The interior of the Cantor set is empty. In other words, the Cantor set contains no nontrivial interval.*

Proof. If $x \in C^\circ$, then $(x-r, x+r) \subset C$ for some positive r. So C must contain an interval. But intervals are connected, while the Cantor set is totally disconnected. ■

***Exercise* 2.127.** Is $\frac{1}{4} \in C$? If yes, is it an endpoint of an interval of some C_n?

Theorem 2.128. *The Cantor set is uncountable.*

Proof. We construct a bijection between C and the set of all infinite sequences consisting of 0's and 2's. The latter set has the cardinality of $\mathbb{R}$. Let $x \in C$. We will assign a sequence $\omega_1\omega_2\omega_3\cdots$ of 0's and 2's, to x. We know that x belongs to each C_n. Consider $n = 1$. If x belongs to the left interval of C_1, i.e. $[0, \frac{1}{3}]$, we set $\omega_1 = 0$. If x belongs to the right interval of C_1, i.e. $[\frac{2}{3}, 1]$, we set $\omega_1 = 2$. Now suppose we have determined $\omega_1\cdots\omega_n$. Let $[a,b]$ be one of the 2^n intervals of C_n that contains x. When we construct C_{n+1}, we remove the open middle third of $[a,b]$ and we obtain two intervals $[a,c_1]$ and $[c_2,b]$. If x belongs to the left subinterval of

$[a, b]$, i.e. $[a, c_1]$, we set $\omega_{n+1} = 0$. If x belongs to the right subinterval of $[a, b]$, i.e. $[c_2, b]$, we set $\omega_{n+1} = 2$.

It is easy to see that this assignment is one-to-one. Consider two points $x, y \in C$ with $x < y$. Note that x, y cannot be in the same interval of C_n for all n, since the length of those intervals goes to zero. Let n be the largest integer for which x, y belong to the same interval of C_n. Then $\omega_{n+1}(x) = 0$ while $\omega_{n+1}(y) = 2$.

Finally, we need to show that this assignment is onto. Suppose we have a sequence $\omega_1\omega_2\omega_3\cdots$ of 0's and 2's. Let $[a_1, b_1]$ be the left or the right interval of C_1 according to whether $\omega_1 = 0$ or $\omega_1 = 2$. Inductively, let $[a_{n+1}, b_{n+1}]$ be the left or the right subinterval of $[a_n, b_n]$ according to whether $\omega_{n+1} = 0$ or $\omega_{n+1} = 2$. Note that $[a_n, b_n]$ is one of the 2^n intervals of C_n. Now we have a decreasing sequence of nonempty compact sets $[a_n, b_n]$, whose diameter, which is $\frac{1}{3^n}$, goes to zero. Thus their intersection is a single point, i.e. $\bigcap_{n\geq 0}[a_n, b_n] = \{x\}$. It is easy to see that the sequence assigned to x is $\omega_1\omega_2\omega_3\cdots$. ■

Remark. We can show that in the above proof we have $x = \sum_{n\geq 1} \frac{\omega_n}{3^n}$. In other words, $(0.\omega_1\omega_2\omega_3\cdots)_3$ is a representation of x in base 3. To see this we can easily prove by induction that $a_n = \sum_{k=1}^{n} \frac{\omega_k}{3^k}$. Then as $x \in [a_n, b_n]$ and $b_n - a_n = \frac{1}{3^n}$ we obtain $0 \leq x - \sum_{k=1}^{n} \frac{\omega_k}{3^k} \leq \frac{1}{3^n}$. Now let $n \to \infty$ and get the desired.

Remark. Note that a point x is an endpoint of an interval of one of the C_n's, if and only if its associated sequence $\omega_1\omega_2\omega_3\cdots$ is constant 0 or constant 2 eventually.

Definition 2.129. A subset A of $\mathbb{R}$ has **measure zero** if for every $\epsilon > 0$ there exist countably many intervals (a_i, b_i) such that $A \subset \bigcup_{i\geq 1}(a_i, b_i)$, and

$$\sum_{i\geq 1} b_i - a_i < \epsilon.$$

Theorem 2.130. *The Cantor set has measure zero.*

Proof. Suppose $\epsilon > 0$ is given. Let n be large enough so that $\frac{2^n}{3^n} < \frac{\epsilon}{2}$. Let $[a_i, b_i]$ for $i = 1, \ldots, 2^n$ be the intervals of C_n. Then we have

$$C \subset \bigcup (a_i - \frac{\epsilon}{2^{n+2}}, b_i + \frac{\epsilon}{2^{n+2}}),$$

and $\sum[b_i + \frac{\epsilon}{2^{n+2}} - (a_i - \frac{\epsilon}{2^{n+2}})] = 2^n(\frac{1}{3^n} + \frac{\epsilon}{2^{n+1}}) < \epsilon$. ■

2.8 Fundamental Theorem of Algebra

Proposition 2.131. *Every complex number has an nth root for any $n \in \mathbb{N}$.*

Proof. Let z be a complex number, and suppose $z = a + ib$ where $a, b \in \mathbb{R}$. If $z = 0$ then $0^n = 0$. So we assume that $z \neq 0$. Then we have $z = r(\cos\theta + i\sin\theta)$, where $r = \sqrt{a^2 + b^2}$ and θ is the unique number in the interval $[0, 2\pi)$ such that $\cos\theta = \frac{a}{r}$ and $\sin\theta = \frac{b}{r}$. As shown in Chapter 6 we have

$$(\cos\alpha + i\sin\alpha)^n = \cos(n\alpha) + i\sin(n\alpha).$$

Hence we have $(\sqrt[n]{r}(\cos\frac{\theta}{n} + i\sin\frac{\theta}{n}))^n = z$. ■

Second Proof. We present another proof of this theorem that avoids using trigonometric functions. As before we can assume that $z = a + ib \neq 0$, where $a, b \in \mathbb{R}$. We are looking for $w \in \mathbb{C}$ that satisfies $w^n = z$. First suppose $n = 2$, and $w = c + id$ for some $c, d \in \mathbb{R}$. Then we have

$$c^2 - d^2 = a, \qquad 2cd = b.$$

If we multiply the first equation with c^2 and use the second equation, we obtain $c^4 - ac^2 - \frac{1}{4}b^2 = 0$. Therefore we get (using a similar equation for d)

$$c = \pm\sqrt{\frac{1}{2}(a + \sqrt{a^2 + b^2})}, \qquad d = \pm\sqrt{\frac{1}{2}(-a + \sqrt{a^2 + b^2})}.$$

(The signs of c, d are the same when $b \geq 0$, and are different otherwise.) Hence every complex number has two square roots. By induction on $k \in \mathbb{N}$ we can show that every complex number has a 2^kth root.
Now suppose n is odd. If u satisfies $u^n = \frac{z}{|z|}$ then $w = \sqrt[n]{|z|}u$ is an nth root of z. Hence we can assume that $|z| = 1$. For $s \in [0, 2]$ let $w_s := 1 - s + i\sqrt{2s - s^2}$ (note that $2s - s^2 \geq 0$ for $0 \leq s \leq 2$). Then as s goes from 0 to 2, w_s moves from 1 to -1 along the upper half of the unit circle in the complex plane. Let

$$r := \sup\{t \in [0, 2] : |w_s^n - 1| \leq |z - 1| \text{ for all } s \in [0, t]\}.$$

Note that $t = 0$ always belongs to the set, so the supremum exists. Now we claim that w_r or $\bar{w}_r$ is an nth root of z. First let us show that $|w_r^n - 1| = |z - 1|$. Due to the continuity we have $|w_r^n - 1| \leq |z - 1|$. Suppose to the contrary that $|w_r^n - 1| < |z - 1|$. But $r < 2$, since n is odd and we have

$$|w_2^n - 1| = |(-1)^n - 1| = 2 \geq 1 + |z| \geq |1 - z|.$$

Then again continuity implies that for all small ϵ we have $|w_{r+\epsilon}^n - 1| < |z - 1|$, which contradicts the fact that r is the supremum. Thus we must have $|w_r^n - 1| = |z - 1|$. On the other hand we have $|w_r| = 1$, which implies $|w_r^n| = 1 = |z|$.

We claim that these two equalities imply that $w_r^n = z$ or $\bar{w}_r^n = z$, as desired. Suppose $w_r^n = c + id$ for some $c, d \in \mathbb{R}$. Then we have

$$\begin{cases} c^2 + d^2 = a^2 + b^2 = 1, \\ (c-1)^2 + d^2 = (a-1)^2 + b^2. \end{cases}$$

By subtracting these two equations we obtain $2c - 1 = 2a - 1$, so $c = a$. Hence

$$b = \pm\sqrt{1-a^2} = \pm\sqrt{1-c^2} = \pm d.$$

Therefore either $w_r^n = z$ or $\bar{w}_r^n = \overline{w_r^n} = z$, and hence z has an nth root. Therefore every complex number has an nth root when n is odd.

Finally suppose n is an arbitrary positive integer. Then $n = 2^k m$ where m is odd. Let u be a complex number that satisfies $u^m = z$. Then let w be such that $w^{2^k} = u$. Hence we have $w^n = (w^{2^k})^m = u^m = z$. ■

Fundamental Theorem of Algebra. *Every nonconstant polynomial with complex coefficients has a root in complex numbers.*

Proof. **(Argand, 1806)** Let the polynomial be

$$p(z) := a_n z^n + \cdots + a_1 z + a_0,$$

with $n \geq 1$, and $a_n \neq 0$. First note that $|p(z)| \to +\infty$ as $|z| \to +\infty$. The reason is that

$$|p(z)| = |z|^n \, |a_n + \frac{a_{n-1}}{z} + \cdots + \frac{a_0}{z^n}| \to +\infty \times |a_n| = +\infty.$$

Thus there is $R > 0$ such that $|p(z)| > |p(0)|$ when $|z| > R$. Hence the continuous function $|p(z)|$ assumes its global minimum on the compact set $\{|z| \leq R\}$. Suppose the minimum occurs at z_0, so that

$$|p(z_0)| \leq |p(z)| \text{ for all } z \in \mathbb{C}.$$

Our goal is to show that $p(z_0) = 0$. Suppose to the contrary that $|p(z_0)| > 0$. We will find another point $\tilde{z}_0 \in \mathbb{C}$ such that $|p(\tilde{z}_0)| < |p(z_0)|$, and arrive at a contradiction. Let

$$q(w) := \frac{1}{p(z_0)} p(w + z_0) = \frac{a_n}{p(z_0)}(w + z_0)^n + \cdots + \frac{a_1}{p(z_0)}(w + z_0) + \frac{a_0}{p(z_0)}.$$

Note that by binomial expansion, q is a polynomial in w of degree n. Also $q(0) = 1$. Thus we have

$$q(w) = 1 + b_k w^k + \cdots + b_n w^n,$$

where k is the smallest positive integer for which $b_k \neq 0$. The idea is that as $w \to 0$ the dominant term is w^k, and we can make $b_k w^k$ a negative real number. Let us make this idea precise.

Let w_0 be a kth root of $-\frac{1}{b_k}$, that we know exists in $\mathbb{C}$. (This is our crucial use of complex numbers in this proof.) Then for $t \in \mathbb{R}$ we have

$$\begin{aligned} q(tw_0) &= 1 + b_k t^k w_0^k + b_{k+1} t^{k+1} w_0^{k+1} + \cdots + b_n t^n w_0^n \\ &= 1 - \quad t^k \quad + t^k (t b_{k+1} w_0^{k+1} + \cdots + t^{n-k} b_n w_0^n). \end{aligned}$$

Now note that the expression in the parentheses goes to 0 as $t \to 0$. Let t_0 be a small enough positive number so that the absolute value of the expression in the parentheses is less than $\frac{1}{2}$. Also assume that $t_0 < 1$. Then we have

$$|q(t_0 w_0)| \leq |1 - t_0^k| + \frac{1}{2}|t_0^k| = 1 - \frac{1}{2} t_0^k < 1.$$

Hence we have

$$|p(t_0 w_0 + z_0)| < |p(z_0)|.$$

This contradiction proves that $p(z_0) = 0$ as desired. ■

Theorem 2.132. *Let p be a polynomial with complex coefficients that has degree $n \geq 1$. Then there are (not necessarily distinct) complex numbers $\lambda_1, \ldots, \lambda_n$, and $c \in \mathbb{C} - \{0\}$, such that*

$$p(z) = c(z - \lambda_1) \cdots (z - \lambda_n).$$

Proof. The proof is by induction on n. For $n = 1$ the claim holds trivially. Now suppose it also holds for polynomials of degree $n - 1$. Then we know that p has a complex root λ_1. Hence there is a polynomial q of degree $n - 1$ such that

$$p(z) = (z - \lambda_1) q(z).$$

Now by the induction hypothesis q has a factorization

$$q(z) = c(z - \lambda_2) \cdots (z - \lambda_n).$$

Thus we get the desired factorization for p. Finally note that if $c = 0$ then $p = 0$, which contradicts the fact that $\deg p \geq 1$. ■

Theorem 2.133. *Suppose $p : \mathbb{R} \to \mathbb{R}$ is a polynomial with odd degree. Then p has a root in $\mathbb{R}$.*

Proof. Let

$$p(x) = a_n x^n + \cdots + a_0,$$

where $a_i \in \mathbb{R}$ and $a_n \neq 0$. Suppose $a_n > 0$, the other case is similar. Then we have

$$\begin{aligned}\lim_{x\to+\infty} p(x) &= \lim_{x\to+\infty} x^n \big(a_n + \frac{a_{n-1}}{x} + \cdots + \frac{a_0}{x^n}\big)\\ &= +\infty(a_n + 0 + \cdots + 0) = +\infty.\end{aligned}$$

This means that for large enough x, $p(x)$ is a large positive number. Similarly, using the fact that p has odd degree, we get $\lim_{x\to-\infty} p(x) = -\infty$. Hence $p(x)$ is negative for large negative x. Therefore by the intermediate value theorem $p(a) = 0$ for some a, since we know that p is a continuous function. ■

Chapter 3

Limits, Sequences, and Series

3.1 Limits

Definition 3.1. Suppose X, Y are metric spaces, and $A \subset X$. Let $a \in X$ be a limit point of A, and let $b \in Y$. Then given a function $f : A \to Y$, we say the **limit** of f as x approaches a is b, and we write $\lim\limits_{x \to a} f(x) = b$, if

$$\begin{aligned} &\forall \epsilon > 0\ \exists \delta > 0 \text{ such that } \forall x \in A \\ &\qquad 0 < d_X(x, a) < \delta \implies d_Y(f(x), b) < \epsilon. \end{aligned}$$

Remark. Note that f need not be defined at a.

Notation.

(i) The symbol $x \to a$ can be read "as x approaches a".

(ii) In this section, we assume that X, Y, Z are metric spaces containing the points a, b, c respectively. We also assume that $A \subset X$.

Theorem 3.2. *Suppose $f : A \to Y$, and a is a limit point of A. Then we have $\lim_{x \to a} f(x) = b$ if and only if the function $g : A \cup \{a\} \to Y$, defined by*

$$g(x) := \begin{cases} f(x) & x \neq a \\ b & x = a, \end{cases}$$

is continuous at a.

Proof. This is a trivial consequence of the definition of limit and the definition of continuity. ■

Theorem 3.3. *Suppose $f : A \to Y$, and $a \in A$ is a limit point of A. Then the function $f : A \to Y$ is continuous at the point a if and only if*

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

Proof. This follows immediately from the previous theorem by setting $g = f$. ■

Remark. The above two theorems enable us to deduce most of the theorems in this section, from the corresponding results about continuous functions.

Definition 3.4. Suppose X, Y are metric spaces, $A \subset X$, and $f : A \to Y$. Then we define **infinite limits** and **limits at infinity** as follows.

(i) Suppose $Y = \mathbb{R}^m$, and $a \in X$ is a limit point of A. We say $\lim_{x\to a} f(x) = \infty$ if

$$\forall M > 0\ \exists \delta > 0 \text{ such that } \forall x \in A$$
$$0 < d_X(x, a) < \delta \implies |f(x)| > M.$$

(ii) Suppose $b \in Y$, $X = \mathbb{R}^n$, and A is unbounded. We say $\lim_{x\to\infty} f(x) = b$ if

$$\forall \epsilon > 0\ \exists N > 0 \text{ such that } \forall x \in A$$
$$|x| > N \implies d_Y(f(x), b) < \epsilon.$$

(iii) Suppose $Y = \mathbb{R}^m$, $X = \mathbb{R}^n$, and A is unbounded. We say $\lim_{x\to\infty} f(x) = \infty$ if

$$\forall M > 0\ \exists N > 0 \text{ such that } \forall x \in A$$
$$|x| > N \implies |f(x)| > M.$$

Remark. For convenience, we say ∞ is a limit point of $A \subset \mathbb{R}^n$, if A is unbounded. Note that this happens if and only if A has elements with arbitrarily large norms.

Proposition 3.5. *Let $\widehat{\mathbb{R}^n}$ be $\mathbb{R}^n \cup \{\infty\}$, where ∞ is an object different from all elements of $\mathbb{R}^n$. Let*

$$\hat{d}(x, y) := \left|\frac{1}{1 + |x|^2} - \frac{1}{1 + |y|^2}\right| + \left|\frac{x}{1 + |x|^2} - \frac{y}{1 + |y|^2}\right|,$$
$$\hat{d}(x, \infty) = \hat{d}(\infty, x) := \frac{1 + |x|}{1 + |x|^2}, \qquad \hat{d}(\infty, \infty) := 0,$$

for $x, y \in \mathbb{R}^n$. Then $\hat{d}$ is a metric on $\widehat{\mathbb{R}^n}$.

Remark. ∞ is called *infinity*. In contrast, the elements of $\mathbb{R}^n$ are called *finite*.

Proof. First note that the denominators in the definition of $\hat{d}$ are all nonzero, so $\hat{d}$ is defined at every two points. It is also obvious that $\hat{d}$ is nonnegative and symmetric, and $\hat{d}(z, z) = 0$ for all $z \in \widehat{\mathbb{R}^n}$. Now note that $\hat{d}(x, \infty) > 0$ for all $x \in \mathbb{R}^n$. Suppose $\hat{d}(x, y) = 0$ for some $x, y \in \mathbb{R}^n$. Then the two terms in the definition of $\hat{d}(x, y)$ are both zero. Hence $|x| = |y|$, and therefore $x = y$. Thus $\hat{d}$ is positive definite too.

It only remains to show that $\hat{d}$ satisfies the triangle inequality. For $x, y, z \in \mathbb{R}^n$, we can add and subtract $\frac{1}{1+|z|^2}$, and $\frac{z}{1+|z|^2}$, in appropriate places, to get

$$\hat{d}(x,y) \leq \hat{d}(x,z) + \hat{d}(z,y).$$

Now when $z = \infty$ we have

$$\hat{d}(x,y) \leq \hat{d}(x,\infty) + \hat{d}(\infty,y),$$

since $|X - Y| \leq |X| + |Y|$ for X, Y in a Euclidean space. When one of the x, y, for example x, is ∞, we have

$$\begin{aligned}\hat{d}(z,y) &\geq \frac{1}{1+|y|^2} - \frac{1}{1+|z|^2} + \frac{|y|}{1+|y|^2} - \frac{|z|}{1+|z|^2} \\ &= \hat{d}(\infty,y) - \hat{d}(\infty,z),\end{aligned}$$

since $|X - Y| \geq |Y| - |X|$ for X, Y in a Euclidean space. Finally, when at least two of the x, y, z are ∞, the triangle inequality holds trivially. ■

Remark. $\widehat{\mathbb{R}^n}$ is actually a compact metric space. It is called the *one-point compactification* of $\mathbb{R}^n$. Also, the metric $\hat{d}$ is induced by the the so-called *stereographic projection* from the sphere S^n onto $\mathbb{R}^n$.

Theorem 3.6. *Suppose X, Y are metric spaces, $A \subset X$, and $f : A \to Y$.*

(i) *Suppose $Y = \mathbb{R}^m$, and $a \in X$ is a limit point of A. Let the function $g : A \to \widehat{\mathbb{R}^m}$ be equal to f at every point. Then we have:*
$\lim_{x\to a} f(x) = \infty$ *if and only if* $\lim_{x\to a} g(x) = \infty$ *as a function between two metric spaces.*

(ii) *Suppose $b \in Y$, $X = \mathbb{R}^n$, and A is unbounded. Let $\hat{A}$ be the subset of $\widehat{\mathbb{R}^n}$ that is equal to A. Let the function $\hat{g} : \hat{A} \to Y$ be equal to f at every point. Then we have:*
$\lim_{x\to\infty} f(x) = b$ *if and only if* $\lim_{x\to\infty} \hat{g}(x) = b$ *as a function between two metric spaces.*

(iii) *Suppose $Y = \mathbb{R}^m$, $X = \mathbb{R}^n$, and A is unbounded. Let $\hat{A}$ be the subset of $\widehat{\mathbb{R}^n}$ that is equal to A. Let the function $\hat{f} : \hat{A} \to \widehat{\mathbb{R}^m}$ be equal to f at every point. Then we have:*
$\lim_{x\to\infty} f(x) = \infty$ *if and only if* $\lim_{x\to\infty} \hat{f}(x) = \infty$ *as a function between two metric spaces.*

Proof. We will only prove (iii). The other two cases are similar. Note that for all $x \in A = \hat{A}$ we have $f(x) = \hat{f}(x)$, so we will denote these values simply by $f(x)$.

First suppose A is unbounded, and we have $\lim_{x\to\infty} f(x) = \infty$. Then ∞ is a limit point of $\hat{A}$. Because if $a_k \in A$ satisfies $|a_k| > k$, then we have

$$\hat{d}(a_k, \infty) = \frac{1 + |a_k|}{1 + |a_k|^2} \leq \frac{1}{|a_k|^2} + \frac{1}{|a_k|} < \frac{1}{k^2} + \frac{1}{k} \leq \frac{2}{k} \xrightarrow[k\to\infty]{} 0.$$

Now for a given $\epsilon > 0$, we want to find $\delta > 0$, such that for all $x \in \hat{A} = A$ with $0 < \hat{d}(x, \infty) < \delta$ we have $\hat{d}(f(x), \infty) < \epsilon$. Suppose $\epsilon < 1$. We know that there is $N > 0$, such that for all $x \in A$ with $|x| > N$ we have $y := |f(x)| > \frac{2}{\epsilon}$. Note that $y > 1$, so $y + 1 < 2y$. Then we have

$$\frac{1 + y^2}{1 + y} > \frac{y^2}{2y} = \frac{y}{2} > \frac{1}{\epsilon} \implies \hat{d}(f(x), \infty) = \frac{1 + y}{1 + y^2} < \epsilon.$$

On the other hand, we can assume that $N > 1$. Then we have

$$\frac{1 + |x|^2}{1 + |x|} > \frac{|x|^2}{2|x|} = \frac{|x|}{2} > \frac{N}{2} \implies 0 < \hat{d}(x, \infty) < \frac{2}{N}.$$

Hence we can take δ to be $\frac{2}{N}$.

Conversely, suppose ∞ is a limit point of $\hat{A}$, and $\lim_{x\to\infty} \hat{f}(x) = \infty$ as a function between two metric spaces. First note that if $a_k \in \hat{A}$ satisfies $\hat{d}(a_k, \infty) < \frac{1}{k}$, then we have

$$1 + |a_k|^2 \geq \frac{1 + |a_k|^2}{1 + |a_k|} > k \implies |a_k| > \sqrt{k - 1}.$$

Then for $k = L^2 + 1$ we have $|a_k| > L$. Hence A is unbounded. Now for a given $M > 0$, we want to find $N > 0$, such that for all $x \in A$ with $|x| > N$ we have $|f(x)| > M$. We know that there is $\delta > 0$, such that for all $x \in \hat{A} = A$ with $0 < \hat{d}(x, \infty) < \delta$ we have $\hat{d}(f(x), \infty) < \frac{1}{1+M^2}$. Then we have

$$1 + |f(x)|^2 \geq \frac{1 + |f(x)|^2}{1 + |f(x)|} > 1 + M^2 \implies |f(x)| > M.$$

As before, we can show that if $|x| > N > \max\{1, \frac{2}{\delta}\}$, then $0 < \hat{d}(x, \infty) < \frac{2}{N} < \delta$. So we get the desired. ■

Remark. The significance of the above theorem is that it allows us to deal with finite limits, limits at infinity, and infinite limits, simultaneously. In particular, **in all of the following theorems a can be ∞ too.**

Theorem 3.7. *Suppose $f : A \to Y$, and a is a limit point of A. If $\lim_{x\to a} f(x)$ exists, it is unique.*

$\boxed{\text{Proof.}}$ Suppose to the contrary that there are two limits b_1, b_2. Then for $x \in A$ near a we must have

$$d_Y(f(x), b_1) < \frac{1}{2} d_Y(b_1, b_2), \qquad d_Y(f(x), b_2) < \frac{1}{2} d_Y(b_1, b_2).$$

But this is in contradiction with the triangle inequality for d_Y. Note that there is at least one $x \in A$ close enough to a, so that $f(x)$ is close enough to $\lim_{x\to a} f(x)$. Because a is a limit point of A. ■

Definition 3.8. Suppose X, Y are metric spaces, $A \subset X$, and $f : A \to Y$. Also suppose a is a limit point of A.

(i) When $X = \mathbb{R}$ and $a \in \mathbb{R}$, we use the notations

$$\lim_{x\to a^+} f(x), \qquad \lim_{x\to a^-} f(x),$$

to denote the limits (if they exist) as $x \to a$ of the functions $f|_{A\cap(a,\infty)}$, and $f|_{A\cap(-\infty,a)}$, respectively. We call these limits the **right-hand limit**, and the **left-hand limit**, respectively. We also refer to these limits as **one-sided limits**.

(ii) When $X = \mathbb{R}$ and $a = \infty$, i.e. for limits at infinity, we use the notations

$$\lim_{x\to+\infty} f(x), \qquad \lim_{x\to-\infty} f(x),$$

to denote the limits at infinity (if they exist) of the functions $f|_{A\cap(0,\infty)}$, and $f|_{A\cap(-\infty,0)}$, respectively.

(iii) Finally suppose $Y = \mathbb{R}$, and $\lim_{x\to a} f(x) = \infty$. Then we say

$$\lim_{x\to a} f(x) = +\infty, \qquad \text{or} \qquad \lim_{x\to a} f(x) = -\infty,$$

if f is respectively positive, or negative, on $A \cap U$ where U is a neighborhood of a. Note that here we can also let a be ∞. In addition, we can also replace $x \to a$ with $x \to a^\pm$, or $x \to \pm\infty$.

Remark. We can easily see that $\lim_{x\to a} f(x) = +\infty$ if and only if for every $M > 0$, we can take x to be close enough to a so that $f(x) > M$. Similarly, $\lim_{x\to a} f(x) = -\infty$ if and only if for every $M > 0$, we can take x to be close enough to a so that $f(x) < -M$.

***Exercise* 3.9.** Consider the set of extended real numbers, $\mathbb{R} \cup \{\pm\infty\}$. Let

$$\tilde{d}(x, y) := \left| \frac{x}{1+|x|} - \frac{y}{1+|y|} \right|, \qquad \tilde{d}(+\infty, -\infty) := 2,$$
$$\tilde{d}(x, +\infty) := 1 - \frac{x}{1+|x|}, \qquad \tilde{d}(x, -\infty) := 1 + \frac{x}{1+|x|},$$

for $x, y \in \mathbb{R}$.

(i) Show that $\tilde{d}$ is a metric on $\mathbb{R} \cup \{\pm\infty\}$.

(ii) Show that a function f into $\mathbb{R}$ has limit $\pm\infty$, if and only if when we consider f as a function into $\mathbb{R} \cup \{\pm\infty\}$, it has the limit $\pm\infty$, as a function between two metric spaces.

(iii) Show that a function g defined on a subset of $\mathbb{R}$ has limit as $x \to \pm\infty$, if and only if when we consider g as a function defined on a subset of $\mathbb{R} \cup \{\pm\infty\}$, it has the same limit as $x \to \pm\infty$, as a function between two metric spaces.

Definition 3.10. We say a sequence (a_n) in $\mathbb{R}^m$ **diverges to infinity**, and we write $\lim a_n = \infty$, or $a_n \to \infty$, if as a sequence in $\widehat{\mathbb{R}^m}$, (a_n) converges to ∞.

Suppose $m = 1$ and $a_n \to \infty$. We say $a_n \to +\infty$, if there is $N \in \mathbb{N}$ such that for $n \geq N$ we have $a_n > 0$. We also say $a_n \to -\infty$, if there is $N \in \mathbb{N}$ such that for $n \geq N$ we have $a_n < 0$.

Proposition 3.11. *Suppose $A \subset \mathbb{R}$, and $f : A \to Y$. Also suppose $a \in \mathbb{R}$ is a limit point of both $A \cap (-\infty, a)$ and $A \cap (a, \infty)$. Then*

$$\lim_{x\to a} f(x) = b \iff \lim_{x\to a^-} f(x) = b = \lim_{x\to a^+} f(x).$$

Proof. First note that a is obviously a limit point of A too. Suppose that both one-sided limits equal b. Then for a given $\epsilon > 0$ there are $\delta_1, \delta_2 > 0$ such that for $x \in A$ we have

$$\begin{aligned} 0 < |x - a| < \delta_1,\ x < a &\implies d_Y(f(x), b) < \epsilon, \\ 0 < |x - a| < \delta_2,\ x > a &\implies d_Y(f(x), b) < \epsilon. \end{aligned} \tag{$*$}$$

Hence for $\delta = \min\{\delta_1, \delta_2\}$ and $x \in A$ we have

$$0 < |x - a| < \delta \implies d_Y(f(x), b) < \epsilon. \tag{$**$}$$

Thus $\lim_{x\to a} f(x) = b$. Conversely, if for a given $\epsilon > 0$ there is $\delta > 0$ such that $(**)$ holds for all $x \in A$, then for $\delta_1, \delta_2 = \delta$ both implications of $(*)$ also hold for all $x \in A$. Therefore both one-sided limits exist and are equal to b. ■

Theorem 3.12. *Suppose $f : A \to Y$, and a is a limit point of A. Then we have $\lim_{x\to a} f(x) = b$ if and only if for any sequence $\{a_n\} \subset A - \{a\}$, where $a_n \to a$, we have $f(a_n) \to b$. Here, a, b can be ∞ or $\pm\infty$ too.*

Proof. Suppose $\lim\limits_{x\to a} f(x) = b$. We know that

$$g(x) := \begin{cases} f(x) & x \neq a \\ b & x = a, \end{cases}$$

is continuous at a. Hence $f(a_n) = g(a_n) \to g(a) = b$. Note that it is crucial that the sequence does not intersect $\{a\}$, since if $f(a) \neq b$ then $f(a_n)$ can have a constant subsequence that does not converge to b.

On the other hand, suppose $f(a_n) \to b$ for any sequence $\{a_n\} \subset A - \{a\}$, where $a_n \to a$. It is enough to show that g is continuous at a. Suppose (c_n) is a sequence in $A \cup \{a\}$ that converges to a. Then we have

$$g(c_n) = \begin{cases} f(c_n) & c_n \neq a, \\ b & c_n = a. \end{cases}$$

Now for large enough n, $f(c_n)$ is close to b when $c_n \neq a$. Also b is close to b when $c_n = a$. Thus $g(c_n) \to b$.

For the infinite limits, the only case that needs special consideration is when $b = \pm\infty$. Because all the other types of infinite limits and limits at infinity, can be considered as limits of functions between two metric spaces. Suppose that for example $b = +\infty$, the other case is similar. Then $\lim_{x\to a} f(x) = \infty$, and $f(x) > 0$ on $A \cap U$ where U is a neighborhood of a. Let $\{a_n\} \subset A - \{a\}$ be a sequence such that $a_n \to a$. Then by the previous part we know that $f(a_n) \to \infty$. Also, $f(a_n) > 0$ for large n, since $a_n \in U \cap A$ for large enough n. Conversely, suppose that $f(a_n) \to \infty$ and $f(a_n) > 0$ for large enough n, for any sequence $\{a_n\} \subset A - \{a\}$ such that $a_n \to a$. Then we know that $\lim_{x\to a} f(x) = \infty$. So we only need to show that $f > 0$ on a neighborhood of a in A. If this does not happen, then there is $a_n \in B_{\frac{1}{n}}(a) \cap A$ such that $f(a_n) < 0$. But then we would have $a_n \to a$, which is a contradiction. ■

Remark. In the above theorem, the assumption $\{a_n\} \subset A - \{a\}$ is essential. For example let

$$F(x) := \begin{cases} 0 & x \neq 0 \\ 1 & x = 0. \end{cases}$$

Then $\lim_{x\to 0} F(x) = 0$. But for the constant sequence $a_n = 0$ we have $F(a_n) = F(0) = 1 \not\to 0$.

Theorem 3.13. *Suppose $f : A \to Y$, and a is a limit point of A. Then $\lim_{x\to a} f(x)$ is in the closure of the image of f.*

Proof. Since a is a limit point of A, there is a sequence (a_n) of distinct points of A that converges to a. Thus a_n belongs to $A - \{a\}$ for large enough n, since at most one them can be equal to a. Then by Theorem 3.12, $\lim_{x\to a} f(x)$ is the limit of the sequence $(f(a_n))$. Hence it belongs to $\overline{f(A)}$. ■

Theorem 3.14. *Suppose $f : A \to Y$, and a is a limit point of A. Let $b = \lim_{x\to a} f(x)$, and suppose $b \in U$, where $U \subset Y$ is an open set. Then there is an open set V containing a, such that $f(x) \in U$ when $x \in A \cap V$.*

Proof. In the definition of limit, let ϵ be small enough so that $B_\epsilon(b) \subset U$, and consider the corresponding δ. Then $V = B_\delta(a)$ has the desired property. ■

Remark. The most useful case of the above theorem is when the range is $\mathbb{R}$, and U is the set of positive or negative numbers. Then we conclude that f is respectively positive or negative, on the intersection of A and an open neighborhood of a.

Theorem 3.15. *Suppose $Y_1, \dots, Y_k$ are metric spaces, and $f_i : A \to Y_i$. Also suppose a is a limit point of A. Let*

$$f = (f_1, \dots, f_k) : A \to Y_1 \times \cdots \times Y_k.$$

Then $\lim_{x\to a} f(x) = (b_1, \dots, b_k)$ if and only if $\lim_{x\to a} f_i(x) = b_i$ for each i.

Proof. Apply Theorem 3.2, and the corresponding result for continuous functions. ■

Theorem 3.16. *Suppose $f : A \to Y$, and a is a limit point of A. Let $b = \lim_{x\to a} f(x)$, and suppose $V \subset Y$ is a neighborhood of b. Also suppose $F : V \to Z$ is continuous at b. Then we have*

$$\lim_{x\to a} F(f(x)) = F\big(\lim_{x\to a} f(x)\big).$$

Proof. We know that

$$g(x) := \begin{cases} f(x) & x \neq a \\ b & x = a \end{cases}$$

is continuous at a. Then $F \circ g$ is continuous at a, i.e.

$$F(g(x)) = \begin{cases} F(f(x)) & x \neq a \\ F(b) & x = a \end{cases}$$

is continuous at a. Hence

$$\lim_{x\to a} F(f(x)) = F(b).$$ ■

Theorem 3.17. *Suppose $f : A \to Y$, and a is a limit point of A. Let $b = \lim_{x\to a} f(x)$, and suppose for $B \subset Y$ we have*

$$f(A - \{a\}) \subset B - \{b\}.$$

Also suppose $F : B \to Z$, and $\lim_{y\to b} F(y) = c$. Then $\lim_{x\to a} F(f(x)) = c$. Here, a, b, c can be ∞ or $\pm\infty$ too.

Proof. Let (a_n) be a sequence in $A - \{a\}$ such that $a_n \to a$. Then $f(a_n) \to b$ by Theorem 3.12, and $f(a_n) \in B - \{b\}$ by assumption. Hence again by Theorem 3.12 we have $F(f(a_n)) \to c$, and consequently $\lim_{x\to a} F(f(x)) = c$. ■

Remark. In the above theorem, the assumption $f(A - \{a\}) \subset B - \{b\}$ is essential. For example $\lim_{x\to 0} x\sin(\frac{1}{x}) = 0$, and for

$$F(y) := \begin{cases} y & y \neq 0 \\ 1 & y = 0 \end{cases}$$

we have $\lim_{y\to 0} F(y) = 0$. But $\lim_{x\to 0} F(x\sin(\frac{1}{x}))$ does not exist, since for $n \in \mathbb{N}$ we have

$$\begin{aligned} F\Big(\frac{1}{n\pi}\sin(n\pi)\Big) &= F(0) = 1, \\ F\Big(\frac{2}{(2n+1)\pi}\sin\big(\frac{(2n+1)\pi}{2}\big)\Big) &= \frac{2}{(2n+1)\pi} \to 0. \end{aligned}$$

Remark. The above theorem allows us to change the variable when we compute a limit. A particular interesting case is when $f : (0, 1) \to (1, +\infty)$ maps x to $y = \frac{1}{x}$. Then for $F : (1, +\infty) \to \mathbb{R}$ we have

$$\lim_{y\to+\infty} F(y) = c \iff \lim_{x\to 0^+} F\big(\frac{1}{x}\big) = c.$$

For the $\Longleftarrow$ implication we have to use $f^{-1} : (1, +\infty) \to (0, 1)$ that maps y to $x = \frac{1}{y}$.

Theorem 3.18. *Let a be a limit point of A, and let $c_1, c_2 \in \mathbb{R}$. Suppose $F, G : A \to \mathbb{R}^m$ and $h : A \to \mathbb{R}$ have finite limits at a. Then*

(i) $$\lim_{x\to a} \big(c_1F(x) + c_2G(x)\big) = c_1 \lim_{x\to a} F(x) + c_2 \lim_{x\to a} G(x).$$

(ii) $$\lim_{x\to a} \big(h(x)F(x)\big) = \big(\lim_{x\to a} h(x)\big)\big(\lim_{x\to a} F(x)\big).$$

(iii) *If $\lim_{x\to a} h(x) \neq 0$ then h is nonzero on a neighborhood of a, and we have*

$$\lim_{x\to a} \frac{1}{h(x)} = \frac{1}{\lim\limits_{x\to a} h(x)}.$$

Proof. Apply Theorem 3.2, and the corresponding results for continuous functions. Note that if the limit of the real-valued function h is positive, or negative, then h is respectively positive, or negative, on a neighborhood of a. ■

Theorem 3.19. *Let a be a limit point of A. Suppose one of the two functions $F : A \to \mathbb{R}^m$ and $h : A \to \mathbb{R}$ is bounded, and the other one has limit 0 as $x \to a$. Then $\lim_{x\to a} h(x)F(x) = 0$.*

Proof. For x close enough to a, one of the $|F(x)|$ or $|h(x)|$ is less than some constant $M > 0$, and the other one can be made less than $\frac{\epsilon}{M}$, for any given $\epsilon > 0$. Hence $|h(x)F(x)| < \epsilon$. ■

Theorem 3.20. *Suppose $f : A \to \mathbb{R}^m$, and a is a limit point of A. Then we have*

(i)
$$\lim_{x\to a} f(x) = 0 \iff \lim_{x\to a} |f(x)| = 0.$$

(ii)
$$\lim_{x\to a} f(x) = \infty \iff \lim_{x\to a} |f(x)| = +\infty.$$

Proof. Both claims can be proved easily from the definition. For (i) we can use $|f(x) - 0| = |f(x)| = \big||f(x)| - 0\big|$. And for (ii) we can use $|f(x)| = \big||f(x)|\big|$. ■

Theorem 3.21. *For the function $\frac{1}{|x|} : \mathbb{R}^n - \{0\} \to \mathbb{R}$ we have*

$$\lim_{x\to\infty} \frac{1}{|x|} = 0, \qquad \lim_{x\to 0} \frac{1}{|x|} = +\infty.$$

And for the function $\frac{1}{x} : \mathbb{R} - \{0\} \to \mathbb{R}$ we have

$$\lim_{x\to-\infty} \frac{1}{x} = \lim_{x\to+\infty} \frac{1}{x} = 0, \qquad \lim_{x\to 0^+} \frac{1}{x} = +\infty, \qquad \lim_{x\to 0^-} \frac{1}{x} = -\infty.$$

Proof. Just note that for $N > 0$ we have $0 < |x| < \frac{1}{N}$ if and only if $\frac{1}{|x|} > N$. The case of $\frac{1}{x}$ is similar. ■

Theorem 3.22. *Let a be a limit point of A. Suppose $f, g, h : A \to \mathbb{R}$, and we have $f \le g \le h$ on $U \cap A$, where U is a neighborhood of a.*

(i) *If f, g have finite limits at a, then we have*

$$\lim_{x\to a} f(x) \le \lim_{x\to a} g(x).$$

(ii) **(Squeeze Theorem)** *Let $b \in \mathbb{R}$ be a finite number. Then*

$$\lim_{x\to a} f(x) = b = \lim_{x\to a} h(x) \implies \lim_{x\to a} g(x) = b.$$

(iii) $\lim_{x\to a} f(x) = +\infty$ *implies* $\lim_{x\to a} g(x) = +\infty$.

(iv) $\lim_{x\to a} h(x) = -\infty$ *implies* $\lim_{x\to a} g(x) = -\infty$.

Proof. (i) We know that $g-f \geq 0$. Since $[0,\infty)$ is closed, it contains the closure of the image of $g-f$. Hence we have

$$\lim_{x\to a} g(x) - \lim_{x\to a} f(x) = \lim_{x\to a}\big(g(x)-f(x)\big) \in [0,\infty).$$

(ii) Suppose $\epsilon > 0$ is given. Then for x close enough to a we have

$$-\epsilon < f(x)-b \leq g(x)-b \leq h(x)-b < \epsilon.$$

(iii) Let $M>0$ be given. Then for x close enough to a we have $g(x) \geq f(x) > M$.

(iv) Suppose $M>0$ is a given positive number. Then for x close enough to a we have $g(x) \leq h(x) < -M$. ■

Theorem 3.23. *Let a be a limit point of A. Suppose $f,g : A \to \mathbb{R}$.*

(i) *We have*

$$\lim_{x\to a} f(x) = +\infty \iff \lim_{x\to a}(-f(x)) = -\infty.$$

(ii) *If $\lim_{x\to a} f(x) = +\infty$, and $\lim_{x\to a} g(x)$ is either a finite number or $+\infty$, then*

$$\lim_{x\to a}\big(f(x)+g(x)\big) = +\infty.$$

(iii) *If $\lim_{x\to a} f(x) = +\infty$, and $\lim_{x\to a} g(x)$ is either a positive finite number or $+\infty$, then*

$$\lim_{x\to a}\big(f(x)g(x)\big) = +\infty.$$

(iv) *If $\lim_{x\to a} f(x) = \infty$ then f is nonzero on a neighborhood of a, and we have*

$$\lim_{x\to a}\frac{1}{f(x)} = 0.$$

(v) *If $\lim_{x\to a} f(x) = 0$, and f is nonzero on a neighborhood of a, then we have*

$$\lim_{x\to a}\frac{1}{f(x)} = \infty.$$

Proof. (i) Just note that for $M>0$ we have $f(x) > M$ if and only if $-f(x) < -M$.

(ii) Note that on a neighborhood of a, g is bounded below. Since otherwise we would get a sequence $a_n \to a$ such that $a_n \neq a$, and $g(a_n) \to -\infty$. But this is in contradiction with our assumption due to the Theorem 3.12. So $g > -C$ for some $C>0$. Thus $f(x)+g(x) > M$ if $f(x) > M+C$.

(iii) Note that on a neighborhood of a, $g > c$ for some $c > 0$. This is an easy consequence of the definition of limit. Hence we have $f(x)g(x) > M$ if $f(x) > \frac{M}{c}$.

(iv) Note that as an easy consequence of the definition of limit, f is nonzero on a neighborhood of a. Now we have $|\frac{1}{f(x)}| < \epsilon$ if $|f(x)| > \frac{1}{\epsilon}$.

(v) We have $|f(x)| > M$ if $|\frac{1}{f(x)}| < \frac{1}{M}$. ■

Remark. In the above theorem, if we combine part (i) with parts (ii) and (iii), we obtain

(i) If $\lim_{x\to a} f(x) = -\infty$, and $\lim_{x\to a} g(x)$ is either a finite number or $-\infty$, then

$$\lim_{x\to a} \big(f(x) + g(x)\big) = -\infty.$$

(ii) If $\lim_{x\to a} f(x) = +\infty$, and $\lim_{x\to a} g(x)$ is either a negative finite number or $-\infty$, then

$$\lim_{x\to a} \big(f(x)g(x)\big) = -\infty.$$

(iii) If $\lim_{x\to a} f(x) = -\infty$, and $\lim_{x\to a} g(x)$ is either a positive finite number or $+\infty$, then

$$\lim_{x\to a} \big(f(x)g(x)\big) = -\infty.$$

(iv) If $\lim_{x\to a} f(x) = -\infty$, and $\lim_{x\to a} g(x)$ is either a negative finite number or $-\infty$, then

$$\lim_{x\to a} \big(f(x)g(x)\big) = +\infty.$$

In addition we have

(v) If $\lim_{x\to a} f(x) = 0$, and f is nonzero on a neighborhood of a, and $\lim_{x\to a} g(x)$ is either a nonzero finite number or ∞, then

$$\lim_{x\to a} \frac{g(x)}{f(x)} = \infty.$$

Because $\lim_{x\to a} \frac{1}{|f(x)|} = +\infty$, and $\lim_{x\to a} |g(x)|$ is either a positive finite number or $+\infty$, hence $\lim_{x\to a} \left|\frac{g(x)}{f(x)}\right| = +\infty$.

(vi) If $\lim_{x\to a} f(x) = \infty$, and $\lim_{x\to a} g(x)$ is a finite number, then

$$\lim_{x\to a} \frac{g(x)}{f(x)} = \lim_{x\to a} \frac{1}{f(x)} \cdot \lim_{x\to a} g(x) = 0 \cdot \lim_{x\to a} g(x) = 0.$$

Thus we can compute the limit of all combinations of functions with finite or infinite limits, except for the famous *indeterminate forms*

$$\frac{0}{0}, \quad \frac{\infty}{\infty}, \quad 0\cdot\infty, \quad \infty - \infty.$$

3.2 Sequences

Proposition 3.24. *Let $\widehat{\mathbb{N}}$ be $\mathbb{N} \cup \{\infty\}$, where ∞ is an object different from all elements of $\mathbb{N}$. Let*

$$\hat{d}(m,n) := \left|\frac{1}{m} - \frac{1}{n}\right|, \qquad \hat{d}(m,\infty) = \hat{d}(\infty,m) := \frac{1}{m}, \qquad \hat{d}(\infty,\infty) := 0,$$

for $m,n \in \mathbb{N}$. Then $\hat{d}$ is a metric on $\widehat{\mathbb{N}}$.

Remark. ∞ is called *infinity*. In contrast, the elements of $\mathbb{N}$ are called *finite*.

Proof. It is obvious that $\hat{d}$ is symmetric and positive definite. So we only have to show that $\hat{d}$ satisfies the triangle inequality. For $m, n, k \in \mathbb{N}$ we have

$$\hat{d}(m,n) = \left|\frac{1}{m} - \frac{1}{n}\right| \le \left|\frac{1}{m} - \frac{1}{k}\right| + \left|\frac{1}{k} - \frac{1}{n}\right| = \hat{d}(m,k) + \hat{d}(k,n).$$

Now when $k = \infty$ we have

$$\hat{d}(m,n) = \left|\frac{1}{m} - \frac{1}{n}\right| \le \frac{1}{m} + \frac{1}{n} = \hat{d}(m,\infty) + \hat{d}(\infty,m).$$

When one of the m, n, for example m, is ∞, we have

$$\hat{d}(\infty,n) - \hat{d}(\infty,k) = \frac{1}{n} - \frac{1}{k} \le \left|\frac{1}{k} - \frac{1}{n}\right| = \hat{d}(k,n).$$

Finally, when at least two of the m, n, k are ∞, the triangle inequality holds trivially. ■

Theorem 3.25. *Let (a_n) be a sequence in a metric space X. Let $f : \mathbb{N} \to X$ be the function that defines the sequence, i.e. $f(n) = a_n$. Then $\lim a_n = a$ if and only if $\lim_{n\to\infty} f(n) = a$ as a function between two metric spaces, when we regard $\mathbb{N}$ as a subset of $\widehat{\mathbb{N}}$.*

Proof. First note that ∞ is a limit point of $\mathbb{N} \subset \widehat{\mathbb{N}}$, since $\lim n = \infty$. Now suppose $\lim a_n = a$. Then for any $\epsilon > 0$ there is $N \in \mathbb{N}$, such that for all $n \ge N$ we have $d_X(a_n, a) < \epsilon$. We want to show that $\lim_{n\to\infty} f(n) = a$, i.e. for any given $\epsilon > 0$ we want to find $\delta > 0$, such that if $\hat{d}(n,\infty) < \delta$ then $d_X(f(n), a) = d_X(a_n, a) < \epsilon$. But

$$\hat{d}(n,\infty) = \frac{1}{n} < \delta \iff n > \frac{1}{\delta}.$$

So it suffices to have $\frac{1}{\delta} \ge N$, or equivalently $\delta \le \frac{1}{N}$. Conversely suppose that $\lim_{n\to\infty} f(n) = a$. This time we want to find N. It is obvious from the above that it suffices to take $N > \frac{1}{\delta}$. ■

Remark. The above theorem enables us to apply all the results about limits of functions, to the limits of sequences.

Theorem 3.26. *Suppose (a_n), (b_n) are convergent sequences in $\mathbb{R}^m$, and (c_n) is a convergent sequence in $\mathbb{R}$. Then*

(i) $\lim(a_n + b_n) = \lim a_n + \lim b_n$.

(ii) $\lim c_n a_n = \lim c_n \lim a_n$.

(iii) *If $\lim c_n \ne 0$ then $c_n \ne 0$ for large n, and we have $\lim \frac{1}{c_n} = \frac{1}{\lim c_n}$.*

Proof. This is an obvious consequence of the corresponding result about limits of functions. For part (iii) note that any neighborhood of $\infty \in \widehat{\mathbb{N}}$, contains an open ball around ∞, and open balls around ∞ are of the form $\{n > N\}$ for some $N \in \mathbb{N}$. ■

Remark. Remember that we say a sequence (a_n) in $\mathbb{R}^m$ **diverges to infinity**, and we write $\lim a_n = \infty$, or $a_n \to \infty$, if as a sequence in $\widehat{\mathbb{R}^m}$, (a_n) converges to ∞.

Also when $m = 1$ and $a_n \to \infty$, we say $a_n \to +\infty$, if there is $N \in \mathbb{N}$ such that for $n \geq N$ we have $a_n > 0$. Similarly, when $m = 1$ and $a_n \to \infty$, we say $a_n \to -\infty$, if there is $N \in \mathbb{N}$ such that for $n \geq N$ we have $a_n < 0$.

By using the fact that sequences that diverge to infinity are special cases of functions that have infinite limit, we obtain the following result for a sequence (a_n) in $\mathbb{R}^m$.

(i) $\lim a_n = \infty$ if and only if

$$\forall M > 0 \, \exists N \in \mathbb{N} \text{ such that } \forall n \geq N \text{ we have } |a_n| > M.$$

Also, when $m = 1$ we have

(ii) $\lim a_n = +\infty$ if and only if

$$\forall M > 0 \, \exists N \in \mathbb{N} \text{ such that } \forall n \geq N \text{ we have } a_n > M.$$

(iii) $\lim a_n = -\infty$ if and only if

$$\forall M > 0 \, \exists N \in \mathbb{N} \text{ such that } \forall n \geq N \text{ we have } a_n < -M.$$

Theorem 3.27. *Let (a_n), (b_n), and (c_n) be sequences in $\mathbb{R}$. Suppose that there is $N \in \mathbb{N}$ such that for $n \geq N$ we have $a_n \leq b_n \leq c_n$.*

(i) *If $(a_n), (b_n)$ converge, then $\lim a_n \leq \lim b_n$.*

(ii) **(Squeeze Theorem)** *Suppose $(a_n), (c_n)$ are convergent, and*

$$\lim a_n = b = \lim c_n$$

for some $b \in \mathbb{R}$. Then (b_n) is also convergent, and we have $\lim b_n = b$.

(iii) $\lim a_n = +\infty$ *implies* $\lim b_n = +\infty$.

(iv) $\lim c_n = -\infty$ *implies* $\lim b_n = -\infty$.

Proof. This is just a special case of Theorem 3.22. Note that any neighborhood of $\infty \in \widehat{\mathbb{N}}$, contains an open ball around ∞, and open balls around ∞ are of the form $\{n > N - 1\}$ for some $N \in \mathbb{N}$. ■

Theorem 3.28. *An increasing sequence in $\mathbb{R}$ is convergent if and only if it is bounded above, and in this case the limit of the sequence is its supremum. Similarly, a decreasing sequence in $\mathbb{R}$ is convergent if and only if it is bounded below, and in this case the limit of the sequence is its infimum.*

Proof. First note that if a sequence converges, its terms are eventually in a neighborhood of its limit, therefore it is bounded. Now for the converse, let (a_n) be a bounded above increasing sequence in $\mathbb{R}$. Let $a := \sup\{a_n\}$. For any $\epsilon > 0$ we know that $a - \epsilon$ is not an upper bound of $\{a_n\}$. Therefore there is N such that $a - \epsilon < a_N$. Hence by monotonicity, for all $n \geq N$ we have

$$a - \epsilon < a_N \leq a_n \leq a < a + \epsilon.$$

Thus $a_n \to a$. The case of decreasing sequences is similar. ■

Theorem 3.29. *An unbounded increasing sequence in $\mathbb{R}$ diverges to $+\infty$, which is also its supremum. And, an unbounded decreasing sequence in $\mathbb{R}$ diverges to $-\infty$, which is also its infimum.*

Proof. Let (a_n) be an unbounded increasing sequence. For any $M > 0$, we know that M is not an upper bound of $\{a_n\}$. Therefore there is N such that $M < a_N$. Hence by monotonicity, for all $n \geq N$ we have $M < a_N \leq a_n$. Thus $a_n \to +\infty$. Note that we also have $\sup\{a_n\} = +\infty$, since $\{a_n\}$ is not bounded above. The case of decreasing sequences is similar. ■

Theorem 3.30. *For $r \in \mathbb{R}$ we have*

$$\lim_{n\to\infty} r^n = \begin{cases} 0 & |r| < 1, \\ \infty & |r| > 1. \end{cases}$$

Proof. If $|r| > 1$, then $b := |r| - 1 > 0$. By an easy induction we can show that for $n > 1$ we have

$$|r|^n = (1+b)^n > 1 + nb.$$

Therefore $|r^n| = |r|^n > M$, for $n > \frac{M-1}{b}$. Therefore $r^n \to \infty$.

If $0 < |r| < 1$, then $\frac{1}{|r|} > 1$. Thus $\frac{1}{|r|^n} \to \infty$. Therefore for any given $\epsilon > 0$, there is $N \in \mathbb{N}$ such that for $n \geq N$ we have $\frac{1}{|r|^n} > \frac{1}{\epsilon}$. So for $n \geq N$ we have $|r^n| = |r|^n < \epsilon$. Hence $r^n \to 0$. ■

Theorem 3.31. *For rational $p > 0$, and $n \in \mathbb{N}$, we have $\lim_{n\to\infty} \frac{1}{n^p} = 0$.*

Proof. Suppose $p = \frac{m}{k}$. Then we have $n^p = (\sqrt[k]{n})^m$. Hence it suffices to show that

$$\lim_{n\to\infty} \frac{1}{\sqrt[k]{n}} = 0,$$

for $k \in \mathbb{N}$. Take $N > (\frac{1}{\epsilon})^k$ by the Archimedean property of real numbers. Then for $n \geq N$ we have $0 < \frac{1}{\sqrt[k]{n}} < \epsilon$, since the root function is strictly increasing. ■

Theorem 3.32. *For $r > 0$ we have $\lim_{n\to\infty} \sqrt[n]{r} = 1$. Also for $n \in \mathbb{N}$ we have $\lim_{n\to\infty} \sqrt[n]{n} = 1$.*

Proof. When $r > 1$ we have $\sqrt[n]{r} > \sqrt[n]{1} = 1$. Let $b := \sqrt[n]{r} - 1 > 0$. By binomial theorem for $n > 1$ we have

$$r = (\sqrt[n]{r})^n = (1+b)^n > nb = n(\sqrt[n]{r} - 1).$$

Therefore $0 < \sqrt[n]{r} - 1 < \frac{r}{n}$. Hence we get the desired result by the squeeze theorem. When $0 < r < 1$, we have $\frac{1}{r} > 1$. Thus

$$\sqrt[n]{r} = \frac{1}{\sqrt[n]{\frac{1}{r}}} \underset{n\to\infty}{\longrightarrow} \frac{1}{1} = 1.$$

For the second limit let $c := \sqrt[n]{n} - 1 > 0$. Then by binomial theorem for $n > 1$ we have

$$n = (\sqrt[n]{n})^n = (1+c)^n > \frac{n(n-1)}{2}c^2 = \frac{n(n-1)}{2}(\sqrt[n]{n} - 1)^2.$$

Hence we get $0 < \sqrt[n]{n} - 1 < \sqrt{\frac{2}{n-1}}$. Thus we get the desired result by the squeeze theorem. ■

3.3 Lim sup and Lim inf

Definition 3.33. Suppose (a_n) is a sequence in $\mathbb{R}$. If (a_n) is bounded above we define

$$\limsup a_n := \lim_{n\to\infty} \sup_{k\geq n} a_k,$$

and if it is not bounded above we define $\limsup a_n := +\infty$. Similarly

$$\liminf a_n := \lim_{n\to\infty} \inf_{k\geq n} a_k,$$

when (a_n) is bounded below, otherwise $\liminf a_n := -\infty$.

Remark. Note that we require the boundedness in the definition, in order to make sure that the supremums and infimums are finite numbers. Also note that the sequences $\sup_{k\geq n} a_k$, and $\inf_{k\geq n} a_k$, are respectively decreasing, and increasing. Therefore by Theorems 3.28, and 3.29, the above limits exist, and we have

$$\limsup a_n = \inf_{n\geq 1} \sup_{k\geq n} a_k, \qquad \liminf a_n = \sup_{n\geq 1} \inf_{k\geq n} a_k.$$

Furthermore, it is obvious that we always have

$$\liminf a_n \leq \limsup a_n.$$

Theorem 3.34. *For a sequence (a_n) in $\mathbb{R}$, $\lim a_n = a$ if and only if*

$$\limsup a_n = a = \liminf a_n.$$

Here a can be $\pm\infty$ too.

Proof. Let $c_n = \sup_{k \geq n} a_k$ and $b_n = \inf_{k \geq n} a_k$. Then $b_n \leq a_n \leq c_n$ for all n. Therefore if $\limsup a_n = a = \liminf a_n$ then both $(b_n), (c_n)$ converge to a, and so does (a_n). Note that for infinite limits we only need one of the inequalities.

Now assume $\lim a_n = a$ and a is finite. Then for a given $\epsilon > 0$ there is $N \in \mathbb{N}$ such that for $n \geq N$ we have $a - \frac{\epsilon}{2} < a_n < a + \frac{\epsilon}{2}$. Therefore we have

$$a - \epsilon < a - \frac{\epsilon}{2} \leq b_n \leq c_n \leq a + \frac{\epsilon}{2} < a + \epsilon,$$

for $n \geq N$. Hence $b_n, c_n \to a$, as desired.

If $a = +\infty$, then $\limsup a_n$ is $+\infty$ by definition. For $\liminf a_n$ note that for large enough n we have $a_n > M + 1$, which implies $b_n \geq M + 1 > M$. Thus $\liminf a_n$ is also $+\infty$, since M was arbitrary. The case of $a = -\infty$ is similar. ■

Theorem 3.35. *Suppose (a_n) is a sequence in $\mathbb{R}$. Then it has subsequences converging to $\limsup a_n$, and to $\liminf a_n$. Furthermore, if a subsequence $a_{n_i} \to a$ we have*

$$\liminf a_n \leq a \leq \limsup a_n.$$

Proof. First suppose (a_n) is bounded above. Let us find a subsequence $a_{m_j} \to \limsup a_n$. Suppose we have chosen $a_{m_1}, \ldots, a_{m_k}$. Then we choose $a_{m_{k+1}}$ with $m_{k+1} > m_k$, so that

$$\big(\sup_{l > m_k} a_l\big) - \frac{1}{k+1} < a_{m_{k+1}} \leq \big(\sup_{l > m_k} a_l\big).$$

Because $(\sup_{l > m_k} a_l) - \frac{1}{k+1}$ is not an upper bound for the set $\{a_n : n > m_k\}$. Therefore by the squeeze theorem, (a_{m_j}) converges to $\limsup a_n$, since $\sup_{l > m_k} a_l$ converges to $\limsup a_n$ as $k \to \infty$. Note that when $\limsup a_n = -\infty$, we only need one of the inequalities.

Next suppose (a_n) is not bounded above, so that $\limsup a_n = +\infty$. Then for any $n \in \mathbb{N}$ there is a_{m_n} such that $a_{m_n} > n$. We can also choose a_{m_n} in a way that $m_{n+1} > m_n$. Because $\{a_k\}_{k > m_n}$ cannot be bounded above, since otherwise $\{a_k\}_{k \geq 1}$ would have been bounded above. Hence (a_{m_n}) is a subsequence of (a_n) that converges to $+\infty$, as desired. The case of $\liminf a_n$ is similar.

Now suppose $a_{n_i} \to a$. If one of the $\limsup a_n$ or $\liminf a_n$ is infinite, then the required inequality holds trivially. So suppose (a_n) is bounded. Then we have

$$\inf_{k \geq n_i} a_k \leq a_{n_i} \leq \sup_{k \geq n_i} a_k.$$

This implies the desired inequality, noting that the subsequences of a convergent sequence converge to the same limit as the original sequence. ■

Remark. Let (a_n) be a sequence in $\mathbb{R}$. We know that the set of all subsequential limits of (a_n), i.e. the limits of all subsequences of (a_n), is a closed set. By the above theorem the supremum of this closed set is $\limsup a_n$, and its infimum is $\liminf a_n$.

3.4 Series

Definition 3.36. Suppose (a_n) is a sequence in $\mathbb{R}^m$. The **series** $\sum_{n=1}^{\infty} a_n$ is the sequence of **partial sums** $S_k = \sum_{n=1}^{k} a_n$. If (S_k) converges, or diverges to infinity, we denote its limit by the same notation $\sum_{n=1}^{\infty} a_n$.

Remark. We may also denote a series or its limit by $\sum_{n\geq 1} a_n$.

Remark. Note that a series in $\mathbb{C}$ is just a series in $\mathbb{R}^2$. So, all of the following theorems also apply to series in $\mathbb{C}$, except those theorems that are only valid for series in $\mathbb{R}$.

Theorem 3.37. *Suppose* $\sum_{n=1}^{\infty} a_n$ *is a series in* $\mathbb{R}^m$*. If* $\sum_{n=1}^{\infty} a_n$ *converges, then we have* $\lim a_n = 0$.

Proof. Let S be the limit of the series, and let S_n be the nth partial sum of the series. Then $a_n = S_n - S_{n-1} \to S - S = 0$ as $n \to \infty$. ■

***Example* 3.38.** For $r \in \mathbb{R}$ with $|r| < 1$ we have

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}.$$

Since $\sum_{n=0}^{k} r^n = \frac{1-r^{k+1}}{1-r} \to \frac{1}{1-r}$. This series is called the *geometric series*.

***Example* 3.39.** The *harmonic series*

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots$$

diverges to infinity. To see this note that

$$\sum_{n=2^j+1}^{2^{j+1}} \frac{1}{n} \geq \sum_{n=2^j+1}^{2^{j+1}} \frac{1}{2^{j+1}} = \frac{2^j}{2^{j+1}} = \frac{1}{2} \implies \sum_{n=1}^{2^k} \frac{1}{n} \geq 1 + \frac{k}{2} \xrightarrow[k\to\infty]{} \infty.$$

Finally note that the sequence of partial sums of the harmonic series is increasing, since the terms of the series are positive. So the divergence of the above subsequence of partial sums implies the divergence of the whole sequence of partial sums.

***Exercise* 3.40.** Show that $\sum_{n=1}^{\infty} \frac{1}{n^p}$ is convergent if and only if $p > 1$.

Theorem 3.41. *A series of nonnegative real numbers is convergent if and only if its sequence of partial sums is bounded above.*

Proof. Just note that the sequence of partial sums is an increasing sequence in $\mathbb{R}$. ■

Theorem 3.42. *Suppose $\sum_{n=1}^{\infty} a_n, \sum_{n=1}^{\infty} b_n$ are convergent series in $\mathbb{R}^m$, and $c \in \mathbb{R}$. Then $\sum_{n=1}^{\infty}(a_n + cb_n)$ is also convergent and we have*

$$\sum_{n=1}^{\infty}(a_n + cb_n) = \sum_{n=1}^{\infty} a_n + c\sum_{n=1}^{\infty} b_n.$$

Proof. The partial sums satisfy $\sum_{n=1}^{k}(a_n + cb_n) = \sum_{n=1}^{k} a_n + c\sum_{n=1}^{k} b_n$. Now let $k \to \infty$. ■

Remark. The following theorem is the Cauchy criterion for the convergence of series.

Theorem 3.43. *Suppose $\sum_{n=1}^{\infty} a_n$ is a series in $\mathbb{R}^m$. Then $\sum_{n=1}^{\infty} a_n$ is convergent if and only if*

$$\forall \epsilon > 0 \, \exists N \in \mathbb{N} \textit{ such that } \forall m \geq n \geq N \textit{ we have } \left|\sum_{k=n}^{m} a_k\right| < \epsilon.$$

Proof. Let (S_n) be the sequence of partial sums. We know that the sequence of partial sums is convergent if and only if it is Cauchy. Now note that $S_m - S_n = \sum_{k=n}^{m} a_k$, when $m \geq n$. ■

Definition 3.44. Suppose $\sum_{n=1}^{\infty} a_n$ is a series in $\mathbb{R}^m$. We say $\sum_{n=1}^{\infty} a_n$ is **absolutely convergent** if $\sum_{n=1}^{\infty} |a_n|$ converges. A convergent series which is not absolutely convergent, is **conditionally convergent**.

Comparison Test. *Suppose (a_n) is a sequence in $\mathbb{R}^m$, and (b_n) is a sequence in $\mathbb{R}$. Also suppose that for some $n_0 \in \mathbb{N}$ we have $|a_n| \leq b_n$ for all $n \geq n_0$. Then if $\sum_{n=1}^{\infty} b_n$ converges, $\sum_{n=1}^{\infty} a_n$ converges too. Equivalently, if $\sum_{n=1}^{\infty} a_n$ diverges then $\sum_{n=1}^{\infty} b_n$ diverges too.*

Proof. For given $\epsilon > 0$ there is $N \geq n_0$ such that for $m \geq n \geq N$ we have

$$\left|\sum_{k=n}^{m} a_k\right| \leq \sum_{k=n}^{m} |a_k| \leq \sum_{k=n}^{m} b_k = \left|\sum_{k=n}^{m} b_k\right| < \epsilon.$$

Therefore $\sum_{n=1}^{\infty} a_n$ converges due to the Cauchy criterion for series. The last claim of the theorem is just the contrapositive of the first claim. ■

Theorem 3.45. *An absolutely convergent series is convergent.*

Proof. Suppose $\sum_{n=1}^{\infty} a_n$ is an absolutely convergent series in $\mathbb{R}^m$. Let $b_n := |a_n|$. Then the result follows by applying the comparison test, noting that $|a_n| \leq b_n$. ■

Limit Comparison Test. *Suppose $(a_n), (b_n)$ are sequences of positive real numbers, and for some $c \in (0, \infty)$ we have $\lim \frac{a_n}{b_n} = c$. Then $\sum_{n=1}^{\infty} a_n$ converges if and only if $\sum_{n=1}^{\infty} b_n$ converges.*

Proof. For sufficiently large n we have $0 < \frac{c}{2} < \frac{a_n}{b_n} < \frac{3c}{2}$. Thus $a_n < \frac{3c}{2} b_n$ and $b_n < \frac{2}{c} a_n$. Now we get the result by applying the comparison test twice. ■

Root Test. *Suppose $\sum_{n=1}^{\infty} a_n$ is a series in $\mathbb{R}^m$ and*

$$r := \limsup \sqrt[n]{|a_n|}.$$

(i) *If $r < 1$, the series converges absolutely.*
(ii) *If $1 < r \leq +\infty$, the series diverges.*

Proof. **(i)** Note that $r \geq 0$. Let $s \in (r, 1)$, so that $r < s < 1$. Then we have $\sup_{n \geq m} \sqrt[n]{|a_n|} < s$ for some large enough m, since the sequence of supremums converges to r. Therefore for $n \geq m$ we have $\sqrt[n]{|a_n|} < s$, and thus $|a_n| < s^n$. But $\sum_{n \geq 1} s^n$ is a convergent geometric series. Hence by applying the comparison test we get the desired result.

(ii) We claim that $\sqrt[n]{|a_n|} > 1$ for infinitely many n. Because otherwise for large n we would have $\sqrt[n]{|a_n|} \leq 1$, which implies that $\sup_{n \geq m} \sqrt[n]{|a_n|} \leq 1$ for all large values of m. Hence we must have $r \leq 1$, which is a contradiction. Thus for infinitely many n we have $\sqrt[n]{|a_n|} > 1$, and therefore for those n we have $|a_n| > 1$. Hence $a_n \not\to 0$. So the series diverges. ■

Ratio Test. *Suppose $\sum_{n=1}^{\infty} a_n$ is a series in $\mathbb{R}^m$ and $|a_n| \neq 0$ for all n. Let*

$$r := \limsup \frac{|a_{n+1}|}{|a_n|}, \qquad s := \liminf \frac{|a_{n+1}|}{|a_n|}.$$

(i) *If $r < 1$, the series converges absolutely.*
(ii) *If $1 < s \leq +\infty$, the series diverges.*

Proof. **(i)** Note that $r \geq s \geq 0$. Let $t \in (r, 1)$, so that $r < t < 1$. Then $\sup_{n \geq m} \frac{|a_{n+1}|}{|a_n|} < t$ for some large enough m, since the sequence of supremums converges to r. Therefore for $n \geq m$ we have $\frac{|a_{n+1}|}{|a_n|} < t$, or $|a_{n+1}| < t|a_n|$. Thus by an easy induction we obtain that

$$|a_n| < |a_m| t^{n-m} = \frac{|a_m|}{t^m} t^n,$$

for $n \geq m$. But the series

$$\sum_{n\geq 1} \frac{|a_m|}{t^m} t^n = \frac{|a_m|}{t^m} \sum_{n\geq 1} t^n$$

is convergent. Thus by applying the comparison test we get the desired result.

(ii) Let $t \in (s, 1)$, so that $s > t > 1$. Then $s \geq \inf_{n\geq m} \frac{|a_{n+1}|}{|a_n|} > t > 1$ for some large enough m, since the sequence of infimums converges to s. Therefore $\frac{|a_{n+1}|}{|a_n|} > t$ for $n \geq m$. Hence as before we have $|a_n| > t^{n-m}|a_m|$, for $n \geq m$. But we have

$$t^{n-m}|a_m| \underset{n\to\infty}{\longrightarrow} \infty.$$

Therefore $a_n \not\to 0$. So the series diverges. ■

***Example* 3.46.** The ratio test and the root test cannot determine the convergence or divergence of the series $\sum_{n=1}^{\infty} \frac{1}{n^p}$ for $p > 0$. Because we have

$$\frac{\frac{1}{(n+1)^p}}{\frac{1}{n^p}} = (\frac{n}{n+1})^p \underset{n\to\infty}{\longrightarrow} 1^p = 1,$$
$$\sqrt[n]{\frac{1}{n^p}} = \frac{1}{(\sqrt[n]{n})^p} \underset{n\to\infty}{\longrightarrow} \frac{1}{1^p} = 1.$$

***Exercise* 3.47.** Give an example of a series for which the ratio test is inconclusive, but the root test implies its convergence. Can you find a series for which the root test is inconclusive, but the ratio test is not?

Alternating Series Test. *Suppose (a_n) is a decreasing sequence of positive real numbers that converges to 0. Then the series $\sum_{n=1}^{\infty}(-1)^n a_n$ converges.*

Proof. Let $S_k := \sum_{n=1}^{k}(-1)^n a_n$. Then for $k \geq 1$ we have

$$\begin{aligned}
&S_{2k+2} - S_{2k} = a_{2k+2} - a_{2k+1} \leq 0,\\
&S_{2k+1} - S_{2k-1} = -a_{2k+1} + a_{2k} \geq 0,\\
&S_{2k} - S_1 = a_{2k} + (-a_{2k-1} + a_{2k-2}) + \cdots + (-a_3 + a_2) \geq 0,\\
&S_{2k-1} = -a_{2k-1} + (a_{2k-2} - a_{2k-3}) + \cdots + (a_2 - a_1) \leq 0.
\end{aligned}$$

Thus (S_{2k}) is a decreasing sequence bounded below by S_1. Hence it is convergent. Similarly, (S_{2k-1}) is an increasing sequence bounded above by 0. Hence it is convergent too. But we have $S_{2k} - S_{2k+1} = a_{2k} \to 0$. Therefore the limits of (S_{2k}) and (S_{2k-1}) are the same. This implies that (S_k) is convergent, as desired. ■

***Example* 3.48.** The series $\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$ converges conditionally by the above theorem.

Definition 3.49. The series $\sum_{n=1}^{\infty} b_n$ is called a **rearrangement** of the series $\sum_{n=1}^{\infty} a_n$, if there exists a bijective map $\sigma : \mathbb{N} \to \mathbb{N}$ such that $b_n = a_{\sigma(n)}$.

Theorem 3.50. *Suppose $\sum_{n=1}^{\infty} a_n$ is an absolutely convergent series in $\mathbb{R}$. Let $\sum_{n=1}^{\infty} b_n$ be a rearrangement of $\sum_{n=1}^{\infty} a_n$. Then $\sum_{n=1}^{\infty} b_n$ is also absolutely convergent, and we have*

$$\sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty} a_n.$$

Proof. Note that $\sum_{n=1}^{\infty} a_n$ is also a rearrangement of $\sum_{n=1}^{\infty} b_n$. Because if $b_n = a_{\sigma(n)}$ for some bijective map σ, then $a_n = b_{\sigma^{-1}(n)}$. First, let us show that $\sum_{n=1}^{\infty} b_n$ is absolutely convergent. We have

$$\sum_{n=1}^{k} |b_n| = \sum_{n=1}^{k} |a_{\sigma(n)}| \leq \sum_{n=1}^{m} |a_n| \leq \sum_{n=1}^{\infty} |a_n| < +\infty,$$

where $m = \max_{n \leq k} \sigma(n)$. Therefore the sequence of partial sums of $\sum_{n=1}^{\infty} |b_n|$ is bounded. Hence it is convergent, and we have

$$\sum_{n=1}^{\infty} |b_n| \leq \sum_{n=1}^{\infty} |a_n|.$$

Since $\sum_{n=1}^{\infty} a_n$ is also a rearrangement of $\sum_{n=1}^{\infty} b_n$, we can switch the roles of $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$, and repeat the above argument, to get

$$\sum_{n=1}^{\infty} |b_n| = \sum_{n=1}^{\infty} |a_n|.$$

Next we show that the two series have the same value. Note that

$$\sum_{n=1}^{\infty} \big||a_n| + a_n\big| \leq 2 \sum_{n=1}^{\infty} |a_n| < +\infty,$$

so $\sum_{n=1}^{\infty}(|a_n| + a_n)$ is absolutely convergent. Also note that $\sum_{n=1}^{\infty}(|b_n| + b_n)$ is a rearrangement of the series $\sum_{n=1}^{\infty}(|a_n| + a_n)$ via the map σ. Thus we have

$$\sum_{n=1}^{\infty}(|b_n| + b_n) = \sum_{n=1}^{\infty} \big||b_n| + b_n\big| = \sum_{n=1}^{\infty} \big||a_n| + a_n\big| = \sum_{n=1}^{\infty}(|a_n| + a_n),$$

since $|r| + r \geq 0$ for any real number r. Hence we have

$$\begin{aligned}\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty}(|a_n| + a_n - |a_n|) &= \sum_{n=1}^{\infty}(|a_n| + a_n) - \sum_{n=1}^{\infty} |a_n| \\ &= \sum_{n=1}^{\infty}(|b_n| + b_n) - \sum_{n=1}^{\infty} |b_n| = \sum_{n=1}^{\infty} b_n.\end{aligned}$$

■

Chapter 4

Differentiation

4.1 Rules of Differentiation

Definition 4.1. Suppose $I \subset \mathbb{R}$ is open and $f : I \to \mathbb{R}$. Then we say f is **differentiable** at a point $x \in I$ if

$$f'(x) := \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}$$

exists. $f'(x)$ is called the **derivative** of f at x.

Remark. This is equivalent to the existence of $a \in \mathbb{R}$ such that for small h we have

$$f(x+h) = f(x) + ah + R(h), \qquad \text{where} \quad \lim_{h\to 0} \frac{R(h)}{|h|} = 0.$$

Theorem 4.2. *The derivative of a constant function exists and equals zero everywhere. For $n \in \mathbb{N}$, the functions $f(x) = x^n$ from $\mathbb{R}$ to $\mathbb{R}$ are differentiable with $f'(x) = nx^{n-1}$. Also $g(x) = \frac{1}{x}$ from $\mathbb{R} - \{0\}$ to $\mathbb{R}$ is differentiable and $g'(x) = \frac{-1}{x^2}$.*

Proof. We have

$$\begin{aligned}&\frac{1}{h}[(x+h)^n - x^n] \\ &\qquad = \frac{1}{h}(x+h-x)[(x+h)^{n-1} + (x+h)^{n-2}x + \cdots + x^{n-1}] \xrightarrow[h\to 0]{} nx^{n-1}.\end{aligned}$$

Also

$$\frac{1}{h}\left(\frac{1}{x+h} - \frac{1}{x}\right) = \frac{-1}{x(x+h)} \xrightarrow[h\to 0]{} \frac{-1}{x^2}.$$

■

Theorem 4.3. *Suppose f is differentiable at x, then it is continuous at x too.*

Proof. $\lim_{y\to x}[f(y)-f(x)] = \lim_{y\to x}[\frac{f(y)-f(x)}{y-x}(y-x)] = f'(x)\times 0 = 0.$ ■

Definition 4.4. We denote by f' or $f^{(1)}$ the function whose value is $f'(x)$, and is defined on the set where f is differentiable. We also define the ***k*th derivative** inductively as

$$f^{(k)} := (f^{(k-1)})',$$

when it exists. We also set the zeroth derivative of f to be $f^{(0)} := f$.

We say f is of **class C^k** if $f^{(k)}$ exists and is continuous on the domain of f. A function is called **infinitely differentiable** or **smooth** or of **class C^∞** if it has derivatives of all orders at every point of its domain. Finally, note that a function is of **class C^0** if it is continuous on its domain.

Remark. It is obvious that C^{k+1} functions are also C^k, since differentiability implies continuity. Thus we have $C^0 \supset C^1 \supset \cdots \supset C^\infty$.

Remark. Note that the $(k+1)$th derivative of a function f equals the kth derivative of its derivative f'. This can be proved by an easy induction on k. Consequently, a function f is C^{k+1} if and only if f' is C^k.

***Example* 4.5.** The function $f(x) = x^n|x|$ is of class C^n but not of class C^{n+1} (why?).

Theorem 4.6. *Suppose $I \subset \mathbb{R}$ is an open set containing a point x, and $f, g : I \to \mathbb{R}$ are differentiable at x. Let $c_1, c_2 \in \mathbb{R}$. Then $c_1 f + c_2 g$ is differentiable at x, and we have*

$$(c_1 f + c_2 g)'(x) = c_1 f'(x) + c_2 g'(x).$$

Proof. We have

$$\lim_{h\to 0}\frac{(c_1 f + c_2 g)(x+h) - (c_1 f + c_2 g)(x)}{h} = c_1 \lim_{h\to 0}\frac{f(x+h)-f(x)}{h} + c_2 \lim_{h\to 0}\frac{g(x+h)-g(x)}{h}.$$ ■

***Example* 4.7.** The derivative of a polynomial $p(x) = a_0 + a_1 x + \cdots + a_n x^n$ is the polynomial

$$p'(x) = a_1 + 2a_2 x + \cdots + n a_n x^{n-1}.$$

Note that if $a_n \neq 0$ then $na_n \neq 0$, so

$$\deg p' = \deg p - 1.$$

Thus $p^{(n)}$ is a polynomial of degree 0, i.e. it is a constant polynomial; and therefore $p^{(m)} \equiv 0$ for all $m \geq n+1$. Hence in particular, we see that polynomials are C^∞ functions.

***Exercise* 4.8.** Show that for a polynomial $p(x) = a_0 + a_1 x + \cdots + a_n x^n$ we have

$$a_k = \frac{p^{(k)}(0)}{k!}.$$

This gives another proof that the coefficients of a polynomial (and hence its degree) are uniquely determined by the polynomial.

Leibniz Rule. *Suppose $I \subset \mathbb{R}$ is an open set containing a point x, and $f, g : I \to \mathbb{R}$ are differentiable at x. Then fg is differentiable at x, and we have*

$$(fg)'(x) = f'(x)g(x) + f(x)g'(x).$$

Proof. We have

$$\begin{aligned}
&\frac{1}{h}[(fg)(x+h) - (fg)(x)] \\
&\qquad = \frac{1}{h}[f(x+h)g(x+h) - f(x+h)g(x) + f(x+h)g(x) - f(x)g(x)] \\
&\qquad = f(x+h)\frac{1}{h}[g(x+h) - g(x)] + \frac{1}{h}[f(x+h) - f(x)]g(x) \\
&\qquad \xrightarrow[h \to 0]{} f(x)g'(x) + f'(x)g(x).
\end{aligned}$$

■

Chain Rule. *Suppose $I \subset \mathbb{R}$ is an open set containing a point a, and $f : I \to \mathbb{R}$ are differentiable at a. Also suppose J is a neighborhood of $f(a)$, and $g : J \to \mathbb{R}$ is differentiable at $f(a)$. Then $g \circ f$ is differentiable at a, and we have*

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

Proof. We have

$$\frac{g(f(x)) - g(f(a))}{x - a} = \frac{f(x) - f(a)}{x - a} \times \begin{cases} \frac{g(f(x)) - g(f(a))}{f(x) - f(a)} & f(x) \neq f(a) \\ g'(f(a)) & f(x) = f(a) \end{cases}.$$

Let $h(x)$ be the function on the right hand side above which is defined by two formulas. We only need to show that

$$\lim_{x \to a} h(x) = g'(f(a)).$$

Suppose $\epsilon > 0$ is given. Then there is $\delta > 0$ so that $0 < |y - f(a)| < \delta$ implies that

$$\left| \frac{g(y) - g(f(a))}{y - f(a)} - g'(f(a)) \right| < \epsilon.$$

Let $\tilde{\delta}$ be small enough so that $|x - a| < \tilde{\delta}$ implies that $|f(x) - f(a)| < \delta$. Then for $|x - a| < \tilde{\delta}$ we will either have $f(x) = f(a)$ in which case $h(x) = g'(f(a))$, or we have $0 < |f(x) - f(a)| < \delta$ which implies $|h(x) - g'(f(a))| < \epsilon$. Hence in both cases we have $|h(x) - g'(f(a))| < \epsilon$ as desired. ■

Theorem 4.9. *Suppose $I \subset \mathbb{R}$ is an open set containing a point x, and $f, g : I \to \mathbb{R}$ are differentiable at x. If $g \neq 0$ on I, then $\frac{f}{g}$ is differentiable at x, and we have*

$$\Big(\frac{f}{g}\Big)'(x) = \frac{g(x)f'(x) - g'(x)f(x)}{(g(x))^2}.$$

Proof. For $x \neq 0$ set $\iota(x) = \frac{1}{x}$. Then if we apply the chain rule to obtain the derivative of the function $\iota(g) = \frac{1}{g}$, we get

$$\Big(\frac{1}{g}\Big)' = (\iota(g))' = g'\iota'(g) = \frac{-g'}{g^2}.$$

Then we can compute the derivative of $\frac{f}{g}$ by using the Leibniz rule. ■

***Exercise* 4.10.** Suppose $f_1, \dots, f_m$ and $g_1, \dots, g_n$ are nonzero differentiable functions. Show that

$$\Big(\frac{f_1 \cdots f_n}{g_1 \cdots g_m}\Big)' = \frac{f_1 \cdots f_n}{g_1 \cdots g_m}\Big(\frac{f_1'}{f_1} + \cdots + \frac{f_n'}{f_n} - \frac{g_1'}{g_1} - \cdots - \frac{g_m'}{g_m}\Big).$$

Theorem 4.11. *Suppose $I \subset \mathbb{R}$ is open, and $f, g : I \to \mathbb{R}$ are C^k functions for some $1 \leq k \leq \infty$. Let J be an open set containing $f(I)$, and suppose $F : J \to \mathbb{R}$ is a C^k function. Also let $c_1, c_2 \in \mathbb{R}$. Then*

(i) *$c_1 f + c_2 g$, fg, and $F \circ f$ are C^k functions.*

(ii) *If $g \neq 0$ on I, then $\frac{f}{g}$ is a C^k function.*

Proof. In the following whenever we divide by g, we assume that it is nonzero. We know that

$$\begin{aligned} (c_1 f + c_2 g)' &= c_1 f' + c_2 g', \\ (fg)' &= f'g + fg', \\ (F \circ f)' &= (F' \circ f)f', \\ (f/g)' &= (f'g - fg')/g^2. \end{aligned} \tag{$*$}$$

For $1 \leq k < \infty$, the proof is by induction on k. When $k = 1$ we know that f, g, F and their derivatives, are continuous. Therefore $c_1 f + c_2 g$, fg, $F \circ f$, and f/g have continuous derivatives by $(*)$. Because the sum, the product, the inverse (when the function is nonzero), and the composition of continuous functions are continuous. Note that constant functions are continuous too.

Now suppose the theorem is true for some $k < \infty$. Then we have to prove the theorem for $k + 1$. Let f, g, F be C^{k+1} functions. Then we know that f, g, F and their derivatives are C^k functions. Hence by the induction hypothesis we know that the derivatives of $c_1 f + c_2 g$, fg, $F \circ f$, and f/g are C^k functions. Because by $(*)$, their derivatives are the linear combination of one or two terms, and those terms are C^k functions. (Note that those terms themselves are either C^k functions, or they are the product of two C^k functions, or they are the product of a C^k function and a function which is the composition of two C^k functions, or they are the product of two C^k functions divided by the product of two nonzero C^k functions.) Therefore $c_1 f + c_2 g$, fg, $F \circ f$, and f/g are C^{k+1} functions.

Finally if the functions f, g, F are C^∞ functions, then they are C^k functions for all $k < \infty$. Therefore $c_1 f + c_2 g$, fg, $F \circ f$, and f/g are C^k functions for all $k < \infty$. Hence they are also C^∞ functions. ■

Theorem 4.12. *Suppose I, J are open intervals, and $f : I \to J$ is an invertible function. Also suppose f has a nonzero derivative at $a \in I$, and f^{-1} is continuous at $f(a)$. Then f^{-1} is differentiable at $f(a)$, and we have*

$$(f^{-1})'(f(a)) = \frac{1}{f'(a)}.$$

Proof. Let $b = f(a)$, and $x = f^{-1}(b + h)$, where h is small enough so that $b + h \in J$. Then we have

$$\frac{f^{-1}(b+h) - f^{-1}(b)}{h} = \frac{f^{-1}(b+h) - a}{b + h - b} = \frac{x - a}{f(x) - f(a)} = \frac{1}{\dfrac{f(x) - f(a)}{x - a}}.$$

When h is small, $x = f^{-1}(b+h)$ is close to $f^{-1}(b) = a$, since f^{-1} is continuous at $b = f(a)$. Hence the above fraction is close to $\frac{1}{f'(a)}$, and therefore f^{-1} is differentiable at b with the desired derivative. ■

***Exercise* 4.13.** Show that for $p \in \mathbb{Q}$ the function $x \mapsto x^p$ is differentiable on $(0, \infty)$, and we have $(x^p)' = px^{p-1}$.

***Exercise* 4.14.** Consider the function $f : (-1, 1) \to (-1, 1)$ defined by

$$f(x) = \begin{cases} \frac{1}{\log(2n+1-2n_0)} & x = \frac{1}{\log n},\ n \geq 2n_0, \\ \frac{1}{\log(2n)} & x = 1 - \frac{1}{2n},\ n \geq n_0, \\ 1 - \frac{1}{n} & x = 1 - \frac{1}{2n+1-2n_0},\ n \geq 2n_0, \\ x & \text{otherwise}, \end{cases}$$

where n_0 is a large enough constant. Show that f is an invertible function, and $f'(0) = 1$, but f^{-1} is not continuous at $f(0)$. Hence f^{-1} cannot be differentiable at $f(0)$ either.

4.2 Extrema of Single Variable Functions

Proposition 4.15. *Suppose f is a function into $\mathbb{R}$ defined on a neighborhood of $a \in \mathbb{R}$. If $f'(a) > 0$ then*

$$f(a-h) < f(a) < f(a+h)$$

for small $h > 0$. When $f'(a) < 0$ the inequalities are reversed.

Proof. Suppose $f'(a) > 0$, the other case is similar. For small $h > 0$, we know that $\frac{1}{\pm h}\big(f(a \pm h) - f(a)\big)$ is close to $f'(a)$. Therefore this fraction is also positive. Hence its numerator and denominator must have the same sign. Thus we get the desired inequalities. ■

Remark. Note that unlike the case of functions whose derivative has a sign on an interval (which we consider below), in the above proposition we are not claiming that f is monotone. The above proposition only says that if $f'(a)$ is nonzero, then we can compare the values of f around a with $f(a)$. As an exercise you can show that the function

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

satisfies $f'(0) = 1$, but it is not monotone on any interval containing 0.

Definition 4.16. Suppose X is a metric space, and $f : X \to \mathbb{R}$ is a function. We say f has a **local maximum** at $y \in X$ if $f(y) \geq f(x)$ for all x in a neighborhood of y. Similarly, we say f has a **local minimum** at $y \in X$ if $f(y) \leq f(x)$ for all x in a neighborhood of y. A **local extremum** of f, is either a local maximum of f, or a local minimum of f.

Theorem 4.17. *Suppose f is a function into $\mathbb{R}$ defined on a neighborhood of $a \in \mathbb{R}$, and it is differentiable at a. If f has a local maximum or minimum at a, then $f'(a) = 0$.*

Proof. Suppose to the contrary that $f'(a) \neq 0$. Let us assume that $f'(a) < 0$, the other case is similar. Then by the last proposition, for all small $h > 0$ we have

$$f(a-h) > f(a) > f(a+h).$$

Thus f cannot have a local maximum, nor a local minimum, at a. ■

Mean Value Theorem. *Suppose $f : [a, b] \to \mathbb{R}$ is continuous, and it is differentiable on (a, b). Then there is $c \in (a, b)$ such that*

$$f(b) - f(a) = f'(c)(b-a).$$

Remark. The special case of the mean value theorem in which $f(b) = f(a)$, and we conclude the existence of c such that $f'(c) = 0$, is known as the **Rolle's theorem**. Note that the general case actually follows from this special case, as we have proved below.

Proof. Let

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

Then $g(a) = g(b) = 0$, g is continuous on $[a, b]$, and it is differentiable on (a, b). If $g = 0$ everywhere then $g' = 0$ everywhere. Otherwise, as g is continuous on a compact set, either its maximum is positive or its minimum is negative. Hence g has a positive maximum or a negative minimum in (a, b), i.e. it has a local extremum. In any case there is $c \in (a, b)$ such that $g'(c) = 0$. Therefore $f'(c) - \frac{f(b)-f(a)}{b-a} = 0$. ■

Theorem 4.18. *Suppose f is differentiable on an open interval with $|f'| \leq M$. Then for any a, b in the interval we have*

$$|f(b) - f(a)| \leq M|b - a|.$$

In particular, if $f' = 0$ on the interval, f is constant.

Proof. By the mean value theorem, for some c between a, b we have

$$|f(b) - f(a)| = |f'(c)(b-a)| \leq M|b-a|.$$

Now if $f' \equiv 0$ then $M = 0$. Hence for any a, b we have $f(b) = f(a)$. ■

Theorem 4.19. *If $f : (a, b) \to \mathbb{R}$ has positive derivative on (a, b), then f is strictly increasing. And if f has negative derivative on (a, b), then f is strictly decreasing.*

Proof. Suppose $f' > 0$. If $x > y$, then there is some c between x, y such that $f(x) - f(y) = f'(c)(x - y) > 0$. The other case is similar. ■

Theorem 4.20. *Suppose $f : (a, b) \to \mathbb{R}$ is continuous, and $c \in (a, b)$.*
(i) *If $f' > 0$ on (a, c) and $f' < 0$ on (c, b), then $f(c)$ is the maximum of f.*
(ii) *If $f' < 0$ on (a, c) and $f' > 0$ on (c, b), then $f(c)$ is the minimum of f.*

Proof. Suppose $f' < 0$ on (c, b). If $x \in (c, b)$, then there is $y \in (c, x)$ such that $f(x) - f(c) = f'(y)(x - c) < 0$. The other cases are similar. ■

Theorem 4.21. *Suppose f is a differentiable function from a neighborhood of $c \in \mathbb{R}$ into $\mathbb{R}$, and $f'(c) = 0$.*
(i) *If $f''(c) > 0$ then f has a local minimum at c.*
(ii) *If $f''(c) < 0$ then f has a local maximum at c.*

Proof. If $f''(c) > 0$, then there is $l > 0$ such that $f'(c-h) < f'(c) = 0 < f'(c+h)$ for all $0 < h < l$. So the hypothesis of the previous theorem is satisfied on the interval $(c-l, c+l)$. The other case is similar. ■

Darboux's Theorem. *Suppose f is differentiable on an interval. Then its derivative has the intermediate value property.*

Proof. Suppose $a < b$, and $f'(a) < r < f'(b)$. Then for small enough $h > 0$ we have

$$\frac{f(a+h)-f(a)}{h} < r, \qquad \frac{f(b-h)-f(b)}{-h} > r.$$

Consider the function $\phi(x) := \frac{f(x+h)-f(x)}{h}$. Now ϕ is continuous, since f is continuous. Hence there is $c \in (a, b-h)$ such that $\phi(c) = r$. This means that

$$\frac{f(c+h)-f(c)}{c+h-c} = r.$$

Therefore, by the mean value theorem there is $\tilde{c} \in (c, c+h)$ such that $f'(\tilde{c}) = r$. ■

Second Proof. Let $g(x) := f(x) - rx$. Then $g'(a) < 0$ and $g'(b) > 0$. Now g is continuous on $[a, b]$, and attains its minimum there. But for small positive ϵ we have $g(a) > g(a+\epsilon)$ and $g(b-\epsilon) < g(b)$. Therefore g attains its minimum at a point $c \in (a, b)$. Thus $g'(c) = f'(c) - r = 0$. ■

Remark. By the above theorem, the derivative of a differentiable function cannot have jump discontinuities. But the derivative can be discontinuous. For example, let

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

Then $f'(x) = \begin{cases} 2x\sin\frac{1}{x} - \cos\frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}$ is not continuous at $x = 0$.

Inverse Function Theorem in Dimension One. *Suppose $f : (a, b) \to \mathbb{R}$ is a differentiable function, and its derivative is nonzero everywhere. Then f is invertible, and f^{-1} is differentiable. If in addition f is C^k for some $1 \leq k \leq \infty$, then f^{-1} is also C^k.*

Proof. Note that f is continuous, since it is differentiable. Also note that f' is either positive everywhere, or negative everywhere; since otherwise f' would have vanished at some point by Darboux's theorem. Hence f is strictly monotone; thus it is one-to-one. Let $I = f((a, b))$ be the image of f. Then I is an interval by the intermediate value theorem. Now let $y \in I$. We know that $y = f(x)$ for some $x \in (a, b)$. Suppose ϵ is small enough so that $x \pm \epsilon \in (a, b)$. Then by the intermediate value theorem and strict monotonicity of f, the set $f((x-\epsilon, x+\epsilon))$ is an open interval with endpoints $f(x \pm \epsilon)$, which contains y. Hence I is an open interval.

Consider $f^{-1} : I \to (a,b)$. We want to show that f^{-1} is continuous at y. So for a given $\epsilon > 0$ we want to find $\delta > 0$ such that if $|w-y| < \delta$ then $|f^{-1}(w)-f^{-1}(y)| < \epsilon$. Let

$$\delta < |y - f(x \pm \epsilon)|.$$

Then if $|w-y| < \delta$ we have $w \in f((x-\epsilon, x+\epsilon))$, since we have seen that the set $f((x-\epsilon, x+\epsilon))$ is an open interval with endpoints $f(x \pm \epsilon)$. Hence there is a unique $z \in (x-\epsilon, x+\epsilon)$ such that $f(z) = w$. Therefore we have

$$|f^{-1}(w) - f^{-1}(y)| = |z - x| < \epsilon,$$

as desired. Thus f^{-1} is continuous at y. Therefore by Theorem 4.12, f^{-1} is differentiable at y, and we have

$$(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))}.$$

So f^{-1} is differentiable on I.

The proof of the last statement of the theorem is by induction on k, when $k < \infty$. If f is C^1, then as f', f^{-1} are continuous and $f' \neq 0$, $(f^{-1})'$ must be continuous too. Thus f^{-1} is also C^1. Now suppose the claim is true for some k. Then if f is C^{k+1}, it is also C^k. Hence f^{-1} is C^k by the induction hypothesis. Then $f' \circ f^{-1}$ is C^k too, since f' is also C^k. Thus $(f^{-1})'$ is also C^k, since $f' \circ f^{-1} \neq 0$. Therefore f^{-1} is C^{k+1}. Finally if f is C^∞, then it is C^k for all $k < \infty$. Therefore f^{-1} is also C^k for all $k < \infty$. Hence f^{-1} is C^∞ too. ■

Remark. In the above theorem, it is essential to assume that $f'(x) \neq 0$ for all x. For example $f(x) = x^3$ is a C^∞ and invertible function on $\mathbb{R}$, but its inverse $f^{-1}(x) = \sqrt[3]{x}$ is not differentiable at $x = 0$.

Remark. As we have seen in the above proof, the strict monotonicity and continuity of f on an interval are sufficient to imply that f is invertible, and f^{-1} is continuous. We have also seen in Exercise 2.84 that continuous one-to-one maps from an interval into $\mathbb{R}$ are strictly monotone. Thus a continuous and one-to-one function $f : (a,b) \to \mathbb{R}$ is invertible, and f^{-1} is continuous (on the image of f). However, this is not true if the codomain of f is not $\mathbb{R}$. For example, as we have shown in Example 2.67, the function

$$\theta \mapsto (\cos\theta, \sin\theta),$$

from $[0, 2\pi)$ to the unit circle S^1, is a continuous bijection whose inverse is not continuous. We can also consider this as a function from the interval $[0, 2\pi)$ into $\mathbb{R}^2$. As another example, you can show that the function

$$\theta \mapsto (|\cos\theta|\sin 2\theta, \sin\theta \sin 2\theta),$$

from $(0, \pi)$ into $\mathbb{R}^2$, is a continuous one-to-one function whose inverse is not continuous. (Hint: consider sequences approaching $0, \pi$.)

4.3 L'Hôpital's Rules

Theorem 4.22. *Suppose $f, g : [a, b] \to \mathbb{R}$ are continuous, and they are differentiable on (a, b). Then there is $c \in (a, b)$ such that*

$$(f(b) - f(a))g'(c) = f'(c)(g(b) - g(a)).$$

Proof. The result follows by applying the mean value theorem to the function

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

Just note that this function satisfies the conditions of the mean value theorem. ■

Remark. The point of the above theorem is that f', g' are evaluated at the same point.

L'Hôpital's Rule. *Suppose f, g are real-valued functions defined on an open interval I; and I either contains the point a, or has a as an endpoint, where a can be $\pm\infty$ too. Suppose g, g' are nonzero on $I - \{a\}$. Also suppose*

$$\lim_{x\to a} f(x) = \lim_{x\to a} g(x) = 0, \qquad \lim_{x\to a} \frac{f'(x)}{g'(x)} = b,$$

where b can be ∞ or $\pm\infty$ too. Then we have

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

Proof. Suppose $\epsilon > 0$ is given. Let $x \in I$, and let t be between x, a. Note that by the mean value theorem we have $g(x) \neq g(t)$, since $g' \neq 0$. We know that $f(t), g(t) \to 0$ as $t \to a$. So for each given x we can find t such that

$$\left| \frac{f(x)}{g(x)} - \frac{f(x) - f(t)}{g(x) - g(t)} \right| < \epsilon.$$

Note that here we have used the continuity of the division. Therefore by the previous theorem, for some z between x, t we have

$$\left| \frac{f(x)}{g(x)} - \frac{f'(z)}{g'(z)} \right| < \epsilon.$$

Note that z is also between x, a. First suppose b is finite. Now when x is close enough to a, then z is also close enough to a, so that

$$\left| \frac{f'(z)}{g'(z)} - b \right| < \epsilon, \text{ and therefore } \left| \frac{f(x)}{g(x)} - b \right| < 2\epsilon.$$

Thus the limit of $\frac{f}{g}$ is b. Next suppose $b = +\infty$. Let $M > 0$ be given. Then when x is close enough to a, z is also close enough to a, so that

$$\frac{f'(z)}{g'(z)} > M, \text{ and therefore } \frac{f(x)}{g(x)} > M - \epsilon.$$

Hence the limit of $\frac{f}{g}$ is $+\infty$ too, since M is arbitrary. The cases of $b = -\infty$ and $b = \infty$ are similar. ■

Second Proof. When a is finite, we can redefine f, g to be zero at a. Then f, g become continuous at a. Now for some z between x, a we have

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

When x is close to a, z is also close to a, and the result follows as before. ■

L'Hôpital's Rule. *Suppose f, g are real-valued functions defined on an open interval I; and I either contains the point a, or has a as an endpoint, where a can be $\pm\infty$ too. Suppose g, g' are nonzero on $I - \{a\}$. Also suppose*

$$\lim_{x \to a} g(x) = \infty, \qquad \lim_{x \to a} \frac{f'(x)}{g'(x)} = b,$$

where b can be ∞ or $\pm\infty$ too. Then we have

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

Proof. Let $t \in I$, and let x be between t, a. Note that by the mean value theorem we have $g(x) \neq g(t)$, since $g' \neq 0$. Then for some z between t, x we have

$$\frac{f'(z)}{g'(z)} = \frac{f(x) - f(t)}{g(x) - g(t)} = \frac{\frac{f(x)}{g(x)} - \frac{f(t)}{g(x)}}{1 - \frac{g(t)}{g(x)}}.$$

Therefore

$$\begin{aligned}\frac{f(x)}{g(x)} &= \frac{f'(z)}{g'(z)}\Big(1-\frac{g(t)}{g(x)}\Big)+\frac{f(t)}{g(x)} \\ &= \frac{f'(z)}{g'(z)}-\frac{f'(z)}{g'(z)}\frac{g(t)}{g(x)}+\frac{f(t)}{g(x)}. \qquad (*)\end{aligned}$$

First suppose b is finite. Suppose $\epsilon > 0$ is given. When t is close enough to a, then z is also close enough to a, so that

$$\left|\frac{f'(z)}{g'(z)}-b\right| < \epsilon.$$

Now fix t, and let x be close enough to a so that

$$\left|\frac{g(t)}{g(x)}\right| < \epsilon, \qquad \left|\frac{f(t)}{g(x)}\right| < \epsilon.$$

Then we have

$$\left|\frac{f'(z)}{g'(z)}\frac{g(t)}{g(x)}\right| < \epsilon(|b|+\epsilon).$$

Note that this bound is independent of z. Finally we have

$$\left|\frac{f(x)}{g(x)}-b\right| < 2\epsilon+\epsilon(|b|+\epsilon),$$

which can be made as small as we want.

Next suppose $b=\infty$. Let x be close enough to a so that

$$1-\frac{g(t)}{g(x)} > \frac{1}{2}, \qquad \left|\frac{f(t)}{g(x)}\right| < 1.$$

Then from $(*)$ we get

$$\left|\frac{f(x)}{g(x)}\right| > \frac{1}{2}\left|\frac{f'(z)}{g'(z)}\right|-1.$$

Let $M>0$ be given. Then when x is close enough to a, z is also close enough to a, so that

$$\left|\frac{f'(z)}{g'(z)}\right| > M, \text{ and therefore } \left|\frac{f(x)}{g(x)}\right| > \frac{1}{2}M-1.$$

Hence the limit of $\frac{f}{g}$ is ∞. Finally note that when $\left|\frac{f'(z)}{g'(z)}\right|$ is large, $\frac{f(x)}{g(x)}$ and $\frac{f'(z)}{g'(z)}$ have the same sign. Hence the case of $b=\pm\infty$ follows too. ■

4.4 Taylor Polynomials

Definition 4.23. Suppose f is a real-valued function on an open interval I and it is n times differentiable at $a \in I$. The nth order **Taylor polynomial** of f at a is

$$P(h) := \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} h^k,$$

and the nth order **Taylor remainder** is $R(h) := f(a+h) - P(h)$.

Remark. Note that P, R are n times differentiable at 0, and $P^{(k)}(0) = f^{(k)}(a)$, $R^{(k)}(0) = 0$ for $k \leq n$.

Theorem 4.24. *Suppose f is a real-valued function on an open interval I, and it is n times differentiable at $a \in I$. Let R be the nth order Taylor remainder of f at a, then*

$$\lim_{h \to 0} \frac{R(h)}{h^n} = 0.$$

*If in addition f has $n+1$ derivatives around a, then we have the **Lagrange form** for the remainder, i.e.*

$$R(h) = \frac{f^{n+1}(\theta)}{(n+1)!} h^{n+1},$$

for some θ between a and $a+h$.

Proof. Remember that R is n times differentiable at $h = 0$, and $R^{(i)}(0) = 0$ for $0 \leq i \leq n$. By the mean value theorem we have

$$\begin{aligned} R(h) = R(h) - R(0) &= (h-0)R'(h_1) \\ &= hh_1R''(h_2) = \cdots = hh_1 \cdots h_{n-2}R^{(n-1)}(h_{n-1}), \end{aligned}$$

for some $0 < |h_{n-1}| < \cdots < |h_1| < |h|$. Therefore

$$\left| \frac{R(h)}{h^n} \right| < \left| \frac{R(h)}{hh_1 \cdots h_{n-1}} \right| = \left| \frac{R^{(n-1)}(h_{n-1})}{h_{n-1}} \right|.$$

But when h is small, h_{n-1} is also small. Thus the last fraction is close to $R^{(n)}(0) = 0$.

For the second part let $g(x) = R(h)x^{n+1} - h^{n+1}R(x)$. Then $g^{(i)}(0) = 0$ for $i = 1, \ldots, n$, and also $g(h) = 0$. Therefore $g'(h_1) = 0$ for some h_1 with $0 < |h_1| < |h|$. This implies that $g''(h_2) = 0$ for some h_2 with $0 < |h_2| < |h_1| < |h|$. Continuing this way we get $g^{(n)}(h_n) = 0$ for some h_n with $0 < |h_n| < |h|$. Then

$$(n+1)!R(h) - h^{n+1}f^{(n+1)}(a + h_{n+1}) = g^{(n+1)}(h_{n+1}) = 0,$$

for some h_{n+1}. Note that $P^{(n+1)} = 0$ as P is a polynomial of degree n. Now let $\theta = a + h_{n+1}$. ■

Remark. It is easy to see that P is the only polynomial of degree at most n for which we have

$$\lim_{h\to 0}\frac{f(a+h)-P(h)}{h^n}=0.$$

Because if there were two such polynomials P_1, P_2, we would have

$$\lim_{h\to 0}\frac{P_1(h)-P_2(h)}{h^n}=0.$$

Now $\deg P_1, \deg P_2 \le n$. So if $P_1 - P_2 \neq 0$ then

$$(P_1-P_2)(h)=a_nh^n+a_{n-1}h^{n-1}\cdots+a_mh^m,$$

where $m \le n$ and $a_m \neq 0$. But we have

$$a_m=\lim_{h\to 0}\frac{(P_1-P_2)(h)}{h^m}=\lim_{h\to 0}h^{n-m}\frac{(P_1-P_2)(h)}{h^n}=0,$$

which is a contradiction. Hence $P_1 = P_2$.

4.5 Convex Functions of One Variable

Definition 4.25. A real-valued function f on an interval I is a **convex** function if for all points $a, b \in I$ we have

$$f((1-t)a+tb)\le(1-t)f(a)+tf(b)$$

for all $0 \le t \le 1$. A function f is **concave** if $-f$ is convex.

Theorem 4.26. *Convex and concave functions are locally Lipschitz continuous in the interior of their domains, i.e. every point in the interior of their domains has a neighborhood on which they are Lipschitz.*

Proof. Suppose f is convex, the case of concave functions is similar. Let a be a point in the domain of f, and suppose $b > a$. For a point $a < c < b$ we have

$$f(c)=f(a+\frac{c-a}{b-a}(b-a))\le\frac{c-a}{b-a}f(b)+\frac{b-c}{b-a}f(a).$$

Therefore

$$m_{ac}:=\frac{f(c)-f(a)}{c-a}\le\frac{f(b)-f(a)}{b-a}=:m_{ab}.$$

Note that m_{ac} is the slope of the line joining $(a, f(a))$ and $(c, f(c))$. Similarly we have $m_{ab} \le m_{cb}$. Now let a be an interior point. Then $m_{da} \le m_{dx} \le m_{ax} \le m_{ab}$ for $d < a < x < b$. Therefore

$$\left|\frac{f(x)-f(a)}{x-a}\right|\le L,$$

where $L = \max\{m_{da}, m_{ab}\}$. The case of $x < a$ is similar. ■

Theorem 4.27. *A continuous function with increasing derivative is convex, and a continuous function with decreasing derivative is concave. In particular, a continuous function with positive second derivative is convex, and a continuous function with negative second derivative is concave.*

Proof. Let $g(t) = -f((1-t)a + tb) + (1-t)f(a) + tf(b)$. Then f is convex if and only if $g \geq 0$. We have

$$g'(t) = -(b-a)f'(a + t(b-a)) - f(a) + f(b).$$

As $g(0) = g(1) = 0$, there is c such that $g'(c) = 0$. But we must have $g' \geq 0$ on $(0, c)$ and $g' \leq 0$ on $(c, 1)$, since f' is increasing. Therefore $g \geq g(0) = 0$ on $(0, c]$, and $g \geq g(1) = 0$ on $[c, 1)$. The other case is similar. ■

Theorem 4.28. *The graph of a convex function lies above any tangent line to it, and the graph of a concave function lies below any tangent line to it.*

Proof. Suppose f is convex. Let a be a point at which f is differentiable, and let $b > a$ be another point in the domain of f. For a point $a < c < b$ we know that

$$\frac{f(c) - f(a)}{c - a} \leq \frac{f(b) - f(a)}{b - a}.$$

Hence we have

$$f'(a) = \lim_{c \to a^+} \frac{f(c) - f(a)}{c - a} \leq \lim_{c \to a^+} \frac{f(b) - f(a)}{b - a} = \frac{f(b) - f(a)}{b - a}.$$

Therefore $f(a) + f'(a)(b-a) \leq f(b)$. The other cases are similar. ■

Chapter 5

Integration

5.1 The Riemann Integral

Definition 5.1. A **partition** P of an interval $[a,b] \subset \mathbb{R}$ is a finite set of points $\{a_0, \ldots, a_n\}$ such that

$$a = a_0 < a_1 < \cdots < a_n = b.$$

The intervals $[a_{i-1}, a_i]$ are called the **subintervals** of the partition P. The **mesh** of the partition P is

$$\|P\| := \max_{i \leq n} |a_i - a_{i-1}|.$$

A **tagged partition** is a partition P with a sequence $T = (x_1, \ldots, x_n)$ of **tags**

$$x_i \in [a_{i-1}, a_i].$$

We say a partition Q is a **refinement** of a partition P if $P \subset Q$. The **common refinement** of two partitions P_1, P_2 is $P_1 \cup P_2$.

Remark. It is easy to see that for a refinement Q of P we have $\|Q\| \leq \|P\|$.

Definition 5.2. Let $f : [a,b] \to \mathbb{R}$. The **Riemann sum** of f corresponding to the tagged partition $P = \{a_0, \ldots, a_n\}$, $T = (x_1, \ldots, x_n)$ is

$$R(f, P, T) := \sum_{i=1}^{n} f(x_i)(a_i - a_{i-1}).$$

Definition 5.3. Let $f : [a,b] \to \mathbb{R}$. We say f is **Riemann integrable** (on $[a,b]$), if there exists $I \in \mathbb{R}$ so that $\forall \epsilon > 0$ $\exists \delta > 0$ such that for all tagged partitions P, T with $\|P\| < \delta$ we have

$$|I - R(f, P, T)| < \epsilon.$$

Note that by the next theorem I is unique. We call I the **Riemann integral** of f (over $[a, b]$) and denote it by

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

In this notation, the function f is also referred to as the *integrand.*

Theorem 5.4. *The integral of an integrable function is unique.*

Proof. Suppose there are two numbers I, J satisfying the above definition. Then for each $\epsilon > 0$ there is a tagged partition P, T such that

$$|I - J| \leq |I - R(f, P, T)| + |J - R(f, P, T)| < \epsilon + \epsilon = 2\epsilon.$$

Hence we must have $I - J = 0$. ■

Remark. Intuitively, the integral of a nonnegative function is the area of the region below its graph. But to make this idea precise, and to prove it rigorously, we have to use more advanced tools. For example, we can do this by using the notion of multiple Riemann integrals. See Theorem 8.47. A different approach, which provides a deeper understanding, is studied in measure theory.

Theorem 5.5. *A Riemann integrable function is bounded.*

Proof. Suppose to the contrary that f is an unbounded Riemann integrable function on $[a, b]$. Then there is $\delta > 0$ such that for all tagged partitions P, T with $\|P\| < \delta$ we have

$$|I - R(f, P, T)| < 1. \tag{$*$}$$

Let $P = \{a_i\}, T = (x_i)$ be a tagged partition with $\|P\| < \delta$. Since f is unbounded, it is unbounded on at least one of the subintervals $[a_{j-1}, a_j]$. Let $x'_i = x_i$ for $i \neq j$. Then we can find $x'_j \in [a_{j-1}, a_j]$ such that for $T' = (x'_i)$ we have

$$|R(f, P, T') - R(f, P, T)| = |f(x'_j) - f(x_j)|(a_j - a_{j-1}) > 2.$$

But this contradicts $(*)$. ■

Theorem 5.6. *Let $f, g : [a, b] \to \mathbb{R}$ be Riemann integrable, and $c, c_1, c_2 \in \mathbb{R}$. Then we have*

(i) *The constant function c is Riemann integrable and $\int_a^b c\,dx = (b - a)c$.*

(ii) *$c_1 f + c_2 g$ is Riemann integrable and*

$$\int_a^b [c_1 f(x) + c_2 g(x)]dx = c_1 \int_a^b f(x)dx + c_2 \int_a^b g(x)dx.$$

(iii) *If $f \leq g$ then*

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

(iv) *If $|f| \leq M$ then*

$$\left|\int_a^b f(x)dx\right| \leq M(b-a).$$

Proof. **(i)** The Riemann sums of a constant function are all $(b-a)c$, hence they converge to this number.

(ii) First note that for any tagged partition P, T we have

$$R(c_1 f + c_2 g, P, T) = c_1 R(f, P, T) + c_2 R(g, P, T).$$

Now let I, J be the integrals of f, g respectively. Let δ be small enough so that for all tagged partitions P, T with $\|P\| < \delta$ we have

$$|I - R(f, P, T)| < \frac{\epsilon}{2|c_1| + 2}, \qquad |J - R(g, P, T)| < \frac{\epsilon}{2|c_2| + 2}.$$

Then

$$\begin{aligned} &|c_1 I + c_2 J - R(c_1 f + c_2 g, P, T)| \\ &\qquad = |c_1 I + c_2 J - c_1 R(f, P, T) - c_2 R(g, P, T)| \\ &\qquad \leq |c_1||I - R(f, P, T)| + |c_2||J - R(g, P, T)| < \epsilon. \end{aligned}$$

(iii) First note that for all tagged partitions P, T we have

$$R(f, P, T) \leq R(g, P, T).$$

Now let I, J be the integrals of f, g respectively. Suppose to the contrary that $J < I$. Let δ be small enough so that for all tagged partitions P, T with $\|P\| < \delta$ we have

$$|I - R(f, P, T)| < \frac{I - J}{2}, \qquad |J - R(g, P, T)| < \frac{I - J}{2}.$$

Then we must have $R(f, P, T) > R(g, P, T)$, which is a contradiction.

(iv) We have $-M \leq f \leq M$. Now the result follows from parts (i) and (iii). ■

***Example* 5.7.** The characteristic function of $\mathbb{Q}$, i.e.

$$\chi_{\mathbb{Q}}(x) := \begin{cases} 1 & x \in \mathbb{Q}, \\ 0 & x \notin \mathbb{Q}, \end{cases}$$

is not Riemann integrable on any interval $[a, b]$. Because for any partition P we can choose a sequence T of rational tags and a sequence T' of irrational tags, so that

$$R(\chi_{\mathbb{Q}}, P, T) = b - a, \qquad R(\chi_{\mathbb{Q}}, P, T') = 0.$$

This function is also known as the **Dirichlet function**.

5.2 Integrable Functions

Definition 5.8. A subset A of $\mathbb{R}$ has **measure zero** if for every $\epsilon > 0$ there exist countably many intervals (a_i, b_i) such that $A \subset \bigcup_{i\geq 1}(a_i, b_i)$, and

$$\sum_{i\geq 1} b_i - a_i < \epsilon.$$

This series is called the **total length** of the family $\{(a_i, b_i)\}_{i\geq 1}$.

We say a property holds **almost everywhere**, abbreviated **a.e.**, if it holds for all points outside a set of measure zero.

Remark. Remember that a countable set is either finite or countably infinite.

Remark. An obvious consequence of the definition is that if A has measure zero and $B \subset A$, then B has measure zero too.

***Example* 5.9.** It is easy to see that every finite subset of $\mathbb{R}$ has measure zero. Also, we have seen that the Cantor set has measure zero.

Theorem 5.10. *Let $\{A_k\}$ be a countable family of sets that have measure zero. Then $\bigcup_k A_k$ has measure zero. In particular, every countable subset of $\mathbb{R}$ has measure zero.*

Proof. Let $\epsilon > 0$ be given. Then we can cover A_k with a countable family of intervals $\{(a_{k,i}, b_{k,i})\}_{i\geq 1}$ such that

$$\sum_{i\geq 1} b_{k,i} - a_{k,i} < \frac{\epsilon}{2^k}.$$

Then $\{(a_{k,i}, b_{k,i})\}_{i,k\geq 1}$ is a countable family of intervals that covers $\bigcup_k A_k$, and

$$\sum_{i,k\geq 1} b_{k,i} - a_{k,i} < \sum_{k\geq 1} \frac{\epsilon}{2^k} \leq \epsilon.$$

■

Remark. If we want to be completely rigorous in the above proof, we have to arrange the family of intervals $\{(a_{k,i}, b_{k,i})\}_{i,k\geq 1}$ into a sequence. Note that different arrangements does not change the total length of the family, since the length of each interval is positive and therefore the series of the total length is absolutely convergent. Now suppose we have arranged the family as the sequence $\{(a_j, b_j)\}_{j\geq 1}$. Then for any $N \in \mathbb{N}$ there is $M \in \mathbb{N}$ such that

$$\{(a_j, b_j)\}_{1\leq j\leq N} \subset \{(a_{k,i}, b_{k,i})\}_{1\leq i,k\leq M}.$$

Then we have

$$\sum_{j\leq N} b_j - a_j \leq \sum_{k\leq M}\sum_{i\leq M} b_{k,i} - a_{k,i} < \sum_{k\leq M} \frac{\epsilon}{2^k} < \epsilon.$$

Now by taking the limit as $N \to \infty$ we get $\sum_{j\geq 1} b_j - a_j \leq \epsilon$ as desired.

***Example* 5.11.** $\mathbb{Q}$ has measure zero, since it is countable.

Definition 5.12. Let I be a subset of $\mathbb{R}$, and let $f : I \to \mathbb{R}$. We define the **oscillation** of f at a point $x \in I$ as

$$\operatorname{osc}_x f := \lim_{r \to 0} \sup\{ |f(z) - f(y)| : z, y \in (x - r, x + r) \cap I\}.$$

Remark. Note that the supremums in the above expression decrease as $r \to 0$, hence the limit exists.

Remark. The oscillation at a point is a measure of the size of the discontinuity at that point. In particular, we can easily show that $\operatorname{osc}_x f = 0$ if and only if f is continuous at x.

Riemann-Lebesgue Theorem. *A function $f : [a, b] \to \mathbb{R}$ is Riemann integrable if and only if it is bounded and its set of discontinuities has measure zero.*

Proof. Let D be the set of discontinuities of f. Suppose D has measure zero. Also suppose that $|f| \le M$ for some $M > 0$. We want to show that f is Riemann integrable. The idea is to show that the Riemann sums of f satisfy a Cauchy criterion. Let $\epsilon > 0$ be given. Then there are countably many open intervals (α_k, β_k) such that

$$D \subset \bigcup (\alpha_k, \beta_k), \qquad \sum \beta_k - \alpha_k < \epsilon.$$

Now, f is continuous at each point of $K := [a, b] - \bigcup(\alpha_k, \beta_k)$. So for every $x \in K$ there is an open interval I_x containing x such that $|f(x) - f(y)| < \epsilon$ whenever $y \in I_x \cap [a, b]$. Then the collection

$$\mathcal{U} := \{(\alpha_k, \beta_k)\}_{k \ge 1} \cup \{I_x\}_{x \in K}$$

is an open covering of the compact set $[a, b]$. Thus it has a Lebesgue number $\delta > 0$, i.e. for every $z \in [a, b]$ there is $U \in \mathcal{U}$ such that $(z - \delta, z + \delta) \subset U$.

Let $P = \{a_i\}, T = (x_i)$ be a tagged partition, with $\|P\| < \delta$. Let $Q = \{b_j\}$ be a refinement of P, and let $S = (y_j)$ be a choice of tags for Q. Then we have

$$R(f, P, T) = \sum f(x_i)(a_i - a_{i-1}) = \sum f(x_j)(b_j - b_{j-1}),$$

where $x_j := x_i$ if $(b_{j-1}, b_j) \subset (a_{i-1}, a_i)$. Let

$$J := \{j : (b_{j-1}, b_j) \subset (a_{i-1}, a_i) \subset I_x \text{ for some } x \in K\}.$$

Then if $j \in J$ we have $x_j, y_j \in I_x$ for some $x \in K$, and therefore $|f(x_j) - f(y_j)| < 2\epsilon$. Now note that for any i and any $z \in (a_{i-1}, a_i)$ we have

$$(a_{i-1}, a_i) \subset (z - \delta, z + \delta) \subset U$$

for some $U \in \mathcal{U}$. Hence if $j \notin J$ we must have

$$(b_{j-1}, b_j) \subset (a_{i-1}, a_i) \subset (\alpha_k, \beta_k)$$

for some k. Thus

$$\sum_{j \notin J} (b_j - b_{j-1}) \leq \sum (\beta_k - \alpha_k) < \epsilon.$$

Note that k can be the same for several distinct $j_1, j_2, \dots$, i.e. we might have $(b_{j_l - 1}, b_{j_l}) \subset (\alpha_{k_0}, \beta_{k_0})$ for some k_0, and $l = 1, 2, \dots$. But this does not affect the inequality, since $\sum_l (b_{j_l} - b_{j_{l-1}}) \leq \beta_{k_0} - \alpha_{k_0}$, and we do not need to add $\beta_{k_0} - \alpha_{k_0}$ several times in the right hand side. Hence we have

$$\begin{aligned} |R(f,P,T) - R(f,Q,S)| &\leq \sum |f(x_j) - f(y_j)|(b_j - b_{j-1}) \\ &= \sum_{j \in J} |f(x_j) - f(y_j)|(b_j - b_{j-1}) \\ &\qquad + \sum_{j \notin J} |f(x_j) - f(y_j)|(b_j - b_{j-1}) \\ &< 2\epsilon \sum_{j \in J} (b_j - b_{j-1}) + 2M \sum_{j \notin J} (b_j - b_{j-1}) \\ &< 2\epsilon(b-a) + 2M\epsilon =: C\epsilon. \end{aligned}$$

Now let P, P' be two partitions with mesh less than δ. Let Q be the common refinement of P, P'. Let T, T', S be choices of tags for P, P', Q respectively. Then by the above inequality we get

$$\begin{aligned} |R(f,P,T) - R(f,P',T')| \leq |R(f,P,T) - R(f,Q,S)| \\ + |R(f,P',T') - R(f,Q,S)| < 2C\epsilon. \end{aligned}$$

This is the Cauchy criterion that we were looking for.

Finally let P_n be the partition that divides $[a, b]$ into n equal subintervals. Let T_n be the sequence of the right endpoints of these subintervals. Then for any $\epsilon > 0$ we can take n to be large enough so that $\|P_n\| = \frac{b-a}{n} < \delta$. Hence we have

$$|R(f, P_m, T_m) - R(f, P_n, T_n)| < 2C\epsilon,$$

for $m \geq n$. Therefore the sequence $R(f, P_n, T_n)$ is Cauchy in $\mathbb{R}$. Thus it converges to some number I. Now let n be large enough so that $\|P_n\| < \delta$ and $|I - R(f, P_n, T_n)| < C\epsilon$. Then for an arbitrary tagged partition P, T with mesh less than δ, we have

$$|I - R(f,P,T)| \leq |I - R(f,P_n,T_n)| + |R(f,P_n,T_n) - R(f,P,T)| < 3C\epsilon.$$

As ϵ is arbitrary we get the desired result. ■

Proof of the Converse. Next suppose f is Riemann integrable and its integral is I. Then we know that f is bounded. Let D be the set of discontinuities of f. We have $D = \bigcup_{k\geq 1} D_k$ where

$$D_k := \{x \in [a,b] : \operatorname{osc}_x f \geq \frac{1}{k}\}.$$

In order to show that D has measure zero, it suffices to show that each D_k has measure zero. Now for any given $\epsilon > 0$ we can find $\delta > 0$ such that for any tagged partition $P = \{a_j\}, T = (x_j)$ with $\|P\| < \delta$ we have

$$|R(f,P,T) - I| < \epsilon.$$

Let $S = (y_j)$ be another sequence of tags for P. Then we have

$$\left|\sum \big(f(x_j) - f(y_j)\big)(a_j - a_{j-1})\right| = |R(f,P,T) - R(f,P,S)| < 2\epsilon.$$

Consider some fixed k. Let $J := \{j : (a_{j-1}, a_j) \cap D_k \neq \emptyset\}$. Note that $\{(a_{j-1}, a_j)\}_{j\in J}$ is a finite family of open intervals that covers D_k, except for possibly finitely many points of $D_k \cap \{a_j\}$. But we can cover those finite points by finitely many open intervals with total length less than ϵ. Thus we only need to show that the total length of $\{(a_{j-1}, a_j)\}_{j\in J}$ is small. Now for $j \in J$ we can choose $x_j, y_j \in (a_{j-1}, a_j)$ such that

$$f(x_j) - f(y_j) \geq \frac{1}{2k}.$$

The reason is that there is $z \in D_k \cap (a_{j-1}, a_j)$, and since $\operatorname{osc}_z f \geq \frac{1}{k}$, we can find points x, y near z inside its open neighborhood (a_{j-1}, a_j) such that $|f(x) - f(y)|$ is as close to $\frac{1}{k}$ as we want. Then for $j \notin J$ we choose $x_j = y_j \in (a_{j-1}, a_j)$, so that $f(x_j) - f(y_j) = 0$. Thus we have

$$\begin{aligned}\frac{1}{2k}\sum_{j\in J}(a_j - a_{j-1}) &\leq \sum_{j\in J}\big(f(x_j) - f(y_j)\big)(a_j - a_{j-1}) \\ &= \sum \big(f(x_j) - f(y_j)\big)(a_j - a_{j-1}) < 2\epsilon.\end{aligned}$$

Hence $\sum_{j\in J}(a_j - a_{j-1}) < 4k\epsilon$. Therefore D_k has measure zero as desired. ■

Remark. An interesting consequence of the Riemann-Lebesgue theorem is that the interval $[a,b]$ does not have measure zero, which is itself a nontrivial fact. The reason is that the characteristic function of $\mathbb{Q}$ restricted to $[a,b]$ is a bounded function whose set of discontinuities is all of $[a,b]$, and it is not Riemann integrable. Therefore if the interval had measure zero we would have a contradiction. From this we can easily conclude that the open interval (a,b) does not have measure zero either (how?).

Remark. Another interesting result is that by the above remark the intervals in $\mathbb{R}$, and hence $\mathbb{R}$ itself, are uncountable. Since otherwise they would have zero measure. Note that this proof of the uncountability of $\mathbb{R}$ avoids any use of Cantor's diagonal argument.

***Example* 5.13.** Consider the function

$$f(x) = \begin{cases} 1 & \text{when } x = 0, \\ \frac{1}{q} & \text{when } x = \frac{p}{q} \in \mathbb{Q}, \\ & \quad \text{where } p \in \mathbb{Z} \text{ and } q \in \mathbb{N} \text{ have no common divisor,} \\ 0 & \text{when } x \notin \mathbb{Q}. \end{cases}$$

Note that any nonzero rational number x can be written uniquely as $\frac{p}{q}$, where p, q are as above. So f is well defined. We will show that f is continuous at every $x \notin \mathbb{Q}$, and it is discontinuous at every $x \in \mathbb{Q}$. Therefore f is Riemann integrable. To prove this, first suppose $x \in \mathbb{Q}$. Then there is a sequence of irrational numbers x_n converging to x. But $f(x_n) = 0$, so $f(x_n) \to 0 \neq f(x)$. Hence f is discontinuous at x.

Next suppose $x \notin \mathbb{Q}$, and $x_n \to x$. We have to show that $f(x_n) \to 0 = f(x)$, in order to conclude that f is continuous at x. Let $\epsilon > 0$ be given. Let $p \in \mathbb{Z}$ and $q \in \mathbb{N}$. We claim that there are at most finitely many $\frac{p}{q} \in \mathbb{Q} \cap (x-1, x+1)$ such that p, q have no common divisor, and $0 < q < \frac{1}{\epsilon}$. The reason is that we have $|\frac{p}{q} - x| < 1$. Hence $|p| < \frac{1}{\epsilon}(|x| + 1)$. Thus there are finitely many choices for p, q, and the claim follows. Let N be large enough so that for $n \geq N$, x_n is closer to x than any of these finitely many rational numbers. Also, we can take N to be large enough so that for $n \geq N$ we have $x_n \neq 0$. This is possible since $x \neq 0$. Now for $n \geq N$ we must have $|f(x_n)| < \epsilon$. Since if $x_n \notin \mathbb{Q}$ then $f(x_n) = 0$. And if $x_n \in \mathbb{Q}$ then we have $x_n = \frac{p}{q}$ where p, q have no common divisor, and $q > \frac{1}{\epsilon}$. Thus $f(x_n) = \frac{1}{q} < \epsilon$. Finally, as ϵ was arbitrary, we get $f(x_n) \to 0$ as desired.

Theorem 5.14. *Continuous functions are Riemann integrable.*

Proof. A continuous function on an interval $[a, b]$ is bounded, and its set of discontinuities is empty. ■

***Exercise* 5.15.** Show that the set of discontinuities of a monotone function is countable, and conclude that monotone functions are Riemann integrable.

Theorem 5.16. *The product of Riemann integrable functions is Riemann integrable.*

Proof. Let $Z(f)$ denote the set of discontinuities of a function f. Now suppose f, g are Riemann integrable. Then $Z(f)$ and $Z(g)$ have measure zero. But

$$Z(fg) \subset Z(f) \cup Z(g),$$

since fg is continuous at the points where both f, g are continuous. Thus $Z(fg)$ has measure zero too. Also, fg is bounded as f, g are bounded. Hence fg is Riemann integrable. ■

Theorem 5.17. *Suppose $f : [a, b] \to [c, d]$ is Riemann integrable and $\phi : [c, d] \to \mathbb{R}$ is continuous. Then $\phi \circ f$ is Riemann integrable.*

Proof. First note that $\phi \circ f$ is bounded, since ϕ is continuous. Let $Z(f)$ denote the set of discontinuities of f. Then $Z(f)$ has measure zero. As ϕ is continuous we have $Z(\phi \circ f) \subset Z(f)$, since $\phi \circ f$ is continuous at the points where f is continuous. Thus $Z(\phi \circ f)$ has measure zero too. Hence $\phi \circ f$ is Riemann integrable. ■

Theorem 5.18. *Suppose $f : [a, b] \to \mathbb{R}$ is Riemann integrable. Then $|f|$ is Riemann integrable and we have*

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

Proof. Since $|\ |$ is a continuous function, $|f|$ is Riemann integrable. The inequality follows from the monotonicity of the integral and $-|f| \leq f \leq |f|$. ■

***Exercise* 5.19.** Suppose f, g are integrable, and $|g| \geq m$ for some $m > 0$. Show that $\frac{f}{g}$ is Riemann integrable.

Theorem 5.20. *Suppose $f : [a, b] \to \mathbb{R}$ is Riemann integrable and $\psi : [c, d] \to [a, b]$ is a bijection such that $\psi^{-1} : [a, b] \to [c, d]$ is Lipschitz, i.e.*

$$|\psi^{-1}(x) - \psi^{-1}(y)| \leq K|x - y|$$

for all $x, y \in [a, b]$ and some $K > 0$. Then $f \circ \psi$ is Riemann integrable.

In particular, if $\psi : [c, d] \to [a, b]$ is a continuous bijection which is differentiable on (c, d) and $|\psi'| > \kappa$ for some $\kappa > 0$, then $f \circ \psi$ is Riemann integrable.

Proof. First note that $f \circ \psi$ is bounded, since f is bounded. Next note that as ψ^{-1} is continuous and $[a, b]$ is compact, $\psi = (\psi^{-1})^{-1}$ is continuous. So ψ, ψ^{-1} are homeomorphisms. Hence by the intermediate value theorem ψ, ψ^{-1} are strictly monotone, as shown in Exercise 2.84. Now let $Z(f)$ be the set of discontinuities of f, and let $Z(f \circ \psi)$ be the set of discontinuities of $f \circ \psi$. Then we have

$$Z(f \circ \psi) \subset \psi^{-1}(Z(f)),$$

since $f \circ \psi$ is continuous at a point x if f is continuous at $\psi(x)$. (Actually, the above two sets are equal due to the continuity of ψ^{-1}, but we do not use this fact). We want to show that $Z(f \circ \psi)$ has measure zero. It suffices to show that $\psi^{-1}(Z(f))$ has measure zero. Since finite sets have measure zero, we can assume that $c, d \notin \psi^{-1}(Z(f))$.

Suppose $\epsilon > 0$ is given. Let $\{(a_i, b_i)\}_{i\geq 1}$ be a countable family of intervals that covers $Z(f)$, with $\sum_{i\geq 1} b_i - a_i < \epsilon$. We can assume that each (a_i, b_i) is inside (a, b). To simplify the notation, we assume that ψ^{-1} is strictly increasing. Then $\{(\psi^{-1}(a_i), \psi^{-1}(b_i))\}_{i\geq 1}$ is a countable family of intervals covering $\psi^{-1}(Z(f))$, whose total length is

$$\sum_{i\geq 1}[\psi^{-1}(b_i) - \psi^{-1}(a_i)] \leq K\sum_{i\geq 1} b_i - a_i < K\epsilon.$$

For the second statement of the theorem, note that ψ^{-1} is continuous since ψ is continuous on the compact set $[c, d]$. Thus as we said before, ψ is strictly monotone. Hence $\psi(\{c, d\}) = \{a, b\}$, and $\psi((c, d)) = (a, b)$. Then ψ^{-1} is differentiable on (a, b) with $|(\psi^{-1})'| < \frac{1}{\kappa}$. Therefore by the mean value theorem we have

$$|\psi^{-1}(x) - \psi^{-1}(y)| \leq \frac{1}{\kappa}|x - y|$$

for all $x, y \in [a, b]$. ■

Theorem 5.21. *Suppose $f : [a, b] \to \mathbb{R}$ is Riemann integrable and ψ is a C^1 function (not necessarily one-to-one) on an open interval containing $[c, d]$. If $\psi([c, d]) \subset [a, b]$, and the set of critical points of ψ in $[c, d]$ i.e.*

$$C := \{x \in [c, d] : \psi'(x) = 0\}$$

has measure zero, then $f \circ \psi$ is Riemann integrable.

Proof. It is enough to show that $f \circ \psi$ is bounded and its set of discontinuities has measure zero. Since f is integrable and therefore bounded, $f \circ \psi$ is bounded too. Let $D \subset [c, d]$ be the set of discontinuities of $f \circ \psi$, and $Z \subset [a, b]$ be the set of discontinuities of f. As ψ is continuous we have $D \subset B := \psi^{-1}(Z)$. So it suffices to show that B has measure zero.

Now note that the set of critical points of ψ in $[c, d]$ is a closed set, since ψ' is continuous. By our assumption C has measure zero. Let I_n be the open set $\{x \in (c, d) : |\psi'(x)| > \frac{1}{n}\}$. Then similarly to the proof of the previous theorem we can show that $B \cap I_n$ has measure zero. Note that I_n is the union of countably many disjoint intervals, and the restriction of ψ to each of these intervals is one-to-one. Now we have

$$B = (B \cap \{c, d\}) \bigcup (B \cap C) \bigcup \bigcup_{n\geq 1} (B \cap I_n).$$

Hence B is the union of countably many sets with measure zero, so B has measure zero too. ■

Remark. The composition of two Riemann integrable functions is not in general Riemann integrable. Even if f is Riemann integrable and ψ is a C^1 bijection, $f \circ \psi$ is not necessarily Riemann integrable. Hence in the above two theorems we either imposed a lower bound on $|\psi'|$, or required the set of critical points of ψ to have measure zero. Interestingly, it can be shown that when ψ is a C^1 bijection whose set of critical points is not of measure zero, then there is a Riemann integrable function f such that $f \circ \psi$ is not Riemann integrable.

Theorem 5.22. *Suppose $f : [a, b] \to \mathbb{R}$ is Riemann integrable. Then the restriction of f to any subinterval of $[a, b]$ is Riemann integrable.*

Proof. The restriction of f to any subinterval is obviously bounded. Let $Z(f)$ denote the set of discontinuities of f. Then $Z(f)$ has measure zero. Suppose $[c, d]$ is a subinterval of $[a, b]$. Since $Z(f|_{[c,d]}) \subset Z(f)$, $Z(f|_{[c,d]})$ has measure zero too. Thus $f|_{[c,d]}$ is Riemann integrable. ■

5.3 Properties of the Integral

Notation. When $b > a$ we use the convention

$$\int_b^a f(x)dx := -\int_a^b f(x)dx,$$

and when $b = a$ we set $\int_a^a f(x)dx := 0$.

Theorem 5.23. *Suppose $f : [a, c] \to \mathbb{R}$ is Riemann integrable, and $b \in (a, c)$. Then we have*

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

Proof. We know that f is Riemann integrable over $[a, b]$ and $[b, c]$. Let P, P' be partitions of $[a, b]$, $[b, c]$ with mesh less than δ. Then $Q = P \cup P'$ is a partition of $[a, c]$ with mesh less than δ. Suppose T, T' are choices of tags for P, P', then $S = T \cup T'$ is a choice of tags for Q. Let I_{yz} be the integral of f over the interval $[y, z]$. Then for δ small enough we have

$$|I_{ac} - R(f, Q, S)| < \epsilon, \ |I_{ab} - R(f, P, T)| < \epsilon, \ |I_{bc} - R(f, P', T')| < \epsilon.$$

Now note that

$$R(f, Q, S) = R(f, P, T) + R(f, P', T').$$

Therefore $|I_{ac} - I_{ab} - I_{bc}| < 3\epsilon$. Since ϵ is arbitrary we must have $I_{ac} = I_{ab} + I_{bc}$ as desired. ■

Remark. It is easy to see that in the above theorem we can allow a, b, c to have any order other than $a < b < c$. We can also allow some of them to be equal. For example when $a < c < b$ we have

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

Remark. As a consequence of the above theorem, we can easily show by induction that if $P = \{a_0, \dots, a_n\}$ is a partition of $[a, c]$ then we have

$$\int_a^c f(x)dx = \sum_{j=1}^n \int_{a_{j-1}}^{a_j} f(x)dx.$$

Fundamental Theorem of Calculus I. *Suppose $f : [a, b] \to \mathbb{R}$ is Riemann integrable. Then*

$$F(x) = \int_a^x f(t)dt$$

is continuous on $[a, b]$. Also F is differentiable at every $x \in (a, b)$ where f is continuous, with derivative $F'(x) = f(x)$.

Remark. The function $F(x) = \int_a^x f(t)dt$ is called the **indefinite integral** of f.

Proof. We know that $|f| \le M$ for some $M > 0$. Then we have

$$|F(x+h) - F(x)| = \left| \int_x^{x+h} f(t)dt \right| \le M|h| \xrightarrow[h \to 0]{} 0.$$

Hence F is continuous.

Now note that we have

$$\frac{1}{h}[F(x+h) - F(x)] = \frac{1}{h} \int_x^{x+h} f(t)dt.$$

To show that F is differentiable, it is enough to prove that the right hand side of the above equation converges to $f(x)$ as $h \to 0$. Consider the constant function $g = f(x)$. Then

$$\begin{aligned}\left| \frac{1}{h} \int_x^{x+h} f(t)dt - f(x) \right| &= \left| \frac{1}{h} \int_x^{x+h} f(t)dt - \frac{1}{h} \int_x^{x+h} g(t)dt \right| \\ &= \frac{1}{|h|} \left| \int_x^{x+h} \big(f(t) - f(x)\big)dt \right| \\ &\le \sup_{|t-x| \le |h|} |f(t) - f(x)| \xrightarrow[h \to 0]{} 0,\end{aligned}$$

due to the continuity of f at x. ■

Remark. The above theorem implies that a continuous function f has an **antiderivative**, i.e. there is a differentiable function F such that $F' = f$ on the interior of the domain of f.

Fundamental Theorem of Calculus II. *Suppose f is defined on an open interval containing $[a, b]$, and is differentiable at every point of $[a, b]$. If f' is Riemann integrable over $[a, b]$, then*

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

Notation. We sometimes write $f(x)\big|_a^b$ instead of $f(b) - f(a)$.

Proof. Let $I = \int_a^b f'(x)dx$, and let P_n be the partition that divides $[a, b]$ into n equal subintervals. Then by the mean value theorem, for some $x_i \in (a_{i-1}, a_i)$ we have

$$\begin{aligned}\big|I - \big(f(b) - f(a)\big)\big| &= \Big|I - \sum \big(f(a_i) - f(a_{i-1})\big)\Big| \\ &= \Big|I - \sum f'(x_i)(a_i - a_{i-1})\Big| = \big|I - R(f', P_n, T_n)\big|.\end{aligned}$$

Where $T_n = (x_i)$. As $n \to \infty$, $R(f', P_n, T_n) \to I$. Thus we must have $I = f(b) - f(a)$. ■

***Example* 5.24.** There exists an increasing continuous function $F : [0, 1] \to \mathbb{R}$, called the **Cantor function**, whose derivative vanishes almost everywhere but it is not constant. The construction of F uses the Cantor set C. Let $x \in C$. We know that $x = \sum_{n\geq 1} \frac{\omega_n}{3^n}$ where each ω_n is either 0 or 2. Here $(0.\omega_1\omega_2\omega_3\cdots)_3$ is a representation of x in base 3. We define

$$F(x) = \sum_{n\geq 1} \frac{\omega_n/2}{2^n}.$$

In other words, we change each 2 in the expansion of x to 1 and let $F(x)$ be the number whose expansion in base 2 is the new sequence, i.e.

$$F(x) = \big(0.\frac{\omega_1}{2}\frac{\omega_2}{2}\frac{\omega_3}{2}\dots\big)_2.$$

For example $\frac{1}{4}$ can be represented in base 3 as $(0.020202\dots)_3$. So the representation of $F(\frac{1}{4})$ in base 2 is $(0.010101\dots)_2$, i.e. $F(\frac{1}{4}) = \frac{1}{3}$. Similarly we can see that $F(0) = 0$ and $F(1) = 1$.

If $x \in [0, 1] - C$ then it belongs to one of the intervals (c_1, c_2) that we removed during the construction of C. Then we have $F(c_1) = F(c_2)$. To see this we can show by induction that if (c_1, c_2) is removed during the nth step of construction of C then the representation of c_1, c_2 in base 3 is of the form

$$c_1 = (0.\omega_1 \cdots \omega_{n-1}1)_3 = (0.\omega_1 \cdots \omega_{n-1}0222\dots)_3, \qquad c_2 = (0.\omega_1 \cdots \omega_{n-1}2)_3,$$

where each ω_i is either 0 or 2. This is clear for the only interval removed during the first step of construction i.e. $(\frac{1}{3}, \frac{2}{3})$. For the induction step note that the endpoints of the closest intervals removed during the $(n+1)$th step of construction of C which are respectively on the left and right hand side of (c_1, c_2) are

$$\begin{aligned} c_1 - \frac{2}{3^{n+1}} &= (0.\omega_1 \cdots \omega_{n-1}01)_3, & c_1 - \frac{1}{3^{n+1}} &= (0.\omega_1 \cdots \omega_{n-1}02)_3,\\ c_2 + \frac{1}{3^{n+1}} &= (0.\omega_1 \cdots \omega_{n-1}21)_3, & c_2 + \frac{2}{3^{n+1}} &= (0.\omega_1 \cdots \omega_{n-1}22)_3. \end{aligned}$$

Therefore by using this representation we get

$$F(c_1) = \big(0.\frac{\omega_1}{2} \cdots \frac{\omega_{n-1}}{2}0111\dots\big)_2 = \big(0.\frac{\omega_1}{2} \cdots \frac{\omega_{n-1}}{2}1\big)_2 = F(c_2).$$

Now for $x \in (c_1, c_2)$ we define $F(x)$ to be the same as $F(c_1) = F(c_2)$. Thus F is constant on the interval (c_1, c_2). This is true for every interval that we removed during the construction of C, i.e. F is constant over each such interval. As a result $F'(x) = 0$ for any $x \in [0, 1] - C$, because x belongs to one of the removed intervals. Therefore F' vanishes almost everywhere, since C has measure zero.

It only remains to show that F is continuous and increasing. Assume that for $x, y \in [0, 1]$ we have $0 < y - x < 3^{-n}$. If $x, y \notin C$ then either x, y belong to the same interval removed during the construction of C in which case $F(x) = F(y)$, or they belong to two different intervals. In the latter case, and in the case that x and/or y belong to C, we have $F(y) = F(c)$ and $F(x) = F(\tilde{c})$, where $x \le \tilde{c} < c \le y$ and $c, \tilde{c} \in C$. But then we have $0 < c - \tilde{c} < 3^{-n}$, so the representation of $c, \tilde{c}$ in base 3 using only $0, 2$ agree on at least the first n terms. In addition if their representations agree on the first $m \ge n$ terms, then the $(m + 1)$th digit in the representation of c is greater than the $(m + 1)$th digit in the representation of $\tilde{c}$. Thus the representation of $F(c), F(\tilde{c})$ in base 2 agree on the first m terms too, and the $(m+1)$th digit in the representation of $F(c)$ is greater than the $(m+1)$th digit in the representation of $F(\tilde{c})$. Hence

$$0 \le F(y) - F(x) \le 2^{-n}.$$

Note that by our construction F is increasing, but not strictly increasing. There are continuous functions that are strictly increasing and their derivatives vanish almost everywhere. But their construction requires more advanced tools. ■

Integration by Parts. *Suppose f, g are differentiable on an open interval containing $[a, b]$, and f', g' are Riemann integrable on $[a, b]$. Then*

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

Proof. Note that f, g are continuous, hence they are Riemann integrable. Also note that fg' and $f'g$ are Riemann integrable. Now we can apply the fundamental theorem of calculus to $F = fg$, noting that $F' = f'g + fg'$. ■

Remark. It is trivial that we can allow $b \leq a$ in the above two theorems. We can also allow $d \leq c$ in the following theorem.

Change of Variable. *Suppose g has a continuous derivative on an open interval containing $[c, d]$. Also suppose that g' is nonzero on (c, d). Let I be the closed interval with endpoints $g(c), g(d)$. Then for a Riemann integrable function $f : I \to \mathbb{R}$ we have*

$$\int_{g(c)}^{g(d)} f(x)dx = \int_c^d f(g(t))g'(t)dt.$$

Remark. In Calculus courses this theorem is also referred to as *integration by substitution.*

Proof. First we assume that g' is nonzero on an open interval containing $[c, d]$. Let $a = g(c)$ and $b = g(d)$. As g' is continuous and nonzero, it is either positive or negative. First suppose $g' > 0$. Note that g is continuous and strictly increasing, thus $g : [c, d] \to [a, b]$ is a bijection by the intermediate value theorem. Now as g' is continuous and positive on the compact set $[c, d]$, there are positive constants K, κ such that

$$0 < \kappa \leq g' \leq K \qquad \text{on } [c, d].$$

Thus $f(g)g'$ is Riemann integrable on $[c, d]$.

Let $P_n = \{c_i\}$ be the partition that divides $[c, d]$ into n equal subintervals. Then there are $x_i \in (c_{i-1}, c_i)$ such that

$$g(c_i) - g(c_{i-1}) = g'(x_i)(c_i - c_{i-1}).$$

Now let $a_i := g(c_i)$. As g is strictly increasing, $Q_n := \{a_i\}$ is a partition of $[a, b]$. We also have

$$\|Q_n\| \leq K\|P_n\|.$$

Set $y_i := g(x_i)$. Then $T_n = (x_i)$ and $S_n = (y_i)$ are choices of tags for P_n, Q_n respectively. Hence we have

$$\begin{aligned} R(f, Q_n, S_n) &= \sum f(y_i)(a_i - a_{i-1}) \\ &= \sum f(g(x_i))g'(x_i)(c_i - c_{i-1}) = R(f(g)g', P_n, T_n). \end{aligned}$$

As $n \to \infty$, $\|P_n\| \to 0$ and $\|Q_n\| \to 0$. Thus the Riemann sums converge to the corresponding integrals, and we get the desired result.

Next suppose $g' < 0$. Then $g : [c, d] \to [b, a]$ is a continuous strictly decreasing bijection. We also have

$$0 < \kappa \leq |g'| \leq K \qquad \text{on } [c, d],$$

for some positive constants K, κ. Furthermore, $g'f(g)$ is Riemann integrable on $[c, d]$. We can repeat the above proof, but this time we define $a_i := g(c_{n-i})$ and $y_i := g(x_{n-i+1})$. Then as g is strictly decreasing, $Q_n := \{a_i\}$ is a partition of $[b, a]$. Therefore

$$\begin{aligned}
-R(f, Q_n, S_n) &= -\sum_{i=1}^{n} f(y_i)(a_i - a_{i-1}) \\
&= -\sum_{i=1}^{n} f(g(x_{n-i+1}))\big(g(c_{n-i}) - g(c_{n-i+1})\big) \\
&= -\sum_{j=1}^{n} f(g(x_j))\big(g(c_{j-1}) - g(c_j)\big) \qquad\qquad (j := n - i + 1) \\
&= \sum_{i=1}^{n} f(g(x_j))g'(x_j)(c_j - c_{j-1}) = R(f(g)g', P_n, T_n).
\end{aligned}$$

Hence in the limit $n \to \infty$ we obtain

$$\int_{g(c)}^{g(d)} f(x)dx = -\int_b^a f(x)dx = \int_c^d f(g(t))g'(t)dt.$$

Now consider the general case in which we only assume that g' is nonzero on (c, d). Let $\epsilon > 0$ be small. Then g has a continuous nonzero derivative on an open interval containing $[c + \epsilon, d - \epsilon]$. Hence we have

$$\int_{g(c+\epsilon)}^{g(d-\epsilon)} f(x)dx = \int_{c+\epsilon}^{d-\epsilon} f(g(t))g'(t)dt. \qquad (*)$$

Suppose $M > 0$ is an upper bound for $|f|$. Then we have

$$\begin{aligned}
\left| \int_{g(c)}^{g(d)} f(x)dx - \int_{g(c+\epsilon)}^{g(d-\epsilon)} f(x)dx \right| &= \left| \int_{g(d-\epsilon)}^{g(d)} f(x)dx + \int_{g(c)}^{g(c+\epsilon)} f(x)dx \right| \\
&\leq \int_{g(d-\epsilon)}^{g(d)} |f(x)|dx + \int_{g(c)}^{g(c+\epsilon)} |f(x)|dx \\
&\leq M\big(|g(d) - g(d - \epsilon)| + |g(c + \epsilon) - g(c)|\big).
\end{aligned}$$

Hence as $\epsilon \to 0$ we have $\int_{g(c+\epsilon)}^{g(d-\epsilon)} f(x)dx \to \int_{g(c)}^{g(d)} f(x)dx$, since g is continuous. Similarly we can show that $\int_{c+\epsilon}^{d-\epsilon} f(g(t))g'(t)dt \to \int_{c}^{d} f(g(t))g'(t)dt$, since g' is bounded too. Thus we get the desired result if we let $\epsilon \to 0$ in $(*)$. ■

Remark. In the above theorem if f is continuous on an open interval containing the image of g, then we only need to assume that g is differentiable and g' is Riemann integrable. Because if F is an antiderivative of f, then for $G = F \circ g$ we have $G' = (F' \circ g)g' = (f \circ g)g'$. Thus the fundamental theorem of calculus implies

$$\int_c^d f(g(t))g'(t)dt = G(d) - G(c) = F(g(d)) - F(g(c)) = \int_{g(c)}^{g(d)} f(x)dx.$$

The hypothesis of continuity of f can be further weakened to mere Riemann integrability of f, but the proof is more involved. Interestingly, in this more general case, $(f \circ g)g'$ is Riemann integrable while $f \circ g$ is not necessarily Riemann integrable.

***Exercise* 5.25.** Suppose f, g are Riemann integrable functions on $[a, b]$, and $f = g$ a.e. Show that

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

Hint: Since intervals do not have measure zero, for every partition of $[a, b]$ we can choose tags at which f, g are equal.

Remark. The assumption of Riemann integrability of both f, g is critical in the above exercise, since for example the characteristic function of $\mathbb{Q}$ is a.e. zero but it is not Riemann integrable.

***Exercise* 5.26.** Show that if f is continuous on $[a, b]$, then there is $c \in [a, b]$ such that

$$f(c) = \frac{1}{b-a}\int_a^b f(x)dx.$$

This is the *mean value theorem for integrals.*

Improper Integrals

Definition 5.27. Suppose $f : [a, b) \to \mathbb{R}$ is integrable on any closed subinterval $[a, c]$. Then the **improper Riemann integral** of f is

$$\int_a^b f(x)dx := \lim_{c \to b^-} \int_a^c f(x)dx$$

if the limit exists. Here b can be $+\infty$ too. We can similarly define the improper integral when $f : (a, b] \to \mathbb{R}$, where here a can be $-\infty$ too. We say an improper integral is **convergent** if its corresponding limit exists and is finite. Finally for $f:\mathbb{R}\to\mathbb{R}$ we define

$$\int_{-\infty}^{\infty} f(x)dx := \int_{-\infty}^{0} f(x)dx + \int_{0}^{\infty} f(x)dx$$

if the two improper integrals on the right hand side converge.

Integral Test. *Suppose $f:[1,\infty)\to\mathbb{R}$ is a decreasing nonnegative function. Then the series $\sum_{n=1}^{\infty} f(n)$ is convergent if and only if $\int_1^\infty f(x)dx$ is convergent.*

Proof. We have $f(k+1)\le\int_k^{k+1} f(x)dx\le f(k)$. Thus

$$\sum_{k=n+1}^{m+1} f(k) \le \int_n^{m+1} f(x)dx \le \sum_{k=n}^{m} f(k),$$
$$\int_n^{m+1} f(x)dx \le \sum_{k=n}^{m} f(k) \le \int_{n-1}^{m} f(x)dx.$$

Now if the improper integral is finite, then $\int_n^m f(x)dx\to 0$ as $m,n\to\infty$. Hence the sequence of the partial sums is Cauchy, and the series is convergent.

Conversely if the series is convergent, then $\sum_{k=n}^{m} f(k)\to 0$ as $m,n\to\infty$. Therefore the sequence

$$a_n := \int_1^n f(x)dx$$

is Cauchy. Thus $a_n\to a$. Then we must have $\int_1^\infty f(x)dx = a$, since

$$\int_1^n f(x)dx \le \int_1^c f(x)dx \le \int_1^{n+1} f(x)dx$$

if $n\le c<n+1$. ■

***Example* 5.28.** The series $\sum_{n=1}^{\infty}\frac{1}{n^p}$ is convergent if and only if $p>1$. Because when $p\ne 1$ we have

$$\int_1^\infty \frac{1}{x^p}dx = \lim_{c\to\infty}\Big(\frac{1}{1-p}c^{1-p}-\frac{1}{1-p}\Big) = \begin{cases} \frac{1}{p-1} & p>1,\\ \infty & p<1.\end{cases}$$

And when $p=1$ we have

$$\int_1^\infty \frac{1}{x}dx = \lim_{c\to\infty}(\log c-\log 1)=\infty.$$

5.4 The Darboux Integral

Definition 5.29. Let $f : [a, b] \to \mathbb{R}$ be a bounded function. The **lower sum** and **upper sum** of f with respect to the partition $P = \{a_0, \dots, a_n\}$ of $[a, b]$ are respectively

$$L(f, P) := \sum_{i=1}^{n} m_i(a_i - a_{i-1}), \qquad U(f, P) := \sum_{i=1}^{n} M_i(a_i - a_{i-1}),$$

where

$$m_i = \inf\{f(x) : a_{i-1} \le x \le a_i\}, \qquad M_i = \sup\{f(x) : a_{i-1} \le x \le a_i\}.$$

Remark. It is obvious that for any choice of tags T we have

$$L(f, P) \le R(f, P, T) \le U(f, P).$$

Also note that $L(f, P)$ and $U(f, P)$ are not necessarily Riemann sums for some choice of tags, because f does not necessarily achieve its infimum or supremum over the subintervals $[a_{i-1}, a_i]$.

Definition 5.30. Let $f : [a, b] \to \mathbb{R}$ be a bounded function. The **lower integral** and **upper integral** of f are respectively

$$\underline{\int_a^b} f(x)dx := \sup_P L(f, P), \qquad \overline{\int_a^b} f(x)dx := \inf_P U(f, P).$$

Here, P ranges over all partitions of $[a, b]$. We say f is **Darboux integrable** (on $[a, b]$) if

$$\underline{\int_a^b} f(x)dx = \overline{\int_a^b} f(x)dx,$$

and in this case we denote this common value by $\int_a^b f(x)dx$ and call it the **Darboux integral** of f (over $[a, b]$).

Proposition 5.31. *Suppose P is a partition of $[a, b]$, and Q is a refinement of P. Then for any bounded function $f : [a, b] \to \mathbb{R}$ we have*

$$L(f, P) \le L(f, Q) \le U(f, Q) \le U(f, P).$$

Proof. Suppose $P = \{a_0, \dots, a_n\}$. It suffices to prove the claim when $Q = P \cup \{c\}$. The general case then follows by an easy induction. Suppose $a_{j-1} < c < a_j$. Let

$$m := \inf\{f(x) : c \le x \le a_j\}, \qquad m' := \inf\{f(x) : a_{j-1} \le x \le c\}.$$

Then we have $m, m' \geq m_j = \inf\{f(x) : a_{j-1} \leq x \leq a_j\}$. Hence

$$\begin{aligned} L(f,Q) - L(f,P) &= m(a_j - c) + m'(c - a_{j-1}) - m_j(a_j - a_{j-1}) \\ &\geq m_j(a_j - c) + m_j(c - a_{j-1}) - m_j(a_j - a_{j-1}) = 0. \end{aligned}$$

The case of upper sums is similar. ■

Remark. Suppose P_1, P_2 are two partitions of $[a,b]$, and $Q = P_1 \cup P_2$ is their common refinement. Then the above proposition implies that

$$L(f,P_1) \leq L(f,Q) \leq U(f,Q) \leq U(f,P_2).$$

Thus any lower sum is less than or equal to any upper sum. As a result for any bounded function f we have

$$\underline{\int_a^b} f(x)dx \leq \overline{\int_a^b} f(x)dx.$$

Theorem 5.32. *A bounded function $f : [a,b] \to \mathbb{R}$ is Darboux integrable if and only if for all $\epsilon > 0$ there exists a partition P of $[a,b]$ such that*

$$U(f,P) - L(f,P) < \epsilon.$$

Proof. Suppose f is Darboux integrable. So we have $\overline{\int_a^b} f(x)dx = \underline{\int_a^b} f(x)dx$. Let $\epsilon > 0$ be given. Since the upper integral is the infimum of the upper sums and the lower integral is the supremum of the lower sums, there are partitions P_1, P_2 such that

$$U(f,P_1) - \overline{\int_a^b} f(x)dx < \frac{\epsilon}{2}, \qquad \underline{\int_a^b} f(x)dx - L(f,P_2) < \frac{\epsilon}{2}.$$

Therefore we have $U(f,P_1) - L(f,P_2) < \epsilon$. Now let P be the common refinement of P_1, P_2. Then we have

$$U(f,P) - L(f,P) \leq U(f,P_1) - L(f,P_2) < \epsilon,$$

because refining a partition causes the upper sum to decrease and the lower sum to increase.

Next suppose f satisfies the specified property in the theorem. Then for all $\epsilon > 0$ we have $0 \leq \overline{\int_a^b} f(x)dx - \underline{\int_a^b} f(x)dx < \epsilon$. Thus $\overline{\int_a^b} f(x)dx = \underline{\int_a^b} f(x)dx$, and f is Darboux integrable. ■

Theorem 5.33. *A function $f : [a,b] \to \mathbb{R}$ is Riemann integrable if and only if it is Darboux integrable. In this case, the Riemann integral of f is the same as its Darboux integral.*

Proof. First suppose f is Riemann integrable and I is its Riemann integral. Then f is bounded. Suppose $\epsilon > 0$ is given. There is $\delta > 0$ such that if P is a partition of $[a,b]$ with $\|P\| < \delta$ then $|R(f,P,T) - I| < \frac{\epsilon}{4}$ for any choice of tags T. Let $P = \{a_0, \dots, a_n\}$ be such a partition, and let m_i, M_i be respectively the infimum and supremum of f on $[a_{i-1}, a_i]$. Then we can choose tags $T_1 = (x_1, \dots, x_n)$ such that $0 \le f(x_i) - m_i < \frac{\epsilon}{4(b-a)}$. Then we have

$$0 \le R(f,P,T_1) - L(f,P) = \sum (f(x_i) - m_i)(a_i - a_{i-1}) < \frac{\epsilon}{4}.$$

Similarly we can choose tags T_2 so that

$$0 \le U(f,P) - R(f,P,T_2) < \frac{\epsilon}{4}.$$

Therefore we have

$$\begin{aligned} U(f,P) - L(f,P) = U(f,P) &- R(f,P,T_2) + R(f,P,T_2) - I \\ &+ I - R(f,P,T_1) + R(f,P,T_1) - L(f,P) < \epsilon. \end{aligned}$$

Hence f is Darboux integrable. Finally note that we also have $|U(f,P) - I| < \frac{\epsilon}{2}$. Therefore $|\overline{\int_a^b} f(x)dx - I| \le \frac{\epsilon}{2}$. Thus the Darboux integral of f is the same as its Riemann integral, since ϵ is arbitrary.

Next suppose f is Darboux integrable. Then f is bounded. Suppose $\epsilon > 0$ is given. Then there is a partition P such that $U(f,P) - L(f,P) < \epsilon$. Since any Riemann sum is between the upper sum and the lower sum, for any choices of tags T, S for P we have

$$|R(f,P,T) - R(f,P,S)| < \epsilon.$$

Now we can repeat the argument given at the end of the proof of Riemann-Lebesgue theorem to conclude that the set of discontinuities of f has measure zero. Hence f is Riemann integrable. Therefore the Riemann integral of f is the same as its Darboux integral as we proved in the last paragraph. ■

***Exercise* 5.34.** Prove that a Darboux integrable function is Riemann integrable, without using the Riemann-Lebesgue theorem.

Theorem 5.35. *A bounded function $f : [a,b] \to \mathbb{R}$ is Riemann integrable if and only if for all $\epsilon > 0$ there exists a partition P of $[a,b]$ such that*

$$U(f,P) - L(f,P) < \epsilon.$$

Proof. This is a consequence of the previous two theorems. ■

Chapter 6

Sequences and Series of Functions

6.1 Uniform Convergence

Definition 6.1. Let X, Y be two metric spaces. A sequence of functions $f_n : X \to Y$ **converges pointwise** to the function $f : X \to Y$ if $f_n(x) \to f(x)$ for all $x \in X$. The sequence (f_n) **converges uniformly** to f if

$$\begin{aligned}&\forall \epsilon > 0\ \exists N \in \mathbb{N} \text{ such that}\\ &\qquad \forall n \geq N\ \forall x \in X \text{ we have } d_Y(f_n(x), f(x)) < \epsilon.\end{aligned}$$

Remark. It is obvious that uniform convergence implies pointwise convergence. The difference between the two modes of convergence is that in uniform convergence the integer N does not depend on $x \in X$.

***Example* 6.2.** Let $f_n : (0,1) \to \mathbb{R}$ be given by $f_n(x) = x^n$. Then f_n converges pointwise to the constant function 0, but not uniformly (why?).

Theorem 6.3. *Suppose the sequence of functions $f_n : X \to Y$ converges uniformly to $f : X \to Y$, and each f_n is continuous at $a \in X$. Then f is continuous at a. In particular, the uniform limit of a sequence of continuous functions is continuous.*

Proof. Given $\epsilon > 0$, let N be large enough such that

$$d_Y(f_n(x), f(x)) < \frac{\epsilon}{3},$$

for all $x \in X$ and $n \geq N$. Now there is $\delta > 0$ so that

$$d_X(x, a) < \delta \implies d_Y(f_N(x), f_N(a)) < \frac{\epsilon}{3}.$$

Then for $d_X(x, a) < \delta$ we have

$$d_Y(f(x), f(a)) \leq d_Y(f(x), f_N(x)) + d_Y(f_N(x), f_N(a)) + d_Y(f_N(a), f(a)) < \epsilon. \quad ■$$

Remark. We can state the above result as

$$\lim_{x\to a}\lim_{n\to\infty} f_n(x) = \lim_{x\to a} f(x) = f(a) = \lim_{n\to\infty} f_n(a) = \lim_{n\to\infty}\lim_{x\to a} f_n(x).$$

In other words, we can interchange the two limits. This is not possible in general, even for sequences. For example

$$\lim_{m\to\infty}\lim_{n\to\infty}\frac{m}{m+n} = \lim_{m\to\infty} 0 = 0 \neq 1 = \lim_{n\to\infty} 1 = \lim_{n\to\infty}\lim_{m\to\infty}\frac{m}{m+n}.$$

***Example* 6.4.** The assumption of uniform convergence is essential in the above theorem. For example let $f_n : [0,1] \to \mathbb{R}$ be given by $f_n(x) = x^n$. Then f_n converges pointwise to the discontinuous function

$$f(x) = \begin{cases} 0 & x \in [0,1), \\ 1 & x = 1. \end{cases}$$

***Example* 6.5.** Let $f_n : [0,1] \to \mathbb{R}$ for $n \geq 2$ be defined as

$$f_n(x) = \begin{cases} nx & 0 \leq x \leq \frac{1}{n}, \\ 2 - nx & \frac{1}{n} \leq x \leq \frac{2}{n}, \\ 0 & \frac{2}{n} \leq x \leq 1. \end{cases}$$

Then f_n converges pointwise to the constant function 0, but not uniformly (why?). This example illustrates the fact that pointwise convergence does not imply uniform convergence, even if the functions and their limit are all continuous and uniformly bounded with a compact domain of definition.

Definition 6.6. Suppose X, Y are metric spaces. We denote by $C^0(X,Y)$ the space of all continuous functions from X to Y. We also use the convention $C^0(X) := C^0(X,\mathbb{R})$. In addition, we denote by $C^0_b(X,Y)$ and $C^0_b(X)$, the subsets of $C^0(X,Y)$ and $C^0(X)$ consisting of bounded continuous functions.

Remark. Note that when X is compact, $C^0_b(X) = C^0(X)$, since all continuous functions on a compact space are bounded.

Theorem 6.7. *Suppose X is a metric space, and $f, g \in C^0_b(X)$. Then*

$$d_{\sup}(f,g) := \sup\{\, |f(x) - g(x)| : x \in X\}$$

*is a metric on $C^0_b(X)$, called the **sup metric**. Furthermore, the convergence with respect to the sup metric is the same as uniform convergence.*

Proof. First note that $|f-g|$ is bounded, since

$$|f-g| \leq |f|+|g| \leq M_1+M_2,$$

where M_1, M_2 are upper bounds for $|f|, |g|$ respectively. Thus $d_{\sup}(f,g)$ is finite. It is obvious that $d_{\sup}$ is positive definite and symmetric. The triangle inequality is also easy to check. For $f,g,h \in C_b^0(X)$ we have

$$\begin{aligned} d_{\sup}(f,h) &= \sup_{x\in X}\{|f(x)-h(x)|\} \\ &\leq \sup_{x\in X}\{|f(x)-g(x)|+|g(x)-h(x)|\} \\ &\leq \sup_{x\in X}\{|f(x)-g(x)|\}+\sup_{x\in X}\{|g(x)-h(x)|\} \\ &= d_{\sup}(f,g)+d_{\sup}(g,h). \end{aligned}$$

Here we used the fact that $\sup_{a\in A,\, b\in B}\{a+b\} \leq \sup_{a\in A}\{a\}+\sup_{b\in B}\{b\}$.

Now suppose (f_n) is a sequence in $C_b^0(X)$ that converges to $f \in C_b^0(X)$ in the sup metric. This means that for every $\epsilon > 0$ there is $N \in \mathbb{N}$ such that for all $n \geq N$ we have

$$\sup_{x\in X}\{|f_n(x)-f(x)|\} = d_{\sup}(f_n,f) < \epsilon.$$

But this means that for all $x \in X$ we have $|f_n(x)-f(x)| < \epsilon$. So (f_n) converges uniformly to f. The converse is also true, since if $|f_n(x)-f(x)| < \epsilon$ for all $x \in X$, then $\sup_{x\in X}\{|f_n(x)-f(x)|\} < \epsilon$ too. ■

Remark. For $f \in C_b^0(X)$ we also define the **sup norm**

$$\|f\|_{\sup} := \sup\{\,|f(x)| : x \in X\},$$

which is obviously finite since f is bounded. Then we have $d_{\sup}(f,g) = \|f-g\|_{\sup}$, and $\|f\|_{\sup} = d_{\sup}(f,0)$. Thus in particular we have

$$\begin{aligned} \|f+g\|_{\sup} &= d_{\sup}(f+g,0) \leq d_{\sup}(f+g,g)+d_{\sup}(g,0) \\ &= \|f+g-g\|_{\sup}+\|g\|_{\sup} = \|f\|_{\sup}+\|g\|_{\sup}. \end{aligned}$$

Theorem 6.8. *Suppose X is a metric space. Then the space $C_b^0(X)$ equipped with the metric $d_{\sup}$ is a complete metric space. In particular, $C^0(X)$ with $d_{\sup}$ is a complete metric space when X is compact.*

Proof. Let (f_n) be a Cauchy sequence in $C_b^0(X)$. Then for each $\epsilon > 0$ there is $N \in \mathbb{N}$ such that for all $m,n \geq N$ we have

$$\sup_{x\in X}\{|f_n(x)-f_m(x)|\} = d_{\sup}(f_n,f_m) < \epsilon.$$

Thus in particular, for each $x \in X$ the sequence $(f_n(x))$ is a Cauchy sequence in $\mathbb{R}$. Hence it converges to a real number that we call $f(x)$. Therefore (f_n) converges pointwise to f. We have to show that $f \in C_b^0(X)$, and (f_n) converges uniformly to f. If we show the latter, then f is automatically continuous since it is the uniform limit of continuous functions f_n. Also, f will be automatically bounded. Because for large n we would have $\|f - f_n\|_{\text{sup}} < 1$. But $\|f_n\|_{\text{sup}} \leq M$ for some $M > 0$. Hence $\|f\|_{\text{sup}} < M + 1$.

Now suppose $\epsilon > 0$ is given. Then there is $N \in \mathbb{N}$ such that for all $m, n \geq N$ and all $x \in X$ we have

$$|f_n(x) - f_m(x)| \leq d_{\text{sup}}(f_n, f_m) < \frac{\epsilon}{2}.$$

In addition for each $x \in X$ we have $f_n(x) \to f(x)$. Thus for each $x \in X$ there is $m(x) \geq N$ such that

$$|f(x) - f_{m(x)}(x)| < \frac{\epsilon}{2}.$$

Therefore for each $x \in X$ and $n \geq N$ we have

$$|f(x) - f_n(x)| < |f(x) - f_{m(x)}(x)| + |f_{m(x)}(x) - f_n(x)| < \epsilon.$$

Hence for all $n \geq N$ we have $d_{\text{sup}}(f, f_n) < \epsilon$ as desired. ■

Definition 6.9. Suppose (f_n) is a sequence of real-valued functions on some metric space. The series $\sum_{n=1}^{\infty} f_n$ is the sequence of partial sums $S_k = \sum_{n=1}^{k} f_n$. If (S_k) converges pointwise, we denote its limit by the same notation $\sum_{n=1}^{\infty} f_n$ and we say the series converges. If (S_k) converges uniformly we say the series $\sum_{n=1}^{\infty} f_n$ converges uniformly. And if $\sum_{n=1}^{\infty} |f_n|$ converges pointwise we say the series $\sum_{n=1}^{\infty} f_n$ converges absolutely.

Remark. A uniformly convergent series of continuous functions is continuous, since each partial sum is continuous.

Weierstrass M-test. *Suppose $\sum_{n=1}^{\infty} M_n$ is a convergent series of real numbers, and $\sum_{n=1}^{\infty} f_n$ is a series of functions in $C_b^0(X)$ where X is a metric space. If $\|f_n\|_{\text{sup}} \leq M_n$ for all n, then $\sum_{n=1}^{\infty} f_n$ converges absolutely and uniformly to a bounded continuous function.*

Proof. Suppose $\epsilon > 0$ is given. Then for large enough m, k with $m > k$ we have

$$\begin{aligned} d_{\text{sup}}\Big(\sum_{n=1}^{m} f_n, \sum_{n=1}^{k} f_n\Big) &= \Big\|\sum_{n=k+1}^{m} f_n\Big\|_{\text{sup}} \\ &\leq \sum_{n=k+1}^{m} \|f_n\|_{\text{sup}} \leq \sum_{n=k+1}^{m} M_n < \epsilon, \end{aligned}$$

since $\sum_{n=1}^{\infty} M_n$ is convergent. Thus the sequence of partial sums of the series $\sum_{n=1}^{\infty} f_n$ is Cauchy in $C_b^0(X)$. Hence the series $\sum_{n=1}^{\infty} f_n$ converges uniformly to a bounded continuous function. To see that it also converges absolutely we can repeat the above argument for the series $\sum_{n=1}^{\infty} |f_n|$ noting that $\||f_n|\|_{\text{sup}} = \|f_n\|_{\text{sup}}$. ■

Theorem 6.10. *Let $f, g : (a, b) \to \mathbb{R}$. Suppose the sequence of differentiable functions $f_n : (a, b) \to \mathbb{R}$ converges pointwise to f, and the sequence (f_n') converges uniformly to g. Then f is differentiable and $f' = g$.*

Proof. Fix some $x \in (a, b)$. We have to show that $\frac{f(x+h)-f(x)}{h} - g(x)$ goes to zero when $h \to 0$. Let

$$r_n(h) := \frac{f_n(x+h) - f_n(x)}{h} - f_n'(x),$$

for h near but not equal to 0. Then (r_n) converges pointwise to $\frac{f(x+h)-f(x)}{h} - g(x)$. But by the mean value theorem applied to the function $f_m - f_n$, we have

$$\begin{aligned} r_m - r_n &= \frac{[f_m(x+h) - f_n(x+h)] - [f_m(x) - f_n(x)]}{h} + f_n'(x) - f_m'(x) \\ &= f_m'(x+ch) - f_n'(x+ch) + f_n'(x) - f_m'(x), \end{aligned}$$

for some $c \in (0, 1)$. Suppose $\epsilon > 0$ is given. Then there is $N \in \mathbb{N}$ such that $|f_n'(z) - g(z)| < \frac{\epsilon}{8}$ for all $z \in (a, b)$ and $n \geq N$. Hence for $m, n \geq N$ we have $|f_n'(z) - f_m'(z)| < \frac{\epsilon}{4}$ for all $z \in (a, b)$. Therefore for such m, n we have $|r_m(h) - r_n(h)| < \frac{\epsilon}{2}$ for all h. Now set $n = N$ and let $m \to \infty$ to get

$$\left| \frac{f(x+h) - f(x)}{h} - g(x) - r_N(h) \right| < \frac{\epsilon}{2},$$

for all h. Then for small enough h we have $|r_N(h)| < \frac{\epsilon}{2}$, since f_N is differentiable at x. Hence

$$\left| \frac{f(x+h) - f(x)}{h} - g(x) \right| < \epsilon,$$

for small enough h. As ϵ is arbitrary we get $f'(x) = g(x)$. ■

Remark. In the above theorem, we can furthermore deduce that (f_n) converges uniformly to f. To see this let $x_0 \in (a, b)$ be a fixed point. Let

$$s_n(x) := f_n(x) - f_n(x_0),$$

for $x \in (a, b)$. Then similarly to the above proof, for a given $\epsilon > 0$ there is $N \in \mathbb{N}$ such that for $m, n \geq N$ we have

$$|s_m(x) - s_n(x)| < \epsilon|x - x_0| < (b - a)\epsilon,$$

for all x. Now let $m \to \infty$ to get

$$\big|f(x) - f(x_0) - \big(f_n(x) - f_n(x_0)\big)\big| < (b-a)\epsilon,$$

for all x. Suppose $N_1 \in \mathbb{N}$ is large enough so that $|f(x_0) - f_n(x_0)| < \epsilon$ for $n \geq N_1$. Then for $n \geq \max\{N, N_1\}$ we have

$$\begin{aligned}|f(x) - f_n(x)| &\leq \big|f(x) - f(x_0) - \big(f_n(x) - f_n(x_0)\big)\big| + |f(x_0) - f_n(x_0)| \\ &< (b-a+1)\epsilon.\end{aligned}$$

Therefore (f_n) converges uniformly to f, since ϵ and x are arbitrary. ■

***Example* 6.11.** The conclusion of the above theorem does not hold without assuming the convergence of (f_n'). For example $f_n(x) = \sqrt{x^2 + 1/n}$ converges uniformly to the nondifferentiable function $f(x) = |x|$ (why?). Even if the limit function is differentiable, the limit of the derivatives need not be equal to the derivative of the limit. For example $f_n(x) = \frac{1}{n}\sin(n^2 x)$ converges uniformly to the constant function 0, but

$$f_n'(0) = n\cos(n^2 0) = n \to +\infty \neq 0.$$

Theorem 6.12. *Suppose the sequence of Riemann integrable functions $f_n : [a,b] \to \mathbb{R}$ converges uniformly to $f : [a,b] \to \mathbb{R}$. Then f is Riemann integrable and we have*

$$\lim_{n\to\infty} \int_a^b f_n(x)dx = \int_a^b f(x)dx.$$

Proof. First note that f is bounded. Because for N large enough we have $|f_N(x) - f(x)| < 1$ for all $x \in [a,b]$. But f_N is bounded being Riemann integrable. So we have $|f| < 1 + M$, where M is an upper bound for $|f_N|$. Now let D_n be the set of discontinuities of f_n. Then D_n has measure zero. Hence $D := \bigcup D_n$ has measure zero. Also each f_n is continuous on $[a,b] - D$. Therefore f is continuous on $[a,b] - D$. Thus the set of discontinuities of f has measure zero, since it is contained in D. Hence f is Riemann integrable. Finally, suppose $\epsilon > 0$ is given. Then for n large enough we have $|f_n(x) - f(x)| < \frac{\epsilon}{b-a}$. Hence

$$\left|\int_a^b f_n(x)dx - \int_a^b f(x)dx\right| = \left|\int_a^b [f_n(x) - f(x)]dx\right| \leq (b-a)\frac{\epsilon}{b-a} = \epsilon. \quad ■$$

***Exercise* 6.13.** The assumption of uniform convergence is essential in the above theorem. For example let $f_n : [0,1] \to \mathbb{R}$ be given by

$$f_n(x) = \begin{cases} nx^n & x \in [0,1), \\ 0 & x = 1. \end{cases}$$

Show that f_n converges pointwise to 0, but we have

$$\int_0^1 f_n(x)dx = \frac{n}{n+1} \to 1 \neq 0 = \int_0^1 0\,dx.$$

***Example* 6.14.** Here is another example that shows the limit of the integral is not necessarily the same as the integral of the limit, if the convergence is not uniform. Let $f_n : [0,1] \to \mathbb{R}$ for $n \geq 2$ be defined as

$$f_n(x) = \begin{cases} n^2 x & 0 \leq x \leq \frac{1}{n}, \\ 2n - n^2 x & \frac{1}{n} \leq x \leq \frac{2}{n}, \\ 0 & \frac{2}{n} \leq x \leq 1. \end{cases}$$

Then f_n converges pointwise to the constant function 0, but we have

$$\int_0^1 f_n(x)dx = 1 \not\to 0 = \int_0^1 0\,dx.$$

Theorem 6.15. *Suppose the sequence of Riemann integrable functions $f_n : [a,b] \to \mathbb{R}$ converges uniformly to $f : [a,b] \to \mathbb{R}$. Then the sequence of indefinite integrals $F_n(x) := \int_a^x f_n(t)dt$ converges uniformly to $F(x) := \int_a^x f(t)dt$.*

Proof. First note that by the last theorem f is Riemann integrable. Now suppose $\epsilon > 0$ is given. Then for large enough n we have $|f_n(t) - f(t)| < \frac{\epsilon}{b-a}$ for all $t \in [a,b]$. Thus for all $x \in [a,b]$ we have

$$|F_n(x) - F(x)| = \left| \int_a^x [f_n(t) - f(t)]dt \right| \leq (x-a)\frac{\epsilon}{b-a} \leq \epsilon. \qquad \blacksquare$$

Term by Term Differentiation and Integration. *Suppose we have a sequence of functions $f_n : [a,b] \to \mathbb{R}$.*

(i) *If each f_n is Riemann integrable and the series $\sum_{n=1}^\infty f_n$ converges uniformly, then $\sum_{n=1}^\infty f_n$ is Riemann integrable and*

$$\int_a^b \sum_{n=1}^\infty f_n(x)\,dx = \sum_{n=1}^\infty \int_a^b f_n(x)dx.$$

Furthermore, the series of indefinite integrals $\sum_{n=1}^\infty \int_a^x f_n(t)dt$ converges uniformly to $\int_a^x \sum_{n=1}^\infty f_n(t)\,dt$.

(ii) *Suppose the series $\sum_{n=1}^\infty f_n$ converges pointwise. If each f_n is differentiable on (a,b) and the series $\sum_{n=1}^\infty f_n'$ converges uniformly on (a,b), then $\sum_{n=1}^\infty f_n$ is differentiable on (a,b) and*

$$\left(\sum_{n=1}^\infty f_n\right)' = \sum_{n=1}^\infty f_n'.$$

Proof. We just need to apply the corresponding results about sequences to the sequence of partial sums. Let $S_k := \sum_{n=1}^{k} f_n$, and $S := \sum_{n=1}^{\infty} f_n$. Then in (i) the sequence (S_k) converges uniformly to S. Hence S is Riemann integrable, and we have

$$\begin{aligned}\int_a^b \sum_{n=1}^{\infty} f_n(x)\,dx &= \int_a^b S(x)dx = \lim_{k\to\infty}\int_a^b S_k(x)dx\\ &= \lim_{k\to\infty}\sum_{n=1}^{k}\int_a^b f_n(x)dx = \sum_{n=1}^{\infty}\int_a^b f_n(x)dx.\end{aligned}$$

The cases of indefinite integrals and derivatives are similar. ■

Theorem 6.16. *There exists a continuous function $f : \mathbb{R} \to \mathbb{R}$ that is nowhere differentiable, i.e. it is not differentiable at any point.*

Proof. Let $\phi(x) := |x|$ for $x \in [-1, 1]$. Extend ϕ to a 2-periodic function on all of $\mathbb{R}$ by setting $\phi(x+2k) := \phi(x)$ for $x \in [-1, 1]$ and $k \in \mathbb{Z}$. Note that ϕ is continuous, and $0 \le \phi \le 1$. Also note that ϕ equals the linear function

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

on the interval $[k, k + 1]$ where $k \in \mathbb{Z}$. In addition, for all $x, y \in \mathbb{R}$ we have

$$|\phi(x) - \phi(y)| \le |x - y|.$$

Because if $|x - y| \ge 1$ or $x, y \in [k, k + 1]$ for some $k \in \mathbb{Z}$, then the inequality holds obviously. Otherwise, we can choose an integer k between x, y, and compare the values of ϕ at these three points, to get the desired inequality. Furthermore, note that when $x, y \in [k, k + 1]$ for some $k \in \mathbb{Z}$, the inequality becomes an equality.

We claim that

$$f(x) := \sum_{n=0}^{\infty}\big(\frac{3}{4}\big)^n\phi(4^n x)$$

is a continuous function that is not differentiable at any x. First note that the numerical series $\sum_{n\ge 0}(\frac{3}{4})^n$ is convergent, so by the M-test of Weierstrass, the above series converges uniformly. Therefore f is continuous.

Now note that $\phi(4^n x)$ is a periodic function of period $\frac{2}{4^n}$. Let $\delta_m = \frac{2}{4^{m+1}}$. Then for a fixed x and every $n > m$ we have $\phi(4^n(x \pm \delta_m)) = \phi(4^n x)$, since $\pm 4^n\delta_m = \pm 4^{n-m-1}\cdot 2$ is an integer multiple of the period of ϕ. We also have $4^m\delta_m = \frac{1}{2}$. Thus, depending on $4^m x$ we can choose $h_m = \delta_m$ or $h_m = -\delta_m$ so that $4^m(x + h_m)$ and $4^m x$ belong to an interval of the form $[k, k+1]$ where $k \in \mathbb{Z}$. Then we have

$$|\phi(4^m(x + h_m)) - \phi(4^m x)| = |4^m h_m| = 4^m\delta_m = \frac{1}{2}.$$

Therefore we have

$$\begin{aligned}|f(x+h_m)-f(x)| &= \Big|\sum_{n\geq 0}(\frac{3}{4})^n(\phi(4^n(x+h_m))-\phi(4^nx))\Big| \\ &= \Big|(\frac{3}{4})^m(\phi(4^m(x+h_m))-\phi(4^mx)) \\ &\qquad +\sum_{n<m}(\frac{3}{4})^n(\phi(4^n(x+h_m))-\phi(4^nx))\Big| \\ &\geq (\frac{3}{4})^m\frac{1}{2}-\sum_{n<m}(\frac{3}{4})^n|4^nh_m| \\ &= \frac{1}{2\cdot 4^m}(3^m-\frac{3^m-1}{3-1}) = \frac{3^m+1}{4^{m+1}}.\end{aligned}$$

Hence

$$\left|\frac{f(x+h_m)-f(x)}{h_m}\right| \geq \frac{3^m+1}{2} \underset{m\to\infty}{\longrightarrow} \infty.$$

But $h_m \to 0$ as $m\to\infty$. So f cannot be differentiable at x, since otherwise $\frac{1}{h_m}(f(x+h_m)-f(x))$ would have converged to $f'(x)$. ■

6.2 Power Series

Definition 6.17. A **power series** is a series of real-valued functions of a real variable that has the form

$$\sum_{n=0}^{\infty} c_n(x-a)^n.$$

The numbers c_n are the coefficients of the power series, and a is its center.

Remark. When we deal with power series, we use the convention $0^0=1$.

Theorem 6.18. *For every power series $\sum_{n=0}^{\infty} c_n(x-a)^n$ there is $R\in[0,\infty]$, called its* ***radius of convergence****, such that the power series converges absolutely for $|x-a|<R$ and diverges for $|x-a|>R$. Furthermore*

$$R=\frac{1}{\limsup\sqrt[n]{|c_n|}}.$$

Proof. By the root test, the convergence and divergence are determined by the value

$$\begin{aligned}\limsup\sqrt[n]{|c_n(x-a)^n|} &= \limsup|x-a|\sqrt[n]{|c_n|} \\ &= |x-a|\limsup\sqrt[n]{|c_n|} = \frac{|x-a|}{R}.\end{aligned}$$ ■

Remark. The interval $(a-R, a+R)$ is called the **interval of convergence** of the power series.

Theorem 6.19. *Suppose the power series* $f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$ *has radius of convergence* $R > 0$. *Then the power series converges uniformly on* $[a-r, a+r]$ *for every* $r < R$.

Proof. Let $S_n(x)$ be the nth partial sum of the power series. Then for $m > n$ and $|x| \leq r$ we have

$$|S_m(x) - S_n(x)| = \left| \sum_{i=n+1}^{m} c_i x^i \right| \leq \sum_{i=n+1}^{m} |c_i| r^i.$$

But the power series is absolutely convergent at $x = r$, so $\sum_{i=n+1}^{m} |c_i| r^i \to 0$ as $m, n \to \infty$. Thus for a given $\epsilon > 0$ we can find N such that for $m > n \geq N$ we have $\sum_{i=n+1}^{m} |c_i| r^i < \epsilon$. Hence $|S_m(x) - S_n(x)| < \epsilon$. Now if we let $m \to \infty$, then for $n \geq N$ and $|x| \leq r$ we have $|f(x) - S_n(x)| \leq \epsilon$. ■

Remark. The above theorem implies that power series are continuous inside their interval of convergence, since their partial sums are continuous polynomials.

Proposition 6.20. *Suppose* $(a_n), (b_n)$ *are sequences in* $\mathbb{R}$, *and* a_n *is convergent with* $\lim a_n > 0$. *Then*

$$\limsup a_n b_n = \lim a_n \limsup b_n.$$

Proof. Let $a = \lim a_n$, $b = \limsup b_n$ and $c = \limsup a_n b_n$. We know that the lim sup of any sequence is the limit of some subsequence of it. We also know that the limit of any subsequence of a sequence is less than or equal to the lim sup of that sequence. Now suppose $b_{n_k} \to b$. Then $a_{n_k} b_{n_k} \to ab$. Hence $ab \leq c$. Conversely, suppose $a_{n_l} b_{n_l} \to c$. Then for large enough l we must have $a_{n_l} > 0$, since $a > 0$. Therefore we have

$$b_{n_l} = \frac{a_{n_l} b_{n_l}}{a_{n_l}} \to \frac{c}{a}.$$

Thus $\frac{c}{a} \leq b$, or $c \leq ab$. So $c = ab$ as desired. ■

Theorem 6.21. *Suppose the power series* $f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$ *has radius of convergence* $R > 0$. *Then* f *is differentiable on* $(a-R, a+R)$ *and we have*

$$f'(x) = \sum_{n=1}^{\infty} n c_n (x-a)^{n-1},$$

$$\int_a^x f(t)dt = \sum_{n=0}^{\infty} \frac{c_n}{n+1}(x-a)^{n+1},$$

for all $x \in (a-R, a+R)$. In addition, the radii of convergence of the power series of $f'(x)$ and $\int_a^x f(t)dt$ are also R.

Remark. In other words, the derivative and the indefinite integral of a power series can be computed term by term inside its interval of convergence.

Proof. For $x \neq a$ we have

$$\begin{aligned}\limsup \sqrt[n]{|nc_n(x-a)^{n-1}|} &= \limsup |x-a|^{\frac{n-1}{n}} \sqrt[n]{n|c_n|} \\ &= |x-a| \lim \left(|x-a|^{\frac{-1}{n}} \sqrt[n]{n}\right) \limsup \sqrt[n]{|c_n|} \\ &= |x-a| \limsup \sqrt[n]{|c_n|} = \frac{|x-a|}{R}.\end{aligned}$$

Thus by the root test, the radius of convergence of the claimed power series of f' is R. Now for some fixed $x \in (a-R, a+R)$ choose $r < R$ so that $x \in [a-r, a+r]$. We know that the power series of f and f' converge uniformly on $[a-r, a+r]$. Also each term of the power series of f, i.e. $c_n(x-a)^n$, is obviously differentiable. Therefore using term by term differentiation we obtain

$$f'(x) = \sum_{n=0}^{\infty} \big(c_n(x-a)^n\big)' = \sum_{n=1}^{\infty} nc_n(x-a)^{n-1}.$$

Similarly, using term by term integration we have

$$\int_a^x f(t)dt = \sum_{n=0}^{\infty} \int_a^x c_n(t-a)^n dt = \sum_{n=0}^{\infty} \frac{c_n}{n+1}(x-a)^{n+1},$$

since each term of the power series of $f(t)$, i.e. $c_n(t-a)^n$, is obviously Riemann integrable.

Finally, note that the radius of convergence of the power series of $\int_a^x f(t)dt$ is also R. Because if it has a bigger radius of convergence, then we can differentiate its power series term by term to conclude that the radius of convergence of the power series of f is bigger R, which is a contradiction. We can also directly compute the radius of convergence of $\int_a^x f(t)dt$ by noting that

$$\begin{aligned}\limsup \sqrt[n]{\left|\frac{c_n}{n+1}(x-a)^{n+1}\right|} &= \limsup |x-a|^{\frac{n+1}{n}} \sqrt[n]{\frac{|c_n|}{n+1}} \\ &= |x-a| \lim \left(|x-a|^{\frac{1}{n}} \frac{1}{\sqrt[n]{n+1}}\right) \limsup \sqrt[n]{|c_n|} \\ &= |x-a| \limsup \sqrt[n]{|c_n|} = \frac{|x-a|}{R},\end{aligned}$$

and employing the root test. ■

Theorem 6.22. *A power series $f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$ with positive radius of convergence is infinitely differentiable inside its interval of convergence and we have*

$$c_n = \frac{f^{(n)}(a)}{n!}.$$

Proof. Since f' is also a power series with the same radius of convergence as f, f'' is also a power series with the same radius of convergence. We can inductively show that each derivative $f^{(k)}$ exists and is a power series with the same radius of convergence as f. Moreover, we can show by induction that

$$f^{(k)}(x) = \sum_{n=k}^{\infty}[n(n-1)\cdots(n-k+1)]c_n(x-a)^{n-k}.$$

Hence we have $f^{(k)}(a) = k(k-1)\cdots(k-k+1)c_k = k!\,c_k$ as desired. ■

Remark. An interesting consequence of the above theorem is that if

$$\sum_{n=0}^{\infty}c_n(x-a)^n = \sum_{n=0}^{\infty}b_n(x-a)^n$$

on an open interval around a, then we have $c_n = b_n$ for all n.

Definition 6.23. Suppose $f : (b,c) \to \mathbb{R}$ is smooth, and $a \in (b,c)$. The **Taylor series** of f at a is the power series

$$\sum_{n=0}^{\infty}\frac{f^{(n)}(a)}{n!}(x-a)^n.$$

Remark. Note that we do not assume anything about the convergence of the Taylor series. It can be divergent for all $x \neq a$. Even if the Taylor series converges, its limit is not necessarily f.

Definition 6.24. A function $f : (b,c) \to \mathbb{R}$ is called **analytic** or of **class C^ω** if it can be expressed locally as a convergent power series, i.e. for every $a \in (b,c)$ there exists a power series $\sum_{n=0}^{\infty} c_n(x-a)^n$ with positive radius of convergence such that for x near a we have

$$f(x) = \sum_{n=0}^{\infty}c_n(x-a)^n.$$

Remark. Note that by Theorem 6.22, an analytic function is smooth and we have $c_n = \frac{f^{(n)}(a)}{n!}$. Therefore if a function can be expressed as a power series, that power series is the Taylor series of the function.

Remark. Power series with positive radius of convergence are analytic inside their interval of convergence. Note that this fact is not trivial, because it is not obvious that a power series around the center a can be expressed (locally) as a power series around another center b.

***Example* 6.25.** There are smooth functions that are not analytic. For example the function

$$f(x) := \begin{cases} e^{-\frac{1}{x}} & x > 0 \\ 0 & x \le 0 \end{cases}$$

is smooth, but is not analytic. It is obvious that f is smooth on the set $\{x \neq 0\}$. We show by induction that f is infinitely differentiable at 0, and we have $f^{(n)}(0) = 0$ for all n. Then it follows that the Taylor series of f at 0 is identically zero, which differs from f for $x > 0$. So f is not analytic. Note that the Taylor series of f at 0 is convergent on $\mathbb{R}$, but its value does not equal f. Now we claim that for each $n \in \mathbb{N}$ there is a polynomial p_n such that

$$f^{(n)}(x) = \begin{cases} p_n(\frac{1}{x})e^{-\frac{1}{x}} & x > 0, \\ 0 & x \le 0. \end{cases}$$

This holds trivially for $n = 0$, and also for $x < 0$ and all n. Suppose the claim is true for some n. Then for $n + 1$, and $x > 0$ we have

$$f^{(n+1)}(x) = \frac{-1}{x^2}p_n'(\frac{1}{x})e^{-\frac{1}{x}} + p_n(\frac{1}{x})(\frac{1}{x^2})e^{-\frac{1}{x}} = p_{n+1}(\frac{1}{x})e^{-\frac{1}{x}},$$

where $p_{n+1}(t) := -t^2p_n'(t) + t^2p_n(t)$. Also for $x = 0$ we have $\lim_{x\to 0^-} \frac{f^{(n)}(x)-0}{x-0} = 0$, and

$$\begin{aligned} \lim_{x\to 0^+} \frac{f^{(n)}(x) - 0}{x - 0} &= \lim_{x\to 0^+} \frac{1}{x}p_n(\frac{1}{x})e^{-\frac{1}{x}} \\ &= \lim_{t\to+\infty} tp_n(t)e^{-t} = \lim_{t\to+\infty} \frac{tp_n(t)}{e^t}. \qquad (t := \tfrac{1}{x}) \end{aligned}$$

Note that $\lim_{t\to+\infty} e^t = +\infty$, so if we apply the L'Hôpital's rule repeatedly we get

$$\lim_{t\to+\infty} \frac{q_n(t)}{e^t} = \lim_{t\to+\infty} \frac{q_n'(t)}{e^t} = \lim_{t\to+\infty} \frac{q_n''(t)}{e^t} = \cdots = \lim_{t\to+\infty} \frac{q_n^{(m+1)}(t)}{e^t} = \lim_{t\to+\infty} \frac{0}{e^t} = 0,$$

where $q_n(t) = tp_n(t)$ and $m = \deg q_n$. Hence $f^{(n+1)}(0) = 0$ as desired. ■

6.3 Exponential and Logarithmic Functions

Theorem 6.26. *The radius of convergence of the power series of the* ***exponential function***

$$\exp(x) := \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

is infinity. Furthermore we have $\exp' = \exp$.

Proof. Apply the ratio test. Then use term by term differentiation. ■

Remark. Note that $\exp(0) = 1$.

Theorem 6.27. *The exponential function* exp *is positive and strictly increasing. It also satisfies*

$$\begin{aligned} \exp(x+y) &= \exp(x)\exp(y), \\ \exp(-x) &= \frac{1}{\exp(x)}, \end{aligned}$$

for all $x, y \in \mathbb{R}$. *Furthermore we have*

$$\lim_{x\to-\infty} \exp(x) = 0, \qquad \lim_{x\to+\infty} \exp(x) = +\infty.$$

Proof. Let

$$\begin{aligned} f(x) &:= \exp(x+y) - \exp(x)\exp(y), \\ g(x) &:= f(x)\exp(-x). \end{aligned}$$

Then $f(0) = 0$, and $f' = f$. Also $g(0) = 0$, and $g' \equiv 0$. Hence $g \equiv 0$. Now, it is obvious from the definition that $\exp(x) \geq 1$ for $x \geq 0$. Hence for $x \leq 0$ we have

$$\exp(-x) > 0 \implies f(x) = 0.$$

Suppose $y \geq 0$, and set $x = -y$. Then we have $f(-y) = 0$. Thus

$$1 = \exp(0) = \exp(-y+y) = \exp(-y)\exp(y).$$

Since $\exp(y) > 0$, we obtain $\exp(-y) > 0$. Therefore exp is positive everywhere. Consequently $f \equiv 0$. Thus the first identity holds for all x, y. The second identity follows easily since for all $x \in \mathbb{R}$ we have $\exp(-x)\exp(x) = \exp(-x+x) = \exp(0) = 1$. Finally note that $\exp' = \exp > 0$. So exp is strictly increasing.

Next we obtain the limits. We can easily deduce from the definition that $\exp(x) > 1 + x$ for $x > 0$. This gives the limit at $+\infty$. For the limit at $-\infty$ we use the inequality

$$0 < \exp(x) = \frac{1}{\exp(-x)} < \frac{1}{1-x},$$

for $x < 0$. ■

Remark. In the above proof, from $f' = f$ and $f(0) = 0$, we concluded that $f \equiv 0$. This is a standard technique in differential equations. The usual trick is to multiply f by $\exp(-x)$, and then to differentiate the product. Here we had to be more careful, since we have not proved the needed properties of the exponential function yet.

Definition 6.28. We define the **logarithm** of $x > 0$ to be

$$\log(x) := \int_1^x \frac{1}{t} dt.$$

Remark. Note that $\log(1) = 0$.

Theorem 6.29. *For all $x, y > 0$ we have*

$$\begin{aligned} \log'(x) &= \frac{1}{x}, \\ \log(xy) &= \log(x) + \log(y), \\ \log\left(\frac{1}{x}\right) &= -\log(x). \end{aligned}$$

In addition, log *is strictly increasing. Furthermore*

$$\lim_{x\to 0^+} \log(x) = -\infty, \qquad \lim_{x\to+\infty} \log(x) = +\infty.$$

Proof. By the fundamental theorem of calculus we have $\log'(x) = \frac{1}{x}$ for $x > 1$. For $x \leq 1$ we have

$$\log(x) = \int_1^x \frac{1}{t} dt = -\int_x^1 \frac{1}{t} dt = \int_c^x \frac{1}{t} dt - \int_c^1 \frac{1}{t} dt,$$

for some fixed $c \in (0, x)$. Again, we obtain the desired formula by the fundamental theorem of calculus. Now it is obvious that log is strictly increasing, since $\frac{1}{x} > 0$ for $x > 0$.

To prove the second identity, we fix some $y > 0$. Then $(\log(xy))' = \frac{y}{xy} = \frac{1}{x}$. So for

$$f(x) := \log(xy) - \log(x) - \log(y)$$

we have $f'(x) \equiv 0$. But $f(1) = 0$, hence $f \equiv 0$ as desired. The third identity follows easily because $\log(x) + \log(\frac{1}{x}) = \log(x\frac{1}{x}) = \log(1) = 0$.

Now we obtain the limits. Note that $\log(2) > \log(1) = 0$. We also have

$$\begin{aligned}\log(2^n) &= \log(2 \times 2^{n-1}) = \log(2) + \log(2^{n-1}) \\ &= 2\log(2) + \log(2^{n-2}) = \cdots = n\log(2).\end{aligned}$$

Now suppose $M > 0$ is given. Then there is a positive integer $n > \frac{M}{\log(2)}$. So for $x > 2^n$ we have

$$\log(x) > \log(2^n) = n\log(2) > M.$$

Hence $\log(x) \to +\infty$ as $x \to +\infty$. To obtain the limit as $x \to 0^+$ we can argue similarly noting that $\log(\frac{1}{2^n}) = -n\log(2)$, and $\frac{1}{2^n} \to 0$. ■

Second Proof. To prove the second identity, we fix some $x, y > 0$. Then we have

$$\log(xy) = \int_1^{xy} \frac{1}{t}dt = \int_1^{x} \frac{1}{t}dt + \int_x^{xy} \frac{1}{t}dt = \log(x) + \int_x^{xy} \frac{1}{t}dt.$$

Thus it suffices to show that $\int_x^{xy} \frac{1}{t}dt = \log(y)$. Let $g(s) := xs$ for $s > 0$. Then by changing the variable we get

$$\int_x^{xy} \frac{1}{t}dt = \int_{g(1)}^{g(y)} \frac{1}{t}dt = \int_1^{y} \frac{1}{g(s)}g'(s)ds = \int_1^{y} \frac{1}{xs}xds = \int_1^{y} \frac{1}{s}ds = \log(y).$$

■

Theorem 6.30. *The logarithm* $\log : (0, +\infty) \to \mathbb{R}$ *is the inverse of the exponential function* $\exp : \mathbb{R} \to (0, +\infty)$*. Consequently, both* $\log$ *and* $\exp$ *are one-to-one and onto.*

Proof. We have

$$[\log(\exp(x))]' = \frac{1}{\exp(x)}\exp'(x) = \frac{1}{\exp(x)}\exp(x) = 1 = x'.$$

Also, we have $\log(\exp(0)) = \log(1) = 0$. Thus $\log(\exp(x)) = x$. This implies that log is onto. Therefore log is the inverse of exp, since log is also one-to-one as it is strictly increasing. ■

Theorem 6.31. *For* $|x| < 1$ *we have* $\log(1 + x) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}x^n$.

Proof. For $|t| < 1$ we have $\frac{1}{1+t} = \frac{1}{1-(-t)} = \sum_{k=0}^{\infty}(-t)^k = \sum_{k=0}^{\infty}(-1)^k t^k$. Thus

$$\int_0^x \frac{1}{1+t}dt = \sum_{k=0}^{\infty} \frac{(-1)^k}{k+1}x^{k+1} = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n}x^n.$$

But $(\log(1+x))' = \frac{1}{1+x}$, so we get the desired result by the fundamental theorem of calculus. ■

Theorem 6.32. $\log, \exp$ *are analytic. As a result, they are both smooth.*

Proof. We have to show that $\log, \exp$ can be expressed locally as a convergent power series around any point in their domains. We do this by employing the functional identities that they satisfy. First consider exp. Let $a \in \mathbb{R}$ be an arbitrary point. Then we have

$$\begin{aligned}\exp(x) &= \exp(a + x - a) = \exp(a)\exp(x-a)\\ &= \exp(a)\sum_{n=0}^{\infty}\frac{(x-a)^n}{n!} = \sum_{n=0}^{\infty}\frac{\exp(a)}{n!}(x-a)^n.\end{aligned}$$

Note that the radius of convergence of this power series is infinity. Now consider log. Let $a > 0$. Then for $|x - a| < a$ we have

$$\begin{aligned}\log(x) &= \log(a + x - a) = \log\big(a(1 + \frac{x-a}{a})\big)\\ &= \log(a) + \log\big(1 + \frac{x-a}{a}\big) = \log(a) + \sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{na^n}(x-a)^n.\end{aligned}$$

■

Real Exponents.

Theorem 6.33. *Let* $x \in (0, \infty)$. *Then for all* $r \in \mathbb{R}$ *we have*

$$\log x^r = r\log x.$$

Proof. Note that $x^r > 0$ when $x > 0$. First suppose that r is rational. When $r = 0$ the equality holds trivially, since $\log 1 = 0$. Next we show by induction that for all positive integers n, the equality holds when $r = n$. The case of $n = 1$ is trivial. For the induction step we have

$$\log x^{n+1} = \log(x^n x) = \log x^n + \log x = n\log x + \log x = (n+1)\log x.$$

Now if $r = \frac{k}{m} \in \mathbb{Q}$ for some positive integers m, k, we have

$$k\log x = \log x^k = \log\big(\sqrt[m]{x^k}\big)^m = m\log\sqrt[m]{x^k} \implies \log x^{\frac{k}{m}} = \frac{k}{m}\log x.$$

Then, suppose $r = -q < 0$ where $q \in \mathbb{Q}$. Then

$$\log x^r = \log x^{-q} = \log\frac{1}{x^q} = -\log x^q = -q\log x = r\log x.$$

Finally for an arbitrary $r \in \mathbb{R}$ we have

$$x^r = \begin{cases} \sup\{x^p : p \in \mathbb{Q}, p \le r\} & x \ge 1, \\ \inf\{x^p : p \in \mathbb{Q}, p \le r\} & 0 < x < 1. \end{cases}$$

Let p, q be rational numbers. When $x = 1$ the desired equality holds trivially, since $1^r = 1$. When $x > 1$ we have $p \log x = \log x^p \le \log x^r$ for all $p \le r$, since log is increasing. Thus $r \log x \le \log x^r$. Because $\log x > \log 1 = 0$, so

$$p \le \frac{\log x^r}{\log x} \text{ for all } p \le r \implies r \le \frac{\log x^r}{\log x}.$$

On the other hand we know that for all $q \ge r$ we have $x^q \ge x^p$, for every $p \le r$. Hence $x^q \ge x^r$. Therefore $q \log x = \log x^q \ge \log x^r$. Thus $r \log x \ge \log x^r$, and we get the desired equality. The case of $0 < x < 1$ is similar. Just note that in this case $\log x < 0$, so we have to reverse the inequalities when we divide and multiply by $\log x$. ■

Theorem 6.34. *For all $x, y > 0$ and all $r, s \in \mathbb{R}$ we have*
(i) $x^r = \exp(r \log x)$.
(ii) $x^r x^s = x^{r+s}$.
(iii) $(x^r)^s = x^{rs}$.
(iv) $(xy)^r = x^r y^r$.

Proof. We have

$$\begin{aligned} x^r &= \exp(\log x^r) = \exp(r \log x), \\ x^r x^s &= \exp(r \log x) \exp(s \log x) \\ &= \exp((r+s) \log x) = x^{r+s}, \\ (x^r)^s &= \exp(s \log x^r) = \exp(rs \log x) = x^{rs}, \\ (xy)^r &= \exp(r \log xy) = \exp(r(\log x + \log y)) \\ &= \exp(r \log x) \exp(r \log y) = x^r y^r. \end{aligned}$$
■

Theorem 6.35. *Suppose $a, r \in \mathbb{R}$, and $a > 0$. Then the functions $x \mapsto x^r$ and $x \mapsto a^x$ are smooth functions on $(0, \infty)$ and $\mathbb{R}$, respectively. In addition we have*

$$(x^r)' = r x^{r-1}, \qquad (a^x)' = (\log a) a^x.$$

As a result, x^r is strictly increasing on $(0, \infty)$ when $r > 0$, and it is strictly decreasing when $r < 0$. Also, a^x is strictly increasing on $\mathbb{R}$ when $a > 1$, and it is strictly decreasing when $0 < a < 1$.

Proof. Note that $x^r = \exp(r \log x)$ and $a^x = \exp(x \log a)$. Hence both x^r and a^x are compositions of smooth functions, so they are smooth. Now we have

$$\begin{aligned}(x^r)' &= [\exp(r\log x)]' = \frac{r}{x}\exp(r\log x) = rx^{-1}x^r = rx^{r-1},\\ (a^x)' &= [\exp(x\log a)]' = (\log a)\exp(x\log a) = (\log a)a^x.\end{aligned}$$

The last part of the theorem follows from the fact that the sign of the derivative determines the monotonicity of the function. Note that the powers of a positive number are always positive. Also note that $\log a > 0 = \log 1$ if and only if $a > 1$. ■

Remark. Suppose f, g are differentiable functions and $f > 0$. Then the derivative of the function f^g can be computed as follows. We have

$$\begin{aligned}\left[f(x)^{g(x)}\right]' &= \left[\exp\left(g(x)\log f(x)\right)\right]'\\ &= \left[g'(x)\log f(x) + g(x)\frac{f'(x)}{f(x)}\right]\exp\left(g(x)\log f(x)\right)\\ &= \left[g'(x)\log f(x) + g(x)\frac{f'(x)}{f(x)}\right]f(x)^{g(x)}.\end{aligned}$$

This is a case of the so-called *logarithmic differentiation.*

Definition 6.36. $e := \exp(1) = \sum_{n=0}^{\infty}\frac{1}{n!}$.

Remark. Note that $e > \sum_{n=0}^{2}\frac{1}{n!} = 2.5$. We also have

$$\log 3 = \int_1^3 \frac{dt}{t} = \sum_{n=4}^{11}\int_{n/4}^{(n+1)/4}\frac{dt}{t} \geq \sum_{n=4}^{11}\frac{1}{4}\frac{4}{n+1} = \frac{1}{5} + \cdots + \frac{1}{12} > 1.$$

Hence $e = \exp(1) < \exp(\log 3) = 3$. Thus $2.5 < e < 3$.

Theorem 6.37. *For all $x \in \mathbb{R}$ we have*

$$\exp(x) = e^x.$$

Proof. Note that $\log e = 1$. Thus $e^x = \exp(x\log e) = \exp(x)$. ■

Theorem 6.38. *We have*

$$e = \lim_{h\to 0^+}(1+h)^{\frac{1}{h}} = \lim_{x\to+\infty}\left(1+\frac{1}{x}\right)^x.$$

Proof. Note that we have

$$1 = \log'(1) = \lim_{h\to 0^+}\frac{\log(1+h)-\log(1)}{h} = \lim_{h\to 0^+}\log(1+h)^{\frac{1}{h}}.$$

Hence as exp is continuous we get

$$\begin{aligned} e = \exp(1) &= \exp\big(\lim_{h\to0^+} \log(1+h)^{\frac{1}{h}}\big) \\ &= \lim_{h\to0^+} \exp\big(\log(1+h)^{\frac{1}{h}}\big) = \lim_{h\to0^+} (1+h)^{\frac{1}{h}}. \end{aligned}$$

By changing the variable to $x = \frac{1}{h}$ in this limit, we also get the second limit. ■

Theorem 6.39. *Suppose $a, r \in \mathbb{R}$, and $a > 0$. Then we have*

$$\begin{array}{rll} \textit{If } r > 0 \textit{ then} & \lim_{x\to0^+} x^r = 0, & \lim_{x\to+\infty} x^r = +\infty, \\ \textit{If } r < 0 \textit{ then} & \lim_{x\to0^+} x^r = +\infty, & \lim_{x\to+\infty} x^r = 0, \\ \textit{If } a > 1 \textit{ then} & \lim_{x\to-\infty} a^x = 0, & \lim_{x\to+\infty} a^x = +\infty, \\ \textit{If } 0 < a < 1 \textit{ then} & \lim_{x\to-\infty} a^x = +\infty, & \lim_{x\to+\infty} a^x = 0. \end{array}$$

Proof. We know that $x^r = \exp(r\log x)$. Hence we can compute all the above limits by using Theorem 3.17, and the value of the limits of exp and log. The same is true for $a^x = \exp(x\log a)$. Note that the sign of r and $\log a$ affect the value of the limits. ■

Theorem 6.40. *For all $r > 0$ and all $a > 1$ we have*

$$\lim_{x\to+\infty} \frac{a^x}{x^r} = +\infty, \qquad \lim_{x\to+\infty} \frac{\log x}{x^r} = 0.$$

Remark. The above limits mean that when $x \to +\infty$, a^x grows faster than any power of x, and any power of x grows faster than $\log x$.

Proof. Suppose n is an integer greater than r. Then it is obvious from the series expansion of $\exp(x)$ that for $x > 0$ we have

$$a^x = \exp(x\log a) > \frac{(\log a)^{n+1}}{(n+1)!}x^{n+1}.$$

Note that $\log a$ is positive too. Hence $\frac{a^x}{x^n} > \frac{(\log a)^{n+1}}{(n+1)!}x \to \infty$ as $x \to \infty$. Therefore

$$\lim_{x\to\infty} \frac{a^x}{x^r} = \lim_{x\to\infty} x^{n-r}\frac{a^x}{x^n} = \infty \cdot \infty = \infty.$$

For the second limit we can use the L'Hôpital's rule to obtain

$$\lim_{x\to\infty} \frac{\log x}{x^r} = \lim_{x\to\infty} \frac{\frac{1}{x}}{rx^{r-1}} = \lim_{x\to\infty} \frac{1}{rx^r} = \frac{1}{\infty} = 0.$$ ■

Theorem 6.41. *Suppose $a, b > 0$, and $b \neq 1$. Then $b^{\frac{\log a}{\log b}} = a$.*

Proof. Note that $\log b \neq 0$. Then we have

$$b^{\frac{\log a}{\log b}} = \exp\Big(\frac{\log a}{\log b}\log b\Big) = \exp(\log a) = a. \qquad ■$$

Remark. Suppose $a, b > 0$, and $b \neq 1$. We can define $\log_b a$ to be the unique positive real number such that $b^{\log_b a} = a$. Then by the above theorem we have $\log_b a = \frac{\log a}{\log b}$.

Definition 6.42. Suppose $r \in \mathbb{R}$ and $n \in \mathbb{N}$. The number

$$\binom{r}{n} := \frac{r(r-1)\cdots(r-n+1)}{n!}$$

is called a **binomial coefficient**. We also set $\binom{r}{0} := 1$.

Remark. Note that when $r = m$ is a positive integer greater than n, we have

$$\frac{m\cdots(m-n+1)}{n!} = \frac{m\cdots(m-n+1)(m-n)\cdots 1}{n!(m-n)!} = \frac{m!}{n!(m-n)!}.$$

So this new notion of binomial coefficient agrees with the old one. (When $n = 0$, or $m = n$, the equality of the two values is also obvious.) Also note that when m, n are positive integers such that $m < n$ then $\binom{m}{n} = 0$.

Proposition 6.43. *For all $r \in \mathbb{R}$ and $n \in \mathbb{N}$ we have*

$$\binom{r}{n} = \frac{r}{n}\binom{r-1}{n-1},$$
$$\binom{r-1}{n} + \binom{r-1}{n-1} = \binom{r}{n}.$$

Proof. We have

$$\frac{r}{n}\binom{r-1}{n-1} = \frac{r}{n}\,\frac{(r-1)\cdots(r-n+1)}{(n-1)!} = \frac{r(r-1)\cdots(r-n+1)}{n!} = \binom{r}{n}.$$

We also have

$$\begin{aligned}\binom{r-1}{n} + \binom{r-1}{n-1} &= \frac{(r-1)\cdots(r-n)}{n!} + \frac{(r-1)\cdots(r-n+1)}{(n-1)!}\\ &= \frac{(r-1)\cdots(r-n+1)}{(n-1)!}\Big(\frac{r-n}{n} + 1\Big)\\ &= \frac{(r-1)\cdots(r-n+1)}{(n-1)!}\,\frac{r}{n}\\ &= \frac{r(r-1)\cdots(r-n+1)}{n!} = \binom{r}{n}.\end{aligned}$$

■

Binomial Series. *Suppose $r \in \mathbb{R}$. Then for $|x| < 1$ we have*

$$(1+x)^r = \sum_{n=0}^{\infty} \binom{r}{n} x^n.$$

Proof. Let $f(x) := \sum_{n=0}^{\infty} \binom{r}{n} x^n$. We have

$$\left| \frac{\binom{r}{n+1} x^{n+1}}{\binom{r}{n} x^n} \right| = \left| \frac{r-n}{n+1} \right| |x| \xrightarrow[n\to\infty]{} |x|.$$

Thus by the ratio test the power series is absolutely convergent for $|x| < 1$, i.e. the radius of convergence of the power series is at least 1. When $r = 0$ the result holds trivially. So suppose $r \neq 0$. Then for $|x| < 1$ we have

$$f'(x) = \sum_{n=1}^{\infty} n \binom{r}{n} x^{n-1} = \sum_{n=1}^{\infty} r \binom{r-1}{n-1} x^{n-1}.$$

Thus

$$\begin{aligned}
\frac{(1+x)}{r} f'(x) &= \sum_{n=1}^{\infty} \binom{r-1}{n-1} x^{n-1} + \sum_{n=1}^{\infty} \binom{r-1}{n-1} x^n \\
&= 1 + \sum_{n=1}^{\infty} \binom{r-1}{n} x^n + \sum_{n=1}^{\infty} \binom{r-1}{n-1} x^n \\
&\qquad \text{(We replaced } n-1 \text{ with } n \text{ in the 1st sum.)} \\
&= 1 + \sum_{n=1}^{\infty} \left[\binom{r-1}{n} + \binom{r-1}{n-1} \right] x^n = 1 + \sum_{n=1}^{\infty} \binom{r}{n} x^n = f(x).
\end{aligned}$$

Hence we have (note that $(1+x)^r > 0$ for $|x| < 1$)

$$\left[\frac{f(x)}{(1+x)^r} \right]' = \frac{f'(x)(1+x)^r - rf(x)(1+x)^{r-1}}{(1+x)^{2r}} = 0.$$

In addition $f(0) = 1 = (1+0)^r$. Therefore $f(x) = (1+x)^r$ for all $|x| < 1$. ■

6.4 Trigonometric Functions

Theorem 6.44. *The radii of convergence of the power series of* ***sine*** *and* ***cosine***

$$\sin(x) := \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1},$$

$$\cos(x) := \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n},$$

are infinity. Furthermore we have $\sin' = \cos$ *and* $\cos' = -\sin$.

Proof. Apply the ratio test. Then use term by term differentiation. ■

Remark. Note that $\sin(0) = 0$ and $\cos(0) = 1$. It is also obvious that

$$\sin(-x) = -\sin(x),$$
$$\cos(-x) = \cos(x).$$

Remark. We usually consider $\cos x$ to be the power series $\sum_{n\geq 0} a_n x^n$, where $a_{2n-1} = 0$ and $a_{2n} = \frac{(-1)^n}{(2n)!}$. But sometimes we prefer to not have the zero coefficients, for example when we apply the ratio test in the above proof. In these cases we consider $\cos x$ to be the numerical series $\sum_{n\geq 0} b_n (x^2)^n$, where $b_n = \frac{(-1)^n}{(2n)!}$. It is easy to see that the two characterizations of cos are identical, simply by noting that the partial sums of the two series have the same value at every x. We also have similar considerations regarding $\sin x$.

Theorem 6.45. *For all $x, y \in \mathbb{R}$ we have*

(i) $\sin^2(x) + \cos^2(x) = 1$.

(ii) $\sin(x + y) = \sin(x)\cos(y) + \cos(x)\sin(y)$.

(iii) $\cos(x + y) = \cos(x)\cos(y) - \sin(x)\sin(y)$.

Proof. (i) We have

$$[\sin^2(x) + \cos^2(x)]' = 2\sin(x)\cos(x) + 2\cos(x)(-\sin(x)) = 0.$$

But $\sin^2(0) + \cos^2(0) = 1$, so the identity holds for all x.

(ii) Fix some y and let

$$f(x) := \sin(x + y) - \sin(x)\cos(y) - \cos(x)\sin(y).$$

Then we can easily see that $f'' = -f$, and $f(0) = f'(0) = 0$. Now set

$$g_1(x) := f(x)\cos(x) - f'(x)\sin(x),$$
$$g_2(x) := f(x)\sin(x) + f'(x)\cos(x).$$

Then

$$g_1' = f'\cos x - f\sin x - f''\sin x - f'\cos x = -f\sin x + f\sin x = 0.$$

Similarly $g_2' = 0$. In addition we have $g_1(0) = g_2(0) = 0$. Therefore $g_1 \equiv 0 \equiv g_2$. Hence we also have

$$\begin{aligned} f &= f[\cos^2 x + \sin^2 x] + f'[-\cos x \sin x + \sin x \cos x] \\ &= g_1 \cos x + g_2 \sin x \equiv 0. \end{aligned}$$

(iii) This follows from (ii) by differentiating with respect to x. ■

Remark. In the above proof, from $f'' = -f$ and $f(0) = f'(0) = 0$, we concluded that $f \equiv 0$. Similarly to the case of the exponential function, the technique used here comes from differential equations.

Definition 6.46. $\pi := 2\int_{-1}^{1} \sqrt{1-x^2}\,dx$.

Remark. The graph of $\sqrt{1-x^2}$ for $x \in [-1,1]$ is a semicircle. Thus the above definition intuitively means that π is the area of a circle of radius one. By approximating the above integral we can estimate the value of π as accurately as we want. It is also easy to see that $\pi := 4\int_{0}^{1} \sqrt{1-x^2}\,dx$.

Theorem 6.47. *We have* $\cos(\frac{\pi}{2}) = 0$ *and* $\sin(\frac{\pi}{2}) = 1$. *Furthermore,* $\frac{\pi}{2}$ *is the smallest positive number whose cosine is zero.*

Proof. As $\cos(0) = 1$ and cos is continuous, $\cos(x) > 0$ when $x > 0$ is small. Suppose to the contrary that $\cos(x) > 0$ for all $x > 0$. Then sin will be strictly increasing on $[0,\infty)$. Let $a > 0$. Then $b := \sin a > \sin 0 = 0$. Hence by the mean value theorem there is $c > a$ such that

$$\cos(a + \frac{\cos a}{b}) - \cos a = \frac{\cos a}{b}(-\sin c) < \frac{\cos a}{b}(-\sin a) = -\cos a.$$

This implies that $\cos(a + \frac{\cos a}{b}) < 0$, which is a contradiction. Therefore $\cos(x) \le 0$ for some $x > 0$. Due to the continuity of cos, the set

$$A := \{x \ge 0 : \cos(x) = 0\} = \{x \ge 0\} \cap \{\cos(x) = 0\}$$

is nonempty and closed. Therefore $\inf A \in A$. Note that $\inf A > 0$, since $\cos 0 = 1$. Now we define

$$\alpha := 2\inf A.$$

Then $\cos(\frac{\alpha}{2}) = 0$. Therefore $\sin(\frac{\alpha}{2}) = \pm 1$ by the first identity in the last theorem. But $\cos > 0$ on $[0, \frac{\alpha}{2})$. Thus sin is strictly increasing and positive on $(0, \frac{\alpha}{2})$. Hence $\sin(\frac{\alpha}{2})$ must be positive.

It only remains to show that $\alpha = \pi$. We do this by changing the variable in the integral $\int_0^1 \sqrt{1-x^2}\,dx$. We know that the derivative of sin is continuous everywhere, and is positive on the interval $(0, \frac{\alpha}{2})$. Thus by the change of variable $x = \sin t$ we obtain

$$\int_{\sin(0)}^{\sin(\frac{\alpha}{2})} \sqrt{1-x^2}\,dx = \int_0^{\frac{\alpha}{2}} \sqrt{1-\sin^2 t}\cos t\,dt = \int_0^{\frac{\alpha}{2}} \sqrt{\cos^2 t}\cos t\,dt.$$

Note that cos is positive on $(0, \frac{\alpha}{2})$, so $\sqrt{\cos^2 t} = \cos t$. We also have $\cos 2t = \cos^2 t - \sin^2 t = 2\cos^2 t - 1$. Therefore

$$\begin{aligned}\int_0^1 \sqrt{1-x^2}\,dx &= \int_0^{\frac{\alpha}{2}} \cos^2 t\,dt = \int_0^{\frac{\alpha}{2}} \frac{\cos 2t + 1}{2}\,dt \\ &= \frac{1}{4}\sin 2t + \frac{t}{2}\Big|_0^{\frac{\alpha}{2}} = \frac{1}{4}\sin(\alpha) + \frac{\alpha}{4}.\end{aligned}$$

But $\sin(\alpha) = 2\sin(\frac{\alpha}{2})\cos(\frac{\alpha}{2}) = 0$. Hence we get

$$\pi = 4\int_0^1 \sqrt{1-x^2}\,dx = \sin(\alpha) + \alpha = \alpha,$$

as desired. ■

Definition 6.48. A function $f : \mathbb{R} \to \mathbb{R}$ is called **periodic** with **period** $p \in \mathbb{R}-\{0\}$, if for all $x \in \mathbb{R}$ we have

$$f(x+p) = f(x).$$

Remark. Note that if f is periodic with period p, then for all $x \in \mathbb{R}$ and all $n \in \mathbb{Z}$ we have

$$f(x+np) = f(x).$$

This can be proved by an easy induction on n when $n \geq 0$. The case of negative n follows immediately by noting that $f(x-p) = f(x-p+p) = f(x)$.

Theorem 6.49. $\sin$ *and* $\cos$ *are periodic with period* 2π*, i.e. for all* $x \in \mathbb{R}$ *we have*

$$\begin{aligned}\sin(x+2\pi) &= \sin x,\\ \cos(x+2\pi) &= \cos x.\end{aligned}$$

Furthermore, the range of $\sin$ *and* $\cos$ *is* $[-1, 1]$.

Proof. We have

$$\begin{aligned}\sin\pi &= \sin(\frac{\pi}{2}+\frac{\pi}{2}) = 2\sin\frac{\pi}{2}\cos\frac{\pi}{2} = 0,\\ \cos\pi &= \cos(\frac{\pi}{2}+\frac{\pi}{2}) = \cos^2\frac{\pi}{2} - \sin^2\frac{\pi}{2} = -1.\end{aligned}$$

Therefore we can similarly show that $\sin 2\pi = 0$ and $\cos 2\pi = 1$. Hence

$$\begin{aligned}\sin(x+2\pi) &= \sin x\cos 2\pi + \cos x\sin 2\pi = \sin x,\\ \cos(x+2\pi) &= \cos x\cos 2\pi - \sin x\sin 2\pi = \cos x.\end{aligned}$$

Next we show that the range of $\sin, \cos$ is $[-1, 1]$. First note that $\sin^2 + \cos^2 = 1$. So $|\sin|, |\cos| \leq 1$, which means the range is contained in $[-1, 1]$. On the other hand we have $\cos 0 = 1, \cos\pi = -1$. Also $\sin\frac{\pi}{2} = 1$, and we can easily show that $\sin\frac{3\pi}{2} = -1$. Thus we get the desired result by the intermediate value theorem. ■

Remark. Similarly to the above proof, we can show that

$$\sin(x+\pi) = -\sin x, \qquad \cos(x+\pi) = -\cos x.$$

Theorem 6.50. sin *and* cos *are analytic functions. As a result, they are both smooth.*

Proof. We have to show that $\sin, \cos$ can be expressed locally as a convergent power series around any point in their domains. We do this by employing the functional identities that they satisfy. Let $a \in \mathbb{R}$. Then we have

$$\begin{aligned}\sin(x) &= \sin(a + x - a) = \sin(a)\cos(x-a) + \cos(a)\sin(x-a)\\ &= \sin(a)\sum_{n=0}^{\infty}\frac{(-1)^n}{(2n)!}(x-a)^{2n} + \cos(a)\sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)!}(x-a)^{2n+1}\\ &= \sum_{n=0}^{\infty}\frac{c_n}{n!}(x-a)^n, \qquad \text{where } c_n := \begin{cases}(-1)^{\frac{n}{2}}\sin(a) & n \text{ is even,}\\ (-1)^{\frac{n-1}{2}}\cos(a) & n \text{ is odd.}\end{cases}\end{aligned}$$

Note that the radius of convergence of this power series is infinity. The case of cos is similar. ■

Geometric Definition of Trigonometric Functions. Let us show that our definitions of sin and cos agree with the familiar geometric definitions presented in high school algebra courses. Let ABC be a right triangle in the plane $\mathbb{R}^2$, with the right angle at B. We assume that the coordinates are chosen so that $A = (0,0)$, $B = (b,0)$, and $C = (b,c)$, where $b, c > 0$. First we have to define the measure of an angle.

Definition. The angle between two rays in $\mathbb{R}^2$ emanating from the origin is t *radians* if the portion of the unit circle S^1 inside that angle has length t.

In order to work with the above definition, we need to define and compute the length of a curve in the plane. First we start with the following definition.

Definition 6.51. Suppose $r : [a,b] \to \mathbb{R}^n$ is continuous. We say r is a **rectifiable path** if there is $C > 0$ such that for every partition $P = \{a_0, \ldots, a_k\}$ of $[a,b]$ we have

$$L(r,P) := \sum_{j=0}^{k-1}|r(a_{j+1}) - r(a_j)| \le C.$$

When r is a rectifiable path, we define its **length** to be

$$L(r) = \sup_P L(r,P),$$

where the supremum is taken over all partitions P of $[a,b]$.

Remark. Note that $L(r,P)$ is the length of a polygonal path with the vertices $r(a_0),\dots,r(a_k)$. So when the mesh of P converges to zero, i.e. when the partition becomes finer, we expect that the corresponding polygonal path approximates the image of r more closely. Thus we expect that for well-behaved paths, $L(r,P)$ converges to the length of r as the mesh of P goes to zero. On the other hand, the straight line segment is the shortest path between two points. Thus when we refine a partition, i.e. when we add more points to it, the length of the corresponding polygonal path increases. In other words, if $Q = P \cup \{c\}$ is a refinement of P, with $a_l < c < a_{l+1}$, then we have

$$\begin{aligned} L(r,P) &= \sum_{j=0}^{k-1} |r(a_{j+1}) - r(a_j)| = |r(a_{l+1}) - r(a_l)| + \sum_{j\neq l} |r(a_{j+1}) - r(a_j)| \\ &\le |r(a_{l+1}) - r(c)| + |r(c) - r(a_l)| + \sum_{j\neq l} |r(a_{j+1}) - r(a_j)| = L(r,Q). \end{aligned}$$

And inductively, we can show that $L(r,P) \le L(r,Q)$, when Q is a refinement of P which has m extra points. Also note that for every two partitions P_1, P_2, there is a partition $Q = P_1 \cup P_2$, aka their common refinement, such that $L(r,P_i) \le L(r,Q)$ for $i = 1,2$. Therefore to compute the limit of $L(r,P)$, when the mesh of P goes to zero, it suffices to take the supremum of $L(r,P)$ over all partitions P. Hence we arrive at the above definition.

Definition 6.52. Let $\mathcal{C} \subset \mathbb{R}^n$. We say $\mathcal{C}$ is a **rectifiable curve** if there is a one-to-one continuous function $r : [a,b] \to \mathbb{R}^n$ which is a rectifiable path, such that $\mathcal{C}$ is the image of r, i.e. $\mathcal{C} = r([a,b])$. In this case we say r is a **parametrization** of $\mathcal{C}$. We also define the **length** of $\mathcal{C}$ to be

$$L(\mathcal{C}) := L(r).$$

Theorem 6.53. *Suppose $r : [a,b] \to \mathbb{R}^n$ is a one-to-one rectifiable path, and $\mathcal{C}$ is the image of r. Also suppose that $\rho : [c,d] \to \mathbb{R}^n$ is a one-to-one continuous function whose image equals $\mathcal{C}$. Then ρ is a rectifiable path, and has the same length as r, i.e. $L(r) = L(\rho)$.*

Remark. This theorem means that the notions of rectifiability and length of a curve do not depend on the particular parametrization. In other words, these notions are invariant under *reparametrization.*

Proof. Consider the function $\alpha := r^{-1} \circ \rho : [c,d] \to \mathcal{C} \to [a,b]$. Since r is a one-to-one continuous function on a compact domain, r^{-1} is continuous too. Therefore α is a one-to-one continuous function between two intervals. Thus it is either strictly increasing or strictly decreasing, as shown in Exercise 2.84. Let us assume that α

is strictly increasing, the other case is similar. Note that α is also onto, since the image of both r, ρ is $\mathcal{C}$. Now suppose $Q = \{c_0, \dots, c_k\}$ is a partition of $[c, d]$. Let $a_j := \alpha(c_j)$. Then $P := \{a_0, \dots, a_k\}$ is a partition of $[a, b]$, because $a_j < a_{j+1}$, and we must have $a_0 = a$, $a_k = b$, due to the fact that α is strictly increasing and onto. (Note that for P to be a partition when α is strictly decreasing, we have to set $a_j := \alpha(c_{k-j})$.)

Now note that $\rho = r \circ \alpha$. Hence $\rho(c_j) = r(a_j)$. Therefore we have

$$L(\rho, Q) = \sum_{j=0}^{k-1} |\rho(c_{j+1}) - \rho(c_j)| = \sum_{j=0}^{k-1} |r(a_{j+1}) - r(a_j)| = L(r, P) \leq L(r).$$

Thus ρ is rectifiable. In addition we have $L(\rho) \leq L(r)$. If we repeat the above argument with the roles of r, ρ switched, we get $L(r) \leq L(\rho)$ too. ■

Theorem 6.54. *Suppose I is an open interval containing $[a, b]$, and $r = (r_1, \dots, r_n) : I \to \mathbb{R}^n$. Suppose that for each i, r_i is differentiable on I, and r_i' is continuous over $[a, b]$. Then $r|_{[a,b]}$ is a rectifiable path, and we have*

$$L(r|_{[a,b]}) = \int_a^b |r'(t)|\, dt = \int_a^b \left(r_1'(t)^2 + \cdots + r_n'(t)^2\right)^{\frac{1}{2}} dt,$$

where $r' := (r_1', \dots, r_n')$.

Proof. It is obvious that r is continuous. Let $x, y \in [a, b]$, and let $z := r(y) - r(x)$. Then for each i we have $z_i = r_i(y) - r_i(x) = \int_x^y r_i'(t) dt$. Hence we have

$$\begin{aligned}|z|^2 &= \sum_{i\leq n} z_i^2 = \sum_{i\leq n} z_i \int_x^y r_i'(t) dt = \int_x^y \sum_{i\leq n} z_i r_i'(t) dt \\ &= \int_x^y z \cdot r'(t) dt \leq \int_x^y |z||r'(t)|\, dt = |z| \int_x^y |r'(t)|\, dt.\end{aligned}$$

Therefore $|r(y) - r(x)| = |z| \leq \int_x^y |r'(t)|\, dt$. Now let $P = \{a_0, \dots, a_k\}$ be a partition of $[a, b]$. Then we have

$$L(r|_{[a,b]}, P) = \sum_{j=0}^{k-1} |r(a_{j+1}) - r(a_j)| \leq \sum_{j=0}^{k-1} \int_{a_j}^{a_{j+1}} |r'(t)|\, dt = \int_a^b |r'(t)|\, dt.$$

Hence $r|_{[a,b]}$ is rectifiable. In addition we have $L(r|_{[a,b]}) \leq \int_a^b |r'(t)|\, dt$. Thus we only need to prove the reverse inequality.

On the other hand we know that r' is uniformly continuous on $[a, b]$. Hence for a given $\epsilon > 0$ there is $\delta > 0$ such that if $|x - y| < \delta$ then $|r'(x) - r'(y)| < \epsilon$. Now let

$P = \{a_0, \dots, a_k\}$ be a partition of $[a, b]$ whose mesh $\|P\| = \max_{j<k} |a_{j+1} - a_j| < \delta$. Then for every $t \in [a_j, a_{j+1}]$ we have $|r'(t) - r'(a_j)| < \epsilon$, so $|r'(t)| < |r'(a_j)| + \epsilon$. Therefore

$$\int_{a_j}^{a_{j+1}} |r'(t)|\, dt \le \big(|r'(a_j)| + \epsilon\big)(a_{j+1} - a_j). \qquad (*)$$

Now for each i there is $\theta_i \in [a_j, a_{j+1}]$ so that

$$r_i(a_{j+1}) - r_i(a_j) = r_i'(\theta_i)(a_{j+1} - a_j).$$

Thus

$$\begin{aligned} &|r_i(a_{j+1}) - r_i(a_j) - r_i'(a_j)(a_{j+1} - a_j)| \\ &\qquad = |r_i'(\theta_i) - r_i'(a_j)|(a_{j+1} - a_j) < \epsilon(a_{j+1} - a_j). \end{aligned}$$

Therefore we have

$$\begin{aligned} &|r(a_{j+1}) - r(a_j) - r'(a_j)(a_{j+1} - a_j)| \\ &\qquad \le \sum_{i \le n} |r_i(a_{j+1}) - r_i(a_j) - r_i'(a_j)(a_{j+1} - a_j)| < n\epsilon(a_{j+1} - a_j). \end{aligned}$$

Hence $|r'(a_j)|(a_{j+1} - a_j) < |r(a_{j+1}) - r(a_j)| + n\epsilon(a_{j+1} - a_j)$. Thus from inequality $(*)$ we get

$$\int_{a_j}^{a_{j+1}} |r'(t)|\, dt \le |r(a_{j+1}) - r(a_j)| + (n + 1)\epsilon(a_{j+1} - a_j).$$

Therefore we have

$$\begin{aligned} \int_a^b |r'(t)|\, dt &= \sum_{j<k} \int_{a_j}^{a_{j+1}} |r'(t)|\, dt \\ &\le \sum_{j<k} |r(a_{j+1}) - r(a_j)| + (n + 1)\epsilon \sum_{j<k} (a_{j+1} - a_j) \\ &= L(r|_{[a,b]}, P) + (n + 1)\epsilon(b - a) \le L(r|_{[a,b]}) + (n + 1)\epsilon(b - a). \end{aligned}$$

Now as ϵ is arbitrary we get $\int_a^b |r'(t)|\, dt \le L(r|_{[a,b]})$, as desired. ■

Finally let us return to our original problem. Recall that ABC is a right triangle in the plane $\mathbb{R}^2$, with the right angle at B, such that $A = (0, 0)$, $B = (b, 0)$, and $C = (b, c)$, where $b, c > 0$. In order to compute the sine and cosine of the angle $\widehat{A}$ at the vertex A, we have to compute the length of the portion of the unit circle that lies between the two rays $\overrightarrow{AB}, \overrightarrow{AC}$. This portion has the endpoints $(1, 0)$ and

$(\frac{b}{\sqrt{b^2+c^2}}, \frac{c}{\sqrt{b^2+c^2}})$. The map $t \mapsto (\sqrt{1-t^2}, t)$ for $t \in [0, \frac{c}{\sqrt{b^2+c^2}}]$ is a one-to-one parametrization of the portion. Let $a := \frac{c}{\sqrt{b^2+c^2}}$. Thus the length of the portion is

$$\int_0^a \sqrt{1+\Big(\frac{-t}{\sqrt{1-t^2}}\Big)^2}dt = \int_0^a \frac{1}{\sqrt{1-t^2}}dt.$$

Since $a < 1$, the integrand is continuous. Also $\sin 0 = 0 < a < 1 = \sin\frac{\pi}{2}$, and sin is strictly increasing on $[0, \frac{\pi}{2})$, since $\cos > 0$ on $[0, \frac{\pi}{2})$. Thus there is a unique $\alpha \in (0, \frac{\pi}{2})$ so that $\sin\alpha = a$. Now we change the variable of integration as $t = \sin s$. Hence we have

$$\int_0^a \frac{1}{\sqrt{1-t^2}}dt = \int_0^\alpha \frac{1}{\sqrt{1-\sin^2 s}}\cos s\, ds = \int_0^\alpha \frac{\cos s}{\sqrt{\cos^2 s}}ds = \int_0^\alpha 1ds = \alpha.$$

Therefore we can finally compute the sine and cosine of the angle $\widehat{A}$. Remember that by definition these are the sine and cosine of the length of the portion of the unit circle that lies between the two rays $\overrightarrow{AB}, \overrightarrow{AC}$, i.e. α. The lengths of the sides of the triangle are $\overline{AB} = b$, $\overline{BC} = c$, and by the Pythagorean theorem $\overline{AC} = \sqrt{b^2+c^2}$. Thus

$$\begin{aligned}\sin\widehat{A} &:= \sin\alpha = a = \frac{c}{\sqrt{b^2+c^2}} = \frac{\overline{BC}}{\overline{AC}},\\ \cos\widehat{A} &:= \cos\alpha = \sqrt{1-\sin^2\alpha} = \sqrt{1-a^2} = \frac{b}{\sqrt{b^2+c^2}} = \frac{\overline{AB}}{\overline{AC}}.\end{aligned}$$

These are the familiar geometric expressions for sine and cosine. ■

Complex Exponents.

Definition 6.55. Let $r \in \mathbb{R}$ be positive, and $z = x + iy \in \mathbb{C}$. We define

$$r^z = r^{x+iy} := r^x\big(\cos(y\log r) + i\sin(y\log r)\big).$$

Remark. Obviously when z is real, r^z has the same value as we defined before.

Theorem 6.56. *Let $r, s \in (0, \infty)$ and $z, w \in \mathbb{C}$. Then we have*

(i) $r^z r^w = r^{z+w}$.

(ii) $(rs)^z = r^z s^z$.

$\boxed{\textbf{Proof.}}$ Suppose $z = x + iy$ and $w = a + ib$.

(i) We have

$$\begin{aligned}
r^z r^w &= r^x\big(\cos(y\log r) + i\sin(y\log r)\big)\, r^a\big(\cos(b\log r) + i\sin(b\log r)\big)\\
&= r^x r^a\,\big[\cos(y\log r)\cos(b\log r) - \sin(y\log r)\sin(b\log r)\\
&\qquad\qquad + i\big(\cos(y\log r)\sin(b\log r) + \sin(y\log r)\cos(b\log r)\big)\big]\\
&= r^{x+a}\,\big[\cos\big((y+b)\log r\big) + i\sin\big((y+b)\log r\big)\big]\\
&= r^{z+w}.
\end{aligned}$$

(ii) We have

$$\begin{aligned}
r^z s^z &= r^x\big(\cos(y\log r) + i\sin(y\log r)\big)\, s^x\big(\cos(y\log s) + i\sin(y\log s)\big)\\
&= r^x s^x\,\big[\cos(y\log r)\cos(y\log s) - \sin(y\log r)\sin(y\log s)\\
&\qquad\qquad + i\big(\cos(y\log r)\sin(y\log s) + \sin(y\log r)\cos(y\log s)\big)\big]\\
&= (rs)^x\,\big[\cos\big(y(\log r + \log s)\big) + i\sin\big(y(\log r + \log s)\big)\big]\\
&= (rs)^x\,\big[\cos\big(y\log(rs)\big) + i\sin\big(y\log(rs)\big)\big]\\
&= (rs)^z.
\end{aligned}$$

■

Theorem 6.57. *Let $r \in (0,\infty)$ and $z = x + iy \in \mathbb{C}$. Then for any $n \in \mathbb{Z}$ and $a \in \mathbb{R}$ we have*

(i) $(r^z)^n = r^{nz}$.

(ii) $(r^a)^z = r^{az}$.

Proof. (i) It is trivial that the equality holds for $n = 0$. For positive n we prove the theorem by induction. The induction step is

$$(r^z)^{n+1} = (r^z)^n r^z = r^{nz} r^z = r^{nz+z} = r^{(n+1)z}.$$

Next consider $n = -1$. Then

$$\begin{aligned}
(r^z)^{-1} &= \big(r^x(\cos(y\log r) + i\sin(y\log r))\big)^{-1}\\
&= r^{-x}\,\frac{\cos(y\log r) - i\sin(y\log r)}{\cos^2(y\log r) + \sin^2(y\log r)}\\
&= r^{-x}\big(\cos(-y\log r) + i\sin(-y\log r)\big) = r^{-z}.
\end{aligned}$$

Now suppose $n = -m < 0$. We have

$$(r^z)^{-m} = ((r^z)^{-1})^m = (r^{-z})^m = r^{-mz}.$$

(ii) We have

$$\begin{aligned}
(r^a)^z &= (r^a)^x\big(\cos(y\log r^a) + i\sin(y\log r^a)\big)\\
&= r^{ax}\big(\cos(ay\log r) + i\sin(ay\log r)\big)\\
&= r^{az}.
\end{aligned}$$

■

Theorem 6.58. *Let $r \in (0,\infty)$ and $z = a+ib \in \mathbb{C}$. Then the functions $t \mapsto t^z$ and $t \mapsto r^{tz}$ are differentiable functions on $(0,\infty)$ and $\mathbb{R}$, respectively. Furthermore we have*

(i) $(t^z)' = \frac{d}{dt}(t^z) = zt^{z-1}$.

(ii) $(r^{tz})' = \frac{d}{dt}(r^{tz}) = (\log r)zr^{tz}$.

Proof. First note that the derivative of a complex-valued function is the function whose real and imaginary parts are respectively the derivative of the real and imaginary parts of the original function.

(i) We have

$$\begin{aligned}
(t^z)' &= \big(t^a\cos(b\log t) + it^a\sin(b\log t)\big)' \\
&= at^{a-1}\cos(b\log t) - t^a\sin(b\log t)\frac{b}{t} \\
&\qquad\qquad + i\big(at^{a-1}\sin(b\log t) + t^a\cos(b\log t)\frac{b}{t}\big) \\
&= (a+ib)\,t^{a-1}\big(\cos(b\log t) + i\sin(b\log t)\big) \\
&= zt^{z-1}.
\end{aligned}$$

(ii) We have

$$\begin{aligned}
(r^{tz})' &= \big(r^{ta}\cos(tb\log r) + ir^{ta}\sin(tb\log r)\big)' \\
&= (\log r)ar^{ta}\cos(tb\log r) - r^{ta}b(\log r)\sin(tb\log r) \\
&\qquad\qquad + i\big((\log r)ar^{ta}\sin(tb\log r) + r^{ta}b(\log r)\cos(tb\log r)\big) \\
&= (\log r)(a+ib)\,r^{ta}\big(\cos(tb\log r) + i\sin(tb\log r)\big) \\
&= (\log r)zr^{tz}.
\end{aligned}$$

Alternatively, we can compute the derivative using the last part as follows

$$(r^{tz})' = ((r^t)^z)' = z(r^t)^{z-1}(r^t)' = (\log r)z(r^t)^{z-1}r^t = (\log r)z(r^t)^z. \qquad ■$$

Remark. Let $z = x+iy \in \mathbb{C}$ and $r = e$. Then we have

$$e^z = e^x(\cos y + i\sin y).$$

In particular for $\theta \in \mathbb{R}$ we have

$$e^{i\theta} = \cos\theta + i\sin\theta.$$

Note that e^z is a periodic function with period $2\pi i$, since

$$e^{z+2\pi i} = e^x\big(\cos(y+2\pi) + i\sin(y+2\pi)\big) = e^x(\cos y + i\sin y) = e^z.$$

As a result we have $e^{i(\theta+2\pi)} = e^{i\theta}$. Another interesting fact is that for $\theta = \pi$ we have

$$e^{i\pi} = \cos\pi + i\sin\pi = -1,$$

or equivalently $e^{i\pi} + 1 = 0$.

Theorem 6.59. *The map $\theta \mapsto e^{i\theta}$ is a one-to-one and onto correspondence between the interval $[0, 2\pi)$ and the unit circle $S^1 \subset \mathbb{C} = \mathbb{R}^2$.*

Proof. Note that we always have

$$|e^{i\theta}| = \sqrt{\cos^2\theta + \sin^2\theta} = 1.$$

Hence the image of the map is inside S^1. Let $a + ib \in S^1$. Then $a^2 + b^2 = 1$, so $-1 \le a, b \le 1$. Also we have $b = \pm\sqrt{1 - a^2}$. Now the range of cos is $[-1, 1]$, and cos is periodic with period 2π. Thus there is $\theta \in [0, 2\pi)$ such that $a = \cos\theta$. If $b, \sin\theta$ have the same sign, then

$$b = \pm\sqrt{1 - a^2} = \pm\sqrt{1 - \cos^2\theta} = \sin\theta.$$

Therefore we have $e^{i\theta} = a + ib$ as desired. Otherwise we can use $2\pi - \theta$ instead of θ, since

$$\cos(2\pi - \theta) = \cos(-\theta) = \cos\theta,$$
$$\sin(2\pi - \theta) = \sin(-\theta) = -\sin\theta.$$

Note that in this case we can assume $\theta \neq 0$, since for $\theta = 0$ we must have $b = 0 = \sin 0$. Also if $\theta \in (0, 2\pi)$ then $2\pi - \theta \in (0, 2\pi)$ too. Hence we have shown that the map is onto.

Next let us show that the map is one-to-one. We know that cos is positive on $(0, \frac{\pi}{2})$, so sin is increasing and therefore positive on this interval. Since $\sin(2\pi - x) = -\sin x$, and $\sin(\pi - x) = \sin x$, we see that sin is positive on $(0, \pi)$, and it is negative on $(\pi, 2\pi)$. Therefore cos is decreasing on $(0, \pi)$, and it is increasing on $(\pi, 2\pi)$, so it is injective on each of these intervals. Also note that cos maps both of these intervals onto $(-1, 1)$, by the intermediate value theorem. Thus every number in the interval $(-1, 1)$ is the cos of two points in $[0, 2\pi)$, one in $(0, \pi)$ and one in $(\pi, 2\pi)$. In addition, the sin of these two points have different signs. Also, the only points in $[0, 2\pi)$ with cos equal to $1, -1$ are respectively $0, \pi$. Hence if for two points $\theta, \phi \in [0, 2\pi)$ we have $e^{i\theta} = e^{i\phi}$, then $\cos\theta = \cos\phi$, and $\sin\theta = \sin\phi$. But by the above arguments, these two equalities imply that $\theta = \phi$. Therefore we get the desired. ■

6.5 Compactness in Function Spaces

Definition 6.60. A metric space is **separable** if it has a countable dense subset.

***Example* 6.61.** $\mathbb{R}$ is separable since $\mathbb{Q}$ is a countable and dense subset of $\mathbb{R}$. More generally any interval in $\mathbb{R}$ is separable, because the rational numbers in the interval form a countable and dense subset of the interval.

Theorem 6.62. *A compact metric space is separable.*

Proof. Suppose X is a compact metric space. Let $n \in \mathbb{N}$. Then the family $\{B_{\frac{1}{n}}(x)\}_{x\in X}$ is an open covering of X. Hence there are finitely many points $x_1, \dots, x_k \in X$ such that

$$X \subset B_{\frac{1}{n}}(x_1) \cup \cdots \cup B_{\frac{1}{n}}(x_k). \tag{$*$}$$

Let us rename these points and call them $x_{n,1}, \dots, x_{n,k_n}$. Now we claim that the set

$$A := \{x_{n,i} : n \in \mathbb{N},\, i \le k_n\}$$

is dense in X. To prove this we have to show that $X = \bar{A}$. Let $x \in X$. Then for every $n \in \mathbb{N}$ there is $x_{n,i_n} \in A$ such that $x \in B_{\frac{1}{n}}(x_{n,i_n})$ by $(*)$. It is easy to see that $x_{n,i_n} \to x$ as $n \to \infty$. Therefore every $x \in X$ is in the closure of A, and consequently A is dense in X. Finally note that A is countable since it is the union of countably many finite sets. ■

Definition 6.63. Let X, Y be two metric spaces. A sequence of continuous functions $f_n : X \to Y$ is **equicontinuous** if

$$\begin{aligned}\forall \epsilon > 0\ \exists \delta > 0 \text{ such that } \forall n\ \ \forall x, y \in X\\ d_X(x,y) < \delta \implies d_Y(f_n(x), f_n(y)) < \epsilon.\end{aligned}$$

Remark. The point of the above definition is that δ does not depend on n. It also does not depend on x, y, so we are tacitly assuming that each f_n is uniformly continuous.

Arzela-Ascoli Theorem. *Suppose X is a compact metric space. Then every bounded equicontinuous sequence of functions in $C^0(X,\mathbb{R})$ has a uniformly convergent subsequence.*

Proof. Let (f_n) be a bounded equicontinuous sequence in $C^0(X)$. So there is $C > 0$ such that $\|f_n\|_{\sup} \le C$ for all n. Suppose $\{x_1, x_2, \dots\}$ is a countable dense subset of X. Then for every i, n we have $|f_n(x_i)| \le \|f_n\|_{\sup} \le C$. Thus the sequence $(f_n(x_1))$ is a bounded sequence in $\mathbb{R}$. Hence it has a subsequence

$(f_{n,1}(x_1))$ that converges to some $y_1 \in \mathbb{R}$. Now consider the sequence $(f_{n,1}(x_2))$. It is also a bounded sequence. So it has a subsequence $(f_{n,2}(x_2))$ that converges to some y_2. We can continue this process inductively and obtain a subsequence $(f_{n,k})$ of $(f_{n,k-1})$ such that $(f_{n,k}(x_k))$ converges to some y_k. Note that by our construction $f_{n,k}(x_i) \to y_i$ for $i \leq k$.

We claim that the diagonal sequence $(f_{k,k})$ is a uniformly convergent subsequence of (f_n). First note that for all i we have $f_{k,k}(x_i) \to y_i$, since $(f_{k,k})$ is a subsequence of $(f_{n,i})$ if we ignore its first $i-1$ terms. Now suppose $\epsilon > 0$ is given. Then equicontinuity implies that there is $\delta > 0$ so that for all $x, y \in X$ and all $n \in \mathbb{N}$ we have

$$d_X(x,y) < \delta \implies |f_n(x) - f_n(y)| < \frac{\epsilon}{3}.$$

Also, for any $x \in X$ there is x_i such that $d_X(x, x_i) < \delta$. This means that the family of open balls $\{B_\delta(x_i)\}$ is an open covering of X. Hence there are finitely many of x_i's namely $x_1, \ldots, x_j$ such that

$$X \subset B_\delta(x_1) \cup \cdots \cup B_\delta(x_j).$$

Then let $N_i \in \mathbb{N}$ be large enough so that for $m, n \geq N_i$ we have $|f_{n,n}(x_i) - f_{m,m}(x_i)| < \frac{\epsilon}{3}$. This is possible since the sequences $(f_{n,n}(x_i))$ are convergent, therefore they are Cauchy. Let $N = \max\{N_1, \ldots, N_j\}$. Then for any $x \in X$ there is x_i with $i \leq j$ so that $d_X(x, x_i) < \delta$. Hence for $m, n \geq N$ we have

$$\begin{aligned} &|f_{n,n}(x) - f_{m,m}(x)| \\ &\qquad \leq |f_{n,n}(x) - f_{n,n}(x_i)| + |f_{n,n}(x_i) - f_{m,m}(x_i)| + |f_{m,m}(x_i) - f_{m,m}(x)| \\ &\qquad < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon. \end{aligned}$$

Therefore $(f_{n,n})$ is a Cauchy sequence in $C^0(X)$. Thus the result follows because $C^0(X)$ is complete, and convergence in $C^0(X)$ is the uniform convergence. ■

***Example* 6.64.** The mere boundedness of (f_n) does not imply that it has a uniformly convergent subsequence. For example the sequence of functions (x^n) in $C^0([0,1])$ is bounded, but it does not have a uniformly convergent subsequence. To see this, suppose to the contrary that a subsequence (x^{n_k}) converges uniformly to a function f. Then it also converges pointwise to f. Thus we must have $f(1) = 1$ and $f(x) = 0$ for $0 \leq x < 1$. Therefore f is not continuous. But this contradicts the fact that the uniform limit of continuous functions is continuous.

Theorem 6.65. *Let (f_n) be a sequence in $C^0([a,b],\mathbb{R})$. Suppose each f_n is differentiable on (a,b), and for some $M > 0$ we have $|f_n'(x)| \leq M$ independently of x, n. Also suppose that for some $x_0 \in [a,b]$ the sequence $(f_n(x_0))$ is bounded. Then (f_n) has a uniformly convergent subsequence.*

Proof. It suffices to show that (f_n) is bounded and equicontinuous. By the mean value theorem we have

$$|f_n(x) - f_n(y)| \leq M|x - y|$$

for all $x, y \in [a, b]$ and all $n \in \mathbb{N}$. Hence to show that (f_n) is equicontinuous, for a given ϵ we can take $\delta = \frac{\epsilon}{M}$. Next to show that (f_n) is bounded in $C^0([a, b])$ we note that for any $x \in [a, b]$ we have

$$|f_n(x)| \leq |f_n(x) - f_n(x_0)| + |f_n(x_0)| \leq M|x - x_0| + C \leq M(b - a) + C,$$

where $C > 0$ is a bound for the sequence $(|f_n(x_0)|)$. ■

6.6 Approximation by Polynomials

Weierstrass Approximation Theorem. *The set of polynomials is dense in $C^0([a, b], \mathbb{R})$, i.e. any real-valued continuous function on $[a, b]$ is the uniform limit of a sequence of polynomials.*

Proof. **(Bernstein, 1912)** It suffices to prove the theorem when $[a, b] = [0, 1]$. Because if $f(x)$ is a continuous function of $x \in [a, b]$, then $f((1 - t)a + tb)$ is a continuous function of $t \in [0, 1]$. Hence if p is a polynomial such that

$$\big|p(t) - f((1 - t)a + tb)\big| < \epsilon,$$

for a given $\epsilon > 0$ and $t \in [0, 1]$, then

$$\Big|p\big(\frac{x - a}{b - a}\big) - f(x)\Big| < \epsilon,$$

for $x \in [a, b]$. Note that by the binomial theorem, $p\big(\frac{x-a}{b-a}\big)$ is a polynomial in x.

So suppose f is a continuous function on $[0, 1]$. Let

$$p_n(x) := \sum_{k=0}^{n} \binom{n}{k} f\big(\frac{k}{n}\big) x^k (1 - x)^{n-k}.$$

p_n is a polynomial called a **Bernstein polynomial**. We claim that p_n converges uniformly to f as $n \to \infty$. First note that by the binomial theorem we have

$$\sum_{k=0}^{n} \binom{n}{k} x^k (1 - x)^{n-k} = (x + 1 - x)^n = 1.$$

We also have

$$\sum_{k=0}^{n} (nx - k)^2 \binom{n}{k} x^k (1 - x)^{n-k} = nx(1 - x). \qquad (*)$$

We will prove the above identity at the end of this proof. To simplify the notation we set

$$r_k(x) := \binom{n}{k} x^k (1-x)^{n-k}.$$

Note that $r_k(x) \geq 0$ for all $x \in [0,1]$.

Now suppose $\epsilon > 0$ is given. We want to find $N \in \mathbb{N}$ such that for $n \geq N$ and $x \in [0,1]$ we have $|p_n(x) - f(x)| < \epsilon$. Since $f(x) = \sum_{k=0}^n f(x) r_k(x)$, we have

$$|p_n(x) - f(x)| \leq \sum_{k=0}^n \big| f\big(\frac{k}{n}\big) - f(x) \big| r_k(x).$$

Due to the uniform continuity of f there is $\delta > 0$ such that $|y - x| < \delta$ implies $|f(y) - f(x)| < \frac{\epsilon}{2}$. Let

$$I := \{0 \leq k \leq n : \Big|\frac{k}{n} - x\Big| < \delta\}, \qquad J := \{0 \leq k \leq n : \Big|\frac{k}{n} - x\Big| \geq \delta\}.$$

Then for $k \in J$ we have $(nx - k)^2 \geq n^2\delta^2$. Thus

$$\begin{aligned}\sum_{k \in J} r_k(x) &= \frac{1}{n^2\delta^2} \sum_{k \in J} n^2\delta^2 r_k(x) \leq \frac{1}{n^2\delta^2} \sum_{k \in J} (nx - k)^2 r_k(x) \\ &\leq \frac{1}{n^2\delta^2} \sum_{k \leq n} (nx - k)^2 r_k(x) = \frac{1}{n\delta^2} x(1-x) \leq \frac{1}{4n\delta^2}.\end{aligned}$$

Note that here we used the fact that the maximum of $x(1-x)$ on $[0,1]$ is $\frac{1}{2}$. Now let $M > 0$ be the maximum of f on $[0,1]$. Then for $n > \frac{M}{\epsilon\delta^2}$ we have

$$\begin{aligned}|p_n(x) - f(x)| &\leq \sum_{k \in I} \big| f\big(\frac{k}{n}\big) - f(x) \big| r_k(x) + \sum_{k \in J} \big| f\big(\frac{k}{n}\big) - f(x) \big| r_k(x) \\ &< \frac{\epsilon}{2} \sum_{k \in I} r_k(x) + 2M \sum_{k \in J} r_k(x) \leq \frac{\epsilon}{2} \sum_{k \leq n} r_k(x) + \frac{M}{2n\delta^2} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.\end{aligned}$$

Proof of the identity $(*)$: We differentiate $(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}$ with respect to x twice to obtain

$$\begin{aligned}n(x+y)^{n-1} &= \sum_{k=0}^n \binom{n}{k} k x^{k-1} y^{n-k}, \\ n(n-1)(x+y)^{n-2} &= \sum_{k=0}^n \binom{n}{k} k(k-1) x^{k-2} y^{n-k}.\end{aligned}$$

Then we multiply these equations by x, x^2 respectively, and we set $y = 1-x$. Hence we get

$$\begin{aligned} nx &= \sum_{k=0}^{n} \binom{n}{k} k x^k (1-x)^{n-k} &&= \sum_{k=0}^{n} k r_k(x), \\ n(n-1)x^2 &= \sum_{k=0}^{n} \binom{n}{k} k(k-1) x^k (1-x)^{n-k} &&= \sum_{k=0}^{n} k(k-1) r_k(x). \end{aligned}$$

Now we have

$$\begin{aligned} \sum_{k=0}^{n} (nx-k)^2 r_k(x) &= n^2x^2 \sum_{k=0}^{n} r_k(x) - 2nx \sum_{k=0}^{n} k r_k(x) + \sum_{k=0}^{n} k^2 r_k(x) \\ &= n^2x^2 - 2n^2x^2 + \sum_{k=0}^{n} k(k-1) r_k(x) + \sum_{k=0}^{n} k r_k(x) \\ &= -n^2x^2 + n(n-1)x^2 + nx = nx(1-x). \qquad \blacksquare \end{aligned}$$

***Exercise* 6.66.** Give an example of a continuous function on $(0,1)$ that is not the uniform limit of any sequence of polynomials. Do the same for $[0,\infty)$.

Chapter 7

Multivariable Differential Calculus

7.1 Derivatives

Definition 7.1. Suppose $U \subset \mathbb{R}^n$ is open and $f : U \to \mathbb{R}^m$. Then we say f is **differentiable** at a point $x \in U$ if there exist an m by n matrix $A \in \mathbb{R}^{m \times n}$ such that for $h \in \mathbb{R}^n$ with $|h|$ small we have

$$f(x+h) = f(x) + Ah + R(h),$$

where R is a function from a neighborhood of the origin of $\mathbb{R}^n$ into $\mathbb{R}^m$ that satisfies

$$\lim_{h \to 0} \frac{R(h)}{|h|} = 0.$$

Remark. A function R satisfying the above property is called **sublinear** (at zero).

Remark. Differentiability of a function means that we can locally approximate the function with a linear function, and the error to this approximation decays faster than a linear function (i.e. it is sublinear).

Remark. There is an equivalent way to formulate differentiability. Let

$$r(h) := \begin{cases} \frac{R(h)}{|h|} & h \neq 0 \\ 0 & h = 0. \end{cases}$$

Then for $h \in \mathbb{R}^n$ with $|h|$ small we have

$$f(x+h) = f(x) + Ah + |h| r(h), \qquad \text{and} \qquad \lim_{h \to 0} r(h) = 0.$$

It is easy to show that f is differentiable at x if and only if for some matrix A and function r the above relations hold.

Remark. Note that R, r are both continuous at 0. We call them the remainders.

Theorem 7.2. *The matrix in the definition of differentiability is unique.*

Proof. Suppose to the contrary that there are two such matrices A_1, A_2 satisfying the differentiability relation. Then for any $h \in \mathbb{R}^n$ and small positive t we have

$$f(x+th) = f(x) + A_i(th) + |th| r_i(th) \qquad i = 1, 2,$$

for some remainders r_1, r_2. If we subtract these two relations we get

$$(A_1 - A_2)(th) = |th|\big(r_2(th) - r_1(th)\big).$$

Thus

$$(A_1 - A_2)h = |h|\big(r_2(th) - r_1(th)\big).$$

But

$$\lim_{t\to 0} r_i(th) = r_i\big(\lim_{t\to 0} th\big) = r_i(0) = 0,$$

since r_i's are continuous at 0. Hence

$$(A_1 - A_2)h = \lim_{t\to 0}(A_1 - A_2)h = \lim_{t\to 0} |h|\big(r_2(th) - r_1(th)\big) = 0.$$

Therefore $A_1 = A_2$. ■

Definition 7.3. The matrix in the definition of differentiability is called the **(total or Frechet) derivative** of f at x and is denoted by $Df(x)$.

Definition 7.4. A map $T : \mathbb{R}^n \to \mathbb{R}^m$ is called **linear** if for every $v, w \in \mathbb{R}^n$ and $c_1, c_2 \in \mathbb{R}$ we have

$$T(c_1 v + c_2 w) = c_1 T(v) + c_2 T(w).$$

Remark. Let A be an $m \times n$ matrix. Then the map $v \mapsto Av$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. Conversely, let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear map. Let $v \in \mathbb{R}^n$. Then we have $v = v_1 e_1 + \cdots + v_n e_n$. Hence

$$T(v) = T(v_1 e_1 + \cdots + v_n e_n) = v_1 T(e_1) + \cdots + v_n T(e_n) = Av,$$

where A is the $m \times n$ matrix whose jth column is $T(e_j)$. In addition, note that if for some matrix B we have $T(v) = Bv$ for every v, then as shown in Appendix A we have

$$B_{\cdot,j} = Be_j = T(e_j) = A_{\cdot,j},$$

i.e. the jth column of B is equal to the jth column of A, for every j. Hence $B = A$. Therefore every linear map between Euclidean spaces is given by the action of a unique matrix.

Theorem 7.5. *A linear map between Euclidean spaces is differentiable everywhere and its derivative is its matrix. Also the derivative of a constant function exists and equals zero everywhere.*

Proof. Just note that the remainder is zero in both cases. ■

Definition 7.6. Suppose $U \subset \mathbb{R}^n$ is open and $f = (f_1, \dots, f_m) : U \to \mathbb{R}^m$. The **directional derivative** of f at $x \in U$ in the direction of $v \in \mathbb{R}^n$ is the vector

$$D_v f(x) := \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t},$$

if the limit exists. Let $\{e_1, \dots, e_n\}$ be the standard basis of $\mathbb{R}^n$. Then we denote $D_{e_j} f(x)$ by

$$D_j f(x),\ D_{x_j} f(x),\ \frac{\partial f}{\partial x_j}(x),\ \text{or}\ f_{x_j}.$$

The **partial derivatives** of f at x are

$$D_j f_i(x) = \lim_{t \to 0} \frac{f_i(x + te_j) - f_i(x)}{t}.$$

Remark. Suppose I is an open subset of $\mathbb{R}$, and $f : I \to \mathbb{R}^m$ is differentiable at x. Then $Df(x) \in \mathbb{R}^{m \times 1} = \mathbb{R}^m$, i.e. the derivative is a vector. It is easy to see that in this case we have

$$Df(x) = D_1 f(x) = f'(x) := \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}.$$

Here $D_1 f(x)$ is the directional derivative of f in the direction of $1 \in \mathbb{R}$.

Theorem 7.7. *Suppose $U \subset \mathbb{R}^n$ is open and $f = (f_1, \dots, f_m) : U \to \mathbb{R}^m$. Then for some $v \in \mathbb{R}^n$ the directional derivative $D_v f(x)$ exists if and only if $D_v f_i(x)$ exists for all i. Furthermore*

$$D_v f(x) = \big(D_v f_1(x), \dots, D_v f_m(x)\big).$$

In particular when $n = 1$ we have

$$f'(x) = \big(f_1'(x), \dots, f_m'(x)\big).$$

Proof. We have

$$\frac{1}{t}[f(x + tv) - f(x)] = \Big(\frac{1}{t}[f_1(x + tv) - f_1(x)], \dots, \frac{1}{t}[f_m(x + tv) - f_m(x)]\Big).$$

Now let $t \to 0$. Note that the limit of a vector function exists if and only if the limit of every component exists. ■

Theorem 7.8. *Suppose $U \subset \mathbb{R}^n$ is open, and $f : U \to \mathbb{R}^m$ is differentiable at x. Then all the directional derivatives of f at x exist and*

$$D_v f(x) = Df(x)v$$

for all $v \in \mathbb{R}^n$. In particular the jth column of $Df(x)$ is $D_j f(x)$, i.e.

$$Df(x) = \Big[\ D_1 f \ \Big|\ \dots\ \Big|\ D_n f\ \Big].$$

Proof. For every vector v and small real number t we can put $h = tv$ in the differentiability relation to obtain

$$f(x + tv) = f(x) + Df(x)(tv) + |tv|r(tv).$$

Therefore we have

$$D_v f(x) = \lim_{t\to 0} \frac{f(x + tv) - f(x)}{t} = \lim_{t\to 0} \big(Df(x)v \pm |v|r(tv)\big) = Df(x)v.$$

The last statement of the theorem follows from the fact that the jth column of a matrix equals the action of that matrix on the standard basis vector e_j. ■

Remark. The above theorem gives a second proof that the matrix in the definition of differentiability is unique. Since if there were two such matrices, their action on every vector v must have been equal to $D_v f$ which is determined uniquely as the value of a limit.

Remark. When $m = 1$, i.e. when f is scalar-valued, $Df(x)$ is a $1 \times n$ row vector. The **gradient** of f at x is the $n \times 1$ column vector defined by $\nabla f(x) := \big(Df(x)\big)^{\mathsf{T}}$. So, the gradient $\nabla f(x)$ is a vector in $\mathbb{R}^n$, while the derivative $Df(x)$ defines a linear map from $\mathbb{R}^n$ to $\mathbb{R}$. Furthermore, for $v \in \mathbb{R}^n$ we have

$$D_v f(x) = Df(x)v = \sum D_j f(x) v_j = v \cdot \nabla f(x).$$

Theorem 7.9. *Suppose $U \subset \mathbb{R}^n$ is open and $f = (f_1, \dots, f_m) : U \to \mathbb{R}^m$ is differentiable at x. Then the ijth entry of $Df(x)$ is $D_j f_i(x)$, i.e.*

$$Df(x) = \begin{bmatrix} D_1 f_1(x) & D_2 f_1(x) & \dots & D_n f_1(x) \\ D_1 f_2(x) & D_2 f_2(x) & \dots & D_n f_2(x) \\ \vdots & \vdots & \ddots & \vdots \\ D_1 f_m(x) & D_2 f_m(x) & \dots & D_n f_m(x) \end{bmatrix}.$$

Remark. The above matrix of partial derivatives is also known as the **Jacobian matrix**. Thus the theorem says that if a function is differentiable, its derivative is equal to its Jacobian matrix.

Proof. The ijth entry of the matrix $Df(x)$ is the ith entry of its jth column. But by the previous theorems the jth column of $Df(x)$ is $D_j f(x)$, and the ith entry of $D_j f(x)$ is $D_j f_i(x)$. ■

***Example* 7.10.** Consider the function $f : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ that maps each matrix A to $f(A) = A^2$. Let us for the moment assume that f is differentiable, and compute Df. First note that we can think of matrices in $\mathbb{R}^{n\times n}$ as vectors in $\mathbb{R}^{n^2}$. Now the ijth entry of $f(A)$, or in other words its ijth component, is

$$(f(A))_{ij} = A_{i1}A_{1j} + A_{i2}A_{2j} + \cdots + A_{in}A_{nj}.$$

Therefore we have

$$\begin{aligned} D_{A_{il}} f_{ij} &= A_{lj}, & D_{A_{ij}} f_{ij} &= A_{ii} + A_{jj}, \\ D_{A_{kj}} f_{ij} &= A_{ik}, & D_{A_{kl}} f_{ij} &= 0, \end{aligned}$$

where $k \neq i, l \neq j$. Note that when $i = j$, the above formula gives the correct value $D_{A_{ii}} f_{ii} = 2A_{ii}$. Also note that Df is an $n^2 \times n^2$ matrix. Next let us compute the value of $Df(A)B$, where $B \in \mathbb{R}^{n\times n}$. Keep in mind that we use double indices ij to denote the components of the n^2-dimensional vector B. The same is true about the matrix $Df(A)$. We have

$$\begin{aligned} (Df(A)B)_{ij} &= \sum_{k,l} (Df(A))_{ij,kl} B_{kl} = \sum_{k,l} D_{A_{kl}} f_{ij}(A) B_{kl} \\ &= \sum_{l\neq j} D_{A_{il}} f_{ij}(A) B_{il} + \sum_{k\neq i} D_{A_{kj}} f_{ij}(A) B_{kj} \\ &\qquad + D_{A_{ij}} f_{ij}(A) B_{ij} + \sum_{k\neq i, l\neq j} D_{A_{kl}} f_{ij}(A) B_{kl} \\ &= \sum_{l\neq j} A_{lj} B_{il} + \sum_{k\neq i} A_{ik} B_{kj} + (A_{ii} + A_{jj}) B_{ij} \\ &= \sum_{l\leq n} A_{lj} B_{il} + \sum_{k\leq n} A_{ik} B_{kj} = (BA)_{ij} + (AB)_{ij}. \end{aligned}$$

Therefore we have

$$Df(A)B = BA + AB.$$

Note that this is the generalization of the familiar formula $(x^2)' = 2x$ to the space of matrices, where the multiplication is not commutative. Finally, we can easily show that f is differentiable; because the remainder

$$f(A + B) - f(A) - Df(A)B = (A + B)^2 - A^2 - BA - AB = B^2$$

is easily seen to be sublinear.

Theorem 7.11. *Suppose f is differentiable at x, then it is continuous at x too.*

Proof. For y near x we have

$$\lim_{y\to x} f(y) = \lim_{y\to x}[f(x) + Df(x)(y-x) + R(y-x)] = f(x).$$

Note that here we used the continuity of R at 0, and the continuity of the linear map defined by the matrix $Df(x)$. (Remember that polynomial functions between Euclidean spaces are continuous.) ■

***Example* 7.12.** A function that has directional derivatives in every direction at a point is not necessarily differentiable there; it is not even necessarily continuous there. As an example, consider $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y) = \begin{cases} \frac{x^2y}{x^4+y^2} & (x,y) \neq (0,0), \\ 0 & (x,y) = (0,0). \end{cases}$$

Then f is not continuous at $(0,0)$, since if we approach the origin along the parabola $y = x^2$ the value of f approaches $\frac{1}{2}$. Thus f cannot be differentiable at $(0,0)$. But f has directional derivatives at the origin in every direction. Let $v = (a,b)$ be an arbitrary nonzero vector in $\mathbb{R}^2$. Then we have

$$D_vf(0,0) = \lim_{t\to 0}\frac{f(ta,tb) - f(0,0)}{t} = \lim_{t\to 0}\frac{t^3a^2b}{t(t^4a^4 + t^2b^2)} = \begin{cases} \frac{a^2}{b} & b \neq 0, \\ 0 & b = 0. \end{cases}$$

Note that unlike differentiable functions, $D_vf(0,0)$ does not depend linearly on the vector v.

Remark. Another interesting property of the function f in the above example is that its restriction to every line passing through the origin is continuous at the origin (since f has directional derivative in every direction), but f is not continuous at the origin as a function of two variables.

***Example* 7.13.** Even if $D_vf(x)$ exists for every v and depends linearly on v, we cannot conclude that f is continuous at x. For example consider

$$f(x,y) = \begin{cases} 1 & y = x^2 \text{ and } x \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then f is clearly discontinuous at $(0,0)$, but $D_vf(0,0) = 0$ for every vector v. A more interesting example is the function

$$g(x,y) = \begin{cases} x & y = x^2, \\ 0 & \text{otherwise,} \end{cases}$$

which is continuous at $(0,0)$ and $D_vg(0,0) = 0$ for every vector v. But g is not differentiable at $(0,0)$ (why?).

Remark. Suppose $D_v f(x)$ exists for every vector v and depends linearly on v. Let A be the matrix whose jth column is $D_j f(x)$. Then if $v = (v_1, \dots, v_n)$ we have

$$D_v f(x) = \sum v_j D_j f(x) = Av.$$

Now set

$$R(h) := f(x+h) - f(x) - Ah.$$

Then when $h = tv$ for some $t > 0$ we have

$$\frac{R(h)}{|h|} = \frac{1}{|v|}\Big(\frac{f(x+tv) - f(x)}{t} - Av\Big) \xrightarrow[t\to 0]{} \frac{1}{|v|}\big(D_v f(x) - Av\big) = 0.$$

Therefore R is sublinear when we approach 0 in the direction of v. In other words, for every $\epsilon > 0$ there is $\delta_v > 0$, such that when $|h| < \delta_v$, and h is a multiple of v, then

$$\left|\frac{R(h)}{|h|}\right| < \epsilon.$$

But this is not enough for f to be differentiable at x. Because δ_v depends on v, while in the definition of differentiability we require $\frac{R(h)}{|h|}$ to go to zero as a function of several variables, i.e. we need to be able to choose a δ that works for all v.

***Exercise* 7.14.** Check that the nondifferentiable function g of the previous example has a remainder at $(0, 0)$ that is sublinear in every direction but is not sublinear as a function of two variables.

Theorem 7.15. *Suppose $U \subset \mathbb{R}^n$ is an open set containing a point x, and $f = (f_1, \dots, f_m) : U \to \mathbb{R}^m$. Then f is differentiable at x if and only if each f_i is differentiable at x. And in this case Df_i is the ith row of Df, i.e.*

$$Df(x) = \begin{bmatrix} Df_1(x) \\ \hline \vdots \\ \hline Df_m(x) \end{bmatrix}.$$

Proof. Suppose that f is differentiable at x. Then there exist an $m \times n$ matrix A and a sublinear function R, so that for small $h \in \mathbb{R}^n$ we have

$$f(x+h) = f(x) + Ah + R(h).$$

Thus for each i, the ith components of both sides are equal, i.e. for small $h \in \mathbb{R}^n$ we have

$$f_i(x+h) = f_i(x) + (Ah)_i + R_i(h) = f_i(x) + A_{i,\cdot}h + R_i(h),$$

where $A_{i,\cdot}$ is the ith row of the matrix A. Note that we also have $\lim_{h\to 0} R_i(h)/|h| = 0$, since $R_i(h)/|h|$ is the ith component of $R(h)/|h|$, and $\lim_{h\to 0} R(h)/|h| = 0$. Hence f_i is differentiable at x, and its derivative is the ith row of $Df(x)$.

Conversely, suppose each f_i is differentiable at x. Let A be the $m \times n$ matrix whose ith row is $Df_i(x)$. Suppose R_i is the remainder in the differentiability relation of f_i. Then let R be the function into $\mathbb{R}^m$ defined on a neighborhood of $0 \in \mathbb{R}^n$, whose ith component is R_i. Note that each R_i is defined on an open ball around $0 \in \mathbb{R}^n$, so R is defined on the open ball with the smallest radius. Now for h in this smallest open ball we have

$$f(x+h) = f(x) + Ah + R(h).$$

Because for each i, the ith components of both side are equal, as we have

$$f_i(x) + (Ah)_i + R_i(h) = f_i(x) + D_i f(x)h + R_i(h) = f_i(x+h).$$

Also similarly to the above, we can show that R is sublinear, since its components are sublinear. Therefore f is differentiable at x, and has the required derivative. ■

Proposition 7.16. *Suppose $x, v \in \mathbb{R}^n$, U is a neighborhood of x, and $f : U \to \mathbb{R}^m$. Also suppose that f has directional derivative in the v direction at the points near x that lie on the line*

$$\{x + sv : s \in \mathbb{R}\}.$$

Let $a \in \mathbb{R}$, and let $g(t) := f(x + atv)$ be a function from a neighborhood of $0 \in \mathbb{R}$ into $\mathbb{R}^m$. Then we have

$$g'(t) = aD_v f(x + atv).$$

Proof. If $a = 0$ then g is constant and the relation holds obviously. When $a \neq 0$ we set $y = x + atv$ and $h = \frac{s}{a}$ to obtain

$$\lim_{h\to 0} \frac{g(t+h) - g(t)}{h} = \lim_{s\to 0} a\,\frac{f(y+sv) - f(y)}{s} = aD_v f(y) = aD_v f(x + atv). \quad ■$$

Remark. In the above proposition, when f is differentiable, the formula for g' is a consequence of the chain rule as we will show later. But here we only assumed that f has directional derivative in one direction. This version with the weaker hypothesis is sometimes useful when we do not know a priori that f is differentiable.

Theorem 7.17. *If a function f has partial derivatives in a neighborhood of a point $x \in \mathbb{R}^n$, and the partial derivatives are all continuous at x; then f is differentiable at x.*

Proof. It is sufficient to prove the differentiability of each component of the function f, so without loss of generality we can assume that f is real-valued. Suppose $h = (h_1, \dots, h_n) \in \mathbb{R}^n$ is a point with small $|h|$. Let $y_j := (h_1, \dots, h_j, 0, \dots, 0)$. Then we have

$$f(x+h) - f(x) - \sum_{j=1}^{n} D_j f(x) h_j = \sum_{j=1}^{n} \big[f(x+y_j) - f(x+y_{j-1}) - D_j f(x) h_j \big].$$

Now let $g_j(t) := f(x + y_{j-1} + t h_j e_j)$, where e_j's are the standard basis vectors of $\mathbb{R}^n$. Then by the mean value theorem for some $t_j \in (0,1)$ we have

$$f(x+y_j) - f(x+y_{j-1}) = g_j(1) - g_j(0) = g_j'(t_j) = h_j D_j f(x + y_{j-1} + t_j h_j e_j).$$

Let $\theta_j := x + y_{j-1} + t_j h_j e_j$. Hence we have

$$\begin{aligned} &\frac{1}{|h|} \Big| f(x+h) - f(x) - \sum D_j f(x) h_j \Big| \\ &\qquad = \Big| \sum \big[D_j f(\theta_j) - D_j f(x) \big] \frac{h_j}{|h|} \Big| \le \sum |D_j f(\theta_j) - D_j f(x)|. \end{aligned}$$

The last expression goes to zero as $h \to 0$, since the partial derivatives are continuous at x, and $\theta_j \to x$ as $h \to 0$. Thus f is differentiable at x, and its derivative is the $1 \times n$ matrix whose jth column is $D_j f(x)$. ■

***Example* 7.18.** The converse of the above theorem is not true. A function can be differentiable at a point while its partial derivatives are discontinuous. For example

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

is differentiable at $(0,0)$, but its partial derivatives are unbounded as we approach the origin.

7.2 Rules of Differentiation

Theorem 7.19. *Suppose $U \subset \mathbb{R}^n$ is an open set containing a point x, and $f, g : U \to \mathbb{R}^m$ are differentiable at x. Let $c_1, c_2 \in \mathbb{R}$. Then $c_1 f + c_2 g$ is differentiable at x and*

$$D(c_1 f + c_2 g)(x) = c_1 Df(x) + c_2 Dg(x).$$

Proof. We have

$$\begin{aligned} (c_1 f + c_2 g)(x+h) &= (c_1 f + c_2 g)(x) \\ &\qquad + \big(c_1 Df(x) + c_2 Dg(x)\big)h + |h|\big(c_1 r_f(h) + c_2 r_g(h)\big), \end{aligned}$$

where r_f, r_g are the remainders in the differentiability relations of f, g, respectively. Then we have

$$\lim_{h\to 0} c_1 r_f(h) + c_2 r_g(h) = 0.$$

Thus the last term is sublinear and $c_1 f + c_2 g$ is differentiable with the desired derivative. ■

Definition 7.20. A map $L : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k} \to \mathbb{R}^m$ is called **multilinear**, or ***k*-linear**, if L is linear with respect to each variable when the other variables are fixed, i.e. for any $i \le k$ we have

$$\begin{aligned} &L(v_1, \dots, v_{i-1}, c_1 v_i + c_2 w_i, v_{i+1}, \dots, v_k) \\ &\qquad = c_1 L(v_1, \dots, v_{i-1}, v_i, v_{i+1}, \dots, v_k) + c_2 L(v_1, \dots, v_{i-1}, w_i, v_{i+1}, \dots, v_k), \end{aligned}$$

for every vectors $v_1, \dots, v_k, w_i$ and scalars $c_1, c_2 \in \mathbb{R}$.

Remark. A 2-linear maps is also known as a **bilinear** map. In addition, note that a 1-linear map is just a linear map.

Theorem 7.21. *Suppose $L : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k} \to \mathbb{R}^m$ is a multilinear map. Then there exists $C > 0$ such that*

$$|L(v_1, \dots, v_k)| \le C|v_1| \cdots |v_k|,$$

for all $(v_1, \dots, v_k) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$. As a result L is continuous.

Remark. A particular case of the above theorem is when $k = 1$ and we have a linear map defined by an $m \times n$ matrix A. Then there is $C > 0$ such that for all $v \in \mathbb{R}^n$ we have

$$|Av| \le C|v|.$$

This implies that for two vectors v, w

$$|Av - Aw| = |A(v - w)| \le C|v - w|,$$

i.e. linear maps are Lipschitz.

Proof. The proof is by induction on k. For the induction base we have $k = 1$, i.e. L is a linear map from $\mathbb{R}^n$ into $\mathbb{R}^m$. Let $e_1, \dots, e_n$ be the standard basis of $\mathbb{R}^n$, and let $v = a_1 e_1 + \cdots + a_n e_n$ be an arbitrary vector in $\mathbb{R}^n$. Then we have

$$|Lv| = |a_1 L e_1 + \cdots + a_n L e_n| \le |a_1||Le_1| + \cdots + |a_n||Le_n| \le C|v|,$$

where $C = |Le_1| + \cdots + |Le_n|$. Note that here we used the special property of the standard norm of $\mathbb{R}^n$ that $|a_i| \le |v|$.

Now suppose $k > 1$ and the claim holds for $(k-1)$-linear maps. Define

$$L_v(v_2, \dots, v_k) := L(v, v_2, \dots, v_k).$$

Then it is easy to see that L_v is a $(k-1)$-linear map. Hence we have

$$|L_v(v_2, \dots, v_k)| \le C_v |v_2| \dots |v_k|,$$

for some constant $C_v > 0$. Now let $v_1 = \sum a_i e_i$ where e_i's are the standard basis of $\mathbb{R}^{n_1}$. We have

$$\begin{aligned} |L(v_1, v_2, \dots, v_k)| &= \Big|\sum a_i L(e_i, v_2, \dots, v_k)\Big| \le \sum |a_i|\, |L_{e_i}(v_2, \dots, v_k)| \\ &\le \sum |v_1| C_{e_i} |v_2| \cdots |v_k| = \big(\sum C_{e_i}\big) |v_1||v_2| \cdots |v_k|. \end{aligned}$$

To prove the continuity suppose $|w_i - v_i| < 1$. So $|w_i| < D := \max_{i \le k}(|v_i| + 1)$. Then we have

$$\begin{aligned} &|L(v_1, v_2, \dots, v_k) - L(w_1, w_2, \dots, w_k)| \\ &\quad = \Big|\sum\nolimits_{i \le k} \big(L(w_1, \dots, w_{i-1}, v_i, v_{i+1}, \dots, v_k) - L(w_1, \dots, w_{i-1}, w_i, v_{i+1}, \dots, v_k)\big)\Big| \\ &\quad \le \sum \big|L(w_1, \dots, w_{i-1}, v_i, v_{i+1}, \dots, v_k) - L(w_1, \dots, w_{i-1}, w_i, v_{i+1}, \dots, v_k)\big| \\ &\quad = \sum \big|L(w_1, \dots, w_{i-1}, v_i - w_i, v_{i+1}, \dots, v_k)\big| \\ &\quad \le \sum C|w_1| \cdots |w_{i-1}||v_{i+1}| \cdots |v_k||v_i - w_i| \le CD^{k-1} \sum |v_i - w_i|. \end{aligned}$$

Thus when w_i is close to v_i for each i, $L(w_1, \dots, w_k)$ is close to $L(v_1, \dots, v_k)$. ■

Leibniz Rule. *Suppose $U \subset \mathbb{R}^n$ is an open set containing a point x, and $f : U \to \mathbb{R}^m$ and $g : U \to \mathbb{R}^p$ are differentiable at x. Also suppose $B : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^q$ is bilinear. Then $B[f, g]$ is differentiable at x, and for every $v \in \mathbb{R}^n$ we have*

$$D(B[f, g])(x)v = B[Df(x)v, g(x)] + B[f(x), Dg(x)v].$$

In particular, when $n = 1$ we have

$$(B[f, g])'(x) = B[f'(x), g(x)] + B[f(x), g'(x)].$$

Remark. If we use the product notation $f \star g := B[f, g]$, and the directional derivative notation, the Leibniz rule can be written in the more familiar form

$$D_v(f \star g)(x) = D_v f(x) \star g(x) + f(x) \star D_v g(x).$$

Proof. Let $F = B[f,g]$, $A = Df(x)$ and $E = Dg(x)$. Let R, S be respectively the remainders in the differentiability relations of f, g. Then

$$\begin{aligned} F(x+h) &= B[f(x+h), g(x+h)] \\ &= B\big[f(x) + Ah + R(h),\, g(x) + Eh + S(h)\big] \\ &= F(x) + B[Ah, g(x)] + B[f(x), Eh] + B[Ah, Eh] \\ &\qquad + B[R(h), g(x+h)] + B[f(x) + Ah, S(h)]. \end{aligned}$$

The 2nd and 3rd terms in the last expression are linear in h, and the matrix of their sum is the required derivative. So, it suffices to show that the last three terms in the above formula are sublinear. As B is bilinear we have $|B[a,b]| \le C|a||b|$ for some constant C. Also there are constants C_A, C_E such that $|Ah| \le C_A|h|$, and $|Eh| \le C_E|h|$. Therefore

$$\frac{|B[Ah, Eh]|}{|h|} \le \frac{C|Ah||Eh|}{|h|} \le CC_AC_E|h| \xrightarrow[h\to 0]{} 0.$$

Thus the forth term is sublinear by the squeeze theorem. For the fifth term we have

$$\frac{1}{|h|}B[R(h), g(x+h)] = B\Big[\frac{R(h)}{|h|}, g(x+h)\Big] \xrightarrow[h\to 0]{} B[0, g(x)] = 0.$$

And for the last term we have

$$\frac{1}{|h|}B[f(x) + Ah, S(h)] = B\Big[f(x) + Ah, \frac{S(h)}{|h|}\Big] \xrightarrow[h\to 0]{} B[f(x), 0] = 0.$$

Note that B is continuous; and it vanishes when one of its components is zero, due to its bilinearity. ■

Remark. Let $e_1, e_2, \dots$ be the standard basis of the Euclidean space $\mathbb{R}^m$. We will use the same notation for these vectors in every dimension. Let $B : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^q$ be a bilinear map. For $k \le m$ and $l \le p$, consider the vector $B[e_k, e_l] \in \mathbb{R}^q$. Then there are unique real numbers B^i_{kl} such that

$$B[e_k, e_l] = \sum_{i\le q} B^i_{kl} e_i.$$

The numbers B^i_{kl} are called the *components* of the bilinear map B. Let $\langle\,,\rangle$ be the standard inner product on Euclidean spaces. Then we have

$$B^i_{kl} = \langle B[e_k, e_l], e_i\rangle.$$

Now for all $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^p$ there are unique real numbers u_k, v_l such that $u = \sum_{k\leq m} u_k e_k$, and $v = \sum_{l\leq p} v_l e_l$. Hence by using the bilinearity of B we get

$$B[u,v] = \sum_{i\leq q}\sum_{k\leq m}\sum_{l\leq p}(B^i_{kl}u_k v_l)e_i.$$

Note that $u_k = \langle u, e_k\rangle$ and $v_l = \langle v, e_l\rangle$. Thus we also have

$$\langle B[u,v], e_i\rangle = \sum_{k\leq m}\sum_{l\leq p} B^i_{kl}u_k v_l = \sum_{k\leq m}\sum_{l\leq p} B^i_{kl}\langle u, e_k\rangle\langle v, e_l\rangle.$$

Therefore in the Leibniz rule, the entries of $D(B[f,g])(x)$ are

$$\begin{aligned}
D_j(B[f,g])_i(x) &= \big\langle D(B[f,g])(x)e_j, e_i\big\rangle \\
&= \big\langle B[Df(x)e_j, g(x)] + B[f(x), Dg(x)e_j], e_i\big\rangle \\
&= \big\langle B[D_jf(x), g(x)] + B[f(x), D_jg(x)], e_i\big\rangle \\
&= \sum_{k\leq m}\sum_{l\leq p} B^i_{kl}\big(D_jf_k(x)g_l(x) + f_k(x)D_jg_l(x)\big).
\end{aligned}$$

Note that $\langle D_jf(x), e_k\rangle = D_jf_k(x)$, and $\langle D_jg(x), e_l\rangle = D_jg_l(x)$.

***Example* 7.22.** Some particular cases of the Leibniz rule are when the bilinear pairing is the standard inner product on $\mathbb{R}^m$, or the exterior product on $\mathbb{R}^3$. An interesting particular case is when $m = pq$, and the bilinear pairing is the action of the $q \times p$ matrix $f(x)$ on the p-dimensional vector $g(x)$.

Chain Rule. *Suppose $U \subset \mathbb{R}^n$ is an open set containing a point x, and $f : U \to \mathbb{R}^m$ is differentiable at x. Also suppose V is a neighborhood of $f(x)$, and $g : V \to \mathbb{R}^p$ is differentiable at $f(x)$. Then $g \circ f$ is differentiable at x and*

$$D(g\circ f)(x) = Dg(f(x))Df(x).$$

Remark. On the right hand side of the above formula we have the product of two matrices. When we think in terms of the linear maps defined by these matrices, their product is just the matrix of the composition of those linear maps. In other words, the chain rule says that the derivative of the composition of two functions is the composition of their derivatives.

Remark. If we use x_j for the coordinates of $\mathbb{R}^n$ and y_k for the coordinates of $\mathbb{R}^m$, then the ijth entry of the product matrix in the chain rule is

$$\frac{\partial}{\partial x_j}(g\circ f)_i = \sum_{k=1}^{m}\frac{\partial g_i}{\partial y_k}\frac{\partial f_k}{\partial x_j},$$

which is the familiar form of chain rule from Calculus courses.

Proof. Let $A = Df(x)$ and $B = Dg(f(x))$. Also let R, s be the remainders in the differentiability relations of f, g, respectively. Then we have

$$g(f(x) + \tilde{h}) = g(f(x)) + B\tilde{h} + |\tilde{h}|s(\tilde{h}).$$

Now by substituting $\tilde{h}$ with $Ah + R(h)$ we get

$$\begin{aligned} g(f(x+h)) &= g\big(f(x) + Ah + R(h)\big) \\ &= g(f(x)) + B(Ah) + B(R(h)) + |Ah + R(h)|\, s(Ah + R(h)). \end{aligned}$$

We have to show that the last two terms are sublinear. For the first one we have

$$\frac{B(R(h))}{|h|} = B\Big(\frac{R(h)}{|h|}\Big) \xrightarrow[h\to 0]{} B(0) = 0.$$

Now for the last term, let C_A be a positive constant such that $|Ah| \leq C_A|h|$. Then

$$\begin{aligned} \frac{1}{|h|}|Ah + R(h)||s(Ah + R(h))| &\leq \big(C_A + \frac{|R(h)|}{|h|}\big)|s(Ah + R(h))| \\ &\xrightarrow[h\to 0]{} (C_A + 0)|s(0)| = 0. \end{aligned}$$

Note that $|s|$ is continuous at zero. Hence we get the desired result by the squeeze theorem. ■

Theorem 7.23. *Suppose $U, V \subset \mathbb{R}^n$ are open, and $f : U \to V$ is invertible. If f is differentiable at x and f^{-1} is differentiable at $f(x)$, then $Df(x)$ is an invertible matrix and we have*

$$Df^{-1}(f(x)) = (Df(x))^{-1}.$$

Proof. Note that $f^{-1} \circ f = \mathrm{id}_U$, and $D(\mathrm{id}_U)(x) = I$ where I is the identity matrix. Then the chain rule implies that

$$Df^{-1}(f(x))Df(x) = I.$$

Hence the theorem follows. ■

Theorem 7.24. *Suppose $U \subset \mathbb{R}^n$ is an open set that contains the closed line segment joining the two points a, b, i.e.*

$$I := \{a + t(b-a) : t \in [0,1]\} \subset U.$$

Also suppose $f : U \to \mathbb{R}^m$ is differentiable at every point of I, and its partial derivatives are bounded on I. Then we have

$$|f(b) - f(a)| \leq M|b-a|,$$

where $M := \sup\limits_{x\in I} \big(\sum\limits_{i,j} |D_j f_i(x)|^2\big)^{\frac{1}{2}}$.

Proof. First note that f is continuous on I since it is differentiable there. Let

$$g(t) := [f(b) - f(a)] \cdot [f(a + t(b - a)) - f(a)].$$

Then g is a real-valued function of a single variable $t \in [0, 1]$ that satisfies the conditions of the mean value theorem. Thus for some $\tau \in (0, 1)$ we have

$$\begin{aligned}
|f(b) - f(a)|^2 &= g(1) - g(0) = g'(\tau) \\
&= [f(b) - f(a)] \cdot [Df(a + \tau(b - a))(b - a)] \\
&\leq |f(b) - f(a)|\,|Df(a + \tau(b - a))(b - a)|.
\end{aligned}$$

If $f(b) - f(a) = 0$ then the estimate holds trivially. Otherwise for $\theta := a + \tau(b - a)$ we have

$$\begin{aligned}
|f(b) - f(a)| &\leq |Df(\theta)(b - a)| \\
&= \Big(\sum_{i=1}^{m} \big(Df_i(\theta) \cdot (b - a)\big)^2\Big)^{\frac{1}{2}} \leq \Big(\sum_{i=1}^{m} \big(|Df_i(\theta)|\,|b - a|\big)^2\Big)^{\frac{1}{2}} \\
&= |b - a|\Big(\sum_{i=1}^{m} |Df_i(\theta)|^2\Big)^{\frac{1}{2}} = |b - a|\Big(\sum_{i=1}^{m}\sum_{j=1}^{n} |D_j f_i(\theta)|^2\Big)^{\frac{1}{2}} \\
&\leq M|b - a|.
\end{aligned}$$

■

Theorem 7.25. *Suppose $U \subset \mathbb{R}^n$ is an open set that contains the closed line segment joining the two points a, b, i.e.*

$$I := \{a + t(b - a) : t \in [0, 1]\} \subset U.$$

Also suppose $f : U \to \mathbb{R}^m$ is differentiable at every point of I, and its partial derivatives are continuous on I. Then we have

$$f(b) - f(a) = \int_0^1 Df\big(a + t(b - a)\big)(b - a)\, dt.$$

Remark. Let $v := b - a$. Note that the integrand in the above formula is the vector $Df(a + tv)v = D_v f(a + tv)$. Also note that as shown in Theorem 8.9, the integral of a vector-valued function is the vector whose ith component is the integral of the ith component of the function.

Proof. Let $g(t) := f(a + t(b - a))$. Then g is a function of a single variable t, whose domain is an open interval containing $[0, 1]$. Also, due to the chain rule, g is differentiable at every point of $[0, 1]$. Let $v := b - a$. Then we have $g'(t) = Df(a + tv)v$. Thus for every $i \leq m$ we have $g_i'(t) = \sum_{j \leq n} D_j f_i(a + tv)v_j$. So g_i'

is continuous on $[0,1]$, since the partial derivatives of f are continuous on the line segment I. Hence by the fundamental theorem of calculus we have

$$g_i(1) - g_i(0) = \int_0^1 g_i'(t)dt.$$

Therefore

$$f(b) - f(a) = g(1) - g(0) = \int_0^1 g'(t)dt = \int_0^1 Df(a+tv)v\,dt. \qquad ■$$

Remark. Note that Df is an $m \times n$ matrix, which can be considered as a vector in $\mathbb{R}^{mn}$. Thus by Theorem 8.9, $A := \int_0^1 Df(a+tv)\,dt$ is also an $m \times n$ matrix whose ijth entry is $\int_0^1 D_j f_i(a+tv)\,dt$. Now, in the above theorem we can also write

$$\int_0^1 Df_i(a+tv)v\,dt = \int_0^1 \sum_{j\leq n} D_j f_i(a+tv)v_j\,dt = \sum_{j\leq n} v_j \int_0^1 D_j f_i(a+tv)\,dt.$$

Hence we have $\int_0^1 Df(a+tv)v\,dt = Av$. Therefore we get

$$f(b) - f(a) = \int_0^1 Df(a+tv)v\,dt = Av = \Big(\int_0^1 Df(a+t(b-a))\,dt\Big)(b-a).$$

Finally, let us mention that we can think of $\int_0^1 Df(a+tv)\,dt$ as the mean of Df along the line segment I.

Theorem 7.26. *Suppose $U \subset \mathbb{R}^n$ is an open set that contains $V \times [a,b]$ where $V \subset \mathbb{R}^{n-1}$ is open. Also suppose that $f : U \to \mathbb{R}$ is continuous, and for some $i < n$, $D_i f$ is continuous. Let $g : V \to \mathbb{R}$ be given by*

$$g(x_1,\dots,x_{n-1}) := \int_a^b f(x_1,\dots,x_{n-1},x_n)\,dx_n.$$

Then $D_i g$ exists and we have

$$D_i g(x_1,\dots,x_{n-1}) = \int_a^b D_i f(x_1,\dots,x_{n-1},x_n)\,dx_n.$$

Remark. In other words, this theorem says that under suitable conditions we can change the order of differentiation and integration, or as it is commonly referred to, we can *differentiate under the integral sign.*

Remark. For simplicity of the notation, we assumed that we are integrating with respect to x_n; but similar results hold when we integrate with respect to x_j, provided that $j \neq i$. Of course, when $j = i$, i.e. when we are integrating and differentiating with respect to the same variable, we can use the fundamental theorem of calculus, since every other variable is fixed in this process.

Proof. We fix every variable other than x_i and x_n, and will suppress them in the notation. Let

$$r(x_i,h) := \frac{1}{h}\Big(g(x_i+h)-g(x_i)-h\int_a^b D_i f(x_i,x_n)\,dx_n\Big).$$

We need to show that $\lim_{h\to 0} r(x_i,h)=0$. First note that for some $\delta_0>0$ we have

$$K := \{(x_1,\ldots,x_{i-1})\}\times[x_i-\delta_0,x_i+\delta_0]\times\{(x_{i+1},\ldots,x_{n-1})\}\times[a,b]\subset U,$$

since $(x_1,\ldots,x_i,\ldots,x_{n-1})\in V$, and V is open. Also note that K is compact. Now for $|h|\leq\delta<\delta_0$ we have

$$\begin{aligned}|r(x_i,h)| &= \Big|\frac{1}{h}\Big(g(x_i+h)-g(x_i)-h\int_a^b D_i f(x_i,x_n)\,dx_n\Big)\Big| \\ &= \Big|\frac{1}{h}\int_a^b f(x_i+h,x_n)-f(x_i,x_n)-hD_i f(x_i,x_n)\,dx_n\Big| \\ &\leq \frac{b-a}{|h|}\max_{|h|\leq\delta,\,x_n\in[a,b]}\big|f(x_i+h,x_n)-f(x_i,x_n)-hD_i f(x_i,x_n)\big|,\end{aligned}$$

since $f(x_i+h,x_n)-f(x_i,x_n)-hD_i f(x_i,x_n)$ is continuous on the compact set K, and hence it is bounded on K.

Next, for some fixed values of h,x_i,x_n, we consider the real-valued function $p(t):=f(x_i+th,x_n)$ of one variable t. Then by the mean value theorem there exist $\tau\in(0,1)$ such that

$$f(x_i+h,x_n)-f(x_i,x_n)=p(1)-p(0)=p'(\tau)=hD_i f(x_i+\tau h,x_n).$$

Since τ can depend on h,x_n, we denote it by $\tau(h,x_n)$. Note that the dependence of τ on x_i does not concern us at this moment, because we want to estimate $r(x_i,h)$ when $h\to 0$ and x_i is fixed. By combining the above relations we get

$$\begin{aligned}|r(x_i,h)| &\leq \frac{b-a}{|h|}\max_{|h|\leq\delta,\,x_n\in[a,b]}\big|hD_i f(x_i+\tau(h,x_n)h,x_n)-hD_i f(x_i,x_n)\big| \\ &= (b-a)\max_{|h|\leq\delta,\,x_n\in[a,b]}\big|D_i f(x_i+\tau(h,x_n)h,x_n)-D_i f(x_i,x_n)\big|.\end{aligned}$$

But the distance of the two points $(x_i+\tau(h,x_n)h,x_n)$ and (x_i,x_n) is at most δ, since $\tau\in(0,1)$ and $|h|<\delta$. On the other hand, $D_i f$ is continuous on the compact set K. Therefore it is uniformly continuous on K. Hence for a given $\epsilon>0$ there exists $\delta<\delta_0$, such that for all $|h|<\delta$ and $x_n\in[a,b]$ we have

$$\big|D_i f(x_i+\tau(h,x_n)h,x_n)-D_i f(x_i,x_n)\big|<\epsilon.$$

Thus $r(x_i,h)\to 0$ as $h\to 0$; and consequently, g has the required ith partial derivative. ■

***Exercise* 7.27.** In the above theorem, it can be shown that $D_i g$ is also continuous. More generally, suppose $U \subset \mathbb{R}^n$ contains $V \times [a,b]$ where $V \subset \mathbb{R}^{n-1}$, and $f : U \to \mathbb{R}$ is continuous. Also suppose $\phi, \psi : V \to [a,b]$ are continuous functions with $\phi \leq \psi$. Show that the function $\tilde{f} : V \to \mathbb{R}$ given by

$$\tilde{f}(x_1, \dots, x_{n-1}) := \int_{\phi(x_1,\dots,x_{n-1})}^{\psi(x_1,\dots,x_{n-1})} f(x_1, \dots, x_{n-1}, x_n)\, dx_n$$

is continuous.

***Exercise* 7.28.** Let

$$f(x,y) = \begin{cases} \frac{x}{|x|} y & 0 < y \leq \sqrt{|x|}, \\ \frac{x}{|x|}(-y + 2\sqrt{|x|}) & 0 < \sqrt{|x|} \leq y \leq 2\sqrt{|x|}, \\ 0 & \text{otherwise.} \end{cases}$$

Show that $\frac{d}{dx} \int_{-1}^{1} f(x,y)dy \neq \int_{-1}^{1} \frac{\partial}{\partial x} f(x,y)dy$ at $x = 0$.

7.3 Higher Derivatives

Definition 7.29. Let $U \subset \mathbb{R}^n$ be an open set and $f : U \to \mathbb{R}^m$. The ***k*th order partial derivatives** of f are inductively defined to be the partial derivatives of the $(k-1)$th order partial derivatives of f. Note that the first order partial derivatives of f are just the partial derivatives of f. The kth order partial derivatives are also called the partial derivatives of order k. We denote them by

$$D^k_{i_1 i_2 \dots i_k} f_j = \frac{\partial^k f_j}{\partial x_{i_1} \partial x_{i_2} \dots \partial x_{i_k}} := D_{i_1}(D_{i_2}(\dots(D_{i_k} f_j))).$$

Notation. Since the order of a partial derivative is apparent from the number of indices in it, we also denote the above partial derivative simply by $D_{i_1 i_2 \dots i_k} f_j$.

Definition 7.30. We say a function f is of **class C^k** if it is continuous, and has continuous partial derivatives of orders $1, 2, \dots, k$ on its domain. The function f is called **infinitely differentiable** or **smooth** or of **class C^∞** if it is continuous, and has continuous partial derivatives of all orders on its domain. Finally, we consider the 0th order partial derivative of f to be f itself, and we say f is of **class C^0** if it is continuous on its domain.

Remark. Note that the $(k+1)$th order partial derivatives of a function f are the kth order partial derivatives of the partial derivatives of f. This can be proved by an easy induction on k. Consequently, a function f is C^{k+1} if and only if its partial derivatives are C^k. Note that a function whose partial derivatives are continuous, is differentiable, and therefore is continuous too. Also, it is trivial that a function f is C^k if and only if each component of f is C^k.

Remark. Suppose $U \subset \mathbb{R}^n$ is open and $f : U \to \mathbb{R}^m$ is differentiable. Then for every $x \in U$, $Df(x)$ is an $m \times n$ matrix, i.e. $Df : U \to \mathbb{R}^{m\times n}$. But $\mathbb{R}^{m\times n}$ as a normed space is equivalent to $\mathbb{R}^{mn}$. Thus we can talk about the continuity and differentiability of the matrix-valued map Df. For example we can say f is C^1 if Df is continuous. But this implies that the entries of the matrix Df are continuous functions, i.e. the partial derivatives of f are continuous. On the other hand if the partial derivatives of f are continuous then we know that f is differentiable. In addition we know that Df is continuous since its entries are continuous. Thus the two definitions of C^1 functions are equivalent.

Let us also check the equivalence of the two definitions when $k = 2$, i.e. for C^2 functions. We can say f is twice differentiable if Df is differentiable. Then we can think of the 2nd derivative at every point as an $mn\times n$ matrix, i.e. $D^2 f : U \to \mathbb{R}^{mn^2}$. Now we can say f is C^2 if $D^2 f$ is continuous. Since we assumed that Df is differentiable, we know that the 2nd partial derivatives of f exist. It can also be checked easily that the 2nd partial derivatives are the entries of $D^2 f$. Hence the 2nd partial derivatives of f are also continuous, and f is C^2 with respect to the first definition. Conversely, if f is C^2 with respect to the first definition then the 2nd partial derivatives of f are continuous. But the 2nd partial derivatives are the partial derivatives of Df, so Df is differentiable. Finally $D^2 f$ is continuous since its entries i.e. the 2nd partial derivatives are continuous.

Definition 7.31. Let $U \subset \mathbb{R}^n$ be an open set, and suppose $f : U \to \mathbb{R}$ has partial derivatives of the second order at some point $x \in U$. Then the **Hessian matrix** of f at x is the $n \times n$ matrix

$$D^2 f(x) := [D_{ij} f(x)] = \begin{bmatrix} D_{11}f(x) & D_{12}f(x) & \dots & D_{1n}f(x) \\ D_{21}f(x) & D_{22}f(x) & \dots & D_{2n}f(x) \\ \vdots & \vdots & \ddots & \vdots \\ D_{n1}f(x) & D_{n2}f(x) & \dots & D_{nn}f(x) \end{bmatrix}.$$

Theorem 7.32. *Suppose $U \subset \mathbb{R}^n$ is open and $f : U \to \mathbb{R}$. If $D_i f$ and $D_j f$ exist on U, and they are both differentiable at a point $a \in U$, then*

$$D_{ij} f(a) = D_{ji} f(a).$$

Remark. Thus when the partial derivatives of f are differentiable at a, the Hessian matrix of f at a is symmetric.

Remark. When $f : U \to \mathbb{R}^m$ we can obviously apply this theorem to every component of f. The same is true for the next theorem.

Proof. For small $h \in \mathbb{R}$ let

$$A(h) := f(a + he_i + he_j) - f(a + he_i) - f(a + he_j) + f(a),$$

where e_i, e_j are the elements of the standard basis of $\mathbb{R}^n$. Consider the function

$$g(t) := f(a + he_i + the_j) - f(a + the_j)$$

of a real variable t. Then by the mean value theorem for some $\tau \in (0, 1)$ we have

$$\begin{aligned} A(h) = g(1) - g(0) &= g'(\tau) \\ &= hD_j f(a + he_i + \tau he_j) - hD_j f(a + \tau he_j). \end{aligned}$$

Note that τ may depend on h. Now the differentiability of $D_j f$ at a implies that for any small $\tilde{h} \in \mathbb{R}^n$ we have

$$D_j f(a + \tilde{h}) = D_j f(a) + DD_j f(a)\tilde{h} + R(\tilde{h}),$$

where R is a sublinear function. If we replace $\tilde{h}$ with $he_i + \tau he_j$ and τhe_j, and subtract the two equations we obtain

$$\begin{aligned} \frac{A(h)}{h} &= D_j f(a + he_i + \tau he_j) - D_j f(a + \tau he_j) \\ &= DD_j f(a)(he_i + \tau he_j) + R(he_i + \tau he_j) - DD_j f(a)(\tau he_j) - R(\tau he_j) \\ &= DD_j f(a)(he_i) + R(he_i + \tau he_j) - R(\tau he_j) \\ &= hD_i D_j f(a) + R(he_i + \tau he_j) - R(\tau he_j). \end{aligned}$$

In addition note that

$$\begin{aligned} \frac{\left|R(he_i + \tau he_j) - R(\tau he_j)\right|}{|h|} &\le \sqrt{1 + \tau^2}\frac{\left|R(he_i + \tau he_j)\right|}{\left|h\sqrt{1 + \tau^2}\right|} + |\tau|\frac{\left|R(\tau he_j)\right|}{|\tau h|} \\ &\le 2\frac{\left|R(he_i + \tau he_j)\right|}{\left|he_i + \tau he_j\right|} + \frac{\left|R(\tau he_j)\right|}{\left|\tau he_j\right|} \xrightarrow[h\to 0]{} 0, \end{aligned}$$

since R is sublinear. Thus we get

$$\begin{aligned} \lim_{h\to 0} \frac{A(h)}{h^2} = D_i D_j f(a) + \lim_{h\to 0} &\frac{R(he_i + \tau he_j) - R(\tau he_j)}{h} \\ &= D_i D_j f(a) + 0 = D_{ij} f(a). \end{aligned}$$

Finally, by switching the role of i, j we similarly get $\lim_{h\to 0} \frac{A(h)}{h^2} = D_{ji} f(a)$. Hence we obtain the desired result, because the value of the limit is unique. ■

Example 7.33. The differentiability of the first order partial derivatives of f is essential for the equality of its mixed second order partial derivatives. For example let

$$f(x, y) := \begin{cases} \frac{xy(x^2 - y^2)}{x^2 + y^2} & (x, y) \ne (0, 0), \\ 0 & (x, y) = (0, 0). \end{cases}$$

Then for $(x, y) \neq (0, 0)$ we have

$$\begin{aligned} D_1 f(x, y) &= \frac{(3x^2 y - y^3)(x^2 + y^2) - 2x(x^3 y - xy^3)}{(x^2 + y^2)^2} = \frac{x^4 y + 4x^2 y^3 - y^5}{(x^2 + y^2)^2}, \\ D_2 f(x, y) &= \frac{(x^3 - 3xy^2)(x^2 + y^2) - 2y(x^3 y - xy^3)}{(x^2 + y^2)^2} = \frac{x^5 - 4x^3 y^2 - xy^4}{(x^2 + y^2)^2}. \end{aligned}$$

And

$$\begin{aligned} D_1 f(0, 0) &= \lim_{t \to 0} \frac{1}{t} [f(t, 0) - f(0, 0)] = 0, \\ D_2 f(0, 0) &= \lim_{t \to 0} \frac{1}{t} [f(0, t) - f(0, 0)] = 0. \end{aligned}$$

Therefore

$$\begin{aligned} D_{12} f(0, 0) &= \lim_{t \to 0} \frac{1}{t} [D_2 f(t, 0) - D_2 f(0, 0)] = \lim_{t \to 0} \frac{1}{t} t = 1, \\ D_{21} f(0, 0) &= \lim_{t \to 0} \frac{1}{t} [D_1 f(0, t) - D_1 f(0, 0)] = \lim_{t \to 0} \frac{1}{t} (-t) = -1. \end{aligned}$$

Hence $D_{12} f(0, 0) \neq D_{21} f(0, 0)$. In addition, note that for $(x, y) \neq (0, 0)$ we have

$$\begin{aligned} D_{12} f(x, y) &= \frac{(5x^4 - 12x^2 y^2 - y^4)(x^2 + y^2) - 4x(x^5 - 4x^3 y^2 - xy^4)}{(x^2 + y^2)^3} \\ &= \frac{x^6 + 9x^4 y^2 - 9x^2 y^4 - y^6}{(x^2 + y^2)^3} = D_{21} f(x, y). \end{aligned}$$

Therefore although $D_{12} f$, $D_{21} f$ exist everywhere and are equal at every point other than $(0, 0)$, they are not continuous at $(0, 0)$.

Theorem 7.34. *Let $U \subset \mathbb{R}^n$ be an open set, and $f : U \to \mathbb{R}$. Suppose the $(k-1)$th order partial derivatives $D_{j_2 \dots j_k} f$ and $D_{i_2 \dots i_k} f$ exist on U, and they are both differentiable at $a \in U$. Also suppose that $j_1, \dots, j_k$ is a permutation of $i_1, \dots, i_k$. Then we have*

$$D_{j_1 \dots j_k} f(a) = D_{i_1 \dots i_k} f(a).$$

Proof. For a function like f and a real number $h \in \mathbb{R}$ we define the function $\mathsf{D}_i^h f$ as follows

$$\mathsf{D}_i^h f(x) := f(x + he_i) - f(x).$$

We can think of D_i^h as a discrete differentiation operator in the ith direction. Note that for any two indices i,j we have

$$\begin{aligned}\big(\mathsf{D}_j^h(\mathsf{D}_i^h f)\big)(x) &= \mathsf{D}_i^h f(x+he_j) - \mathsf{D}_i^h f(x)\\ &= f(x+he_j+he_i) - f(x+he_j) - \big(f(x+he_i)-f(x)\big)\\ &= f(x+he_j+he_i) - f(x+he_j) - f(x+he_i) + f(x)\\ &= f(x+he_i+he_j) - f(x+he_i) - \big(f(x+he_j)-f(x)\big)\\ &= \mathsf{D}_j^h f(x+he_i) - \mathsf{D}_j^h f(x) = \big(\mathsf{D}_i^h(\mathsf{D}_j^h f)\big)(x).\end{aligned}$$

Therefore D_i^h and D_j^h commute. We also have

$$D_j(\mathsf{D}_i^h f)(x) = D_j\big(f(x+he_i)-f(x)\big) = D_j f(x+he_i) - D_j f(x) = \mathsf{D}_i^h(D_j f)(x).$$

Thus D_i^h and D_j commute too. In addition note that $\mathsf{D}_i^h\mathsf{D}_j^h f$ equals $A(h)$ in the proof of last theorem. We are going to use a higher-dimensional version of that proof here. Note that for any function like g we have

$$\mathsf{D}_j^h g(x) = g(x+he_j) - g(x) = hD_j g(x+\tau he_j)$$

for some $\tau\in(0,1)$, provided that $D_j g$ exists on a neighborhood of x, and h is small enough.

Now consider $\mathsf{D}_{i_k}^h(\mathsf{D}_{i_{k-1}}^h(\dots(\mathsf{D}_{i_1}^h f)))$. Let $g := \mathsf{D}_{i_{k-1}}^h(\dots(\mathsf{D}_{i_1}^h f))$. We have

$$\begin{aligned}\mathsf{D}_{i_k}^h(\mathsf{D}_{i_{k-1}}^h(\dots(\mathsf{D}_{i_1}^h f)))(a) &= \mathsf{D}_{i_k}^h g(a)\\ &= g(a+he_{i_k}) - g(x) = hD_{i_k} g(a+\tau_k he_{i_k})\\ &= hD_{i_k}(\mathsf{D}_{i_{k-1}}^h(\dots(\mathsf{D}_{i_1}^h f)))(a+\tau_k he_{i_k})\\ &= h\mathsf{D}_{i_{k-1}}^h(\dots(\mathsf{D}_{i_1}^h(D_{i_k} f)))(a+\tau_k he_{i_k}),\end{aligned}$$

where $\tau_k\in(0,1)$. If we repeat the above computation with $D_{i_k}f$ instead of f and $a+\tau_k he_{i_k}$ instead of a we get

$$\begin{aligned}&\mathsf{D}_{i_{k-1}}^h(\dots(\mathsf{D}_{i_1}^h(D_{i_k} f)))(a+\tau_k he_{i_k})\\ &\qquad = h\mathsf{D}_{i_{k-2}}^h(\dots(\mathsf{D}_{i_1}^h(D_{i_{k-1}i_k} f)))(a+\tau_k he_{i_k}+\tau_{k-1}he_{i_{k-1}}).\end{aligned}$$

Hence we can continue inductively and obtain

$$\mathsf{D}_{i_k}^h(\mathsf{D}_{i_{k-1}}^h(\dots(\mathsf{D}_{i_1}^h f)))(a) = h^{k-1}\mathsf{D}_{i_1}^h(D_{i_2\dots i_k} f)(a+\theta),$$

where $\theta = \tau_k he_{i_k} + \dots + \tau_2 he_{i_2}$ for some $\tau_k,\dots,\tau_2\in(0,1)$. Note that $|\theta|\le|h|$. Now by using the differentiability of $D_{i_2\dots i_k}f$ at a we have

$$\begin{aligned}\mathsf{D}_{i_1}^h(D_{i_2\dots i_k} f)(a+\theta) &= D_{i_2\dots i_k}f(a+\theta+he_{i_1}) - D_{i_2\dots i_k}f(a+\theta)\\ &= D_{i_2\dots i_k}f(a) + DD_{i_2\dots i_k}f(a)(\theta+he_{i_1}) + R(\theta+he_{i_1})\\ &\qquad - D_{i_2\dots i_k}f(a) - DD_{i_2\dots i_k}f(a)(\theta) - R(\theta)\\ &= DD_{i_2\dots i_k}f(a)(he_{i_1}) + R(\theta+he_{i_1}) - R(\theta)\\ &= hD_{i_1i_2\dots i_k}f(a) + R(\theta+he_{i_1}) - R(\theta),\end{aligned}$$

where R is a sublinear remainder. Therefore we get

$$\begin{aligned}\frac{1}{h^k}\mathsf{D}^h_{i_k}(\mathsf{D}^h_{i_{k-1}}(\dots(\mathsf{D}^h_{i_1}f)))(a) &= \frac{1}{h}\mathsf{D}^h_{i_1}(D_{i_2\dots i_k}f)(a+\theta)\\ &= D_{i_1 i_2\dots i_k}f(a) + \frac{R(\theta+he_{i_1})-R(\theta)}{h}.\end{aligned}$$

Thus

$$\lim_{h\to 0}\frac{1}{h^k}\mathsf{D}^h_{i_k}(\mathsf{D}^h_{i_{k-1}}(\dots(\mathsf{D}^h_{i_1}f)))(a) = D_{i_1 i_2\dots i_k}f(a),$$

since $|\theta|\le|h|$, $|\theta+he_{i_1}|\le|h|$, and R is sublinear.

However note that D^h_i and D^h_j commute for any i,j. So by rearranging $i_1,\dots,i_k$ we get

$$\mathsf{D}^h_{i_k}(\mathsf{D}^h_{i_{k-1}}(\dots(\mathsf{D}^h_{i_1}f))) = \mathsf{D}^h_{j_k}(\mathsf{D}^h_{j_{k-1}}(\dots(\mathsf{D}^h_{j_1}f))).$$

Hence we have

$$\begin{aligned}D_{i_1 i_2\dots i_k}f(a) &= \lim_{h\to 0}\frac{1}{h^k}\mathsf{D}^h_{i_k}(\mathsf{D}^h_{i_{k-1}}(\dots(\mathsf{D}^h_{i_1}f)))(a)\\ &= \lim_{h\to 0}\frac{1}{h^k}\mathsf{D}^h_{j_k}(\mathsf{D}^h_{j_{k-1}}(\dots(\mathsf{D}^h_{j_1}f)))(a) = D_{j_1\dots j_k}f(a),\end{aligned}$$

as desired. ■

Remark. Note that the above proof works because $\mathsf{D}^h_i, \mathsf{D}^h_j$ commute for any i,j. However, in the previous theorem we have shown that D_i, D_j commute too. So we must be able to prove the above theorem by rearranging $D_{i_1}\dots D_{i_k}$ to obtain $D_{j_1}\dots D_{j_k}$. But this requires strengthening the assumptions of the theorem, because the commutativity of D_i, D_j holds under some assumptions, unlike the commutativity of $\mathsf{D}^h_i, \mathsf{D}^h_j$, which is a mere algebraic property and does not require special differentiability assumptions to hold.

Nevertheless, it is instructive to see how the other approach works. Let us try to show that $D_{ijk}f(a) = D_{jki}f(a)$. First note that we have

$$D_{ijk}f(a) = D_{ij}D_kf(a) = D_{ji}D_kf(a) = D_{jik}f(a)$$

provided that $D_{jk}f, D_{ik}f$ are differentiable at a. Next we need to show that $D_{jik}f(a)$ and $D_{jki}f(a)$ are equal. We know that $D_{ki}f = D_{ik}f$ if the partial derivatives of f are differentiable. And if we differentiate both sides of this equality we get the desired. But in order to differentiate both sides of $D_{ki}f = D_{ik}f$ we need to know that it holds on a neighborhood of a. Hence in this approach we need to assume that partial derivatives of f are differentiable on a neighborhood of a and 2nd order partial derivatives of f are differentiable at a to conclude the symmetry of the 3rd order partial derivatives of f at a. But as we have seen in the above proof, we can obtain the symmetry of the 3rd order partial derivatives of f by merely assuming the differentiability of the 2nd order partial derivatives of f at a, and no differentiability assumption on a neighborhood of a is necessary. ■

Theorem 7.35. *Suppose $U \subset \mathbb{R}^n$ is open. Let $f : U \to \mathbb{R}^m$ and $g : U \to \mathbb{R}^p$ be C^k functions for some $1 \le k \le \infty$. Let $V \subset \mathbb{R}^m$ be an open set containing $f(U)$, and suppose $F : V \to \mathbb{R}^q$ is a C^k function. Also suppose $B : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^q$ is a bilinear map, and $c_1, c_2 \in \mathbb{R}$. Then*

(i) *If $p = m$ then $c_1 f + c_2 g$ is a C^k function.*

(ii) *$B[f, g]$ and $F \circ f$ are C^k functions.*

(iii) *If $p = m = 1$ then fg is a C^k function. If in addition $g \neq 0$ on U, then $\frac{f}{g}$ is also a C^k function.*

Proof. Let $B^i_{\kappa l}$ be the components of B, as defined in the remark after the Leibniz rule. Then we have

$$\begin{aligned} D_j(c_1 f + c_2 g)_i &= c_1 D_j f_i + c_2 D_j g_i, \\ D_j(B[f, g])_i &= \textstyle\sum_{\kappa\le m} \sum_{l\le p} B^i_{\kappa l}\big(g_l D_j f_\kappa + f_\kappa D_j g_l\big), \\ D_j(F \circ f)_i &= \textstyle\sum_{l\le m} (D_j f_l)((D_l F_i) \circ f). \end{aligned} \qquad (*)$$

For $1 \le k < \infty$, the proof is by induction on k. When $k = 1$ we know that the components of f, g, F and their partial derivatives, are continuous. Therefore $c_1 f + c_2 g$, $B[f, g]$, and $F \circ f$ are continuous, and have continuous partial derivatives by $(*)$. Because the sum, the product, and the composition of continuous functions are continuous. Note that constant functions and bilinear maps are continuous. Also the function $x \mapsto (f(x), g(x))$ from U into $\mathbb{R}^m \times \mathbb{R}^p$ is continuous.

Now suppose the theorem is true for some $k < \infty$. Then we have to prove the theorem for $k + 1$. Let f, g, F be C^{k+1} functions. Then we know that the components of f, g, F and their partial derivatives, are C^k functions. Hence by the induction hypothesis we know that the partial derivatives of $c_1 f + c_2 g$, $B[f, g]$, and $F \circ f$ are C^k functions. Because by $(*)$, their partial derivatives can be expressed as a linear combination of C^k functions. (Note that these functions are either C^k, or they are the product of two C^k functions, or they are the product of a C^k function and a function which is the composition of two C^k functions. Also, note that the multiplication of two real numbers is a bilinear map.) Notice that here we are using the induction hypothesis applied to linear combinations with more than two terms; but this more general case follows by an easy induction on the number of terms, from the case of linear combinations with only two terms. Therefore $c_1 f + c_2 g$, $B[f, g]$, and $F \circ f$ are C^{k+1} functions.

Finally if the functions f, g, F are C^∞ functions, then they are C^k functions for all $k < \infty$. Therefore $c_1 f + c_2 g$, $B[f, g]$, and $F \circ f$ are C^k functions for all $k < \infty$. Hence they are also C^∞ functions.

Part (iii) follows easily from the previous two parts. We only need to note that the multiplication of real numbers is a bilinear map. Also if $g \neq 0$ then $\frac{1}{g}$ is C^k. Because the map $x \mapsto x$ is obviously a C^∞ function from $\mathbb{R} - \{0\}$ into $\mathbb{R}$. Hence $x \mapsto \frac{1}{x}$ is also a C^∞ function by the one-dimensional version of this theorem, i.e. Theorem 4.11. Therefore the composition of $x \mapsto \frac{1}{x}$ and g, i.e. $\frac{1}{g}$, is a C^k function. ■

Definition 7.36. A **multi-index** is a vector $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{Z}^n$ which has nonnegative components, i.e. $\alpha_i \geq 0$. The **order** of α is

$$|\alpha| := \alpha_1 + \cdots + \alpha_n.$$

We also define

$$\alpha! := \alpha_1! \cdots \alpha_n!.$$

For $h \in \mathbb{R}^n$ we define

$$h^\alpha := h_1^{\alpha_1} \cdots h_n^{\alpha_n},$$

where we interpret $0^0 = 1$. Also for a function f defined on an open set in $\mathbb{R}^n$ we define

$$D^\alpha f := D_1^{\alpha_1} \cdots D_n^{\alpha_n} f,$$

where $D_i^k = \overbrace{D_i D_i \cdots D_i}^{k \text{ times}}$. When $\alpha = 0 = (0, \dots, 0)$ we use the convention $D^0 f = f$.

Definition 7.37. Suppose $U \subset \mathbb{R}^n$ is open and $f : U \to \mathbb{R}^m$ has partial derivatives up to order k at $x \in U$. The kth order **Taylor polynomial** of f at x is

$$P(h) = \sum_{|\alpha| \leq k} \frac{1}{\alpha!} D^\alpha f(x) h^\alpha,$$

and the kth order **Taylor remainder** is $R(h) = f(x+h) - P(h)$.

Remark. In fact the kth order Taylor polynomial is

$$\sum_{j \leq k} \Big[\sum_{i_1 \leq n} \cdots \sum_{i_j \leq n} \frac{1}{j!} D_{i_1 \dots i_j} f(x) h_{i_1} \cdots h_{i_j} \Big].$$

But because of the symmetry of mixed partial derivatives we can see that for any multi-index α with $|\alpha| = j$ there are $\frac{j!}{\alpha!}$ partial derivatives of order j which are equal to $D^\alpha f(x)$. Hence we arrive at the initial formula for the Taylor polynomial.

***Exercise* 7.38.** Show that a multivariable polynomial

$$p(x_1, \dots, x_n) = \sum_{k_i \leq m_i} c_{k_1 \dots k_n} x_1^{k_1} \cdots x_n^{k_n},$$

is a C^∞ function from $\mathbb{R}^n$ to $\mathbb{R}$. In addition, show that

$$c_{k_1\dots k_n} = \frac{1}{k_1!\cdots k_n!} D_1^{k_1}\cdots D_n^{k_n} p.$$

This also proves that the coefficients of a multivariable polynomial are uniquely determined by the polynomial.

Theorem 7.39. *Let $U \subset \mathbb{R}^n$ be an open set, and $f : U \to \mathbb{R}^m$. Suppose all the partial derivatives of f of orders less than or equal to $k-1$ exist on U, and they are all differentiable at some $x \in U$. Let R be the kth order Taylor remainder of f at x. Then*

$$\lim_{h\to 0} \frac{R(h)}{|h|^k} = 0.$$

Remark. Note that the kth order partial derivatives of f at x exist too. Also note that by Theorem 7.34 all the mixed partial derivatives of f at x up to order k are symmetric.

Remark. It is actually sufficient to just assume that all the $(k-1)$th order partial derivatives of f are differentiable at x. Since then their continuity at x implies the differentiability of the $(k-2)$th order partial derivatives at x, and so on.

Proof. We only need to prove the limit is zero for each component of f, so we assume that $m = 1$. We proceed by induction on k. When $k = 1$ the assumption is that the 0th order partial derivative of f, i.e. f itself, is differentiable at x. Hence

$$R(h) = f(x+h) - f(x) - \sum_i D_i f(x) h_i = f(x+h) - f(x) - Df(x)h$$

is sublinear as desired. Next, for the induction step, suppose the theorem is true for some k. Note that the induction hypothesis is that the conclusion of the theorem holds for any function (not just f) that satisfies theorem's assumptions for this particular value of k. And in fact we are going to apply the induction hypothesis to the derivatives of f.

Let f be a function whose partial derivatives up to order k exist on U, and they are all differentiable at x. Then the $(k+1)$th order Taylor remainder of f at x is

$$R(h) = f(x+h) - \sum_{|\alpha|\le k+1} \frac{1}{\alpha!} D^\alpha f(x) h^\alpha.$$

Since the right hand side is differentiable, R is differentiable too. Thus if we differentiate with respect to h_j we obtain

$$\begin{aligned} D_jR(h) &= D_jf(x+h) - \sum_{|\alpha|\le k+1} \frac{1}{\alpha!} D^\alpha f(x) \frac{\partial}{\partial h_j}(h^\alpha) \\ &= D_jf(x+h) - \sum_{|\alpha|\le k+1} \frac{1}{\alpha!} D^\alpha f(x) \alpha_j h^{\alpha - e_j}. \end{aligned}$$

Here $e_j \in \mathbb{Z}^n$ is the vector whose jth component is 1 and its other components are zero. If $\alpha_j = 0$ then the term containing $D^\alpha f(x)$ vanishes. So the above sum is actually a sum over those multi-indices α for which we have $\alpha_j \neq 0$. For such an α let $\beta := \alpha - e_j$. Note that as α runs through all multi-indices of order at most $k+1$ with $\alpha_j \neq 0$, β runs through all multi-indices of order at most k. Hence we have

$$D_jR(h) = D_jf(x+h) - \sum_{|\beta|\le k} \frac{1}{\beta!} D^\beta D_j f(x) h^\beta.$$

Note that here we used the facts that $\frac{\alpha_j}{\alpha!} = \frac{1}{\beta!}$, and $D^\alpha f(x) = D^\beta D_j f(x)$ due to the symmetry of mixed partial derivatives of order at most $k+1$ at x. Therefore D_jR is the kth order Taylor remainder of D_jf at x. Hence by the induction hypothesis we have

$$\lim_{h\to 0} \frac{D_jR(h)}{|h|^k} = 0.$$

Let us define $r_j(h) := |D_jR(h)|/|h|^k$ for $h \neq 0$, and $r_j(0) := 0$. Then r_j is continuous at $h = 0$. Now let $y_j := (h_1, \dots, h_j, 0, \dots, 0)$. Then, by the mean value theorem, for some $t_j \in (0,1)$ we have

$$\begin{aligned} R(h) = R(h) - R(0) &= \sum_{j=1}^n R(y_j) - R(y_{j-1}) \\ &= \sum_{j=1}^n h_j D_j R(h_1, \dots, h_{j-1}, t_j h_j, 0 \dots, 0). \end{aligned}$$

Let $\theta_j := (h_1, \dots, h_{j-1}, t_j h_j, 0 \dots, 0)$. Then $|\theta_j| \le |h|$. Hence

$$\frac{|R(h)|}{|h|^{k+1}} \le \sum \frac{|h_j|}{|h|} \frac{|D_jR(\theta_j)|}{|h|^k} \le \sum \frac{|D_jR(\theta_j)|}{|h|^k} = \sum \frac{r_j(\theta_j)|\theta_j|^k}{|h|^k} \le \sum r_j(\theta_j).$$

But $\theta_j \to 0$ as $h \to 0$, because $|\theta_j| \le |h|$. Thus we get

$$\lim_{h\to 0} r_j(\theta_j) = \lim_{\theta_j \to 0} r_j(\theta_j) = 0 \implies \lim_{h\to 0} \sum r_j(\theta_j) = 0.$$

Hence $\lim_{h\to 0} \frac{|R(h)|}{|h|^{k+1}} = 0$ as required. ■

7.4 Extrema of Multivariable Functions

Theorem 7.40. *Suppose $U \subset \mathbb{R}^n$ is an open connected set and $f : U \to \mathbb{R}^m$ is differentiable on U. If $Df = 0$ on U then f is constant.*

Remark. If U is disconnected then f can have different constant values on different components of U, so the connectedness of U is essential in the theorem.

Proof. It suffices to show that each component of f is constant; so we assume that $m = 1$. Let c be a value in the image of f. Then as f is continuous, the set

$$A := \{x \in U : f(x) = c\}$$

is closed in U. If we show that A is also open in U, then as A is obviously nonempty we must have $A = U$, and consequently f is constant. To do this let $a \in A$ be an arbitrary point. Since U is open there is an open ball $B_r(a) \subset U$. Let $b \in B_r(a)$, and consider the function $g(t) := f(a + t(b - a))$. Then for some $\tau \in (0, 1)$ we have

$$f(b) - f(a) = g(1) - g(0) = g'(\tau) = Df(a + \tau(b - a))(b - a) = 0.$$

Hence $f(b) = c$ too. Thus $B_r(a) \subset A$, and A is open. ■

Definition 7.41. Suppose X is a metric space, and $f : X \to \mathbb{R}$ is a function. We say f has a **local maximum** at $y \in X$ if $f(y) \geq f(x)$ for all x in a neighborhood of y. Similarly, we say f has a **local minimum** at $y \in X$ if $f(y) \leq f(x)$ for all x in a neighborhood of y. A **local extremum** of f, is either a local maximum of f, or a local minimum of f.

Theorem 7.42. *Suppose f is a real-valued function defined on an open neighborhood of $a \in \mathbb{R}^n$, and it is differentiable at a. If f has a local maximum or minimum at a, then $Df(a) = 0$.*

Proof. For a fixed $v \in \mathbb{R}^n$ and small $t \in \mathbb{R}$, let $g(t) := f(a + tv)$. Then it is obvious that g has a local extremum at $t = 0$. Hence

$$0 = g'(0) = D_v f(a) = Df(a)v.$$

Since v is arbitrary we must have $Df(a) = 0$. ■

Second Derivative Test. *Suppose f is a real-valued function defined on an open neighborhood of $x \in \mathbb{R}^n$. Also suppose f has first order partial derivatives around x, and its partial derivatives are differentiable at x.*

(i) *Suppose $Df(x) = 0$. If the Hessian matrix $D^2 f(x)$ is positive definite then f has a local minimum at x, and if $D^2 f(x)$ is negative definite then f has a local maximum at x.*

(ii) *If f has a local minimum at x then $D^2f(x)$ is positive semi-definite, and if f has a local maximum at x then $D^2f(x)$ is negative semi-definite.*

Proof. First note that as the partial derivatives D_if are differentiable at x, they are also continuous at x. Hence f is differentiable at x, and $Df(x)$ exists. Also note that by our hypothesis $A := D^2f(x)$ is a symmetric $n\times n$ matrix. Hence there is an orthonormal basis $v_1,\dots,v_n$ for $\mathbb{R}^n$ consisting of the eigenvectors of A with corresponding eigenvalues $\lambda_1,\dots,\lambda_n$, i.e. we have

$$\langle v_i, v_j\rangle = \begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases}, \qquad \text{and} \qquad Av_i = \lambda_i v_i.$$

Here $\langle\,,\rangle$ is the standard inner product of $\mathbb{R}^n$. The positive definiteness of A means that $\lambda_i > 0$ for all i. Similarly A is negative definite, positive semi-definite, or negative semi-definite if respectively all λ_i's are negative, nonnegative, or nonpositive. Let $h = (h_1,\dots,h_n)$ be an arbitrary vector in $\mathbb{R}^n$, and let $(l_1,\dots,l_n)$ be the coordinates of h in the eigenvectors basis, i.e. $h = l_1v_1 + \cdots + l_nv_n$. Then we have

$$\begin{aligned}\sum_{i,j} A_{ij}h_ih_j &= \sum_i h_i(Ah)_i = \langle Ah, h\rangle \\ &= \Big\langle \sum_i l_iAv_i, \sum_j l_jv_j \Big\rangle = \sum_{i,j}\lambda_il_il_j\langle v_i, v_j\rangle = \sum_i \lambda_i l_i^2.\end{aligned}$$

We also have

$$|h|^2 = \sum_i h_i^2 = \langle h, h\rangle = \Big\langle \sum_i l_iv_i, \sum_j l_jv_j \Big\rangle = \sum_{i,j} l_il_j\langle v_i, v_j\rangle = \sum_i l_i^2.$$

In the following we give the proofs for the case of a minimum, the case of a maximum is similar.

(i) Suppose $h\in\mathbb{R}^n$ and $|h|$ is small. We use the above notations for the coordinates of h. Let R be the second order Taylor remainder of f at x. Then by Theorem 7.39 we have (note that we are not using the multi-index notation here)

$$\begin{aligned} f(x+h) - f(x) &= Df(x)h + \frac{1}{2}\sum_{i,j} D_{ij}f(x)h_ih_j + R(h) \\ &= \frac{1}{2}\sum_{i,j} A_{ij}h_ih_j + R(h) \\ &= \frac{1}{2}\sum_i \lambda_i l_i^2 + R(h) \geq \frac{1}{2}\lambda_1|h|^2 + R(h),\end{aligned}$$

where λ_1 is the smallest eigenvalue. Now since $\lim_{h\to 0}\frac{|R(h)|}{|h|^2} = 0$, when $|h|$ is small enough we have $\frac{|R(h)|}{|h|^2} < \frac{1}{4}\lambda_1$. Hence

$$\frac{f(x+h)-f(x)}{|h|^2} \geq \frac{1}{2}\lambda_1 - \frac{1}{4}\lambda_1 > 0.$$

Thus $f(x+h) > f(x)$ when $|h|$ is small enough. Therefore f has a local minimum at x.

(ii) Since f has a local minimum at x we know that $Df(x) = 0$. Similarly to the above we have

$$\frac{1}{2}\sum \lambda_i l_i^2 + R(h) = f(x+h) - f(x) \geq 0$$

Let us set $h = \frac{1}{k}v_j$, where k is a large integer. That is we set $l_j = \frac{1}{k}$, and all other l_i's equal to 0. Then we have

$$\frac{1}{2}\lambda_j \geq -\frac{R(\frac{1}{k}v_j)}{\frac{1}{k^2}} = -\frac{R(\frac{1}{k}v_j)}{\frac{1}{k^2}|v_j|^2} = -\frac{R(h)}{|h|}.$$

Now as $k \to \infty$ we have $h \to 0$; so the right hand side of the above inequality goes to zero. Therefore $\lambda_j \geq 0$ as desired. ■

***Example* 7.43.** Positive semi-definiteness of the Hessian matrix is not sufficient to ensure that we have a local minimum. For example $g(t) = t^3$ does not have a local minimum at $t = 0$, even though its first derivative vanishes at $t = 0$ and its second derivative is nonnegative there. A more interesting example is the function

$$f(x, y) = (y - x^2)(y - 2x^2).$$

It is easy to see that $Df(0, 0) = 0$ and $D^2f(0, 0)$ is positive semi-definite. But f does not have a local minimum at the origin, because it is negative on the region between the two parabolas $y = x^2$ and $y = 2x^2$ in its domain, and is positive outside of that region. However, the restriction of f to every line passing through the origin has a local minimum at the origin. Because the two parabolas are tangent to the line $y = 0$ at the origin; so, near the origin, no other line passing through the origin can stay in the region between the two parabolas. And this is also true for the line $y = 0$ itself.

7.5 The Inverse and Implicit Function Theorems

Inverse Function Theorem. *Suppose $U \subset \mathbb{R}^n$ is open, and $f : U \to \mathbb{R}^n$ is C^k for some $1 \leq k \leq \infty$. Also suppose that for some $a \in U$ the matrix $Df(a)$ is invertible. Then there is an open set V containing a such that*

(i) $f\big|_V$ *is one-to-one, and*

(ii) $f(V)$ *is open.*

Furthermore, if $g : f(V) \to V$ is the inverse of $f\big|_V$, then

(iii) g *is C^k, and for $z \in f(V)$ we have*

$$Dg(z) = \big(Df(g(z))\big)^{-1}.$$

Proof. We break the proof into several parts to make it more comprehensible, although the parts are intertwined.

(i) First let us show that f is one-to-one on a neighborhood of a. Let $A := Df(a)$. Then by our assumption A is an invertible matrix. Let C be a positive constant such that $|A^{-1}h| \leq C|h|$ for every $h \in \mathbb{R}^n$. Let

$$F(x) := x - A^{-1}f(x).$$

Then F is differentiable, and we have $DF(x) = I - A^{-1}Df(x)$. So DF is continuous since Df is continuous. In addition we have

$$DF(a) = I - A^{-1}Df(a) = I - A^{-1}A = 0.$$

Hence, when r is small enough, for $x \in B_r(a)$ we must have $\big(\sum_{i,j} |D_jF_i(x)|^2\big)^{\frac{1}{2}} \leq \frac{1}{2}$. Thus by Theorem 7.24 we have

$$|F(x) - F(y)| \leq \frac{1}{2}|x - y|,$$

for $x, y \in B_r(a)$ (note that the line segment joining x, y is also inside $B_r(a)$). Therefore we obtain

$$\begin{aligned} |A^{-1}(f(x) - f(y))| &= |x - y - (F(x) - F(y))| \\ &\geq |x - y| - |F(x) - F(y)| \geq |x - y| - \frac{1}{2}|x - y| = \frac{1}{2}|x - y|. \end{aligned}$$

Hence we get

$$|f(x) - f(y)| \geq \frac{1}{C}|A^{-1}(f(x) - f(y))| \geq \frac{1}{2C}|x - y|, \qquad (*)$$

for every $x, y \in B_r(a)$. The above relation clearly implies that $f|_{B_r(a)}$ is one-to-one, because from $f(x) = f(y)$ we can conclude that $|x - y| = 0$, and thus $x = y$.

(ii) Next note that $Df(x)$ is also invertible when x is near a. Because we know that $\det Df(a) \neq 0$. Therefore we must have $\det Df(x) \neq 0$ for x close to a, since det and Df are continuous functions. Thus we can assume that r is small enough so that $Df(x)$ is invertible for $x \in B_r(a)$.

Now let us show that the image of $f|_{B_r(a)}$ contains an open ball $B_s(f(a))$ around $f(a)$. Fix a point $y_0 \in B_s(f(a))$, and consider the function $\varphi : \overline{B_{\frac{r}{2}}(a)} \to \mathbb{R}$ defined by

$$\varphi(x) := |f(x) - y_0|^2.$$

Then the continuous function φ attains its minimum on its compact domain $\overline{B_{\frac{r}{2}}(a)}$ at some point x_0. Now note that by $(*)$ for $x \in \partial B_{\frac{r}{2}}(a)$ we have

$$|f(x) - f(a)| \geq \frac{1}{2C}|x - a| = \frac{r}{4C}.$$

Hence

$$\varphi(x) = |f(x) - y_0|^2 \geq \big(|f(x) - f(a)| - |f(a) - y_0|\big)^2 \geq \big(\frac{r}{4C} - s\big)^2.$$

On the other hand, $\varphi(a) = |f(a) - y_0|^2 < s^2$. Thus if we have $s < \frac{r}{4C} - s$, or equivalently $s < \frac{r}{8C}$, then $\varphi(a) < \varphi(x)$. Hence φ does not attain its minimum on $\partial B_{\frac{r}{2}}(a)$. So we must have $x_0 \in B_{\frac{r}{2}}(a)$. Therefore φ has a local minimum at x_0, and consequently we get $D\varphi(x_0) = 0$. But $\varphi(x) = |f(x) - y_0|^2 = \langle f(x) - y_0, f(x) - y_0\rangle$, where $\langle\, ,\rangle$ is the standard inner product of $\mathbb{R}^n$. Hence by Leibniz rule we get

$$D_j\varphi(x) = 2\langle D_jf(x), f(x) - y_0\rangle.$$

Thus for every j we have

$$\big(Df(x_0)^{\mathsf{T}}(f(x_0) - y_0)\big)_j = \langle D_jf(x_0), f(x_0) - y_0\rangle = 0.$$

However, the matrix $Df(x_0)$ is invertible, so its transpose is also invertible. Therefore we must have $f(x_0) - y_0 = 0$, i.e. $f(x_0) = y_0$. Hence when s is small enough, $B_s(f(a))$ is in the image of $f|_{B_r(a)}$, as desired.

Now let $V := B_r(a) \cap f^{-1}(B_s(f(a)))$. Then V is open since f is continuous. Also $f(V) = B_s(f(a))$, since $B_s(f(a))$ is contained in the image of $f|_{B_r(a)}$. Thus $f(V)$ is open. In addition, $f|_V$ is one-to-one because $f|_{B_r(a)}$ is one-to-one. Hence parts (i) and (ii) of the theorem are proved.

(iii) Next let us show that g, the inverse of $f|_V$, is C^k. Let $z, w \in f(V)$. Then we have $z = f(x)$ and $w = f(y)$ for some $x, y \in V$. Equivalently we have $g(z) = x$ and $g(w) = y$. Hence by $(*)$ we get

$$|g(w) - g(z)| = |y - x| \leq 2C|f(y) - f(x)| = 2C|w - z|.$$

Therefore g is continuous. In addition, we know that

$$f(y) - f(x) = \tilde{A}(y-x) + R(y-x),$$

where $\tilde{A} = Df(x)$, and R is sublinear. Thus we get

$$\begin{aligned} g(w) - g(z) - \tilde{A}^{-1}(w-z) &= y - x - \tilde{A}^{-1}\big(f(y) - f(x)\big) \\ &= y - x - \tilde{A}^{-1}\big(\tilde{A}(y-x) + R(y-x)\big) \\ &= y - x - (y-x) - \tilde{A}^{-1}\big(R(y-x)\big) \\ &= -\tilde{A}^{-1}\big(R(g(w)-g(z))\big). \end{aligned}$$

Let $\tilde{C}$ be a positive constant such that $|\tilde{A}^{-1}h| \le \tilde{C}|h|$ for every $h \in \mathbb{R}^n$. Then by $(*)$ we obtain

$$\begin{aligned} \frac{\big|\tilde{A}^{-1}\big(R(g(w)-g(z))\big)\big|}{|w-z|} &\le \tilde{C}\frac{\big|R(g(w)-g(z))\big|}{|w-z|} \\ &= \tilde{C}\frac{\big|R(y-x)\big|}{|f(y)-f(x)|} \le 2C\tilde{C}\frac{\big|R(y-x)\big|}{|y-x|} \xrightarrow[y\to x]{} 0. \end{aligned}$$

However when $w \to z$ we also have $y \to x$, because by $(*)$ we have $|y-x| \le 2C|w-z|$. Hence $-\tilde{A}^{-1}\big(R(g(w)-g(z))\big)$ is sublinear, and therefore g is differentiable at z. We also have

$$Dg(z) = \tilde{A}^{-1} = \big(Df(x)\big)^{-1} = \big(Df(g(z))\big)^{-1},$$

as desired. Now note that $Df(g(z))$ is continuous in z, since g, Df are continuous. In addition, due to Cramer's rule, the entries of the inverse of an invertible matrix are rational functions (i.e. quotients of polynomials) of the entries of the matrix. Thus $\big(Df(g(z))\big)^{-1}$ is continuous in z. Therefore Dg is continuous, and g is C^1.

Finally we show by induction that for finite k, g is C^k when f is C^k. The case of $k = 1$ is already proved. Suppose the claim holds for $k-1$. Let f be C^k. Then f is also C^{k-1}; so by induction hypothesis g is C^{k-1}. Since the entries of the inverse of an invertible matrix are rational functions of the entries of the matrix, and rational functions are smooth, $Dg = (Df(g))^{-1}$ is C^{k-1}. Therefore g is C^k, as desired. At the end note that when f is C^∞ then it is C^k for every finite k. Hence g is C^k for every finite k. Thus g is also C^∞. ■

In the next theorem we identify $\mathbb{R}^{n+m}$ with $\mathbb{R}^n \times \mathbb{R}^m$; so we denote the points of $\mathbb{R}^{n+m}$ by (x, y) where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$.

Implicit Function Theorem. *Suppose $U \subset \mathbb{R}^{n+m}$ is open, and $f : U \to \mathbb{R}^m$ is C^k for some $1 \le k \le \infty$. Let $(a, b) \in U$, and suppose $f(a, b) = c$. Consider the level set*

$$\Gamma := \{(x, y) \in U : f(x, y) = c\}.$$

Suppose the $m \times m$ matrix

$$\left[\frac{\partial f_i}{\partial y_j}(a,b)\right]$$

is invertible. Then

(i) *There is an open set V in $\mathbb{R}^{n+m}$ containing (a,b), an open set W in $\mathbb{R}^n$ containing a, and a unique function $g : W \to \mathbb{R}^m$ such that*

$$\Gamma \cap V = \{(x, g(x)) : x \in W\}.$$

(ii) *Furthermore, g is a C^k function that satisfies $g(a) = b$, and for every $x \in W$ we have $f(x, g(x)) = c$, and*

$$Dg(x) = -\left[\frac{\partial f_i}{\partial y_j}(x, g(x))\right]^{-1}\left[\frac{\partial f_i}{\partial x_j}(x, g(x))\right].$$

Remark. In other words, near the point (a,b) the level set $\Gamma = \{f = c\}$ is the graph of a C^k function g.

Alternatively, we can say that near (a,b) we can solve the equation $f(x,y) = c$ for y to obtain $y = g(x)$.

Proof. Consider the function $F : U \to \mathbb{R}^n \times \mathbb{R}^m$ which is defined by

$$F(x,y) := (x, f(x,y)).$$

It is obvious that F is C^k, since its components are C^k. Now we have

$$DF = \begin{bmatrix} I & 0 \\ D_x f & D_y f \end{bmatrix},$$

where I is the $n \times n$ identity matrix, 0 is the $n \times m$ zero matrix, and $D_x f, D_y f$ are the $m \times n$ and $m \times m$ matrices

$$\left[\frac{\partial f_i}{\partial x_j}\right], \left[\frac{\partial f_i}{\partial y_j}\right],$$

respectively. Hence we have $\det DF = \det I \cdot \det D_y f = \det D_y f$. Therefore $\det DF \neq 0$ at (a,b); so DF is invertible at (a,b). Thus we can apply the inverse function theorem to conclude that there is an open set V containing (a,b) such that $F\big|_V$ is one-to-one, $F(V)$ is open, and G, the inverse of $F\big|_V$, is C^k. Note that we have $F(a,b) = (a, f(a,b)) = (a,c)$. So $G(a,c) = (a,b)$. We also know that

$$(x,y) = F(G(x,y)) = \big(G_1(x,y), \dots, G_n(x,y), f(G(x,y))\big). \tag{$*$}$$

Thus the first n components of $G(x,y)$ is simply x, i.e. we have

$$G(x,y) = (x, H(x,y)),$$

for some C^k function $H : F(V) \to \mathbb{R}^m$. Hence we get

$$(x, y) = G(F(x, y)) = G(x, f(x, y)) = (x, H(x, f(x, y))).$$

Therefore we must have

$$H(x, f(x, y)) = y, \tag{$**$}$$

for every $(x, y) \in V$.

Now note that

$$W := \{x : (x, c) \in F(V)\}$$

is open, because it is the inverse image of the open set $F(V)$ under the continuous function $x \mapsto (x, c)$. We also have $a \in W$, since $(a, c) = F(a, b) \in F(V)$. Now suppose $f(x, y) = c$ for some $(x, y) \in V$. Then by $(**)$ we have $H(x, c) = y$. Keeping this in mind, for $x \in W$ we define

$$g(x) := H(x, c).$$

Then when $(x, y) \in \Gamma \cap V$, i.e. when $f(x, y) = c$, we have

$$(x, c) = (x, f(x, y)) = F(x, y) \in F(V);$$

so $x \in W$. And we also have $y = H(x, c) = g(x)$. On the other hand, for $x \in W$ we have $(x, g(x)) = (x, H(x, c)) = G(x, c) \in V$, and

$$f(x, g(x)) = f(x, H(x, c)) = f(G(x, c)) = c, \tag{$***$}$$

due to $(*)$. Hence we have shown that

$$\Gamma \cap V = \{(x, g(x)) : x \in W\}.$$

This equality also implies that g is unique, because it says that $\Gamma \cap V$ is the graph of the function g, and the graph of a function uniquely determines the function. It is obvious too that g is C^k, since it is the composition of the C^k functions H and $x \mapsto (x, c)$. Furthermore, by $(**)$ we have $g(a) = H(a, c) = H(a, f(a, b)) = b$.

Finally, if by using chain rule we differentiate the equation $(***)$, we obtain

$$Df(x, g(x))D\hat{g}(x) = Dc = 0,$$

where $\hat{g}(x) := (x, g(x))$. The above equality can be rewritten as

$$0 = \begin{bmatrix} D_x f(x, g(x)) & D_y f(x, g(x)) \end{bmatrix} \begin{bmatrix} I \\ Dg(x) \end{bmatrix} = D_x f(x, g(x)) + D_y f(x, g(x))Dg(x),$$

which implies the desired formula for $Dg(x)$. Note that as we have shown in the proof of inverse function theorem, we can assume that DF is invertible on V, which implies that $D_y f$ is invertible on V, since $\det DF = \det D_y f$. Hence $D_y f(x, g(x))$ is invertible for $x \in W$, because as we have shown above $(x, g(x)) \in V$. ■

7.6 Lagrange Multipliers

Theorem 7.44. *Let $U \subset \mathbb{R}^n$ be an open set, and $f : U \to \mathbb{R}$ be a C^1 function. Also, let $g : U \to \mathbb{R}^k$ be a C^1 function, where $k < n$. Consider the level set*

$$\Gamma := \{x \in U : g(x) = 0\}.$$

If $f|_\Gamma$ has a local extremum at $a \in \Gamma$, and $Dg(a)$ has full rank k, then we must have

$$Df(a) = \lambda_1 Dg_1(a) + \cdots + \lambda_k Dg_k(a),$$

for some constants $\lambda_1, \dots, \lambda_k \in \mathbb{R}$.

Remark. The constants $\lambda_1, \dots, \lambda_k$ are called **Lagrange multipliers**.

Proof. Since the $k \times n$ matrix $Dg(a)$ has rank k, it must have k linearly independent columns. To simplify the notation let us assume that the last k columns of $Dg(a)$ are linearly independent. Also let us denote the points in $\mathbb{R}^n$ by (z, y), where $z \in \mathbb{R}^{n-k}$ and $y \in \mathbb{R}^k$. In this notation we denote $a = (b, c)$. Then the $k \times k$ matrix

$$D_y g(a) := \left[\frac{\partial g_i}{\partial y_j}(a)\right]_{1 \le i,j \le k}$$

is invertible, since its columns are linearly independent. Hence by the implicit function theorem, there is an open set $V \subset \mathbb{R}^n$ containing a, an open set $W \subset \mathbb{R}^{n-k}$ containing b, and a unique C^1 function $h : W \to \mathbb{R}^k$ such that

$$\Gamma \cap V = \{(z, h(z)) : z \in W\}.$$

Note that $h(b) = c$, and

$$Dh(b) = -\big(D_y g(a)\big)^{-1} D_z g(a).$$

Now consider the function $\phi : W \to \mathbb{R}$ defined by $\phi(z) := f(z, h(z))$. Then ϕ has a local extremum at b. Thus we must have

$$\begin{aligned} 0 = D\phi(b) &= D_z f(b, h(b)) + D_y f(b, h(b)) Dh(b) \\ &= D_z f(a) - D_y f(a)\big(D_y g(a)\big)^{-1} D_z g(a). \end{aligned}$$

Hence we have $D_z f(a) = D_y f(a)\big(D_y g(a)\big)^{-1} D_z g(a)$. In addition, we have

$$D_y f(a) = D_y f(a) I = D_y f(a)\big(D_y g(a)\big)^{-1} D_y g(a).$$

Therefore $Df(a) = D_y f(a)\big(D_y g(a)\big)^{-1} Dg(a)$, since the columns of Df, Dg, i.e. $D_{z_i} f, D_{y_j} f$, $D_{z_i} g, D_{y_j} g$, satisfy this equation. Now note that $D_y f$ is a $1 \times k$ matrix and $(D_y g)^{-1}$ is a $k \times k$ matrix; so $D_y f(D_y g)^{-1}$ is a $1 \times k$ matrix. Let us denote the $1 \times k$ matrix $D_y f(D_y g)^{-1}$ by $\begin{bmatrix}\lambda_1 & \dots & \lambda_k\end{bmatrix}$. Then we get

$$Df(a) = \begin{bmatrix}\lambda_1 & \dots & \lambda_k\end{bmatrix} Dg(a) = \begin{bmatrix}\lambda_1 & \dots & \lambda_k\end{bmatrix} \begin{bmatrix}Dg_1(a) \\ \vdots \\ Dg_k(a)\end{bmatrix},$$

which gives the desired. ■

When $k = 1$ we must have

$$Df(a) = \lambda Dg(a)$$

for some $\lambda \in \mathbb{R}$, under the assumption $Dg(a) \neq 0$. Let us present another proof for this case, which provides a geometric intuition for why the theorem holds. Note that in this case $h : W \to \mathbb{R}$ for some open set $W \subset \mathbb{R}^{n-1}$, and Γ is the graph of h. Suppose to the contrary that $Df(a)$ is not a multiple of $Dg(a)$. Then the idea is that since the projection of $Df(a)$ on Γ is nonzero, if on Γ we move along the direction of the projection of $Df(a)$ and its opposite direction, the value of f will increase and decrease; and therefore $f|_\Gamma$ cannot have a local extremum at a.

Now since $g(z, h(z)) = 0$, for $j < n$ we have

$$D_j g + D_n g D_j h = 0 \implies Dg \cdot (e_j, D_j h) = 0.$$

We intuitively know that the vectors $(e_j, D_j h)$ are tangent to Γ at a, and so $Dg(a)$ is normal to its level set Γ at a (for more details see Section 9.3). Let us simplify the notation by setting

$$u := \big(Dg(a)\big)^{\mathsf{T}}, \qquad v := \big(Df(a)\big)^{\mathsf{T}}.$$

Then $\frac{v \cdot u}{|u|^2} u$ is the orthogonal projection of v on u. Thus

$$w := v - \frac{v \cdot u}{|u|^2} u$$

is the orthogonal projection of v on the $(n-1)$-dimensional plane tangent to Γ at a. Note that since we assumed that v is not a multiple of u, we have $w \neq 0$. Also note that w is orthogonal to u; so $u \cdot w = 0$.

Let us denote a vector in $\mathbb{R}^n$ like w by $(\tilde{w}, w_n)$, where $\tilde{w} \in \mathbb{R}^{n-1}$. Now consider the points $b_t := b + t\tilde{w}$, which for small t lie in W. Then $(b_t, h(b_t)) \in \Gamma$. Remember that $\phi(\cdot) = f(\cdot, h(\cdot))$. So for some sublinear function ρ we have

$$\begin{aligned}\phi(b_t) &= \phi(b) + D\phi(b)(b_t - b) + \rho(b_t - b) \\ &= \phi(b) + \big(D_z f(a) - D_y f(a)\big(D_y g(a)\big)^{-1} D_z g(a)\big)(b_t - b) + \rho(b_t - b) \\ &= f(a) + t(\tilde{v} - \tfrac{v_n}{u_n}\tilde{u}) \cdot \tilde{w} + \rho(t\tilde{w}).\end{aligned}$$

Note that $\tilde{u} \cdot \tilde{w} = u \cdot w - u_n w_n = 0 - u_n w_n = -u_n w_n$. Hence

$$(\tilde{v} - \tfrac{v_n}{u_n}\tilde{u}) \cdot \tilde{w} = \tilde{v} \cdot \tilde{w} - \tfrac{v_n}{u_n}\tilde{u} \cdot \tilde{w} = \tilde{v} \cdot \tilde{w} + v_n w_n = v \cdot w > 0,$$

because $v \cdot w = |v|^2 - \frac{(v \cdot u)^2}{|u|^2} \geq 0$ by Cauchy-Schwarz inequality; and $v \cdot w \neq 0$, since otherwise we would get $w \cdot w = 0$ (note that we also have $w \cdot u = 0$), which would have implied $w = 0$, contrary to our assumption. Finally since ρ is sublinear, for small t we have $|\rho(t\tilde{w})| < |t|(\tilde{v} - \frac{v_n}{u_n}\tilde{u}) \cdot \tilde{w}$. Therefore $f(b_t, h(b_t)) = \phi(b_t)$ will be larger and smaller than $f(a)$ for positive and negative values of t, respectively. Hence f cannot have a local extremum at a.

7.7 Holomorphic Functions and Cauchy-Riemann Equations

Definition 7.45. Suppose $U \subset \mathbb{C}$ is open. A function $f : U \to \mathbb{C}$ is **complex differentiable** at $z \in U$ if

$$f'(z) := \lim_{w \to z} \frac{f(w) - f(z)}{w - z}$$

exists. $f'(z)$ is called the **(complex) derivative** of f at z.

Remark. The proof of the following three theorems goes along the same lines as their corresponding theorems regarding functions of one real variable.

Theorem 7.46. *The complex derivative of a constant function equals zero everywhere. Also for $n \in \mathbb{N}$, the functions $f(z) = z^n$ from $\mathbb{C}$ to $\mathbb{C}$ are complex differentiable with $f'(z) = nz^{n-1}$. In addition, $g(z) = \frac{1}{z}$ from $\mathbb{C} - \{0\}$ to $\mathbb{C}$ is complex differentiable with $g'(z) = \frac{-1}{z^2}$.*

Theorem 7.47. *Suppose f is complex differentiable at z, then f is continuous at z.*

Theorem 7.48. *Suppose f, g are complex differentiable at z, F is complex differentiable at $f(z)$, and $c_1, c_2 \in \mathbb{C}$. Then $c_1 f + c_2 g$, fg and $F \circ f$ are complex differentiable at z with derivatives*

$$\begin{aligned}
&(c_1 f + c_2 g)'(z) = c_1 f'(z) + c_2 g'(z),\\
&(fg)'(z) = f'(z)g(z) + f(z)g'(z),\\
&(F \circ f)'(z) = F'(f(z))f'(z).
\end{aligned}$$

If in addition $g \neq 0$, then $\frac{f}{g}$ is also complex differentiable at z and

$$(\frac{f}{g})'(z) = \frac{g(z)f'(z) - g'(z)f(z)}{(g(z))^2}.$$

***Example* 7.49.** A polynomial function $p(z) = a_0 + a_1 z + \cdots + a_n z^n$ is complex differentiable at every point, and we have $p'(z) = a_1 + 2a_2 z + \cdots + n a_n z^{n-1}$.

Theorem 7.50. *Suppose f is a function from an open subset of $\mathbb{C}$ into $\mathbb{C}$. We can also regard f as a function from an open subset of $\mathbb{R}^2$ into $\mathbb{R}^2$. Then, f is complex differentiable at a point z, with $f'(z) = a + ib$ for some $a, b \in \mathbb{R}$, if and only if f is differentiable at z with*

$$Df(z) = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}.$$

Proof. Let $h = c + id$ be a complex number. We know that h is the same as the vector (c, d) in $\mathbb{R}^2$. Now we have

$$\begin{aligned} f'(z)h &= (a + ib)(c + id) = ac - bd + i(ad + bc) \\ &= (ac - bd, ad + bc) = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = Df(z)h. \end{aligned}$$

Thus we have

$$\left| \frac{f(z + h) - f(z) - Df(z)h}{|h|} \right| = \left| \frac{f(z + h) - f(z) - f'(z)h}{h} \right|.$$

Hence when $h \to 0$, one of the above expressions goes to zero if and only if the other one goes to zero, and therefore we get the desired result. ■

Cauchy-Riemann Equations. *Suppose the function $f = u + iv$ is complex differentiable at the point $c + id$, where $c, d \in \mathbb{R}$ and u, v are real-valued. Then*

$$u_x(c, d) = v_y(c, d), \qquad u_y(c, d) = -v_x(c, d).$$

Proof. By the previous theorem we know that $f = u + iv = (u, v)$ is also differentiable at $c + id = (c, d)$. In addition, we know that if $f'(c + id) = a + ib$ then

$$\begin{bmatrix} u_x(c, d) & u_y(c, d) \\ v_x(c, d) & v_y(c, d) \end{bmatrix} = Df(c, d) = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}.$$

Hence we must have $u_x(c, d) = a = v_y(c, d)$, and $u_y(c, d) = -b = -v_x(c, d)$, as desired. ■

Theorem 7.51. *Suppose the function $f = u + iv$ is complex differentiable at the point z, where u, v are real-valued. Then*

$$f'(z) = u_x(z) + iv_x(z) = v_y(z) - iu_y(z).$$

Proof. By Theorem 7.50 we know that $f = u + iv = (u, v)$ is also differentiable at z. In addition, we know that if $f'(z) = a + ib$ then

$$\begin{bmatrix} u_x(z) & u_y(z) \\ v_x(z) & v_y(z) \end{bmatrix} = Df(z) = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}.$$

Hence we must have $a = u_x(z) = v_y(z)$, and $b = v_x(z) = -u_y(z)$. Therefore we get the desired formulas for $f'(z)$. ■

Definition 7.52. Suppose $U \subset \mathbb{C}$ is open. A function $f : U \to \mathbb{C}$ is **holomorphic** on U if it is complex differentiable at every point of U. We denote by f', the function from U into $\mathbb{C}$ that takes $z \mapsto f'(z)$.

Remark. It is an obvious consequence of Theorem 7.48 that if f, g are holomorphic on an open set $U \subset \mathbb{C}$, and F is holomorphic on an open set containing $f(U)$, then $c_1 f + c_2 g$, fg, and $F \circ f$ are holomorphic on U. If in addition $g \neq 0$ on U, then $\frac{f}{g}$ is also holomorphic on U.

Theorem 7.53. *A function on an open set with continuous partial derivatives satisfying Cauchy-Riemann equations, is holomorphic.*

Proof. Having continuous partial derivatives implies differentiability, and satisfying the Cauchy-Riemann equations implies complex differentiability by Theorem 7.50. ■

Theorem 7.54. *Suppose U, V are open subsets of $\mathbb{C}$, and $f : U \to V$ is an invertible function. If f has a nonzero complex derivative at a, and f^{-1} is continuous at $f(a)$, then f^{-1} is complex differentiable at $f(a)$ and*

$$(f^{-1})'(f(a)) = \frac{1}{f'(a)}.$$

Proof. Let $b = f(a)$, and $z = f^{-1}(b + h)$, where h is small enough so that $b + h \in V$. Then we have

$$\frac{f^{-1}(b+h) - f^{-1}(b)}{h} = \frac{f^{-1}(b+h) - a}{b + h - b} = \frac{z - a}{f(z) - f(a)} = \frac{1}{\frac{f(z) - f(a)}{z - a}}.$$

When h is small, $z = f^{-1}(b+h)$ is close to $f^{-1}(b) = a$, since f^{-1} is continuous at $b = f(a)$. Hence the above fraction is close to $\frac{1}{f'(a)}$, and therefore f^{-1} is differentiable at b with the desired derivative. ■

***Example* 7.55.** The function e^z is holomorphic on $\mathbb{C}$, and we have $(e^z)' = e^z$. To see this note that for $z = x + iy$ we have

$$e^z = e^x \cos y + ie^x \sin y =: u + iv.$$

Therefore $u_x = e^x \cos y = v_y$, and $u_y = -e^x \sin y = -v_x$. So the Cauchy-Riemann equations are satisfied. Since the partial derivatives of e^z are obviously continuous, e^z is holomorphic. Furthermore we have $(e^z)' = u_x + iv_x = e^z$, as desired.

Theorem 7.56. *A holomorphic function f with zero complex derivative on an open connected set, is constant on that set.*

Proof. If $f' = 0$ then $Df = 0$. Hence f is constant by Theorem 7.40. ■

Chapter 8

Multiple Integrals

8.1 Multiple Riemann Integrals

Definition 8.1. A **closed rectangle** in $\mathbb{R}^n$ is a product of n bounded closed intervals, i.e. it is a set of the form

$$[a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n.$$

Similarly, an **open rectangle** in $\mathbb{R}^n$ is a product of n bounded open intervals, i.e. it is a set of the form

$$(a_1, b_1) \times \cdots \times (a_n, b_n) \subset \mathbb{R}^n.$$

In general, a **rectangle** R in $\mathbb{R}^n$ is a product of n bounded intervals, i.e. there are bounded intervals $I_1, I_2, \ldots, I_n \subset \mathbb{R}$ such that

$$R := I_1 \times I_2 \times \cdots \times I_n \subset \mathbb{R}^n.$$

Each interval I_i can be closed, open, or half-open. The intervals I_i are called the **edges** of R. Let a_i, b_i be the left endpoint and the right endpoint of I_i respectively. Then $b_i - a_i$ is the length of the interval I_i. The rectangle R is called a **cube** if $b_i - a_i = b_1 - a_1$ for all $i \leq n$. When $n = 2$, cubes are called **squares**. The **volume** of the rectangle R is the positive real number

$$|R| := (b_1 - a_1) \cdots (b_n - a_n).$$

When $n = 1, 2$, the volume is called the **length** or the **area**, respectively. The points $(c_1, \ldots, c_n)$ where each c_i is either a_i or b_i, are called the **vertices** of the rectangle R.

Remark. Note that a closed rectangle is closed, being a product of closed sets; and an open rectangle is open, being a product of open sets. Let R be the rectangle in the above definition. Then we have

$$\overline{R} = [a_1, b_1] \times \cdots \times [a_n, b_n], \qquad R^\circ = (a_1, b_1) \times \cdots \times (a_n, b_n).$$

To see this let O, C denote the above open rectangle and closed rectangle respectively. Then O is open, C is closed, and $O \subset R \subset C$. Hence $O \subset R^\circ$ and $\overline{R} \subset C$. Also, it is easy to see that each point of $C - O$ is the limit of a sequence of points of O, and is also the limit of a sequence of points of C^c. Therefore $C \subset \overline{O} \subset \overline{R}$, and $(C - O) \cap R^\circ = \emptyset$. Hence $\overline{R} = C$, and $R^\circ = O$, as desired. In other words, the closure of a rectangle is a closed rectangle, and the interior of a rectangle is an open rectangle. As a result we also have

$$\partial R = \overline{R} - R^\circ = \bigcup_{i \le n} [a_1, b_1] \times \cdots \times \{a_i, b_i\} \times \cdots \times [a_n, b_n].$$

Definition 8.2. A **partition** P of an interval $[a, b] \subset \mathbb{R}$ is a finite set of points $\{c_0, \ldots, c_m\}$ such that

$$a = c_0 < c_1 < \cdots < c_m = b.$$

The interval $[c_{i-1}, c_i]$ is called the ith **subinterval** of the partition P. The **mesh** of the partition P is

$$\|P\| := \max_{i \le m} |c_i - c_{i-1}|.$$

Definition 8.3. A **partition** of the closed rectangle

$$R = [a_1, b_1] \times \cdots \times [a_n, b_n],$$

is a cartesian product $P := P_1 \times \cdots \times P_n$, where each P_i is a partition of the interval $[a_i, b_i]$. Suppose $[c_i, d_i]$ is a subinterval of the partition P_i, then the closed rectangle

$$[c_1, d_1] \times \cdots \times [c_n, d_n]$$

is called a **subrectangle** of the partition P. If P_i divides $[a_i, b_i]$ into N_i subintervals, then P divides R into $N_1 \cdots N_n$ subrectangles. We denote these subrectangles by R_α, where α is the multi-index $(\alpha_1, \ldots, \alpha_n)$ such that $1 \le \alpha_i \le N_i$. In this notation, R_α denotes the subrectangle $I_{\alpha_1} \times \cdots \times I_{\alpha_n}$, where I_{α_i} is the α_ith subinterval of P_i. We sometimes abuse the notation and write $P = \{R_\alpha\}$.

The **mesh** of the partition P is

$$\|P\| := \left(\|P_1\|^2 + \cdots + \|P_n\|^2\right)^{\frac{1}{2}},$$

i.e. $\|P\|$ is the largest diameter of the subrectangles of P. A **tagged partition** is a partition $P = \{R_\alpha\}$ with a finite sequence $T = (x_\alpha)$ of **tags**

$$x_\alpha \in R_\alpha.$$

We say a partition Q is a **refinement** of a partition P if $P_i \subset Q_i$ for each i. The **common refinement** of two partitions P, P^* is

$$(P_1 \cup P_1^*) \times \cdots \times (P_n \cup P_n^*).$$

Remark. It is easy to see that for a refinement Q of P we have $\|Q\| \le \|P\|$.

Theorem 8.4. *Let $R \subset \mathbb{R}^n$ be a closed rectangle.*

(i) *Suppose $P = \{R_\alpha\}$ is a partition of R. Then we have*

$$|R| = \sum |R_\alpha|.$$

(ii) *Suppose $R_1, \dots, R_m \subset R$ are closed rectangles that have pairwise disjoint interiors, i.e. $R_i^\circ \cap R_k^\circ = \emptyset$ for every $i \neq k$. Then*

$$\sum |R_i| \le |R|.$$

Proof. Let $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$.

(i) Suppose $P = P_1 \times \cdots \times P_n$. Let us denote the subrectangles of P by $R_\alpha = I_{\alpha_1} \times \cdots \times I_{\alpha_n}$, where I_{α_i} is a subinterval of P_i and $1 \le \alpha_i \le N_i$. Then we have

$$\sum |R_\alpha| = \sum_{\alpha_1=1}^{N_1} \cdots \sum_{\alpha_n=1}^{N_n} |I_{\alpha_1}| \cdots |I_{\alpha_n}| = \prod_{i=1}^{n} \Big(\sum_{\alpha_i=1}^{N_i} |I_{\alpha_i}| \Big) = \prod_{i=1}^{n} (b_i - a_i) = |R|.$$

(ii) Suppose

$$R_i = [a_{i1}, b_{i1}] \times \cdots \times [a_{in}, b_{in}].$$

Let Q_j be a partition of $[a_j, b_j]$ that contains a_{ij}, b_{ij} for all $i \le m$. Then for a fixed i, $Q_j \cap [a_{ij}, b_{ij}]$ is a partition of $[a_{ij}, b_{ij}]$. Now $Q = \prod_{j \le n} Q_j$ is a partition of R that contains all the vertices of every R_i. Then note that

$$Q \cap R_i = \Big(\prod_{j \le n} Q_j \Big) \bigcap \Big(\prod_{j \le n} [a_{ij}, b_{ij}] \Big) = \prod_{j \le n} (Q_j \cap [a_{ij}, b_{ij}])$$

is a partition of R_i. Suppose $\{S_\alpha\}$ is the set of subrectangles of Q. Then the set of subrectangles of $Q \cap R_i$ is

$$\{S_\alpha : S_\alpha \subset R_i\}.$$

To see this note that each subinterval of $Q_j \cap [a_{ij}, b_{ij}]$ is also a subinterval of Q_j. Thus the subrectangles of $Q \cap R_i$ belong to the set $\{S_\alpha\}$. The subrectangles of $Q \cap R_i$ are also obviously subsets of R_i, so they belong to $\{S_\alpha : S_\alpha \subset R_i\}$. On the other hand, if S_α is a subrectangle of Q such that $S_\alpha = \prod_{j \le n} I_{\alpha_j} \subset R_i$, then I_{α_j} is a subinterval of $Q_j \cap [a_{ij}, b_{ij}]$. Therefore S_α is a subrectangle of $Q \cap R_i$ as desired. As a result we have

$$\sum_{S_\alpha \subset R_i} |S_\alpha| = |R_i|.$$

In addition, for $i \neq k$ we have

$$\{S_\alpha : S_\alpha \subset R_i\} \cap \{S_\alpha : S_\alpha \subset R_k\} = \emptyset.$$

Because if $S_\alpha \subset R_i$ then $S_\alpha^\circ \subset R_i^\circ$. Hence $S_\alpha^\circ \cap R_k^\circ = \emptyset$. So we cannot have $S_\alpha \subset R_k$, since S_α° is nonempty. Therefore we finally get

$$\sum_{i=1}^{m} |R_i| = \sum_{i=1}^{m} \sum_{S_\alpha \subset R_i} |S_\alpha| \leq \sum_{\text{all } \alpha} |S_\alpha| = |R|.$$

Note that in the left hand side of the above inequality, no $|S_\alpha|$ can appear more than once. ■

Remark. Similarly to the above theorem we can show that if $R, R_1, \dots, R_m$ are closed rectangles in $\mathbb{R}^n$ such that $R \subset \bigcup R_i$, then we have

$$|R| \leq \sum |R_i|.$$

Note that this fact is nontrivial even when $n = 1$. If in addition $R_1, \dots, R_m$ have pairwise disjoint interiors, i.e. $R_i^\circ \cap R_k^\circ = \emptyset$ for every $i \neq k$; and if $R = \bigcup R_i$, then we have

$$|R| = \sum |R_i|.$$

We should mention that we do not use these facts in the development of multiple Riemann integrals.

Definition 8.5. Let R be a closed rectangle in $\mathbb{R}^n$, and let $f : R \to \mathbb{R}^m$. The **Riemann sum** of f corresponding to the tagged partition $P = \{R_\alpha\}, T = (x_\alpha)$ of R is

$$R(f, P, T) := \sum_{\alpha} f(x_\alpha)|R_\alpha|.$$

Definition 8.6. Let R be a closed rectangle in $\mathbb{R}^n$, and let $f : R \to \mathbb{R}^m$. We say f is **Riemann integrable** (on R), if there exists $I \in \mathbb{R}^m$ so that $\forall \epsilon > 0\ \exists \delta > 0$ such that for all tagged partitions P, T with $\|P\| < \delta$ we have

$$|I - R(f, P, T)| < \epsilon.$$

Note that by the next theorem I is unique. We call I the **Riemann integral** of f (over R) and denote it by

$$\int_R f(x)dx = \int_R f(x_1, \dots, x_n)\, dx_1 \dots dx_n.$$

In this notation, the function f is also referred to as the *integrand.*

Remark. Note that in this definition, if a partition P satisfies $\|P\| < \delta$, then for any choice of tags T for P we have $|I - R(f, P, T)| < \epsilon$. In other words, for a partition with small enough mesh, the choice of tags does not affect the closeness of the Riemann sum to the integral of a Riemann integrable function.

Remark. When $n = 1$ and R is the closed interval $[a, b]$, we use the usual notation $\int_a^b$ instead of $\int_{[a,b]}$.

Theorem 8.7. *The integral of an integrable function is unique.*

Proof. Suppose there are two vectors $I, J \in \mathbb{R}^m$ satisfying the above definition. Then for each $\epsilon > 0$ there is a tagged partition P, T such that

$$|I - J| \leq |I - R(f, P, T)| + |J - R(f, P, T)| < \epsilon + \epsilon = 2\epsilon.$$

Hence we must have $I - J = 0$. ■

Theorem 8.8. *A Riemann integrable function is bounded.*

Proof. Suppose to the contrary that f is an unbounded Riemann integrable function on the rectangle R. Then there is $\delta > 0$ such that for all tagged partitions P, T with $\|P\| < \delta$ we have

$$|I - R(f, P, T)| < 1. \tag{$*$}$$

Let $P = \{R_\alpha\}, T = (x_\alpha)$ be a tagged partition with $\|P\| < \delta$. Since f is unbounded, it is unbounded on at least one of the subrectangles R_β. Let $x'_\alpha = x_\beta$ for $\alpha \neq \beta$. Then we can find $x'_\beta \in R_\beta$ such that for $T' = (x'_\alpha)$ we have

$$|R(f, P, T') - R(f, P, T)| = |f(x'_\beta) - f(x_\beta)||R_\beta| > 2.$$

But this contradicts $(*)$. ■

Theorem 8.9. *Let R be a closed rectangle in $\mathbb{R}^n$, and let $f = (f_1, \dots, f_m) : R \to \mathbb{R}^m$. Then f is Riemann integrable if and only if each f_i is Riemann integrable, and in this case we have*

$$\int_R f(x)dx = \Big(\int_R f_1(x)dx, \dots, \int_R f_m(x)dx\Big).$$

Proof. For each tagged partition P, T of R we have

$$R(f, P, T) = \big(R(f_1, P, T), \dots, R(f_m, P, T)\big).$$

Thus $R(f, P, T)$ is close to $I = (I_1, \dots, I_m)$ if and only if each $R(f_i, P, T)$ is close to I_i. ■

Remark. By the above theorem, in the remainder of this chapter we can assume that $m = 1$ without any loss of generality.

Theorem 8.10. *Let R be a closed rectangle in $\mathbb{R}^n$, and let $f, g : R \to \mathbb{R}$ be Riemann integrable. Then for $c, c_1, c_2 \in \mathbb{R}$ we have*

(i) *The constant function c is Riemann integrable and $\int_R c\, dx = c|R|$.*

(ii) *$c_1 f + c_2 g$ is Riemann integrable and*

$$\int_R [c_1 f(x) + c_2 g(x)] dx = c_1 \int_R f(x) dx + c_2 \int_R g(x) dx.$$

(iii) *If $f \le g$ then*

$$\int_R f(x) dx \le \int_R g(x) dx.$$

(iv) *If $|f| \le M$ then*

$$\left| \int_R f(x) dx \right| \le M|R|.$$

$\boxed{\textbf{Proof.}}$ **(i)** The Riemann sums of the constant function c corresponding to a partition $P = \{R_\alpha\}$ are all equal to $\sum c|R_\alpha| = c|R|$, because of Theorem 8.4. Hence the Riemann sums converge to $c|R|$.

(ii) First note that for any tagged partition P, T we have

$$R(c_1 f + c_2 g, P, T) = c_1 R(f, P, T) + c_2 R(g, P, T).$$

Now let I, J be the integrals of f, g respectively. Let δ be small enough so that for all tagged partitions P, T with $\|P\| < \delta$ we have

$$|I - R(f, P, T)| < \frac{\epsilon}{2|c_1| + 2}, \qquad |J - R(g, P, T)| < \frac{\epsilon}{2|c_2| + 2}.$$

Then

$$\begin{aligned} &|c_1 I + c_2 J - R(c_1 f + c_2 g, P, T)| \\ &\quad = |c_1 I + c_2 J - c_1 R(f, P, T) - c_2 R(g, P, T)| \\ &\quad \le |c_1||I - R(f, P, T)| + |c_2||J - R(g, P, T)| < \epsilon. \end{aligned}$$

(iii) First note that for all tagged partitions P, T we have

$$R(f, P, T) \le R(g, P, T).$$

Now let I, J be the integrals of f, g respectively. Suppose to the contrary that $J < I$. Let δ be small enough so that for all tagged partitions P, T with $\|P\| < \delta$ we have

$$|I - R(f, P, T)| < \frac{I - J}{2}, \qquad |J - R(g, P, T)| < \frac{I - J}{2}.$$

Then we must have $R(f, P, T) > R(g, P, T)$, which is a contradiction.

(iv) We have $-M \le f \le M$. Now the result follows from parts (i) and (iii). ■

***Example* 8.11.** Consider the characteristic function of $\mathbb{Q}^n$, i.e.

$$\chi_{\mathbb{Q}^n}(x) := \begin{cases} 1 & x \in \mathbb{Q}^n, \\ 0 & x \notin \mathbb{Q}^n. \end{cases}$$

Then for any partition P of a rectangle R we can choose a sequence T of tags with rational coordinates, and a sequence T' of tags with irrational coordinates so that

$$R(\chi_{\mathbb{Q}^n}, P, T) = |R|, \qquad R(\chi_{\mathbb{Q}^n}, P, T') = 0.$$

Hence $\chi_{\mathbb{Q}^n}$ is not Riemann integrable on any rectangle R. It should also be noted that $\chi_{\mathbb{Q}^n}$ is not continuous at any point. (why?)

Definition 8.12. A subset A of $\mathbb{R}^n$ has **measure zero** if for every $\epsilon > 0$ there exist a countable family of open cubes in $\mathbb{R}^n$, $\{Q_i\}$, such that $A \subset \bigcup_{i\geq 1} Q_i$, and

$$\sum_{i\geq 1} |Q_i| < \epsilon.$$

We say a property holds **almost everywhere**, abbreviated **a.e.**, if it holds for all points outside a set of measure zero.

Remark. Remember that a countable set is either finite or countably infinite.

Remark. An obvious consequence of the definition is that if A has measure zero and $B \subset A$, then B has measure zero too.

Proposition 8.13. *A set $A \subset \mathbb{R}^n$ has measure zero if and only if for every $\epsilon > 0$ there exist a countable family of open rectangles in $\mathbb{R}^n$, $\{R_i\}$, such that $A \subset \bigcup_{i\geq 1} R_i$, and*

$$\sum_{i\geq 1} |R_i| < \epsilon.$$

Proof. If A has measure zero, then for every positive ϵ there is obviously a family of open rectangles with the desired properties, namely the family of open cubes in the definition of zero measure. Conversely, suppose that for every positive ϵ there is a family of open rectangles $\{R_i\}$ that covers A, such that $\sum_{i\geq 1} |R_i| < \epsilon$. Consider a fixed i, and suppose we have

$$R_i = (a_1, b_1) \times \cdots \times (a_n, b_n).$$

Let $l_j := b_j - a_j$, and let l be a positive number less than $\min_{j\leq n} l_j$. Then we have $k_j := \lfloor \frac{l_j}{l} \rfloor \in \mathbb{N}$. We also have $k_j l \leq l_j < (k_j + 1)l = k_j l + l$. Now consider the open rectangle

$$S_i = (a_1, a_1 + k_1 l + l) \times \cdots \times (a_n, a_n + k_n l + l).$$

Obviously we have $R_i \subset S_i$. In addition we have

$$|S_i| = \prod_{j\le n}(k_j l + l) \le \prod_{j\le n}(l_j + l) \le \prod_{j\le n} 2l_j = 2^n |R_i|.$$

Now each interval $[a_j, a_j + k_j l + l]$ has a partition with $k_j + 1$ closed subintervals of length l. Then we get a partition of $\overline{S}_i$ with $N_i := \prod_{j\le n}(k_j + 1)$ subrectangles, which are all closed cubes with volume l^n. Note that by Theorem 8.4, the volume of $\overline{S}_i$ is the sum of the volumes of these closed cubes; hence $|S_i| = |\overline{S}_i| = N_i l^n$. We can cover each of these closed cubes by an open cube whose volume is $2l^n$. Call these open cubes Q_{ij}, where $j \le N_i$. Then we have $\sum_{j\le N_i} |Q_{ij}| = 2|S_i|$.

We can repeat the above construction for every i, to get a countable family $\{S_i\}$ of open rectangles that covers A, such that

$$\sum_{i\ge 1} |S_i| \le \sum_{i\ge 1} 2^n |R_i| < 2^n \epsilon.$$

Now $\{Q_{ij} : i \ge 1,\ j \le N_i\}$ is a family of open cubes that covers A; and it is also countable, since it is the union of countably many finite families. We consider this family with the order

$$Q_{11}, Q_{12}, \dots, Q_{1N_1}, Q_{21}, \dots, Q_{2N_2}, \dots, Q_{m1}, \dots, Q_{mN_m}, \dots.$$

Let us denote the kth cube in this sequence by Q_k. Then for $N \le N_1 + \cdots + N_m$ we have

$$\sum_{k=1}^{N} |Q_k| \le \sum_{i=1}^{m} \sum_{j\le N_i} |Q_{ij}| = 2\sum_{i=1}^{m} |S_i| \le 2\sum_{i\ge 1} |S_i| < 2^{n+1}\epsilon.$$

By taking the limit as $N \to \infty$ we obtain $\sum_{k\ge 1} |Q_k| \le 2^{n+1}\epsilon$. Now as ϵ is arbitrary, we get the desired result. ■

Definition 8.14. Let $\{R_i\}_{i\ge 1}$ be a countable family of rectangles in $\mathbb{R}^n$. The series $\sum_{i\ge 1} |R_i|$ is called the **total volume** of the family. When $n = 1, 2$, the total volume is called the **total length** or the **total area**, respectively. Note that the order of rectangles has no effect on the total volume of the family, since the volume of each rectangle is positive and therefore the series of the total volume is absolutely convergent.

Remark. Note that the concept of zero measure depends on the dimension n. For example the interval $[0, 1]$ does not have measure zero as a subset of $\mathbb{R}$, but if we regard it as the subset $[0, 1] \times \{0\}$ of $\mathbb{R}^2$, it has measure zero. To see this, notice that

$$[0, 1] \times \{0\} \subset R := (-1, 2) \times (-\epsilon, \epsilon),$$

and $|R| = 6\epsilon$.

Theorem 8.15. *Let $\{A_k\}$ be a countable family of subsets of $\mathbb{R}^n$ that have measure zero. Then $\bigcup_k A_k$ has measure zero. In particular, every countable subset of $\mathbb{R}^n$ has measure zero.*

Proof. Let $\epsilon > 0$ be given. Then we can cover A_k with a countable family of open rectangles $\{R_{ki}\}_{i\geq 1}$ such that

$$\sum_{i\geq 1}|R_{ki}| < \frac{\epsilon}{2^k}.$$

Then $\{R_{ki}\}_{i,k\geq 1}$ is a countable family of open rectangles that covers $\bigcup_k A_k$, and

$$\sum_{i,k\geq 1}|R_{ki}| < \sum_{k\geq 1}\frac{\epsilon}{2^k} \leq \epsilon.$$

The final statement of the theorem follows from the trivial fact that a set with one element has measure zero. ■

Remark. If we want to be completely rigorous in the above proof, we have to arrange the family of rectangles $\{R_{ki}\}_{i,k\geq 1}$ into a sequence. Note that different arrangements does not change the total volume of the family, since the volume of each rectangle is positive and therefore the series of the total volume is absolutely convergent. Now suppose we have arranged the family as the sequence $\{R_j\}_{j\geq 1}$. Then for any $N \in \mathbb{N}$ there is $M \in \mathbb{N}$ such that

$$\{R_j\}_{1\leq j\leq N} \subset \{R_{ki}\}_{1\leq i,k\leq M}.$$

Then we have

$$\sum_{j\leq N}|R_j| \leq \sum_{k\leq M}\sum_{i\leq M}|R_{ki}| < \sum_{k\leq M}\frac{\epsilon}{2^k} < \epsilon.$$

Now by taking the limit as $N \to \infty$ we get $\sum_{j\geq 1}|R_j| \leq \epsilon$ as desired.

Proposition 8.16. *Let $a \in \mathbb{R}$. Then the $(n-1)$-dimensional plane*

$$\{x \in \mathbb{R}^n : x_i = a\}$$

has measure zero in $\mathbb{R}^n$. Also, for every rectangle $R \subset \mathbb{R}^n$, ∂R has measure zero.

Proof. Let P be the described plane. Then we have $P \subset \bigcup_{j\geq 1} R_j$, where R_j is the rectangle $\prod_{k\leq n} I_k$ in which $I_k = (-2^{j-1}, 2^{j-1})$ for $k \neq i$, and

$$I_i = (a - \epsilon 2^{-nj-1}, a + \epsilon 2^{-nj-1}).$$

Now we have $|R_j| = \epsilon 2^{-nj}2^{j(n-1)} = \epsilon 2^{-j}$. Hence

$$\sum_{j\geq 1} |R_j| = \epsilon \sum_{j\geq 1} 2^{-j} = \epsilon.$$

Therefore P has measure zero, since ϵ is arbitrary. Finally note that ∂R is contained in the union of finitely many planes, so it has measure zero. ■

Definition 8.17. Let R be a subset of $\mathbb{R}^n$, and let $f : R \to \mathbb{R}$. We define the **oscillation** of f at a point $x \in R$ as

$$\operatorname{osc}_x f := \lim_{r\to 0} \sup\{\, |f(z) - f(y)| : z, y \in B_r(x) \cap R\}.$$

Remark. Note that the supremums in the above expression decrease as $r \to 0$, hence the limit exists.

Remark. The oscillation at a point is a measure of the size of the discontinuity at that point. In particular, we can easily see that $\operatorname{osc}_x f = 0$ if and only if f is continuous at x.

Riemann-Lebesgue Theorem (Rectangular Domains). *Let R be a closed rectangle in $\mathbb{R}^n$, and let $f : R \to \mathbb{R}$. Then f is Riemann integrable if and only if it is bounded and its set of discontinuities has measure zero.*

Proof. Let D be the set of discontinuities of f. Suppose D has measure zero. Also suppose that $|f| \leq M$ for some $M > 0$. We want to show that f is Riemann integrable. The idea is to show that the Riemann sums of f satisfy a Cauchy criterion. Let $\epsilon > 0$ be given. Then there are countably many open rectangles A_k such that

$$D \subset \bigcup A_k, \qquad \sum |A_k| < \epsilon.$$

Now, f is continuous at each point of $K := R - \bigcup A_k$. So for every $x \in K$ there is an open rectangle I_x containing x such that $|f(x) - f(y)| < \epsilon$ whenever $y \in I_x \cap R$. Then the collection

$$\mathcal{U} := \{A_k\}_{k\geq 1} \cup \{I_x\}_{x\in K}$$

is an open covering of the compact set R. Thus it has a Lebesgue number $\delta > 0$, i.e. for every $z \in R$ there is $U \in \mathcal{U}$ such that $B_\delta(z) \subset U$.

Let $P = \{R_\alpha\}, T = (x_\alpha)$ be a tagged partition of R, with $\|P\| < \delta$. Let $P' = \{R'_\beta\}$ be a refinement of P, and let $T' = (x'_\beta)$ be a choice of tags for P'. Then we have

$$R(f, P, T) = \sum_\alpha f(x_\alpha)|R_\alpha| = \sum_\alpha \Big(f(x_\alpha) \sum_{R'_\beta \subset R_\alpha} |R'_\beta|\Big) = \sum_\beta f(x_\beta)|R'_\beta|,$$

where $x_\beta := x_\alpha$ if $R'_\beta \subset R_\alpha$. Note that $|R_\alpha| = \sum_{R'_\beta \subset R_\alpha} |R'_\beta|$, since $\{R'_\beta : R'_\beta \subset R_\alpha\}$ is a partition of R_α.

Let

$$J := \{\beta : R'_\beta \subset R_\alpha \subset I_x \text{ for some } x \in K\}.$$

Then if $\beta \in J$ we have $x_\beta, x'_\beta \in I_x$ for some $x \in K$, and therefore $|f(x_\beta) - f(x'_\beta)| < 2\epsilon$. Now note that for any α and any $z \in R_\alpha$ we have

$$R_\alpha \subset B_\delta(z) \subset U$$

for some $U \in \mathcal{U}$. Hence if $\beta \notin J$ we must have

$$R'_\beta \subset R_\alpha \subset A_k$$

for some k. Thus

$$\sum_{\beta \notin J} |R'_\beta| \leq \sum |A_k| < \epsilon.$$

Note that k can be the same for several distinct $\beta_1, \beta_2, \dots$, i.e. we might have $R'_{\beta_j} \subset A_{k_0}$ for some k_0, and $j = 1, 2, \dots$. But this does not affect the inequality, since $\sum_j |R'_{\beta_j}| \leq |A_{k_0}|$. Thus we do not need to add $|A_{k_0}|$ several times in the right hand side. Hence we have

$$\begin{aligned}
|R(f,P,T) - R(f,P',T')| &\leq \sum \big|f(x_\beta) - f(x'_\beta)\big||R'_\beta| \\
&= \sum_{\beta \in J} \big|f(x_\beta) - f(x'_\beta)\big||R'_\beta| \\
&\qquad + \sum_{\beta \notin J} \big|f(x_\beta) - f(x'_\beta)\big||R'_\beta| \\
&< 2\epsilon \sum_{\beta \in J} |R'_\beta| + 2M \sum_{\beta \notin J} |R'_\beta| \\
&< 2\epsilon|R| + 2M\epsilon =: C\epsilon.
\end{aligned}$$

Now let P, P^* be two partitions with mesh less than δ. Let Q be the common refinement of P, P^*. Let T, T^*, S be choices of tags for P, P^*, Q, respectively. Then by the above inequality we get

$$\begin{aligned}
|R(f,P,T) - R(f,P^*,T^*)| &\leq |R(f,P,T) - R(f,Q,S)| \\
&\qquad + |R(f,P^*,T^*) - R(f,Q,S)| < 2C\epsilon.
\end{aligned}$$

This is the Cauchy criterion that we were looking for.

Finally let P_N be the partition that divides R into N^n equal subrectangles. Let T_N be the sequence of the midpoints of these subrectangles. Then for any $\epsilon > 0$ we can take N to be large enough so that $\|P_N\| = \frac{d}{N} < \delta$, where d is the diameter of R. Hence we have

$$|R(f, P_{N'}, T_{N'}) - R(f, P_N, T_N)| < 2C\epsilon,$$

for $N' \geq N$. Therefore the sequence $R(f, P_N, T_N)$ is Cauchy in $\mathbb{R}$. Thus it converges to some number I. Now let N be large enough so that $\|P_N\| < \delta$ and $|I - R(f, P_N, T_N)| < C\epsilon$. Then for an arbitrary tagged partition P, T with mesh less than δ, we have

$$|I - R(f, P, T)| \leq |I - R(f, P_N, T_N)| + |R(f, P_N, T_N) - R(f, P, T)| < 3C\epsilon.$$

As ϵ is arbitrary we get the desired result. ■

Proof of the Converse. Next suppose f is Riemann integrable on R, and its integral is I. Then we know that f is bounded. Let D be the set of discontinuities of f. We have $D = \bigcup_{k\geq 1} D_k$ where

$$D_k := \{x \in R : \operatorname{osc}_x f \geq \frac{1}{k}\}.$$

In order to show that D has measure zero, it suffices to show that each D_k has measure zero. Now for any given $\epsilon > 0$ we can find $\delta > 0$ such that for any tagged partition $P = \{R_\alpha\}, T = (x_\alpha)$ with $\|P\| < \delta$ we have

$$|R(f, P, T) - I| < \epsilon.$$

Let $S = (y_\alpha)$ be another sequence of tags for P. Then we have

$$\Big|\sum \big(f(x_\alpha) - f(y_\alpha)\big)|R_\alpha|\Big| = |R(f, P, T) - R(f, P, S)| < 2\epsilon.$$

Consider some fixed k. Let $J := \{\alpha : R_\alpha^\circ \cap D_k \neq \emptyset\}$, where R_α° is the interior of R_α. Note that $\{R_\alpha^\circ\}_{\alpha\in J}$ is a finite family of open rectangles that covers D_k, except for its subset $D_k \cap \bigcup \partial R_\alpha$. But $\bigcup \partial R_\alpha$ is the union of finitely many sets of measure zero, so it has measure zero. Hence it can be covered by a countable family of open rectangles whose total volume is as small as we want. Thus, in order to show that D_k has measure zero, we only need to show that the total volume of $\{R_\alpha^\circ\}_{\alpha\in J}$ is small. Now for $\alpha \in J$ we can choose $x_\alpha, y_\alpha \in R_\alpha$ such that

$$f(x_\alpha) - f(y_\alpha) \geq \frac{1}{2k}.$$

The reason is that there is $z \in D_k \cap R_\alpha^\circ$, and since $\operatorname{osc}_z f \geq \frac{1}{k}$, we can find points x, y near z inside its open neighborhood R_α° such that $|f(x) - f(y)|$ is as close to $\frac{1}{k}$

as we want. Then for $\alpha \notin J$ we choose $x_\alpha = y_\alpha \in R_\alpha$, so that $f(x_\alpha) - f(y_\alpha) = 0$. Thus we have

$$\begin{aligned} \frac{1}{2k}\sum_{\alpha\in J}|R_\alpha^\circ| &\le \sum_{\alpha\in J}\big(f(x_\alpha) - f(y_\alpha)\big)|R_\alpha| \\ &= \sum_{\text{all } \alpha}\big(f(x_\alpha) - f(y_\alpha)\big)|R_\alpha| < 2\epsilon. \end{aligned}$$

Hence $\sum_{\alpha\in J}|R_\alpha^\circ| < 4k\epsilon$. Therefore D_k has measure zero as desired. ■

Proposition 8.18. *The closed and open rectangles in $\mathbb{R}^n$ do not have measure zero.*

Proof. Let

$$R = [a_1, b_1] \times \cdots \times [a_n, b_n]$$

be a closed rectangle. Then the characteristic function of $\mathbb{Q}^n$ restricted to R is a bounded function whose set of discontinuities is all of R, and it is not Riemann integrable. Therefore if the rectangle had measure zero we would have a contradiction with the Riemann-Lebesgue theorem. The result for open rectangles follows immediately, since every open rectangle contains a smaller closed rectangle. ■

8.2 Multiple Darboux Integrals

Definition 8.19. Suppose $R \subset \mathbb{R}^n$ is a closed rectangle, and $f : R \to \mathbb{R}$ is a bounded function. Let $P = \{R_\alpha\}$ be a partition of R. The **lower sum** and **upper sum** of f with respect to the partition P are respectively

$$L(f,P) := \sum_\alpha m_\alpha|R_\alpha|, \qquad U(f,P) := \sum_\alpha M_\alpha|R_\alpha|,$$

where

$$m_\alpha = \inf\{f(x) : x \in R_\alpha\}, \qquad M_\alpha = \sup\{f(x) : x \in R_\alpha\}.$$

Remark. It is obvious that for any choice of tags T we have

$$L(f,P) \le R(f,P,T) \le U(f,P).$$

Definition 8.20. Suppose $R \subset \mathbb{R}^n$ is a closed rectangle, and $f : R \to \mathbb{R}$ is a bounded function. The **lower integral** and **upper integral** of f are respectively

$$\underline{\int_R} f(x)dx := \sup_P L(f,P), \qquad \overline{\int_R} f(x)dx := \inf_P U(f,P).$$

Here, P ranges over all partitions of R. We say f is **Darboux integrable** (on R) if

$$\underline{\int_R} f(x)dx = \overline{\int_R} f(x)dx,$$

and in this case we denote this common value by $\int_R f(x)dx$ and call it the **Darboux integral** of f (over R).

Proposition 8.21. *Suppose P is a partition of the closed rectangle $R \subset \mathbb{R}^n$, and Q is a refinement of P. Then for any bounded function $f : R \to \mathbb{R}$ we have*

$$L(f,P) \le L(f,Q) \le U(f,Q) \le U(f,P).$$

Proof. Suppose $P = P_1 \times \cdots \times P_n$, and for some k we have $P_k = \{a_0, \ldots, a_m\}$. It suffices to prove the claim for a refinement $Q = Q_1 \times \cdots \times Q_n$, where $Q_i = P_i$ for $i \ne k$, and $Q_k = P_k \cup \{b\}$. The general case then follows by an easy induction. Suppose $a_{j-1} < b < a_j$. Then the subrectangles of P and Q are the same except for those subrectangles of P whose kth factor is $[a_{j-1}, a_j]$. These subrectangles split into two subrectangles of Q with the kth factor $[a_{j-1}, b]$ and $[b, a_j]$. Let $\{R_\alpha\}_{\alpha \in I}$ be the set of all such subrectangles of P. For $\alpha \in I$ set

$$\begin{aligned} m'_\alpha &:= \inf\{f(x) : x \in R_\alpha,\ a_{j-1} \le x_k \le b\}, \\ m''_\alpha &:= \inf\{f(x) : x \in R_\alpha,\ b \le x_k \le a_j\}. \end{aligned}$$

Then we have $m'_\alpha, m''_\alpha \ge m_\alpha = \inf\{f(x) : x \in R_\alpha\}$. Note that for each $\alpha \in I$, the volume of the two subrectangles of Q that are obtained from R_α are

$$|R_\alpha| \frac{b - a_{j-1}}{a_j - a_{j-1}}, \qquad |R_\alpha| \frac{a_j - b}{a_j - a_{j-1}}.$$

Hence we have

$$\begin{aligned} L(f,Q) - L(f,P) &= \sum_{\alpha \in I} \Big(m'_\alpha |R_\alpha| \frac{b - a_{j-1}}{a_j - a_{j-1}} + m''_\alpha |R_\alpha| \frac{a_j - b}{a_j - a_{j-1}} - m_\alpha |R_\alpha| \Big) \\ &\ge \sum_{\alpha \in I} \Big(m_\alpha |R_\alpha| \frac{b - a_{j-1}}{a_j - a_{j-1}} + m_\alpha |R_\alpha| \frac{a_j - b}{a_j - a_{j-1}} - m_\alpha |R_\alpha| \Big) = 0. \end{aligned}$$

The case of upper sums is similar. ■

Remark. Suppose P, P^* are two partitions of R; and Q is their common refinement. Then the above proposition implies that

$$L(f,P) \le L(f,Q) \le U(f,Q) \le U(f,P^*).$$

Thus any lower sum is less than or equal to any upper sum. As a result for any bounded function f we have

$$\underline{\int_R} f(x)dx \leq \overline{\int_R} f(x)dx.$$

***Exercise* 8.22.** Suppose $R \subset \mathbb{R}^n$ is a closed rectangle, and $f, g : R \to \mathbb{R}$ are bounded functions such that $f \leq g$. Show that for any partition P of R we have

$$L(f,P) \leq L(g,P), \qquad U(f,P) \leq U(g,P).$$

Also show that

$$\underline{\int_R} f(x)dx \leq \underline{\int_R} g(x)dx, \qquad \overline{\int_R} f(x)dx \leq \overline{\int_R} g(x)dx.$$

Theorem 8.23. *Let $R \subset \mathbb{R}^n$ be a closed rectangle. A bounded function $f : R \to \mathbb{R}$ is Darboux integrable if and only if for all $\epsilon > 0$ there exists a partition P of R such that*

$$U(f,P) - L(f,P) < \epsilon.$$

Proof. Suppose f is Darboux integrable. So we have $\overline{\int_R} f(x)dx = \underline{\int_R} f(x)dx$. Let $\epsilon > 0$ be given. Since the upper integral is the infimum of the upper sums and the lower integral is the supremum of the lower sums, there are partitions P, P^* such that

$$U(f,P) - \overline{\int_R} f(x)dx < \frac{\epsilon}{2}, \qquad \underline{\int_R} f(x)dx - L(f,P^*) < \frac{\epsilon}{2}.$$

Therefore we have $U(f,P) - L(f,P^*) < \epsilon$. Now let Q be the common refinement of P, P^*. Then we have

$$U(f,Q) - L(f,Q) \leq U(f,P) - L(f,P^*) < \epsilon,$$

because refining a partition causes the upper sum to decrease and the lower sum to increase.

Next suppose f satisfies the specified property in the theorem. Then for all $\epsilon > 0$ we have $0 \leq \overline{\int_R} f(x)dx - \underline{\int_R} f(x)dx < \epsilon$. Thus $\overline{\int_R} f(x)dx = \underline{\int_R} f(x)dx$, and f is Darboux integrable. ■

Theorem 8.24. *Let $R \subset \mathbb{R}^n$ be a closed rectangle. A function $f : R \to \mathbb{R}$ is Riemann integrable if and only if it is Darboux integrable. In this case, the Riemann integral of f is the same as its Darboux integral.*

Proof. First suppose f is Riemann integrable and I is its Riemann integral. Then f is bounded. Suppose $\epsilon > 0$ is given. There is $\delta > 0$ such that if P is a partition of R with $\|P\| < \delta$ then $|R(f, P, T) - I| < \frac{\epsilon}{4}$ for any choice of tags T. Let $P = \{R_\alpha\}$ be such a partition, and let m_α, M_α be respectively the infimum and supremum of f over R_α. Then we can choose tags $T' = (x_\alpha)$ such that $0 \le f(x_\alpha) - m_\alpha < \frac{\epsilon}{4|R|}$. Then we have

$$0 \le R(f, P, T') - L(f, P) = \sum (f(x_\alpha) - m_\alpha)|R_\alpha| < \frac{\epsilon}{4|R|} \sum |R_\alpha| = \frac{\epsilon}{4}.$$

Similarly we can choose tags T'' so that

$$0 \le U(f, P) - R(f, P, T'') < \frac{\epsilon}{4}.$$

Therefore we have

$$\begin{aligned} U(f, P) - L(f, P) = U(f, P) - R(f, P, T'') + R(f, P, T'') - I \\ + I - R(f, P, T') + R(f, P, T') - L(f, P) < \epsilon. \end{aligned}$$

Hence f is Darboux integrable. Finally note that we also have $|U(f, P) - I| < \frac{\epsilon}{2}$. Therefore $\left|\overline{\int_R} f(x)dx - I\right| \le \frac{\epsilon}{2}$. Thus the Darboux integral of f is the same as its Riemann integral, since ϵ is arbitrary.

Next suppose f is Darboux integrable. Then f is bounded. Suppose $\epsilon > 0$ is given. Then there is a partition P such that $U(f, P) - L(f, P) < \epsilon$. Since any Riemann sum is between the upper sum and the lower sum, for any choices of tags T, S for P we have

$$|R(f, P, T) - R(f, P, S)| < \epsilon.$$

Now we can repeat the argument given at the end of the proof of Riemann-Lebesgue theorem to conclude that the set of discontinuities of f has measure zero. Hence f is Riemann integrable. Therefore the Riemann integral of f is the same as its Darboux integral as we proved in the last paragraph. ■

***Exercise* 8.25.** Prove that a Darboux integrable function is Riemann integrable, without using the Riemann-Lebesgue theorem.

Theorem 8.26. *Let $R \subset \mathbb{R}^n$ be a closed rectangle. A bounded function $f : R \to \mathbb{R}$ is Riemann integrable if and only if for all $\epsilon > 0$ there exists a partition P of R such that*

$$U(f, P) - L(f, P) < \epsilon.$$

Proof. This is a consequence of the previous two theorems. ■

8.3 Integration over General Domains

Definition 8.27. A set $S \subset \mathbb{R}^n$ is called **Jordan measurable** if it is bounded, and its boundary, ∂S, has measure zero.

Proposition 8.28. *A bounded set $S \subset \mathbb{R}^n$ is Jordan measurable if and only if its characteristic function*

$$\chi_S(x) := \begin{cases} 1 & x \in S \\ 0 & x \notin S \end{cases}$$

is Riemann integrable over some closed rectangle R containing S.

Proof. First we show that the set of discontinuities of χ_S is ∂S. To see this note that if a point x belongs to the interior of S or the interior of S^c, then χ_S is constant 1 or 0 on a neighborhood of x, hence χ_S is continuous at x. Otherwise, every open neighborhood of x must intersect both S and S^c, which means $x \in \partial S$. Then x is the limit of some sequence in S and some sequence in S^c, since $\partial S = \overline{S} \cap \overline{S^c}$. Therefore χ_S cannot be continuous at x.

Now as χ_S is bounded, we conclude that χ_S is Riemann integrable if and only if its set of discontinuities, i.e. ∂S, has measure zero. ■

Definition 8.29. When $S \subset \mathbb{R}^n$ is Jordan measurable we set

$$|S| := \int_R \chi_S(x)dx,$$

where R is a closed rectangle containing S. We call $|S|$ the **(Jordan) content** or the **volume** of S. When $n = 1, 2$, the volume is called the **length** or the **area**, respectively.

Definition 8.30. Suppose $S \subset \mathbb{R}^n$ is Jordan measurable, and $f : S \to \mathbb{R}$. We say f is Riemann integrable over S if for some closed rectangle R containing S the function

$$F(x) := \begin{cases} f(x) & x \in S \\ 0 & x \in R - S \end{cases}$$

is Riemann integrable over R. If this happens, we define the Riemann integral of f over S to be

$$\int_S f(x)dx := \int_R F(x)dx.$$

Remark. It must be checked that the above two definitions do not depend on the choice of the rectangle R. First note that by the generalized Riemann-Lebesgue theorem proved below (whose proof does not rely on the independence of these notions from R), the integrability of f over S is completely determined by the behavior of f over S. Hence the integrability of f does not depend on R. We should mention that the fact that ∂S has measure zero is needed in the proof of the generalized Riemann-Lebesgue theorem.

Next, to show that the value of the integral is independent of R, it suffices to consider two rectangles $R_1 \subset R_2$. Because for any two arbitrary rectangles we can choose a larger rectangle containing both of them, and then we can compare the integral over each rectangle to the integral over the larger rectangle. Let F_i be the extension of f by zero to R_i. Consider a sequence of partitions P_j of R_2 whose meshes approach zero, and they all contain the vertices of R_1. Let T_j be a choice of tags for P_j that has no intersection with ∂R_1. Then they induce tagged partitions of R_1, which we continue to denote them by P_j, T_j. Now we have

$$R(F_1, P_j, T_j) = R(F_2, P_j, T_j),$$

since F_2 is zero at the points of T_j that are outside R_1. As $j \to \infty$ the Riemann sums converge to the corresponding integrals, hence we get the desired equality of the integrals.

Finally note that the volume of a Jordan measurable set S is independent of the rectangle R containing S, since the value of the integral $\int_R \chi_S(x)dx$ is independent of R.

Remark. Another issue that we must be careful about, when we generalize a notion, is that the new definition is compatible with the old one. Here we have to check that the new definitions of integrability, integral, and volume, are the same as the old definitions, when S is a closed rectangle. First note that rectangles are obviously Jordan measurable, so the old notions are special cases of the new notions. Now we can take the R in the new definitions to be the same as S, as we have showed that we can change R freely. Then we have retrieved the old definitions of integrability and integral. For the volume of the closed rectangle S we have

$$|S|_{\text{new}} = \int_S \chi_S(x)dx = \int_S 1\, dx = |S|_{\text{old}}.$$

Finally, when S is an open rectangle, its volume has been defined to be the same as the volume of its closure $\overline{S}$, which is a closed rectangle. Let R_i be an increasing sequence of closed rectangles inside S that converge to S. Then we have

$$\begin{aligned}|\overline{S}| = \int_{\overline{S}} \chi_{\overline{S}}(x)dx \geq \int_{\overline{S}} \chi_S(x)dx = |S|_{\text{new}}& \\ \geq \int_{\overline{S}} \chi_{R_i}(x)dx = |R_i| \to |\overline{S}|.&\end{aligned}$$

Note that here we have used the continuity of the volume of closed rectangles, which is evident from its old definition.

***Example* 8.31.** Let $A = \mathbb{Q} \cap [0,1] \subset \mathbb{R}$. Then A is not Jordan measurable, since

$$\partial A = \bar{A} - A^\circ = [0,1] - \emptyset = [0,1]$$

does not have measure zero. On the other hand, the Cantor set $C \subset \mathbb{R}$ is Jordan measurable, since

$$\partial C = \overline{C} - C^\circ = C - \emptyset = C$$

has measure zero.

Remark. There are open subsets of $\mathbb{R}^n$ that are not Jordan measurable, i.e. their boundaries have positive measure.

Remark. Let S be a Jordan measurable subset of $\mathbb{R}^n$, and let R be a closed rectangle containing S. Let $P = \{R_\alpha\}$ be a partition of R. We know that the Jordan content of S is given by

$$|S| = \int_R \chi_S(x)dx.$$

The lower and upper sums for the above integral with respect to the partition P are respectively

$$L(\chi_S, P) = \sum_\alpha m_\alpha |R_\alpha|, \qquad U(\chi_S, P) = \sum_\alpha M_\alpha |R_\alpha|,$$

where

$$m_\alpha = \inf\{\chi_S(x) : x \in R_\alpha\}, \qquad M_\alpha = \sup\{\chi_S(x) : x \in R_\alpha\}.$$

If $R_\alpha \subset S$ then χ_S is 1 over it; so $m_\alpha = M_\alpha = 1$. And if $R_\alpha \cap S = \emptyset$ then $m_\alpha = M_\alpha = 0$. But if R_α intersects both S and $R - S$, we have $m_\alpha = 0$ and $M_\alpha = 1$. Hence the above lower and upper sums can be written as

$$L(\chi_S, P) = \sum_{R_\alpha \subset S} |R_\alpha|, \qquad U(\chi_S, P) = \sum_{R_\alpha \cap S \neq \emptyset} |R_\alpha|.$$

Since an integral is between its lower and upper sums we have

$$\sum_{R_\alpha \subset S} |R_\alpha| \le |S| \le \sum_{R_\alpha \cap S \neq \emptyset} |R_\alpha|.$$

We also know that the lower and upper sums converge to $|S|$ as $\|P\| \to 0$ (see the proof of Theorem 8.24). Therefore to compute the volume of S we can either approximate it from below with the volume of finitely many disjoint rectangles inside S, or approximate it from above with the volume of finitely many rectangles covering S. Also note that for the integral to exist, or equivalently, for S to be Jordan measurable, the difference between upper and lower sums must go to zero as the partition P becomes finer. In other words, for S to be Jordan measurable, the difference between the approximations of its volume from outside and inside must go to zero.

Theorem 8.32. *Let S be a Jordan measurable subset of $\mathbb{R}^n$, and let $f, g : S \to \mathbb{R}$ be Riemann integrable. Then for $c, c_1, c_2 \in \mathbb{R}$ we have*

(i) *The constant function c is Riemann integrable over S and $\int_S c\,dx = c|S|$.*

(ii) *$c_1 f + c_2 g$ is Riemann integrable over S and*

$$\int_S [c_1 f(x) + c_2 g(x)]dx = c_1 \int_S f(x)dx + c_2 \int_S g(x)dx.$$

(iii) *If $f \leq g$ then*

$$\int_S f(x)dx \leq \int_S g(x)dx.$$

(iv) *If $|f| \leq M$ then*

$$\left|\int_S f(x)dx\right| \leq M|S|.$$

Remark. As a result, if $|S| = 0$ then $\int_S f(x)dx = 0$.

Proof. Suppose R is a closed rectangle that contains S. Let F, G be the extensions of f, g by zero to R, respectively. Then by the assumption we know that F, G are integrable over R.

(i) The integrability of c over S is equivalent to the integrability of $c\chi_S$ over R, which is equivalent to Jordan measurability of S. Then we have

$$\int_S c\,dx = \int_R c\,\chi_S\,dx = c\int_R \chi_S\,dx = c|S|.$$

(ii) The integrability of $c_1 f + c_2 g$ over S is equivalent to the integrability of $c_1 F + c_2 G$ over R, since $c_1 F + c_2 G$ is the extension of $c_1 f + c_2 g$ by zero to R. Then we have

$$\begin{aligned}\int_S [c_1 f(x) + c_2 g(x)]dx &= \int_R [c_1 F(x) + c_2 G(x)]dx \\ &= c_1 \int_R F(x)dx + c_2 \int_R G(x)dx \\ &= c_1 \int_S f(x)dx + c_2 \int_S g(x)dx.\end{aligned}$$

(iii) It is obvious that $F \leq G$ on R. Thus

$$\int_S f(x)dx = \int_R F(x)dx \leq \int_R G(x)dx = \int_S g(x)dx.$$

(iv) We have $-M \leq f \leq M$. Now the result follows from parts (i) and (iii). ■

Riemann-Lebesgue Theorem (General Domains). *Let S be a Jordan measurable subset of $\mathbb{R}^n$, and let $f : S \to \mathbb{R}$. Then f is Riemann integrable if and only if f is bounded on S, and its set of discontinuities in the interior of S has measure zero.*

Remark. Since $S \subset \overline{S} = S^\circ \cup \partial S$, and ∂S has measure zero, the set of discontinuities of f in S° has measure zero if and only if the set of discontinuities of f in S has measure zero.

Proof. Suppose R is a closed rectangle that contains S. Let F be the extension of f by zero to R. Then by definition, f is integrable over S if and only if F is integrable over R. Since F is zero outside S, the boundedness of F on R is equivalent to the boundedness of f on S. Let D be the set of discontinuities of f in the interior of S, and let Z be the set of discontinuities of F in R.

First suppose D has measure zero, and f is bounded. Then F is bounded. Let $x \in Z$. Then x cannot belong to $R - \overline{S}$, since $R - \overline{S}$ is an open set in R, and F is zero (and hence continuous) over it. Thus either $x \in S^\circ$, or $x \in \partial S$. If $x \in S^\circ$ then $F = f$ on a neighborhood of x, hence f is also discontinuous at x. Therefore we have $Z \subset D \cup \partial S$. Thus Z has measure zero too, since S is Jordan measurable. Hence F is integrable, and by definition f is integrable too.

Now suppose f is integrable. Then F is integrable. Therefore F is bounded, and Z has measure zero. Thus f is bounded too. Let $y \in D$. Then $y \in S^\circ$, and $F = f$ on a neighborhood of y; hence F is also discontinuous at y. Therefore we have $D \subset Z$. Thus D has measure zero as desired. ■

Remark. In the remark after the Theorem 8.4 we stated that if $R, R_1, \dots, R_k$ are closed rectangles in $\mathbb{R}^n$, and $R \subset \bigcup R_i$, then

$$|R| \le \sum |R_i|.$$

An interesting proof of this fact, using the Riemann integral, is as follows. First note that $\chi_R \le \sum \chi_{R_i}$. Because if $\chi_R(x) = 1$ then $x \in R$. Thus $x \in R_i$ for some i. Hence $\sum \chi_{R_i}(x) \ge 1$. Now let $\hat{R}$ be a closed rectangle containing R and R_i's. Then χ_R and χ_{R_i}'s are integrable over $\hat{R}$, since rectangles are Jordan measurable. Hence we have

$$|R| = \int_{\hat{R}} \chi_R(x)dx \le \sum \int_{\hat{R}} \chi_{R_i}(x)dx = \sum |R_i|.$$

Theorem 8.33. *Continuous functions on a compact Jordan measurable set are Riemann integrable.*

Proof. A continuous function on a compact set is bounded, and its set of discontinuities is empty. ■

Theorem 8.34. *Suppose $S \subset \mathbb{R}^n$ is Jordan measurable, and $f, g : S \to \mathbb{R}$ are Riemann integrable. Then their product, fg, is Riemann integrable.*

Proof. First note that fg is bounded on S, since f, g are bounded. Let $Z(f), Z(g)$ be the set of discontinuities of f, g, respectively. Then $Z(f)$ and $Z(g)$ have measure zero. But for $Z(fg)$, the set of discontinuities of fg, we have

$$Z(fg) \subset Z(f) \cup Z(g),$$

since fg is continuous at the points where both f, g are continuous. Thus $Z(fg)$ has measure zero too. Hence fg is Riemann integrable. ■

Theorem 8.35. *Suppose $S \subset \mathbb{R}^n$ is Jordan measurable, and $f : S \to \mathbb{R}$ is Riemann integrable. Also suppose that $f(S) \subset [a, b]$, and $\phi : [a, b] \to \mathbb{R}$ is continuous. Then $\phi \circ f$ is Riemann integrable.*

Proof. First note that ϕ is bounded, since it is continuous. Hence $\phi \circ f$ is bounded. Let $Z(f), Z(\phi \circ f)$ be the set of discontinuities of $f, \phi \circ f$, respectively. Then $Z(f)$ has measure zero. As ϕ is continuous we have $Z(\phi \circ f) \subset Z(f)$, since $\phi \circ f$ is continuous at the points where f is continuous. Thus $Z(\phi \circ f)$ has measure zero too. Therefore $\phi \circ f$ is Riemann integrable. ■

Theorem 8.36. *Suppose $S \subset \mathbb{R}^n$ is Jordan measurable, and $f : S \to \mathbb{R}^m$ is Riemann integrable. Then $|f|$ is Riemann integrable and we have*

$$\left| \int_S f(x)dx \right| \le \int_S |f(x)|dx.$$

Proof. Suppose $f = (f_1, \dots, f_m)$. Then each f_i is integrable. Thus $|f| = \sqrt{f_1^2 + \cdots + f_m^2}$ is integrable, since the functions $t \mapsto t^2, \sqrt{t}$ are continuous. When $m = 1$, the inequality for the integrals follows from the monotonicity of the integral and $-|f| \le f \le |f|$. For the case of $m > 1$ let $z := \int_S f(x)dx$. Then for each i we have $z_i = \int_S f_i(x)dx$. Now we have

$$\begin{aligned} |z|^2 = \sum_{i \le m} z_i^2 = \sum_{i \le m} z_i \int_S f_i(x)dx = \int_S \sum_{i \le m} z_i f_i(x)dx \\ = \int_S z \cdot f(x)dx \le \int_S |z||f(x)|\,dx = |z| \int_S |f(x)|\,dx. \end{aligned}$$

Therefore if $|z| \ne 0$ we get $\left| \int_S f(x)dx \right| = |z| \le \int_S |f(x)|\,dx$. And if $|z| = 0$ we have $\int_S |f(x)|\,dx \ge \int_S 0\,dx = 0 = |z|$. ■

Remark. The above theorem is one of the few results in this chapter in which the case of $m > 1$ does not follow trivially form the case of $m = 1$. The reason is that the dependence of $|f|$ on $|f_i|$'s is not linear.

Theorem 8.37. *Suppose $S \subset \mathbb{R}^n$ is Jordan measurable, and $f, g : S \to \mathbb{R}$ are Riemann integrable functions such that $f = g$ a.e. on S. Then we have*

$$\int_S f(x)dx = \int_S g(x)dx.$$

$\boxed{\text{Proof.}}$ Let R be a closed rectangle containing S, and let F, G be the extensions of f, g by zero to R. Then we have $F = G$ a.e. on R, since F, G are zero on $R - S$. Consider a sequence of partitions $P_j = \{R_{j,\alpha}\}$ of R whose meshes approach zero. Let $T_j = (x_{j,\alpha})$ be a choice of tags for P_j such that for any tag $x_{j,\alpha}$ in T_j we have $F(x_{j,\alpha}) = G(x_{j,\alpha})$. Note that this is possible since $F = G$ outside a set of measure zero, and the subrectangles of P_j do not have measure zero. So each subrectangle $R_{j,\alpha}$ contains a point at which F, G are equal. Therefore we have

$$R(F, P_j, T_j) = \sum_\alpha F(x_{j,\alpha})|R_{j,\alpha}| = \sum_\alpha G(x_{j,\alpha})|R_{j,\alpha}| = R(G, P_j, T_j).$$

Then as $j \to \infty$ the Riemann sums converge to the corresponding integrals, hence we get

$$\int_S f(x)dx = \int_R F(x)dx = \int_R G(x)dx = \int_S g(x)dx. \qquad \blacksquare$$

Remark. The assumption of Riemann integrability of both f, g is critical in the above theorem, since for example the characteristic function of $\mathbb{Q}^n$ is a.e. zero but it is not Riemann integrable.

Theorem 8.38. *Suppose $S \subset \mathbb{R}^n$ is Jordan measurable, and $f : S \to \mathbb{R}$ is a Riemann integrable function such that $\int_S f(x)dx = 0$ and $f \geq 0$. Then $f = 0$ a.e. on S.*

$\boxed{\text{Proof.}}$ Suppose to the contrary that the set $C = \{x \in S : f(x) \neq 0\}$ does not have measure zero. We claim that f is continuous at some point of $C \cap S^\circ$. Otherwise we would have $C \subset Z \cup \partial S$, where Z is the set of discontinuities of f. But both $Z, \partial S$ have measure zero, so this implies that C has measure zero, contrary to our assumption. Now let $a \in C \cap S^\circ$ be a point at which f is continuous. Suppose for example that $f(a) > 0$. Then by continuity, $f \geq \epsilon > 0$ on a rectangle $R \subset S^\circ$ that has a in its interior. We also have $f \geq f\chi_R$, since $f \geq 0$. Hence

$$\int_S f(x)dx \geq \int_S f(x)\chi_R \, dx = \int_R f(x)dx \geq \epsilon|R| > 0,$$

which is a contradiction. So C must have measure zero as desired. Note that the equality in the above formula follows from extending $f\chi_R$ by zero to a closed rectangle containing S, and observing that the extended function is also the extension of $f|_R$. $\blacksquare$

Theorem 8.39. *Suppose $S \subset \mathbb{R}^n$ is Jordan measurable. Then S has measure zero if and only if $|S| = 0$.*

Proof. Let R be a closed rectangle containing S. Note that R is Jordan measurable, and χ_S is Riemann integrable over R. Now if S has measure zero, then $\chi_S = 0$ a.e. on R. Hence we have

$$|S| = \int_R \chi_S \, dx = \int_R 0 \, dx = 0.$$

Conversely suppose $|S| = 0$. Then $\int_R \chi_S \, dx = 0$. But $\chi_S \geq 0$, so we must have $\chi_S = 0$ a.e. on R. Therefore S has measure zero, since χ_S is nonzero on S. ■

Remark. A set of measure zero is not necessarily Jordan measurable. For example $\mathbb{Q} \cap [0, 1]$ has measure zero in $\mathbb{R}$, but it is not Jordan measurable. Because its boundary is all of $[0, 1]$, which does not have measure zero.

Theorem 8.40. *Suppose $S_1, S_2 \subset \mathbb{R}^n$ are Jordan measurable. Then $S_1 \cup S_2$, $S_1 \cap S_2$, and $S_2 - S_1$ are Jordan measurable too. We also have*

$$|S_1 \cup S_2| = |S_1| + |S_2| - |S_1 \cap S_2|.$$

In addition, if $S_1 \subset S_2$ then we have $|S_1| \leq |S_2|$, and

$$|S_2 - S_1| = |S_2| - |S_1|.$$

Remark. As a result we have

$$|S_1 \cup S_2| \leq |S_1| + |S_2|.$$

By an easy induction, we can show that the union and the intersection of finitely many Jordan measurable sets are Jordan measurable, and the above inequality also holds for more than two Jordan measurable sets.

Proof. Let R be a closed rectangle containing both S_1, S_2. Then R will also contain $S_1 \cup S_2$, $S_1 \cap S_2$, and $S_2 - S_1$. So these sets are bounded. It is easy to see that $\chi_{S_1 \cap S_2} = \chi_{S_1} \chi_{S_2}$. Since χ_{S_1}, χ_{S_2} are Riemann integrable over R, $\chi_{S_1 \cap S_2}$ is also Riemann integrable over R. Thus $S_1 \cap S_2$ is Jordan measurable. Similarly we can easily show that

$$\chi_{S_1 \cup S_2} = \chi_{S_1} + \chi_{S_2} - \chi_{S_1 \cap S_2}, \qquad \chi_{S_2 - S_1} = \chi_{S_2} - \chi_{S_1 \cap S_2}.$$

Hence $S_1 \cup S_2$ and $S_2 - S_1$ are Jordan measurable too, since their characteristic functions are Riemann integrable. Now if we integrate the equation involving $\chi_{S_1 \cup S_2}$

we get

$$\begin{aligned}|S_1 \cup S_2| &= \int_R \chi_{S_1 \cup S_2} dx \\ &= \int_R \chi_{S_1} dx + \int_R \chi_{S_2} dx - \int_R \chi_{S_1 \cap S_2} dx \\ &= |S_1| + |S_2| - |S_1 \cap S_2|.\end{aligned}$$

To prove the final statement in the theorem, note that if $S_1 \subset S_2$ then we have

$$S_1 \cup (S_2 - S_1) = S_2, \qquad S_1 \cap (S_2 - S_1) = \emptyset.$$

So we get $|S_2| = |S_1| + |S_2 - S_1| - |\emptyset| = |S_1| + |S_2 - S_1|$. Hence we obtain

$$|S_2| - |S_1| = |S_2 - S_1| \geq 0,$$

as desired. ■

***Exercise* 8.41.** For two Jordan measurable sets $S_1, S_2 \subset \mathbb{R}^n$ prove that $S_1 \cup S_2$, $S_1 \cap S_2$, and $S_1 - S_2$ are Jordan measurable, by showing directly that their boundaries have measure zero.

***Exercise* 8.42.** Suppose S is Jordan measurable. Show that $\overline{S}$, S°, and ∂S are also Jordan measurable.

Theorem 8.43. *Suppose $S_1, S_2 \subset \mathbb{R}^n$ are Jordan measurable sets that have disjoint interiors, i.e. $S_1^\circ \cap S_2^\circ = \emptyset$. Then a function $f : S_1 \cup S_2 \to \mathbb{R}$ is Riemann integrable if and only if $f\big|_{S_1}, f\big|_{S_2}$ are Riemann integrable. Also, in this case we have*

$$\int_{S_1 \cup S_2} f(x)dx = \int_{S_1} f(x)dx + \int_{S_2} f(x)dx.$$

Remark. As we will see in the following proof, the above additivity property of integral with respect to the domain of integration follows from the additivity of integral with respect to the integrand.

Remark. By an easy induction, this theorem can be generalized to the case of more than two Jordan measurable sets $S_1, \dots, S_k$ with pairwise disjoint interiors. (Note that under this assumption, S_k° cannot intersect

$$(S_1 \cup \cdots \cup S_{k-1})^\circ \subset S_1 \cup \cdots \cup S_{k-1} \subset (S_1^\circ \cup \cdots \cup S_{k-1}^\circ) \cup (\partial S_1 \cup \cdots \cup \partial S_{k-1}).$$

Since otherwise S_k° would have to intersect ∂S_j for some j, which implies that S_k° would intersect S_j°.) In particular, if we set $f = 1$ we obtain

$$|S_1 \cup \cdots \cup S_k| = |S_1| + \cdots + |S_k|,$$

when $S_1, \dots, S_k$ are Jordan measurable sets with pairwise disjoint interiors.

Proof. First note that $S_1 \cup S_2$ is Jordan measurable. Also note that f is bounded on $S_1 \cup S_2$ if and only if it is bounded on both S_1, S_2. Now let Z_i be the set of discontinuities of $f\big|_{S_i}$ in S_i, and let Z be the set of discontinuities of f in $S_1 \cup S_2$. Then

$$Z_i \subset \partial S_i \cup Z.$$

Because if $x \in Z_i - \partial S_i$ then $x \in S_i^\circ$, and we have $f\big|_{S_i} = f$ on an open neighborhood of x. So $x \in Z$. Now when f is integrable, Z has measure zero. Then Z_i has measure zero too, since S_i is Jordan measurable and its boundary has measure zero. Therefore $f\big|_{S_i}$ is integrable. Similarly, we can show that

$$Z \subset Z_1 \cup Z_2 \cup \partial S_1 \cup \partial S_2.$$

Thus when $f\big|_{S_1}, f\big|_{S_2}$ are integrable, Z_1, Z_2 have measure zero. Then Z has measure zero too, since S_1, S_2 are Jordan measurable and their boundaries have measure zero.

Next let R be a closed rectangle containing $S_1 \cup S_2$, and let F be the extension of f by zero to R. Then the extension of $f\big|_{S_i}$ by zero to R equals $F\chi_{S_i}$. It is easy to see that

$$\chi_{S_1\cup S_2} = \chi_{S_1} + \chi_{S_2} - \chi_{S_1\cap S_2}.$$

Note that $S_1 \cap S_2$ is Jordan measurable. In addition, by the theorem's hypothesis we have $S_1 \cap S_2 \subset \partial S_1 \cup \partial S_2$. Because if $x \in S_1 \cap S_2 - \partial S_1$ then $x \in S_1^\circ$. Thus $x \notin S_2^\circ$, and we must have $x \in \partial S_2$. Therefore $S_1 \cap S_2$ has measure zero. Hence $\chi_{S_1\cap S_2} = 0$ a.e. Now we have

$$\begin{aligned}\int_{S_1\cup S_2} f(x)dx &= \int_R F(x)dx = \int_R F(x)\chi_{S_1\cup S_2}\,dx \\ &= \int_R F(x)\chi_{S_1}\,dx + \int_R F(x)\chi_{S_2}\,dx - \int_R F(x)\chi_{S_1\cap S_2}\,dx \\ &= \int_{S_1} f(x)dx + \int_{S_2} f(x)dx,\end{aligned}$$

as desired. Note that all the functions in the above formula are Riemann integrable. Also we have used the fact that $F\chi_{S_1\cap S_2} = 0$ a.e., so its integral is zero. ■

Remark. Note that if $f \geq 0$ then $\int_S f(x)dx \geq \int_S 0\,dx = 0$. Now if $S_1 \subset S_2$, by the above theorem we get

$$\int_{S_1} f(x)dx \leq \int_{S_1} f(x)dx + \int_{S_2 - S_1} f(x)dx = \int_{S_2} f(x)dx.$$

So, the integral of a nonnegative integrable function is nonnegative, and it increases as we enlarge the domain of integration.

8.4 Iterated Integrals

Fubini's Theorem. *Suppose $R_1 \subset \mathbb{R}^n$, $R_2 \subset \mathbb{R}^m$ are closed rectangles, and $f : R_1 \times R_2 \to \mathbb{R}$ is Riemann integrable. We denote the elements of $R_1 \times R_2$ by (x, y) where $x \in R_1$, $y \in R_2$. For each $y \in R_2$ let $f_y : R_1 \to \mathbb{R}$ be defined by $f_y(x) := f(x, y)$. Then the two functions*

$$\underline{F}(y) := \underline{\int_{R_1}} f_y(x)dx = \underline{\int_{R_1}} f(x, y)dx, \qquad \overline{F}(y) := \overline{\int_{R_1}} f_y(x)dx = \overline{\int_{R_1}} f(x, y)dx,$$

are Riemann integrable over R_2, and we have

$$\int_{R_1 \times R_2} f(x, y)dxdy = \int_{R_2} \underline{F}(y)dy = \int_{R_2} \overline{F}(y)dy.$$

Proof. Let $P = \{R_\alpha\}$ and $Q = \{S_\beta\}$ be partitions of R_1, R_2 respectively. Then $P \times Q = \{R_\alpha \times S_\beta\}$ is a partition of $R_1 \times R_2$. For any $y_0 \in S_\beta$ we have

$$m_{\alpha\beta} := \inf\{f(x, y) : (x, y) \in R_\alpha \times S_\beta\} \le \inf\{f(x, y_0) : x \in R_\alpha\} =: m_\alpha(y_0).$$

Hence by the definition of lower integral we have

$$\sum_\alpha m_{\alpha\beta}|R_\alpha| \le \sum_\alpha m_\alpha(y_0)|R_\alpha| = L(f_{y_0}, P) \le \underline{\int_{R_1}} f(x, y_0)dx = \underline{F}(y_0).$$

Since y_0 is arbitrary we get $\sum_\alpha m_{\alpha\beta}|R_\alpha| \le \inf_{S_\beta} \underline{F}$. Therefore

$$\begin{aligned} L(f, P \times Q) = \sum_{\alpha,\beta} m_{\alpha\beta}|R_\alpha \times S_\beta| &= \sum_\beta \Big(\sum_\alpha m_{\alpha\beta}|R_\alpha|\Big)|S_\beta| \\ &\le \sum_\beta \Big(\inf_{S_\beta} \underline{F}\Big)|S_\beta| = L(\underline{F}, Q) \le \underline{\int_{R_2}} \underline{F}(y)dy. \end{aligned}$$

If we take the supremum over all partitions $P \times Q$ we obtain

$$\int_{R_1 \times R_2} f(x, y)dxdy \le \underline{\int_{R_2}} \underline{F}(y)dy. \qquad (*)$$

Note that by definition every partition of $R_1 \times R_2$ is of the form $P \times Q$, so the supremum is indeed over all partitions of $R_1 \times R_2$. Similarly we can show that

$$\overline{\int_{R_2}} \overline{F}(y)dy \le U(f, P \times Q).$$

But $\underline{F} \le \overline{F}$, so

$$\overline{\int_{R_2}} \underline{F}(y)dy \le \overline{\int_{R_2}} \overline{F}(y)dy \le U(f, P \times Q).$$

Now if we take the infimum over all partitions $P \times Q$, and combine the result with inequality $(*)$, we obtain

$$\int_{R_1 \times R_2} f(x,y)dxdy \le \underline{\int_{R_2}} \underline{F}(y)dy \le \overline{\int_{R_2}} \underline{F}(y)dy \le \int_{R_1 \times R_2} f(x,y)dxdy.$$

Therefore $\underline{F}$ is integrable over R_2, and its integral equals the integral of f over $R_1 \times R_2$. The same is true for $\overline{F}$ by a similar reasoning. ■

Remark. Since $\int_{R_2} \underline{F}(y)dy = \int_{R_2} \overline{F}(y)dy$, and $\underline{F} \le \overline{F}$, we must have $\underline{F}(y) = \overline{F}(y)$ for a.e. $y \in R_2$. In other words $f(x,y)$ is an integrable function of x over R_1 for a.e. fixed $y \in R_2$.

Remark. When f is continuous, $f(x,y)$ is an integrable function of x over R_1 for every fixed $y \in R_2$. Hence

$$\underline{F}(y) = \overline{F}(y) = \int_{R_1} f(x,y)dx,$$

for all $y \in R_2$. We can also switch the role of x, y, and apply the Fubini's theorem twice to obtain

$$\int_{R_1 \times R_2} f(x,y)dxdy = \int_{R_2} \Big(\int_{R_1} f(x,y)dx \Big) dy = \int_{R_1} \Big(\int_{R_2} f(x,y)dy \Big) dx.$$

The two integrals on the right hand side of the above formula are called **iterated integrals**. Note that as a result of Fubini's theorem, we can change the order of integration in iterated integrals, when f is continuous on the rectangle $R_1 \times R_2$.

Theorem 8.44. *Suppose $S \subset \mathbb{R}^{n-1}$ is a compact set, and $\psi : S \to \mathbb{R}$ is a continuous function. Then the graph of ψ i.e.*

$$G := \{(x, \psi(x)) \in \mathbb{R}^n : x \in S\}$$

has measure zero in $\mathbb{R}^n$.

Proof. Let R be a closed rectangle containing S (note that S is bounded). We know that ψ is uniformly continuous, since S is compact. Now for a given $\epsilon > 0$ there is $\delta > 0$ such that for $x, y \in S$ with $|x - y| < \delta$ we have $|\psi(y) - \psi(x)| < \epsilon$. Let P be a partition of R whose mesh is less than δ. Then whenever $x, y \in S$ belong to the same subrectangle of P, we have $|\psi(y) - \psi(x)| < \epsilon$. Let $R_1, \dots, R_m$ be subrectangles of P that intersect S, and let x_i be a point in $R_i \cap S$. Let R_i' be an open rectangle containing R_i whose volume is twice the volume of R_i. Then we have

$$G \subset \bigcup_{i\leq m} R_i \times \big(\psi(x_i)-\epsilon, \psi(x_i)+\epsilon\big) \subset \bigcup_{i\leq m} R'_i \times \big(\psi(x_i)-\epsilon, \psi(x_i)+\epsilon\big).$$

The total volume of this open covering of G is

$$\sum_{i\leq m} 2\epsilon|R'_i| = 4\epsilon \sum_{i\leq m} |R_i| \leq 4\epsilon|R|.$$

Now as ϵ is arbitrary we can conclude that G has measure zero. ■

Remark. We can also show that the graph of a continuous function over an open set has measure zero. First let V be a bounded open set in $\mathbb{R}^{n-1}$. Then V is the union of countably many compact sets $K_j = V - \bigcup_{x\in\partial V} B_{1/j}(x)$. To see this first note that $V = \bigcup_j K_j$, since the distance of a point $z \in V$ from the points on the compact set ∂V has a positive minimum, and thus z must belong to some K_j. On the other hand, K_j is closed, because for the open set $B = \bigcup_{x\in\partial V} B_{1/j}(x)$ we have

$$\overline{V} \cap B^c = (V \cup \partial V) \cap B^c = (V \cap B^c) \cup (\partial V \cap B^c) = V \cap B^c = K_j,$$

since $\partial V \subset B$. So K_j is compact, being a subset of the bounded set V. Now note that any open set U in $\mathbb{R}^{n-1}$ is the union of countably many bounded open sets $V_i := U \cap B_i(0)$. Thus U is also the union of countably many compact sets K_{ij}, since the union of countably many countable families is countable. Therefore the graph of a continuous function ψ over U is the union of countably many graphs of continuous functions $\psi|_{K_{ij}}$, which have measure zero; so the graph of ψ also has measure zero.

Theorem 8.45. *Suppose $S \subset \mathbb{R}^{n-1}$ is a compact Jordan measurable set. Let $\phi, \psi : S \to \mathbb{R}$ be continuous functions such that $\phi \leq \psi$. Then*

$$C := \{(x, y) \in \mathbb{R}^n : x \in S,\ \phi(x) \leq y \leq \psi(x)\}$$

is a compact Jordan measurable set, and for any continuous function $f : C \to \mathbb{R}$ we have

$$\int_C f(x, y)dxdy = \int_S \Big(\int_{\phi(x)}^{\psi(x)} f(x, y)dy\Big)dx.$$

Remark. For simplicity of the notation, we assumed that y is the nth component of the point $(x, y) \in \mathbb{R}^n$; but similar results hold when y is any other component.

Proof. Suppose $R \subset \mathbb{R}^{n-1}$ is a closed rectangle containing S. Since ϕ, ψ are continuous on a compact set, there are $m, M \in \mathbb{R}$ such that $m < \phi \le \psi < M$. Then $R \times [m, M]$ is a closed rectangle containing C, so C is bounded. Let F be the extension of f by zero to this rectangle. Suppose for now that C is a compact Jordan measurable set; then f is Riemann integrable, and by Fubini's theorem we have

$$\begin{aligned}\int_C f(x,y)dxdy &= \int_{R\times[m,M]} F(x,y)dxdy \\ &= \int_R \Big(\overline{\int_m^M} F(x,y)dy \Big) dx = \int_R \Big(\int_m^M F(x,y)dy \Big) dx. \qquad (*)\end{aligned}$$

Note that for every fixed $x \in R - S$, $F(x,y) = 0$; and for every fixed $x \in S$, $F(x,y)$ is at most discontinuous at two points $y = \phi(x)$ and $y = \psi(x)$. Thus $F(x,y)$ is a Riemann integrable function of y for every fixed x. Also note that $\int_m^M F(x,y)dy$ is a Riemann integrable function of x by Fubini's theorem. Now as $R - S$ is Jordan measurable, and does not intersect S, we have

$$\int_R \int_m^M F dydx = \int_{R-S} \int_m^M F dydx + \int_S \int_m^M F dydx = \int_S \int_m^M F dydx, \qquad (**)$$

since $\int_m^M F(x,y)dy = 0$ for every fixed $x \in R - S$. In addition for every fixed $x \in S$ we have

$$\begin{aligned}\int_m^M F(x,y)dy = \int_m^{\phi(x)} F(x,y)dy + \int_{\phi(x)}^{\psi(x)} F(x,y)dy + \int_{\psi(x)}^M F(x,y)dy \\ = \int_{\phi(x)}^{\psi(x)} f(x,y)dy,\end{aligned}$$

since $F(x,y) = 0$ a.e. on $[m, \phi(x)]$ and a.e. on $[\psi(x), M]$. Now we can integrate this relation over S, and use $(*), (**)$, to get the desired equation. Note that the last term of the above formula is a Riemann integrable function of x, since it equals the first term which is Riemann integrable.

Therefore it only remains to show that C is compact, and ∂C has measure zero. Note that C is the image of the continuous map $\Phi : S \times [0,1] \to \mathbb{R}^n$ defined by

$$\Phi(x,t) = \big(x, \phi(x) + t(\psi(x) - \phi(x))\big).$$

Therefore C is compact, since $S \times [0,1]$ is compact. In particular C is closed, and we have $\partial C \subset C$. Now if $a \in S^\circ$ and $\phi(a) < b < \psi(a)$, then $(a,b) \in C^\circ$. To see this note that there is $\epsilon > 0$ such that $\phi(a) + 2\epsilon < b < \psi(a) - 2\epsilon$. Then there is $\delta > 0$ such that $B_\delta(a) \subset S$, and for $x \in B_\delta(a)$ we have $\phi(x) + \epsilon < b < \psi(x) - \epsilon$ due to continuity. Thus $B_\delta(a)\times(b-\epsilon,b+\epsilon)$ is an open subset of C containing (a,b), and therefore $(a,b)\in C^\circ$. Hence we must have

$$\begin{aligned}\partial C \subset \{(x,y) : x\in\partial S,\ \phi(x)\le y\le\psi(x)\}&\\ \cup\{(x,\phi(x)) : x\in S\}\cup\{(x,\psi(x)) : x\in S\}&\\ \subset(\partial S\times[m,M])\cup\{(x,\phi(x)) : x\in S\}\cup\{(x,\psi(x)) : x\in S\}&.\end{aligned}$$

Therefore to show that ∂C has measure zero it suffices to show that $\partial S\times[m,M]$ has measure zero, because the graphs of continuous functions have measure zero. But ∂S has measure zero in $\mathbb{R}^{n-1}$. Thus for any $\epsilon>0$ there is a family of open rectangles $R_i\subset\mathbb{R}^{n-1}$ such that $\partial S\subset\bigcup_{i\ge1}R_i$, and $\sum_{i\ge1}|R_i|<\epsilon$. Then we have

$$\partial S\times[m,M]\subset\bigcup_{i\ge1}R_i\times(m-1,M+1),$$

and

$$\sum_{i\ge1}|R_i\times(m-1,M+1)|<\epsilon(M-m+2).$$

Hence $\partial S\times[m,M]$ has measure zero as desired. ■

Remark. In practice, when we want to compute a multiple integral, we reduce the dimension either by employing the Fubini's theorem when the domain is rectangular, or by employing the above theorem when the domain is more general. Then, after repeating this process several times, we finally arrive at one-dimensional integrals, and we can compute them by the fundamental theorem of calculus. Note that sometimes we have to break our domain into several smaller parts in order to be able to apply the above theorem to each part.

***Example* 8.46.** As an example consider the three-dimensional region

$$C=\{(x,y,z)\in\mathbb{R}^3 : a\le x\le b,\ g(x)\le y\le h(x),\ \phi(x,y)\le z\le\psi(x,y)\}.$$

Let $S=\{(x,y)\in\mathbb{R}^2 : a\le x\le b,\ g(x)\le y\le h(x)\}$. Suppose g,h,ϕ,ψ are continuous functions satisfying $g\le h$ and $\phi\le\psi$. Then by the above theorem S is a compact Jordan measurable set. We also have

$$C=\{(\tilde{x},z)\in\mathbb{R}^3 : \tilde{x}=(x,y)\in S,\ \phi(\tilde{x})\le z\le\psi(\tilde{x})\}.$$

Hence, by repeatedly applying the above theorem, we obtain

$$\begin{aligned}\int_C f(x,y,z)dxdydz=\int_C f(\tilde{x},z)d\tilde{x}dz&=\int_S\Big(\int_{\phi(\tilde{x})}^{\psi(\tilde{x})}f(\tilde{x},z)dz\Big)d\tilde{x}\\ &=\int_S\Big(\int_{\phi(x,y)}^{\psi(x,y)}f(x,y,z)dz\Big)dxdy=\int_S\tilde{f}(x,y)dxdy\end{aligned}$$

$$= \int_{[a,b]} \Big[\int_{g(x)}^{h(x)} \tilde{f}(x,y)dy \Big] dx$$
$$= \int_a^b \Big[\int_{g(x)}^{h(x)} \Big(\int_{\phi(x,y)}^{\psi(x,y)} f(x,y,z)dz \Big) dy \Big] dx,$$

which is a familiar formula from calculus. Note that $\tilde{f}(x,y) := \int_{\phi(x,y)}^{\psi(x,y)} f(x,y,z)dz$ is a continuous function by Exercise 7.27. We can similarly compute integrals over higher-dimensional regions of the same type.

Theorem 8.47. *Suppose $S \subset \mathbb{R}^{n-1}$ is a compact Jordan measurable set, and $\psi : S \to \mathbb{R}$ is a nonnegative continuous function. Then the set under the graph of ψ i.e.*

$$C := \{(x,y) \in \mathbb{R}^n : x \in S,\ 0 \le y \le \psi(x)\}$$

is Jordan measurable, and

$$|C| = \int_S \psi(x)dx.$$

In other words, the volume of the set under the graph of ψ equals the integral of ψ.

Proof. This is a particular case of the previous theorem. Let $f = 1$, and $\phi = 0$ in that theorem. Then we have

$$|C| = \int_C 1\, dxdy = \int_S \int_0^{\psi(x)} 1\, dydx = \int_S \psi(x)dx. \qquad ■$$

8.5 Change of Variables

Definition 8.48. Suppose $A \subset \mathbb{R}^n$. A function $F : A \to \mathbb{R}^m$ is called **locally Lipschitz** if for every $a \in A$ there are $K, r > 0$ such that

$$|F(x) - F(y)| \le K|x - y|$$

for all $x, y \in A \cap B_r(a)$, where $B_r(a)$ is the open ball of radius r around a.

Remark. In other words, a function F is locally Lipschitz if every point in its domain has a neighborhood on which F is Lipschitz.

Remark. It is easy to see that locally Lipschitz functions are continuous. It is also obvious that a Lipschitz function is locally Lipschitz; but the converse is not true. For example the function $x \mapsto x^2$ is a locally Lipschitz function from $\mathbb{R}$ to $\mathbb{R}$ which is not Lipschitz.

***Example* 8.49.** Suppose $U \subset \mathbb{R}^n$ is open, and $f : U \to \mathbb{R}^m$ is differentiable. If the partial derivatives of f are locally bounded (i.e. every point in U has a neighborhood over which the partial derivatives are bounded), then f is locally Lipschitz. In particular, if f is C^1 then it is locally Lipschitz, because continuous partial derivatives are bounded on closed balls inside the domain. To prove the claim let $a \in U$, and suppose the partial derivatives of f are bounded on $B_r(a) \subset U$. Let

$$M_a = \sup_{x \in B_r(a)} \Big(\sum_{i,j} |D_j f_i(x)|^2\Big)^{\frac{1}{2}}.$$

Then by Theorem 7.24, for every $x, y \in B_r(a)$ we have

$$|f(x) - f(y)| \le M_a |x - y|,$$

since $B_r(a)$ contains the line segment joining x, y.

***Exercise* 8.50.** Show that the composition of locally Lipschitz functions is locally Lipschitz.

Theorem 8.51. *Suppose $Z \subset \mathbb{R}^n$ has measure zero, and $F : Z \to \mathbb{R}^n$ is locally Lipschitz. Then $F(Z)$ has measure zero.*

Proof. Every $a \in Z$ has a neighborhood $B_r(a)$ such that F is Lipschitz on $Z \cap B_r(a)$. Then the family $\{B_r(a) : a \in Z\}$ is an open covering of Z. By Theorem 11.57 every open covering of a subset of $\mathbb{R}^n$ has a countable subcovering. Let us denote this countable subcovering by $\{B_i\}$. Then we have

$$F(Z) = \bigcup_{i \ge 1} F(Z \cap B_i) = \bigcup_{i \ge 1} F|_{Z \cap B_i}(Z \cap B_i).$$

But $F|_{Z \cap B_i}$ is Lipschitz, and $Z \cap B_i$ has measure zero. Therefore it suffices to prove the theorem for Lipschitz maps. Because then it follows that each $F|_{Z \cap B_i}(Z \cap B_i)$ has measure zero. And as $F(Z)$ is the union of countably many sets of measure zero, it also has measure zero as desired.

So we assume that F is Lipschitz, and Z has measure zero. We want to show that $F(Z)$ has measure zero. For any $\epsilon > 0$ there is a family of open cubes $Q_i \subset \mathbb{R}^n$ such that $Z \subset \bigcup_{i \ge 1} Q_i$, and $\sum_{i \ge 1} |Q_i| < \epsilon$. Suppose the length of the edges of Q_i is l_i. It is easy to see that the diameter of Q_i, i.e. the maximum distance between points of Q_i, is $l_i \sqrt{n}$. But the diameter of $Z \cap Q_i$ is less than or equal to the diameter of Q_i. Hence the diameter of $F(Z \cap Q_i)$ is at most $K l_i \sqrt{n}$, since the distance of two points $F(x), F(y)$ is less than or equal to K times the distance of x, y. Therefore $F(Z \cap Q_i) \subset R_i$, where R_i is an open cube whose edges are of length $3K l_i \sqrt{n}$, and is centered at some point $z \in F(Z \cap Q_i)$. Because for any other point

$z' \in F(Z \cap Q_i)$ we have $|z' - z| \le K l_i \sqrt{n}$; so the absolute value of each coordinate of $z' - z$ is less than or equal to $K l_i \sqrt{n}$, which is strictly less than $\frac{3}{2} K l_i \sqrt{n}$. Now we have

$$F(Z) = \bigcup_{i \ge 1} F(Z \cap Q_i) \subset \bigcup_{i \ge 1} R_i,$$

and

$$\sum_{i \ge 1} |R_i| = 3^n K^n \sqrt{n^n} \sum_{i \ge 1} l_i^n = 3^n K^n \sqrt{n^n} \sum_{i \ge 1} |Q_i| < 3^n K^n \sqrt{n^n} \epsilon.$$

Thus as ϵ is arbitrary, $F(Z)$ has measure zero. ■

Remark. Note that in the above proof we cannot use rectangles instead of cubes. Because the diameter of a rectangle cannot be estimated by the volume of the rectangle. In other words, a rectangle with small volume can have a very large diameter.

Remark. If we merely assume that F is continuous, then $F(Z)$ does not necessarily have measure zero when Z has measure zero. For example the Cantor function (Example 5.24) is a continuous function that maps the standard Cantor set, which has measure zero, onto $[0, 1]$.

Theorem 8.52. *Suppose $S_1, S_2 \subset \mathbb{R}^n$ are Jordan measurable, and $f : S_1 \to \mathbb{R}$ is Riemann integrable. Let $\psi : S_2 \to S_1$ be a homeomorphism such that $\psi^{-1} : S_1 \to S_2$ is locally Lipschitz. Then $f \circ \psi$ is Riemann integrable.*

Proof. First note that $f \circ \psi$ is bounded, since f is bounded. Thus we only need to show that $Z(f \circ \psi)$, the set of discontinuities of $f \circ \psi$, has measure zero. Let $Z(f)$ be the set of discontinuities of f. Then we have

$$Z(f \circ \psi) \subset \psi^{-1}(Z(f)),$$

since $f \circ \psi$ is continuous at a point x if f is continuous at $\psi(x)$. (Actually, the above two sets are equal due to the continuity of ψ^{-1}, but we do not use this fact). Now, $Z(f)$ has measure zero, and ψ^{-1} is locally Lipschitz. Thus $\psi^{-1}(Z(f))$ has measure zero. Therefore $Z(f \circ \psi)$ has measure zero too. ■

Theorem 8.53. *Suppose $S \subset \mathbb{R}^n$ is bounded, and $U \subset \mathbb{R}^n$ is an open set containing $\overline{S}$. Let $\phi : U \to \mathbb{R}^n$ be a one-to-one C^1 map such that $D\phi(x)$ is an invertible matrix for every $x \in U$. Then we have*

$$\partial(\phi(S)) = \phi(\partial S).$$

As a result, if S is Jordan measurable, $\phi(S)$ is Jordan measurable too.

Proof. First note that $\phi(U)$ is open, and $\phi^{-1} : \phi(U) \to U$ is C^1. To see this let $b \in \phi(U)$. Then there is $a \in U$ such that $b = \phi(a)$. Now by the inverse function theorem a has an open neighborhood V such that $\phi(V)$ is open, and the inverse of $\phi|_V$, which is $\phi^{-1}|_{\phi(V)}$, is C^1. Hence $\phi(V) \subset \phi(U)$ is an open neighborhood of b, and the claim follows. Thus, in particular, ϕ^{-1} and ϕ have the same properties.

Now let us show that $\partial(\phi(S)) = \phi(\partial S)$. Consider S°, the interior of S. Then similarly to the above we can show that $\phi(S^\circ)$ is an open set, since S° is an open set. In addition, we have

$$\phi(S^\circ) \subset \phi(S) \subset \phi(\overline{S}) = \phi(S^\circ \cup \partial S) = \phi(S^\circ) \cup \phi(\partial S).$$

Thus $\phi(S^\circ) \subset (\phi(S))^\circ$. On the other hand note that $\phi(\overline{S})$ is closed, since $\overline{S}$ is compact. So $\overline{\phi(S)} \subset \phi(\overline{S})$. Hence

$$\partial(\phi(S)) = \overline{\phi(S)} - (\phi(S))^\circ \subset \phi(\overline{S}) - \phi(S^\circ) \subset \phi(\partial S).$$

Similarly for ϕ^{-1} we have $\partial(\phi^{-1}(C)) \subset \phi^{-1}(\partial C)$, where $C \subset \phi(U)$. Thus we get

$$\partial S = \partial\big(\phi^{-1}(\phi(S))\big) \subset \phi^{-1}\big(\partial(\phi(S))\big).$$

Hence we obtain $\phi(\partial S) \subset \partial(\phi(S))$. Therefore the two sets are equal.

Finally, using the equality $\partial(\phi(S)) = \phi(\partial S)$, we can conclude that if S is Jordan measurable, then $\partial(\phi(S))$ has measure zero too; because ϕ is locally Lipschitz (since it is C^1), and ∂S has measure zero. In addition, $\overline{S}$ is compact, since it is bounded as S is bounded. Thus $\phi(\overline{S})$ is compact, and therefore it is also bounded. So $\phi(S) \subset \phi(\overline{S})$ is bounded too. Hence $\phi(S)$ is Jordan measurable. ■

Remember that any invertible matrix is the product of several *elementary matrices*, which have one of the following forms

$$\begin{array}{r} \\ i\text{-th row} \rightarrow \\ \vdots \\ j\text{-th row} \rightarrow \\ \end{array}
\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix},
\quad
\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & a & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix},
\quad
\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & c & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix},$$

where $a, c \in \mathbb{R}$, and $a \neq 0$. So, the first matrix is obtained by interchanging the ith row and the jth row of the identity matrix; the second matrix is obtained by multiplying the ith row of the identity matrix by the nonzero constant a; and the third matrix is obtained by adding c times the jth row of the identity matrix to its ith row.

Lemma 8.54. *Suppose $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear map given by an elementary matrix, and $R \subset \mathbb{R}^n$ is a closed rectangle. Then $T(R)$ is Jordan measurable, and we have*

$$|T(R)| = |\det T|\,|R|.$$

Proof. Remember that a linear map T between Euclidean spaces is given by the action of a unique matrix, which we still denote by T. Let $R = \prod_{k\leq n}[a_k, b_k]$. Suppose T is an elementary matrix of the first kind. Then we have

$$T(x_1, \dots, x_i, \dots, x_j, \dots, x_n) = (x_1, \dots, x_j, \dots, x_i, \dots, x_n).$$

Hence $T(R) = [a_1, b_1] \times \cdots \times [a_j, b_j] \times \cdots \times [a_i, b_i] \times \cdots \times [a_n, b_n]$. Therefore

$$|T(R)| = \prod_{k\leq n}(b_k - a_k) = |R| = |-1|\,|R| = |\det T|\,|R|.$$

Next suppose T is an elementary matrix of the second kind. Then we have

$$T(x_1, \dots, x_i, \dots, x_n) = (x_1, \dots, ax_i, \dots, x_n).$$

Therefore $T(R) = [a_1, b_1] \times \cdots \times [aa_i, ab_i] \times \cdots \times [a_n, b_n]$ when $a > 0$, and $T(R) = [a_1, b_1] \times \cdots \times [ab_i, aa_i] \times \cdots \times [a_n, b_n]$ when $a < 0$. Thus

$$|T(R)| = |ab_i - aa_i| \prod_{k\neq i}(b_k - a_k) = |a| \prod_{k\leq n}(b_k - a_k) = |a|\,|R| = |\det T|\,|R|.$$

Note that in both cases, $T(R)$ is a closed rectangle; hence it is Jordan measurable.

Finally suppose that T is an elementary matrix of the third kind. Then we have

$$T(x_1, \dots, x_i, \dots, x_n) = (x_1, \dots, x_i + cx_j, \dots, x_n) =: (y_1, \dots, y_n).$$

Note that for $k \neq i$ we have $y_k = x_k \in [a_k, b_k]$, and for each fixed $y_j = x_j$ we have $y_i - cy_j = x_i + cx_j - cx_j = x_i \in [a_i, b_i]$. Hence

$$T(R) = \{y \in \mathbb{R}^n : y_k \in [a_k, b_k] \text{ for } k \neq i, \text{ and } y_i \in [a_i + cy_j, b_i + cy_j]\}.$$

Therefore by Theorem 8.45, $T(R)$ is Jordan measurable, and we have

$$\begin{aligned}
|T(R)| &= \int_{T(R)} 1\, dy_1 \dots dy_{i-1} dy_i dy_{i+1} \dots dy_n \\
&= \int_{\prod_{k\neq i}[a_k,b_k]} \Big(\int_{a_i+cy_j}^{b_i+cy_j} 1\, dy_i \Big) dy_1 \dots dy_{i-1} dy_{i+1} \dots dy_n \\
&= \int_{\prod_{k\neq i}[a_k,b_k]} (b_i - a_i)\, dy_1 \dots dy_{i-1} dy_{i+1} \dots dy_n \\
&= (b_i - a_i) \prod_{k\neq i}(b_k - a_k) = \prod_{k\leq n}(b_k - a_k) = 1\,|R| = |\det T|\,|R|,
\end{aligned}$$

as desired. ■

Theorem 8.55. *Suppose $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear map, and $S \subset \mathbb{R}^n$ is Jordan measurable. Then $T(S)$ is also Jordan measurable, and we have*

$$|T(S)| = |\det T|\,|S|.$$

Remark. This theorem provides a geometric interpretation for the (absolute value of) determinant of a linear map, as the factor by which the linear map changes the volume.

Remark. As a result, if $\det T = \pm 1$ then T preserves the volume. In particular, if T is an orthogonal linear map (i.e. it preserves the distance, which is equivalent to $T^{\intercal}T = I$), then T also preserves the volume.

Proof. If T is not invertible then its image is inside an $(n-1)$-dimensional plane Γ, which has measure zero in $\mathbb{R}^n$. Thus $T(S)$ has measure zero too. Note that $T(S)$ is Jordan measurable, since its boundary is also a subset of Γ, because Γ is closed. Hence we have $|T(S)| = 0$. Therefore

$$0 = |T(S)| = 0\,|S| = |\det T|\,|S|,$$

as desired. So, in the rest of the proof, we can assume that T is invertible. Then by Theorem 8.53, it follows that $T(S)$ is Jordan measurable, since T is one-to-one and C^1, and its derivative, which is T itself, is invertible.

Next note that if $T = T_1T_2\cdots T_k$, and the theorem holds for each T_i, then the theorem also holds for T. Because we have

$$\begin{aligned}
|T(S)| = |T_1\cdots T_k(S)| &= |T_1(T_2\cdots T_k(S))| \\
&= |\det T_1|\,|T_2\cdots T_k(S)| \\
&= |\det T_1|\,|\det T_2|\,|T_3\cdots T_k(S)| \\
&= |\det T_1T_2|\,|T_3\cdots T_k(S)| \\
&\quad\vdots \\
&= |\det T_1T_2\cdots T_k|\,|S| = |\det T|\,|S|.
\end{aligned}$$

Now since any invertible matrix is the product of several elementary matrices, it suffices to prove the theorem for linear maps given by elementary matrices.

So, let T be a linear map given by an elementary matrix. Let R be a closed rectangle containing S, and let $P = \{R_\alpha\}$ be a partition of R. We know that the volume of S is given by $|S| = \int_R \chi_S(x)dx$, and as we have seen in Section 8.3, the lower and upper sums for this integral with respect to the partition P are

$$L(\chi_S, P) = \sum_{R_\alpha\subset S} |R_\alpha|, \qquad U(\chi_S, P) = \sum_{R_\alpha\cap S\neq\emptyset} |R_\alpha|,$$

respectively. For a given $\epsilon > 0$ we choose P so that

$$|S| - \epsilon < \sum_{R_\alpha \subset S} |R_\alpha| \leq |S| \leq \sum_{R_\alpha \cap S \neq \emptyset} |R_\alpha| < |S| + \epsilon. \tag{$*$}$$

Now note that $R_\alpha \subset S$ if and only if $T(R_\alpha) \subset T(S)$, and $R_\alpha \cap S \neq \emptyset$ if and only if $T(R_\alpha) \cap T(S) \neq \emptyset$, since T is an invertible function. Thus we have

$$\bigcup_{R_\alpha \subset S} T(R_\alpha) \subset T(S) \subset \bigcup_{R_\alpha \cap S \neq \emptyset} T(R_\alpha).$$

Therefore by Theorem 8.40 we have

$$\Big|\bigcup_{R_\alpha \subset S} T(R_\alpha)\Big| \leq |T(S)| \leq \Big|\bigcup_{R_\alpha \cap S \neq \emptyset} T(R_\alpha)\Big| \leq \sum_{R_\alpha \cap S \neq \emptyset} |T(R_\alpha)|.$$

However, note that two different subrectangles R_α, R_β can only intersect at their boundaries. Hence $T(R_\alpha), T(R_\beta)$ too can only intersect at their boundaries. Because if $x \in T(R_\alpha) \cap T(R_\beta)$ then $T^{-1}x \in R_\alpha \cap R_\beta$, and therefore $T^{-1}x \in \partial R_\alpha \cap \partial R_\beta$. On the other hand, by Theorem 8.53 we have $\partial(T(A)) = T(\partial A)$ for any set A. So

$$x \in T(\partial R_\alpha) \cap T(\partial R_\beta) = \partial(T(R_\alpha)) \cap \partial(T(R_\beta)),$$

as desired. Hence if we use $f = 1$ in Theorem 8.43 we get

$$\Big|\bigcup_{R_\alpha \subset S} T(R_\alpha)\Big| = \sum_{R_\alpha \subset S} |T(R_\alpha)|.$$

Therefore we have shown that

$$\sum_{R_\alpha \subset S} |T(R_\alpha)| \leq |T(S)| \leq \sum_{R_\alpha \cap S \neq \emptyset} |T(R_\alpha)|.$$

Finally note that $|T(R_\alpha)| = |\det T|\,|R_\alpha|$ by the previous lemma. Thus by $(*)$ we get

$$\begin{aligned}
|\det T|(|S| - \epsilon) &< |\det T| \sum_{R_\alpha \subset S} |R_\alpha| = \sum_{R_\alpha \subset S} |\det T||R_\alpha| \\
&= \sum_{R_\alpha \subset S} |T(R_\alpha)| \leq |T(S)| \leq \sum_{R_\alpha \cap S \neq \emptyset} |T(R_\alpha)| \\
&= \sum_{R_\alpha \cap S \neq \emptyset} |\det T||R_\alpha| = |\det T| \sum_{R_\alpha \cap S \neq \emptyset} |R_\alpha| < |\det T|(|S| + \epsilon).
\end{aligned}$$

Hence $\big||T(S)| - |\det T|\,|S|\big| < |\det T|\,\epsilon$; and since ϵ is arbitrary, we get the desired result. ■

The above theorem can also be interpreted as

$$\int_{T(S)} 1\, dy = |T(S)| = |\det T|\, |S| = \int_S |\det T|\, dx = \int_S 1 \circ T\, |\det T|\, dx,$$

which is the change of variables theorem for the constant function $f = 1$ and the linear change of variables $\phi = T$. In general, suppose $S \subset \mathbb{R}^n$ is Jordan measurable, and $U \subset \mathbb{R}^n$ is an open set containing $\overline{S}$. Let $\phi : U \to \mathbb{R}^n$ be a one-to-one C^1 map such that $D\phi(x)$ is an invertible matrix for every $x \in U$. We have seen that $\phi(S)$ is a Jordan measurable set. Let $f : \phi(S) \to \mathbb{R}$ be a Riemann integrable function. Then the function $(f \circ \phi)\, |\det D\phi|$ is also Riemann integrable, and the change of variables theorem says that

$$\int_{\phi(S)} f(y) dy = \int_S (f \circ \phi)(x)\, |\det D\phi(x)|\, dx.$$

Let us provide a heuristic argument for the validity of the above equality. For simplicity suppose S is a rectangle. Let $P = \{R_\alpha\}$, $\{x_\alpha\}$ be a tagged partition of S. Then $\{\phi(R_\alpha)\}$ is a partition of $\phi(S)$ into deformed rectangles (which are still Jordan measurable). And the points $y_\alpha = \phi(x_\alpha)$ are tags for this deformed partition. Consider the linear functions

$$\phi_\alpha(x) := \phi(x_\alpha) + D\phi(x_\alpha)(x - x_\alpha).$$

Then ϕ_α approximates ϕ over R_α, provided that $\|P\|$ is small enough. We denote this by writing $\phi_\alpha \approx \phi$. Note that by the above theorem we have $|\phi_\alpha(R_\alpha)| = |\det D\phi(x_\alpha)||R_\alpha|$, since translation by the constant vector $\phi(x_\alpha)$ does not change the volume. Now if we approximate integrals by (deformed) Riemann sums we get

$$\begin{aligned}\int_{\phi(S)} f(y) dy &\approx \sum f(y_\alpha)|\phi(R_\alpha)| \\ &\approx \sum f(y_\alpha)|\phi_\alpha(R_\alpha)| \\ &= \sum f(\phi(x_\alpha))\, |\det D\phi(x_\alpha)||R_\alpha| \\ &\approx \int_S (f \circ \phi)(x)\, |\det D\phi(x)|\, dx,\end{aligned}$$

as wanted.

Lemma 8.56. *Suppose $S \subset \mathbb{R}^n$ is Jordan measurable, and $U \subset \mathbb{R}^n$ is an open set containing $\overline{S}$. Let $\phi : U \to \mathbb{R}^n$ be a one-to-one C^1 map such that $D\phi(x)$ is an invertible matrix for every $x \in U$. Then $\phi(S)$ is also Jordan measurable, and its volume satisfies*

$$|\phi(S)| \le \int_S |\det D\phi(x)|\, dx.$$

Remark. In fact, the change of variables theorem will imply that $|\phi(S)|$ is equal to the above integral.

Proof. We have already shown in Theorem 8.53 that $\phi(S)$ is Jordan measurable. Also, as we have seen in the proof of Theorem 8.53, ϕ^{-1} is C^1 due to the inverse function theorem. Thus both ϕ, ϕ^{-1} are locally Lipschitz. We break the proof into several parts to make it more comprehensible, although the parts are intertwined.

(i) Let Q be a closed cube containing S, and let $P = \{Q_\alpha\}$ be a partition of Q into smaller cubes. (Note that in this proof we work with cubes, not rectangles.) We know that the volume of S is given by $|S| = \int_Q \chi_S(x)dx$, and as we have seen in Section 8.3, the upper sum for this integral with respect to the partition P satisfies

$$|S| \le U(\chi_S, P) = \sum_{Q_\alpha \cap S \neq \emptyset} |Q_\alpha|.$$

Now note that by Exercise 2.111, the distance between the points of ∂U and $\overline{S}$ has a positive lower bound, since they are disjoint sets, ∂U is closed, and $\overline{S}$ is compact. Therefore when $\|P\|$ is small enough, $Q_\alpha \cap S \neq \emptyset$ implies that $Q_\alpha \subset U$. Let us assume that this holds for the partition P.

Next note that by Theorem 8.53 each $\phi(Q_\alpha)$ is Jordan measurable. In addition, $Q_\alpha \cap S \neq \emptyset$ if and only if $\phi(Q_\alpha) \cap \phi(S) \neq \emptyset$, since ϕ is a one-to-one function. Thus we have

$$\phi(S) \subset \bigcup_{Q_\alpha \cap S \neq \emptyset} \phi(Q_\alpha).$$

Therefore by Theorem 8.40 we have

$$|\phi(S)| \le \Big| \bigcup_{Q_\alpha \cap S \neq \emptyset} \phi(Q_\alpha) \Big| \le \sum_{Q_\alpha \cap S \neq \emptyset} |\phi(Q_\alpha)|. \tag{$*$}$$

Thus it suffices to estimate $\phi(Q_\alpha)$ when Q_α is a closed cube with small diameter. Let x_α be the center point of the cube Q_α. We claim that for every small enough $r > 0$ there is $\delta > 0$ such that if $\|P\| < \delta$, then independently of α we have

$$|\phi(Q_\alpha)| \le (1 + r)^n |\det D\phi(x_\alpha)| |Q_\alpha|. \tag{$**$}$$

To prove $(**)$, it suffices to show that

$$\phi(Q_\alpha) \subset \phi_\alpha\big((1 + r)Q_\alpha\big), \tag{$***$}$$

where $\phi_\alpha(x) := \phi(x_\alpha) + D\phi(x_\alpha)(x - x_\alpha)$, and cQ_α is the closed cube with center x_α whose edges' length is c times the length of the edges of Q_α. It is easy to see that $|cQ_\alpha| = c^n |Q_\alpha|$. Then by Theorem 8.55, $\phi_\alpha\big((1 + r)Q_\alpha\big)$ is Jordan measurable, and we get

$$|\phi(Q_\alpha)| \le \big|\phi_\alpha\big((1 + r)Q_\alpha\big)\big| = (1 + r)^n |\det D\phi(x_\alpha)| |Q_\alpha|,$$

as desired.

(ii) Now let us prove $(***)$. First, for $z \in \mathbb{R}^n$ we define

$$|z|_\infty := \max_{i \le n} |z_i|.$$

Note that $|z|_\infty$ is the distance from z to 0 with respect to the metric defined by taking the maximum of the distances of the components. Thus in particular, the triangle inequality $|z+w|_\infty \le |z|_\infty + |w|_\infty$ holds. It is also easy to see that

$$|z|_\infty \le |z| \le \sqrt{n}\,|z|_\infty.$$

Let $x \in Q_\alpha$. Note that we have $|x - x_\alpha|_\infty \le s/2$, where s is the length of the edges of Q_α. In addition, note that the line segment joining x, x_α is inside Q_α. Hence by Theorem 7.25 we have

$$\begin{aligned}
&\phi(x) - \phi(x_\alpha) - D\phi(x_\alpha)(x - x_\alpha) \\
&\qquad = \int_0^1 D\phi\big(x_\alpha + t(x - x_\alpha)\big)(x - x_\alpha)\,dt - D\phi(x_\alpha)(x - x_\alpha) \\
&\qquad = \int_0^1 \Big(D\phi\big(x_\alpha + t(x - x_\alpha)\big) - D\phi(x_\alpha)\Big)(x - x_\alpha)\,dt.
\end{aligned}$$

Let $A_\alpha := D\phi(x_\alpha)$. Then we have

$$\begin{aligned}
&A_\alpha^{-1}\big(\phi(x) - \phi(x_\alpha)\big) - (x - x_\alpha) \\
&\qquad = \int_0^1 A_\alpha^{-1}\Big(D\phi\big(x_\alpha + t(x - x_\alpha)\big) - D\phi(x_\alpha)\Big)(x - x_\alpha)\,dt,
\end{aligned}$$

since A_α does not depend on t, and integration commutes with forming linear combinations. Next note that due to Cramer's rule, the entries of the inverse of an invertible matrix are rational functions (i.e. quotients of polynomials) of the entries of the matrix. Thus $(D\phi)^{-1}$ is continuous. Hence there exists $C > 0$ such that on the compact set $\bigcup_{Q_\alpha \cap S \neq \emptyset} Q_\alpha \subset U$ we have $\big(\sum_{i,j} |(D\phi)^{-1}_{ij}|^2\big)^{\frac{1}{2}} \le C$. Then it easily follows that for every z we have $|(D\phi)^{-1} z| \le C|z|$ (see the proof of Theorem 7.24). Hence we have $|A_\alpha^{-1} z| \le C|z|$, independently of α.

On the other hand, we have $x_\alpha(t) := x_\alpha + t(x - x_\alpha) \in Q_\alpha$. Therefore

$$|x_\alpha(t) - x_\alpha| \le \|P\| < \delta.$$

Also, since $D\phi$ is continuous on the compact set $\bigcup_{Q_\alpha \cap S \neq \emptyset} Q_\alpha \subset U$, it is uniformly continuous there. Thus we can choose δ small enough so that

$$\Big(\sum_{i,j} |D\phi_i(x_\alpha(t)) - D\phi_i(x_\alpha)|^2\Big)^{\frac{1}{2}} \le \epsilon,$$

independently of α. Then it follows that

$$\left|\Big(D\phi\big(x_\alpha(t)\big) - D\phi(x_\alpha)\Big)(x - x_\alpha)\right| \le \epsilon|x - x_\alpha| \le \epsilon\sqrt{n}\,|x - x_\alpha|_\infty.$$

Hence we can conclude that

$$\begin{aligned}
&\big|A_\alpha^{-1}\big(\phi(x) - \phi(x_\alpha)\big) - (x - x_\alpha)\big|_\infty \\
&\qquad \le \big|A_\alpha^{-1}\big(\phi(x) - \phi(x_\alpha)\big) - (x - x_\alpha)\big| \\
&\qquad = \left|\int_0^1 A_\alpha^{-1}\Big(D\phi\big(x_\alpha(t)\big) - D\phi(x_\alpha)\Big)(x - x_\alpha)\,dt\right| \\
&\qquad \le \int_0^1 \left|A_\alpha^{-1}\Big(D\phi\big(x_\alpha(t)\big) - D\phi(x_\alpha)\Big)(x - x_\alpha)\right| dt \\
&\qquad \le \int_0^1 C\left|\Big(D\phi\big(x_\alpha(t)\big) - D\phi(x_\alpha)\Big)(x - x_\alpha)\right| dt \\
&\qquad \le \int_0^1 C\epsilon\sqrt{n}\,|x - x_\alpha|_\infty\,dt = C\epsilon\sqrt{n}\,|x - x_\alpha|_\infty \le r|x - x_\alpha|_\infty
\end{aligned}$$

for $\epsilon \le r/(C\sqrt{n})$. Then we get

$$\big|A_\alpha^{-1}\big(\phi(x) - \phi(x_\alpha)\big)\big|_\infty - |x - x_\alpha|_\infty \le r|x - x_\alpha|_\infty,$$

since $|w|_\infty - |z|_\infty \le |w - z|_\infty$ due to the triangle inequality. Thus we have shown that for $x \in Q_\alpha$ we have

$$\big|A_\alpha^{-1}\big(\phi(x) - \phi(x_\alpha)\big)\big|_\infty \le (1 + r)|x - x_\alpha|_\infty,$$

where r does not depend on α. Finally, let $y = \phi(x) \in \phi(Q_\alpha)$. Then we have

$$\big|x_\alpha + A_\alpha^{-1}\big(y - \phi(x_\alpha)\big) - x_\alpha\big|_\infty \le (1 + r)|x - x_\alpha|_\infty \le (1 + r)s/2,$$

where s is the length of the edges of Q_α. So $\tilde{x} := x_\alpha + A_\alpha^{-1}\big(y - \phi(x_\alpha)\big) \in (1 + r)Q_\alpha$. Thus

$$\begin{aligned}
y &= A_\alpha\tilde{x} - A_\alpha x_\alpha + \phi(x_\alpha) \\
&\qquad\qquad = \phi(x_\alpha) + D\phi(x_\alpha)(\tilde{x} - x_\alpha) = \phi_\alpha(\tilde{x}) \in \phi_\alpha\big((1 + r)Q_\alpha\big).
\end{aligned}$$

Therefore we have shown that the inclusion $(*{*}*)$ holds.

(iii) Finally, if we use the estimate $(**)$ in the bound $(*)$ we get

$$|\phi(S)| \le \sum_{Q_\alpha \cap S \ne \emptyset} |\phi(Q_\alpha)| \le (1 + r)^n \sum_{Q_\alpha \cap S \ne \emptyset} |\det D\phi(x_\alpha)||Q_\alpha|.$$

Note that for a given $r > 0$ and partition P, the above bound holds provided that $\|P\| < \delta$ for some small enough δ. Now note that the rightmost term above looks like $(1+r)^n$ times a Riemann sum for the integral $\int_S |\det D\phi(x)|\,dx$, with respect to the tagged partition $P, (x_\alpha)$. However, in order for this to be true we must have $x_\alpha \in S \cap Q_\alpha$ (which might not be true); because we are actually integrating $\chi_S |\det D\phi|$ over the cube Q. But $|\det D\phi|$ is continuous on the compact set $\bigcup_{Q_\alpha \cap S \neq \emptyset} Q_\alpha \subset U$; so it is uniformly continuous there. Thus for a given $\tilde{\epsilon} > 0$ we can choose δ small enough so that for $x \in Q_\alpha$ we have

$$\big||\det D\phi(x)| - |\det D\phi(x_\alpha)|\big| \le \tilde{\epsilon},$$

independently of α. Let $\tilde{x}_\alpha \in S \cap Q_\alpha$. Then we have

$$\begin{aligned}
|\phi(S)| &\le (1+r)^n \sum_{Q_\alpha \cap S \neq \emptyset} |\det D\phi(x_\alpha)||Q_\alpha| \\
&\le (1+r)^n \sum_{Q_\alpha \cap S \neq \emptyset} \big(|\det D\phi(\tilde{x}_\alpha)| + \tilde{\epsilon}\big)|Q_\alpha| \\
&= (1+r)^n \sum_{Q_\alpha \cap S \neq \emptyset} |\det D\phi(\tilde{x}_\alpha)||Q_\alpha| + (1+r)^n \tilde{\epsilon} \sum_{Q_\alpha \cap S \neq \emptyset} |Q_\alpha| \\
&\le (1+r)^n R\big(\chi_S |\det D\phi|, P, (\tilde{x}_\alpha)\big) + (1+r)^n \tilde{\epsilon} |Q|,
\end{aligned}$$

where R is the Riemann sum for the integral $\int_S |\det D\phi(x)|\,dx$, with respect to the tagged partition $P, (\tilde{x}_\alpha)$. Hence as $\|P\|$ goes to zero, the Riemann sum must converge to the integral, since $|\det D\phi|$ is Riemann integrable and S is Jordan measurable. Therefore we get

$$|\phi(S)| \le (1+r)^n \int_S |\det D\phi(x)|\,dx + (1+r)^n \tilde{\epsilon} |Q|.$$

In addition, $r, \tilde{\epsilon}$ are arbitrary small positive numbers. So if we let $r, \tilde{\epsilon} \to 0$ we get the desired bound for $|\phi(S)|$. ■

In the following proof, we need to first consider the case of nonnegative functions, and then use that to deduce the result for general functions. To that end, for a real-valued function f we set

$$f^+ := \max\{f, 0\} = \begin{cases} f & \text{if } f \ge 0, \\ 0 & \text{if } f < 0, \end{cases} \qquad f^- := -\min\{f, 0\} = \begin{cases} 0 & \text{if } f \ge 0, \\ -f & \text{if } f < 0. \end{cases}$$

Note that both $f^+, f^- \ge 0$, and we have $f = f^+ - f^-$. In addition, note that if f is Riemann integrable, then both $f^\pm$ are also Riemann integrable. Because we have $f^+ = x^+ \circ f$ and $f^- = x^- \circ f$, where

$$x^+ := \max\{x, 0\} = \begin{cases} x & \text{if } x \ge 0, \\ 0 & \text{if } x < 0, \end{cases} \qquad x^- := -\min\{x, 0\} = \begin{cases} 0 & \text{if } x \ge 0, \\ -x & \text{if } x < 0, \end{cases}$$

and both $x^{\pm}$ are continuous functions. Other useful properties of $f^{\pm}$, which we are not going to employ now, are that $|f| = f^{+} + f^{-}$ and $f^{\pm} \le |f|$.

Change of Variables. *Suppose $S \subset \mathbb{R}^n$ is Jordan measurable, and $U \subset \mathbb{R}^n$ is an open set containing $\overline{S}$. Let $\phi : U \to \mathbb{R}^n$ be a one-to-one C^1 map such that $D\phi(x)$ is an invertible matrix for every $x \in U$. Then $\phi(S)$ is a Jordan measurable set. Moreover, for every Riemann integrable function $f : \phi(S) \to \mathbb{R}$ the function $(f \circ \phi)\,|\det D\phi|$ is Riemann integrable on S, and we have*

$$\int_{\phi(S)} f(y)dy = \int_S (f \circ \phi)(x)\,|\det D\phi(x)|\,dx.$$

Proof. We have already shown in Theorem 8.53 that $\phi(S)$ is Jordan measurable. Also, as we have seen in the proof of Theorem 8.53, due to the inverse function theorem, $\phi(U)$ is open, and ϕ^{-1} is C^1. Thus in particular, both ϕ, ϕ^{-1} are locally Lipschitz. Furthermore, $f \circ \phi$ is Riemann integrable on S, since $\phi : S \to \phi(S)$ is a homeomorphism, and ϕ^{-1} is locally Lipschitz. In addition, $|\det D\phi|$ is a continuous function on the compact set $\overline{S}$. Thus it is continuous and bounded on S, and therefore it is Riemann integrable on S. Hence $(f\circ\phi)\,|\det D\phi|$ is Riemann integrable on S.

First let us assume that $f \ge 0$. Let R be a closed rectangle containing $\phi(S)$, and let $P = \{R_\alpha\}$ be a partition of R. Then the lower sum for the integral $\int_{\phi(S)} f(y)dy$ with respect to the partition P is

$$L(\chi_{\phi(S)}f, P) = \sum_{\alpha} m_\alpha |R_\alpha|, \qquad \text{where } \ m_\alpha = \inf_{R_\alpha} \chi_{\phi(S)}f.$$

If $R_\alpha \cap \phi(S) = \emptyset$ then $\chi_{\phi(S)} = 0$ over R_α; so $m_\alpha = 0$. Hence the lower sum can be written as

$$L = \sum_{R_\alpha \cap \phi(S) \neq \emptyset} m_\alpha |R_\alpha|.$$

Now note that by Exercise 2.111, the distance between the points of $\partial(\phi(U))$ and $\phi(\overline{S})$ has a positive lower bound, since they are disjoint sets, $\partial(\phi(U))$ is closed, and $\phi(\overline{S})$ is compact. Therefore when $\|P\|$ is small enough, $R_\alpha \cap \phi(S) \neq \emptyset$ implies that $R_\alpha \subset \phi(U)$. Let us choose a partition P so that this holds.

Next note that by Theorem 8.53 each $\phi^{-1}(R_\alpha)$ is Jordan measurable. Also, $R_\alpha \cap \phi(S) \neq \emptyset$ if and only if $\phi^{-1}(R_\alpha) \cap S \neq \emptyset$, since ϕ is a one-to-one function. In addition we have $S \subset \bigcup_{R_\alpha \cap \phi(S) \neq \emptyset} \phi^{-1}(R_\alpha)$, because $\phi(S) \subset \bigcup R_\alpha$. Furthermore, two different subrectangles R_α, R_β can only intersect at their boundaries. Hence, similarly to the proof of Theorem 8.55, we can show that $\phi^{-1}(R_\alpha), \phi^{-1}(R_\beta)$ too can only intersect at their boundaries. Thus $\phi^{-1}(R_\alpha) \cap S$ and $\phi^{-1}(R_\beta) \cap S$ have disjoint interiors. Hence, by Theorem 8.43, for any integrable function g we get

$$\int_S g\,dx = \sum_{R_\alpha\cap\phi(S)\neq\emptyset} \int_{\phi^{-1}(R_\alpha)\cap S} g\,dx = \sum_{R_\alpha\cap\phi(S)\neq\emptyset} \int_{\phi^{-1}(R_\alpha)} \chi_S g\,dx,$$

because $\chi_S g$ vanishes on the Jordan measurable set $\phi^{-1}(R_\alpha) - S$. Also note that if $x \in \phi^{-1}(R_\alpha)$ then $\phi(x) \in R_\alpha$; so $m_\alpha \leq (\chi_{\phi(S)} f)(\phi(x))$. Now, by the previous lemma, and the fact that $m_\alpha \geq 0$, we have

$$\begin{aligned}
L = \sum_{R_\alpha\cap\phi(S)\neq\emptyset} m_\alpha|R_\alpha| &= \sum_{R_\alpha\cap\phi(S)\neq\emptyset} m_\alpha|\phi(\phi^{-1}(R_\alpha))| \\
&\leq \sum_{R_\alpha\cap\phi(S)\neq\emptyset} m_\alpha \int_{\phi^{-1}(R_\alpha)} |\det D\phi(x)|\,dx \\
&= \sum_{R_\alpha\cap\phi(S)\neq\emptyset} \int_{\phi^{-1}(R_\alpha)} m_\alpha |\det D\phi|\,dx \\
&\leq \sum_{R_\alpha\cap\phi(S)\neq\emptyset} \int_{\phi^{-1}(R_\alpha)} (\chi_{\phi(S)} f)(\phi(x))\,|\det D\phi|\,dx \\
&= \sum_{R_\alpha\cap\phi(S)\neq\emptyset} \int_{\phi^{-1}(R_\alpha)} \chi_S (f\circ\phi)\,|\det D\phi|\,dx = \int_S (f\circ\phi)\,|\det D\phi|\,dx.
\end{aligned}$$

Note that we have used the fact that $\chi_{\phi(S)}(\phi(x)) = \chi_S(x)$. Therefore we obtain

$$\int_{\phi(S)} f\,dy = \sup_P L \leq \int_S (f\circ\phi)\,|\det D\phi|\,dx.$$

On the other hand, note that $\phi^{-1} : \phi(U) \to U$ is also a one-to-one C^1 map such that $D\phi^{-1}(y) = \big(D\phi(\phi^{-1}(y))\big)^{-1}$ is an invertible matrix for every $y \in \phi(U)$. In addition, $\overline{\phi(S)} \subset \phi(\overline{S}) \subset \phi(U)$, since $\phi(\overline{S})$ is a compact set containing $\phi(S)$. Hence we can apply the above inequality to the Riemann integrable and nonnegative function $(f\circ\phi)\,|\det D\phi|$ on the Jordan measurable set $S = \phi^{-1}(\phi(S))$, and obtain

$$\begin{aligned}
&\int_S (f\circ\phi)(x)\,|\det D\phi(x)|\,dx \\
&\qquad = \int_{\phi^{-1}(\phi(S))} (f\circ\phi)(x)\,|\det D\phi(x)|\,dx \\
&\qquad \leq \int_{\phi(S)} (f\circ\phi)(\phi^{-1}(y))\,|\det D\phi(\phi^{-1}(y))|\,|\det D\phi^{-1}(y)|\,dy \\
&\qquad = \int_{\phi(S)} f(y)\,|\det I|\,dy = \int_{\phi(S)} f\,dy.
\end{aligned}$$

Thus we get the desired equality when $f \geq 0$.

Finally, in general, when f is not necessarily a nonnegative function, we can write $f = f^+ - f^-$. Remember that $f^\pm$ are nonnegative Riemann integrable functions. So, by what we have proved so far, we get

$$\begin{aligned}\int_{\phi(S)} f(y)dy &= \int_{\phi(S)} f^+(y)dy - \int_{\phi(S)} f^-(y)dy \\ &= \int_S (f^+ \circ \phi)\,|\det D\phi|\,dx - \int_S (f^- \circ \phi)\,|\det D\phi|\,dx \\ &= \int_S \big((f^+ - f^-) \circ \phi\big)\,|\det D\phi|\,dx = \int_S (f \circ \phi)\,|\det D\phi|\,dx,\end{aligned}$$

as desired. ■

Remark. We can also deduce the change of variables theorem for a general function f from the case of nonnegative functions as follows. First note that since f is Riemann integrable it is bounded, so we have $f \geq -M$ for some $M \geq 0$. Hence the change of variables theorem can be applied to $f + M \geq 0$. Thus we have

$$\begin{aligned}\int_{\phi(S)} f\,dy + \int_{\phi(S)} M\,dy = \int_{\phi(S)} f + M\,dy &= \int_S \big((f + M) \circ \phi\big)\,|\det D\phi|\,dx \\ &= \int_S (f \circ \phi)\,|\det D\phi|\,dx + \int_S (M \circ \phi)\,|\det D\phi|\,dx \\ &= \int_S (f \circ \phi)\,|\det D\phi|\,dx + M\int_S |\det D\phi|\,dx.\end{aligned}$$

But $\int_{\phi(S)} M\,dy = M\int_{\phi(S)} 1\,dy = M\int_S |\det D\phi|\,dx$, because the change of variables theorem holds for the constant nonnegative function 1. Therefore we obtain the desired result for f.

8.6 Improper Integrals

Remember that for a function $f : [a, b) \to \mathbb{R}$, which is unbounded and/or has an unbounded domain (and hence is not Riemann integrable), we can define its improper integral by using limits:

$$\int_a^b f(x)dx := \lim_{c \to b^-} \int_a^c f(x)dx,$$

provided that the limit exists. Let us rewrite this definition as follows

$$\int_{[a,b)} f(x)dx = \lim_{c \to b^-} \int_{[a,c]} f(x)dx.$$

In other words, if as we enlarge the domain of integration the values of the integrals converge to a limit, then that limit will by definition be the value of the integral over the limiting domain. We can use the same idea in higher dimensions, to define the improper integral of functions which are not Riemann integrable in the proper sense. Note that in higher dimensions, unlike integrals over intervals, there is another way that a function can cease to be Riemann integrable: when the domain of integration is not Jordan measurable. For example, there are open sets which are not Jordan measurable. (Such open sets also exist in one dimension, but they are less likely to be encountered.) So the need for an extended notion of integration in higher dimensions is at least as essential as in one dimension, if not even more.

In higher dimensions we can also use limits to define improper integrals. However, unlike intervals, there is no canonical way to approximate a domain in higher dimensions using smaller domains. For example, to approximate a square in two dimensions, we can use smaller squares inside that square, or we can use smaller squares with round corners. Depending on the application, either of these approximation methods can be preferable. To overcome this difficulty, we have to show that different methods of approximating the domain of integration result in the same value for the improper integral. Although it is possible to show this for a suitable class of functions, an easier approach is to use supremums instead of limits. Let us consider the one-dimensional improper integral again. This time let us first assume that $f \geq 0$. Then it is easy to see that

$$\int_{[a,b)} f(x)dx = \lim_{c \to b^-} \int_{[a,c]} f(x)dx = \sup_{[a,c] \subset [a,b)} \int_{[a,c]} f(x)dx,$$

because the values of the integrals of a nonnegative function increase as we enlarge the domain of integration. This observation motivates the following definition.

Definition 8.57. Let $A \subset \mathbb{R}^n$ and $f : A \to \mathbb{R}$. Suppose f is Riemann integrable on every compact Jordan measurable set $S \subset A$. When $f \geq 0$, provided that the following supremum is finite, the **improper Riemann integral** of f over A is

$$\int_A f(x)dx := \sup_{S \subset A} \int_S f(x)dx,$$

where the supremum is over all compact Jordan measurable subsets of A.

In general, when f is not necessarily nonnegative, the **improper Riemann integral** of $f : A \to \mathbb{R}$ is

$$\int_A f(x)dx := \int_A f^+(x)dx - \int_A f^-(x)dx,$$

provided that the improper integrals of f^+, f^- exist. If the improper Riemann integral of f exists we say f is **integrable** over A.

Remark. Remember that $f^{\pm} = \max\{\pm f, 0\}$ are nonnegative functions such that $f = f^{+} - f^{-}$. Hence, Riemann integrability of $f^{\pm}$ on every compact Jordan measurable set $S \subset A$ implies Riemann integrability of f on them. Conversely, since $f^{\pm} = x^{\pm} \circ f$ and $x^{\pm} = \max\{\pm x, 0\}$ are continuous functions, Riemann integrability of f on every compact Jordan measurable set $S \subset A$ implies Riemann integrability of $f^{\pm}$ on them.

Notation. When $f \geq 0$ and the above supremum is infinite we set $\int_A f(x)dx = \infty$. In general, if only one of the integrals of $f^{\pm}$ is infinite and the other one is finite, we set $\int_A f(x)dx = \pm\infty$, where the sign of ∞ is determined based on the one integral which is infinite. (And if both the integrals of $f^{\pm}$ are infinite, we do not assign any value to the integral of f, since the difference of two infinities has no well-defined value.) But, keep in mind that these assignments are mere notational conventions, and when we talk about the integrability of a function we mean that its improper integral exists as a finite value.

Remark. Since Jordan measurable sets are bounded, the supremum in the above definition can also be taken over all closed Jordan measurable subsets of A. However, note that we do not take the supremum over all Jordan measurable subsets of A. Because in that case if A happens to be Jordan measurable itself, then we would be assuming that f is Riemann integrable over A, whereas there are functions over Jordan measurable sets whose improper integral exists but they are not Riemann integrable in the proper sense; see Example 8.66.

Remark. Note that we do not impose any assumptions on A. In particular, it need not be bounded or Jordan measurable. But when A is Jordan measurable, and f is Riemann integrable, we need to check that these new definitions of integral and integrability are compatible with the old ones. This will be done in Theorem 8.65.

On the other hand, although the definition does not require it, to get a sensible definition of integral we need to restrict the domain A to be a member of a suitable class of subsets of $\mathbb{R}^n$. An appropriate such class is the class of Lebesgue measurable sets; see Section 10.2 for the definition. For example, all open sets, closed sets, or sets with measure zero in $\mathbb{R}^n$ are Lebesgue measurable. Since all the sets that we will encounter in the sequel are Lebesgue measurable, we are not going to concern ourselves with checking this condition when we use improper integrals.

Theorem 8.58. *Let $A \subset \mathbb{R}^n$, and let $f, g : A \to \mathbb{R}$ be integrable functions. Then we have*

(i) *$f + g$ is integrable and*

$$\int_A [f(x) + g(x)]dx = \int_A f(x)dx + \int_A g(x)dx.$$

(ii) *cf is integrable for $c \in \mathbb{R}$ and we have*

$$\int_A cf(x)dx = c\int_A f(x)dx.$$

(iii) *If $f \leq g$ then*

$$\int_A f(x)dx \leq \int_A g(x)dx.$$

Proof. Let S be an arbitrary compact Jordan measurable subset of A. By assumption, f, g are Riemann integrable on S. Hence $f+g$, cf, and $|f|$ are all Riemann integrable on S.

(i) First assume that $f, g \geq 0$. Then $f+g \geq 0$ too. We have

$$\begin{aligned}\int_S f(x)+g(x)\,dx &= \int_S f(x)dx + \int_S g(x)dx \\ &\leq \sup_{S\subset A}\int_S f(x)dx + \sup_{S\subset A}\int_S g(x)dx.\end{aligned}$$

Hence

$$\sup_{S\subset A}\int_S f(x)+g(x)\,dx \leq \sup_{S\subset A}\int_S f(x)dx + \sup_{S\subset A}\int_S g(x)dx < \infty, \qquad (*)$$

and therefore $f+g$ is integrable. It is worth noting that although the supremum of the sum is less than or equal to the sum of the supremums, the equality does not hold in general. But, as we will see, in this particular case the equality holds, since the integral of a nonnegative function increases as we enlarge the domain of integration.

Now let S_0 be an arbitrary, but fixed, compact Jordan measurable subset of A. Then for compact Jordan measurable sets $\tilde{S}$ satisfying $S_0 \subset \tilde{S} \subset A$ we have

$$\begin{aligned}\int_{S_0} f(x)dx + \int_{\tilde{S}} g(x)dx &\leq \int_{\tilde{S}} f(x)dx + \int_{\tilde{S}} g(x)dx \\ &= \int_{\tilde{S}} f(x)+g(x)\,dx \leq \sup_{S\subset A}\int_S f(x)+g(x)\,dx.\end{aligned}$$

By taking supremum over $\tilde{S}$ on the left hand side we get

$$\int_{S_0} f(x)dx + \sup_{S_0\subset\tilde{S}\subset A}\int_S g(x)dx \leq \sup_{S\subset A}\int_S f(x)+g(x)\,dx.$$

However, note that for every S we have $S \subset S_0\cup S$, and $S_0\cup S$ is a compact Jordan measurable set containing S_0; so

$$\int_S g(x)dx \leq \int_{S_0\cup S} g(x)dx \leq \sup_{S_0\subset\tilde{S}\subset A}\int_{\tilde{S}} g(x)dx.$$

Thus $\sup_{S\subset A}\int_S g(x)dx \le \sup_{S_0\subset\tilde{S}\subset A}\int_{\tilde{S}} g(x)dx$, and therefore the two supremums are equal. Hence we have shown that

$$\int_{S_0} f(x)dx + \sup_{S\subset A}\int_S g(x)dx \le \sup_{S\subset A}\int_S f(x)+g(x)\,dx.$$

Now by taking supremum over S_0 we obtain

$$\sup_{S_0\subset A}\int_{S_0} f(x)dx + \sup_{S\subset A}\int_S g(x)dx \le \sup_{S\subset A}\int_S f(x)+g(x)\,dx,$$

which together with $(*)$ implies that

$$\sup_{S\subset A}\int_S f(x)+g(x)\,dx = \sup_{S\subset A}\int_S f(x)dx + \sup_{S\subset A}\int_S f(x)dx,$$

or equivalently

$$\int_A f(x)+g(x)\,dx = \int_A f(x)dx + \int_A g(x)dx.$$

Hence we have proved this part for $f,g\ge 0$.

Next, let us consider general f,g. We have $f=f^+-f^-$ and $g=g^+-g^-$. Thus

$$(f+g)^+ - (f+g)^- = f+g = f^+ - f^- + g^+ - g^-.$$

So

$$(f+g)^+ + f^- + g^- = f^+ + g^+ + (f+g)^-. \tag{$**$}$$

But $f^\pm, g^\pm \ge 0$, and they are integrable by assumption; so $f^+ + g^+$ and $f^- + g^-$ are integrable by the above argument. On the other hand we have

$$\begin{aligned}\int_S (f+g)^\pm dx \le \int_S |f+g|\,dx \le \int_S |f|+|g|\,dx &= \int_S f^+ + f^- + g^+ + g^- dx\\ &= \int_S f^+dx + \int_S f^-dx + \int_S g^+dx + \int_S g^-dx\\ &\le \int_A f^+dx + \int_A f^-dx + \int_A g^+dx + \int_A g^-dx < \infty\end{aligned}$$

for every S. Hence by taking supremum over S we can see that $(f+g)^\pm$, and therefore $f+g$, are integrable over A. So, all the functions in the equation $(**)$ are nonnegative and integrable; thus after integration we obtain

$$\int_A (f+g)^+dx + \int_A f^-dx + \int_A g^-dx = \int_A f^+dx + \int_A g^+dx + \int_A (f+g)^-dx.$$

Then by rearranging the terms we get

$$\begin{aligned}\int_A (f+g)dx &= \int_A (f+g)^+dx - \int_A (f+g)^-dx\\ &= \int_A f^+dx - \int_A f^-dx + \int_A g^+dx - \int_A g^-dx = \int_A fdx + \int_A gdx,\end{aligned}$$

as desired.

(ii) First assume that $c > 0$ and $f \geq 0$. Then $cf \geq 0$ too. We have

$$\int_S cf(x)\,dx = c\int_S f(x)dx \leq c\sup_{S\subset A}\int_S f(x)dx = c\int_A f(x)dx.$$

Hence $\sup_{S\subset A}\int_S cf(x)\,dx \leq c\int_A f(x)dx < \infty$, and therefore cf is integrable. In addition we have

$$\int_A cf(x)dx \leq c\int_A f(x)dx.$$

On the other hand, by applying the above inequality to the integrable function cf and the constant $\frac{1}{c}$ we obtain

$$\int_A f(x)dx = \int_A \frac{1}{c}\big(cf(x)\big)dx \leq \frac{1}{c}\int_A cf(x)dx \implies c\int_A f(x)dx \leq \int_A cf(x)dx.$$

Therefore $\int_A cf(x)dx = c\int_A f(x)dx$, and thus we have proved this part in this case.

Next, let us consider general c, f. We have $f = f^+ - f^-$. When $c > 0$ we can easily see that $(cf)^\pm = cf^\pm$, which are integrable functions by the above argument. Hence we have

$$\begin{aligned}\int_A cfdx &= \int_A (cf)^+dx - \int_A (cf)^-dx\\ &= \int_A cf^+dx - \int_A cf^-dx = c\int_A f^+dx - c\int_A f^-dx = c\int_A fdx.\end{aligned}$$

Similarly, when $c < 0$ we can easily see that $(cf)^\pm = -cf^\mp$, which are again integrable functions by the above argument. Then we have

$$\begin{aligned}\int_A cfdx &= \int_A (cf)^+dx - \int_A (cf)^-dx\\ &= \int_A -cf^-dx - \int_A -cf^+dx = -c\int_A f^-dx + c\int_A f^+dx = c\int_A fdx.\end{aligned}$$

Finally, when $c = 0$ we have $cf = 0$, which is an integrable function. Also, $\int_A 0f(x)dx = 0 = 0\int_A f(x)dx$, as desired.

(iii) First suppose $f, g \geq 0$. Then we have

$$\int_S f(x)dx \leq \int_S g(x)dx \leq \int_A g(x)dx.$$

Hence by taking supremum over S we get $\int_A f(x)dx \leq \int_A g(x)dx$. Now for general f, g we have

$$f^+ - f^- = f \leq g = g^+ - g^-.$$

Hence $f^+ + g^- \leq g^+ + f^-$. Then by the above argument and part (i) we obtain

$$\int_A f^+dx + \int_A g^-dx = \int_A f^+ + g^-dx \leq \int_A g^+ + f^-dx = \int_A g^+dx + \int_A f^-dx.$$

Hence by rearranging the terms we get

$$\int_A fdx = \int_A f^+dx - \int_A f^-dx \leq \int_A g^+dx - \int_A g^-dx = \int_A gdx,$$

as desired. ■

Theorem 8.59. *Let $A \subset \mathbb{R}^n$, and suppose $f : A \to \mathbb{R}$ is Riemann integrable on every compact Jordan measurable subset of A. Then the existence of the improper integral $\int_A f(x)dx$ is equivalent to the existence of the improper integral $\int_A |f(x)|dx$. And in this case we have*

$$\left|\int_A f(x)dx\right| \leq \int_A |f(x)|dx.$$

Remark. In particular, if f is integrable over A then $|f|$ is also integrable over A. However, the integrability of $|f|$ alone does not imply the integrability of f. As an example, consider $f : [0, 1] \to \mathbb{R}$ which has value 1 on rational points, and value -1 on irrational points. Then $|f|$ is integrable while f is not, since it is discontinuous at every point. In comparison, what this theorem states is that under the assumption of integrability of f on compact Jordan measurable subsets of A, the finiteness of $\int_A f(x)dx$ is equivalent to the finiteness of $\int_A |f(x)|dx$.

Remark. This result is in contrast with the properties of one-dimensional improper integrals. For example, it can be shown that $\int_0^\infty \frac{\sin(x)}{x}\,dx$ exists while $\int_0^\infty |\frac{\sin(x)}{x}|\,dx = \infty$. The difference originates from the different definitions being used: using limits in one dimension and using supremums in higher dimensions. Although it is possible to define improper integrals in higher dimensions by using limits, and avoid separating the case of positive functions and general ones in the definition, the resulting theory is needlessly complicated for our purposes.

$\boxed{\textbf{Proof.}}$ Note that $|f|$ is also Riemann integrable on every compact Jordan measurable set $S \subset A$. First suppose $\int_A f(x)dx$ exists. We have $f = f^+ - f^-$, and by assumption $f^\pm$ are integrable. On the other hand we know that $|f| = f^+ + f^-$. Hence, by part (i) of the previous theorem, $|f|$ is integrable and we have

$$\int_A |f(x)|dx = \int_A f^+(x)dx + \int_A f^-(x)dx.$$

Now, noting that $\int_A f^\pm(x)dx \geq 0$, we have

$$\begin{aligned}\left|\int_A f(x)dx\right| &= \left|\int_A f^+(x)dx - \int_A f^-(x)dx\right| \\ &\leq \left|\int_A f^+(x)dx\right| + \left|\int_A f^-(x)dx\right| \\ &= \int_A f^+(x)dx + \int_A f^-(x)dx = \int_A |f(x)|\,dx,\end{aligned}$$

as desired. Conversely, suppose $\int_A |f(x)|dx$ exists. We know that $f^\pm \leq |f|$, so

$$\int_S f^\pm dx \leq \int_S |f|dx \leq \sup_{S\subset A} \int_S |f|dx = \int_A |f|dx < \infty$$

for every S. Hence by taking supremum over S we can see that $f^\pm$, and therefore f, are integrable over A. ■

Theorem 8.60. *Let $A \subset \mathbb{R}^n$, and suppose $f, g : A \to \mathbb{R}$ are Riemann integrable on every compact Jordan measurable subset of A. Suppose $|f| \leq g$. If the improper integral $\int_A g(x)dx$ exists then the improper integral $\int_A f(x)dx$ exists too, and we have*

$$\left|\int_A f(x)dx\right| \leq \int_A g(x)dx.$$

Remark. Equivalently, if the improper integral $\int_A f(x)dx$ does not exist, then the improper integral $\int_A g(x)dx$ does not exist either (it is infinite).

Remark. As a result, if the integral $\int_A 1\,dx$ exists (in other words, if A has finite volume), then all bounded continuous functions are integrable over A. Because a bounded continuous function like f is Riemann integrable on every Jordan measurable subset of A, and satisfies $|f| \leq C$ for some constant C. In addition, the constant function $C = C \cdot 1$ is integrable over A.

$\boxed{\textbf{Proof.}}$ We only need to show that $|f|$ is integrable over A. Because then by the previous theorem f is also integrable over A, and in addition we have

$$\left|\int_A f(x)dx\right| \leq \int_A |f(x)|dx \leq \int_A g(x)dx.$$

Now for any compact Jordan measurable set $S \subset A$ we have

$$\int_S |f(x)|dx \leq \int_S g(x)dx \leq \sup_{S\subset A} \int_S g(x)dx = \int_A g(x)dx < \infty.$$

So $\sup_{S\subset A} \int_S |f(x)|dx$ is also finite, and therefore $|f|$ is integrable over A. ■

Besides its theoretical implications, the following theorem also provides a method for computing improper integrals, as we will see in Examples 8.66 and 8.67.

Theorem 8.61. *Let $A \subset \mathbb{R}^n$ and $f : A \to \mathbb{R}$, and suppose f is Riemann integrable on every compact Jordan measurable subset of A. Let S_k be a sequence of compact Jordan measurable subsets of A such that $S_k \subset S_{k+1}^\circ$ and $A = \bigcup_{k\geq 1} S_k$. Then the improper integral $\int_A f(x)dx$ exists if and only if*

$$\sup_{k\geq 1} \int_{S_k} |f(x)|dx < \infty.$$

And in this case we have

$$\int_A f(x)dx = \lim_{k\to\infty} \int_{S_k} f(x)dx.$$

Remark. Note the absolute value in the above condition for the existence of the improper integral.

Remark. In fact, in this theorem we only need S_k to be a subset of the interior of S_{k+1} as a subspace of A. The interior of S_{k+1} as a subspace of A can be strictly larger than S_{k+1}°, which is the interior of S_{k+1} in $\mathbb{R}^n$. For example if $A = [0,1]^2$, then the interior of $[0,1/2]^2$ as a subspace of A is $[0,1/2)^2$, since points on $\{0\} \times (0,1/2)$ and $(0,1/2) \times \{0\}$ have a neighborhood in A which completely lies in $[0,1/2]^2$.

Proof. First let us assume that $f \geq 0$. Now if the improper integral $\int_A f(x)dx$ exists, we have

$$\sup_{k\geq 1} \int_{S_k} |f(x)|dx = \sup_{k\geq 1} \int_{S_k} f(x)dx \leq \sup_{S\subset A} \int_S f(x)dx = \int_A f(x)dx < \infty, \qquad (\star)$$

where the rightmost supremum is over all compact Jordan measurable subsets of A. Conversely, assume that $\sup_{k\geq 1} \int_{S_k} |f(x)|dx < \infty$. We have to show that $\sup_{S\subset A} \int_S f(x)dx$ is also finite. We will show that in fact

$$\sup_{S\subset A} \int_S f(x)dx \leq \sup_{k\geq 1} \int_{S_k} f(x)dx. \qquad (*)$$

Let $S \subset A$ be a compact Jordan measurable set. If $S \subset S_m$ for some m, we have

$$\int_S f(x)dx \leq \int_{S_m} f(x)dx \leq \sup_{k\geq 1} \int_{S_k} f(x)dx. \qquad (**)$$

Suppose to the contrary that $S \not\subset S_k$ for any k. Then for every k there exists $x_k \in S - S_k$. The sequence x_k lies in the compact set S, so it has a subsequence, which we still denote by x_k, converging to $x^* \in S \subset A \subset \bigcup S_k$. Thus there is m such that $x^* \in S_m \subset S^\circ_{m+1}$. Now since x^* is an interior point of S_{m+1}, an open neighborhood B of x^* is a subset of S_{m+1}. But due to the convergence of x_k to x^*, B contains every x_k for large enough k. (Note that for this to be true, B only needs to be an open neighborhood of x^* in A, since the sequence x_k and its limit x^* lie in A. So we only need x^* to be an interior point of S_{m+1} as a subspace of A.) Therefore, for large enough k we get $x_k \in B \subset S_{m+1} \subset S_k$, which is a contradiction. Hence we must have $S \subset S_m$ for some m. And thus the inequality $(**)$ holds for every such S. Now we can take the supremum over S in the inequality $(**)$ to obtain $(*)$, as desired. Then, by combining the inequalities $(\star)$ and $(*)$ we get

$$\int_A f(x)dx = \sup_{S\subset A} \int_S f(x)dx = \sup_{k\geq 1} \int_{S_k} f(x)dx = \lim_{k\to\infty} \int_{S_k} f(x)dx,$$

since $f \geq 0$ and therefore the sequence of integrals $\int_{S_k} f(x)dx$ increases and converges to its finite supremum.

Next, for a general function f we have to consider $f^\pm$. By the above argument, the integrability of $f^\pm$ is equivalent to

$$\sup_{k\geq 1} \int_{S_k} f^\pm(x)dx = \sup_{k\geq 1} \int_{S_k} |f^\pm(x)|dx < \infty.$$

Now if f is integrable then by definition $f^\pm$ are integrable. Hence the above two supremums are finite. However, we have $|f| = f^+ + f^-$; so

$$\begin{aligned}\int_{S_k} |f(x)|dx &= \int_{S_k} f^+(x)dx + \int_{S_k} f^-(x)dx \\ &\leq \sup_{k\geq 1} \int_{S_k} f^+(x)dx + \sup_{k\geq 1} \int_{S_k} f^-(x)dx,\end{aligned}$$

which implies $\sup_{k\geq 1} \int_{S_k} |f(x)|dx < \infty$, as wanted. On the other hand, if we have $\sup_{k\geq 1} \int_{S_k} |f(x)|dx < \infty$, then by using the fact that $f^\pm \leq |f|$ we get

$$\int_{S_k} f^\pm(x)dx \leq \int_{S_k} |f(x)|dx \leq \sup_{k\geq 1} \int_{S_k} |f(x)|dx,$$

which implies $\sup_{k\geq 1}\int_{S_k} f^{\pm}(x)dx < \infty$. Thus both $f^{\pm}$ are integrable. Hence, by definition, f is integrable too. Finally, we have

$$\begin{aligned}\int_A f(x)dx &= \int_A f^{+}(x)dx - \int_A f^{-}(x)dx\\ &= \lim_{k\to\infty}\int_{S_k} f^{+}(x)dx - \lim_{k\to\infty}\int_{S_k} f^{-}(x)dx\\ &= \lim_{k\to\infty}\Big(\int_{S_k} f^{+}(x)dx - \int_{S_k} f^{-}(x)dx\Big)\\ &= \lim_{k\to\infty}\int_{S_k} f^{+}(x) - f^{-}(x)\,dx = \lim_{k\to\infty}\int_{S_k} f(x)dx,\end{aligned}$$

as desired. ■

The assumptions of the above theorem are in particular satisfied when the domain of integration is open, as we will show in the next theorem. In the following proof we will employ the distance function to the boundary

$$\operatorname{dist}(x,\partial U) := \inf_{y\in\partial U}|x-y|.$$

***Exercise* 8.62.** Let $U\subset\mathbb{R}^n$. Show that $\operatorname{dist}(x,\partial U)$ is a continuous function.

Solution. We will show that in fact dist is a Lipschitz function satisfying

$$|\operatorname{dist}(x,\partial U)-\operatorname{dist}(z,\partial U)| \leq |x-z|.$$

To see this note that for $y\in\partial U$ we have

$$\operatorname{dist}(x,\partial U) \leq |x-y| \leq |x-z|+|z-y|.$$

By taking infimum over y on the right hand side we get

$$\operatorname{dist}(x,\partial U) \leq |x-z| + \inf_{y\in\partial U}|z-y| = |x-z| + \operatorname{dist}(z,\partial U).$$

Now by switching the role of x, z we can deduce the desired inequality. ■

Theorem 8.63. *Let $U\subset\mathbb{R}^n$ be an open set. Then there is a sequence of compact Jordan measurable sets $S_k\subset U$ such that $S_k\subset S_{k+1}^{\circ}$ and $U=\bigcup_{k\geq 1}S_k$.*

Proof. Consider the sets

$$U_k := \{x\in U : |x|<k \quad\text{and}\quad \operatorname{dist}(x,\partial U) > \frac{1}{k}\}.$$

Since the distance function is continuous, each U_k is an open set. In addition we have $\overline{U}_k \subset U_{k+1}$, because for any limit point z of U_k we must have $|z| \leq k$ and $\text{dist}(z, \partial U) \geq 1/k$; hence $z \in U_{k+1}$. In addition we have

$$U = \bigcup_{k \geq 1} U_k.$$

The reason is that for each $x \in U$ there is $r > 0$ such that $B_r(x) \subset U$; hence $\text{dist}(x, \partial U) > r$. Thus if we choose k large enough so that $k > \frac{1}{r}$ and $k > |x|$ then we have $x \in U_k$.

However, U_k's are not necessarily Jordan measurable. To construct Jordan measurable subsets of U we argue as follows. For each point $x \in \overline{U}_k \subset U_{k+1}$ there is an open rectangle $R_x \subset U_{k+1}$ containing x, since U_{k+1} is open. We can make R_x smaller if necessary to ensure that in fact $\overline{R}_x \subset U_{k+1}$. Now note that $U_k \subset B_k(0)$ is bounded, so $\overline{U}_k$ is compact. Hence finitely many of these open rectangles, namely $R_1, \dots, R_m$ cover $\overline{U}_k$. Let $S_k := \bigcup_{j=1}^{m} \overline{R}_j$. Then S_k is compact, being the union of finitely many compact sets. Also, by Exercise 2.36, ∂S_k is contained in $\bigcup_{j \leq m} \partial R_j$, which has measure zero. So ∂S_k has measure zero too, and thus S_k is Jordan measurable. In addition we have

$$\overline{U}_k \subset \bigcup_{j=1}^{m} R_j \subset S_k \subset U_{k+1}.$$

Now note that we have $S_k \subset U_{k+1} \subset S^{\circ}_{k+1}$, since U_{k+1} is an open subset of S_{k+1}. Furthermore,

$$U = \bigcup_{k \geq 1} U_k \subset \bigcup_{k \geq 1} S_k \subset U.$$

So $U = \bigcup_{k \geq 1} S_k$, as desired. ■

Theorem 8.64. *Suppose $U \subset \mathbb{R}^n$ is an open set and $f : U \to \mathbb{R}$ is integrable. Then the set of discontinuities of f in U has measure zero.*

Remark. However, note that unlike the case of (proper) Riemann integrable functions, here f need not be bounded.

Proof. Let S_k be a sequence of compact Jordan measurable subsets of U given by the previous theorem. Let Z_k be the set of discontinuities of $f|_{S_k}$. Since $f|_{S_k}$ is Riemann integrable by assumption, Z_k has measure zero by Riemann-Lebesgue theorem. Let Z be the set of discontinuities of f. Then we must have $Z = \bigcup_{k \geq 1} Z_k$. To see this suppose f is discontinuous at some $x \in U = \bigcup_{k \geq 1} S_k$. Then $x \in S_j \subset S^{\circ}_{j+1}$ for some j. Hence $f|_{S_{j+1}}$ is also discontinuous at x, since f and its restriction $f|_{S_{j+1}}$ are equal on a neighborhood of x. Conversely, if $f|_{S_j}$ is discontinuous at x

for some j, then f is also discontinuous at x, because restrictions of a continuous function are continuous. Hence $Z = \bigcup_{k\geq 1} Z_k$, and therefore Z has measure zero, since it is the union of countably many sets with measure zero. ■

Theorem 8.65. *Suppose $A \subset \mathbb{R}^n$ is Jordan measurable and $f : A \to \mathbb{R}$ is Riemann integrable on A. Then the improper integral of f over A exists and is equal to the (proper) Riemann integral of f over A.*

Proof. In this proof we denote improper integrals by $\int$ to avoid any possible confusions with proper integrals denoted by $\int$. Let S be an arbitrary compact Jordan measurable subset of A. Note that f is Riemann integrable on every such S. First suppose $f \geq 0$. Then we have

$$\int_S f(x)dx \leq \int_A f(x)dx \implies \sup_{S\subset A} \int_S f(x)dx \leq \int_A f(x)dx < \infty. \qquad (*)$$

Hence the improper integral of f over A exists, and is less than or equal to its proper integral over A. On the other hand, we have $A = A^\circ \cup \partial A$, and we know that ∂A has measure zero. In addition, by Exercise 8.42, we know that both ∂A and A° are Jordan measurable. Thus we have $\int_{\partial A} f(x)dx = 0$, and therefore

$$\int_A f(x)dx = \int_{A^\circ} f(x)dx.$$

Next let us show that for every ϵ there is a compact Jordan measurable set $S \subset A^\circ$ such that $|A^\circ - S| < \epsilon$. We know that ∂A° has measure zero, since A° is Jordan measurable. Thus for a given $\epsilon > 0$ there exist a countable family of open cubes $\{Q_k\}$ that covers ∂A°, and $\sum |Q_k| < \epsilon$. Since A° is bounded, ∂A° is compact; so we can assume that the family of cubes is finite. Let

$$W := A^\circ \cap \big(\bigcup Q_k\big), \qquad S := A^\circ - W.$$

Then by Theorem 8.40, W is Jordan measurable, and its volume satisfies

$$|W| \leq \big|\bigcup Q_k\big| \leq \sum |Q_k| < \epsilon.$$

Therefore S is Jordan measurable too, and we have $|A^\circ - S| = |W| < \epsilon$. Note that S is also closed, because W^c is closed, and we have

$$\begin{aligned} \overline{A^\circ} \cap W^c &= (A^\circ \cup \partial A^\circ) \cap W^c \\ &= (A^\circ \cap W^c) \cup (\partial A^\circ \cap W^c) = (A^\circ \cap W^c) \cup \emptyset = A^\circ - W = S, \end{aligned}$$

since $\partial A^\circ \subset W$. Thus S is compact as it is bounded.

Now we have

$$\left|\int_{A^\circ} f(x)dx - \int_S f(x)dx\right| = \left|\int_{A^\circ - S} f(x)dx\right| \le M\epsilon,$$

where M is an upper bound for the Riemann integrable function $|f|$. Thus, since $\int_A f(x)dx = \int_{A^\circ} f(x)dx$, we have shown that there are compact Jordan measurable subsets of A over which the integral of f is arbitrarily close to $\int_A f(x)dx$. Therefore we must have

$$\int\limits_{A} f(x)dx = \sup_{S \subset A} \int_S f(x)dx \ge \int_A f(x)dx.$$

Hence, by combining this inequality with $(*)$, we see that the improper integral of the nonnegative function f over A is equal to its proper integral over A.

Finally, for a general f we have $f = f^+ - f^-$. We know that $f^\pm$ are also Riemann integrable over A. Then, since $f^\pm \ge 0$, by the above argument their improper integrals over A exist and are equal to their proper integrals. Hence the improper integral of f over A exists, and we have

$$\begin{aligned}\int\limits_{A} f(x)dx &= \int\limits_{A} f^+(x)dx - \int\limits_{A} f^-(x)dx \\ &= \int_A f^+(x)dx - \int_A f^-(x)dx = \int_A f^+(x) - f^-(x)\,dx = \int_A f(x)dx,\end{aligned}$$

as desired. ■

***Example* 8.66.** Consider the function $f(x,y) = \frac{1}{\sqrt{xy}}$ on $(0,1]^2 \subset \mathbb{R}^2$. Then we have

$$\begin{aligned}\int_{[1/k,\,1]^2} f(x,y)dxdy &= \int_{1/k}^1 \int_{1/k}^1 \frac{1}{\sqrt{xy}}\,dx\,dy = \int_{1/k}^1 \frac{1}{\sqrt{y}}\,dy \int_{1/k}^1 \frac{1}{\sqrt{x}}\,dx \\ &= \Big(2\sqrt{y}\,\Big|_{y=1/k}^1\Big)\Big(2\sqrt{x}\,\Big|_{x=1/k}^1\Big) = \big(2 - \frac{2}{\sqrt{k}}\big)^2 \le 4.\end{aligned}$$

Note that the sequence of compact sets $[1/k, 1]^2$ satisfy the assumptions of Theorem 8.61; so f is integrable over $(0,1]^2$, and we have

$$\int_{(0,1]^2} f(x,y)dxdy = \lim_{k\to\infty} \int_{[1/k,\,1]^2} f(x,y)dxdy = 4.$$

Also, notice that $(0,1]^2$ is Jordan measurable, however, f is not Riemann integrable on $(0,1]^2$ in the proper sense, since f is not bounded there.

On the other hand, for $f^2 = \frac{1}{xy}$ we have

$$\begin{aligned}\int_{[1/k,\,1]^2} f^2(x,y)dxdy &= \int_{1/k}^{1}\int_{1/k}^{1} \frac{1}{xy}\,dx\,dy = \int_{1/k}^{1} \frac{1}{y}\,dy \int_{1/k}^{1} \frac{1}{x}\,dx \\ &= \Big(\log y\,\Big|_{y=1/k}^{1}\Big)\Big(\log x\,\Big|_{x=1/k}^{1}\Big) = \big(\log k\big)^2 \underset{k\to\infty}{\longrightarrow} \infty.\end{aligned}$$

So f^2 is not integrable over $(0,1]^2$. In fact we have $\int_{(0,1]^2} f^2(x,y)dxdy = \infty$, since the supremum of integrals over all compact Jordan measurable subsets of $(0,1]^2$ is no less than the supremum of integrals over the sequence of sets $[1/k,1]^2$. As a result, we see that the product of two integrable functions can fail to be integrable. Also, we see that the composition of a continuous function with an integrable function ($t \mapsto t^2$ and f) can fail to be integrable. In both these cases, the problem is that the integrals may become infinite after the operation (multiplication or composition).

Remark. Although the product of two integrable functions may fail to be integrable, it will be integrable when one of the functions is bounded. To see this suppose f,g are integrable over A and $|g| \le C$. Then we have $|fg| \le C|f|$. By Theorem 8.59 we know that $|f|$ is integrable. Thus $C|f|$ is also integrable, and therefore fg is integrable by Theorem 8.60.

***Example* 8.67.** We can also integrate functions over unbounded domains using improper integrals. For example, consider the function $f(x,y) = \frac{1}{x^2y^2}$ on $[1,\infty)^2 \subset \mathbb{R}^2$. Then we have

$$\begin{aligned}\int_{[1,\,k]^2} f(x,y)dxdy &= \int_{1}^{k}\int_{1}^{k} \frac{1}{x^2y^2}dxdy = \int_{1}^{k} \frac{1}{y^2}\Big(\frac{-1}{x}\Big|_{x=1}^{k}\Big)dy \\ &= \int_{1}^{k} \frac{1}{y^2}\big(\frac{-1}{k}+1\big)dy = \Big(\frac{-1}{y}\Big|_{y=1}^{k}\Big)\big(\frac{-1}{k}+1\big) = \big(\frac{-1}{k}+1\big)^2 \le 1.\end{aligned}$$

Note that the sequence of compact sets $[1,k]^2$ satisfy the assumptions of Theorem 8.61; so f is integrable over $[1,\infty)^2$, and we have

$$\int_{[1,\infty)^2} f(x,y)dxdy = \lim_{k\to\infty} \int_{[1,\,k]^2} f(x,y)dxdy = 1.$$

Now let us prove a version of Theorem 8.43 for improper integrals, which will be needed in the next chapter.

Theorem 8.68. *Let $U_1, U_2, U \subset \mathbb{R}^n$ be open sets such that $U_1 \cap U_2 = \emptyset$ and $U = U_1 \cup U_2 \cup Z$, where $Z = U - (U_1 \cup U_2)$ has measure zero. Suppose $f : U \to \mathbb{R}$ is bounded on compact subsets of U. Then f is integrable over U if and only if it is integrable over U_1, U_2, and in this case we have*

$$\int_U f(x)dx = \int_{U_1} f(x)dx + \int_{U_2} f(x)dx.$$

Remark. This theorem can be generalized to the case where U_1, U_2, U are not necessarily open, but merely Lebesgue measurable sets. However, the proof in this more general case requires tools that are not within our reach yet.

Remark. By an easy induction, we can generalize this theorem to the case of more than two pairwise disjoint open sets $U_1, \dots, U_k$.

Remark. If f is integrable over U then it is automatically bounded on compact subsets of U. Because, in this case, by definition f is Riemann integrable, and hence bounded, on every compact Jordan measurable subset of U. Now note that any compact subset K of U is bounded, and has a positive distance from ∂U due to Exercise 2.111. Thus, by the construction of the sequence of compact Jordan measurable subsets in the proof of Theorem 8.63, there is a compact Jordan measurable subset of U that contains K. Therefore f is bounded on K too.

Proof. First suppose $f \ge 0$. Suppose f is integrable over U. Let $S_1 \subset U_1$, $S_2 \subset U_2$, and $S \subset U$ be arbitrary compact Jordan measurable sets. Then $S_1 \cap S_2 = \emptyset$, and $S_1 \cup S_2$ is a compact Jordan measurable subset of U. Hence we have

$$\begin{aligned} \int_{S_1} f(x)dx + \int_{S_2} f(x)dx = \int_{S_1 \cup S_2} f(x)dx \le \sup_{S \subset U} \int_S f(x)dx = \int_U f(x)dx \\ \implies \sup_{S_1 \subset U_1} \int_{S_1} f(x)dx + \sup_{S_2 \subset U_2} \int_{S_2} f(x)dx \le \int_U f(x)dx. \end{aligned}$$

Hence f is integrable over U_1, U_2, and we have

$$\int_{U_1} f(x)dx + \int_{U_2} f(x)dx \le \int_U f(x)dx. \tag{$*$}$$

Now suppose f is integrable over U_1, U_2. Then since the sets of discontinuities of f in U_1, U_2 have measure zero by Theorem 8.64, and Z has measure zero, the set of discontinuities of f in U has measure zero too. Consider a compact Jordan measurable set $S \subset U$. Then f is Riemann integrable on S, because by assumption f is bounded on S, and as we have seen its set of discontinuities has measure zero. Let $K_j \subset U_1$ and $\tilde{K}_j \subset U_2$ be sequences of compact Jordan measurable sets given by Theorem 8.63. Then the sets $S \cap K_j$ and $S \cap \tilde{K}_j$ are compact Jordan measurable subsets of U_1, U_2, respectively. Hence $S \cap K_j$ and $S \cap \tilde{K}_j$ are disjoint, and therefore we have

$$\begin{aligned} \int_S f(x)dx &= \int_{S \cap K_j} f(x)dx + \int_{S \cap \tilde{K}_j} f(x)dx + \int_{S - (K_j \cup \tilde{K}_j)} f(x)dx \\ &\le \int_{U_1} f(x)dx + \int_{U_2} f(x)dx + \int_{S - (K_j \cup \tilde{K}_j)} f(x)dx. \end{aligned} \tag{$\star$}$$

Let us show that as $j \to \infty$ the last integral goes to zero. Consider the set $S \cap Z$. First note that this set is closed. To see this note that $\overline{Z} \subset Z \cup \partial U$, because $\overline{Z} \subset \overline{U} = U \cup \partial U = U_1 \cup U_2 \cup Z \cup \partial U$, while no limit point of Z can belong to the open sets U_1, U_2 because Z does not intersect them. Thus $S \cap \overline{Z} = S \cap Z$, since $S \subset U$ does not intersect ∂U as U is open. So $S \cap Z$ is a closed subset of the compact set S; and thus it is compact too. In addition, $S \cap Z$ has measure zero as it is a subset of Z. So, for any given $\epsilon > 0$, we can cover $S \cap Z$ with a finite family of open cubes R_i such that $\sum_i |R_i| < \epsilon$. Note that for each of the cubes we also have $|R_i| < \epsilon$, so their diameter is also small. On the other hand, $S \cap Z \subset S$ has a positive distance from ∂U by Exercise 2.111. Hence, for small enough ϵ, all these cubes and therefore their union $V := \bigcup R_i$ lie in U. Also note that V is Jordan measurable, being the union of finitely many rectangles, and we have $|V| \leq \sum_i |R_i| < \epsilon$.

Next let us show that for large enough j we have $S - (K_j \cup \tilde{K}_j) \subset V$. Suppose to the contrary that for every j there is $x_j \in S - (K_j \cup \tilde{K}_j)$ such that $x_j \notin V$. Then the sequence x_j lies in the compact set S, so it has a subsequence, which we still denote by x_j, converging to $x^* \in S$. Note that x^* cannot belong to $S \cap Z \subset V$, because otherwise the open set V would contain every x_j for large enough j. Hence x^* must belong to

$$S - Z = S \cap (U_1 \cup U_2) = S \cap \bigcup_{j \geq 1} (K_j \cup \tilde{K}_j).$$

Thus x^* belongs to $S \cap (K_m \cup \tilde{K}_m)$ for some m; so $x^* \in K_{m+1}^\circ \cup \tilde{K}_{m+1}^\circ$ due to the properties of the sequences of compact sets $K_j, \tilde{K}_j$. But then, for large enough j, x_j must belong to the open set $K_{m+1}^\circ \cup \tilde{K}_{m+1}^\circ \subset K_j \cup \tilde{K}_j$, which is a contradiction.

Therefore, for large enough j we have

$$\int_{S-(K_j \cup \tilde{K}_j)} f(x)dx \leq \big|S - (K_j \cup \tilde{K}_j)\big| \sup_S f \leq |V| \sup_S f \leq \epsilon \sup_S f.$$

Note that f is bounded on S by our assumption. Combining this estimate with inequality ($\star$), we conclude that for every $\epsilon > 0$ we have

$$\begin{aligned} \int_S f(x)dx &\leq \int_{U_1} f(x)dx + \int_{U_2} f(x)dx + \epsilon \sup_S f \\ &\implies \int_S f(x)dx \leq \int_{U_1} f(x)dx + \int_{U_2} f(x)dx. \end{aligned}$$

Therefore, by taking supremum over S, we conclude that f is integrable over U, and

$$\int_U f(x)dx = \sup_S \int_S f(x)dx \leq \int_{U_1} f(x)dx + \int_{U_2} f(x)dx,$$

which combined with $(*)$ gives the desired equality for nonnegative f. Thus we have shown that for $f \geq 0$, the integrability of f over U is equivalent to its integrability over U_1, U_2, and the desired equality holds for integrals.

Now for a general function f, by the above argument, we know that the integrability of $f^{\pm}$ over U is equivalent to their integrability over U_1, U_2 (note that the boundedness of f on compact subsets of U implies the boundedness of $f^{\pm}$ on them). So, by definition, the integrability of f over U is equivalent to its integrability over U_1, U_2 too. Furthermore, in this case we have

$$\begin{aligned}
\int_U f(x)dx &= \int_U f^+(x)dx - \int_U f^-(x)dx \\
&= \int_{U_1} f^+(x)dx + \int_{U_2} f^+(x)dx - \int_{U_1} f^-(x)dx - \int_{U_2} f^-(x)dx \\
&= \int_{U_1} f^+(x)dx - \int_{U_1} f^-(x)dx + \int_{U_2} f^+(x)dx - \int_{U_2} f^-(x)dx \\
&= \int_{U_1} f(x)dx + \int_{U_2} f(x)dx,
\end{aligned}$$

as desired. ■

Next we prove another version of the change of variables theorem for improper integrals, in which we integrate over all of the domain of the change of variables map ϕ. This version is in particular useful when we integrate over open sets.

Change of Variables (Open Domains). *Suppose $U \subset \mathbb{R}^n$ is an open set. Let $\phi : U \to \mathbb{R}^n$ be a one-to-one C^1 map such that $D\phi(x)$ is an invertible matrix for every $x \in U$. Then $\phi(U)$ is an open set. Also, for every integrable function $f : \phi(U) \to \mathbb{R}$ the function $(f \circ \phi)\,|\det D\phi|$ is integrable over U, and we have*

$$\int_{\phi(U)} f(y)dy = \int_U (f \circ \phi)(x)\,|\det D\phi(x)|\,dx.$$

If $\int_{\phi(U)} f(y)dy$ is infinite, then the other integral is also infinite, and the equality remains valid.

Remark. Note that in contrast to the other version of change of variables theorem, here we do not assume that ϕ is defined on an open neighborhood of $\overline{U}$; so the assumptions of this theorem are weaker in this respect.

Proof. As we have seen in the proof of Theorem 8.53, using inverse function theorem we can show that $\phi(U)$ is open, and ϕ^{-1} is C^1. Thus in particular, both ϕ, ϕ^{-1} are locally Lipschitz. Let $S \subset U$ be an arbitrary compact Jordan measurable set. Then, by Theorem 8.53, $\phi(S)$ is a compact Jordan measurable subset of $\phi(U)$;

so f is Riemann integrable on $\phi(S)$. Hence $f \circ \phi$ is Riemann integrable on S, since $\phi : U \to \phi(U)$ is a homeomorphism and ϕ^{-1} is locally Lipschitz. In addition, $|\det D\phi|$ is a continuous function. Therefore $(f \circ \phi)\,|\det D\phi|$ is Riemann integrable on S too.

Let $S_k \subset U$ be a sequence of compact Jordan measurable sets given by Theorem 8.63, which satisfy $S_k \subset S_{k+1}^\circ$ and $U = \bigcup_{k \geq 1} S_k$. Then, by Theorem 8.53, $\phi(S_k)$ is a sequence of compact Jordan measurable subsets of $\phi(U)$. In addition we have $\phi(U) = \phi\big(\bigcup_{k\geq 1} S_k\big) = \bigcup_{k\geq 1} \phi(S_k)$. Furthermore, $\phi(S_k) \subset \phi(S_{k+1}^\circ)$, and as shown in the proof of Theorem 8.53, $\phi(S_{k+1}^\circ) \subset (\phi(S_{k+1}))^\circ$. So, by Theorem 8.61, $\sup_{k\geq 1} \int_{\phi(S_k)} |f(y)|dy < \infty$ and

$$\int_{\phi(U)} f(y)dy = \lim_{k\to\infty} \int_{\phi(S_k)} f(y)dy.$$

But, by the change of variables theorem, for every k we have

$$\int_{S_k} \big|(f \circ \phi)(x)\,|\det D\phi(x)|\big|\,dx = \int_{S_k} (|f| \circ \phi)(x)\,|\det D\phi(x)|\,dx = \int_{\phi(S_k)} |f(y)|dy.$$

So $\sup_{k\geq 1} \int_{S_k} \big|(f \circ \phi)(x)\,|\det D\phi(x)|\big|\,dx < \infty$, and therefore $(f \circ \phi)\,|\det D\phi|$ is integrable over U. In addition, by the change of variables theorem, we have

$$\begin{aligned}\int_U (f \circ \phi)(x)\,|\det D\phi(x)|\,dx &= \lim_{k\to\infty} \int_{S_k} (f \circ \phi)(x)\,|\det D\phi(x)|\,dx \\ &= \lim_{k\to\infty} \int_{\phi(S_k)} f(y)dy = \int_{\phi(U)} f(y)dy,\end{aligned}$$

as desired.

Now suppose $\int_{\phi(U)} f(y)dy$ is infinite, say $\int_{\phi(U)} f(y)dy = -\infty$ (the other case is similar). Then we have $\int_{\phi(U)} f^+(y)dy < \infty$ and $\int_{\phi(U)} f^-(y)dy = \infty$. Thus we must have $\sup_{k\geq 1} \int_{\phi(S_k)} f^-(y)dy = \infty$. So, by the change of variables theorem, we have

$$\sup_{k\geq 1} \int_{S_k} (f^- \circ \phi)(x)\,|\det D\phi(x)|\,dx = \sup_{k\geq 1} \int_{\phi(S_k)} f^-(y)dy = \infty.$$

Therefore

$$\int_U \big((f \circ \phi)(x)\,|\det D\phi(x)|\big)^-\,dx = \int_U (f^- \circ \phi)(x)\,|\det D\phi(x)|\,dx = \infty,$$

since the supremum of integrals over all compact Jordan measurable sets $S \subset U$ is no less than the supremum of integrals over the sequence of sets S_k. We can similarly show that $\int_U \big((f \circ \phi)(x)\,|\det D\phi(x)|\big)^+\,dx < \infty$. So the integral of $(f \circ \phi)\,|\det D\phi|$ over U is also equal to $-\infty$. ■

***Example* 8.69.** Let us integrate $f(x,y) = \frac{1}{\sqrt{x^2+y^2}}$ over the open half-disk

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

Consider the polar coordinates on $\mathbb{R}^2$ defined by

$$(x,y) = \phi(r,\theta) = (r\cos\theta, r\sin\theta),$$

for $r \geq 0$ and $0 \leq \theta < 2\pi$. Then it is easy to see that U is the image of the open rectangle $V = \{(r,\theta) : 0 < r < 1,\ \frac{\pi}{2} < \theta < \frac{3\pi}{2}\}$. We have

$$\det D\phi = \det \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} = r.$$

Hence by the change of variables theorem for open domains we get

$$\int_U \frac{1}{\sqrt{x^2+y^2}} dxdy = \int_V \frac{1}{\sqrt{r^2}} rdrd\theta = \int_V drd\theta = \int_0^1 dr \int_{\frac{\pi}{2}}^{\frac{3\pi}{2}} d\theta = \pi.$$

Note that we can apply Fubini's theorem to $\int_V drd\theta$, because the boundary of V has measure zero and thus $\int_V drd\theta = \int_{\overline{V}} drd\theta$. Also, notice that ϕ is not one-to-one at $r = 0$, and its derivative is not invertible there. Therefore the previous version of the change of variables theorem cannot be applied, since ϕ does not satisfy its hypotheses on any open set containing $\overline{V}$. In addition, note that f is not bounded near the origin, so it is not Riemann integrable in the proper sense.

***Example* 8.70.** We can also use the above theorem to deal with cases in which the domain of integration is not necessarily open. Let us again consider the polar coordinates on $\mathbb{R}^2$ defined by $(x,y) = \phi(r,\theta) = (r\cos\theta, r\sin\theta)$ for $r \geq 0$ and $0 \leq \theta < 2\pi$. Remember that $\det D\phi = r$. Suppose we want to integrate the bounded continuous function $f(x,y)$ on the disk $D = \{x^2+y^2 \leq a^2\} = \overline{B_a(0)}$. This disk in the (x,y)-plane is the image of the rectangle $R = \{0 \leq r \leq a,\ 0 \leq \theta < 2\pi\}$ in the (r,θ)-plane. Let us consider the open rectangle

$$R^\circ = \{0 < r < a,\ 0 < \theta < 2\pi\}.$$

Then $\phi(R^\circ)$ is equal to $B_a(0) - \{(x,0) : 0 \leq x \leq a\}$. Now we can apply the change of variables theorem for open domains to get

$$\int_{\phi(R^\circ)} f(x,y)dxdy = \int_{R^\circ} f(r,\theta)\, rdrd\theta.$$

However, we have $R = R^\circ \sqcup \partial R$ and $D = \phi(R^\circ) \sqcup A$, where

$$A = \{(x,0) : 0 \leq x \leq a\} \cup \partial B_a(0).$$

Note that both $\partial R, A$ are Jordan measurable sets with measure zero; therefore integrals over them are zero. Hence we have (noting that all integrals are in the proper sense, since f is bounded and continuous)

$$\begin{aligned}\int_D f(x,y)dxdy &= \int_{\phi(R^\circ)} f(x,y)dxdy + \int_A f(x,y)dxdy \\ &= \int_{\phi(R^\circ)} f(x,y)dxdy = \int_{R^\circ} f(r,\theta)\, rdrd\theta \\ &= \int_{R^\circ} f(r,\theta)\, rdrd\theta + \int_{\partial R} f(r,\theta)\, rdrd\theta \\ &= \int_R f(r,\theta)\, rdrd\theta.\end{aligned}$$

We can prove similar results for spherical and cylindrical coordinates on $\mathbb{R}^3$ too.

Chapter 9

Integration over Surfaces and Rectifiable Sets

9.1 Integration over a Surface Patch

In this chapter we want to define the integral of a function over a "k-dimensional surface" in $\mathbb{R}^n$. Before going into the details of what a "k-dimensional surface" is, let us notice that we expect them to have measure zero in $\mathbb{R}^n$. For example, a curve has measure zero in $\mathbb{R}^2$. Therefore if we try to integrate a function over a surface by the methods of the previous chapter we would get zero. So we need a new definition of integral over surfaces which gives sensible results. Let us start by defining a simple case of "k-dimensional surfaces".

Definition 9.1. Let $V \subset \mathbb{R}^k$ be an open set, and $\phi : V \to \mathbb{R}^n$ be a function, where $k < n$. Suppose

(i) ϕ is one-to-one.

(ii) ϕ is a C^1 function, and $D\phi(x)$ has rank k at every $x \in V$.

(iii) $\phi^{-1} : \phi(V) \to V$ is continuous.

Then we say $\phi(V) \subset \mathbb{R}^n$ is a ***k*-dimensional surface patch**, or simply a ***k*-surface patch**. The map ϕ is called a **parametrization** of the k-surface patch $\phi(V)$, and the function $\sqrt{\det(D\phi^\intercal D\phi)}$ is called the **volume factor**.

Remark. A 1-surface patch is called a **curve**. In addition, when $k = 1, 2$, the volume factor may also be called the *length factor* or the *area factor*, respectively (the motivation for these names will be clarified shortly). We will also see that $\det(D\phi^\intercal D\phi)$ is nonnegative; so its square root is real.

Remark. When the open set V is bounded, an easy way to check that ϕ^{-1} is continuous is to show that ϕ extends to a continuous one-to-one function defined on the compact set $\overline{V}$.

Remark. The requirements that $D\phi$ has full rank k, and ϕ^{-1} is continuous, are to ensure that the k-surface patch has no *singularity*. They also allow us to prove that integration over a k-surface patch does not depend on the parametrization, as we see below. In the next section, we will also see how to integrate over surfaces which have some singularities.

Theorem 9.2. *The dimension of a surface patch does not depend on its parametrization. In other words, if $S = \phi(V) = \psi(U)$ is a surface patch in $\mathbb{R}^n$, where $V \subset \mathbb{R}^k$ and $U \subset \mathbb{R}^p$ are open sets, and ϕ, ψ are parametrizations of S, then we must have $p = k$. In addition, we have*

(i) $\phi^{-1} \circ \psi : U \to V$ *is a C^1 function.*

(ii) $\phi^{-1} : \phi(V) \to V$ *is locally Lipschitz.*

Remark. The function $\phi^{-1}\circ\psi$ is called a **change of coordinates** or a **transition function**.

Proof. To simplify the notation let $g := \phi^{-1} \circ \psi : U \to V$. Note that g and its inverse $g^{-1} = \psi^{-1} \circ \phi : V \to U$ are one-to-one and onto continuous functions. If we show that g, g^{-1} are differentiable, then we have $Dg^{-1}(g(x))\, Dg(x) = I_p$, because $g^{-1}(g(x)) = x$ for $x \in U$. Therefore $Dg(x)$ defines a one-to-one linear map from $\mathbb{R}^p$ to $\mathbb{R}^k$; hence $p \le k$. Similarly we have $Dg(g^{-1}(y))\, Dg^{-1}(y) = I_k$ for $y \in V$. Let $y = g(x)$. Then we get $Dg(x)\, Dg^{-1}(g(x)) = I_k$. Thus $Dg(x)$ defines an onto linear map from $\mathbb{R}^p$ to $\mathbb{R}^k$; hence $p \ge k$. Therefore we must have $p = k$, as desired.

To show that g, g^{-1} are C^1, it suffices to show that they are C^1 on an open neighborhood of each point in their domains. We show this for g; the case of g^{-1} is similar. Let $x_0 \in V$. We know that $D\phi(x_0)$ is an $n \times k$ matrix with rank k. Hence it has k linearly independent rows. Let $P : \mathbb{R}^n \to \mathbb{R}^k$ be the projection onto the coordinates determined by the indices of the k linearly independent rows of $D\phi(x_0)$ (note that P may depend on x_0). Now consider $P \circ \phi : V \to \mathbb{R}^k$. We have $D(P\circ\phi)(x_0) = PD\phi(x_0)$, since P is a linear map (we have denoted the matrix of P simply by P). But it is easy to see that $PD\phi(x_0)$ is a $k \times k$ matrix whose rows are the linearly independent rows of $D\phi(x_0)$. Hence $PD\phi(x_0)$ is an invertible matrix. Therefore, by the inverse function theorem, there is an open neighborhood $V_{x_0} \subset V$ of x_0 such that $P \circ \phi$ is one-to-one on it, $P \circ \phi\,(V_{x_0})$ is open, and $\big(P \circ \phi\,|_{V_{x_0}}\big)^{-1}$ is a C^1 function on $P \circ \phi\,(V_{x_0})$.

Now suppose we want to show that $g = \phi^{-1}\circ\psi$ is C^1 on an open neighborhood of $y_0 \in U$. Let $x_0 = g(y_0)$, and consider V_{x_0}. Then $\phi(V_{x_0})$ is an open subset of S, since ϕ^{-1} is continuous. Hence $U_{y_0} := \psi^{-1}(\phi(V_{x_0}))$ is an open neighborhood of y_0, since ψ is continuous. Then for every $y \in U_{y_0}$ there is $x \in V_{x_0}$ such that $\psi(y) = \phi(x)$, because $\psi(U_{y_0}) = \phi(V_{x_0})$. Therefore

$$P(\psi(y)) = P(\phi(x)) \in P \circ \phi\,(V_{x_0}).$$

Hence $P(\psi(y))$ is in the domain of the C^1 function $\big(P\circ\phi\,|_{V_{x_0}}\big)^{-1}$, and we have

$$\begin{aligned}\big(P\circ\phi\,|_{V_{x_0}}\big)^{-1}\big(P(\psi(y))\big) &= \big(P\circ\phi\,|_{V_{x_0}}\big)^{-1}\big(P(\phi(x))\big) = x \qquad (\text{since } x\in V_{x_0})\\ &= \phi^{-1}(\phi(x)) = \phi^{-1}(\psi(y)) = g(y).\end{aligned}$$

Thus on U_{y_0} we have $g = \big(P\circ\phi\,|_{V_{x_0}}\big)^{-1}\circ P\circ\psi$. In other words, on U_{y_0}, g is equal to the composition of several C^1 functions. Hence g is C^1 on U_{y_0}, as desired.

Finally, let us show that ϕ^{-1} is locally Lipschitz. Let $z_0=\phi(x_0)$. Note that there is r such that $B_r(z_0)\cap S\subset\phi(V_{x_0})$, since $\phi(V_{x_0})$ is an open subset of S. Now for $z\in B_r(z_0)\cap S$ there is a unique $x\in V_{x_0}$ such that $z=\phi(x)$. In addition, $P(\phi(x))$ is in the domain of $\big(P\circ\phi\,|_{V_{x_0}}\big)^{-1}$. Hence we have

$$\phi^{-1}(z) = x = (P\circ\phi)^{-1}\big(P(\phi(x))\big) = (P\circ\phi)^{-1}\big(P(z)\big).$$

Thus on $B_r(z_0)\cap S$ we have $\phi^{-1} = \big(P\circ\phi\,|_{V_{x_0}}\big)^{-1}\circ P$. So, locally, ϕ^{-1} is equal to the composition of two C^1 functions, which are also locally Lipschitz; therefore ϕ^{-1} is locally Lipschitz too. ■

Next we want to define the integral of a function over a k-surface patch $S\subset\mathbb{R}^n$. Suppose we have $S=\phi(V)$. Let $f:S\to\mathbb{R}$ be a function which we want to integrate over S. To motivate the following definition, we pursue the idea of partitioning S into small pieces, and then we use the corresponding Riemann sums to approximate the desired integral. To simplify this heuristic argument let us assume that V is an open rectangle in $\mathbb{R}^k$. An easy way for partitioning S is to consider a partition of V, and map it by ϕ into S. Let $\{R_\alpha\}$ be a partition of V. Then we can consider $\{\phi(R_\alpha)\}$ as a partition of S. If we choose some tags $x_\alpha\in R_\alpha$ then the points $\phi(x_\alpha)\in\phi(R_\alpha)$ can play the role of tags for the partition $\{\phi(R_\alpha)\}$. Now we can form a Riemann sum that approximates the integral of f over S as follows:

$$\sum_\alpha f(\phi(x_\alpha))\,|\phi(R_\alpha)|, \tag{$*$}$$

where $|\phi(R_\alpha)|$ is the "k-dimensional volume" of the k-surface patch $\phi(R_\alpha)$ in $\mathbb{R}^n$.

Hence we need to somehow approximate the volume $|\phi(R_\alpha)|$ without integration (because we want to use these values to define the integral over a k-surface patch!) Let us assume that the mesh of the partition $\{R_\alpha\}$ is so small that over the rectangle R_α, ϕ is almost equal to the linear function defined by its derivative $A_\alpha := D\phi(x_\alpha)$. Then the volume $|\phi(R_\alpha)|$ is almost equal to the volume of $A_\alpha(R_\alpha)$, which is a k-dimensional parallelepiped in $\mathbb{R}^n$. If we had $k=n$ then the volume of $A_\alpha(R_\alpha)$ would have been equal to $|\det(A_\alpha)|\,|R_\alpha|$, as we have seen in the last chapter. However in the actual case where $k<n$, the matrix A_α is not a square matrix, and we have not even defined the volume of a parallelepiped in higher-dimensional ambient space.

But intuitively we know that the volume does not change under rigid motions; in particular, an orthogonal linear map cannot change the volume of k-dimensional parallelepipeds.

Now note that by polar decomposition of matrices we have $A_\alpha = O_\alpha P_\alpha$, where O_α is an $n \times k$ orthogonal matrix (i.e. $O_\alpha^{\intercal} O_\alpha = I_k$, which implies that O_α preserves the distances), and P_α is a $k \times k$ positive matrix (i.e. it is a symmetric matrix with nonnegative eigenvalues). Since we intuitively know that O_α does not change the volume, we must have

$$|A_\alpha(R_\alpha)| = |O_\alpha(P_\alpha(R_\alpha))| = |P_\alpha(R_\alpha)| = |\det(P_\alpha)|\,|R_\alpha|,$$

where the last equality holds because P_α is a linear map on $\mathbb{R}^k$, and R_α is a rectangle in $\mathbb{R}^k$. However we also have

$$A_\alpha^{\intercal} A_\alpha = P_\alpha^{\intercal} O_\alpha^{\intercal} O_\alpha P_\alpha = P_\alpha^{\intercal} I_k P_\alpha = P_\alpha^{\intercal} P_\alpha = P_\alpha P_\alpha = P_\alpha^2.$$

Therefore

$$(\det P_\alpha)^2 = \det(P_\alpha^2) = \det(A_\alpha^{\intercal} A_\alpha) = \det\big(D\phi(x_\alpha)^{\intercal} D\phi(x_\alpha)\big).$$

Thus the volume of $A_\alpha(R_\alpha)$ is equal to the volume of R_α times the volume factor at the point x_α. (We can also see that $\det(D\phi^{\intercal} D\phi)$ is nonnegative.) Hence we can finally approximate the Riemann sum $(*)$ as follows

$$\begin{aligned}\sum_\alpha f(\phi(x_\alpha))\,|\phi(R_\alpha)| &\approx \sum_\alpha f(\phi(x_\alpha))\,|A_\alpha(R_\alpha)| \\ &= \sum_\alpha f(\phi(x_\alpha))\,|\det(P_\alpha)|\,|R_\alpha| \\ &= \sum_\alpha f(\phi(x_\alpha))\,\sqrt{\det\big(D\phi(x_\alpha)^{\intercal} D\phi(x_\alpha)\big)}\,|R_\alpha|.\end{aligned}$$

But the last expression is just a Riemann sum for the function $(f\circ\phi)\sqrt{\det(D\phi^{\intercal} D\phi)}$ over the flat domain V; and this can be used to define the integral of f over the k-surface patch S. Hence we arrive at the following definition.

Definition 9.3. Let $S = \phi(V)$ be a k-surface patch in $\mathbb{R}^n$, where $V \subset \mathbb{R}^k$ is an open set, and ϕ is a parametrization of S. Consider the function $f : S \to \mathbb{R}$. We say f is integrable over S if $(f\circ\phi)\sqrt{\det(D\phi^{\intercal} D\phi)}$ is integrable over V. And in this case we define the integral of f over S to be

$$\int_S f\,d\sigma := \int_V f(\phi(x))\,\sqrt{\det\big(D\phi(x)^{\intercal} D\phi(x)\big)}\,dx.$$

Notation. The notation $d\sigma$ in the integral $\int_S f\,d\sigma$ is incorporated to indicate that we are integrating over a surface. Another common notation for $d\sigma$ is dS, which should not be confused with the notation that we used for the surface S over which we are integrating. Also, when $k = 1$, we usually use ds instead of $d\sigma$.

Remark. We can similarly define the integral of vector-valued functions over a surface patch. Then it would easily follow that the integrability of such a function is equivalent to the integrability of its components, and its integral can be computed componentwise.

Theorem 9.4. *The integrability and the integral of a function over a k-surface patch do not depend on the parametrization.*

Proof. Let $S = \phi(V) = \psi(U)$ be a k-surface patch in $\mathbb{R}^n$, where $V, U \subset \mathbb{R}^k$ are open sets, and ϕ, ψ are parametrizations of S. Let $f : S \to \mathbb{R}$. To simplify the notation we will use the following convention throughout this proof:

$$J_\phi = \sqrt{\det(D\phi^\intercal D\phi)}, \qquad\qquad J_\psi = \sqrt{\det(D\psi^\intercal D\psi)}.$$

We need to show that the integrability of $(f \circ \phi)J_\phi$ over V is equivalent to the integrability of $(f \circ \psi)J_\psi$ over U, and when they are integrable, their integrals are equal. Note that $g := \phi^{-1} \circ \psi : U \to V$ and its inverse $g^{-1} = \psi^{-1} \circ \phi : V \to U$ are one-to-one and onto C^1 functions. Thus Dg, Dg^{-1} are invertible matrices. Hence by the change of variables theorem, the integrability of $(f \circ \phi)J_\phi$ over $V = g(U)$ implies the integrability of

$$(f \circ \phi \circ g)\sqrt{\det(D\phi(g)^\intercal D\phi(g))}\,|\det(Dg)| \tag{$*$}$$

over U, and the two integrals are equal.

Now note that we have $\phi \circ g = \phi \circ \phi^{-1} \circ \psi = \psi$. Hence $f \circ \phi \circ g = f \circ \psi$. In addition, we have $D\psi = D\phi(g)\,Dg$. Therefore

$$\begin{aligned} J_\psi^2 = \det((D\psi)^\intercal D\psi) &= \det\big((D\phi(g)\,Dg)^\intercal D\phi(g)\,Dg\big) \\ &= \det\big(Dg^\intercal D\phi(g)^\intercal D\phi(g)\,Dg\big) \\ &= \det(Dg^\intercal)\det(D\phi(g)^\intercal D\phi(g))\det(Dg) \\ &= \det(D\phi(g)^\intercal D\phi(g))\det(Dg)^2, \end{aligned}$$

since a square matrix and its transpose have the same determinant. Hence the function $(*)$ is equal to

$$(f \circ \psi)\sqrt{\det(D\phi(g)^\intercal D\phi(g))\det(Dg)^2} = (f \circ \psi)\sqrt{J_\psi^2} = (f \circ \psi)J_\psi.$$

Thus we have shown that the integrability of $(f \circ \phi)J_\phi$ over $V = g(U)$ implies the integrability of $(f \circ \psi)J_\psi$ over U, and their integrals are equal. Conversely, by repeating the above argument using g^{-1} instead of g, we can show that the integrability of $(f \circ \psi)J_\psi$ implies the integrability of $(f \circ \phi)J_\phi$. ■

Definition 9.5. Let S be a k-surface patch in $\mathbb{R}^n$. The **(k-dimensional) volume** of S is

$$\mathrm{vol}(S) := \int_S 1\, d\sigma,$$

provided that the integral exists. When $k = 1, 2$, the volume is called the **length** or the **area**, respectively.

Remark. Note that since the volume factor is positive, if the above integral does not exist then its value will be ∞.

Remark. If S has finite volume, then any bounded continuous function like f is integrable over S. Because, in this case, for a parametrization $\phi : V \to S$, the function $\sqrt{\det\big(D\phi^{\mathsf{T}} D\phi\big)}$ is integrable over V due to the finiteness of the volume. Thus the function $f \circ \phi \sqrt{\det\big(D\phi^{\mathsf{T}} D\phi\big)}$ is also integrable over V, since $f \circ \phi$ is bounded and continuous (see the remarks after Theorem 8.60 and Example 8.66).

Remark. In Definition 6.52 we defined rectifiable curves and their lengths. It is easy to see that by Theorem 6.54, a curve with a C^1 parametrization (i.e. a 1-surface patch) is also rectifiable, and both the above definition and Definition 6.52 give the same value for its length.

***Example* 9.6.** Let $V \subset \mathbb{R}^k$ be an open set, and $f : V \to \mathbb{R}$ be a C^1 function. Then the graph of f, i.e. the set

$$S = \{(x, f(x)) : x \in V\},$$

is a k-surface patch in $\mathbb{R}^{k+1}$ with $\phi(x) := (x, f(x))$ as a parametrization. To see this note that ϕ is obviously one-to-one and C^1. Also, it is apparent from the following formula for $D\phi$ that it has rank k. In addition, ϕ^{-1} is just the projection on the first k components; so it is continuous. Now let us compute the volume of S. We have

$$D\phi = \begin{bmatrix} I_k \\ Df \end{bmatrix} \implies D\phi^{\mathsf{T}} D\phi = \begin{bmatrix} I & Df^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} I \\ Df \end{bmatrix} = I + Df^{\mathsf{T}} Df.$$

Note that Df^{T} is a (column) vector. Now note that for any vector a, the matrix aa^{T} has rank one, and its action on any vector x is

$$(aa^{\mathsf{T}})x = a(a^{\mathsf{T}}x) = a(a \cdot x) = (a \cdot x)a.$$

Hence we have $(I + aa^{\mathsf{T}})x = x + (a \cdot x)a$. Thus for $x = a$ we get

$$(I + aa^{\mathsf{T}})a = a + (a \cdot a)a = (1 + |a|^2)a.$$

And if x is orthogonal to a we have $(I + aa^{\mathsf{T}})x = x + 0a = x$. Therefore the $k \times k$ matrix $I + aa^{\mathsf{T}}$ has (at least) $k - 1$ eigenvalues equal to 1, and an eigenvalue equal to $1+|a|^2$. (Note that if $a=0$ then the eigenvalue $1+|a|^2$ is also equal to 1.) Hence we have $\det(I+aa^{\intercal})=1+|a|^2$, since the determinant is equal to the product of all eigenvalues. So

$$\det\big(D\phi(x)^{\intercal}D\phi(x)\big)=\det\big(I+Df^{\intercal}Df\big)=1+|Df|^2.$$

Thus the volume of S, i.e. the volume of the graph of f, is

$$\int_S 1\,d\sigma=\int_V\sqrt{\det\big(D\phi(x)^{\intercal}D\phi(x)\big)}\,dx=\int_V\sqrt{1+|Df|^2}\,dx.$$

***Example* 9.7.** Let us provide a parametrization of a two-dimensional hemisphere in $\mathbb{R}^3$ with radius r, and compute its area. Consider the parametrization $\phi:(0,\pi)\times(0,\pi)\to\mathbb{R}^3$ defined by

$$(\theta,\varphi)\mapsto(r\sin\theta\cos\varphi,r\sin\theta\sin\varphi,r\cos\theta).$$

It is apparent that $\phi_1^2+\phi_2^2+\phi_3^2=r^2$, and $\phi_2>0$; so ϕ parametrizes part of a hemisphere. On the other hand, it is easy to show that each point on that hemisphere is the image of a uniquely determined (θ,φ). We also have

$$D\phi=\begin{bmatrix}r\cos\theta\cos\varphi & -r\sin\theta\sin\varphi\\ r\cos\theta\sin\varphi & r\sin\theta\cos\varphi\\ -r\sin\theta & 0\end{bmatrix}.$$

Note that $D\phi$ has full rank, since $\sin\theta,\sin\varphi$ never vanish on the domain of ϕ. Hence the volume factor is

$$\big(\det(D\phi^{\intercal}D\phi)\big)^{1/2}=\left(\det\begin{bmatrix}r^2 & 0\\ 0 & r^2\sin^2\theta\end{bmatrix}\right)^{1/2}=r^2\sin\theta.$$

Therefore the area of a two-dimensional hemisphere with radius r is

$$\begin{aligned}\int_{(0,\pi)\times(0,\pi)}r^2\sin\theta\,d\theta d\varphi&=\int_{[0,\pi]\times[0,\pi]}r^2\sin\theta\,d\theta d\varphi\\&=\int_0^{\pi}\int_0^{\pi}r^2\sin\theta\,d\theta d\varphi=2\pi r^2.\end{aligned}$$

Note that the second equality holds by Fubini's theorem. And the first equality follows from the fact that the boundary of the closed rectangle $[0,\pi]\times[0,\pi]$ has measure zero; so its inclusion does not alter the integral of the integrable function $r^2\sin\theta$. More precisely, the function $r^2\sin\theta$ on $[0,\pi]^2$ is a.e. equal to the function which equals $r^2\sin\theta$ on $(0,\pi)^2$ and vanishes on its boundary. Hence by Theorem 8.37 their integrals on $[0,\pi]^2$ are equal. Note that both these functions are integrable, since they are bounded and continuous a.e.

9.2 Integration over a Rectifiable Set

In the last section we have computed the area of a two-dimensional hemisphere. Intuitively, we know that we can double that amount to obtain the area of the two-dimensional sphere. However, note that since we are integrating over open sets, then we would have excluded a great circle of the sphere from the domain of integration. In addition, it is not possible to cover the whole sphere with a single 2-surface patch, or to cover it with several nonoverlapping (open) 2-surface patches.

Thus we need another way to rigorously compute the volume of the sphere, or more generally, to compute the integral of a function over a sphere. The good news is that the great circle has "two-dimensional measure zero" (note that it resides in $\mathbb{R}^3$). So its exclusion from the domain of integration should not affect the value of the integral. In this section, we first make the idea of lower-dimensional zero measure precise. Then we will use this notion to define integration over more general "surfaces".

Remember that we say $A \subset \mathbb{R}^n$ has measure zero if for every $\epsilon > 0$ there exist a countable family of open cubes $\{Q_i\}$ such that $A \subset \bigcup_{i\geq 1} Q_i$, and

$$\sum_{i\geq 1} |Q_i| < \epsilon.$$

Let l_i be the length of the edges of Q_i. Then the above inequality becomes $\sum_{i\geq 1} l_i^n < \epsilon$. For generalization, it is better to express this inequality in terms of the diameter of Q_i, i.e. the maximum distance between points of Q_i. We can easily see that the diameter of Q_i is $\operatorname{diam} Q_i = l_i\sqrt{n}$. Then we can say A has measure zero if and only if for every $\epsilon > 0$ there exist a countable family of open cubes covering it such that

$$\sum_{i\geq 1} (\operatorname{diam} Q_i)^n < \epsilon.$$

Note that we have absorbed the factor $n^{-\frac{n}{2}}$ into ϵ, since $n^{-\frac{n}{2}}$ is constant, and ϵ is arbitrary. Now we use the idea of replacing the power n with k in the above inequality to define subsets of $\mathbb{R}^n$ which have "k-dimensional measure zero".

Definition 9.8. Let $A \subset \mathbb{R}^n$, and suppose $k < n$. We say A has ***k*-measure zero** if for every $\epsilon > 0$ there exist a countable family of open cubes in $\mathbb{R}^n$, $\{Q_i\}$, such that $A \subset \bigcup_{i\geq 1} Q_i$, and

$$\sum_{i\geq 1} (\operatorname{diam} Q_i)^k < \epsilon.$$

Remark. For compatibility, if a subset of $\mathbb{R}^n$ has measure zero in the ordinary sense, we may also say that it has n-measure zero.

Remark. An obvious consequence of the definition is that if A has k-measure zero and $B \subset A$, then B has k-measure zero too.

Remark. Suppose $k < l \leq n$. Then if A has k-measure zero, it also has l-measure zero. To see this suppose $\epsilon < 1$, and $\{Q_i\}$ is a countable family of open cubes coveting A such that $\sum_{i\geq 1}(\operatorname{diam} Q_i)^k < \epsilon$. Then for each i we have $\operatorname{diam} Q_i < \epsilon^{\frac{1}{k}} < 1$. Hence $(\operatorname{diam} Q_i)^l \leq (\operatorname{diam} Q_i)^k$. Therefore we also have

$$\sum_{i\geq 1}(\operatorname{diam} Q_i)^l \leq \sum_{i\geq 1}(\operatorname{diam} Q_i)^k < \epsilon,$$

as desired.

Theorem 9.9. *Let $\{A_j\}$ be a countable family of subsets of $\mathbb{R}^n$ that have k-measure zero. Then $\bigcup_j A_j$ has k-measure zero. In particular, every countable subset of $\mathbb{R}^n$ has k-measure zero for every $1 \leq k \leq n$.*

Proof. Let $\epsilon > 0$ be given. Then we can cover A_j with a countable family of open cubes $\{Q_{ji}\}_{i\geq 1}$ such that

$$\sum_{i\geq 1}(\operatorname{diam} Q_{ji})^k < \frac{\epsilon}{2^j}.$$

Then $\{Q_{ji}\}_{i,j\geq 1}$ is a countable family of open cubes that covers $\bigcup_j A_j$, and

$$\sum_{i,j\geq 1}(\operatorname{diam} Q_{ji})^k < \sum_{j\geq 1}\frac{\epsilon}{2^j} \leq \epsilon.$$

The final statement of the theorem follows from the trivial fact that a set with one element has k-measure zero. ■

Theorem 9.10. *Suppose $k \leq n, m$, and $A \subset \mathbb{R}^n$ has k-measure zero. Also suppose $F : A \to \mathbb{R}^m$ is locally Lipschitz. Then $F(A)$ has k-measure zero.*

Proof. The proof is an easy modification of the proof of Theorem 8.51, which considers the case $k = m = n$. Every $a \in A$ has a neighborhood $B_r(a)$ such that F is Lipschitz on $A \cap B_r(a)$. Then the family $\{B_r(a) : a \in A\}$ is an open covering of A. By Theorem 11.57, every open covering of a subset of $\mathbb{R}^n$ has a countable subcovering. Let us denote this countable subcovering by $\{B_i\}$. Then we have

$$F(A) = \bigcup_{i\geq 1} F(A \cap B_i) = \bigcup_{i\geq 1} F|_{A\cap B_i}(A \cap B_i).$$

But $F|_{A\cap B_i}$ is Lipschitz, and $A\cap B_i$ has k-measure zero. Therefore it suffices to prove the theorem for Lipschitz maps. Because then it follows that each $F|_{A\cap B_i}(A \cap B_i)$

has k-measure zero. And as $F(A)$ is the union of countably many sets of k-measure zero, it also has k-measure zero as desired.

So we assume that F is Lipschitz, and A has k-measure zero. We want to show that $F(A)$ has k-measure zero. For any $\epsilon > 0$ there is a family of open cubes $Q_i \subset \mathbb{R}^n$ such that $A \subset \bigcup_{i\geq 1} Q_i$, and $\sum_{i\geq 1}(\operatorname{diam} Q_i)^k < \epsilon$. But the diameter of $A \cap Q_i$, i.e. the maximum distance between its points, is less than or equal to the diameter of Q_i. Hence the diameter of $F(A \cap Q_i)$ is at most $K\operatorname{diam} Q_i$, where K is a constant satisfying $|F(x) - F(y)| \leq K|x - y|$ for any two points x, y. Therefore $F(A \cap Q_i) \subset R_i$, where R_i is an open cube whose edges are of length $3K\operatorname{diam} Q_i$, and is centered at some point $z \in F(A \cap Q_i)$. Because for any other point $z' \in F(A \cap Q_i)$ we have $|z' - z| \leq K\operatorname{diam} Q_i$; so the absolute value of each coordinate of $z' - z$ is less than or equal to $K\operatorname{diam} Q_i$, which is strictly less than $\frac{3}{2}K\operatorname{diam} Q_i$. Now, $\operatorname{diam} R_i = 3\sqrt{n}K\operatorname{diam} Q_i$. We also have

$$F(A) = \bigcup_{i\geq 1} F(A \cap Q_i) \subset \bigcup_{i\geq 1} R_i,$$

and

$$\sum_{i\geq 1}(\operatorname{diam} R_i)^k = 3^k n^{\frac{k}{2}} K^k \sum_{i\geq 1}(\operatorname{diam} Q_i)^k < 3^k n^{\frac{k}{2}} K^k \epsilon.$$

Thus as ϵ is arbitrary, $F(A)$ has k-measure zero. ■

***Example* 9.11.** Remember that if we regard $\mathbb{R}^{k-1}$ as the subset of $\mathbb{R}^k$ on which $x_k = 0$, then $\mathbb{R}^{k-1}$ has measure zero in $\mathbb{R}^k$. Thus any subset of $\mathbb{R}^{k-1}$ has k-measure zero in $\mathbb{R}^k$. Hence the above theorem implies that the image of an open subset of $\mathbb{R}^{k-1}$ under a C^1 function into some $\mathbb{R}^n$ has k-measure zero, since C^1 functions are locally Lipschitz. In particular, $(k-1)$-surface patches have k-measure zero.

We next define a more general class of "k-dimensional surfaces" in $\mathbb{R}^n$, which are also allowed to have some singularities.

Definition 9.12. We say $S \subset \mathbb{R}^n$ is ***k*-rectifiable** if S can be written as a union of pairwise disjoint sets

$$S = A \sqcup S_1 \sqcup \cdots \sqcup S_m,$$

such that each S_j is a k-surface patch which is an open set in S, and A has k-measure zero.

Remark. We can extend the above definition, and allow countably many k-surface patches. However, the simpler case of finitely many k-surface patches is still quite general, and provides many interesting examples.

Next let us introduce one of the most important examples of rectifiable sets.

Definition 9.13. We say $M \subset \mathbb{R}^n$ is a ***k*-dimensional manifold** if for every point $a \in M$ there exists an open ball $B_r(a)$ such that $B_r(a) \cap M$ is a k-surface patch in $\mathbb{R}^n$. (Note that $r > 0$, and can depend on the point a.)

Theorem 9.14. *Every compact manifold is a rectifiable set.*

Proof. Let M be a compact k-dimensional manifold in $\mathbb{R}^n$. We know that for every $a \in M$ there is an open ball $B_r(a)$ such that $B_r(a) \cap M$ is a k-surface patch in $\mathbb{R}^n$. Now note that for each a there is an open set $V_a \subset \mathbb{R}^k$, and a parametrization $\phi_a : V_a \to \mathbb{R}^n$ such that

$$\phi_a(V_a) = B_r(a) \cap M.$$

Consider an open rectangle R_a containing $\phi_a^{-1}(a)$, such that $\overline{R}_a \subset V_a$. Then the collection $\{\phi_a(R_a)\}_{a \in M}$ covers M. Also, due to the continuity of ϕ_a^{-1}, each $\phi_a(R_a)$ is open in $B_r(a) \cap M$, hence it is open in M. Therefore, by compactness of M, finitely many of these open sets, namely $\phi_1(R_1), \dots, \phi_m(R_m)$, cover M (note that we suppressed the notation a_j to j for simplicity). Each $\phi_j(R_j)$ is a k-surface patch, with parametrization $\phi_j|_{R_j}$. However, they may have nonempty intersections.

To construct disjoint k-surface patches out of $\phi_1(R_1), \dots, \phi_m(R_m)$ we proceed as follows. Set $M_1 := \phi_1(R_1)$. Then set $M_2 := \phi_2(R_2) - \phi_1(\overline{R}_1)$. Note that $\phi_1(\overline{R}_1)$ is closed, since $\overline{R}_1$ is compact. Hence M_2 is open in M. Also note that $M_1 \cap M_2 = \emptyset$, because $M_1 \subset \phi_1(\overline{R}_1)$. We continue inductively, and set

$$M_{j+1} := \phi_{j+1}(R_{j+1}) - \big(\phi_1(\overline{R}_1) \cup \cdots \cup \phi_j(\overline{R}_j)\big).$$

Note that each M_j is open in M, and by definition, they are pairwise disjoint. (Some of the M_j's might be empty, in which case we simply discard them). Furthermore, M_j is a k-surface patch with parametrization $\phi_j|_{W_j}$, where $W_j := \phi_j^{-1}(M_j)$.

Now note that $\overline{R}_j = R_j \sqcup \partial R_j$, since R_j is open. Hence we have

$$\phi_j(\overline{R}_j) = \phi_j(R_j) \sqcup \phi_j(\partial R_j),$$

since ϕ_j is one-to-one. On the other hand, ∂R_j has measure zero in $\mathbb{R}^k$, and ϕ_j is locally Lipschitz (since it is C^1); therefore

$$\tilde{A} := \phi_1(\partial R_1) \cup \cdots \cup \phi_m(\partial R_m)$$

has k-measure zero. In addition, we have $A := M - \bigcup_{j \le m} M_j \subset \tilde{A}$. Because every $a \in M$ belongs to some $\phi_j(R_j)$. Let j be the smallest number for which this happens. Now if $a \notin M_j$ then for some $i < j$ we must have $a \in \phi_i(\partial R_i)$, since $a \notin \phi_i(R_i)$ by our assumption about j. Thus, in particular, A has k-measure zero.

Hence we have shown that the k-dimensional manifold M can be written as a union of pairwise disjoint k-surface patches M_j, which are open in M, and a set A with k-measure zero. Therefore M is a k-rectifiable set, as desired. ■

Remark. In the above proof, notice that the volume factor of each M_j is bounded, because ϕ_j is C^1 on the compact set $\overline{R}_j \supset \phi_j^{-1}(M_j)$. Furthermore, the open sets W_j are Jordan measurable. To see this note that $\partial W_j \subset \overline{W}_j \subset \overline{R}_j$, and

$$\phi_j(\partial W_j) \cap M_j = \phi_j(\partial W_j) \cap \phi_j(W_j) = \emptyset,$$

since ϕ_j is one-to-one. Furthermore, we must also have $\phi_j(\partial W_j) \cap M_i = \emptyset$ for $i \neq j$. Because otherwise $\phi_j^{-1}(M_i)$ would intersect ∂W_j. But $\phi_j^{-1}(M_i)$ is an open set, since M_i is open in M. Therefore $\phi_j^{-1}(M_i)$ would intersect W_j; which implies that M_i would intersect $\phi_j(W_j) = M_j$, contradicting their disjointness. Hence $\phi_j(\partial W_j) \subset A$. However, A has k-measure zero; so $\phi_j(\partial W_j)$ has k-measure zero too. On the other hand, ϕ_j^{-1} is locally Lipschitz. Therefore ∂W_j has measure zero, as desired.

Remark. If we allow countably many sets in our definition of rectifiable sets then we can similarly show that noncompact manifolds are also rectifiable. In this case we could use the fact that the covering $\{\phi_a(R_a)\}_{a\in M}$ has a countable subcovering (see Theorem 11.57).

***Example* 9.15.** An easy way for constructing manifolds is to consider the level sets of C^1 functions whose derivative has full rank. Suppose $f : U \to \mathbb{R}^m$ is a C^1 function, where $U \subset \mathbb{R}^n$ is an open set, and $m < n$. Suppose the level set

$$\Gamma := \{x \in U : f(x) = c\}$$

is nonempty. Also suppose that for every $x \in \Gamma$ the $m \times n$ matrix $Df(x)$ has rank m. Then $Df(x)$ must have m linearly independent columns. To simplify the notation let us assume that the last m columns of $Df(x)$ are linearly independent. Also let us denote the points in $\mathbb{R}^n$ by (z, y), where $z \in \mathbb{R}^{n-m}$ and $y \in \mathbb{R}^m$. Then the $m \times m$ matrix

$$\left[\frac{\partial f_i}{\partial y_j}(x)\right]_{1\leq i,j\leq m}$$

is invertible, since its columns are linearly independent. Hence by the implicit function theorem, there is an open set $V \subset \mathbb{R}^n$ containing x, an open set $W \subset \mathbb{R}^{n-m}$, and a unique C^1 function $g : W \to \mathbb{R}^m$ such that

$$\Gamma \cap V = \{(z, g(z)) : z \in W\}.$$

It is easy to see that the function $\phi : W \to \mathbb{R}^n$ defined by $\phi(z) := (z, g(z))$ is a parametrization of $\Gamma \cap V$. Because ϕ is obviously one-to-one and C^1, and ϕ^{-1} is continuous, since it is just the projection on the first $n - m$ components. In addition, we have

$$D\phi = \begin{bmatrix} I_{n-m} \\ Dg \end{bmatrix};$$

so its rank is $n - m$. Hence $\Gamma \cap V$ is an $(n-m)$-surface patch. Thus the level set Γ is an $(n-m)$-dimensional manifold in $\mathbb{R}^n$.

Definition 9.16. Let $S = A \sqcup S_1 \sqcup \cdots \sqcup S_m$ be a k-rectifiable set in $\mathbb{R}^n$, and $f : S \to \mathbb{R}$. We say f is integrable over S if each $f|_{S_j}$ is integrable over S_j, and in this case the integral of f over S is

$$\int_S f\,d\sigma := \int_{S_1} f\,d\sigma + \cdots + \int_{S_m} f\,d\sigma.$$

Remark. As before, we can also define the volume of S as

$$\mathrm{vol}(S) := \int_S 1\,d\sigma.$$

Remark. It is easy to see that if S has finite volume, then any bounded continuous function is integrable over S. In fact, in this case, any bounded function which is continuous on each S_i is integrable over S. (The boundedness on all of S, and not just on each S_i, will assure that the integrability of the function does not depend on the particular decomposition of S, as we will see below.)

Remark. As we mentioned in the last section, we can also define the integral of vector-valued functions over a surface patch, and hence over a rectifiable set. Then it would easily follow that the integrability of such a function is equivalent to the integrability of its components, and its integral can be computed componentwise.

Suppose we have a parametrization $S_i = \phi_i(V_i)$. Let K be a compact subset of $S_i \subset S$. Then $\phi_i^{-1}(K)$ is a compact subset of V_i, since ϕ_i^{-1} is continuous. Now note that $\phi_i^{-1}(K)$ is bounded and has a positive distance from ∂V_i. Thus, by the construction of the sequence of compact Jordan measurable subsets in the proof of Theorem 8.63, there is a compact Jordan measurable set $\tilde{K} \subset V_i$ that contains $\phi_i^{-1}(K)$. Then, by definition, $(f \circ \phi_i)\sqrt{\det(D\phi_i^{\intercal} D\phi_i)}$ is Riemann integrable and therefore bounded on $\tilde{K}$; hence it is bounded on $\phi_i^{-1}(K)$. But $\sqrt{\det(D\phi_i^{\intercal} D\phi_i)}$ is continuous and positive on the compact set $\phi_i^{-1}(K)$, so it has a positive minimum and a positive maximum there. Thus $f \circ \phi_i$ is bounded on $\phi_i^{-1}(K)$, and hence f is bounded on the compact set $K \subset S_i$.

Therefore, for f to be integrable over S with respect to the decomposition $S = A \sqcup S_1 \sqcup \cdots \sqcup S_m$, it must be bounded on compact subsets of S that lie entirely in one of the surface patches S_i. This restriction is imposed on f since we are employing the Riemann integral; however, if more general theories of integration are employed such restrictions will not arise.

A rectifiable set may have different decompositions into surface patches. For example, we can decompose a sphere into two open hemispheres and a great circle, and use the above definition to integrate a function over the sphere. However, there are infinitely many great circles in a sphere, and we can divide the sphere along any one of them. Hence we need to make sure that the value of an integral does not depend on a specific decomposition of a rectifiable set.

Theorem 9.17. *The integrability and the integral of a function f over a k-rectifiable set do not depend on the decomposition of the k-rectifiable set into the union of several k-surface patches and a set with k-measure zero, provided that f is bounded on compact subsets of the k-rectifiable set which lie entirely in one of the surface patches of that decomposition.*

Remark. Note that the assumption of boundedness on compact subsets is automatically satisfied when f is continuous.

Proof. Suppose $S \subset \mathbb{R}^n$ is a k-rectifiable set with two decompositions

$$A \sqcup S_1 \sqcup \cdots \sqcup S_m = S = A' \sqcup S'_1 \sqcup \cdots \sqcup S'_{m'},$$

where each S_j, S'_i is a k-surface patch which is open in S, and A, A' have k-measure zero. Suppose the function $f : S \to \mathbb{R}$ is integrable with respect to the first decomposition, i.e. $f|_{S_j}$ is integrable over S_j for each j. By assumption, f is also bounded on any compact subset of S which lies entirely in S'_i for some i. We know that S_j is a k-surface patch; hence there exists an open set $V_j \subset \mathbb{R}^k$ and a parametrization $\phi_j : V_j \to \mathbb{R}^n$ such that $S_j = \phi_j(V_j)$. Then by definition $(f \circ \phi_j)\sqrt{\det(D\phi_j^\intercal D\phi_j)}$ is integrable over V_j, and

$$\int_{S_j} f \, d\sigma = \int_{V_j} f(\phi_j(x)) \sqrt{\det\big(D\phi_j(x)^\intercal D\phi_j(x)\big)}\, dx.$$

Similarly we have $S'_i = \psi_i(V'_i)$, where $V'_i \subset \mathbb{R}^k$ is an open set, and $\psi_i : V'_i \to \mathbb{R}^n$ is a parametrization. We first need to show that for each i the function $(f \circ \psi_i)\sqrt{\det(D\psi_i^\intercal D\psi_i)}$ is integrable over V'_i. And then show that the integral of f over S does not depend on the decomposition of S, i.e. $\sum_j \int_{S_j} f\, d\sigma = \sum_i \int_{S'_i} f\, d\sigma$.

Now note that $S_j \cap S'_i$ is open in S; so $V_{ji} := \phi_j^{-1}(S_j \cap S'_i)$ is open in V_j, hence it is an open set (some of them might be empty, in which case we can simply discard them). Also note that the sets $V_{j1}, \dots, V_{jm'}, \phi_j^{-1}(A')$ are pairwise disjoint and their union is V_j. In addition, $\phi_j^{-1}(A')$ has measure zero in $\mathbb{R}^k$, since ϕ_j^{-1} is locally Lipschitz and A' has k-measure zero.

So far we have shown that each V_{ji} is an open subset of V_j, and $V_j - \bigcup_i V_{ji} = \phi_j^{-1}(A')$ is a set with k-measure zero in $\mathbb{R}^k$. Hence by applying Theorem 8.68 to the integrable function $g = (f \circ \phi_j)\sqrt{\det(D\phi_j^\intercal D\phi_j)}$ on V_j, we conclude that $g|_{V_{ji}}$ are also integrable, and we have

$$\int_{V_j} g(x)dx = \int_{V_{j1}} g(x)dx + \cdots + \int_{V_{jm'}} g(x)dx.$$

Now note that $S_j \cap S'_i$ is itself a k-surface patch with parametrization $\phi_j|_{V_{ji}}$. Hence we have

$$\int_{S_j \cap S'_i} f\,d\sigma = \int_{V_{ji}} f(\phi_j(x)) \sqrt{\det\left(D\phi_j(x)^\intercal D\phi_j(x)\right)}\,dx.$$

Therefore we have actually shown that $\int_{S_j} f\,d\sigma = \sum_{i\le m'} \int_{S_j\cap S'_i} f\,d\sigma$.

We can similarly show that the sets $V'_{ij} := \psi_i^{-1}(S_j \cap S'_i)$ are open, and together with $\psi_i^{-1}(A)$, they form a partition of V'_i into pairwise disjoint sets. And in addition, $\psi_i^{-1}(A)$ has k-measure zero in $\mathbb{R}^k$. Now note that $\psi_i|_{V'_{ij}}$ is also a parametrization for $S_j \cap S'_i$. Hence $(f\circ\psi_i)\sqrt{\det(D\psi_i^\intercal D\psi_i)}$ is integrable over V'_{ij}, and we also have

$$\int_{S_j \cap S'_i} f\,d\sigma = \int_{V'_{ij}} f(\psi_i(x)) \sqrt{\det\left(D\psi_i(x)^\intercal D\psi_i(x)\right)}\,dx.$$

Next note that that $h = (f\circ\psi_i)\sqrt{\det(D\psi_i^\intercal D\psi_i)}$ is bounded on any compact set $K \subset V'_i$, since f is bounded on the compact set $\psi_i(K) \subset S'_i$ by our assumption, and $\sqrt{\det(D\psi_i^\intercal D\psi_i)}$ is continuous on the compact set K. Therefore by Theorem 8.68, h is integrable over V'_i, and we have

$$\int_{V'_i} h(x)dx = \sum_{j\le m} \int_{V'_{ij}} h(x)dx = \sum_{j\le m} \int_{S_j\cap S'_i} f\,d\sigma.$$

Hence we have shown that f is integrable over S'_i, and we have

$$\int_{S'_i} f\,d\sigma = \int_{V'_i} h(x)dx = \sum_{j\le m} \int_{S_j\cap S'_i} f\,d\sigma.$$

Finally, we can show that the integral of f over S does not depend on its decomposition into k-surface patches. We have

$$\begin{aligned}\int_S f\,d\sigma = \sum_{j\le m}\int_{S_j} f\,d\sigma &= \sum_{j\le m}\sum_{i\le m'}\int_{S_j\cap S'_i} f\,d\sigma\\ &= \sum_{i\le m'}\sum_{j\le m}\int_{S_j\cap S'_i} f\,d\sigma = \sum_{i\le m'}\int_{S'_i} f\,d\sigma,\end{aligned}$$

as desired. ■

Theorem 9.18. *Let S be a k-rectifiable subset of $\mathbb{R}^n$, and suppose $f, g : S \to \mathbb{R}$ are integrable over S. Let $c_1, c_2, C \in \mathbb{R}$. Then we have*

(i) *$c_1 f + c_2 g$ is integrable over S and*

$$\int_S [c_1 f + c_2 g] d\sigma = c_1 \int_S f\,d\sigma + c_2 \int_S g\,d\sigma.$$

(ii) *If $f \leq g$ then*

$$\int_S f\,d\sigma \leq \int_S g\,d\sigma.$$

(iii) *If $|f| \leq C$ then*

$$\left|\int_S f\,d\sigma\right| \leq C\,\mathrm{vol}(S),$$

provided that the volume of S is finite.

Proof. Suppose $S = A \sqcup S_1 \sqcup \cdots \sqcup S_m$, where each S_j is a k-surface patch which is open in S, and A has k-measure zero. Suppose $\phi_j : V_j \to S_j$ is a parametrization of S_j. Then by the assumption we know that $f|_{S_j}, g|_{S_j}$ are integrable over S_j, which means $(f \circ \phi_j)J_j, (g \circ \phi_j)J_j$ are integrable over V_j, where $J_j = \sqrt{\det(D\phi_j^\intercal D\phi_j)}$ is the volume factor of ϕ_j.

(i) Note that $((c_1 f + c_2 g) \circ \phi_j)J_j = c_1(f \circ \phi_j)J_j + c_2(g \circ \phi_j)J_j$ is integrable over V_j for each j. Hence $c_1 f + c_2 g$ is integrable over S. Now we have

$$\begin{aligned}\int_S [c_1 f + c_2 g] d\sigma &= \sum_{j \leq m} \int_{S_j} [c_1 f + c_2 g] d\sigma \\ &= \sum_{j \leq m} \int_{V_j} [c_1 f(\phi_j(x)) + c_2 g(\phi_j(x))] J_j(x) dx \\ &= c_1 \sum_{j \leq m} \int_{V_j} f(\phi_j(x)) J_j(x) dx + c_2 \sum_{j \leq m} \int_{V_j} g(\phi_j(x)) J_j(x) dx \\ &= c_1 \sum_{j \leq m} \int_{S_j} f\,d\sigma + c_2 \sum_{j \leq m} \int_{S_j} g\,d\sigma = c_1 \int_S f\,d\sigma + c_2 \int_S g\,d\sigma.\end{aligned}$$

(ii) Note that we have $(f \circ \phi_j)J_j \leq (g \circ \phi_j)J_j$ for each j, since the volume factor is positive. Thus

$$\begin{aligned}\int_S f\,d\sigma = \sum_{j \leq m} \int_{S_j} f\,d\sigma &= \sum_{j \leq m} \int_{V_j} f(\phi_j(x)) J_j(x) dx \\ &\leq \sum_{j \leq m} \int_{V_j} g(\phi_j(x)) J_j(x) dx = \sum_{j \leq m} \int_{S_j} g\,d\sigma = \int_S g\,d\sigma.\end{aligned}$$

(iii) We have $-C \leq f \leq C$. Also note that constant functions are integrable over S, since they are a constant times the constant function 1, and by our assumption the constant function 1 is integrable over S. Hence by the previous parts we have

$$-\int_S C\,d\sigma = \int_S (-C)\,d\sigma \leq \int_S f\,d\sigma \leq \int_S C\,d\sigma = C \int_S 1\,d\sigma = C\,\mathrm{vol}(S),$$

which gives the desired. ■

9.3 Domains with Almost C^1 Boundary

Definition 9.19. Let $U \subset \mathbb{R}^n$ be an open set, where $n > 1$. Then we say U has **C^1 boundary** if for every point $a \in \partial U$ there exists an open ball $B_r(a)$ and a C^1 function $g : \mathbb{R}^{n-1} \to \mathbb{R}$, such that for some $j \in \{1, \dots, n\}$ and some $\varepsilon \in \{\pm 1\}$ we have

$$U \cap B_r(a) = \{x \in B_r(a) : \varepsilon x_j > g(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n)\}.$$

(Note that $r > 0$, and can depend on the point a.)

Remark. To simplify the notation we usually assume that $j = n$ and $\varepsilon = 1$, so that we have

$$U \cap B_r(a) = \{x : x_n > g(x_1, \dots, x_{n-1})\}.$$

Also note that the domain of g can be an open subset of $\mathbb{R}^{n-1}$, and need not be all of $\mathbb{R}^{n-1}$.

Proposition 9.20. *In the above definition we have*

$$\partial U \cap B_r(a) = \{x \in B_r(a) : x_n = g(x_1, \dots, x_{n-1})\}.$$

Hence $\partial U \cap B_r(a)$ is an $(n-1)$-surface patch with parametrization

$$\phi(x_1, \dots, x_{n-1}) = \big(x_1, \dots, x_{n-1}, g(x_1, \dots, x_{n-1})\big).$$

And consequently, ∂U is an $(n-1)$-dimensional manifold.

Proof. To simplify the notation we set $B = B_r(a)$. First suppose for some $x \in B$ we have $x_n = g(\tilde{x})$, where $\tilde{x} = (x_1, \dots, x_{n-1})$. We want to show that $x \in \partial U$. Since B is an open set, there is an open rectangle $R \subset B$ that contains x. Now note that for large enough m the points $z_m^{\pm} := (\tilde{x}, g(\tilde{x}) \pm \frac{1}{m})$ belong to R. Thus by our assumption about U we have $z_m^{+} \in U \cap B$ and $z_m^{-} \in U^c \cap B$. In addition we have $z_m^{\pm} \to x$; so $x \in \partial U$, as desired.

Conversely, suppose $x \in \partial U \cap B$. We want to show that $x_n = g(\tilde{x})$. If $x_n > g(\tilde{x})$ then we must have $x \in U$, which contradicts the fact that U is open. And if $x_n < g(\tilde{x})$ then there is an open ball $B_s(x) \subset B$ such that for every $y \in B_s(x)$ we have $y_n < g(\tilde{y})$. Because the inverse image of the open interval $(-\infty, 0)$ under the continuous function $y \mapsto y_n - g(\tilde{y})$ is an open neighborhood of x. Thus it would follow that an open neighborhood $B_s(x)$ of x does not intersect U; which implies that x cannot belong to ∂U. Therefore we must have $x_n = g(\tilde{x})$, as desired.

Hence $\partial U \cap B$ is the graph of the function g. Thus, as we have seen in Example 9.15, $\partial U \cap B$ is a surface patch with parametrization ϕ. Note that the domain of ϕ is the (bounded) open set $\phi^{-1}(B) \subset P(B)$, where P is the projection on the first $n-1$ components (actually we have $\phi^{-1} = P|_{\phi^{-1}(B)}$). ■

Remark. Let U be an open set with C^1 boundary. We know that locally U is the region above (or below) the graph of a C^1 function. We have also shown that ∂U is a manifold. So ∂U can be locally parametrized by a C^1 parametrization. However, note that this last property alone does not imply that an open set has C^1 boundary. Because an open set with C^1 boundary must also lie on one side of its boundary.

Definition 9.21. Let $U \subset \mathbb{R}^n$ be an open set, where $n > 1$. Then we say U has **almost C^1 boundary** if ∂U can be written as a union of pairwise disjoint sets

$$\partial U = A \sqcup S_1 \sqcup \cdots \sqcup S_m,$$

such that A has $(n-1)$-measure zero, and each S_i is an $(n-1)$-surface patch which is an open set in ∂U, and U lies on one side of it. More precisely, for each i there is an open set U_i, and a C^1 function $g : \mathbb{R}^{n-1} \to \mathbb{R}$, such that for some $j \in \{1, \dots, n\}$ and some $\varepsilon \in \{\pm 1\}$ we have

$$U \cap U_i = \{x \in U_i : \varepsilon x_j > g(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n)\},$$

and

$$S_i = \partial U \cap U_i = \{x \in U_i : x_j = \varepsilon g(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n)\}.$$

Remark. It is obvious from the above definition that if U has almost C^1 boundary, then ∂U is an $(n-1)$-rectifiable set. The parametrization of S_i corresponding to εg is

$$\phi(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n) := (x_1, \dots, x_{j-1}, \varepsilon g, x_{j+1}, \dots, x_n),$$

and is defined on the open set $\phi^{-1}(U_i)$.

To simplify the notation we usually assume that $j = n$ and $\varepsilon = 1$, so that we have

$$U \cap U_i = \{x : x_n > g(x_1, \dots, x_{n-1})\},$$
$$S_i = \partial U \cap U_i = \{x : x_n = g(x_1, \dots, x_{n-1})\},$$

and

$$\phi(x_1, \dots, x_{n-1}) = \big(x_1, \dots, x_{n-1}, g(x_1, \dots, x_{n-1})\big).$$

Also note that the domain of g can be an open subset of $\mathbb{R}^{n-1}$, and need not be all of $\mathbb{R}^{n-1}$. In fact, if we let $P : (x_1, \dots, x_{n-1}, x_n) \mapsto (x_1, \dots, x_{n-1})$ be the projection on $\mathbb{R}^{n-1}$, then we can assume that the domain of g is $P(U_i)$. Note that $P(U_i)$ is an open set in $\mathbb{R}^{n-1}$. Because we can easily see that the projection of any open ball in $\mathbb{R}^n$ is an open ball in $\mathbb{R}^{n-1}$. And since the image of the union of some open balls under P is the union of the image of those open balls, the image of U_i must be an open set.

In addition, we can easily change the coordinates to locally straighten the boundary of U. Consider the C^1 functions Φ, Ψ defined as

$$\begin{cases} y_i = \Phi_i(x) := x_i & \text{for} \quad i < n, \\ y_n = \Phi_n(x) := x_n - g(x_1, \dots, x_{n-1}), \end{cases}$$

and

$$\begin{cases} x_i = \Psi_i(y) := y_i & \text{for} \quad i < n, \\ x_n = \Psi_n(y) := y_n + g(y_1, \dots, y_{n-1}). \end{cases}$$

Note that the domain of Φ, Ψ is the open set $P^{-1}(V)$, where V is the domain of g. In particular, the domain of Φ, Ψ contains U_i. It is easy to see that $\Psi = \Phi^{-1}$, and

$$\Phi(U \cap U_i) \subset \{y_n > 0\}, \qquad \Phi(\partial U \cap U_i) \subset \{y_n = 0\}.$$

Also note that

$$D\Phi = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -g_{x_1} & -g_{x_2} & \dots & 1 \end{bmatrix};$$

so $D\Phi$ is invertible, and we have $\det D\Phi = 1$. The same is true about $D\Psi$.

Proposition 9.22. *Let $U \subset \mathbb{R}^n$ be a bounded open set with C^1 boundary. Then U also has almost C^1 boundary.*

Proof. The proof is similar to the proof of Theorem 9.14, in which we showed that a manifold is rectifiable. We know that for every point $a \in \partial U$ there is an open ball $B_r(a)$ and a C^1 function g (which depends on a) such that

$$U \cap B_r(a) = \{x \in B_r(a) : x_n > g(x_1, \dots, x_{n-1})\},$$
$$\partial U \cap B_r(a) = \{x \in B_r(a) : x_n = g(x_1, \dots, x_{n-1})\}.$$

Let ϕ_a be the parametrization of $\partial U \cap B_r(a)$ corresponding to g, and let $V_a := \phi_a^{-1}(B_r(a))$ be its domain. Now consider an open rectangle R_a containing $\phi_a^{-1}(a)$, such that $\overline{R}_a \subset V_a$. Then the collection $\{\phi_a(R_a)\}_{a \in \partial U}$ covers ∂U. Also, due to the continuity of ϕ_a^{-1}, each $\phi_a(R_a)$ is open in $B_r(a) \cap \partial U$, hence it is open in ∂U. On the other hand, ∂U is closed, and as U is bounded, ∂U is also bounded. Hence ∂U is compact. Therefore finitely many of these open sets, namely $\phi_1(R_1), \dots, \phi_m(R_m)$, cover ∂U.

Next, similarly to the proof of Theorem 9.14, we can construct disjoint sets $S_1, \dots, S_m$ out of $\phi_1(R_1), \dots, \phi_m(R_m)$ such that

$$\partial U = A \sqcup S_1 \sqcup \dots \sqcup S_m,$$

where A has $(n-1)$-measure zero, and each S_i is an $(n-1)$-surface patch which is an open set in ∂U. (Some of the S_i's might be empty, in which case we simply discard them). Now since S_i is open in ∂U, we have $S_i = \partial U \cap \tilde{U}_i$ for some open set $\tilde{U}_i$. We also have $S_i \subset \phi_i(R_i) \subset B_{r_i}(a_i)$. Set $U_i := \tilde{U}_i \cap B_{r_i}(a_i)$. Then we have

$$S_i = S_i \cap B_{r_i}(a_i) = \partial U \cap \tilde{U}_i \cap B_{r_i}(a_i) = \partial U \cap U_i.$$

To simplify the notation, let us denote the points of $\mathbb{R}^n$ by $x = (\tilde{x}, x_n)$. Then

$$\begin{aligned} S_i = \partial U \cap U_i = \partial U \cap B_{r_i}(a_i) \cap \tilde{U}_i &= \{x \in B_{r_i}(a_i) : x_n = g(\tilde{x})\} \cap \tilde{U}_i \\ &= \{x \in B_{r_i}(a_i) \cap \tilde{U}_i : x_n = g(\tilde{x})\} = \{x \in U_i : x_n = g(\tilde{x})\}. \end{aligned}$$

We also have

$$\begin{aligned} U \cap U_i = U \cap B_{r_i}(a_i) \cap \tilde{U}_i &= \{x \in B_{r_i}(a_i) : x_n > g(\tilde{x})\} \cap \tilde{U}_i \\ &= \{x \in B_{r_i}(a_i) \cap \tilde{U}_i : x_n > g(\tilde{x})\} = \{x \in U_i : x_n > g(\tilde{x})\}. \end{aligned}$$

Therefore U has almost C^1 boundary, as desired. ■

Proposition 9.23. *Let $U \subset \mathbb{R}^n$ be a bounded open set with almost C^1 boundary. Then U is Jordan measurable.*

Proof. We only need to show that ∂U has measure zero in $\mathbb{R}^n$. We know that ∂U consists of a set A with $(n-1)$-measure zero, and several parts S_i which are graphs of C^1 functions over some open sets. As we have shown in Theorem 8.44 and the remark after it, the graph of a continuous function over an open set has measure zero. So each S_i has measure zero. On the other hand, A has n-measure zero in $\mathbb{R}^n$ too, since it has $(n-1)$-measure zero. Therefore ∂U is the union of finitely many sets with measure zero; so it also has measure zero. ■

Next let us define the normal to an almost C^1 boundary. Suppose

$$\begin{aligned} U \cap U_i &= \{x : x_n > g(x_1, \dots, x_{n-1})\}, \\ \partial U \cap U_i &= \{x : x_n = g(x_1, \dots, x_{n-1})\}, \end{aligned}$$

for some C^1 function g. Let $\tilde{x} = (x_1, \dots, x_{n-1})$. Then for small h, the points $(\tilde{x}, g(\tilde{x}))$ and $(\tilde{x} + he_k, g(\tilde{x} + he_k))$ belong to ∂U. So their difference $(he_k, g(\tilde{x} + he_k) - g(\tilde{x}))$ is almost tangent to ∂U. If we normalize this vector by dividing it by h, and let $h \to 0$, then we intuitively know that the limit, i.e. $(e_k, D_k g(\tilde{x}))$, must be tangent to ∂U. It is easy to see that the vector $(Dg(\tilde{x}), -1)$ is orthogonal to $(e_k, D_k g(\tilde{x}))$ for every k. Hence $(Dg(\tilde{x}), -1)$ is normal to ∂U at the point $(\tilde{x}, g(\tilde{x}))$. In addition, note that $(Dg(\tilde{x}), -1)$ is pointing to the outside of U. Because if we slightly move alongside of it we reach the point

$$(\tilde{x}, g(\tilde{x})) + t(Dg(\tilde{x}), -1) = \big(\tilde{x} + tDg(\tilde{x}), g(\tilde{x}) - t\big).$$

Now note that $g(\tilde{x}+tDg(\tilde{x})) = g(\tilde{x})+tDg(\tilde{x})\cdot Dg(\tilde{x})+r(t)$, where $r(t)$ is a sublinear function. Thus for small enough positive t we have $|r(t)| \le t|Dg(\tilde{x})|^2$ (provided that $Dg(\tilde{x})$ is nonzero; otherwise the following desired inequality holds trivially). Hence $g(\tilde{x}+tDg(\tilde{x})) \ge g(\tilde{x}) > g(\tilde{x}) - t$. Therefore the above point belongs to U^c.

Motivated by the above argument, we define the **unit outward normal** to ∂U at the point $a = (\tilde{x}_0, g(\tilde{x}_0))$ to be

$$\nu(a) := \frac{(Dg(\tilde{x}_0), -1)}{\sqrt{1+|Dg(\tilde{x}_0)|^2}} = \frac{1}{\sqrt{1+|Dg(\tilde{x}_0)|^2}}\big(D_1g(\tilde{x}_0), \dots, D_{n-1}g(\tilde{x}_0), -1\big).$$

Note that ν is defined at every point of ∂U except on the set of possible singularities of ∂U, which by assumption has $(n-1)$-measure zero. In other words, ν is defined at *almost every* point of ∂U. Also note that on the nonsingular part of ∂U, which by assumption is an open subset of ∂U, ν is continuous. (When ∂U is C^1, the normal ν is defined and is continuous everywhere.) Of course we also need to check that ν does not depend on the particular representation of ∂U around a as a graph of a function.

To prove the independence of ν from the representation of ∂U, let us assume that for some $j \in \{1, \dots, n\}$, $\varepsilon \in \{\pm 1\}$, and C^1 function f we also have

$$U \cap U'_i = \{y : \varepsilon y_j > f(y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_n)\},$$

and $a \in \partial U \cap U'_i$, where

$$\partial U \cap U'_i = \{y : y_j = \varepsilon f(y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_n)\}.$$

Let $\hat{y} = (y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_n)$. Then the points $(\varepsilon f(\hat{y}), \hat{y})$ are on ∂U, where $(\varepsilon f(\hat{y}), \hat{y})$ is a shorthand notation for $(y_1, \dots, y_{j-1}, \varepsilon f(\hat{y}), y_{j+1}, \dots, y_n)$. Now for some $\hat{y}_0$ we have $a = (\varepsilon f(\hat{y}_0), \hat{y}_0)$. Then if we repeat the above argument, we arrive at the following definition for the unit outward normal to the boundary

$$\nu(a) = \frac{1}{\sqrt{1+|Df(\hat{y}_0)|^2}}\big(D_1f(\hat{y}_0), \dots, D_{j-1}f(\hat{y}_0), -\varepsilon, D_{j+1}f(\hat{y}_0), \dots, D_nf(\hat{y}_0)\big).$$

Let us show that these two values for ν are the same. First note that both of these values are vectors with norm 1. In addition, for every $\tilde{x}$ in a neighborhood of $\tilde{x}_0$ we have $(\tilde{x}, g(\tilde{x})) \in \partial U \cap U'_i$; so there is $\hat{y}$ such that $(\tilde{x}, g(\tilde{x})) = (\varepsilon f(\hat{y}), \hat{y})$. Hence we have

$$\begin{cases} x_k = y_k & k \ne j, n, \\ x_j = \varepsilon f(\hat{y}), \\ y_n = g(\tilde{x}). \end{cases}$$

Thus

$$y_n = g(\tilde{x}) = g(x_1, \dots, x_{n-1}) = g(y_1, \dots, y_{j-1}, \varepsilon f(\hat{y}), y_{j+1}, \dots, y_{n-1}).$$

Therefore $1 = \frac{\partial y_n}{\partial y_n} = \varepsilon D_j g \frac{\partial f}{\partial y_n} = \varepsilon D_j g D_n f$. In particular we get $D_j g \neq 0$. And for $k \neq j, n$ we have

$$0 = \frac{\partial y_n}{\partial y_k} = D_k g + \varepsilon D_j g D_k f \implies D_k f = -\varepsilon \frac{D_k g}{D_j g}.$$

Hence

$$\begin{aligned}&\big(D_1 f, \dots, D_{j-1} f, -\varepsilon, D_{j+1} f, \dots, D_n f\big)\\ &\qquad = \big(-\varepsilon \frac{D_1 g}{D_j g}, \dots, -\varepsilon \frac{D_{j-1} g}{D_j g}, -\varepsilon \frac{D_j g}{D_j g}, -\varepsilon \frac{D_{j+1} g}{D_j g}, \dots, \varepsilon \frac{1}{D_j g}\big)\\ &\qquad = \frac{-\varepsilon}{D_j g}\big(D_1 g, \dots, D_{j-1} g, D_j g, D_{j+1} g, \dots, -1\big).\end{aligned}$$

Now if we show that $\frac{-\varepsilon}{D_j g} > 0$, then we can normalize both sides of the above equation to have norm 1, and conclude that the two formulas for ν give the same value. And it will follow that ν is well defined. To see that $\frac{-\varepsilon}{D_j g} > 0$ note that we obtained the formulas for ν by using the fact that $a + t\nu \in U^c$ for small $t > 0$. We can similarly show that $a - t\nu \in U$. Hence the two formulas for ν cannot point in opposite directions, and $\frac{-\varepsilon}{D_j g}$ must be positive. ■

Sometimes part of ∂U is parametrized, but the parametrization does not come from representing that part of ∂U as the graph of a function. Suppose ψ is a parametrization of an open subset of ∂U containing the point a. In this case we can still find $\nu(a)$ up to multiplication by ±1, using $D_1\psi, \dots, D_{n-1}\psi$. Note that these vectors are linearly independent, since ψ is a parametrization and thus $D\psi$ has rank $n-1$. Let us show that the subspace spanned by $D_1\psi, \dots, D_{n-1}\psi$ does not depend on the parametrization ψ. Let ϕ be another parametrization. Then we know that $\phi^{-1} \circ \psi$ is a C^1 function. We have $\psi = \phi \circ (\phi^{-1} \circ \psi)$. Hence

$$D_k \psi = \sum_l D_l \phi \, D_k (\phi^{-1} \circ \psi)_l = \sum_l a_{kl} D_l \phi,$$

where $a_{kl} = D_k(\phi^{-1} \circ \psi)_l$. So each $D_k\psi$ is a linear combination of $D_1\phi, \dots, D_{n-1}\phi$. Conversely we can show that each $D_k\phi$ is a linear combination of $D_1\psi, \dots, D_{n-1}\psi$. Thus the subspace spanned by $D_1\psi, \dots, D_{n-1}\psi$ does not depend on the parametrization ψ. This subspace is known as the **tangent space** to ∂U at the point a.

Let $\phi(\tilde{x}) = (\tilde{x}, g(\tilde{x}))$ be a parametrization of a neighborhood of a in ∂U, such that U lies above the graph of the function g. As we have explained before the definition of normal to the boundary, ν is orthogonal to $D_1\phi, \dots, D_{n-1}\phi$. Hence $\nu(a)$ is orthogonal to the tangent space to ∂U at a. Now suppose we are given an arbitrary parametrization ψ of ∂U. Then we know that ν is a vector with norm one, and belongs to the one-dimensional orthogonal complement of the tangent space, which is spanned by the vectors $D_1\psi, \dots, D_{n-1}\psi$. Thus to determine ν we only need to determine its direction, so that ν points to the outside of U. However, an arbitrary parametrization like ψ does not provide sufficient information for determining the direction of ν, and we need to additionally know U lies on which side of ∂U.

Another case of interest is when (part of) ∂U is determined as the level set $\{G = c\}$ of a C^1 function G from a subset of $\mathbb{R}^n$ into $\mathbb{R}$ whose derivative does not vanish. Then by the implicit function theorem, $\{G = c\} = \{(\tilde{x}, g(\tilde{x}))\}$ for some C^1 function g. (To simplify the notation we assumed that $D_nG \neq 0$, so that x_n becomes a function of $\tilde{x} = (x_1, \dots, x_{n-1})$.) We also know that

$$G(\tilde{x}, g(\tilde{x})) = c \implies D_kG + D_nGD_kg = 0.$$

Therefore $D_kg = -\frac{D_kG}{D_nG}$. Hence

$$(D_1g, \dots, D_{n-1}g, -1) = \frac{-1}{D_nG}(D_1G, \dots, D_{n-1}G, D_nG) = \frac{-1}{D_nG}DG.$$

Thus ν is proportional to DG; so we have

$$\nu = \pm\frac{DG}{|DG|},$$

since ν has norm one. As before, to determine the sign we need to additionally know U lies on which side of ∂U. The above observation can also be expressed by saying that DG is orthogonal to the level sets of G. ■

9.4 The Divergence Theorem

Let $R = \prod_{i\leq n}(a_i^-, a_i^+)$ be an open rectangle. Then the functions $g_i^\pm = a_i^\pm$ on suitable domains define ∂R. In addition, R lies below the boundary parts corresponding to $g_i^+ = a_i^+$, and above the boundary parts corresponding to $g_i^- = a_i^-$. Hence the unit outward normal ν to ∂R corresponding to these boundary parts are $\pm e_i$, respectively. Thus, in particular, for some $j \leq n$ we have $\nu_j = \pm 1$ on the parts of ∂R corresponding to $g_j^\pm = a_j^\pm$, and $\nu_j = 0$ elsewhere on ∂R. Now let us decompose R into $(a_j^-, a_j^+) \times \hat{R}$, where $\hat{R} = \prod_{i\neq j}(a_i^-, a_i^+)$ is an $(n-1)$-dimensional open rectangle. Let f be a C^1 function from an open set containing R into $\mathbb{R}$. Then we have

$$\begin{aligned}\int_R D_jf\,dx &= \int_{\hat{R}}\int_{a_j^-}^{a_j^+} D_jf\,dx_jd\hat{x}\\ &= \int_{\hat{R}} f(a_j^+, \hat{x}) - f(a_j^-, \hat{x})\,d\hat{x} = \int_{\partial R} f\nu_j\,d\sigma.\end{aligned}$$

This is a special case of the divergence theorem.

Next note that if we have two adjacent rectangles, then the outward normals to their common boundary part will be opposite to each other. Therefore when we integrate $f\nu_j$ over the union of their boundaries, the integrals over the common part of their boundaries cancel each other. By changing the coordinates, we can also show that the same conclusions hold for deformed "curved rectangles". So if a domain can be decomposed as the union of several "curved rectangles" which only intersect at their boundaries, then the divergence theorem holds for that domain. Because the integral of $D_j f$ over the domain is the sum of its integrals over the "curved rectangles", and the sum of the integrals of $f\nu_j$ over the boundaries of the "curved rectangles" is equal to its integral over the boundary of the domain, since the integrals over the common boundary parts cancel each other.

To rigorously implement the above idea for the proof of the divergence theorem, we need to be able to decompose a domain as the union of several "curved rectangles" which only intersect at their boundaries. However, even for nice domains, proving this property is not easy. Thus we follow another approach in which we prove the divergence theorem by first working locally, and then patching things together using *partitions of unity*, which we are going to introduce.

Definition 9.24. Let $\zeta : \mathbb{R}^n \to \mathbb{R}^m$. The **support** of ζ is the closure of the set over which $\zeta \neq 0$, i.e.

$$\operatorname{spt}(\zeta) := \overline{\{x \in \mathbb{R}^n : \zeta(x) \neq 0\}}.$$

A function with compact support is a function whose support is a compact subset of $\mathbb{R}^n$.

Theorem 9.25. *Let $U \subset \mathbb{R}^n$ be an open set, and let $K \subset U$ be a compact set. Then there is a C^∞ function $\zeta : \mathbb{R}^n \to \mathbb{R}$ with compact support, such that* $\operatorname{spt}(\zeta) \subset U$*,* $0 \le \zeta \le 1$*, and* $\zeta = 1$ *on* K*.*

Proof. Consider the function

$$f(t) := \begin{cases} e^{-\frac{1}{t}} & t > 0, \\ 0 & t \le 0. \end{cases}$$

In Example 6.25 we showed that f is a C^∞ function (which is not analytic). Let

$$g(t) := f(1-t)f(t).$$

Then g is a C^∞ function which is positive on $(0, 1)$, and is zero everywhere else. Now consider an open rectangle $R \subset \mathbb{R}^n$ whose ith edge is (a_i, b_i). Set

$$\eta(x) := g\Big(\frac{x_1 - a_1}{b_1 - a_1}\Big) \cdots g\Big(\frac{x_n - a_n}{b_n - a_n}\Big).$$

Then $\eta : \mathbb{R}^n \to \mathbb{R}$ is a C^∞ function which is positive on R, and is zero everywhere else; so $\operatorname{spt}(\eta) = \overline{R}$.

Since $K \subset U$, and U is open, for every $x \in K$ there is an open rectangle R_x containing x, such that $\overline{R}_x \subset U$. And since K is compact, finitely many of these open rectangles, namely $R_1, \ldots, R_m$, cover K. Let η_i be the smooth function constructed above, whose support is $\overline{R}_i$. Then $\gamma := \eta_1 + \cdots + \eta_m$ is a C^∞ function which is positive on $\bigcup_i R_i \supset K$, and is zero outside of it. Since K is compact, we have $\gamma \geq \epsilon$ on K, for some $\epsilon > 0$.

Now consider the function

$$h(t) := \frac{1}{c} \int_0^t g\big(\frac{\tau}{\epsilon}\big) d\tau,$$

where $c = \int_0^\epsilon g(\frac{\tau}{\epsilon}) d\tau$. Then $h'(t) = \frac{1}{c} g(\frac{t}{\epsilon})$; so h is also C^∞. Note that $g(\frac{t}{\epsilon})$ is positive on $(0, \epsilon)$, and zero elsewhere. In addition, for $t < 0$ we have

$$h(t) = \frac{-1}{c} \int_t^0 g\big(\frac{\tau}{\epsilon}\big) d\tau = \frac{-1}{c} \int_t^0 0 \, d\tau = 0;$$

and for $t \geq \epsilon$ we have

$$h(t) = \frac{1}{c} \int_0^t g\big(\frac{\tau}{\epsilon}\big) d\tau = \frac{1}{c} \int_0^\epsilon g\big(\frac{\tau}{\epsilon}\big) d\tau + \frac{1}{c} \int_\epsilon^t g\big(\frac{\tau}{\epsilon}\big) d\tau = \frac{1}{c} c + \frac{1}{c} \int_\epsilon^t 0 \, d\tau = 1.$$

Also, since $g(\frac{\cdot}{\epsilon})$ is positive on $(0, \epsilon)$, h is positive there; and for $t \in (0, \epsilon)$ we obviously have $h(t) < 1$, because $\int_0^t g(\frac{\tau}{\epsilon}) d\tau < \int_0^\epsilon g(\frac{\tau}{\epsilon}) d\tau$.

Finally, let $\zeta := h \circ \gamma$. Then ζ is C^∞, and $0 \leq \zeta \leq 1$, since $0 \leq h \leq 1$. In addition, for $x \in K$ we have

$$\zeta(x) = h(\gamma(x)) = 1,$$

because $\gamma(x) \geq \epsilon$. Also, for $x \notin \bigcup_i R_i$ we have $\zeta(x) = 0$, since $\gamma(x) = 0$. On the other hand, by Exercise 2.36, the closure of the union of finitely many sets is the union of their closures. Hence we have

$$\operatorname{spt}(\zeta) \subset \overline{\bigcup R_i} = \bigcup \overline{R}_i \subset U.$$

Furthermore, $\operatorname{spt}(\zeta)$ is compact, since it is closed, and $\bigcup \overline{R}_i$ is compact. ■

Theorem 9.26. *Let $K \subset \mathbb{R}^n$ be a compact set, and suppose $U_1, \ldots, U_m \subset \mathbb{R}^n$ form an open covering of K. Then there are C^∞ functions $\zeta_1, \ldots, \zeta_m : \mathbb{R}^n \to \mathbb{R}$ with compact support, such that* $\operatorname{spt}(\zeta_i) \subset U_i$, $0 \leq \zeta_i \leq 1$, *and for every $x \in K$ we have*

$$\zeta_1(x) + \cdots + \zeta_m(x) = 1.$$

Remark. The functions $\zeta_1, \dots, \zeta_m$ are called a **partition of unity** subordinate to the open covering $U_1, \dots, U_m$.

Proof. For every $x \in K$ there is U_i such that $x \in U_i$. Hence there is an open rectangle R_x containing x, such that $\overline{R_x} \subset U_i$, because U_i is open. And since K is compact, finitely many of these open rectangles, namely $R_1, \dots, R_l$, cover K. Let V_i be the union of all open rectangles R_j for which we have $\overline{R_j} \subset U_i$. (A rectangle $\overline{R_j}$ can also be a subset of another open set $U_{i'}$, and in this case we include R_j in $V_{i'}$ too.) Note that the open sets $V_1, \dots, V_m$ also cover K. Let $\gamma_i : \mathbb{R}^n \to \mathbb{R}$ be a C^∞ function with compact support $\operatorname{spt}(\gamma_i) \subset U_i$, whose values are in $[0, 1]$, and is 1 on $\overline{V}_i$. Now set $\zeta_1 := \gamma_1$. And for $i > 1$ set

$$\zeta_i := (1 - \gamma_1) \cdots (1 - \gamma_{i-1})\gamma_i.$$

Note that ζ_i is a C^∞ function whose values are in $[0, 1]$, and vanishes outside the support of γ_i. So $\operatorname{spt}(\zeta_i) \subset \operatorname{spt}(\gamma_i) \subset U_i$; and thus $\operatorname{spt}(\zeta_i)$ is compact too.

Finally note that for every k we have

$$\zeta_1 + \cdots + \zeta_k = 1 - (1 - \gamma_1) \cdots (1 - \gamma_k).$$

Because for $k = 1$ the equality holds trivially. And if it holds for k, then for $k + 1$ we have

$$\begin{aligned}\zeta_1 + \cdots + \zeta_k + \zeta_{k+1} &= 1 - (1 - \gamma_1) \cdots (1 - \gamma_k) + (1 - \gamma_1) \cdots (1 - \gamma_k)\gamma_{k+1} \\ &= 1 - (1 - \gamma_1) \cdots (1 - \gamma_k)(1 - \gamma_{k+1}),\end{aligned}$$

as desired. Hence in particular $\zeta_1 + \cdots + \zeta_m = 1 - (1 - \gamma_1) \cdots (1 - \gamma_m)$. Now note that for $x \in K$ we have $x \in V_i$ for some i. Hence we have $\gamma_i(x) = 1$. Therefore we get

$$\zeta_1(x) + \cdots + \zeta_m(x) = 1 - 0 = 1,$$

as wanted. ■

Remark. Suppose that in the above theorem $U_1 = R_1, \dots, U_k = R_k$ are open cubes. Let $\tilde{R}_1, \dots, \tilde{R}_k$ be open cubes with the same center and twice the side length. Then $\tilde{R}_1, \dots, \tilde{R}_k, U_{k+1}, \dots, U_m$ is also an open covering of K. Let us construct a partition of unity subordinate to this partition, which has special properties, and will be needed in the proof of the divergence theorem. First note that if we enlarge one of the sets V_i in the above proof, and use the corresponding γ_i, then the construction of the partition of unity will still work. Let us use R_i instead of V_i. In this case $\operatorname{spt}(\gamma_i)$ is not a subset of R_i anymore, but we can still make sure that $\operatorname{spt}(\gamma_i) \subset \tilde{R}_i$.

Let us recall how we constructed γ_i. Suppose the jth edge of the open cube R_i is (a_j, b_j). Then $b_j - a_j = l$, where l is the side length of R_i. Then the jth edge of $\tilde{R}_i$ is $(a_j - l/2, b_j + l/2)$. Now consider the function

$$\eta(x) := g\Big(\frac{x_1 - (a_1 - l/4)}{\frac{3}{2}l}\Big) \cdots g\Big(\frac{x_n - (a_n - l/4)}{\frac{3}{2}l}\Big),$$

where g is defined in the proof of Theorem 9.25. Then η is a C^∞ function which is positive on the open rectangle whose jth edge is $(a_j - l/4, b_j + l/4)$, and vanishes elsewhere. Let d be the minimum of g on $[\frac{1}{6}, \frac{5}{6}]$. Then $\eta \ge d^n$ on R_i, since for $x \in R_i$ we have $\frac{x_j - (a_j - l/4)}{3l/2} \in [\frac{1}{6}, \frac{5}{6}]$. If we set $\epsilon = d^n$, and use the function

$$h(t) := \frac{1}{c}\int_0^t g\big(\frac{\tau}{\epsilon}\big)d\tau,$$

with $c = \int_0^\epsilon g(\frac{\tau}{\epsilon})d\tau$, then as we showed in the proof of Theorem 9.25, $\gamma_i := h \circ \eta$ is a C^∞ function with values in $[0, 1]$, which is 1 on R_i, and its support is in $\tilde{R}_i$.

Now let us compute $D_j\gamma_i$. We have

$$D_j\gamma_i = h'(\eta)D_j\eta = \frac{1}{c}g\big(\frac{\eta}{\epsilon}\big)\frac{2}{3l}g'\Big(\frac{x_j - (a_j - l/4)}{\frac{3}{2}l}\Big)\prod_{j' \ne j} g\Big(\frac{x_{j'} - (a_{j'} - l/4)}{\frac{3}{2}l}\Big).$$

Let C_0, C_1 be the maximum of g, g' respectively. Then we have

$$|D_j\gamma_i| \le \frac{2C_0^nC_1}{3c}\frac{1}{l} \implies |D\gamma_i| \le \frac{C}{\operatorname{diam}\tilde{R}_i},$$

for some constant C which only depends on the function g, and the dimension n; because the diameter of a cube can be computed in terms of its side length.

Next remember that

$$\zeta_1 + \cdots + \zeta_k = 1 - (1 - \gamma_1)\cdots(1 - \gamma_k).$$

Hence we have $D_j(\zeta_1 + \cdots + \zeta_k) = \sum_{i=1}^k D_j\gamma_i \prod_{i' \ne i}(1 - \gamma_{i'})$. Therefore

$$\begin{aligned}\big|D_j(\zeta_1 + \cdots + \zeta_k)\big| &\le \sum_{i \le k}|D_j\gamma_i|\prod_{i' \ne i}|1 - \gamma_{i'}|\\ &\le \sum_{i \le k}|D_j\gamma_i| = \sum_{i \le k}|D_j\gamma_i|\chi_{\tilde{R}_i} \le \sum_{i \le k}\frac{C}{\operatorname{diam}\tilde{R}_i}\chi_{\tilde{R}_i},\end{aligned}$$

where χ denotes the characteristic function of a set. Note that $|D_j\gamma_i| = |D_j\gamma_i|\chi_{\tilde{R}_i}$ because $D\gamma_i$ is zero outside of $\tilde{R}_i$. ■

Lemma 9.27. *Let $V, A \subset \mathbb{R}^n$. Suppose V is open and $V \subset A$. Also suppose $f : A \to \mathbb{R}$ is integrable over compact Jordan measurable subsets of A, and* $\operatorname{spt}(f) \subset V$. *Then f is integrable over A if and only if it is integrable over V, and in this case we have*

$$\int_A f(x)dx = \int_V f(x)dx.$$

Proof. First suppose $f \geq 0$. Let $S \subset V$ and $K \subset A$ be arbitrary compact Jordan measurable sets. Suppose f is integrable over A. Then since $S \subset A$ we have

$$\int_S f(x)dx \leq \sup_{K \subset A} \int_K f(x)dx = \int_A f(x)dx < \infty.$$

Hence, by taking supremum over S, we conclude that f is integrable over V and we have $\int_V f(x)dx \leq \int_A f(x)dx$.

Now suppose f is integrable over V. Let $S_k \subset V$ be a sequence of compact Jordan measurable sets given by Theorem 8.63, which satisfy $S_k \subset S_{k+1}^\circ$ and $V = \bigcup_{k \geq 1} S_k$. Now $\operatorname{spt}(f) \cap K$ is a compact subset of V, since $\operatorname{spt}(f)$ is closed. So it is bounded and has a positive distance from ∂V. Thus, by the construction of the sequence S_k in the proof of Theorem 8.63, there is m such that $\operatorname{spt}(f) \cap K \subset S_m$. Then, by Theorem 8.43 we get

$$\int_K f(x)dx = \int_{S_m} f(x)dx + \int_{K - S_m} f(x)dx = \int_{S_m} f(x)dx,$$

since f is zero on $K - S_m$. Hence we have

$$\int_K f(x)dx = \int_{S_m} f(x)dx \leq \sup_{S \subset V} \int_S f(x)dx = \int_V f(x)dx.$$

Therefore, by taking supremum over K, we conclude that f is integrable over A and we have $\int_A f(x)dx \leq \int_V f(x)dx$. Thus for $f \geq 0$, the integrability of f over A, V are equivalent, and the integrals are equal.

Finally, for general f, the above argument implies that the integrability of $f^\pm$ over A, V are equivalent, and we have $\int_A f^\pm(x)dx = \int_V f^\pm(x)dx$. Hence the integrability of f over A, V are equivalent too, and

$$\begin{aligned}\int_A f(x)dx &= \int_A f^+(x)dx - \int_A f^-(x)dx \\ &= \int_V f^+(x)dx - \int_V f^-(x)dx = \int_V f(x)dx,\end{aligned}$$

as desired. ■

Next let us prove a special case of the divergence theorem, which will also be used in the proof of the general form of the theorem.

Lemma 9.28. *Let $U \subset \mathbb{R}^n$ be a bounded open set with almost C^1 boundary. Suppose f is a C^1 function from an open set containing $\overline{U}$ into $\mathbb{R}$. Also suppose the support of f is compact, and is contained in one of the open sets U_i such that $S_i = \partial U \cap U_i$ is the graph of a C^1 function. Then we have*

$$\int_U D_j f\, dx = \int_{S_i} f\nu_j\, d\sigma,$$

where ν is the unit outward normal to ∂U.

Remark. Note that $D_j f$ is integrable on U, since it is a bounded continuous function and U is Jordan measurable. Also, $f\nu_j$ is bounded and continuous, and since f has compact support, we can ensure that the domain of integration of $f\nu_j$ has finite volume (as we will see in the following proof); so it is integrable over S_i.

Proof. We know that

$$U \cap U_i = \{x : x_n > g(x_1, \dots, x_{n-1})\},$$
$$\partial U \cap U_i = \{x : x_n = g(x_1, \dots, x_{n-1})\},$$

for some C^1 function g (which depends on i). Consider the C^1 change of coordinates Φ, Ψ which straighten $\partial U \cap U_i$. Then we have

$$\Phi(U \cap U_i) \subset \{y_n > 0\}, \qquad \Phi(\partial U \cap U_i) \subset \{y_n = 0\}.$$

Note that the support of f is contained in U_i (which is itself contained in the domain of Φ). Let us find a bounded open set that contains $\text{spt}(f)$ and its closure is in U_i, such that its image under Φ is also bounded. Note that $\partial U_i, \text{spt}(f)$ are disjoint closed sets and one of them is compact. Hence, by Exercise 2.111, the distance between their points has a positive minimum d. Now consider the open cubes with diameter $\frac{d}{2}$ whose center is a point of $\text{spt}(f)$. Then finitely many of these open cubes cover $\text{spt}(f)$. Then the union of these cubes is our desired bounded open set. Let us call this bounded open set W_i (note that W_i is also Jordan measurable). So $U \cap W_i$ is bounded too. Note that $\overline{U \cap W_i}$ is contained in $\overline{W}_i \subset U_i$; so it is contained in the domain of Φ. In addition, $\Phi(U \cap W_i)$ is also bounded, since it is a subset of the compact set $\Phi(\overline{W}_i)$.

Hence by the change of variables theorem and the previous lemma we have

$$\begin{aligned}\int_U D_j f\, dx &= \int_{U\cap W_i} D_j f\, dx && (\text{since } \text{spt}(f) \subset W_i)\\ &= \int_{\Psi(\Phi(U\cap W_i))} D_j f\, dx = \int_{\Phi(U\cap W_i)} (D_j f)\circ\Psi\, |\det D\Psi|\, dy\\ &= \int_{\Phi(U\cap W_i)} (D_j f)\circ\Psi\, dy. && (\text{since } \det D\Psi = 1)\end{aligned}$$

Remember that

$$\begin{cases} x_i = \Psi_i(y) := y_i & \text{for} \quad i < n, \\ x_n = \Psi_n(y) := y_n + g(y_1, \dots, y_{n-1}). \end{cases}$$

Now note that

$$D_n(f \circ \Psi) = \sum\nolimits_k (D_k f) \circ \Psi \cdot D_n \Psi_k = (D_n f) \circ \Psi. \tag{$*$}$$

And thus when $j \neq n$ we have

$$\begin{aligned} D_j(f \circ \Psi) &= \sum\nolimits_k (D_k f) \circ \Psi \cdot D_j \Psi_k \\ &= (D_j f) \circ \Psi + (D_n f) \circ \Psi \cdot D_j g = (D_j f) \circ \Psi + D_j g D_n(f \circ \Psi). \end{aligned}$$

So

$$(D_j f) \circ \Psi = D_j(f \circ \Psi) - D_j g D_n(f \circ \Psi). \tag{$**$}$$

Let $R \subset \{y_n > 0\}$ be a rectangle containing $\Phi(U \cap W_i)$, such that $\partial R \cap \{y_n = 0\}$ is one of the faces of ∂R, and contains $\Phi(\partial U \cap W_i)$. We can decompose R into $[a_j, b_j] \times \hat{R}$, where $\hat{R}$ is the $(n-1)$-dimensional rectangle which is the product of all the edges of R other than its jth edge $[a_j, b_j]$. We can also decompose R into $\tilde{R} \times [0, b_n]$, where $\tilde{R}$ is the $(n-1)$-dimensional rectangle which is the product of all the edges of R other than its nth edge $[0, b_n]$. Note that $\tilde{R} \times \{0\} = \partial R \cap \{y_n = 0\}$ is the face of ∂R described above. Therefore, by $(**)$, for $j \neq n$ we have

$$\begin{aligned} \int_{U \cap W_i} D_j f \, dx &= \int_{\Phi(U \cap W_i)} (D_j f) \circ \Psi \, dy \\ &= \int_{\Phi(U \cap W_i)} D_j(f \circ \Psi) - D_j g D_n(f \circ \Psi) \, dy \\ &= \int_R D_j(f \circ \Psi) - D_j g D_n(f \circ \Psi) \, dy, \end{aligned}$$

where the last equality follows from the previous lemma, noting that $\text{spt}(f \circ \Psi) \subset \Psi^{-1}(\text{spt}(f)) = \Phi(\text{spt}(f)) \subset \Phi(U \cap W_i)$. Now we have

$$\begin{aligned} \int_R D_j(f \circ \Psi) \, dy &= \int_{\hat{R}} \int_{a_j}^{b_j} D_j(f \circ \Psi) \, dy_j d\hat{y} \\ &= \int_{\hat{R}} f(\Psi(b_j, \hat{y})) - f(\Psi(a_j, \hat{y})) \, d\hat{y} = \int_{\hat{R}} 0 \, d\hat{y} = 0, \end{aligned}$$

since $f \circ \Psi$ vanishes on a neighborhood of the faces of ∂R other than $\partial R \cap \{y_n = 0\}$, and the points $(b_j, \hat{y}), (a_j, \hat{y})$ belong to those faces. Note that $D_j(f \circ \Psi)$ is continuous and R is Jordan measurable, so the above integral is a proper Riemann integral, and we can apply Fubini's theorem. Similarly, since g does not depend on y_n we have

$$\begin{aligned}
-\int_R D_j g D_n(f\circ\Psi)\,dy &= -\int_{\tilde{R}} D_j g\int_0^{b_n} D_n(f\circ\Psi)\,dy_n d\tilde{y}\\
&= -\int_{\tilde{R}} D_j g\cdot\big(f(\Psi(\tilde{y},b_n))-f(\Psi(\tilde{y},0))\big)\,d\tilde{y}\\
&= \int_{\tilde{R}} D_j g\cdot f(\Psi(\tilde{y},0))\,d\tilde{y}\\
&= \int_{\tilde{R}} \tfrac{D_j g(\tilde{y})}{\sqrt{1+|Dg(\tilde{y})|^2}}\sqrt{1+|Dg(\tilde{y})|^2}\cdot f(\tilde{y},g(\tilde{y}))\,d\tilde{y}\\
&= \int_{\tilde{R}} \nu_j(\phi(\tilde{y}))J\cdot f(\phi(\tilde{y}))\,d\tilde{y},
\end{aligned}$$

where ν is the unit outward normal to S_i, and J is the volume factor of the parametrization $\phi(\tilde{y})=(\tilde{y},g(\tilde{y}))=\Psi(\tilde{y},0)$ of S_i. Note that Fubini's theorem implies that the integral over $\tilde{R}$ exists.

Now note that $\phi^{-1}=P$, where P is the projection on the first $n-1$ components. Furthermore, $\phi^{-1}(W_i)=\phi^{-1}(\partial U\cap W_i)$ since the image of ϕ is inside ∂U. Hence we have

$$\phi^{-1}(W_i)=\phi^{-1}(\partial U\cap W_i)=P(\partial U\cap W_i)=P(\Phi(\partial U\cap W_i))\subset\tilde{R},$$

where the last equality follows from the fact that $P(\Phi(y))=P(y)=\tilde{y}$ for every point y. Also note that we have $\operatorname{spt}(f\circ\phi)\subset\phi^{-1}(W_i)\subset\phi^{-1}(U_i)$ since $\operatorname{spt}(f)\subset W_i\subset U_i$. Therefore by applying the previous lemma twice and using the definition of integral over the surface patch S_i we get

$$\begin{aligned}
\int_{\tilde{R}}\nu_j f(\phi(\tilde{y}))J\,d\tilde{y} &= \int_{\phi^{-1}(W_i)}\nu_j f(\phi(\tilde{y}))J\,d\tilde{y}\\
&= \int_{\phi^{-1}(U_i)}\nu_j f(\phi(\tilde{y}))J\,d\tilde{y}=\int_{S_i}\nu_j f\,d\sigma.
\end{aligned}$$

Note that the existence of integrals over $\phi^{-1}(W_i)$ and $\phi^{-1}(U_i)$ follows from the previous lemma too.

On the other hand, by $(*)$, when $j=n$ we have

$$\begin{aligned}
\int_R (D_n f)\circ\Psi\,dy &= \int_R D_n(f\circ\Psi)\,dy\\
&= \int_{\tilde{R}}\int_0^{b_n} D_n(f\circ\Psi)\,dy_n d\tilde{y}\\
&= \int_{\tilde{R}} f(\Psi(\tilde{y},b_n))-f(\Psi(\tilde{y},0))\,d\tilde{y}=-\int_{\tilde{R}} f(\Psi(\tilde{y},0))\,d\tilde{y}
\end{aligned}$$

$$\begin{aligned}&= \int_{\tilde{R}} \frac{-1}{\sqrt{1+|Dg|^2}} \sqrt{1+|Dg|^2} \cdot f(\tilde{y}, g(\tilde{y}))\, d\tilde{y} \\ &= \int_{\tilde{R}} \nu_n f(\phi(\tilde{y})) J\, d\tilde{y} = \int_{S_i} \nu_n f\, d\sigma = \int_{S_i} \nu_j f\, d\sigma.\end{aligned}$$

Hence in either case we have shown that

$$\int_U D_j f\, dx = \int_{S_i} f \nu_j\, d\sigma,$$

as desired. ■

Suppose $U \subset \mathbb{R}^n$ is a bounded open set with C^1 boundary. Then every $a \in \partial U$ has a neighborhood U_a such that $U \cap U_a = \{x_n > g(\tilde{x})\}$ and $\partial U \cap U_a = \{x_n = g(\tilde{x})\}$ for some C^1 function g (here $\tilde{x} = (x_1, \ldots, x_{n-1})$). Since ∂U is compact, finitely many of these open neighborhoods, namely $U_1, \ldots, U_m$, cover ∂U. Then $U, U_1, \ldots, U_m$ is an open covering of the compact set $\overline{U}$. Let $\zeta_0, \zeta_1, \ldots, \zeta_m$ be a partition of unity subordinate to this open covering whose sum is 1 on $\overline{U}$. We have

$$\begin{aligned}\int_U D_j f\, dx = \int_U 1 \cdot D_j f\, dx &= \int_U \textstyle\sum_{i=0}^m \zeta_i D_j f\, dx \\ &= \sum \int_U \zeta_i D_j f\, dx = \sum \int_U D_j(\zeta_i f) - f D_j \zeta_i\, dx \\ &= \sum \int_U D_j(\zeta_i f)\, dx - \sum \int_U f D_j \zeta_i\, dx \\ &= \sum \int_U D_j(\zeta_i f)\, dx - \int_U f D_j \big(\sum \zeta_i\big)\, dx \\ &= \sum \int_U D_j(\zeta_i f)\, dx - \int_U f D_j(1)\, dx = \sum \int_U D_j(\zeta_i f)\, dx.\end{aligned}$$

Let Q be a rectangle containing U. Let us denote a point $x \in \mathbb{R}^n$ by $(x_j, \hat{x})$, where $\hat{x}$ is the vector of all the components of x other than x_j. We can also decompose Q into $[a_j, b_j] \times \hat{Q}$, where $\hat{Q}$ is the $(n-1)$-dimensional rectangle which is the product of all the edges of Q other than its jth edge $[a_j, b_j]$. Then we have

$$\begin{aligned}\int_U D_j(\zeta_0 f)\, dx = \int_Q D_j(\zeta_0 f)\, dx &= \int_{\hat{Q}} \int_{a_j}^{b_j} D_j(\zeta_0 f)\, dx_j d\hat{x} \\ &= \int_{\hat{Q}} (\zeta_0 f)(b_j, \hat{x}) - (\zeta_0 f)(a_j, \hat{x})\, d\hat{x} = \int_{\hat{Q}} 0\, d\hat{x} = 0,\end{aligned}$$

since $\zeta_0 f$ vanishes on ∂Q, and the points $(b_j, \hat{x}), (a_j, \hat{x})$ belong to ∂Q.

Next consider a fixed $i \geq 1$. From the proofs of Proposition 9.22 and Theorem 9.14 it is easy to see that we can decompose ∂U into several disjoint $(n-1)$-surface patches $S_1, \dots, S_l$ and a set A with $(n-1)$-measure zero such that $S_1 = \partial U \cap U_i$. Now we know that $\operatorname{spt}(\zeta_i f)$ is compact and is contained in U_i. Hence by the above lemma we have

$$\int_U D_j(\zeta_i f)\,dx = \int_{S_1} \zeta_i f \nu_j\,d\sigma.$$

However by the definition of integral over a rectifiable set we have

$$\int_{\partial U} \zeta_i f \nu_j\,d\sigma = \sum_{i'=1}^{l} \int_{S_{i'}} \zeta_i f \nu_j\,d\sigma = \int_{S_1} \zeta_i f \nu_j\,d\sigma,$$

since $\zeta_i f$ is zero on $S_2, \dots, S_l$. (Note that for each i we need a different decomposition of ∂U to prove the above equality.) Therefore we get

$$\begin{aligned}\int_U D_j f\,dx = \sum_{i\le m} \int_U D_j(\zeta_i f)\,dx &= \sum_{i\le m} \int_{\partial U} \zeta_i f \nu_j\,d\sigma \\ &= \int_{\partial U} f\nu_j \sum_{i\le m} \zeta_i\,d\sigma = \int_{\partial U} f\nu_j\,d\sigma,\end{aligned}$$

since $\sum_{i\le m} \zeta_i = 1$ on ∂U. Thus we have proved the divergence theorem for domains with C^1 boundary. But when the boundary is almost C^1 and has some singularities we also need to analyze the integrals around the singular points of the boundary. This is done in the next theorem.

The Divergence Theorem. *Let $U \subset \mathbb{R}^n$ be a bounded open set with almost C^1 boundary, and suppose* $\operatorname{vol}(\partial U) < \infty$. *Let f be a C^1 function from an open set containing $\overline{U}$ into $\mathbb{R}$. Then we have*

$$\int_U D_j f\,dx = \int_{\partial U} f\nu_j\,d\sigma,$$

where ν is the unit outward normal to ∂U. As a result, if F is a C^1 function from an open set containing $\overline{U}$ into $\mathbb{R}^n$, we have

$$\int_U \operatorname{div} F\,dx = \int_{\partial U} F\cdot\nu\,d\sigma,$$

where $\operatorname{div} F := D_1F_1 + D_2F_2 + \cdots + D_nF_n$ *is the* ***divergence*** *of F.*

Remark. Note that $D_j f$ is integrable on U, since it is a bounded continuous function and U is Jordan measurable. Also, $f\nu_j$ is a bounded function which is continuous on every surface patch of any decomposition of ∂U; thus since ∂U has finite volume, $f\nu_j$ is integrable over ∂U.

Proof. First note that the second part of the theorem follows from its first part; since for a vector-valued function F we have

$$\int_U \operatorname{div} F\,dx = \int_U \sum D_i F_i\,dx = \int_{\partial U} \sum F_i \nu_i\,d\sigma = \int_{\partial U} F\cdot\nu\,d\sigma.$$

So we only need to prove the first statement. We break the proof into several parts to make it more comprehensible, although the parts are intertwined. The idea is to use a partition of unity subordinate to an open covering of $\overline{U}$, such that the singular part of ∂U lies in some open sets whose total volume is small, and their intersection with ∂U has small "area".

(i) Suppose $\partial U = A \sqcup S_1 \sqcup \cdots \sqcup S_m$, where A has $(n-1)$-measure zero, and each S_i is an $(n-1)$-surface patch. We know that there are open sets U_i such that $S_i = \partial U \cap U_i$ is the graph of a C^1 function, and $U\cap U_i$ is the region above or below that graph. Let $\phi_i : V_i \to \mathbb{R}^n$ be the corresponding parametrization of S_i, where $V_i = \phi_i^{-1}(U_i)$ is an open set in $\mathbb{R}^{n-1}$. We know that the volume of S_i is finite, so $\operatorname{vol}(S_i) = \int_{V_i} J_i(x)dx < \infty$, where J_i is the volume factor of ϕ_i. Thus, by the definition of the integral of the nonnegative function J_i, there is a compact Jordan measurable set $K \subset V_i$ such that

$$\int_{V_i} J_i(x)dx - \epsilon < \int_K J_i(x)dx \le \int_{V_i} J_i(x)dx,$$

for a given $\epsilon > 0$. Consider the open set $W_i := V_i - K$. Then we have

$$\int_{W_i} J_i(x)dx < \epsilon, \tag{$\star$}$$

because for any compact Jordan measurable set $\tilde{K} \subset W_i$ we have $\tilde{K}\cap K = \emptyset$, and thus

$$\int_{\tilde{K}} J_i(x)dx + \int_K J_i(x)dx = \int_{\tilde{K}\cup K} J_i(x)dx \le \int_{V_i} J_i(x)dx,$$

since $\tilde{K}\cup K$ is a compact Jordan measurable subset of V_i. Hence

$$\int_{\tilde{K}} J_i(x)dx \le \int_{V_i} J_i(x)dx - \int_K J_i(x)dx < \epsilon.$$

Therefore, by taking supremum over $\tilde{K}$, we can conclude that J_i is integrable over W_i, and its integral satisfies the desired estimate $(\star)$.

Now note that $V_i - W_i = K$ is compact. Therefore $\phi_i(V_i - W_i)$ is a compact subset of ∂U. Hence $\partial U - \bigcup_i \phi_i(V_i - W_i)$ is an open subset of ∂U. Thus there is an open set $W \subset \mathbb{R}^n$ such that $\partial U \cap W = \partial U - \bigcup_i \phi_i(V_i - W_i)$. Note that $\partial U\cap W$ contains A, since $\phi_i(V_i - W_i) \subset S_i$. In other words, the possible singularities of ∂U

are contained in $\partial U \cap W$. We can also assume that W is bounded, since we can consider its intersection with a bounded open set containing $\overline{U}$. In addition note that $S_i \cap W = \phi_i(W_i)$. So its $(n-1)$-dimensional volume satisfies

$$\mathrm{vol}(S_i \cap W) = \int_{W_i} J_i(x)dx < \epsilon, \tag{$*$}$$

due to the estimate $(\star)$.

On the other hand, since A has $(n-1)$-measure zero, for $\delta > 0$ there is a countable family of open cubes $\{R_k\}$ in $\mathbb{R}^n$ that covers A, and satisfies

$$\sum_{k\geq 1} (\mathrm{diam}\, R_k)^{n-1} < \delta.$$

Note that $\mathrm{diam}\, R_k \leq \delta^{\frac{1}{n-1}}$. Now note that A is closed in ∂U. So A is compact. Thus we can assume that the family of cubes has finitely many elements, namely $R_1, \dots, R_l$. Let $\tilde{R}_k$ be the cube with the same center as R_k, and twice the side length. Then $\tilde{R}_1, \dots, \tilde{R}_l$ also cover A, and we have

$$\sum_{k\geq 1} (\mathrm{diam}\, \tilde{R}_k)^{n-1} < 2^{n-1}\delta. \tag{$**$}$$

In addition, note that $A \subset W$, and W is bounded. Hence A and ∂W are disjoint compact sets. Thus the distance between their points has a positive minimum. Therefore we can assume δ is small enough so that each $\tilde{R}_k$ is inside W. We also assume that $2^{n-1}\delta < \epsilon^{n-1}$, so that $\mathrm{diam}\, \tilde{R}_k < \epsilon$.

(ii) Now the open sets $R_1, \dots, R_l, U_1, \dots, U_m, U$ cover the compact set $U \cup \partial U = \overline{U}$. Hence the open sets $\tilde{R}_1, \dots, \tilde{R}_l, U_1, \dots, U_m, U$ also cover $\overline{U}$. Let $\zeta_1, \dots, \zeta_{l+m}, \zeta_0$ be a partition of unity subordinate to the second open covering whose sum is 1 on $\overline{U}$. As explained in the remark after Theorem 9.26, we can furthermore assume that

$$\Big| D_j\Big(\sum_{k\leq l} \zeta_k\Big)\Big| \leq \sum_{k\leq l} \frac{C}{\mathrm{diam}\, \tilde{R}_k} \chi_{\tilde{R}_k}, \tag{$***$}$$

where the constant C only depends on the dimension n.

As we have shown before this theorem we have

$$\int_U D_j f\, dx = \sum_{i=0}^{l+m} \int_U D_j(\zeta_i f)\, dx.$$

We have also seen that $\int_U D_j(\zeta_0 f)\, dx = 0$. Next consider $\int_U D_j(\zeta_{l+i} f)\, dx$, where $1 \leq i \leq m$. Note that the support of $\zeta_{l+i} f$ is compact and is contained in U_i. Hence by the previous lemma we have

$$\int_U D_j(\zeta_{l+i} f)\, dx = \int_{S_i} \zeta_{l+i} f \nu_j\, d\sigma,$$

for $1 \le i \le m$.

Finally note that

$$\begin{aligned}\int_{\partial U} f\nu_j \, d\sigma &= \sum_{i=0}^{l+m} \int_{\partial U} \zeta_i f \nu_j \, d\sigma \\ &= \sum_{i=1}^{m} \int_{\partial U} \zeta_{l+i} f\nu_j \, d\sigma + \sum_{k=1}^{l} \int_{\partial U} \zeta_k f \nu_j \, d\sigma \\ &= \sum_{i=1}^{m} \sum_{i'=1}^{m} \int_{S_{i'}} \zeta_{l+i} f\nu_j \, d\sigma + \sum_{k=1}^{l} \int_{\partial U} \zeta_k f \nu_j \, d\sigma \\ &= \sum_{i=1}^{m} \int_{S_i} \zeta_{l+i} f\nu_j \, d\sigma + \sum_{k=1}^{l} \int_{\partial U} \zeta_k f \nu_j \, d\sigma,\end{aligned}$$

since ζ_{l+i} is zero on $S_{i'}$ for $i' \neq i$, and ζ_0 is zero on ∂U.

(iii) From what we have shown so far we can conclude that

$$\int_U D_j f \, dx - \int_{\partial U} f\nu_j \, d\sigma = \sum_{k \le l} \int_U D_j(\zeta_k f) \, dx - \sum_{k \le l} \int_{\partial U} \zeta_k f \nu_j \, d\sigma.$$

Let C_0, C_1 be upper bounds for $|f|, |Df|$ on $\overline{U}$, respectively. Then by inequalities $(**), (***)$, and the fact that $\operatorname{diam} R_k < \epsilon$ we have

$$\begin{aligned}\left|\sum_{k \le l} \int_U D_j(\zeta_k f) \, dx\right| &= \left|\int_U D_j\Big(\sum_{k\le l} \zeta_k f\Big) \, dx\right| \\ &= \left|\int_U f D_j\Big(\sum_{k\le l} \zeta_k\Big) + \Big(\sum_{k\le l}\zeta_k\Big) D_j f \, dx\right| \\ &\le \int_U |f| \Big|D_j\Big(\sum \zeta_k\Big)\Big| \, dx + \int_U \sum \zeta_k |D_j f| \, dx \\ &\le \int_U |f| \sum \frac{C}{\operatorname{diam} \tilde{R}_k} \chi_{\tilde{R}_k} \, dx + \int_U \sum \chi_{\tilde{R}_k} |D_j f| \, dx \\ &\le \sum \int_{\tilde{R}_k} |f| \frac{C}{\operatorname{diam} \tilde{R}_k} \, dx + \sum \int_{\tilde{R}_k} |D_j f| \, dx \\ &\le \sum |\tilde{R}_k| \big(\frac{C_0 C}{\operatorname{diam} \tilde{R}_k} + C_1\big) \\ &\le \sum \big(C_0 C (\operatorname{diam} \tilde{R}_k)^{n-1} + C_1 (\operatorname{diam} \tilde{R}_k)^n\big) \\ &\le (C_0 C + C_1 \epsilon) \sum (\operatorname{diam} \tilde{R}_k)^{n-1} < (C_0 C + C_1 \epsilon) 2^{n-1} \delta.\end{aligned}$$

Note that we have also used the fact that $0 \leq \zeta_k \leq \chi_{\tilde{R}_k}$, which holds because $\zeta_k \leq 1$ and vanishes outside $\tilde{R}_k$. Also note that since we do not have an upper bound for the number of cubes R_k, we cannot obtain a bound for $\sum(\operatorname{diam}\tilde{R}_k)^{n-1}$ just by knowing that $\operatorname{diam}\tilde{R}_k$ is small for each k; and thus it is essential to have the inequality $(**)$.

On the other hand, we have $\sum_{k\leq l}\zeta_k \leq 1$, and $\sum_{k\leq l}\zeta_k$ is zero outside W. Remember that $S_i\cap W = \phi_i(W_i)$, and W_i is an open subset of $V_i = \phi_i^{-1}(U_i)$. Let J_i be the volume factor of the parametrization ϕ_i. Then, by Lemma 9.27, we have

$$\begin{aligned}\int_{S_i} f\nu_j \sum\zeta_k\, d\sigma &= \int_{V_i}\big(f\nu_j\sum\zeta_k\big)(\phi_i(x))J_i\,dx\\ &= \int_{W_i}\big(f\nu_j\sum\zeta_k\big)(\phi_i(x))J_i\,dx = \int_{S_i\cap W} f\nu_j\sum\zeta_k\,d\sigma,\end{aligned}$$

because $\operatorname{spt}\big((f\nu_j\sum\zeta_k)\circ\phi_i\big)\subset\operatorname{spt}\big((\sum\zeta_k)\circ\phi_i\big)\subset W_i$ as ϕ_i is one-to-one. Therefore since $|\nu_j\sum\zeta_k|\leq 1$ we get

$$\left|\int_{S_i} f\nu_j\sum\zeta_k\,d\sigma\right| = \left|\int_{S_i\cap W} f\nu_j\sum\zeta_k\,d\sigma\right| \leq C_0\operatorname{vol}(S_i\cap W) < C_0\epsilon,$$

where the bound for $\operatorname{vol}(S_i\cap W)$ is given by inequality $(*)$. So

$$\left|\int_{\partial U} f\nu_j\sum\zeta_k\,d\sigma\right| < mC_0\epsilon.$$

Thus we have shown that

$$\left|\int_U D_jf\,dx - \int_{\partial U} f\nu_j\,d\sigma\right| < (C_0C + C_1\epsilon)2^{n-1}\delta + mC_0\epsilon \to 0,$$

as $\epsilon,\delta\to 0$. Hence we get the desired result. ■

Integration by Parts. *Let $U\subset\mathbb{R}^n$ be a bounded open set with almost C^1 boundary. Let f,g be C^1 functions from an open set containing $\overline{U}$ into $\mathbb{R}$. Then we have*

$$\int_U gD_jf\,dx = -\int_U fD_jg\,dx + \int_{\partial U} fg\nu_j\,d\sigma,$$

where ν is the unit outward normal to ∂U.

Remark. A particularly useful case of integration by parts is when $f|_{\partial U} = 0$ or $g|_{\partial U} = 0$, in which case we have

$$\int_U gD_jf\,dx = -\int_U fD_jg\,dx.$$

Proof. By the divergence theorem we have

$$\int_U gD_jf\,dx + \int_U fD_jg\,dx = \int_U D_j(fg)\,dx = \int_{\partial U} fg\nu_j\,d\sigma,$$

which gives the desired. ■

Suppose $n = 2$. Let

$$(x, y) \mapsto (x, y)^\perp = (-y, x)$$

denote the 90 degrees counterclockwise rotation around the origin in $\mathbb{R}^2$. Note that for $z, w \in \mathbb{R}^2$ we have $(z^\perp)^\perp = -z$, and $z^\perp \cdot w^\perp = z \cdot w$. Let $D \subset \mathbb{R}^2$ be a bounded open set with almost C^1 boundary, and let $F = (f, g)$ be a C^1 function from an open neighborhood of $\overline{D}$ into $\mathbb{R}^2$. Then $F^\perp = (-g, f)$. Hence by the divergence theorem we have

$$\begin{aligned}\int_{\partial D} F\cdot\nu^\perp\,ds &= \int_{\partial D} F^\perp\cdot(\nu^\perp)^\perp\,ds = -\int_{\partial D} F^\perp\cdot\nu\,ds\\ &= -\iint_D \operatorname{div} F^\perp\,dxdy = \iint_D g_x - f_y\,dxdy.\end{aligned}$$

Now note that $\mathrm{T} := \nu^\perp$ is tangent to ∂D. If $\phi : I \to \mathbb{R}^2$ is a parametrization of ∂D then ϕ' is also tangent to ∂D. Thus $\frac{\phi'}{|\phi'|} = \pm\mathrm{T}$. Let us assume that the direction along which ϕ traverses ∂D is such that $\frac{\phi'}{|\phi'|} = \mathrm{T}$ holds (otherwise we can consider $\phi(-t)$ instead). Note that the length factor corresponding to ϕ is $\sqrt{\det(\phi'^\intercal\phi')} = \sqrt{|\phi'|^2} = |\phi'|$. Thus we have

$$\begin{aligned}\int_{\partial D} F\cdot\mathrm{T}\,ds &= \int_I F(\phi(t))\cdot\frac{\phi'(t)}{|\phi'(t)|}\,|\phi'(t)|dt\\ &= \int_I F(\phi(t))\cdot\phi'(t)\,dt = \int_I f(\phi(t))\phi_1'(t) + g(\phi(t))\phi_2'(t)\,dt.\end{aligned}$$

Motivated by this formula we define

$$\int_{\partial D} f dx + g dy := \int_{\partial D} F\cdot\mathrm{T}\,ds = \int_{\partial D} F\cdot\nu^\perp\,ds.$$

Therefore we have shown that

$$\int_{\partial D} f dx + g dy = \iint_D g_x - f_y\,dxdy,$$

which is known as the **Green's theorem**.

9.5 Cauchy Integral Theorem and Formula

Let f be a function from a subset of $\mathbb{C}$ into $\mathbb{C}$. Then we have

$$f(z) = u(x,y) + iv(x,y),$$

where $z = x + iy$, and u, v are functions from a subset of $\mathbb{R}^2 = \mathbb{C}$ into $\mathbb{R}$. Let us formally set $dz = dx + idy$, and formally compute

$$f(z)dz = (u+iv)(dx+idy) = udx - vdy + i(vdx + udy).$$

Now let $D \subset \mathbb{C}$ be a bounded open set with almost C^1 boundary. Then, motivated by the above computation, we define

$$\begin{aligned}\int_{\partial D} f(z)dz &:= \int_{\partial D} udx - vdy + i\int_{\partial D} vdx + udy\\ &= \int_{\partial D} (u,-v)\cdot\nu^{\perp}\,ds + i\int_{\partial D}(v,u)\cdot\nu^{\perp}\,ds,\end{aligned}$$

where ν is the unit outward normal to ∂D, and $\nu^{\perp} = (-\nu_2, \nu_1)$ is the unit tangent vector to ∂D. Let $\phi : I \to \mathbb{C}$ be a parametrization of ∂D such that $\frac{\phi'}{|\phi'|} = \mathrm{T} = \nu^{\perp}$. Note that the length factor corresponding to ϕ is $|\phi'|$. Then we have

$$\begin{aligned}\int_{\partial D} f(z)dz &= \int_I \Big((u,-v)\cdot\frac{\phi'(t)}{|\phi'(t)|} + i\,(v,u)\cdot\frac{\phi'(t)}{|\phi'(t)|}\Big)\,|\phi'(t)|dt\\ &= \int_I (u,-v)\cdot\phi'(t) + i\,(v,u)\cdot\phi'(t)\,dt\\ &= \int_I u(\phi(t))\phi_1'(t) - v(\phi(t))\phi_2'(t) + i\big(v(\phi(t))\phi_1'(t) + u(\phi(t))\phi_2'(t)\big)\,dt\\ &= \int_I u(\phi(t))\big(\phi_1'(t) + i\phi_2'(t)\big) + iv(\phi(t))\big(\phi_1'(t) + i\phi_2'(t)\big)\,dt\\ &= \int_I \big(u(\phi(t)) + iv(\phi(t))\big)\big(\phi_1'(t) + i\phi_2'(t)\big)\,dt;\end{aligned}$$

which is a useful formula for computing the integral $\int_{\partial D} f(z)dz$.

Cauchy Integral Theorem. *Let $D \subset \mathbb{C}$ be a bounded open set with almost C^1 boundary. Suppose $f = u + iv$ is a holomorphic function on D, and furthermore, u, v are C^1 function from an open set containing $\overline{D}$ into $\mathbb{R}$. Then we have*

$$\int_{\partial D} f(z)dz = 0.$$

Proof. We have

$$\begin{aligned}\int_{\partial D} f(z)dz &= \int_{\partial D} udx - vdy + i\int_{\partial D} vdx + udy && \text{(by definition)}\\ &= \iint_D (-v)_x - u_y\, dxdy && \text{(by Green's theorem)}\\ &\qquad + i\iint_D u_x - v_y\, dxdy\\ &= \iint_D 0\, dxdy + i\iint_D 0\, dxdy && \text{(by Cauchy-Riemann equations)}\\ &= 0 + i0 = 0,\end{aligned}$$

as desired. ■

Cauchy Integral Formula. *Let $D \subset \mathbb{C}$ be a bounded open set with almost C^1 boundary. Suppose $f = u + iv$ is a holomorphic function on D, and furthermore, u, v are C^1 function from an open set containing $\overline{D}$ into $\mathbb{R}$. Let $z_0 \in D$. Then we have*

$$f(z_0) = \frac{1}{2\pi i}\int_{\partial D} \frac{f(z)}{z - z_0}\, dz.$$

Proof. Since D is open we have $B_{2r}(z_0) \subset D$ for some r. Then $\overline{B}_r(z_0) \subset D$. Let $U = D - \overline{B}_r(z_0)$. Then $\partial U = \partial D \sqcup \partial B_r(z_0)$. Note that the circle $\partial B_r(z_0)$ is 1-rectifiable; so ∂U is almost C^1. In addition, $\frac{f(z)}{z-z_0}$ is holomorphic on $D - \{z_0\}$. Thus by Cauchy integral theorem we have

$$\int_{\partial U} \frac{f(z)}{z - z_0}\, dz = 0.$$

Now note that if we remove one point from the circle $\partial B_r(z_0)$, we can parametrize the remaining part with a C^1 parametrization. We can also decompose ∂D into several C^1 curves and a set with 1-measure zero. Therefore by the definition of integral over rectifiable sets we have

$$0 = \int_{\partial U} \frac{f(z)}{z - z_0}\, dz = \int_{\partial D} \frac{f(z)}{z - z_0}\, dz + \int_{\partial B} \frac{f(z)}{z - z_0}\, dz,$$

where $B = B_r(z_0)$. Note that in both of the above integrals we use tangent vectors which result from 90 degrees counterclockwise rotation of the outward normal to ∂U. In particular, the normal to $\partial B \subset \partial U$ must point to the inside of B, i.e. toward z_0. If we consider ∂B as the boundary of B, and use the outward normal to ∂B which points to the outside of B, then the sign of the corresponding integral will flip, and we obtain

$$\int_{\partial D} \frac{f(z)}{z - z_0}\, dz = \int_{\partial B} \frac{f(z)}{z - z_0}\, dz.$$

Now let $t \mapsto z_0 + re^{it}$ for $t \in (0, 2\pi)$ be a parametrization of $\partial B - \{z_0 + r\}$. Note that the derivative of this parametrization has the desired direction, compatible with the outward normal. Then we have

$$\int_{\partial B} \frac{f(z)}{z - z_0}\, dz = \int_0^{2\pi} \frac{f(z_0 + re^{it})}{re^{it}}\, ire^{it} dt = i \int_0^{2\pi} f(z_0 + re^{it}) dt.$$

Now note that $\int_0^{2\pi} f(z_0) dt = f(z_0) \int_0^{2\pi} 1\, dt = 2\pi f(z_0)$. Hence

$$\begin{aligned} \left| \int_0^{2\pi} f(z_0 + re^{it}) dt - 2\pi f(z_0) \right| &= \left| \int_0^{2\pi} \big(f(z_0 + re^{it}) - f(z_0) \big) dt \right| \\ &\leq \int_0^{2\pi} \big| f(z_0 + re^{it}) - f(z_0) \big| dt \\ &\leq 2\pi \max_{0 \leq t \leq 2\pi} \big| f(z_0 + re^{it}) - f(z_0) \big| \underset{r \to 0}{\longrightarrow} 0, \end{aligned}$$

since f is continuous. Therefore we get the desired formula. ■

Chapter 10

Lebesgue Measure

10.1 Outer Measure

Definition 10.1. A **closed rectangle** in $\mathbb{R}^n$ is a product of n bounded closed intervals, i.e. it is a set of the form

$$[a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n.$$

Similarly, an **open rectangle** in $\mathbb{R}^n$ is a product of n bounded open intervals, i.e. it is a set of the form

$$(a_1, b_1) \times \cdots \times (a_n, b_n) \subset \mathbb{R}^n.$$

In general, a **rectangle** R in $\mathbb{R}^n$ is a product of n bounded intervals, i.e. there are bounded intervals $I_1, I_2, \ldots, I_n \subset \mathbb{R}$ such that

$$R := I_1 \times I_2 \times \cdots \times I_n \subset \mathbb{R}^n.$$

Each interval I_i can be closed, open, or half-open. The intervals I_i are called the **edges** of R. Let a_i, b_i be the left endpoint and the right endpoint of I_i respectively. Then $b_i - a_i$ is the length of the interval I_i. The rectangle R is called a **cube** if $b_i - a_i = b_1 - a_1$ for all $i \leq n$. When $n = 2$, cubes are called **squares**. The **volume** of the rectangle R is the positive real number

$$|R| := (b_1 - a_1) \cdots (b_n - a_n).$$

When $n = 1, 2$, the volume is called the **length** or the **area**, respectively. The points $(c_1, \ldots, c_n)$ where each c_i is either a_i or b_i, are called the **vertices** of the rectangle R.

Remark. Note that a closed rectangle is closed, being a product of closed sets; and an open rectangle is open, being a product of open sets. Let R be the rectangle in the above definition. It is easy to see that

$$\overline{R} = [a_1, b_1] \times \cdots \times [a_n, b_n], \qquad R^\circ = (a_1, b_1) \times \cdots \times (a_n, b_n).$$

In other words, the closure of a rectangle is a closed rectangle, and the interior of a rectangle is an open rectangle. As a result we have

$$\partial R = \overline{R} - R^\circ = \bigcup_{i \le n} [a_1, b_1] \times \cdots \times \{a_i, b_i\} \times \cdots \times [a_n, b_n].$$

Definition 10.2. A **partition** P of an interval $[a, b] \subset \mathbb{R}$ is a finite set of points $\{c_0, \ldots, c_m\}$ such that

$$a = c_0 < c_1 < \cdots < c_m = b.$$

The interval $[c_{i-1}, c_i]$ is called the ith **subinterval** of the partition P.

Definition 10.3. A **partition** of the closed rectangle

$$R = [a_1, b_1] \times \cdots \times [a_n, b_n],$$

is a cartesian product $P := P_1 \times \cdots \times P_n$, where each P_i is a partition of the interval $[a_i, b_i]$. Suppose $[c_i, d_i]$ is a subinterval of the partition P_i, then the closed rectangle

$$[c_1, d_1] \times \cdots \times [c_n, d_n]$$

is called a **subrectangle** of the partition P. If P_i divides $[a_i, b_i]$ into N_i subintervals, then P divides R into $N_1 \cdots N_n$ subrectangles. We denote these subrectangles by R_α, where α is the multi-index $(\alpha_1, \ldots, \alpha_n)$ such that $1 \le \alpha_i \le N_i$. In this notation, R_α denotes the subrectangle $I_{\alpha_1} \times \cdots \times I_{\alpha_n}$, where I_{α_i} is the α_ith subinterval of P_i. We sometimes abuse the notation and write $P = \{R_\alpha\}$.

Theorem 10.4. *Let $R \subset \mathbb{R}^n$ be a rectangle.*

(i) *Suppose $P = \{R_\alpha\}$ is a partition of $\overline{R}$. Then we have*

$$|R| = \sum |R_\alpha|.$$

(ii) *Suppose $R_1, \ldots, R_m$ are rectangles in $\mathbb{R}^n$, and $R \subset \bigcup R_i$. Then*

$$|R| \le \sum |R_i|.$$

(iii) *Suppose $R_1, \ldots, R_m \subset R$ are rectangles that have pairwise disjoint interiors, i.e. $R_i^\circ \cap R_k^\circ = \emptyset$ for every $i \ne k$. Then*

$$\sum |R_i| \le |R|.$$

If in addition we have $\overline{R} = \bigcup \overline{R}_i$, then

$$|R| = \sum |R_i|.$$

$\boxed{\textbf{Proof.}}$ Let $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$. In the following, we will use the fact that the closure of a rectangle is a closed rectangle that has the same volume as the original rectangle.

(i) Suppose $P = P_1 \times \cdots \times P_n$. Let us denote the subrectangles of P by $R_\alpha = I_{\alpha_1} \times \cdots \times I_{\alpha_n}$, where I_{α_i} is a subinterval of P_i and $1 \leq \alpha_i \leq N_i$. Then we have

$$\sum |R_\alpha| = \sum_{\alpha_1=1}^{N_1} \cdots \sum_{\alpha_n=1}^{N_n} |I_{\alpha_1}| \cdots |I_{\alpha_n}| = \prod_{i=1}^{n} \Big(\sum_{\alpha_i=1}^{N_i} |I_{\alpha_i}| \Big) = \prod_{i=1}^{n} (b_i - a_i) = |R|.$$

(ii) Let us denote R by R_0 to simplify the notation. First we assume that all the rectangles are closed. Suppose

$$R_i = [a_{i1}, b_{i1}] \times \cdots \times [a_{in}, b_{in}].$$

Let

$$\hat{R} = I_1 \times \cdots \times I_n.$$

be a closed rectangle that contains $R_0, R_1, \dots, R_m$. Let Q_j be a partition of I_j that contains a_{ij}, b_{ij} for all $i \leq m$. Then for a fixed i, $Q_j \cap [a_{ij}, b_{ij}]$ is a partition of $[a_{ij}, b_{ij}]$. Now $Q = \prod_{j \leq n} Q_j$ is a partition of $\hat{R}$ that contains all the vertices of every R_i. Then note that

$$Q \cap R_i = \Big(\prod_{j \leq n} Q_j \Big) \bigcap \Big(\prod_{j \leq n} [a_{ij}, b_{ij}] \Big) = \prod_{j \leq n} (Q_j \cap [a_{ij}, b_{ij}])$$

is a partition of R_i. Suppose $\{S_\alpha\}$ is the set of subrectangles of Q. Then the set of subrectangles of $Q \cap R_i$ is

$$\{S_\alpha : S_\alpha \subset R_i\}.$$

To see this note that each subinterval of $Q_j \cap [a_{ij}, b_{ij}]$ is also a subinterval of Q_j. Thus the subrectangles of $Q \cap R_i$ belong to the set $\{S_\alpha\}$. The subrectangles of $Q \cap R_i$ are also obviously subsets of R_i, so they belong to $\{S_\alpha : S_\alpha \subset R_i\}$. On the other hand, suppose S_α is a subrectangle of Q such that $S_\alpha = \prod_{j \leq n} I_{\alpha_j} \subset R_i$. Then the endpoints of I_{α_j} are between a_{ij}, b_{ij}, and also belong to Q_j. So I_{α_j} is a subinterval of $Q_j \cap [a_{ij}, b_{ij}]$. Therefore S_α is a subrectangle of $Q \cap R_i$ as desired. As a result we have

$$\sum_{S_\alpha \subset R_i} |S_\alpha| = |R_i|.$$

Now we have

$$\{S_\alpha : S_\alpha \subset R_0\} \subset \bigcup_{i=1}^{m} \{S_\alpha : S_\alpha \subset R_i\}. \tag{$*$}$$

Because if $S_\alpha = \prod_{j\le n} I_{\alpha_j} \subset R_0$, then S_α° intersects R_i for some i. But we have $R_i = \overline{R_i^\circ}$. Therefore for $x \in S_\alpha^\circ \cap R_i$, there is a sequence $x_l \in R_i^\circ$ that converges to x. But then we must have $x_l \in S_\alpha^\circ$ for large enough l, since S_α° is an open neighborhood of x. Thus S_α° intersects R_i°. So I_{α_j} intersects (a_{ij}, b_{ij}) for all $j \le n$. But we cannot have $I_{\alpha_j} \not\subset [a_{ij}, b_{ij}]$, since that would imply $a_{ij} \in I_{\alpha_j}^\circ$ or $b_{ij} \in I_{\alpha_j}^\circ$, which is impossible due to the fact that a subinterval of a partition cannot contain a point of the partition in its interior. Hence $I_{\alpha_j} \subset [a_{ij}, b_{ij}]$ for all $j \le n$, and therefore $S_\alpha \subset R_i$. Thus we finally obtain that

$$|R| = |R_0| = \sum_{S_\alpha \subset R_0} |S_\alpha| \le \sum_{i=1}^m \sum_{S_\alpha \subset R_i} |S_\alpha| = \sum_{i=1}^m |R_i|.$$

Note that the volume of each $S_\alpha \subset R_0$ appears at least once in the right hand side of the above inequality, due to $(*)$. At the end, suppose that the rectangles are not necessarily closed. Then we have $\overline{R} \subset \bigcup_{i\le m} \overline{R}_i$, since $\bigcup_{i\le m} \overline{R}_i$ is a closed set that contains R. Hence we have

$$|R| = |\overline{R}| \le \sum |\overline{R}_i| = \sum |R_i|.$$

(iii) First we assume that all the rectangles are closed. Suppose

$$R_i = [a_{i1}, b_{i1}] \times \cdots \times [a_{in}, b_{in}].$$

Similarly to the above, let Q be a partition of R that contains all the vertices of every R_i. Then $Q \cap R_i$ is a partition of R_i whose set of subrectangles is $\{S_\alpha : S_\alpha \subset R_i\}$. Hence we have

$$\sum_{S_\alpha \subset R_i} |S_\alpha| = |R_i|.$$

In addition, for $i \ne k$ we have

$$\{S_\alpha : S_\alpha \subset R_i\} \cap \{S_\alpha : S_\alpha \subset R_k\} = \emptyset.$$

Because if $S_\alpha \subset R_i$ then $S_\alpha^\circ \subset R_i^\circ$. Hence $S_\alpha^\circ \cap R_k^\circ = \emptyset$. So we cannot have $S_\alpha \subset R_k$, since S_α° is nonempty. Therefore we get

$$\sum_{i=1}^m |R_i| = \sum_{i=1}^m \sum_{S_\alpha \subset R_i} |S_\alpha| \le \sum_{\text{all } \alpha} |S_\alpha| = |R|.$$

Note that in the left hand side of the above inequality, no $|S_\alpha|$ can appear more than once. Now suppose that the rectangles are not necessarily closed. Then we have $\overline{R}_i \subset \overline{R}$. We also have

$$(\overline{R}_i)^\circ \cap (\overline{R}_k)^\circ = R_i^\circ \cap R_k^\circ = \emptyset,$$

since the interior of the closure of a rectangle equals the corresponding open rectangle, which is equal to the interior of the original rectangle. Hence we have

$$\sum |R_i| = \sum |\overline{R}_i| \leq |\overline{R}| = |R|.$$

Finally, the last statement of the theorem follows easily from the previous parts. Because by part (ii) we have

$$|R| = |\overline{R}| \leq \sum |\overline{R}_i| = \sum |R_i|.$$

Also, by the first statement of part (iii) we have $|R| \geq \sum |R_i|$. Hence the result follows. ■

Remark. Note that part (ii) of the above theorem is nontrivial even when $n = 1$. Also note that when $n > 1$, in the last statement of part (iii), $\{R_i\}$ is not necessarily the set of subrectangles of some partition of R. So it is not apparent how to deduce this fact, although it is geometrically obvious.

Notation. Let $A \subset \mathbb{R} \cup \{\infty\}$ be a nonempty set. If $A - \{\infty\} \neq \emptyset$ we define

$$\inf A := \inf(A - \{\infty\}).$$

Otherwise, if $A = \{\infty\}$ we define $\inf A := \infty$.

Definition 10.5. The **(Lebesgue) outer measure** of $A \subset \mathbb{R}^n$ is

$$m^*(A) := \inf \Bigg\{ \sum_{i\geq 1} |Q_i| \,:\, \{Q_i\}_{i\geq 1} \text{ is a countable family} \\ \text{of open cubes such that } A \subset \bigcup_{i\geq 1} Q_i \Bigg\}.$$

Remark. Remember that a countable set is either finite or countably infinite. Also, note that for any countable family of open cubes $\{Q_i\}$ that covers A, we have $0 \leq \sum |Q_i| \leq \infty$. Hence $0 \leq m^*(A) \leq \infty$.

Remark. Note that the concept of outer measure depends on the dimension n. For example the interval $[0, 1]$ has outer measure one as a subset of $\mathbb{R}$, but if we regard it as the subset $[0, 1] \times \{0\}$ of $\mathbb{R}^2$, it has outer measure zero. We will prove these later in this section.

Proposition 10.6. *Let $A \subset \mathbb{R}^n$. Then we have*

$$m^*(A) := \inf \Bigg\{ \sum_{i\geq 1} |R_i| \,:\, \{R_i\}_{i\geq 1} \textit{ is a countable family} \\ \textit{of open rectangles such that } A \subset \bigcup_{i\geq 1} R_i \Bigg\}.$$

Remark. Note that in the definition of outer measure, and also in the above proposition, we have only considered cubes and rectangles whose edges are parallel to the coordinate axes.

Remark. In the above proposition, instead of open rectangles, we can use closed rectangles or general rectangles. The proof is similar to what follows. But the above version is more useful. Note that the above version is not a trivial consequence of the similar statement for general rectangles.

Proof. Let a be the infimum of $\sum |R_i|$, where $\{R_i\}$ is a countable family of open rectangles that covers A. It is obvious that $a \le m^*(A)$, since every open cube is also an open rectangle. In particular, if $a = \infty$ then $m^*(A) = \infty = a$. So suppose that $a < \infty$. We need to show that $m^*(A) \le a$. Let $\{R_i\}$ be a family of open rectangles that covers A, such that $\sum_{i\ge1} |R_i| < a + \epsilon$, for a given $\epsilon > 0$. Note that this is possible since a is the infimum of the sum of the volumes of rectangles, and therefore $a + \epsilon$ is not a lower bound for them. Consider a fixed i, and suppose we have

$$R_i = (a_1, b_1) \times \cdots \times (a_n, b_n).$$

Let $l_j := b_j - a_j$, and let l be a positive number less than $\min_{j\le n} l_j$. Then we have $k_j := \lfloor \frac{l_j}{l} \rfloor \in \mathbb{N}$. We also have $k_j l \le l_j < (k_j + 1)l = k_j l + l$. Now consider the open rectangle

$$S_i = (a_1, a_1 + k_1 l + l) \times \cdots \times (a_n, a_n + k_n l + l).$$

Obviously we have $R_i \subset S_i$. Furthermore, due to the continuity of the multiplication, we can take l to be small enough so that

$$|S_i| = \prod_{j\le n} (k_j l + l) \le \prod_{j\le n} (l_j + l) \le \Big(\prod_{j\le n} l_j \Big) + \frac{\epsilon}{2^i} = |R_i| + \frac{\epsilon}{2^i}.$$

Now each interval $[a_j, a_j + k_j l + l]$ has a partition with $k_j + 1$ closed subintervals of length l. Then we get a partition of $\overline{S}_i$ with $N_i := \prod_{j\le n}(k_j + 1)$ subrectangles, which are all closed cubes with volume l^n. Note that by Theorem 10.4, the volume of S_i is the sum of the volume of these closed cubes, i.e. it is $N_i l^n$. We can cover each of these closed cubes by an open cube whose volume is less than $l^n + \frac{\epsilon}{N_i 2^i}$. Call these open cubes Q_{ij}, where $j \le N_i$. Then we have $\sum_{j\le N_i} |Q_{ij}| < |S_i| + \frac{\epsilon}{2^i}$.

We can repeat the above construction for every i, to get a countable family $\{S_i\}$ of open rectangles that covers A, such that

$$\sum_{i\ge1} |S_i| \le \sum_{i\ge1} |R_i| + \sum_{i\ge1} \frac{\epsilon}{2^i} < a + 2\epsilon.$$

Now $\{Q_{ij} : i \ge 1,\ j \le N_i\}$ is a family of open cubes that covers A; and it is also countable, since it is the union of countably many finite families. We consider this family with the order

$$Q_{11}, Q_{12}, \dots, Q_{1N_1}, Q_{21}, \dots, Q_{2N_2}, \dots, Q_{m1}, \dots, Q_{mN_m}, \dots.$$

Let us denote the kth cube in this sequence by Q_k. Then for $N \le N_1 + \dots + N_m$ we have

$$\sum_{k=1}^{N} |Q_k| \le \sum_{i=1}^{m} \sum_{j \le N_i} |Q_{ij}| < \sum_{i=1}^{m} |S_i| + \sum_{i=1}^{m} \frac{\epsilon}{2^i} < \sum_{i \ge 1} |S_i| + \sum_{i \ge 1} \frac{\epsilon}{2^i} < a + 3\epsilon.$$

By taking the limit as $N \to \infty$ we obtain $\sum_{k\ge1} |Q_k| \le a + 3\epsilon$. Thus we have $m^*(A) \le a + 3\epsilon$. Now as ϵ is arbitrary, we get $m^*(A) \le a$ as desired. ■

Proposition 10.7. *Let $a \in \mathbb{R}$. Then the $(n-1)$-dimensional plane $\{x \in \mathbb{R}^n : x_i = a\}$ has outer measure zero in $\mathbb{R}^n$.*

Proof. Let P be the described plane. Then we have $P \subset \bigcup_{j\ge1} R_j$, where R_j is the open rectangle $\prod_{k\le n} I_k$ in which $I_k = (-2^{j-1}, 2^{j-1})$ for $k \ne i$, and

$$I_i = (a - \epsilon 2^{-nj-1}, a + \epsilon 2^{-nj-1}).$$

Now we have $|R_j| = \epsilon 2^{-nj} 2^{j(n-1)} = \epsilon 2^{-j}$. Hence

$$\sum_{j\ge1} |R_j| = \epsilon \sum_{j\ge1} 2^{-j} = \epsilon.$$

Therefore P has outer measure zero, since ϵ is arbitrary. ■

Notation. $[0, \infty] := [0, \infty) \cup \{\infty\}$.

Remark. In the rest of this chapter, we have to deal with series of the form $\sum a_k$ where $a_k \in [0, \infty]$. Now if $a_k = \infty$ for some k, then the series diverges to ∞ by definition. Otherwise, the series converges to a finite nonnegative number if its partial sums are bounded, and diverges to ∞ if its partial sums are unbounded. So such series either converge, or diverge to ∞.

Theorem 10.8. *Let $\mathcal{P}(\mathbb{R}^n)$ be the set of all subsets of $\mathbb{R}^n$. Then the Lebesgue outer measure is a function $m^* : \mathcal{P}(\mathbb{R}^n) \to [0, \infty]$ that satisfies*

(i) *The outer measure of the empty set is zero, i.e. $m^*(\emptyset) = 0$.*

(ii) *The outer measure is **monotone**, i.e. if A, B are subsets of $\mathbb{R}^n$ such that $A \subset B$, then $m^*(A) \le m^*(B)$.*

(iii) *The outer measure is **countably subadditive**, i.e. if $\{A_k\}_{k\ge1}$ is a countable family of subsets of $\mathbb{R}^n$, then*

$$m^*\Big(\bigcup_{k\ge1} A_k\Big) \le \sum_{k\ge1} m^*(A_k).$$

Remark. Note that the countable family $\{A_k\}$ can be finite too. In this case, the countable subadditivity is called *(finite) subadditivity.*

$\boxed{\text{Proof.}}$ **(i)** Any open cube covers $\emptyset$. Hence we can cover $\emptyset$ by open cubes with arbitrarily small volume. Thus $m^*(\emptyset) = 0$.

(ii) Any countable family of open cubes that covers B also covers A. Hence we get the desired.

(iii) If $\sum_k m^*(A_k) = \infty$, then the inequality holds trivially. So suppose that $\sum_k m^*(A_k) < \infty$. Then we have $m^*(A_k) < \infty$ for all k. Let $\epsilon > 0$ be given. Then we can cover each A_k with a countable family of open cubes $\{Q_{ki}\}_{i\geq 1}$ such that

$$\sum_{i\geq 1} |Q_{ki}| < m^*(A_k) + \frac{\epsilon}{2^k}.$$

Then $\{Q_{ki}\}_{i,k\geq 1}$ is a countable family of open cubes that covers $\bigcup_k A_k$, and

$$\sum_{i,k\geq 1} |Q_{ki}| \leq \sum_{k\geq 1} m^*(A_k) + \sum_{k\geq 1} \frac{\epsilon}{2^k} \leq \sum_{k\geq 1} m^*(A_k) + \epsilon.$$

Therefore $m^*(\bigcup A_k) \leq \sum_{k\geq 1} m^*(A_k) + \epsilon$, and as ϵ is arbitrary we get the desired. Note that $m^*(\bigcup A_k)$ can be ∞ too. ■

Remark. If we want to be completely rigorous in the above proof, we have to arrange the family of open cubes $\{Q_{ki}\}_{i,k\geq 1}$ into a sequence. Note that different arrangements do not change the sum of the volumes of the family, since the volume of each cube is positive and therefore their series is absolutely convergent. Now suppose we have arranged the family as the sequence $\{Q_j\}_{j\geq 1}$. Then for any $N \in \mathbb{N}$ there is $M \in \mathbb{N}$ such that

$$\{Q_j\}_{1\leq j\leq N} \subset \{Q_{ki}\}_{1\leq i,k\leq M}.$$

Then we have

$$\sum_{j\leq N} |Q_j| \leq \sum_{k\leq M} \sum_{i\leq M} |Q_{ki}| < \sum_{k\leq M} m^*(A_k) + \sum_{k\leq M} \frac{\epsilon}{2^k} < \sum_{k\geq 1} m^*(A_k) + \epsilon.$$

Now by taking the limit as $N \to \infty$ we get $\sum_{j\geq 1} |Q_j| \leq \sum_{k\geq 1} m^*(A_k) + \epsilon$ as desired.

Remark. It is a trivial consequence of the above theorem that if $A \subset \mathbb{R}^n$, and $\{A_k\}_{k\geq 1}$ is a countable family of subsets of $\mathbb{R}^n$ such that $A \subset \bigcup_{k\geq 1} A_k$, then we have

$$m^*(A) \leq \sum_{k\geq 1} m^*(A_k).$$

***Example* 10.9.** It is easy to see that the outer measure of a single point is zero, since we can cover that point by open cubes with arbitrarily small volume. Then the above theorem implies that the outer measure of any countable subset of $\mathbb{R}^n$ is also zero.

Theorem 10.10. *Let R be a rectangle in $\mathbb{R}^n$. Then we have*

$$m^*(R) = |R|.$$

Proof. First suppose that R is closed. Now note that $m^*(R) \le |R|$. Because for any $\epsilon > 0$, we can cover R by a single open rectangle whose volume is less than $|R| + \epsilon$. Thus $m^*(R) \le |R| + \epsilon$, and since ϵ is arbitrary we get the desired. In particular we see that $m^*(R)$ is finite. To show the reverse inequality, let $\{R_i\}$ be a family of open rectangles that covers R, such that for a given $\epsilon > 0$ we have $\sum_{i\ge1} |R_i| < m^*(R)+\epsilon$. Then finitely many of R_i's will cover R, since R is compact. Therefore there is N such that $R \subset \bigcup_{i=1}^{N} R_i$. Thus by Theorem 10.4 we have

$$|R| \le \sum_{i=1}^{N} |R_i| \le \sum_{i\ge1} |R_i| < m^*(R) + \epsilon.$$

As ϵ is arbitrary we get $|R| \le m^*(R)$, as desired.

Next suppose R is a general rectangle. Then we have

$$m^*(R) \le m^*(\overline{R}) = |\overline{R}| = |R|.$$

On the other hand, for every $\epsilon > 0$ there is a closed rectangle $S \subset R$ such that $|S| > |R| - \epsilon$. Hence

$$m^*(R) \ge m^*(S) = |S| > |R| - \epsilon.$$

Thus we get $m^*(R) \ge |R|$, since ϵ was arbitrary. ■

***Example* 10.11.** As a particular case of the above theorem, we see that the outer measure of a bounded interval in $\mathbb{R}$ equals its length. It is also easy to see that the outer measure of an unbounded interval is ∞, since an unbounded interval contains bounded intervals with arbitrarily large lengths.

Theorem 10.12. *Suppose $A \subset \mathbb{R}^n$. Then for any $x \in \mathbb{R}^n$ we have*

$$m^*(A + x) = m^*(A),$$

where $A + x := \{a + x : a \in A\}$.

Proof. First note that if Q is an open cube, then $Q+x$ is also an open cube such that $|Q+x| = |Q|$. Let $\{Q_i\}$ be a countable family of open cubes that covers A. Then $\{Q_i + x\}$ is a countable family of open cubes that covers $A+x$. Hence we have

$$m^*(A+x) \leq \sum |Q_i + x| = \sum |Q_i|.$$

By taking the infimum over all families $\{Q_i\}$ covering A, we get $m^*(A+x) \leq m^*(A)$. The reverse inequality follows similarly since $A = (A+x) + (-x)$. ■

Theorem 10.13. *Let $A \subset \mathbb{R}^n$. Then we have*

$$m^*(A) = \inf\{m^*(U) : U \textit{ is open, and } U \supset A\}.$$

Proof. If $U \supset A$ then $m^*(A) \leq m^*(U)$. Therefore $m^*(A) \leq \inf\{m^*(U)\}$. Now if $m^*(A) = \infty$ then the equality holds trivially. So suppose $m^*(A) < \infty$. Let $\epsilon > 0$. Then there is a countable family of open cubes $\{Q_i\}$ that covers A, and $\sum |Q_i| < m^*(A) + \epsilon$. Now $\tilde{U} := \bigcup Q_i$ is an open set containing A such that

$$m^*(\tilde{U}) \leq \sum m^*(Q_i) = \sum |Q_i| < m^*(A) + \epsilon.$$

Hence $\inf\{m^*(U)\} < m^*(A)+\epsilon$; and as ϵ is arbitrary we have $\inf\{m^*(U)\} \leq m^*(A)$ as desired. ■

Definition 10.14. Let $A, B \subset \mathbb{R}^n$. The distance of the two sets A, B is

$$d(A,B) := \inf\{|a-b| : a \in A, b \in B\}.$$

Remark. Recall that the diameter of a nonempty set $A \subset \mathbb{R}^n$ is

$$\text{diam}(A) := \sup\{|x-y| : x, y \in A\}.$$

We can easily show that the diameter of a rectangle whose edges have lengths $l_1, \dots, l_n$ is $\sqrt{l_1^2 + \cdots + l_n^2}$. We will use this fact in the proof of the next theorem.

Theorem 10.15. *Let $A, B \subset \mathbb{R}^n$. If $d(A,B) > 0$ then*

$$m^*(A \cup B) = m^*(A) + m^*(B).$$

Remark. Note that when $d(A,B) > 0$ then $A \cap B = \emptyset$. But by merely assuming that $A \cap B = \emptyset$ we cannot deduce that $m^*(A \cup B) = m^*(A) + m^*(B)$. The counterexamples will be discussed later.

Proof. If $m^*(A)$ or $m^*(B)$ is ∞, then $m^*(A \cup B) = \infty$ due to the monotonicity of the outer measure. Hence the equality holds trivially. So suppose $m^*(A)$ and $m^*(B)$ are finite. Then

$$m^*(A \cup B) \leq m^*(A) + m^*(B) < \infty.$$

To prove the reverse inequality, let $\{Q_i\}$ be a countable family of open cubes that covers $A \cup B$. Furthermore suppose that

$$\sum |Q_i| < m^*(A \cup B) + \epsilon,$$

for a given $\epsilon > 0$. Now for each i there is a partition of $\overline{Q}_i$ such that each subrectangle of the partition has diameter less than some given $\delta > 0$. This can be achieved by simply dividing each edge of Q_i into subintervals of equal length less than $\frac{\delta}{\sqrt{n}}$. Let R_{ij} for $j \leq N_i$ denote the subrectangles of this partition of $\overline{Q}_i$. Then we have

$$\sum_{j \leq N_i} |R_{ij}| = |Q_i|.$$

Now $\{R_{ij} : i \geq 1, j \leq N_i\}$ is a countable family of closed cubes that covers $A \cup B$. We consider this family with the order

$$R_{11}, R_{12}, \dots, R_{1N_1}, R_{21}, \dots, R_{2N_2}, \dots, R_{m1}, \dots, R_{mN_m}, \dots.$$

Let us denote the kth cube in this sequence by R_k. Then for $N \leq N_1 + \dots + N_m$ we have

$$\sum_{k=1}^{N} |R_k| \leq \sum_{i=1}^{m} \sum_{j \leq N_i} |R_{ij}| = \sum_{i=1}^{m} |Q_i|.$$

Thus $\sum_{k=1}^{N} |R_k| \leq \sum_{i \geq 1} |Q_i|$, and therefore $\sum_{k \geq 1} |R_k| \leq \sum_{i \geq 1} |Q_i|$. Similarly for $N \geq N_1 + \dots + N_{m-1}$ we have $\sum_{k=1}^{N} |R_k| \geq \sum_{i=1}^{m-1} |Q_i|$. Hence we can obtain similarly that $\sum_{k \geq 1} |R_k| \geq \sum_{i \geq 1} |Q_i|$. Therefore

$$\sum_{k \geq 1} |R_k| = \sum_{i \geq 1} |Q_i|.$$

Now let $J := \{k : R_k \cap A \neq \emptyset\}$. We claim that

$$A \subset \bigcup_{k \in J} R_k, \qquad B \subset \bigcup_{k \notin J} R_k.$$

The first inclusion is obvious since $\{R_k\}_{k \geq 1}$ covers $A \cup B$. Hence those cubes that intersect A must cover A. For the second inclusion we have to use the fact that

$d(A,B) > 0$. Note that we have not used this fact so far. Since $d(A,B) > 0$, there is $\delta > 0$ such that $d(A,B) > \delta$. Suppose we have used this δ in the above construction. Then no R_k can intersect both A and B, since otherwise the diameter of that R_k would have been greater than δ, contrary to our assumption. Thus if $R_k \cap B \neq \emptyset$ then $R_k \cap A = \emptyset$, and thus $k \notin J$. But those cubes that intersect B must cover B. Hence those cubes that do not intersect A will also cover B. Finally, due to the subadditivity of the outer measure we have

$$\begin{aligned} m^*(A) + m^*(B) &\le \sum_{k\in J} m^*(R_k) + \sum_{k\notin J} m^*(R_k) \\ &= \sum_{k\in J} |R_k| + \sum_{k\notin J} |R_k| \\ &= \sum_{k\ge 1} |R_k| = \sum_{i\ge 1} |Q_i| < m^*(A\cup B) + \epsilon. \end{aligned}$$

As ϵ was arbitrary we get $m^*(A) + m^*(B) \le m^*(A\cup B)$, as desired. ■

Remark. In the above proof we used the equality

$$\sum_{k\in J} |R_k| + \sum_{k\notin J} |R_k| = \sum_{k\ge 1} |R_k|.$$

First note that if $k_1 < k_2 < \dots$ denote the elements of J, then $\sum_{k\in J} |R_k|$ is by definition $\sum_{j\ge 1} |R_{k_j}|$. We can similarly define $\sum_{k\notin J} |R_k|$. It is also easy to see that these series are convergent, since their sequences of partial sums are sequences of positive numbers, and are bounded by $\sum_{k\ge 1} |R_k|$.

Now note that $\sum_{k\in J} |R_k| = \sum_{k\ge 1} a_k$, and $\sum_{k\notin J} |R_k| = \sum_{k\ge 1} b_k$, where

$$a_k := \begin{cases} |R_k| & k\in J, \\ 0 & k\notin J, \end{cases} \qquad b_k := \begin{cases} 0 & k\in J, \\ |R_k| & k\notin J. \end{cases}$$

To see this, let $L := \sum_{k\in J} |R_k|$, and let s_i denotes the ith partial sum of $\sum a_k$. Then for $k_m \le i < k_{m+1}$ we have

$$s_i = \sum_{k=1}^{i} a_k = \sum_{j=1}^{m} |R_{k_j}|.$$

Thus if M is large enough so that for $m \ge M$ we have

$$\left| \sum_{j=1}^{m} |R_{k_j}| - L \right| < \epsilon,$$

for a given $\epsilon > 0$, then for $i \geq k_m$ we have $|s_i - L| < \epsilon$. Hence $s_i \to L$. The case of $\sum b_k$ is similar. Finally we have

$$\sum_{k\in J} |R_k| + \sum_{k\notin J} |R_k| = \sum_{k\geq 1} a_k + \sum_{k\geq 1} b_k = \sum_{k\geq 1} (a_k + b_k) = \sum_{k\geq 1} |R_k|.$$

10.2 Measurable Sets

Suppose A is a bounded subset of $\mathbb{R}^n$, and R is an open rectangle containing A. We can define the *inner measure* of A to be

$$m_*(A) := |R| - m^*(R - A).$$

It can be shown that this definition does not depend on the rectangle R. When the set A is not too complex, we expect that its inner and outer measures are equal, i.e.

$$m^*(A) = m_*(A).$$

Lebesgue took this as the definition of measurability. If A has this property we have

$$m^*(R \cap A) + m^*(R \cap A^c) = m^*(A) + m^*(R - A) = |R| = m^*(R).$$

Caratheodory modified this definition, and allowed arbitrary sets in place of R. We can think of it as a localization of Lebesgue's definition. Caratheodory's definition is easier to work with, does not assume the boundedness of A, and is more suitable for generalization. We will use his definition throughout this chapter.

Definition 10.16. A set $A \subset \mathbb{R}^n$ is called **(Lebesgue) measurable** if for every $X \subset \mathbb{R}^n$ we have

$$m^*(X) = m^*(X \cap A) + m^*(X \cap A^c).$$

Remark. Note that we only need to check that $m^*(X) \geq m^*(X \cap A) + m^*(X \cap A^c)$, since the reverse inequality always holds due to the subadditivity of the outer measure. Also note that this inequality is trivially true when $m^*(X) = \infty$. So without loss of generality we can always assume that $m^*(X) < \infty$.

Proposition 10.17. *If $A \subset \mathbb{R}^n$ is measurable then A^c is measurable too.*

Proof. For every $X \subset \mathbb{R}^n$ we have

$$m^*(X) = m^*(X \cap A) + m^*(X \cap A^c) = m^*(X \cap A^c) + m^*(X \cap (A^c)^c),$$

since $(A^c)^c = A$. ■

Proposition 10.18. *Suppose $A \subset B \subset \mathbb{R}^n$, and A is measurable. If $m^*(A) < \infty$, then we have*

$$m^*(B - A) = m^*(B) - m^*(A).$$

Remark. Note that we do not need the measurability of B for this equation to hold. But the measurability of A is necessary, as we will show later. Also note that the equation holds if in particular we have $m^*(B) < \infty$, since that implies $m^*(A) \leq m^*(B) < \infty$.

Proof. The measurability of A implies

$$m^*(B) = m^*(B \cap A) + m^*(B \cap A^c) = m^*(A) + m^*(B - A).$$

Now we can subtract $m^*(A)$ from both sides, since it is finite. ■

Definition 10.19. Let $\mathcal{M}$ be the set of all measurable subsets of $\mathbb{R}^n$. The **(Lebesgue) measure** is

$$m := m^*|_{\mathcal{M}} : \mathcal{M} \to [0, \infty].$$

So if $A \subset \mathbb{R}^n$ is measurable, its (Lebesgue) measure is $m(A) := m^*(A)$.

Theorem 10.20. *Suppose $Z \subset \mathbb{R}^n$, and $m^*(Z) = 0$. Then Z is measurable.*

Remark. In particular, the countable subsets of $\mathbb{R}^n$ are measurable. Also, the empty set $\emptyset$ is measurable too.

Proof. For every $X \subset \mathbb{R}^n$ we have $m^*(X \cap Z) \leq m^*(Z) = 0$, since $X \cap Z \subset Z$. Thus $m^*(X \cap Z) = 0$. Hence we have

$$m^*(X \cap Z^c) + m^*(X \cap Z) = m^*(X \cap Z^c) \leq m^*(X),$$

since $X \cap Z^c \subset X$. Therefore Z is measurable. ■

Theorem 10.21. *Suppose $A, B \subset \mathbb{R}^n$ are measurable. Then $A \cup B$, $A \cap B$, and $A - B$ are measurable too.*

Remark. As a consequence, we can show by a simple induction that if $A_1, \ldots, A_k$ are measurable, then $\bigcup_{j=1}^k A_j$ and $\bigcap_{j=1}^k A_j$ are also measurable.

Proof. Let $X \subset \mathbb{R}^n$. Then

$$m^*(X) = m^*(X \cap A) + m^*(X \cap A^c). \tag{$\star$}$$

Now we can use the measurability of B to obtain

$$\begin{aligned} m^*(X \cap A) &= m^*(X \cap A \cap B) + m^*(X \cap A \cap B^c), && (*) \\ m^*(X \cap A^c) &= m^*(X \cap A^c \cap B) + m^*(X \cap A^c \cap B^c). && (**) \end{aligned}$$

On the other hand

$$\begin{aligned} X \cap (A \cup B) &= X \cap \big((A - B) \cup (A \cap B) \cup (B - A)\big) \\ &= \big(X \cap (A \cap B^c)\big) \cup \big(X \cap (A \cap B)\big) \cup \big(X \cap (B \cap A^c)\big). \end{aligned}$$

Hence we have

$$\begin{aligned} m^*(X \cap (A \cup B)) &\leq m^*(X \cap A \cap B^c) + m^*(X \cap A \cap B) \\ &\qquad + m^*(X \cap B \cap A^c) \\ &= m^*(X \cap A) + m^*(X \cap B \cap A^c). && \text{by } (*) \end{aligned}$$

We also have $X \cap (A \cup B)^c = X \cap A^c \cap B^c$. Therefore if we add the outer measure of this set to both sides of the above inequality, we get

$$\begin{aligned} & m^*(X \cap (A \cup B)) + m^*(X \cap (A \cup B)^c) \\ &\qquad \leq m^*(X \cap A) + m^*(X \cap B \cap A^c) + m^*(X \cap A^c \cap B^c) \\ &\qquad = m^*(X \cap A) + m^*(X \cap A^c) && \text{by } (**) \\ &\qquad = m^*(X). && \text{by } (\star) \end{aligned}$$

Thus $A \cup B$ is measurable. Now as A^c, B^c are measurable $A^c \cup B^c$ is measurable too. Hence $A \cap B = (A^c \cup B^c)^c$ is also measurable. Finally, $A - B = A \cap B^c$ is measurable too. ■

Remark. Suppose $A_1 \subset \mathbb{R}^n$ is measurable. Then $A_2 := A_1^c$ is measurable too, and we have $A_1 \cap A_2 = \emptyset$ and $A_1 \cup A_2 = \mathbb{R}^n$. Furthermore, for any $X \subset \mathbb{R}^n$ we have

$$m^*(X) = m^*(X \cap A_1) + m^*(X \cap A_2).$$

The following theorem is a generalization of this property.

Theorem 10.22. *Suppose $\{A_j\}_{j\geq 1}$ is a countable family of measurable subsets of $\mathbb{R}^n$, which are pairwise disjoint i.e. $A_i \cap A_k = \emptyset$ for every $i \neq k$. Let $A := \bigcup_{j\geq 1} A_j$. Then for any $X \subset \mathbb{R}^n$ we have*

$$m^*(X \cap A) = \sum_{j\geq 1} m^*(X \cap A_j).$$

Remark. Note that in this theorem, X need not be measurable.

Proof. Let $B_k := \bigcup_{j=1}^{k} A_j$. First we show by induction that for all $k \in \mathbb{N}$ we have

$$m^*(X \cap B_k) = \sum_{j=1}^{k} m^*(X \cap A_j).$$

The case of $k = 1$ is trivial. So suppose the claim holds for some k. Then by using the measurability of A_{k+1} we get

$$\begin{aligned} m^*(X \cap B_{k+1}) &= m^*\big(X \cap (B_k \cup A_{k+1})\big) \\ &= m^*\big(X \cap (B_k \cup A_{k+1}) \cap A_{k+1}\big) + m^*\big(X \cap (B_k \cup A_{k+1}) \cap A_{k+1}^c\big). \end{aligned}$$

Now we can use the following set theoretic identities

$$(B_k \cup A_{k+1}) \cap A_{k+1} = A_{k+1}, \qquad (B_k \cup A_{k+1}) \cap A_{k+1}^c = B_k \cap A_{k+1}^c.$$

In addition we have $B_k \subset A_{k+1}^c$, since $B_k \cap A_{k+1} = \emptyset$. Hence we have $B_k \cap A_{k+1}^c = B_k$. Therefore

$$\begin{aligned} m^*(X \cap B_{k+1}) &= m^*(X \cap A_{k+1}) + m^*(X \cap B_k) \\ &= m^*(X \cap A_{k+1}) + \sum_{j=1}^{k} m^*(X \cap A_j) = \sum_{j=1}^{k+1} m^*(X \cap A_j). \end{aligned}$$

If $\{A_j\}$ is a finite family, there is nothing left to prove. So let us assume that $\{A_j\}$ is a countably infinite family. Now note that

$$X \cap A = X \cap \bigcup_{j\geq 1} A_j = \bigcup_{j\geq 1} (X \cap A_j).$$

Thus $m^*(X \cap A) \leq \sum_{j=1}^{\infty} m^*(X \cap A_j)$ due to the countable subadditivity of the outer measure. If $m^*(X \cap A) = \infty$ then the equality holds trivially. So suppose $m^*(X \cap A) < \infty$. Thus $m^*(X \cap A_j) < \infty$ for all j, since $X \cap A_j \subset X \cap A$. On the other hand we have $X \cap B_k \subset X \cap A$ for all k. Hence

$$\sum_{j=1}^{k} m^*(X \cap A_j) = m^*(X \cap B_k) \leq m^*(X \cap A).$$

By taking the limit as $k \to \infty$ we obtain $\sum_{j=1}^{\infty} m^*(X \cap A_j) \leq m^*(X \cap A)$. Note that the series is convergent, since its terms are nonnegative real numbers, and its partial sums are bounded. Thus we finally get the desired equality. ■

Theorem 10.23. *Suppose $\{A_j\}_{j\geq 1}$ is a countable family of measurable subsets of $\mathbb{R}^n$. Then $\bigcup_{j\geq 1} A_j$ and $\bigcap_{j\geq 1} A_j$ are measurable too.*

Proof. When the family $\{A_j\}$ is finite, we have seen that the union and the intersection of its elements are measurable. So we assume that $\{A_j\}$ is a countably infinite family. We inductively define $B_1 := A_1$, and $B_{k+1} := A_{k+1} - \bigcup_{j\leq k} A_j$. Since the union of finitely many measurable sets is measurable, and the difference of two measurable sets is measurable, $\{B_j\}_{j\geq 1}$ is a countably infinite family of measurable sets. In addition, note that $B_i \cap B_k = \emptyset$ when $i \neq k$. Because we have $B_i \subset A_i$, so if $i < k$ then $B_i \cap B_k \subset A_i \cap B_k = \emptyset$. Furthermore, we have

$$\bigcup_{j\geq 1} B_j = \bigcup_{j\geq 1} A_j.$$

It is obvious that $\bigcup B_j \subset \bigcup A_j$, since $B_j \subset A_j$ for each j. For the reverse inclusion, let $x \in \bigcup A_j$. Then $x \in A_j$ for some j. Let k be the smallest positive integer such that $x \in A_k$. If $k = 1$ then $x \in A_1 = B_1 \subset \bigcup B_j$. If $k > 1$ then $x \notin A_j$ for $j < k$. Hence $x \in A_k - \bigcup_{j<k} A_j = B_k \subset \bigcup B_j$. Thus we have $\bigcup A_j \subset \bigcup B_j$ as desired.

Now suppose $X \subset \mathbb{R}^n$. Let $A := \bigcup_{j\geq 1} A_j$. We want to show that

$$m^*(X) \geq m^*(X \cap A) + m^*(X \cap A^c).$$

We can assume that $m^*(X) < \infty$, since otherwise the above inequality holds trivially. We showed that $A = \bigcup_{j\geq 1} B_j$, where $\{B_j\}$ is a countable family of pairwise disjoint measurable sets. Let $C_k := \bigcup_{j=1}^{k} B_j$. Then by Theorem 10.22 we have

$$m^*(X \cap A) = \sum_{j\geq 1} m^*(X \cap B_j), \qquad m^*(X \cap C_k) = \sum_{j=1}^{k} m^*(X \cap B_j).$$

In addition we have $m^*(X \cap A^c) \leq m^*(X \cap C_k^c)$, since $C_k \subset A$, and therefore $A^c \subset C_k^c$. But C_k is measurable. Hence

$$\begin{aligned} m^*(X) &= m^*(X \cap C_k) + m^*(X \cap C_k^c) \\ &\geq \sum_{j=1}^{k} m^*(X \cap B_j) + m^*(X \cap A^c). \end{aligned}$$

Now by taking the limit as $k \to \infty$ we obtain

$$m^*(X) \geq \sum_{j=1}^{\infty} m^*(X \cap B_j) + m^*(X \cap A^c) = m^*(X \cap A) + m^*(X \cap A^c).$$

Note that the above series is convergent, since its terms are nonnegative finite real numbers, and its partial sums are bounded. Thus A is measurable as desired. Finally, for the intersection note that we have

$$\bigcap_{j\geq 1} A_j = \Big(\bigcup_{j\geq 1} A_j^c\Big)^c.$$

So the measurability of $\bigcap A_j$ follows, since the complement of a measurable set is measurable too. ■

Definition 10.24. Let $\mathcal{A}$ be a family of subsets of a set X. We say $\mathcal{A}$ is a ***σ*-algebra** on X, if it satisfies the following conditions.

(i) $\phi \in \mathcal{A}$.

(ii) $\mathcal{A}$ is closed under complement, i.e. if $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$.

(iii) $\mathcal{A}$ is closed under countable union, i.e. if $\{A_j\}_{j\geq 1}$ is a countable family of elements of $\mathcal{A}$, then $\bigcup_{j\geq 1} A_j \in \mathcal{A}$.

Remark. Note that a σ-algebra is also closed under countable intersection, since $\bigcap_{j\geq 1} A_j = \big(\bigcup_{j\geq 1} A_j^c\big)^c$. So in particular, a σ-algebra is closed under finite union and finite intersection too. But in general, a σ-algebra is not closed under uncountable union or uncountable intersection.

***Example* 10.25.** Let $\mathcal{M}$ be the family of measurable subsets of $\mathbb{R}^n$. What we have proved so far implies that $\mathcal{M}$ is a σ-algebra.

Theorem 10.26. *Let $\mathcal{M}$ be the set of all measurable subsets of $\mathbb{R}^n$. Then the Lebesgue measure is a function $m : \mathcal{M} \to [0, \infty]$ that satisfies*

(i) *The measure of the empty set is zero, i.e. $m(\emptyset) = 0$.*

(ii) *The measure is **countably additive**, i.e. if $\{A_j\}_{j\geq 1} \subset \mathcal{M}$ is a countable family of measurable sets, which are pairwise disjoint i.e. $A_i \cap A_k = \emptyset$ for every $i \neq k$, then we have*

$$m\Big(\bigcup_{j\geq 1} A_j\Big) = \sum_{j\geq 1} m(A_j).$$

Remark. Note that the countable family $\{A_j\}$ can be finite too. In this case, the countable additivity is called *(finite) additivity.*

Proof. For (i) just note that $\emptyset$ is measurable, and $m(\emptyset) = m^*(\emptyset) = 0$. Now to prove (ii) let $A := \bigcup_{j\geq 1} A_j$. We know that A is measurable. By Theorem 10.22, we also know that for any $X \subset \mathbb{R}^n$ we have $m^*(X \cap A) = \sum_{j\geq 1} m^*(X \cap A_j)$. If we substitute $X = A$ in this equation we get

$$\begin{aligned} m(A) = m^*(A) = m^*(A \cap A) &= \sum_{j\geq 1} m^*(A \cap A_j) \\ &= \sum_{j\geq 1} m^*(A_j) = \sum_{j\geq 1} m(A_j). \end{aligned}$$

Note that $A_j \cap A = A_j$, since $A_j \subset A$. ■

Remark. A sequence of sets $\{A_j\}_{j=1}^{\infty}$ is called *increasing* if $A_{j+1} \subset A_j$ for all j, and it is called *decreasing* if $A_{j+1} \subset A_j$ for all j.

Theorem 10.27. *Let $\{A_j\}$ be a sequence of measurable subsets of $\mathbb{R}^n$.*

(i) *If $\{A_j\}$ is an increasing sequence, then we have*

$$m\Big(\bigcup_{j\geq 1} A_j\Big) = \lim_{j\to\infty} m(A_j).$$

(ii) *If $\{A_j\}$ is a decreasing sequence, and $m(A_1) < \infty$, then we have*

$$m\Big(\bigcap_{j\geq 1} A_j\Big) = \lim_{j\to\infty} m(A_j).$$

Remark. The above theorem is known as the *continuity of measure*. The reason is that we can consider the union of an increasing sequence of sets as their limit. Similarly, we can consider the intersection of a decreasing sequence of sets as their limit.

Proof. (i) Let $B_1 := A_1$, and $B_{j+1} := A_{j+1} - A_j$ for $j \geq 1$. Note that $A_j = \bigcup_{i\leq j} A_i$, since $\{A_j\}$ is an increasing sequence. Then as we saw in the proof of Theorem 10.23, $\{B_j\}$ is a sequence of pairwise disjoint measurable sets, such that

$$A := \bigcup_{j\geq 1} A_j = \bigcup_{j\geq 1} B_j.$$

If $m(A_k) = \infty$ for some k, then $m(A_j) = \infty$ for all $j \geq k$, since $A_k \subset A_j$. Also we have $m(A) = \infty$. Thus $m(A_j) \to m(A)$ as desired. So suppose $m(A_j) < \infty$ for all j. Then $m(B_j) < \infty$ for all j, since $B_j \subset A_j$. We also have

$$m(A) = \sum_{j\geq 1} m(B_j).$$

But it is easy to see that $A_j = \bigcup_{i\leq j} B_i$. Thus $m(A_j) = \sum_{i=1}^{j} m(B_i)$. Therefore $m(A_j)$ is the jth partial sum of the series $\sum m(B_i)$. Hence $m(A_j)$ converges to the limit of the series, i.e. $m(A_j) \to m(A)$.

(ii) Let $C_j := A_1 - A_j$ for $j \geq 1$. Then $\{C_j\}$ is an increasing sequence of measurable sets, because $\{A_j\}$ is a decreasing sequence. We also have $\bigcup_{j\geq 1} C_j = A_1 - A$, where $A := \bigcap_{j\geq 1} A_j$. To see this note that $C_j = A_1 - A_j \subset A_1 - A$. Hence $\bigcup C_j \subset A_1 - A$. On the other hand, if $x \in A_1 - A$ then $x \notin A$. Thus there is k such that $x \notin A_k$. So $x \in A_1 - A_k = C_k \subset \bigcup C_j$. Therefore $A_1 - A \subset \bigcup C_j$ as desired. Hence by the previous part we get

$$\lim m(C_j) = m(A_1 - A) = m(A_1) - m(A).$$

Note that we have used Proposition 10.18 and the fact that $m(A_1) < \infty$. For the same reason we have $m(C_j) = m(A_1) - m(A_j)$. Also, note that $m(A), m(A_j), m(C_j)$

are all finite, since A, A_j, C_j are all subsets of A_1. Therefore we have

$$\begin{aligned}\lim m(A_j) &= \lim \big(m(A_1) - m(C_j)\big)\\ &= \lim m(A_1) - \lim m(C_j)\\ &= m(A_1) - \big(m(A_1) - m(A)\big) = m(A).\end{aligned}$$ ■

Remark. Note that in the second part of the above theorem we can replace the assumption of $m(A_1) < \infty$ with $m(A_j) < \infty$ for some j. The proof is the same, since we can simply ignore the sets $A_1, \dots, A_{j-1}$ in the sequence. But unlike the first part, the second part of the above theorem does not hold without assuming that some A_j has finite measure. For example the intervals (n, ∞) have infinite measure, while their intersection $\bigcap_{n\geq 1}(n, \infty) = \emptyset$ has measure zero.

Theorem 10.28. *Suppose $A \subset \mathbb{R}^n$ is measurable. Then for any $x \in \mathbb{R}^n$, $A + x$ is measurable too, and we have*

$$m(A + x) = m(A).$$

Proof. It suffices to show that $A+x$ is measurable, then the equality of measures of $A, A+x$ follows from the equality of their outer measures. Let us write $B := A+x$ to simplify the notation. Let $X \subset \mathbb{R}^n$. We have to show that

$$m^*(X) = m^*(X \cap B) + m^*(X \cap B^c). \tag{$*$}$$

Let $Y := X + (-x)$. Then we know that

$$m^*(Y) = m^*(Y \cap A) + m^*(Y \cap A^c), \tag{$**$}$$

since A is measurable. But we have

$$(Y \cap A) + x = X \cap B, \qquad (Y \cap A^c) + x = X \cap B^c.$$

Let us prove the second equality, the first one is similar. Let $y \in Y \cap A^c$. Then $y \notin A$, and $y = z - x$ for some $z \in X$. Hence $y + x = z \in X$. Also we cannot have $y + x \in B$, since that would have implied $y = (y + x) - x \in B + (-x) = A$. Thus $y + x \in B^c$. Therefore we have $(Y \cap A^c) + x \subset X \cap B^c$. The reverse inclusion can be proved similarly. Finally we get

$$m^*(Y \cap A) = m^*(X \cap B), \qquad m^*(Y \cap A^c) = m^*(X \cap B^c),$$

since translations preserve the outer measure. For the same reason we also have $m^*(Y) = m^*(X)$. Now if we plug these values into $(**)$ we obtain $(*)$ as desired. ■

Proposition 10.29. *Let $a \in \mathbb{R}$. Then the open half space $\{x \in \mathbb{R}^n : x_i > a\}$ is measurable.*

Proof. Let $X \subset \mathbb{R}^n$. We have to show that

$$m^*(X) \geq m^*(X \cap \{x_i > a\}) + m^*(X \cap \{x_i \leq a\}).$$

We can assume that $m^*(X) < \infty$, since otherwise the above inequality holds trivially. First we assume that $X \cap \{x_i = a\} = \emptyset$. Let

$$X^+ := X \cap \{x_i > a\}, \qquad X^- := X \cap \{x_i < a\} = X \cap \{x_i \leq a\}.$$

Let $\{R_j\}$ be a countable family of open rectangles that covers X, such that

$$\sum_{j \geq 1} |R_j| < m^*(X) + \epsilon,$$

for a given $\epsilon > 0$. For each j let $R_j^+ := R_j \cap \{x_i > a\}$, and $R_j^- := R_j \cap \{x_i < a\}$. It is obvious that $R_j^\pm$ are also open rectangles. In fact, their closures are subrectangles of a partition of $\overline{R}_j$. Hence $|R_j| = |R_j^+| + |R_j^-|$. Furthermore we have

$$X^+ = \{x_i > a\} \cap X \subset \{x_i > a\} \cap \bigcup_{j \geq 1} R_j = \bigcup_{j \geq 1} (\{x_i > a\} \cap R_j) = \bigcup_{j \geq 1} R_j^+.$$

Similarly we have $X^- \subset \bigcup_{j \geq 1} R_j^-$. Therefore

$$\begin{aligned} m^*(X^+) + m^*(X^-) &\leq \sum_{j \geq 1} |R_j^+| + \sum_{j \geq 1} |R_j^-| \\ &= \sum_{j \geq 1} (|R_j^+| + |R_j^-|) = \sum_{j \geq 1} |R_j| < m^*(X) + \epsilon. \end{aligned}$$

As ϵ is arbitrary we get the desired.

Now suppose $X \cap \{x_i = a\} \neq \emptyset$. Let $Z := X \cap \{x_i = a\}$, and $Y := X \cap \{x_i \neq a\}$. Then Z has measure zero, since the $(n-1)$-dimensional plane $\{x_i = a\}$ has measure zero. Also $Y \cap \{x_i = a\} = \emptyset$, so

$$m^*(Y \cap \{x_i > a\}) + m^*(Y \cap \{x_i < a\}) \leq m^*(Y).$$

Therefore by the subadditivity of the outer measure we get

$$\begin{aligned} &m^*(X \cap \{x_i > a\}) + m^*(X \cap \{x_i \leq a\}) \\ &\qquad \leq m^*(X \cap \{x_i > a\}) + m^*(X \cap \{x_i < a\}) + m^*(Z) \\ &\qquad = m^*(Y \cap \{x_i > a\}) + m^*(Y \cap \{x_i < a\}) \\ &\qquad \leq m^*(Y) \leq m^*(X). \end{aligned}$$ ■

Proposition 10.30. *The open rectangles in $\mathbb{R}^n$ are measurable.*

Proof. Let $R = (a_1, b_1) \times \cdots \times (a_n, b_n)$ be an open rectangle. Then we have

$$R = \bigcap_{i=1}^{n} (\{x_i > a_i\} \cap \{x_i < b_i\}).$$

Now each open half space $\{x_i > a_i\}$ or $\{x_i > b_i\}$ is measurable. Hence $\{x_i \le b_i\} = \{x_i > b_i\}^c$ is measurable too. In addition, $\{x_i = b_i\}$ is measurable, since it has measure zero. Thus $\{x_i < b_i\} = \{x_i \le b_i\} - \{x_i = b_i\}$ is also measurable. Therefore R is measurable, since it is the intersection of finitely many measurable sets. ■

Theorem 10.31. *The open and closed subsets of $\mathbb{R}^n$ are measurable.*

Proof. Let $U \subset \mathbb{R}^n$ be an open set. Then for any $x \in U$ there is an open ball B_x whose center is x, such that $B_x \subset U$. Let Q_x be an open cube centered at x that is contained in B_x. Hence $x \in Q_x \subset U$. Now $\{Q_x : x \in U\}$ is an open covering of U, i.e. $U \subset \bigcup_{x \in U} Q_x$. Thus by theorem 11.57, this open covering has a countable subcovering, i.e. there are countably many points $x_1, x_2, \dots \in U$ such that $U \subset \bigcup_{j \ge 1} Q_{x_j}$. But $Q_{x_j} \subset U$ for each j. Hence $U = \bigcup_{j \ge 1} Q_{x_j}$. Therefore U is the union of countably many open cubes, which are measurable. So U is also measurable.

Finally, closed sets are the complements of open sets, which are measurable. Hence closed sets are measurable too. ■

Remark. The Jordan measurable subsets of $\mathbb{R}^n$ are also Lebesgue measurable. Because a set S is Jordan measurable if its boundary ∂S has measure zero. But we have

$$S = S^\circ \cup (\partial S \cap S),$$

where S° is the interior of S. Then S° is Lebesgue measurable since it is open. Also, $\partial S \cap S$ is Lebesgue measurable since it has measure zero. Hence S is Lebesgue measurable too.

***Example* 10.32.** Every rectangle $R \subset \mathbb{R}^n$ is Lebesgue measurable, since it is Jordan measurable. Because ∂R is contained in the union of finitely many $(n-1)$-dimensional planes, so it has measure zero. Note that we do not require the rectangle to be open or closed. As a consequence we have

$$m(R) = m^*(R) = |R|.$$

In other words, the Lebesgue measure of a rectangle is its volume.

Proposition 10.33. *The intersection of a nonempty family of σ-algebras on a set X is a σ-algebra on X.*

Proof. The intersection obviously contains $\emptyset$. It is also easy to check that the intersection is closed under complement and countable union. ■

Definition 10.34. The σ-algebra **generated** by a family $\mathcal{A}$ of subsets of a set X is the intersection of all σ-algebras containing $\mathcal{A}$.

Remark. Note that the power set of X is a σ-algebra containing $\mathcal{A}$, so the family of all σ-algebras containing $\mathcal{A}$ is nonempty, and its intersection is defined.

Proposition 10.35. *Suppose $\mathcal{F}$ is the σ-algebra generated by a family $\mathcal{A}$. Then $\mathcal{F}$ is contained in any σ-algebra containing $\mathcal{A}$. In other words, $\mathcal{F}$ is the smallest σ-algebra with respect to inclusion that contains $\mathcal{A}$.*

Proof. $\mathcal{F}$ is the intersection of all σ-algebras containing $\mathcal{A}$, therefore it is a subset of any one of them. ■

Definition 10.36. A subset of $\mathbb{R}^n$ is called $\boldsymbol{G_\delta}$ if it is the intersection of countably many open sets. And a subset of $\mathbb{R}^n$ is called $\boldsymbol{F_\sigma}$ if it is the union of countably many closed sets.

Definition 10.37. The **Borel $\boldsymbol{\sigma}$-algebra** is the σ-algebra generated by the family of open subsets of $\mathbb{R}^n$, i.e. it is the smallest σ-algebra that contains all open sets. A **Borel set** is a subset of $\mathbb{R}^n$ that belongs to the Borel σ-algebra.

Remark. Note that every closed set is Borel, since the Borel σ-algebra is closed under complement. Also, every G_δ set or F_σ set is Borel, since Borel σ-algebra is closed under countable union and countable intersection. But there are Borel sets that are not among any of these sets. On the other hand, every Borel set is measurable. Because the family of measurable sets is a σ-algebra that contains the open sets. Hence it contains the smallest σ-algebra containing open sets, i.e. the Borel σ-algebra.

Theorem 10.38. *Let $A \subset \mathbb{R}^n$. Then the following assertions are equivalent.*

(i) *A is measurable.*

(ii) *For every $\epsilon > 0$ there is an open set $U \supset A$ such that $m^*(U - A) < \epsilon$.*

(iii) *There is a G_δ set $G \supset A$ such that $m^*(G - A) = 0$.*

(iv) *For every $\epsilon > 0$ there is a closed set $C \subset A$ such that $m^*(A - C) < \epsilon$.*

(v) *There is an F_σ set $F \subset A$ such that $m^*(A - F) = 0$.*

Remark. The above properties are known as the *regularity of Lebesgue measure.*

Remark. Note that by Theorem 10.13, for any set $A \subset \mathbb{R}^n$ and every $\epsilon > 0$, there is an open set $U \supset A$ such that $m^*(U) \leq m^*(A) + \epsilon$. But this does not imply part (ii) of the above theorem, even if $m^*(A) < \infty$. Because in general we only have

$$m^*(U - A) \geq m^*(U) - m^*(A),$$

and the equality requires the measurability of A.

$\boxed{\textbf{Proof.}}$ **(i) $\implies$ (ii):** First we assume that $m^*(A) < \infty$. Then by Theorem 10.13 we know that for every $\epsilon > 0$, there is an open set $U \supset A$ such that $m^*(U) < m^*(A) + \epsilon$. Hence by Proposition 10.18 we have

$$m^*(U - A) = m^*(U) - m^*(A) < \epsilon,$$

since A is measurable and has finite measure. Now suppose that $m^*(A) = \infty$. Let $A_j := A \cap R_j$, where R_j is the open cube centered at the origin whose edges have length j. Note that each A_j is measurable and we have $A = \bigcup_{j\geq 1} A_j$. Also note that $m^*(A_j) \leq m^*(R_j) < \infty$, since $A_j \subset R_j$. Hence there is an open set $U_j \supset A_j$ such that $m^*(U_j - A_j) < \frac{\epsilon}{2^j}$. Now let $U := \bigcup_{j\geq 1} U_j$. Then U is open, and we have $A = \bigcup_{j\geq 1} A_j \subset \bigcup_{j\geq 1} U_j = U$. Furthermore

$$U - A = \Big(\bigcup_{j\geq 1} U_j\Big) \cap A^c = \bigcup_{j\geq 1} (U_j \cap A^c) = \bigcup_{j\geq 1} (U_j - A_j).$$

Hence by subadditivity of the outer measure we get

$$m^*(U - A) \leq \sum_{j\geq 1} m^*(U_j - A_j) < \sum_{j\geq 1} \frac{\epsilon}{2^j} \leq \epsilon.$$

(ii) $\implies$ (iii): Note that here we do not assume that A is measurable. For every $j \in \mathbb{N}$ there is an open set $U_j \supset A$ such that $m^*(U_j - A) < \frac{1}{j}$. Let $G := \bigcap_{j\geq 1} U_j$. Then G is a G_δ set containing A. We also have $G - A \subset U_j - A$ for every j. Hence

$$m^*(G - A) \leq m^*(U_j - A) < \frac{1}{j} \implies m^*(G - A) = 0.$$

(iii) $\implies$ (i): We have $A = G - Z$, where $Z := G - A$. Then $m^*(Z) = 0$, so Z is measurable. On the other hand G is measurable too, since it is G_δ. Thus A is also measurable.

(i) $\implies$ (iv): We know that A^c is measurable too. Thus there is an open set $U \supset A^c$ such that $m^*(U - A^c) < \epsilon$, for a given $\epsilon > 0$. Because we have shown that (i) is equivalent to (ii). Now let $C := U^c$. Then C is closed, and $C \subset (A^c)^c = A$. We also have

$$A - C = A \cap C^c = A \cap U = U \cap (A^c)^c = U - A^c.$$

Hence $m^*(A - C) < \epsilon$.

(iv) $\implies$ (v): For every $j \in \mathbb{N}$ there is a closed set $C_j \subset A$ such that $m^*(A - C_j) < \frac{1}{j}$. Let $F := \bigcup_{j\geq 1} C_j$. Then F is an F_σ set contained in A. We also have $A - F \subset A - C_j$ for every j. Hence

$$m^*(A - F) \leq m^*(A - C_j) < \frac{1}{j} \implies m^*(A - F) = 0.$$

(v) $\Longrightarrow$ (i): We have $A = F \cup Z$, where $Z := A - F$. Then $m^*(Z) = 0$, so Z is measurable. On the other hand F is measurable too, since it is F_σ. Thus A is also measurable. ■

Theorem 10.39. *Suppose $A \subset \mathbb{R}^n$ is measurable. Then we have*

$$m(A) = \sup\{m(C) : C \textit{ is closed, and } C \subset A\}.$$

$\boxed{\textbf{Proof.}}$ This is a simple consequence of the previous theorem. Let

$$\mathcal{C} := \{m(C) : C \text{ is closed, and } C \subset A\}.$$

We know that for every $\epsilon > 0$ there is a closed set $C \subset A$ such that $m(A - C) < \epsilon$. But we have $A = C \cup (A - C)$. Hence $m(A) = m(C) + m(A - C)$, since C and $A - C$ are disjoint and measurable. Now if $m(A) = \infty$ then $m(C) = \infty$, since $m(A - C)$ is finite. Thus $m(A) = \sup \mathcal{C}$ as desired. So let us assume that $m(A) < \infty$. Then $m(C) < \infty$ too, and we have

$$m(C) = m(A) - m(A - C) > m(A) - \epsilon.$$

Therefore $m(A) - \epsilon$ cannot be an upper bound for $\mathcal{C}$. Hence $\sup \mathcal{C} \geq m(A)$. On the other hand, for every closed set $C \subset A$ we have $m(C) \leq m(A)$. Thus $\sup \mathcal{C} \leq m(A)$, and therefore the two values are equal. ■

Remark. If we use $m^*(A)$ instead of $m(A)$, then unlike Theorem 10.13, the above theorem is not true for sets A that are not measurable. In fact when $m^*(A) < \infty$, A is measurable if and only if

$$m^*(A) = \sup\{m(C) : C \text{ is closed, and } C \subset A\}. \tag{$*$}$$

To see this, it suffices to show the if part, as we have already proved the only if part. Now, for any $\epsilon > 0$ there is a closed set $C \subset A$ such that $m(C) > m^*(A) - \epsilon$. But we have $m^*(A - C) = m^*(A) - m(C)$, since C is measurable and has finite measure. Hence $m^*(A - C) < \epsilon$. Thus A is measurable due to the regularity of Lebesgue measure. On the other hand, when $m^*(A) = \infty$, the property $(*)$ does not imply the measurability of A. The counterexamples will be discussed later.

Theorem 10.40. *Suppose $A \subset \mathbb{R}^n$ is measurable, and $m(A) < \infty$. Then*

(i) *For every $\epsilon > 0$ there are finitely many open cubes $Q_1, \dots, Q_k$ such that for $Q := \bigcup_{j=1}^k Q_j$ we have*

$$m(Q \Delta A) < \epsilon,$$

where $Q \Delta A := (Q - A) \cup (A - Q)$.

(ii) *For every $\epsilon > 0$ there is a compact set $K \subset A$ such that $m(A - K) < \epsilon$.*

Proof. (i) Let $\{Q_j\}$ be a countable family of open cubes that covers A, and

$$\sum_{j\geq 1}|Q_j| < m(A)+\frac{\epsilon}{2},$$

for a given $\epsilon > 0$. Then the series $\sum_{j\geq 1}|Q_j|$ is convergent, since $m(A) < \infty$. Hence there is $k \in \mathbb{N}$ such that $\sum_{j>k}|Q_j| < \frac{\epsilon}{2}$. Let $Q := \bigcup_{j\leq k} Q_j$. Then note that $A - Q \subset \bigcup_{j>k} Q_j$. Thus $m(A-Q) \leq \sum_{j>k} m(Q_j) < \frac{\epsilon}{2}$. On the other hand $Q \subset \bigcup_{j\geq 1} Q_j$. Therefore

$$m(Q) \leq \sum_{j\geq 1} m(Q_j) < m(A)+\frac{\epsilon}{2}.$$

Hence $m(Q-A) = m(Q)-m(A) < \frac{\epsilon}{2}$, since A is measurable and has finite measure. Thus we have

$$m(Q\Delta A) \leq m(Q-A)+m(A-Q) < \frac{\epsilon}{2}+\frac{\epsilon}{2} = \epsilon.$$

(ii) We know that there is a closed set $C \subset A$ such that $m(A-C) < \frac{\epsilon}{2}$. Let $K_j := R_j \cap C$, where R_j is the closed cube centered at the origin whose edges have length j. Note that for every j, K_j is closed and bounded, so it is compact. It is also obvious that $K_j \subset K_{j+1}$. We can also easily show that $C = \bigcup_{j\geq 1} K_j$. Hence we have $m(C) = \lim m(K_j)$. Thus there is a large enough k such that

$$m(C-K_k) = m(C)-m(K_k) < \frac{\epsilon}{2}.$$

Note that we have used the measurability of C, K_k. Finally we have $K_k \subset C \subset A$, and $A - K_k = (A-C)\cup(C-K_k)$. Therefore

$$m(A-K_k) \leq m(A-C)+m(C-K_k) < \epsilon. \qquad ■$$

Suppose $A \subset \mathbb{R}^n$ is bounded, and $R \supset A$ is an open rectangle. Remember that Lebesgue's original definition of measurability of A was $m^*(A) = m_*(A)$, where $m_*(A) := |R| - m^*(R-A)$ is the inner measure of A. This condition follows easily from Caratheodory's definition of measurability, i.e. the definition that we used in this section. The next theorem implies that the two definitions are actually equivalent.

Theorem 10.41. *Suppose $A, B \subset \mathbb{R}^n$ are bounded, and $A \subset B$. Also suppose that B is measurable, and*

$$m^*(A)+m^*(B-A) = m(B).$$

Then A is measurable.

Remark. Note that the boundedness of B is essential, since, for example, the above equality holds for every set A when $B = \mathbb{R}^n$.

Proof. Note that $\overline{B}$ is also bounded. Let R be an open rectangle containing $\overline{B}$. Then we have

$$(R - A) \cap B = R \cap A^c \cap B = B - A, \qquad (R - A) \cap B^c = R \cap A^c \cap B^c = R - B.$$

Hence the measurability of B implies that

$$\begin{aligned} m^*(R - A) = m^*((R - A) \cap B) + m^*((R - A) \cap B^c) \\ = m^*(B - A) + m^*(R - B). \end{aligned}$$

Therefore by theorem's assumption we get

$$\begin{aligned} m^*(A) + m^*(R - A) = m^*(A) + m^*(B - A) + m^*(R - B) \\ = m(B) + m(R - B) = m(R), \end{aligned} \qquad (*)$$

since $B, R - B$ are disjoint measurable sets whose union is R.

Now consider $R - A$. Then for a given $\epsilon > 0$ there is an open set $U \supset R - A$ such that $m(U) < m^*(R - A) + \epsilon$. We can assume $U \subset R$, since we can consider $U \cap R$ instead of U. Let $C := R - U$. Then $C \subset R - (R - A) = A$. Hence $\overline{C} \subset \bar{A} \subset \overline{B} \subset R$. We also have $\overline{C} \subset U^c$, since U^c is a closed set that contains C. Thus $\overline{C} \subset R \cap U^c = C$. So C is closed. We also have $m(C) + m(U) = m(R)$, since C, U are disjoint measurable sets whose union is R. Therefore by $(*)$ we get

$$\begin{aligned} m^*(A) - m(C) = m(R) - m^*(R - A) - \big(m(R) - m(U)\big) \\ = m(U) - m^*(R - A) < \epsilon. \end{aligned}$$

Note that here we are using the fact that $m(R)$, and consequently the (outer) measure of the other sets in the above equation, are finite. Now we know that $m^*(A - C) = m^*(A) - m(C)$, since C is a measurable set that has finite measure. Hence $m^*(A - C) < \epsilon$. Thus by the regularity of Lebesgue measure A is measurable.

■

Chapter 11

Topological Spaces

11.1 Topology and Basis

Definition 11.1. A **topology** on a set X is a family of its subsets $\mathcal{T}$ satisfying the following axioms

(i) $\emptyset, X \in \mathcal{T}$.

(ii) $\mathcal{T}$ is closed under arbitrary union, i.e. if $\{U_\alpha\}_{\alpha \in I} \subset \mathcal{T}$ then

$$\bigcup_{\alpha \in I} U_\alpha \in \mathcal{T}.$$

(iii) $\mathcal{T}$ is closed under finite intersection, i.e. if $U_1, \ldots, U_n \in \mathcal{T}$ then

$$\bigcap_{i=1}^{n} U_i \in \mathcal{T}.$$

A set equipped with a topology is called a **topological space**.

Remark. We refer to the elements of a topological space as *points* of the topological space.

Remark. Note that by an easy inductive argument, closedness under finite intersection follows from closedness under binary intersection.

Definition 11.2. Elements of $\mathcal{T}$ are called **open** sets. A subset of X is called **closed** if its complement is open. A **neighborhood** of a point $a \in X$ is a set that contains an open set containing a.

Theorem 11.3. *The family of closed sets has the following properties*

(i) $\emptyset, X$ *are closed sets.*

(ii) *The intersection of any collection of closed sets is a closed set.*

(iii) *The union of finitely many closed sets is a closed set.*

Proof. Take the complement of the respective properties for open sets, and use De Morgan's laws. ■

Proposition 11.4. *The intersection of a nonempty family of topologies on a set X is a topology on X.*

Proof. The intersection obviously contains $\emptyset, X$. It is also easy to check that the intersection is closed under arbitrary union and finite intersection. ■

Definition 11.5. The topology **generated** by a family $\mathcal{A}$ of subsets of a set X is the intersection of all topologies containing $\mathcal{A}$.

Remark. Note that the power set of X is a topology containing $\mathcal{A}$, so the family of all topologies containing $\mathcal{A}$ is nonempty, and its intersection is defined.

Proposition 11.6. *Suppose $\mathcal{T}$ is the topology generated by $\mathcal{A}$. Then $\mathcal{T}$ is contained in any topology containing $\mathcal{A}$. In other words, $\mathcal{T}$ is the smallest topology with respect to inclusion that contains $\mathcal{A}$.*

Proof. $\mathcal{T}$ is the intersection of all topologies containing $\mathcal{A}$, therefore it is a subset of any one of them. ■

Definition 11.7. Let $\mathcal{B}$ be a collection of subsets of X, and let $\mathcal{T}$ be the topology generated by $\mathcal{B}$. Then $\mathcal{B}$ is called a **basis** for $\mathcal{T}$, if it satisfies

(i) $\bigcup_{B\in\mathcal{B}} B = X$.

(ii) If $B_1, B_2 \in \mathcal{B}$ and $x \in B_1 \cap B_2$, then there is $B_3 \in \mathcal{B}$ such that

$$x \in B_3 \subset B_1 \cap B_2.$$

Theorem 11.8. *Suppose $\mathcal{B}$ is a basis for a topology $\mathcal{T}$ on a set X. Then*

$$\mathcal{T} = \{\bigcup_{\alpha} B_\alpha : \{B_\alpha\} \subset \mathcal{B}\}.$$

In other words, every open set in $\mathcal{T}$ is the union of some sets in the basis $\mathcal{B}$.

Proof. It is obvious that $\mathcal{T}$ contains the described family, since $\mathcal{T}$ is a topology that contains $\mathcal{B}$. Hence it is enough to show that the described family is a topology. Then the result follows from the minimality of $\mathcal{T}$. First note that $\mathcal{T}$ contains X by the first property of a basis. It also contains $\emptyset$, since $\emptyset$ is the union of the empty subset of $\mathcal{B}$. In addition, $\mathcal{T}$ is closed under unions by definition. Finally suppose $\{B_\alpha\}, \{B_\beta\} \subset \mathcal{B}$. It suffices to show that

$$[\bigcup_{\alpha} B_\alpha] \cap [\bigcup_{\beta} B_\beta] = \bigcup_{\gamma} B_\gamma,$$

for some $\{B_\gamma\} \subset \mathcal{B}$. We have

$$[\bigcup_\alpha B_\alpha] \cap [\bigcup_\beta B_\beta] = \bigcup_\alpha \bigcup_\beta [B_\alpha \cap B_\beta].$$

But by the second property of a basis, $B_\alpha \cap B_\beta = \bigcup_{\gamma \in I_{\alpha\beta}} B_\gamma$ for some $\{B_\gamma\}_{\gamma \in I_{\alpha\beta}} \subset \mathcal{B}$. Now we have

$$[\bigcup_\alpha B_\alpha] \cap [\bigcup_\beta B_\beta] = \bigcup_{\alpha,\beta} \bigcup_{\gamma \in I_{\alpha\beta}} B_\gamma. \qquad ■$$

Remark. The notion of a basis for a topology, is a generalization of the family of open balls in a metric space.

Remark. Note that the concept of basis in topology is different than the concept of basis in linear algebra. Here, a basis is merely a suitable generator of the topology, and there is no corresponding notion of uniqueness of the representation of an open set in terms of the basis.

Remark. If $\mathcal{B}$ is a basis for the topology on X. Then every open set U of X is the union of some elements of $\mathcal{B}$. So for every $a \in U$ there is $B \in \mathcal{B}$ such that $a \in B \subset U$.

Theorem 11.9. *Let $\mathcal{B}$ be a collection of open subsets of a space X. Then $\mathcal{B}$ is a basis for the topology of X if and only if for every open subset U of X and every $a \in U$ there is $B \in \mathcal{B}$ such that $a \in B \subset U$.*

Proof. If $\mathcal{B}$ is a basis, the claim holds as indicated in the previous remark. Conversely, suppose the family $\mathcal{B}$ has the described property. First note that any open set U can be written as a union of elements of $\mathcal{B}$, namely $U = \bigcup_{a \in U} B_a$ where $B_a \in \mathcal{B}$ satisfies $a \in B_a \subset U$. Thus all the open subsets of X are contained in any topology containing $\mathcal{B}$, especially the topology generated by $\mathcal{B}$. Hence the topology generated by $\mathcal{B}$ equals the topology of X, since the latter obviously contains $\mathcal{B}$.

As a special case of the above, X itself can be written as a union of elements of $\mathcal{B}$. So, all that is left is to check the second property of a basis for $\mathcal{B}$. Let $B_1, B_2 \in \mathcal{B}$, and $x \in B_1 \cap B_2$. Then as B_1, B_2 are open, so is $B_1 \cap B_2$. Hence there is $B_3 \in \mathcal{B}$ such that $x \in B_3 \subset B_1 \cap B_2$, as desired. ■

***Example* 11.10.** The standard topology on $\mathbb{R}$ is the topology generated by bounded open intervals. Note that they form a basis for the topology.

***Example* 11.11.** Remember that when (X, d) is a metric space, a set $U \subset X$ is open if for every $x \in U$ there is $r > 0$ such that the open ball of radius r around x is in U, i.e.

$$B_r(x) := \{y \in X : d(y, x) < r\} \subset U.$$

We always equip metric spaces with this topology and consider them as a special kind of topological spaces.

Proposition 11.12. *The family of open balls is a basis for the topology of a metric space.*

Proof. Since open balls are open sets, and for every open set U and every $a \in U$ there is an open ball $B_r(a) \subset U$, the collection of open balls is a basis. ■

Second Proof. Here we give a direct proof for the theorem. The union of all open balls is the whole space obviously. Now consider two points x, y in the metric space and let $z \in B_r(x) \cap B_s(y)$. Then

$$B_t(z) \subset B_r(x) \cap B_s(y),$$

where $t = \min\{r - d(x,z), s - d(y,z)\}$. Thus the family of open balls is a basis. Finally note that every open set in a metric space is a union of a collection of open balls, so every open set is contained in any topology containing the open balls. Therefore the family of open balls generates the topology of the metric space. ■

11.2 Sequences and Limit Points

Definition 11.13. A sequence (a_n) in a set X is a function

$$\begin{aligned} \mathbb{N} &\to X \\ n &\mapsto a_n \end{aligned}.$$

We also use the notation $(a_n)_{n\in\mathbb{N}}$. A sequence $(b_k)_{k\in\mathbb{N}}$ is called a **subsequence** of $(a_n)_{n\in\mathbb{N}}$ if $b_k = a_{n_k}$, for a strictly increasing sequence $n_1 < n_2 < \cdots$ of positive integers.

Definition 11.14. Suppose (a_n) is a sequence in the topological space X. We say the sequence (a_n) converges to the **limit** $a \in X$, and write $\lim a_n = a$ or $a_n \to a$, if for every open set U containing a there exists an $N \in \mathbb{N}$ such that $a_n \in U$ for all $n \geq N$.

Theorem 11.15. *Let (X, d) be a metric space. Then a sequence (a_n) in X converges to a if and only if*

$$\forall \epsilon > 0 \, \exists N \in \mathbb{N} \textit{ such that } \forall n \geq N \textit{ we have } d(a_n, a) < \epsilon.$$

Proof. Suppose $a_n \to a$. Then for large enough n we have $a_n \in B_\epsilon(a)$, since $B_\epsilon(a)$ is an open set containing a. But we have

$$a_n \in B_\epsilon(a) \iff d(a_n, a) < \epsilon.$$

For the converse, let U be an open set containing a. Then as open balls form a basis for the topology of a metric space, there is an open ball $B_\epsilon(a)$ such that $B_\epsilon(a) \subset U$. Hence for large enough n we have $d(a_n, a) < \epsilon$ thus $a_n \in B_\epsilon(a) \subset U$, as desired. ■

Theorem 11.16. *Suppose $\mathcal{B}$ is a basis for the topology on X, and (a_n) is a sequence in X. Then $a_n \to a$ if and only if for every open set $B \in \mathcal{B}$ that contains a there exists an $N \in \mathbb{N}$ such that $a_n \in B$ for all $n \geq N$.*

Proof. If $a_n \to a$ then the claim holds by the definition of convergence, since the elements of $\mathcal{B}$ are open sets. Conversely, suppose the definition of convergence holds for open sets in $\mathcal{B}$. Let U be an open set containing a. Then there is $B \in \mathcal{B}$ such that $a \in B \subset U$. Hence there is $N \in \mathbb{N}$ such that for all $n \geq N$ we have $a_n \in B \subset U$. Therefore $a_n \to a$. ■

Definition 11.17. A topological space X is called a **Hausdorff** space if for every two distinct points $x, y \in X$ there exist open sets U, V such that

$$x \in U, \ y \in V, \quad \text{and } U \cap V = \emptyset.$$

Theorem 11.18. *Metric spaces are Hausdorff.*

Proof. If $x \neq y$ then $d(x, y) > 0$. Let $r < \frac{1}{2}d(x, y)$, then by the triangle inequality $B_r(x) \cap B_r(y) = \emptyset$. ■

Theorem 11.19. *Every convergent sequence in a Hausdorff space has a unique limit.*

Proof. Suppose to the contrary that a sequence (a_n) converges to two distinct points a, b. Let U, V be open sets containing a, b respectively, such that $U \cap V = \emptyset$. Then for large enough n we must have $a_n \in U$ and $a_n \in V$, which is impossible. ■

Theorem 11.20. *Finite subsets of a Hausdorff space are closed.*

Proof. The empty set is obviously closed. Let $\{a\}$ be a one element subset. Then for any $b \in \{a\}^c$ there are open sets U, V such that $b \in U$, $a \in V$, and $U \cap V = \emptyset$. In particular $a \notin U$, thus $U \subset \{a\}^c$. Thus $\{a\}^c$ is a union of open sets, hence it is open. So $\{a\}$ is closed. Finally, any finite subset is a union of finitely many closed one element subsets, hence it is closed. ■

Theorem 11.21. *Every subsequence of a convergent sequence converges to the same limit(s) as the original sequence.*

Proof. Suppose $a_n \to a$ and $b_k = a_{n_k}$. Let U be an open neighborhood of a. Then there is N so that $a_n \in U$ for $n \geq N$. Now since $n_k \geq k$, the same N works for (b_k). ■

Definition 11.22. Suppose X is a topological space, and $A \subset X$. The **closure** of A, denoted by $\bar{A}$, is the intersection of all closed sets containing A.

Remark. Note that X is a closed set containing A, so the family of all closed sets containing A is nonempty and its intersection is defined.

Theorem 11.23. *Suppose X is a topological space, and $A \subset X$. Then $\bar{A}$ is the smallest closed set that contains A, i.e. it is closed, contains A, and is contained in any closed set containing A. As a result, $\overline{C} = C$ for any closed set C.*

Proof. $\bar{A}$ is the intersection of all closed sets containing A, therefore it is closed, and is a subset of any closed set containing A. It also contains A obviously. Now if C is closed, then C is the smallest closed set containing C. Hence $\overline{C} = C$. ■

Theorem 11.24. *Suppose X is a topological space, and $A \subset X$. If a sequence (a_n) of points in A converges to a, then $a \in \bar{A}$.*

Proof. If $a \notin \bar{A}$, then $a \in (\bar{A})^c$. Now as $(\bar{A})^c$ is open, for large enough n we must have $a_n \in (\bar{A})^c$. But this is impossible since $(\bar{A})^c \subset A^c$. ■

Definition 11.25. A point $x \in X$ is called a **limit point** or an **accumulation point** of $A \subset X$, if every open set containing x intersects A in a point other than x.

Theorem 11.26. *The closure of a set is the union of the set and its limit points.*

Proof. Suppose $x \in \bar{A} - A$. Let U be an open set containing x. Suppose to the contrary that $U \cap A = \emptyset$. Then $U \subset A^c$, and consequently $A \subset U^c$. Now as U^c is closed, we have $\bar{A} \subset U^c$. But this is in contradiction with the fact that $x \in \bar{A}$. ■

Proposition 11.27. *The closure of a set is the set of points that any open neighborhood of them intersects the set.*

Proof. By the last theorem, every point of $\bar{A}$ is either in A or is a limit point of A. Any open neighborhood of these points clearly intersects A. So we only need to show that no other point has this property. This is easy since if $x \notin \bar{A}$, then $x \in (\bar{A})^c$. But $(\bar{A})^c$ is open and does not intersect A. ■

Definition 11.28. Suppose X is a topological space, and $A \subset X$. The **interior** of A is the union of all open subsets of X contained in A. We denote it by A°. The **boundary** of A is $\partial A := \bar{A} - A^\circ$.

Proposition 11.29. *Suppose X is a topological space, and $A \subset X$. Then*

$$\partial A = \bar{A} \cap \overline{A^c}.$$

As a result, ∂A is closed, and equals the set of points that any open neighborhood of them intersects both A, A^c.

Proof. We have $\partial A = \bar{A} - A^\circ = \bar{A} \cap (A^\circ)^c$. So it suffices to show that $(A^\circ)^c = \overline{A^c}$. If $x \notin A^\circ$, then for every open set $U \subset A$ we have $x \notin U$. Let C be a closed set containing A^c. Then C^c is an open set contained in A. Thus $x \notin C^c$. Hence $x \in C$, and as C is arbitrary we get $x \in \overline{A^c}$.

Conversely suppose $x \in \overline{A^c}$. Then for every closed set $C \supset A^c$ we have $x \in C$. Let U be an open set contained in A. Then U^c is a closed set containing A^c. Therefore $x \in U^c$. So $x \notin U$. Hence $x \notin A^\circ$.

Thence we have $\partial A = \bar{A} \cap \overline{A^c}$. Thus ∂A is closed, since it is the intersection of two closed sets. Finally, the last statement of the theorem follows from the previous proposition about characterizing the closure of a set. ■

Proposition 11.30. *Suppose X is a topological space, and $A \subset X$. Then A° is the largest open set contained in A, i.e. it is open, is contained in A, and contains any open subset of A.*

Proof. A° is the union of all open sets contained in A, therefore it is open, and contains any open set contained in A. It is also contained in A obviously. ■

11.3 Subspaces

Definition 11.31. Let X be a topological space, and $Y \subset X$. The **subspace topology** on Y is the topology

$$\{U \cap Y : U \text{ is open in } X\}.$$

Remark. In the above definition, it must be checked that the specified family is a topology.

Theorem 11.32. *Suppose $\mathcal{B}$ is a basis for the topology on X, and $Y \subset X$. Then*

$$\{B \cap Y : B \in \mathcal{B}\}$$

is a basis for the subspace topology on Y.

Proof. First note that the union of the specified family is Y. Now let

$$x \in (B_1 \cap Y) \cap (B_2 \cap Y) = (B_1 \cap B_2) \cap Y.$$

Then there is B_3 such that $x \in B_3 \subset B_1 \cap B_2$. Hence

$$x \in B_3 \cap Y \subset (B_1 \cap Y) \cap (B_2 \cap Y).$$

Finally note that for any open set $U \cap Y$ in the subspace topology, there is a family $\{B_\alpha\}$ such that $U = \bigcup B_\alpha$. Therefore

$$U \cap Y = [\bigcup_\alpha B_\alpha] \cap Y = \bigcup_\alpha [B_\alpha \cap Y].$$ ■

Theorem 11.33. *Suppose Y is a subspace of X. Then $C \subset Y$ is closed in Y if and only if there is a closed subset $D \subset X$ such that $C = D \cap Y$.*

Proof. Let C be a closed set in Y. Then the complement of C in Y is open in Y, i.e. $Y - C$ is open in Y. Thus there is an open set $U \subset X$ such that $Y - C = U \cap Y$. Hence

$$U^c \cap Y = Y - U = Y - (U \cap Y) = Y - (Y - C) = Y \cap C = C.$$

Conversely, let D be a closed subset of X. The D^c is open. Hence $D^c \cap Y$ is open in Y. Now we have

$$Y - (D \cap Y) = Y - D = Y \cap D^c.$$

Thus $D \cap Y$ is closed in Y. ■

Theorem 11.34. *Suppose Y is a subspace of X, and (a_n) is a sequence in Y. Then (a_n) converges to $a \in Y$ as a sequence in Y if and only if $a_n \to a$ as a sequence in X.*

Proof. Any open neighborhood V of a in Y is of the form $Y \cap U$ where U is an open neighborhood of a in X. Now for large enough n, $a_n \in U$ if and only if $a_n \in V$, since $a_n \in Y$. Thus the two convergences are equivalent. ■

Theorem 11.35. *Let $Z \subset Y \subset X$. Then the subspace topology that Z inherits from Y is the same as the subspace topology that Z inherits from X.*

Proof. Let U be an open subset of X. Then the claim follows easily from

$$U \cap Z = (U \cap Y) \cap Z.$$ ■

Theorem 11.36. *The subspaces of a Hausdorff space are Hausdorff.*

Proof. Suppose X is Hausdorff and $Y \subset X$. Let $x, y \in Y$ be distinct points. Then there are open subsets U, V of X containing x, y respectively, such that $U \cap V = \emptyset$. But then $U \cap Y$ and $V \cap Y$ are open neighborhoods of x, y in Y, and

$$(U \cap Y) \cap (V \cap Y) = \emptyset.$$ ■

Theorem 11.37. *Suppose (X, d) is a metric space and $Y \subset X$. Then $d|_{Y \times Y}$ is a metric on Y that induces the same topology as the subspace topology.*

Proof. It is obvious that $d|_{Y \times Y}$ is a metric on Y. We use the notations $B_r(x, X)$ and $B_r(x, Y)$ for the open balls around x in X, Y respectively. Note that we have

$$B_r(x, Y) = \{y \in Y : d(y, x) < r\} = B_r(x, X) \cap Y.$$

Now let V be an open subset of Y in the metric topology. Then $V = \bigcup_{x \in V} B_{r_x}(x, Y)$. Hence

$$V = \bigcup_{x \in V} [B_{r_x}(x, X) \cap Y] = [\bigcup_{x \in V} B_{r_x}(x, X)] \cap Y.$$

But $\bigcup_{x \in V} B_{r_x}(x, X)$ is open in X, since it is a union of open balls. Therefore V is open in the subspace topology.

Conversely suppose V is an open subset of Y in the subspace topology. Then $V = U \cap Y$ for some U which is open in X. Then we have $U = \bigcup_{x \in U} B_{r_x}(x, X)$. Hence

$$V = [\bigcup_{x \in U} B_{r_x}(x, X)] \cap Y = \bigcup_{x \in U} [B_{r_x}(x, X) \cap Y] = \bigcup_{x \in U} B_{r_x}(x, Y).$$

Therefore V is also open in the metric topology. ■

Theorem 11.38. *Suppose A is a subspace of X. Then*
(i) *Open sets in A are open in X when A is open in X.*
(ii) *Closed sets in A are closed in X when A is closed in X.*

Proof. Open sets in A are of the form $U \cap A$ where U is open in X. Hence if A is open in X, $U \cap A$ is also open in X. The proof is the same for closed sets. ■

11.4 Product Spaces

Definition 11.39. Suppose $X_1, \ldots, X_n$ are topological spaces. The **product topology** on $\prod_{i=1}^n X_i = X_1 \times \cdots \times X_n$ is the topology generated by

$$\{U_1 \times \cdots \times U_n : U_i \text{ is open in } X_i\}.$$

Remark. It is easy to see that the product of open sets is indeed a basis for the product topology.

Theorem 11.40. *Suppose $\mathcal{B}_i$ is a basis for the topology on X_i. Then*

$$\{B_1 \times \cdots \times B_n : B_i \in \mathcal{B}_i\}$$

is a basis for the product topology on $\prod_{i=1}^n X_i$.

Proof. First note that the union of the specified family is $\prod X_i$, since for any $x = (x_1, \ldots, x_n)$ in the product space we have $x_i \in B_i$ for some $B_i \in \mathcal{B}_i$, hence $x \in \prod B_i$. Now let

$$x = (x_1, \ldots, x_n) \in \Big(\prod_{i \leq n} B_i\Big) \cap \Big(\prod_{i \leq n} B_i'\Big) = \prod_{i \leq n} (B_i \cap B_i').$$

Then there are B_i'' such that $x_i \in B_i'' \subset B_i \cap B_i'$. Hence

$$x \in \prod_{i\leq n} B_i'' \subset \Big(\prod_{i\leq n} B_i\Big) \cap \Big(\prod_{i\leq n} B_i'\Big).$$

Finally to see that the specified family generates the product topology, it is enough to show that each open set of the form $U_1 \times \cdots \times U_n$ is the union of some elements of the family. The reason is that any open set in the product topology is a union of some open sets of the form $U_1 \times \cdots \times U_n$. Now we have $U_i = \bigcup_{\alpha_i} B_{\alpha_i}$, where $B_{\alpha_i} \in \mathcal{B}_i$. Therefore

$$\begin{aligned} U_1 \times \cdots \times U_n &= \bigcup_{\alpha_1} B_{\alpha_1} \times \cdots \times \bigcup_{\alpha_n} B_{\alpha_n} \\ &= \bigcup_{\alpha_1,\ldots,\alpha_n} B_{\alpha_1} \times \cdots \times B_{\alpha_n}. \end{aligned}$$ ■

***Exercise* 11.41.** Show that the product of closed sets is a closed subset of the product space.

Theorem 11.42. *Suppose (X_i, d_i) are metric spaces. Then, the equivalent metrics*

$$\begin{aligned} d(x,y) &:= \left[\sum_{i=1}^{n} d_i(x_i,y_i)^2\right]^{\frac{1}{2}}, \\ d_{\text{sum}}(x,y) &:= \sum_{i=1}^{n} d_i(x_i,y_i), \\ d_{\max}(x,y) &:= \max_{i\leq n}\{d_i(x_i,y_i)\}, \end{aligned}$$

induce the product topology on $\prod_{i=1}^n X_i$.

Proof. These metrics are all equivalent as we have seen before, so they induce the same topology by Theorem 2.39. On the other hand, let $x = (x_1, \ldots, x_n) \in \prod X_i$. We use the notation $B_r(z, d)$ for the ball of radius r around z with respect to the metric d. Then it is easy to see that

$$B_r(x, d_{\max}) = \prod_{i=1}^{n} B_r(x_i, d_i).$$

Thus the open balls with respect to $d_{\max}$ form a basis for the product topology, hence $d_{\max}$ induces the product topology. ■

Theorem 11.43. *Suppose that $Y_i \subseteq X_i$. Then the product of subspace topologies on $\prod_{i=1}^n Y_i$ is the same as the topology it inherits as a subspace of $\prod_{i=1}^n X_i$.*

Proof. The sets $\prod_{i=1}^{n}(U_i \cap Y_i)$, where U_i is an open subset of X_i, form a basis for the product of subspace topologies. On the other hand, the sets $(\prod_{i=1}^{n} U_i) \cap (\prod_{i=1}^{n} Y_i)$ form a basis for the subspace topology inherited from $\prod_{i=1}^{n} X_i$. But we have

$$\prod_{i=1}^{n}(U_i \cap Y_i) = \Big(\prod_{i=1}^{n} U_i\Big) \cap \Big(\prod_{i=1}^{n} Y_i\Big).$$ ■

Theorem 11.44. *A sequence $(a_n)_{n\in\mathbb{N}} = \big((a_{n,1}, \ldots, a_{n,k})\big)_{n\in\mathbb{N}}$ in $X_1 \times \cdots \times X_k$ converges to $a = (a_1, \ldots, a_k)$ if and only if $a_{n,i} \to a_i$ for each i.*

Proof. Let U be an open neighborhood of a. Then there open subsets U_i of X_i such that

$$a \in U_1 \times \cdots \times U_k \subset U,$$

since the sets $\prod U_i$ form a basis for the product topology. Now if $a_{n,i} \to a_i$ for each i, then for large enough n we have $a_{n,i} \in U_i$ for each i. Hence

$$a_n \in U_1 \times \cdots \times U_k \subset U.$$

For the converse note that if $a_n \to a$, then for large enough n we have $a_n \in \prod U_i$. Therefore $a_{n,i} \in U_i$ for each i. ■

Theorem 11.45. *The product of finitely many Hausdorff spaces is Hausdorff.*

Proof. Suppose $x, y \in \prod_{i\leq n} X_i$, and $x \neq y$. Then $x_j \neq y_j$ for some j. Now there are open sets U_j, V_j in X_j, containing x_j, y_j respectively, such that $U_j \cap V_j = \emptyset$. Then

$$\begin{aligned} x &\in X_1 \times \cdots \times X_{j-1} \times U_j \times X_{j+1} \times \cdots \times X_n \\ y &\in X_1 \times \cdots \times X_{j-1} \times V_j \times X_{j+1} \times \cdots \times X_n, \end{aligned}$$

and

$$[X_1 \times \cdots \times U_j \times \cdots \times X_n] \cap [X_1 \times \cdots \times V_j \times \cdots \times X_n] = \emptyset.$$ ■

11.5 The Countability Axioms

Definition 11.46. A **first countable** topological space X, is a space that has a countable basis at each point, i.e. for every point $x \in X$ there exists a countable family $\{U_n\}$ of open sets containing x, such that any open set that contains x contains one of the U_n's.

Remark. We can redefine U_n to be $\bigcap_{i=1}^{n} U_i$ and assume that the above family is decreasing.

Theorem 11.47. *Metric spaces are first countable.*

Proof. $\{B_{\frac{1}{k}}(x) : k \in \mathbb{N}\}$ is a countable basis at x. ■

Theorem 11.48. *Suppose that A is a subset of a first countable space, and $x \in \bar{A}$. Then there is a sequence in A that converges to x.*

Proof. Let $\{U_n\}$ be a decreasing countable basis at x. Then there are points $a_n \in A \cap U_n$, since $x \in \bar{A}$. Now we have $a_n \to x$. Because for every open neighborhood U of x, there is N such that $U_n \subset U$ for $n \geq N$. Hence $a_n \in U$ for $n \geq N$. ■

Definition 11.49. A **second countable** topological space is a space that has a countable basis.

Remark. Second countable spaces are obviously first countable too.

Theorem 11.50. *Subspaces of first or second countable spaces are respectively first or second countable. Also, products of finitely many first or second countable spaces are respectively first or second countable.*

Proof. First suppose $Y \subset X$, and $x \in Y$. Let $\{U_n\}$ be a countable basis at x. Then it is easy to see that $\{U_n \cap Y\}$ is a countable basis at x with respect to the subspace topology on Y. Similarly, if $\{U_n\}$ is a countable basis for X, then $\{U_n \cap Y\}$ is a countable basis for Y. Note that the restriction of a basis is a basis for the subspace.

Next suppose that $x = (x_1, \dots, x_k) \in \prod_{i \leq k} X_i$. Let $\{U_{n,i}\}_{n\in\mathbb{N}}$ be a decreasing countable basis at x_i. Then $\{\prod_{i\leq k} U_{n,i}\}_{n\in\mathbb{N}}$ is a countable basis at x. The reason is that if U is an open neighborhood of x, then there are open sets V_i in X_i such that

$$x \in \prod_{i\leq k} V_i \subset U.$$

Now for each i there is n_i such that $x \in U_{n_i,i} \subset V_i$. Hence for $n = \max_{i\leq k}\{n_i\}$ we have

$$x \in \prod_{i\leq k} U_{n,i} \subset U.$$

Finally, let $\mathcal{B}_i$ be a countable basis for X_i. Then

$$\{B_1 \times \cdots \times B_n : B_i \in \mathcal{B}_i\}$$

is a countable basis for $\prod_{i\leq k} X_i$. Note that the product of finitely many countable collections is countable. ■

Theorem 11.51. $\mathbb{R}^n$ *and its subspaces are second countable.*

Proof. The bounded intervals with rational endpoints form a countable basis for $\mathbb{R}$ (why?). Therefore their products, i.e. open rectangles whose vertices have rational coordinates, form a countable basis for $\mathbb{R}^n$. Finally, the subspaces of second countable spaces are second countable. ■

Definition 11.52. A subset of a topological space is **dense** if its closure is the whole space. A topological space is called **separable** if it has a countable dense subset.

***Example* 11.53.** $\mathbb{Q}^n$ is a countable dense subset of $\mathbb{R}^n$, so $\mathbb{R}^n$ is separable. By the next theorem, the subspaces of $\mathbb{R}^n$ are also separable, since they are second countable.

Theorem 11.54. *A second countable space is separable.*

Proof. Let $\{U_i\}$ be a countable basis for X. We can assume that all U_i's are nonempty. Let $a_i \in U_i$. Then we claim that $\{a_i\}$ is dense in X. Let x be an arbitrary point of X. Suppose U is an open neighborhood of x. Then for some j we have $x \in U_j \subset U$. Hence $a_j \in U$. Therefore any open neighborhood of x intersects $\{a_i\}$. Hence $x \in \overline{\{a_i\}}$. ■

Theorem 11.55. *A separable metric space is second countable.*

Proof. Let $\{a_n\}$ be a countable dense subset. We claim that the family

$$\mathcal{B} = \{B_{\frac{1}{k}}(a_n) : n, k \in \mathbb{N}\}$$

is a countable basis. First note that $\mathcal{B}$ is countable, since the union of countably many countable sets is countable. Next, we need to show that $\mathcal{B}$ is a basis. Let x be an arbitrary point of the space. Then $B_{\frac{1}{2}}(x)$ contains some a_n, since it is an open set that intersects the closure of $\{a_n\}$, i.e. the whole space. Hence $x \in B_1(a_n)$. Thus $\bigcup \mathcal{B}$ is the whole space.

Now suppose $x \in B_{\frac{1}{k}}(a_n) \cap B_{\frac{1}{l}}(a_m)$. Then for some large enough N we have

$$B_{\frac{1}{N}}(x) \subset B_{\frac{1}{k}}(a_n) \cap B_{\frac{1}{l}}(a_m).$$

Again, $B_{\frac{1}{2N}}(x)$ contains some a_j. Hence

$$x \in B_{\frac{1}{2N}}(a_j) \subset B_{\frac{1}{k}}(a_n) \cap B_{\frac{1}{l}}(a_m).$$

Thus $\mathcal{B}$ is a basis.

Finally, we have to show that $\mathcal{B}$ generates the topology of the space. Suppose U is an open set, and $x \in U$. Then $B_{\frac{1}{N}}(x) \subset U$ for some large enough N. Also, there is some $a_n \in B_{\frac{1}{2N}}(x)$. Therefore $x \in B_{\frac{1}{2N}}(a_n) \subset U$. ■

Definition 11.56. An **open covering** of a subset A of a topological space X, is a family $\mathcal{U}$ of open subsets of X such that

$$\bigcup \mathcal{U} = \bigcup_{U \in \mathcal{U}} U \supset A.$$

A **subcovering** $\mathcal{V}$ of $\mathcal{U}$ is a subfamily of $\mathcal{U}$ which is itself an open covering of A.

Theorem 11.57. *Every open covering of a second countable space has a countable subcovering.*

Proof. Suppose $\mathcal{U}$ is an open covering of X. Let $\mathcal{B}$ be a countable basis for X. For every $x \in X$ there is $U_x \in \mathcal{U}$ such that $x \in U_x$. Also there is $B_i \in \mathcal{B}$ such that $x \in B_i \subset U_x$. Let U_i be one of the U_x's that contain B_i. Since B_i's form a countable family, so do U_i's. Now for every x there is B_i such that $x \in B_i \subset U_i$. Hence $\{U_i\}$ is a countable subcovering of $\mathcal{U}$. ■

Remark. A space whose every open covering has a countable subcovering is called a **Lindelof** space. By the above theorem second countable spaces, in particular $\mathbb{R}^n$ and its subspaces, are Lindelof.

11.6 Continuous Functions

Definition 11.58. Let $f : X \to Y$, and $A \subset Y$. Then the **preimage** or the **inverse image** of A is

$$f^{-1}(A) := \{x \in X : f(x) \in A\}.$$

Note that we do not require the map f to be invertible.

Definition 11.59. A function $f : X \to Y$ is called **continuous** if for each open subset U of Y, the set $f^{-1}(U)$ is an open subset of X. A function that is not continuous is called **discontinuous**.

Theorem 11.60. *Suppose $\mathcal{B}$ is a basis for the topology of Y. Then a map $f : X \to Y$ is continuous if for each open set $B \in \mathcal{B}$, the set $f^{-1}(B)$ is an open subset of X.*

Proof. Suppose U is an open subset of Y. Then $U = \bigcup_\alpha B_\alpha$ for some $B_\alpha \in \mathcal{B}$. Now we have

$$f^{-1}(U) = \{x \in X : f(x) \in \bigcup_\alpha B_\alpha\} = \bigcup_\alpha \{x \in X : f(x) \in B_\alpha\} = \bigcup_\alpha f^{-1}(B_\alpha).$$

Thus $f^{-1}(U)$ is a union of open sets, hence it is open. ■

Theorem 11.61. *A map $f : X \to Y$ is continuous if and only if for each closed subset C of Y, the set $f^{-1}(C)$ is a closed subset of X.*

$\boxed{\textbf{Proof.}}$ We know that C^c is an open subset of Y. Then

$$f^{-1}(C^c) = \{x \in X : f(x) \in C^c\} = \{x \in X : f(x) \notin C\} = (f^{-1}(C))^c.$$

Thus $(f^{-1}(C))^c$ is open, hence $f^{-1}(C)$ is closed. The converse is similar. ■

Proposition 11.62. *The constant functions are continuous. The identity map of any space*

$$\begin{aligned} \mathrm{id}_X : X &\to X \\ x &\mapsto x \end{aligned}$$

is continuous. Also, the projections

$$\begin{aligned} \pi_i : X_1 \times \cdots \times X_n &\to X_i \\ (x_1, \ldots, x_n) &\mapsto x_i \end{aligned}$$

from a product space to any of its components are continuous.

$\boxed{\textbf{Proof.}}$ The first two cases are obvious. For the projections note that if U is an open subset of X_i, then

$$\pi_i^{-1}(U) = X_1 \times \cdots \times X_{i-1} \times U \times X_{i+1} \times \cdots \times X_n$$

is an open subset of $\prod X_j$. ■

Theorem 11.63. *The composition of continuous functions is continuous.*

$\boxed{\textbf{Proof.}}$ Suppose $f : X \to Y$ and $g : Y \to Z$ are continuous. We want to show that $g \circ f : X \to Z$ is continuous. Let U be an open subset of Z. Then

$$\begin{aligned} (g \circ f)^{-1}(U) &= \{x \in X : g(f(x)) \in U\} \\ &= \{x \in X : f(x) \in g^{-1}(U)\} = f^{-1}(g^{-1}(U)). \end{aligned}$$

Hence $(g \circ f)^{-1}(U)$ is an open subset of X, since $g^{-1}(U)$ is an open subset of Y. ■

Theorem 11.64. *Suppose that $f : X \to Y$ is continuous. Then for every convergent sequence $x_n \to x$ in X, we have $f(x_n) \to f(x)$. The converse holds if X is first countable.*

Proof. Let U be an open set containing $f(x)$. Then $f^{-1}(U)$ is an open set containing x. Hence there is a positive integer N such that for all $n \geq N$ we have $x_n \in f^{-1}(U)$. Therefore for $n \geq N$ we have $f(x_n) \in U$ as desired.

For the converse, suppose to the contrary that f is not continuous. Then there is an open set U such that $f^{-1}(U)$ is not open. Therefore $(f^{-1}(U))^c$ is not closed. Let

$$x \in \overline{(f^{-1}(U))^c} - (f^{-1}(U))^c.$$

Note that $x \in f^{-1}(U)$, hence $f(x) \in U$. Now as X is first countable, there is a sequence (a_n) in $(f^{-1}(U))^c$ that converges to x. Then as $a_n \to x$ we have $f(a_n) \to f(x)$. Therefore for large enough n we must have $f(a_n) \in U$. But this is impossible since $a_n \notin f^{-1}(U)$. ■

Remark. Remember that when X, Y are metric spaces, a map $f : X \to Y$ is continuous if and only if

$$\begin{aligned}&\forall x \in X\ \forall \epsilon > 0\ \exists \delta > 0 \text{ such that} \\ &\qquad \forall y \in X\ d_X(y, x) < \delta \implies d_Y(f(y), f(x)) < \epsilon.\end{aligned}$$

Theorem 11.65. *Suppose $f : X \to Y$ is continuous, $A \subset X$, and $f(A) \subset B \subset Y$. Then*

$$\begin{aligned} f|_A : A &\to B \\ x &\mapsto f(x) \end{aligned}$$

is continuous.

Proof. Let V be an open subset of B. Then there is an open subset U of Y such that $V = U \cap B$. Now

$$(f|_A)^{-1}(V) = f^{-1}(U \cap B) \cap A = f^{-1}(U \cap f(A)) \cap A = f^{-1}(U) \cap A.$$ ■

Remark. Let A be a subspace of X. Then the inclusion map

$$\begin{aligned} j_A : A &\to X \\ x &\mapsto x \end{aligned}$$

is continuous, since it is the restriction of id_X.

Theorem 11.66. *For functions into product spaces we have*

(i) *The function*

$$\begin{aligned} f = (f_1, \ldots, f_n) : X &\longrightarrow Y_1 \times \cdots \times Y_n \\ x &\mapsto (f_1(x), \ldots, f_n(x)) \end{aligned}$$

is continuous if and only if each $f_i : X \to Y_i$ is continuous.

(ii) *The function*

$$\begin{aligned} f = f_1 \times \cdots \times f_n : X_1 \times \cdots \times X_n &\longrightarrow Y_1 \times \cdots \times Y_n \\ (x_1, \ldots, x_n) &\mapsto (f_1(x_1), \ldots, f_n(x_n)) \end{aligned}$$

is continuous if each $f_i : X_i \to Y_i$ is continuous. The converse holds if in addition each X_i is nonempty.

$\boxed{\textbf{Proof.}}$ Let U_i be an open subset of Y_i. Note that for the continuity of f it is enough to show that $f^{-1}(U_1 \times \cdots \times U_n)$ is open for all choice of U_i's.

(i) If each f_i is continuous, then $f_i^{-1}(U_i)$ is open for all i. Hence

$$f^{-1}(U_1 \times \cdots \times U_n) = f_1^{-1}(U_1) \cap \cdots \cap f_n^{-1}(U_n)$$

is also open. Conversely if f is continuous we have

$$f_i^{-1}(U_i) = f^{-1}(Y_1 \times \cdots \times Y_{i-1} \times U_i \times Y_{i+1} \times \cdots \times Y_n).$$

(ii) If each f_i is continuous, then $f_i^{-1}(U_i)$ is open for all i. Hence

$$f^{-1}(U_1 \times \cdots \times U_n) = f_1^{-1}(U_1) \times \cdots \times f_n^{-1}(U_n)$$

is also open. Conversely, suppose f is continuous. Let $a_k \in X_k$ be a fixed element. Then by part (i) the map

$$\begin{aligned} j_i : X_i &\longrightarrow X_1 \times \cdots \times X_i \times \cdots \times X_n \\ x_i &\mapsto (a_1, \ldots, a_{i-1}, x_i, a_{i+1}, \ldots, a_n) \end{aligned}$$

is continuous. Now f_i can be written as a composition of continuous functions as follows

$$f_i = \pi_i \circ f \circ j_i,$$

where $\pi_i : Y_1 \times \cdots \times Y_n \to Y_i$ is the projection on the ith component. ■

Pasting Lemma. *Suppose $X = \bigcup_{\alpha \in I} A_\alpha$ and $f : X \to Y$ is a map such that $f|_{A_\alpha} : A_\alpha \to Y$ is continuous. Then f is continuous if one of the following holds*

(i) *Each A_α is open, or*

(ii) *Each A_α is closed, and I is finite.*

$\boxed{\textbf{Proof.}}$ (i) Let U be an open subset of Y. Then

$$f^{-1}(U) = f^{-1}(U) \cap \left(\bigcup_{\alpha \in I} A_\alpha \right) = \bigcup_{\alpha \in I} (f^{-1}(U) \cap A_\alpha) = \bigcup_{\alpha \in I} (f|_{A_\alpha})^{-1}(U).$$

Since A_α is open in X and $(f|_{A_\alpha})^{-1}(U)$ is open in A_α, $(f|_{A_\alpha})^{-1}(U)$ is also open in X. Now the result follows from the fact that the union of a family of open sets is open.

(ii) Let C be a closed subset of Y. Then

$$f^{-1}(C) = f^{-1}(C) \cap \left(\bigcup_{\alpha \in I} A_\alpha \right) = \bigcup_{\alpha \in I} (f^{-1}(C) \cap A_\alpha) = \bigcup_{\alpha \in I} (f|_{A_\alpha})^{-1}(C).$$

Since A_α is closed in X and $(f|_{A_\alpha})^{-1}(C)$ is closed in A_α, $(f|_{A_\alpha})^{-1}(C)$ is also closed in X. Now the result follows from the fact that the union of finitely many closed sets is closed. ■

Remark. If we have functions $f_\alpha : A_\alpha \to Y$ such that

$$f_\alpha|_{A_\alpha \cap A_\beta} = f_\beta|_{A_\alpha \cap A_\beta}$$

for all $\alpha, \beta \in I$, then we can define $f : X \to Y$ as $f|_{A_\alpha} := f_\alpha$. We can use this to give a different formulation of the pasting lemma.

Definition 11.67. A continuous bijective function whose inverse is also continuous is called a **homeomorphism**. Two spaces are said to be **homeomorphic** if there exists a homeomorphism between them.

Remark. It is easy to check that being homeomorphic is an equivalence relation.

***Example* 11.68.** The interval $(0, 1)$ is homeomorphic to $\mathbb{R}$. For example

$$x \mapsto \tan(\frac{\pi}{2}(2x - 1))$$

is a homeomorphism from $(0, 1)$ onto $\mathbb{R}$.

Proposition 11.69. *Suppose $f : X \to Y$ is a homeomorphism. Then f induces a bijection between the topology of X and the topology of Y.*

Proof. Since f^{-1} is continuous, $f(U)$ is open for every open set $U \subset X$. It is easy to see that the map $U \mapsto f(U)$ is a bijection between the topologies. ■

Proposition 11.70. *The map*

$$\begin{aligned} (X_1 \times \cdots \times X_k) \times (X_{k+1} \times \cdots \times X_n) &\to X_1 \times \cdots \times X_n \\ \big((x_1, \ldots, x_k)\ ,\ (x_{k+1}, \ldots, x_n)\big) &\mapsto (x_1, \ldots, x_n) \end{aligned}$$

is a homeomorphism.

Proof. Let U_i be an open subset of X_i. Then the restriction of the described map is a bijection between $(U_1 \times \cdots \times U_k) \times (U_{k+1} \times \cdots \times U_n)$ and $(U_1 \times \cdots \times U_n)$. These sets form bases for the topology of their corresponding spaces. Hence we get the desired result. ■

Definition 11.71. A map $f : X \to Y$ is called **continuous at a point** $x \in X$ if for each open set U containing $f(x)$, the set $f^{-1}(U)$ is a neighborhood of x.

Remark. The above theorems can be generalized to the case of maps that are continuous at a point.

Theorem 11.72. *A map is continuous if and only if it is continuous at every point.*

Proof. It is easy to see that continuity of a map implies its continuity at every point. For the converse, suppose that $f : X \to Y$ is continuous at every $x \in X$. Let U be an open subset of Y. Then for every $x \in f^{-1}(U)$ we have $f(x) \in U$. Thus $f^{-1}(U)$ must be a neighborhood of all of its elements. Therefore $f^{-1}(U)$ is a union of open sets, hence it is open. ■

11.7 Connectedness

Definition 11.73. A **separation** of a topological space X is a pair of nonempty open subsets A, B of X such that

$$X = A \cup B, \text{ and } A \cap B = \emptyset.$$

A topological space X is **disconnected** if there exists a separation of X.

A nonempty topological space is **connected** if it is not disconnected, i.e. there does not exist a separation of it. A nonempty subset of a space is connected if it is connected as a topological space with the subspace topology.

Notation. We use the notation $A \sqcup B$ for $A \cup B$, when $A \cap B = \emptyset$.

Theorem 11.74. *A space X is connected if and only if the only subsets of X that are both open and closed in X are $\emptyset, X$.*

Proof. Any other closed and open subset A, has a nonempty open complement A^c; and together they form a separation of X. ■

Theorem 11.75. *A nonempty subset of $\mathbb{R}$ is connected if and only if it is an interval, or has only one element.*

Proof. This is the same as Theorem 2.82. ■

Intermediate Value Theorem. *Suppose $f : X \to \mathbb{R}$ is continuous and X is connected. If $a, b \in f(X)$ and $a < c < b$ then $c \in f(X)$.*

Proof. If $c \notin f(X)$, then $f^{-1}((-\infty, c)) \sqcup f^{-1}((c, +\infty))$ is a separation of X. ■

Theorem 11.76. *The continuous image of a connected set is connected.*

$\boxed{\text{Proof.}}$ Suppose X is connected and $f : X \to Y$ is continuous. We want to show that $f(X)$ is connected. Let A be a nonempty open and closed subset of $f(X)$. It is enough to show that $A = f(X)$. Now $f^{-1}(A)$ is both open and closed. It is also nonempty, since A is nonempty and contains elements of the image of f. Thus $f^{-1}(A) = X$ since X is connected. Hence we have $A = f(f^{-1}(A)) = f(X)$ as desired. ■

Remark. An immediate consequence of the above theorem is that a space which is homeomorphic to a connected space, is connected.

Theorem 11.77. *The union of a family of connected sets that have a point in common, is connected.*

$\boxed{\text{Proof.}}$ Suppose $\{X_\alpha\}$ is a family of connected sets. Let $X = \bigcup X_\alpha$, and suppose $p \in \bigcap X_\alpha$. Suppose that X has a nonempty closed and open subset U. We either have $p \in U$ or $p \in U^c$. Suppose U contains p, otherwise we can work with the nonempty closed and open subset U^c. Then $U \cap X_\alpha$ is nonempty for all α. Also as $X_\alpha \subset X$, $U \cap X_\alpha$ is a closed and open subset of X_α. Hence by connectedness of X_α, we have $U \cap X_\alpha = X_\alpha$. Therefore U contains all X_α's and we have $U = X$. ■

Theorem 11.78. *Suppose A is a connected subset of X, and $A \subset B \subset \bar{A}$. Then B is also connected. In particular, the closure of a connected set is connected.*

$\boxed{\text{Proof.}}$ Suppose V is a nonempty closed and open subset of B. Then there is an open set $U \subset X$ such that $V = U \cap B$. Now

$$V \cap A = (U \cap B) \cap A = U \cap A$$

is a closed and open subset of A. Let us show that $U \cap A$ is nonempty. We know that there is $b \in U \cap B$. If $b \in A$ we are done. Otherwise we have $b \in \bar{A} - A$. Therefore b is a limit point of A. Thus as U is an open set containing b, U must intersect A. Hence by connectedness of A we get $U \cap A = A$.

Therefore V contains A. Since V is also closed in B, there is a closed subset C of X such that $V = C \cap B$. But then C contains A, and as C is closed we have $C \supset \bar{A}$. Consequently $V = C \cap B = B$. Thus B is connected. ■

Theorem 11.79. *The product of finitely many connected spaces is connected.*

$\boxed{\text{Proof.}}$ It is sufficient to prove the theorem for the product of two spaces. The general result follows by induction. Suppose X, Y are connected. For every $y \in Y$, the map that takes $x \mapsto (x, y)$ from $X \to X \times Y$ is continuous. Also, for every $x \in X$, the map that takes $y \mapsto (x, y)$ from $Y \to X \times Y$ is continuous. Therefore the images of theses maps, i.e. $X \times \{y\}$ and $\{x\} \times Y$ are connected, for all $x \in X$ and $y \in Y$.

Let $a \in X$ be a fixed element. Then for all $y \in Y$, $X \times \{y\} \cup \{a\} \times Y$ is the union of two connected sets having (a, y) in common, so it is connected. Therefore $X \times Y$ is the union of connected subsets

$$X \times Y = \bigcup_{y \in Y} (X \times \{y\} \cup \{a\} \times Y),$$

that have (a, b) in common, where $b \in Y$ is a fixed element. Hence $X \times Y$ is connected. ■

Theorem 11.80. *Suppose X is a topological space, and $x, y \in X$. Define the relation $x \sim y$ if there exists a connected subset that contains both x and y. Then $\sim$ is an equivalence relation. The equivalence classes of $\sim$ are connected and are called the* **(connected) components** *of X.*

Proof. It is obvious that $x \sim x$, since $\{x\}$ is connected. Also $x \sim y$ implies $y \sim x$ by the very definition of $\sim$. Now suppose $x \sim y$ and $y \sim z$. Then there are connected sets A, B such that $x, y \in A$ and $y, z \in B$. Since A, B both contain y, $A \cup B$ is connected. But $x, z \in A \cup B$, hence $x \sim z$.

Next suppose A is the equivalence class of x. Then for every $y \in A$ there is a connected set A_y such that $x, y \in A_y$. Note that for any $z \in A_y$ we have $z \sim x$. Therefore $A_y \subset A$, and consequently $A = \bigcup_{y \in A} A_y$. But $x \in \bigcap_{y \in A} A_y$, thus A is connected. ■

Theorem 11.81. *If a connected subset of a space intersects a component, it is contained in that component.*

Proof. Suppose B is a connected subset that intersects the component A. Let $x \in B \cap A$. Then for any $y \in B$ we have $x \sim y$, since $x, y \in B$. Thus $y \in A$ too, and hence $B \subset A$. ■

Theorem 11.82. *Components are closed.*

Proof. Suppose A is a path component. Then $\bar{A}$ is connected, since A is connected. On the other hand $\bar{A}$ intersects A, so we must have $\bar{A} \subset A$. Hence $A = \bar{A}$, and A is closed. ■

Definition 11.83. A topological space is called **locally connected** if every open neighborhood of every point contains a connected open neighborhood of that point.

Theorem 11.84. *The components of a locally connected space are both open and closed.*

Proof. Let A be a component. Then any $x \in A$ has a connected open neighborhood U_x. Since U_x intersects A, we have $U_x \subset A$. Thus $A = \bigcup_{x \in A} U_x$, and therefore A is open. Now, A^c is also open. Because A^c is the union of other components, which are all open. Hence A is closed too. ■

11.8 Path Connectedness

Definition 11.85. Let X be a topological space, and $x, y \in X$. A **path** in X from x to y, is a continuous function f from an interval $[a, b]$ to X such that $f(a) = x$ and $f(b) = y$. A nonempty topological space X is called **path connected** if for any two points $x, y \in X$ there exists a path from x to y. A nonempty subset A of a space X is path connected if it is path connected as a topological space with the subspace topology, in other words between any two points of A there is a path inside A.

Remark. Note that by a linear change of variable, we can assume that the domain of a path is any given closed interval.

Theorem 11.86. *A path connected space is connected.*

Proof. Suppose to the contrary that X is path connected, and there is a separation $A \sqcup B$ of X. Let $x \in A$ and $y \in B$. Then there is a path $f : [a, b] \to X$ from x to y. But it is easy to see that $f^{-1}(A) \sqcup f^{-1}(B)$ is a separation of $[a, b]$, which is a contradiction. ■

Theorem 11.87. *The continuous image of a path connected set is path connected.*

Proof. Suppose X is path connected and $f : X \to Y$ is continuous. Let $x, y \in f(X)$. Then there are $z, w \in X$ such that $f(z) = x$ and $f(w) = y$. Now let $g : [a, b] \to X$ be a path from z to w. Then $f \circ g : [a, b] \to f(X)$ is a path from x to y. ■

Remark. An immediate consequence of the above theorem is that a space which is homeomorphic to a path connected space, is path connected.

Theorem 11.88. *The union of a family of path connected sets that have a point in common, is path connected.*

Proof. Suppose $\{X_\alpha\}$ is a family of path connected sets. Let $X = \bigcup X_\alpha$, and suppose $p \in \bigcap X_\alpha$. Let $x_1, x_2 \in X$. Suppose $x_1 \in X_1$ and $x_2 \in X_2$. Then there are paths $f_i : [a, b] \to X_i$ from x_i to p. Now by pasting lemma

$$f(t) := \begin{cases} f_1(t) & t \in [a, b] \\ f_2(2b - t) & t \in [b, 2b - a] \end{cases}$$

is a continuous path from x_1 to x_2. ■

Theorem 11.89. *The product of finitely many path connected spaces is path connected.*

Proof. It is sufficient to prove the theorem for the product of two spaces. The general result follows by induction. Suppose X, Y are path connected spaces. Let $(x_1, y_1), (x_2, y_2) \in X \times Y$. Then there are paths $f : [a, b] \to X$ from x_1 to x_2, and $g : [a, b] \to Y$ from y_1 to y_2. Now

$$\begin{aligned}(f, g) : [a, b] &\longrightarrow X \times Y \\ t &\mapsto (f(t), g(t))\end{aligned}$$

is a continuous path from (x_1, y_1) to (x_2, y_2). ■

Theorem 11.90. *Suppose X is a topological space, and $x, y \in X$. Define the relation $x \sim y$ if there exists a path connected subset that contains both x and y. Then $\sim$ is an equivalence relation. The equivalence classes of $\sim$ are path connected and are called the* ***path components*** *of X.*

Proof. It is obvious that $x \sim x$, since $\{x\}$ is path connected. Also $x \sim y$ implies $y \sim x$ by the very definition of $\sim$. Now suppose $x \sim y$ and $y \sim z$. Then there are path connected sets A, B such that $x, y \in A$ and $y, z \in B$. Since A, B both contain y, $A \cup B$ is path connected. But $x, z \in A \cup B$, hence $x \sim z$.

Next suppose A is the equivalence class of x. Then for every $y \in A$ there is a path connected set A_y such that $x, y \in A_y$. Note that for any $z \in A_y$ we have $z \sim x$. Therefore $A_y \subset A$, and consequently $A = \bigcup_{y \in A} A_y$. But $x \in \bigcap_{y \in A} A_y$, thus A is path connected. ■

Remark. Since path components are path connected, they are connected. Thus every path component is contained in the component of its points by Theorem 11.81.

Theorem 11.91. *The path component of X containing a point x is the set of all points $y \in X$ such that there is a path from x to y.*

Proof. If there is path from x to y then the image of that path is a path connected subset that contains both x, y, hence $x \sim y$. Conversely, if $x \sim y$ then x, y belong to a path connected subset of X, hence there is a path from x to y. ■

Remark. We could have defined the path components using the equivalence relation of the existence of a path between any two given points. Then the above theorem implies that we would have obtained the same classes.

Theorem 11.92. *If a path connected subset of a space intersects a path component, it is contained in that path component.*

Proof. Suppose B is a path connected subset that intersects the path component A. Let $x \in B \cap A$. Then for any $y \in B$ we have $x \sim y$, since $x, y \in B$. Thus $y \in A$ too, and hence $B \subset A$. ■

***Example* 11.93.** Not every connected set is path connected. For example the *topologist's sine curve*

$$X := \{(x,y) : y = \sin\frac{1}{x},\ x \in (0,1]\} \bigcup \{(0,y) : y \in [-1,1]\}$$

is a connected subset of $\mathbb{R}^2$ which is not path connected. Let G be the graph of $\sin\frac{1}{x}$ for $x \in (0,1]$. First note that G is path connected, since any two points $\sin\frac{1}{a}$ and $\sin\frac{1}{b}$ on it can be joined by the continuous path $\sin\frac{1}{x}|_{[a,b]}$. So G is also connected. Now note that $X = \overline{G}$. Because if

$$(x_n, \sin\frac{1}{x_n}) \to (t,s)$$

and $(t,s) \notin G$, then $t = 0$, since otherwise (t,s) would belong to G due to the continuity of $\sin\frac{1}{x}$ on its domain. Then as $|\sin| \le 1$, we have $|s| \le 1$. Hence $\overline{G} \subset X$. Conversely, for any $y \in [-1,1]$ we have

$$\big(\frac{1}{c+2n\pi}, \sin(c+2n\pi)\big) \to (0,y),$$

where $c \in (2\pi, 4\pi)$ satisfies $\sin c = y$. Thus $\overline{G} = X$. Therefore X is connected.

Now suppose to the contrary that X is path connected. Let f be a continuous path from the point $(0,0)$ to the point $(1, \sin 1)$, defined on the interval $[0,1]$. Let $S := X - G$ be the closed line segment $\{0\} \times [-1,1]$. Let $A := f^{-1}(S) \subset [0,1]$. Note that A is closed, since S is closed. We will show that A is also open. Let $\tau \in A$, and suppose $z = f(\tau)$. Then $z \in S$. Let B be an open box around z. Then $B \cap X$ consists of infinitely many disjoint small paths on G and a small line segment $B \cap S$ containing z. These are the path components of $B \cap X$. Also, $f^{-1}(B \cap X)$ is an open neighborhood of τ in $[0,1]$. Now $f : f^{-1}(B \cap X) \to B \cap X$ is continuous. Let $I \subset f^{-1}(B \cap X)$ be a small connected open interval around τ. Then $f(I)$ is a path connected subset of $B \cap X$, since I is path connected. So $f(I)$ is contained in one of the path components of $B \cap X$. But $f(I)$ must contain $z = f(\tau)$. Hence $f(I)$ is contained in the line segment $B \cap S$. Thus $I \subset f^{-1}(S) = A$. Therefore A is open, since τ was arbitrary. So as A is nonempty, and $[0,1]$ is connected, we must have $A = [0,1]$. But this means that the image of f is completely inside S, which is a contradiction.

As an interesting consequence, we see that G, i.e. the graph of $\sin\frac{1}{x}$, is a path connected set whose closure is not path connected. ■

Definition 11.94. A topological space is called **locally path connected** if every open neighborhood of every point contains a path connected open neighborhood of that point.

Remark. It is obvious that a locally path connected space is locally connected.

Theorem 11.95. *In a space where every point has a path connected open neighborhood, components and path components are the same. In particular, the components and path components of a locally path connected space are the same.*

Proof. Let A be a path component. Then any $x \in A$ has a path connected open neighborhood U_x. Since U_x intersects A, we have $U_x \subset A$. Thus $A = \bigcup_{x \in A} U_x$, and therefore A is open. Now, A^c is also open. Because A^c is the union of other path components, which are all open. Hence A is closed too. Finally, let B be the component containing $x \in A$. Then we know that $A \subset B$. But A is both closed and open, and B is connected, hence $A = B$ as desired. ■

***Example* 11.96.** A connected space that is not path connected cannot be locally path connected. Because it has only one component, and if it was locally path connected it would have only one path component. So it must have been path connected, contrary to the assumption.

11.9 Compactness

Definition 11.97. An **open covering** of a subset A of a topological space X, is a family $\mathcal{U}$ of open subsets of X such that

$$\bigcup \mathcal{U} = \bigcup_{U \in \mathcal{U}} U \supset A.$$

A **subcovering** $\mathcal{V}$ of $\mathcal{U}$ is a subfamily of $\mathcal{U}$ which is itself an open covering of A.

Definition 11.98. A subset A of a topological space X is **compact** if every open covering of A has a finite subcovering.

Remark. Note that we do not say that A has a finite open covering. Rather, we say that from any open covering of A we can choose finitely many open sets whose union covers A.

Remark. It is easy to see that a subset of a space is compact if and only if it is compact as a space equipped with the subspace topology.

Remark. Note that we consider the empty set $\emptyset$ to be compact. In all of the following theorems, you should check that the claim holds for the empty compact set trivially.

Theorem 11.99. *Closed subsets of a compact space are compact.*

Proof. Let A be a closed subset of the compact space X. Let $\{U_\alpha\}$ be an open covering of A. Then $\{U_\alpha, A^c\}$ is an open covering of X. Thus we have a finite subcovering $\{U_{\alpha_1}, \ldots, U_{\alpha_n}, A^c\}$ of X. Then $\{U_{\alpha_1}, \ldots, U_{\alpha_n}\}$ is a finite subcovering of A. ■

Theorem 11.100. *Compact subsets of a Hausdorff space are closed.*

Proof. Let C be a compact subset of X. If $C = X$ then it is closed. Otherwise, we will show that C^c is open. Let a be an arbitrary element of C^c. It suffices to show that a has an open neighborhood contained in C^c. For every $x \in C$ there are disjoint open sets U_x, V_x such that $x \in U_x$ and $a \in V_x$. Then $\{U_x\}_{x \in C}$ is an open covering of C. Hence it has a finite subcovering $\{U_1, \dots, U_n\}$. Now $\bigcap_{i \le n} V_i$ is an open neighborhood of a that does not intersect $\bigcup_{i \le n} U_i$, hence it does not intersect C. Therefore C^c is open. ■

Theorem 11.101. *Compact subsets of a metric space are closed and bounded.*

Proof. Metric spaces are Hausdorff, so their compact subsets are closed. To prove the boundedness, we can cover the compact subset by the open covering $\{B_n(a)\}_{n \in \mathbb{N}}$ for some a in the space. Now a finite subcovering of this covering must cover the compact subset. Hence it is bounded. ■

Theorem 11.102. *The continuous image of a compact set is compact.*

Proof. Suppose $f : X \to Y$ is continuous, and A is a compact subset of X. Let $\{U_\alpha\}$ be an open covering of $f(A)$. Then $\{f^{-1}(U_\alpha)\}$ is an open covering of A. Thus we have a finite subcovering $\{f^{-1}(U_{\alpha_1}), \dots, f^{-1}(U_{\alpha_n})\}$ of A. Then $\{U_{\alpha_1}, \dots, U_{\alpha_n}\}$ is a finite subcovering of $f(A)$. ■

Remark. An immediate consequence of the above theorem is that a space which is homeomorphic to a compact space, is compact.

Extreme Value Theorem. *A continuous function from a nonempty compact set into $\mathbb{R}$ is bounded, and achieves its maximum and minimum values.*

Proof. Let $f : X \to \mathbb{R}$ be continuous, and suppose X is compact. Then $f(X)$ is compact in the metric space $\mathbb{R}$. Hence $f(X)$ is nonempty, bounded and closed. Every nonempty, closed and bounded subset of $\mathbb{R}$ contains its finite supremum and infimum. Therefore the supremum and infimum of $f(X)$ are achieved by f, i.e. there are $x_1, x_2 \in X$ such that

$$f(x_2) = \sup\{f(x) : x \in X\}, \qquad f(x_1) = \inf\{f(x) : x \in X\}.$$

These are the maximum and the minimum of f respectively. ■

Theorem 11.103. *Closed bounded intervals in $\mathbb{R}$ are compact.*

Proof. The empty set is obviously compact. Let $\mathcal{U}$ be an open covering of $[a, b]$. Set

$$S = \{x \in [a, b] : [a, x] \text{ is covered by a finite subcovering of } \mathcal{U}\},$$

and $c = \sup S$. Note that an interval of the form $[a, a+\delta]$ is contained in some open set in $\mathcal{U}$, so S is nonempty and $a < c$. Let U be an open set in $\mathcal{U}$ that contains c. Then U contains $[c-\epsilon, c]$ for some positive ϵ. If $c \neq b$, then U will contain $[c, c+\epsilon]$ too.

Since $c-\epsilon$ is not an upper bound of S, there is $s \in S$ between $c-\epsilon$ and c. Hence $[a, s]$ is covered by a finite subcovering of $\mathcal{U}$. Therefore by adding U to that finite subcovering, we see that $[a, c]$ is also covered by a finite subcovering of $\mathcal{U}$. Thus $c \in S$. If $c \neq b$ then this finite subcovering of $[a, c]$ that contains U, will also cover $[a, c+\epsilon]$. This contradicts the fact that c is an upper bound of S. Therefore $c = b$ and $b \in S$. ■

Theorem 11.104. *The product of finitely many compact spaces is compact.*

Proof. It is enough to prove the theorem for the product of two spaces. The general result follows by induction. If one of the spaces is empty, the product is empty and compact. Suppose X, Y are nonempty compact spaces. For every $y \in Y$, the map that takes $x \mapsto (x, y)$ from $X \to X \times Y$ is continuous. Also, for every $x \in X$, the map that takes $y \mapsto (x, y)$ from $Y \to X \times Y$ is continuous. Therefore the images of theses maps, i.e. the slices $X \times \{y\}$ and $\{x\} \times Y$ are compact, for all $x \in X$ and $y \in Y$.

Consider an open covering $\mathcal{U}$ of $X \times Y$. It has a finite subcovering $\mathcal{U}'$ that covers a slice $\{x_0\} \times Y$ for a fixed $x_0 \in X$. Now, any point (x_0, y) in the slice has an open neighborhood of the form $U_y \times V_y$ that is contained in $\bigcup \mathcal{U}'$. Since $\{V_y\}_{y \in Y}$ covers Y, it has a finite subset $\{V_1, \ldots, V_n\}$ that covers Y too. Therefore

$$(\bigcap_{i \leq n} U_i) \times Y \subset \bigcup \mathcal{U}'.$$

Hence for any $x \in X$ there is an open neighborhood U_x of x, such that $U_x \times Y$ is covered by a finite subcovering of $\mathcal{U}$. Now $\{U_x\}_{x \in X}$ covers X. Hence it has a finite subset $\{U_1, \ldots, U_m\}$ that covers X too. Therefore $X \times Y$ is covered by the union of finite subcoverings of $\mathcal{U}$ that contain $U_j \times Y$ for $j = 1, \ldots, m$. ■

***Example* 11.105.** The product of closed bounded intervals $[a_1, b_1] \times \cdots \times [a_n, b_n]$ is compact in $\mathbb{R}^n$.

Heine-Borel Theorem. *A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded in the Euclidean metric.*

Proof. Any bounded subset is contained in the product of some closed bounded intervals, which is a compact set. Hence if the subset is closed it is compact. For the converse note that $\mathbb{R}^n$ is a metric space, so its compact subsets are closed and bounded. ■

Theorem 11.106. *An infinite subset of a compact space has a limit point.*

Proof. Suppose to the contrary that A is an infinite subset of the compact space X that does not have a limit point. Then any $a \in A$ has an open neighborhood U_a such that $U_a \cap A = \{a\}$, since a is not a limit point of A. On the other hand, $\bar{A}$ is the union of A and its limit points. Thus $\bar{A} = A$. Hence A is closed in X. Therefore A is compact. Now $\{U_a\}_{a \in A}$ is an open covering of A. Thus for some $a_1, \dots, a_n \in A$, $U_{a_1}, \dots, U_{a_n}$ will cover A. But then we must have

$$A \subset A \cap \bigcup_{i \le n} U_{a_i} = \bigcup_{i \le n} (A \cap U_{a_i}) = \bigcup_{i \le n} \{a_i\},$$

which is a contradiction. ■

Theorem 11.107. *Any sequence in a compact first countable space has a convergent subsequence.*

Proof. Let (a_n) be a sequence in the compact set A. If the set $\{a_n\}$ is finite, then obviously (a_n) has a constant subsequence, and the constant subsequence is convergent. So we assume that the set $\{a_n\}$ is infinite. Now suppose to the contrary that no subsequence of (a_n) converges to a point of A. For any $x \in A$ let $\{U_{x,n}\}_{n \in \mathbb{N}}$ be a decreasing countable basis at x. Then for any $x \in A$ there is n_x such that U_{x,n_x} contains at most a finite number of points of $\{a_n\}$. Since otherwise there is $a \in A$ such that for all m the set $U_{a,m}$ contains infinitely many points of $\{a_n\}$. Then for each m we can choose $a_{n_m} \in U_{a,m}$ such that $n_m > n_{m-1}$. Now it is easy to see that the subsequence (a_{n_m}) converges to a, contrary to our assumption.

Thus $U_{x,n_x} \cap \{a_n\}$ is a finite set for every $x \in A$. Now $\{U_{x,n_x} : x \in A\}$ is an open covering of the compact set A. Hence it has a finite subcovering, namely

$$A \subset U_{x_1,n_{x_1}} \cup \cdots \cup U_{x_k,n_{x_k}},$$

for some $x_1, \dots, x_k \in A$. But this implies that infinitely many points of $\{a_n\}$ must belong to at least one of the open sets $U_{x_i,n_{x_i}}$, which is a contradiction. Consequently (a_n) has a convergent subsequence in A. ■

Bolzano-Weierstrass Theorem. *A bounded sequence in $\mathbb{R}^n$ has a convergent subsequence.*

Proof. Any bounded sequence is contained in the product of some closed bounded intervals, which is a compact set in the first countable space $\mathbb{R}^n$. ■

Theorem 11.108. *A continuous bijection from a compact space into a Hausdorff space is a homeomorphism.*

Proof. Let $f : X \to Y$ be a continuous bijection, and suppose X is compact. We have to show that f^{-1} is continuous. For any closed set $C \subset X$ we have $(f^{-1})^{-1}(C) = f(C)$. But C is compact, hence $f(C)$ is compact too. Thus $f(C)$ is closed, and f^{-1} is continuous. ■

Theorem 11.109. *Compact metric spaces are second countable.*

Proof. Suppose X is a compact metric space. Let $n \in \mathbb{N}$. Then the family $\{B_{\frac{1}{n}}(x)\}_{x\in X}$ is an open covering of X. Hence there are finitely many points $x_1, \dots, x_k \in X$ such that

$$X \subset B_{\frac{1}{n}}(x_1) \cup \cdots \cup B_{\frac{1}{n}}(x_k). \tag{$*$}$$

Let us rename these points and call them $x_{n,1}, \dots, x_{n,k_n}$. Now we claim that the family of open balls

$$\mathcal{B} := \{B_{\frac{1}{n}}(x_{n,i}) : n \in \mathbb{N},\, i \le k_n\}$$

is a countable basis for X. To prove this it suffices to show that for every open set $U \subset X$ and every $a \in U$ there is $B \in \mathcal{B}$ such that $a \in B \subset U$. There is $r > 0$ so that $B_r(a) \subset U$. Now for every $n \in \mathbb{N}$ there is x_{n,i_n} such that $a \in B_{\frac{1}{n}}(x_{n,i_n})$ by $(*)$. It is easy to see that $x_{n,i_n} \to a$ as $n \to \infty$. Therefore for large enough n we have $x_{n,i_n} \in B_{\frac{r}{2}}(a)$. Let n be large enough so that $\frac{1}{n} < \frac{r}{2}$. Then we have

$$a \in B_{\frac{1}{n}}(x_{n,i_n}) \subset B_r(a) \subset U,$$

as desired. Finally note that $\mathcal{B}$ is countable since it is the union of countably many finite collections. ■

Definition 11.110. A **locally compact** topological space is a space in which every point has a compact neighborhood.

Appendix A

Matrices

Definition A.1. Let F be a field, and $m, n \in \mathbb{N}$. An $m \times n$ **matrix** with entries in F is a function

$$A : \{(i, j) : i, j \in \mathbb{N},\ i \leq m,\ j \leq n\} \to F.$$

We denote by A_{ij} (or $A_{i,j}$) the value of A at (i, j), and call it the ijth **entry** of A. The matrix A is usually denoted as a rectangular array of elements of F with m rows and n columns

$$A = [A_{ij}] = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix}.$$

The $1 \times n$ matrix $[A_{i1}, \dots, A_{in}]$ is called the ith **row** of A, and is denoted by $A_{i,\cdot}$. Also, the $m \times 1$ matrix

$$\begin{bmatrix} A_{1j} \\ \vdots \\ A_{mj} \end{bmatrix}$$

is called the jth **column** of A, and is denoted by $A_{\cdot,j}$. A $1 \times n$ matrix is also called a **row vector**, and an $m \times 1$ matrix is also called a **column vector**. The set of $m \times n$ matrices with entries in F is denoted by $F^{m\times n}$. The **size** of a matrix $A \in F^{m\times n}$ is $m \times n$.

Remark. We know that F^n is the set of *ordered n-tuples* of elements of F. In order to make this precise, we can define F^n to be the set of functions

$$a : \{1, 2, \dots, n\} \to F.$$

Then we denote by a_i the value of a at i, and we call it the ith **component** of a. We will also denote a by the following familiar notation

$$a = (a_1, \dots, a_n),$$

and we also call it a **vector**. We can identify F^n with both $F^{1\times n}$ and $F^{n\times 1}$ via the maps

$$\begin{aligned}(a_1,\dots,a_n) &\mapsto [a_1,\dots,a_n],\\ (a_1,\dots,a_n) &\mapsto \begin{bmatrix} a_1\\ \vdots\\ a_n\end{bmatrix}.\end{aligned}$$

In particular, we always identify F with $F^{1\times 1}$. We also refer to the i,1th entry of a column vector, or the 1,ith entry of a row vector, as the ith **component** of them.

Remark. Note that as matrices are functions into F, it suffices to define them by specifying their ijth entry for every i,j. Also, when we want to show that two matrices are equal, it is enough to check the equality of their ijth entry for each i,j. The same things apply to the elements of F^n.

Definition A.2. Let F be a field, and $m,n\in\mathbb{N}$. The $m\times n$ **zero matrix** is a matrix whose entries are all zero. We often denote the zero matrix simply by 0. A **square matrix** is a matrix for which $m=n$, i.e. a matrix that has the same number of rows and columns. The **(main) diagonal** of a square matrix A is the n-tuple $(A_{11},A_{22},\dots,A_{nn})\in F^n$. The entries A_{ii} are referred to as the *diagonal entries* of A. The square matrix A is called **upper triangular** if $A_{ij}=0$ for $j<i$. In other words, the entries of A below its main diagonal are zero, so A has the form

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n}\\ 0 & A_{22} & \dots & A_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & A_{nn}\end{bmatrix}.$$

Similarly, a square matrix A is called **lower triangular** if $A_{ij}=0$ for $j>i$. A **diagonal matrix** is a square matrix A for which $A_{ij}=0$ when $i\neq j$, so it has the form

$$\begin{bmatrix} A_{11} & 0 & \cdots & 0\\ 0 & A_{22} & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & A_{nn}\end{bmatrix}.$$

A special diagonal matrix is the $n\times n$ **identity matrix**, which is defined by

$$I_{ij}=(I_n)_{ij}:=\begin{cases}0 & i\neq j,\\ 1 & i=j.\end{cases}$$

Definition A.3. Let F be a field, and $m, n \in \mathbb{N}$. The **addition** of two $m \times n$ matrices A, B with entries in F, is defined by

$$(A+B)_{ij} := A_{ij} + B_{ij}.$$

The **multiplication** of an $m \times n$ matrix A with an $n \times l$ matrix B is an $m \times l$ matrix AB, which is defined by

$$(AB)_{ij} := \sum_{k=1}^{n} A_{ik} B_{kj}.$$

The **scalar multiplication** of $a \in F$ and $A \in F^{m \times n}$ is defined by

$$(aA)_{ij} := aA_{ij}.$$

The **transpose** of an $m \times n$ matrix A is the $n \times m$ matrix A^{T} that satisfies

$$(A^{\mathsf{T}})_{ij} := A_{ji}.$$

Notation. For a matrix A we set $-A := (-1)A$, so $(-A)_{ij} = -A_{ij}$. Also, for two $m \times n$ matrices A, B we set $A - B := A + (-B)$.

Remark. Remember that we can identify F^n with both $F^{n\times 1}$ and $F^{1\times n}$. These identifications allow us to apply the above operations to the elements of F^n. In particular the addition and scalar multiplication on F^n are defined as follows

$$\begin{aligned} (a_1, \dots, a_n) + (b_1, \dots, b_n) &:= (a_1 + b_1, \dots, a_n + b_n), \\ a(a_1, \dots, a_n) &:= (aa_1, \dots, aa_n), \end{aligned}$$

where $a \in F$ and $(a_1, \dots, a_n), (b_1, \dots, b_n) \in F^n$. In addition, the zero vector is $0 = (0, \dots, 0)$, and we set $-(a_1, \dots, a_n) := (-a_1, \dots, -a_n)$. Note that these operations will also have the properties stated in the next theorem, since they are equivalent to the operations on matrices.

Remark. Let $A, B \in F^{m\times n}$ and $a \in F$. It is easy to show that for every i, j we have

$$\begin{aligned} (A+B)_{i,\cdot} &= A_{i,\cdot} + B_{i,\cdot}, & (aA)_{i,\cdot} &= aA_{i,\cdot}, & (A_{i,\cdot})^{\mathsf{T}} &= A^{\mathsf{T}}_{\cdot,i}, \\ (A+B)_{\cdot,j} &= A_{\cdot,j} + B_{\cdot,j}, & (aA)_{\cdot,j} &= aA_{\cdot,j}, & (A_{\cdot,j})^{\mathsf{T}} &= A^{\mathsf{T}}_{j,\cdot}. \end{aligned}$$

Theorem A.4. *Let F be a field. Then for all $L \in F^{p\times m}$, $A, B, E \in F^{m\times n}$, $C \in F^{n\times l}$, and $a, b \in F$ we have*

(i) *The addition of matrices is associative and commutative, i.e.*

$$A + (B + E) = (A + B) + E, \qquad A + B = B + A.$$

(ii) *Let* $0 \in F^{m\times n}$ *be the zero matrix, then*

$$A + 0 = A, \qquad A + (-A) = 0.$$

(iii) $1A = A$*, and* $I_m A = A = AI_n$.

(iv) *We have*

$$L(A+B) = LA + LB, \qquad (A+B)C = AC + BC.$$

(v) *We have*

$$\begin{aligned} a(A+B) &= aA + aB, \qquad & (a+b)A &= aA + bA, \\ (aA)C = a(AC) &= A(aC), \qquad & a(bA) &= (ab)A. \end{aligned}$$

(vi) *If* A *or* C *is the zero matrix, then* AC *is the zero matrix. Also, if* a *is zero, or* A *is the zero matrix, then* aA *is the zero matrix.*

(vii) *We have*

$$\begin{aligned} (A+B)^{\mathsf{T}} &= A^{\mathsf{T}} + B^{\mathsf{T}}, \qquad & (aA)^{\mathsf{T}} &= aA^{\mathsf{T}}, \\ (AC)^{\mathsf{T}} &= C^{\mathsf{T}}A^{\mathsf{T}}, \qquad & (A^{\mathsf{T}})^{\mathsf{T}} &= A. \end{aligned}$$

Proof. **(i)** For each i, j we have

$$\begin{aligned} \big(A + (B+E)\big)_{ij} &= A_{ij} + (B+E)_{ij} = A_{ij} + (B_{ij} + E_{ij}) \\ &= (A_{ij} + B_{ij}) + E_{ij} = (A+B)_{ij} + E_{ij} = \big((A+B)+E\big)_{ij}. \end{aligned}$$

The other one is similar.

(ii) This is similar to (i).

(iii) It is obvious that $1A = A$. For the second part we have

$$(I_m A)_{ij} = \sum_{k\le m} (I_m)_{ik} A_{kj} = 0A_{1j} + \cdots + 1A_{ij} + \cdots + 0A_{mj} = A_{ij}.$$

The other half is similar.

(iv) We have

$$\begin{aligned} \big((A+B)C\big)_{ij} &= \sum_{k\le n} (A+B)_{ik} C_{kj} = \sum_{k\le n} (A_{ik} + B_{ik}) C_{kj} \\ &= \sum_{k\le n} A_{ik} C_{kj} + \sum_{k\le n} B_{ik} C_{kj} = (AC)_{ij} + (BC)_{ij}. \end{aligned}$$

The other one is similar.

(v) We only prove $a(AC) = A(aC)$, the others can be proved similarly. We have

$$\begin{aligned}\big(A(aC)\big)_{ij} &= \sum_{k\leq n} A_{ik}(aC)_{kj} = \sum_{k\leq n} A_{ik}aC_{kj} \\ &= a\sum_{k\leq n} A_{ik}C_{kj} = a(AC)_{ij} = \big(a(AC)\big)_{ij}.\end{aligned}$$

(vi) These are all easy to show.
(vii) We have $\big((A^{\intercal})^{\intercal}\big)_{ij} = (A^{\intercal})_{ji} = A_{ij}$. Also

$$\big((aA)^{\intercal}\big)_{ij} = (aA)_{ji} = aA_{ji} = a(A^{\intercal})_{ij} = (aA^{\intercal})_{ij}.$$

In addition, we have

$$\begin{aligned}\big((AC)^{\intercal}\big)_{ij} &= (AC)_{ji} = \sum_{k\leq n} A_{jk}C_{ki} \\ &= \sum_{k\leq n} C_{ki}A_{jk} = \sum_{k\leq n}(C^{\intercal})_{ik}(A^{\intercal})_{kj} = (C^{\intercal}A^{\intercal})_{ij}.\end{aligned}$$

The other one is similar. ■

Remark. As a consequence of the above theorem, we can easily show by induction that if $A_1, \dots, A_k \in F^{n\times n}$ then we have

$$(A_1 \cdots A_k)^{\intercal} = A_k^{\intercal} \cdots A_1^{\intercal}.$$

Theorem A.5. *The multiplication of matrices is associative, i.e. for any field F and all matrices $A \in F^{p\times m}$, $B \in F^{m\times n}$, and $C \in F^{n\times l}$, we have*

$$(AB)C = A(BC).$$

Proof. We have

$$\begin{aligned}\big((AB)C\big)_{ij} &= \sum_{k=1}^{n}(AB)_{ik}C_{kj} = \sum_{k=1}^{n}\Big(\sum_{l=1}^{m} A_{il}B_{lk}\Big)C_{kj} \\ &= \sum_{k=1}^{n}\sum_{l=1}^{m} A_{il}B_{lk}C_{kj} = \sum_{l=1}^{m}\sum_{k=1}^{n} A_{il}B_{lk}C_{kj} \\ &= \sum_{l=1}^{m} A_{il}\Big(\sum_{k=1}^{n} B_{lk}C_{kj}\Big) = \sum_{l=1}^{m} A_{il}(BC)_{lj} = \big(A(BC)\big)_{ij}.\end{aligned}$$ ■

***Example* A.6.** Let $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ be matrices in $F^{2\times 2}$, for some field F. Then we have

$$AB = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \neq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = BA.$$

Hence the multiplication of matrices is not in general commutative. This example also shows that the product of two nonzero matrices can be zero. Hence the cancellation law does not hold for matrix multiplication, i.e. for $A, B, C \in F^{n\times n}$

$$AC = BC \;\not\Longrightarrow\; A = B.$$

Theorem A.7. *Suppose F is a field, and $A \in F^{m\times n}$, $C \in F^{n\times l}$. Then we have*

$$(AC)_{ij} = A_{i,\cdot}C_{\cdot,j}, \qquad (AC)_{\cdot,j} = AC_{\cdot,j}, \qquad (AC)_{i,\cdot} = A_{i,\cdot}C.$$

Remark. In other words, the jth column of AC is the product of A and the jth column of C. And the ith row of AC is the product of the ith row of A, and C.

Proof. Since $A_{i,\cdot}$ and $C_{\cdot,j}$ are respectively $1 \times n$ and $n \times 1$ matrices, their product is a 1×1 matrix, i.e. an element of F; and we have

$$(A_{i,\cdot}C_{\cdot,j})_{1,1} = \sum_{k\le n}(A_{i,\cdot})_{1,k}(C_{\cdot,j})_{k,1} = \sum_{k\le n} A_{i,k}C_{k,j} = (AC)_{ij}.$$

Similarly, $(AC)_{\cdot,j}$ and $(AC)_{i,\cdot}$ are respectively $m \times 1$ and $1 \times l$ matrices. Hence we have

$$\big((AC)_{\cdot,j}\big)_{i,1} = (AC)_{i,j} = \sum_{k\le n} A_{ik}C_{kj} = \sum_{k\le n} A_{ik}(C_{\cdot,j})_{k,1} = (AC_{\cdot,j})_{i,1},$$
$$\big((AC)_{i,\cdot}\big)_{1,j} = (AC)_{i,j} = \sum_{k\le n} A_{ik}C_{kj} = \sum_{k\le n}(A_{i,\cdot})_{1,k}C_{kj} = (A_{i,\cdot}C)_{1,j}.$$ ■

***Exercise* A.8.** Suppose $A \in F^{m\times n}$ and $C \in F^{n\times l}$. Show that

$$AC = \sum_{k\le n} A_{\cdot,k}C_{k,\cdot}.$$

Solution. Note that $A_{\cdot,k}$ and $C_{k,\cdot}$ are respectively $m \times 1$ and $1 \times l$ matrices. Hence their product is an $m \times l$ matrix. Now we have

$$\begin{aligned}\Big(\sum_{k\le n} A_{\cdot,k}C_{k,\cdot}\Big)_{i,j} &= \sum_{k\le n}\big(A_{\cdot,k}C_{k,\cdot}\big)_{i,j} = \sum_{k\le n}\sum_{s=1}^{1}(A_{\cdot,k})_{i,s}(C_{k,\cdot})_{s,j} \\ &= \sum_{k\le n}(A_{\cdot,k})_{i,1}(C_{k,\cdot})_{1,j} = \sum_{k\le n} A_{i,k}C_{k,j} = (AC)_{i,j},\end{aligned}$$

as desired. ■

Notation. Let $j, n \in \mathbb{N}$, and suppose $j \leq n$. We denote by e_j the column vector in $F^{n\times 1}$ whose components are all zero except for its jth component which is one, i.e.

$$e_j := \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow j\text{-th row.}$$

We call this the jth vector of the *standard basis* of $F^{n\times 1}$. We also have

$$e_j^{\mathsf{T}} = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \in F^{1\times n}.$$

We call this the jth vector of the standard basis of $F^{1\times n}$. We also sometimes abuse the notation and call e_j's or e_j^{T}'s the standard basis vectors of F^n. Note that we use the same notation for every n. For example e_1 can be any of the followings

$$[1], \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \dots .$$

But this should cause no confusion, since the value of n is usually evident from the context.

Remark. Let I be the identity matrix. Then we have $I_{i,.} = e_i^{\mathsf{T}}$, and $I_{.,j} = e_j$.

Theorem A.9. *Suppose F is a field, and $A \in F^{m\times n}$. Let $x = [x_1, \dots, x_n]^{\mathsf{T}} \in F^{n\times 1}$ be a column vector, and let $y = [y_1, \dots, y_m] \in F^{1\times m}$ be a row vector. Then we have*

$$Ax = \sum_{j\leq n} x_j A_{.,j}, \qquad yA = \sum_{i\leq m} y_i A_{i,.}.$$

In particular for $e_j \in F^{n\times 1}$ and $e_i^{\mathsf{T}} \in F^{1\times m}$ we have

$$Ae_j = A_{.,j}, \qquad e_i^{\mathsf{T}} A = A_{i,.}.$$

Remark. We say that Ax is a *linear combination* of the columns of A, and yA is a linear combination of the rows of A.

Proof. We know that Ax and yA are respectively $m\times 1$ and $1\times n$ matrices. Then we have

$$(Ax)_{i,1} = \sum_{j\leq n} A_{ij}x_j = \sum_{j\leq n} x_j (A_{.,j})_{i,1} = \Big(\sum_{j\leq n} x_j A_{.,j}\Big)_{i,1},$$

$$(yA)_{1,j} = \sum_{i\leq m} y_i A_{ij} = \sum_{i\leq m} y_i (A_{i,.})_{1,j} = \Big(\sum_{i\leq m} y_i A_{i,.}\Big)_{1,j}.$$

The final statement of the theorem is a trivial consequence of the above relations, and the special form of the standard basis vectors. ■

Definition A.10. Let F be a field. A square matrix $A \in F^{n\times n}$ is called **invertible** if there is $B \in F^{n\times n}$ such that

$$AB = I_n = BA.$$

We say B is an **inverse** of A. Also, we say two matrices $A, C \in F^{n\times n}$ **commute** if

$$AC = CA.$$

Theorem A.11. *Suppose F is a field, and $A, C \in F^{n\times n}$ are invertible matrices. Then*

(i) *The inverse of A is unique, and we denote it by A^{-1}.*

(ii) *A^{-1} and $A^{\intercal}$ are also invertible, and*

$$(A^{-1})^{-1} = A, \qquad (A^{\intercal})^{-1} = (A^{-1})^{\intercal}.$$

(iii) *AC is also invertible, and*

$$(AC)^{-1} = C^{-1}A^{-1}.$$

Proof. **(i)** Suppose that A has two inverses B, E. Then

$$B = BI = B(AE) = (BA)E = IE = E.$$

(ii) Since $A^{-1}A = I = AA^{-1}$, A^{-1} is invertible, and we must have $(A^{-1})^{-1} = A$ due to the uniqueness of the inverse. We also have

$$(A^{-1})^{\intercal}A^{\intercal} = (AA^{-1})^{\intercal} = I^{\intercal} = I.$$

Similarly we have $A^{\intercal}(A^{-1})^{\intercal} = I$. Hence we get the desired due to the uniqueness of the inverse.

(iii) First note that

$$\begin{aligned}(C^{-1}A^{-1})(AC) &= C^{-1}\big(A^{-1}(AC)\big) = C^{-1}\big((A^{-1}A)C\big)\\ &= C^{-1}(IC) = C^{-1}C = I.\end{aligned}$$

Similarly $(AC)(C^{-1}A^{-1}) = I$. Therefore AC is invertible. Now the result follows from the uniqueness of the inverse of a matrix. ■

Remark. As a consequence of the above theorem, we can easily show by induction that if $A_1, \dots, A_k \in F^{n\times n}$ are invertible then $A_1 \cdots A_k$ is also invertible, and

$$(A_1 \cdots A_k)^{-1} = A_k^{-1} \cdots A_1^{-1}.$$

***Example* A.12.** Let $a, b, c, d \in F$. Consider the following 2×2 matrices

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \; B = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Then it is easy to show by direct computation that

$$AB = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} ad - bc & 0 \\ 0 & ad - bc \end{bmatrix} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = BA.$$

Now suppose $ad - bc \neq 0$. Then the above equation implies that A is invertible, and

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

On the other hand, if $ad - bc = 0$ then we have $AB = 0$. Therefore A cannot be invertible, since otherwise we would have $B = IB = A^{-1}AB = A^{-1}0 = 0$. But this implies that $A = 0$, and hence $I = A^{-1}A = A^{-1}0 = 0$, which is a contradiction.

Appendix B

Normed Spaces

B.1 Normed Spaces

Definition B.1. Suppose V is a vector space over $\mathbb{R}$ or $\mathbb{C}$. A **norm** on V is a map

$$\|\ \| : V \to \mathbb{R}$$

that satisfies

(i) $\|\ \|$ is *positive definite*, i.e. for every $v \in V$ we have

$$\|v\| \geq 0, \text{ and } \|v\| = 0 \implies v = 0.$$

(ii) $\|\ \|$ is *homogeneous*, i.e. for every vectors v and scalars c we have

$$\|cv\| = |c|\|v\|.$$

(iii) $\|\ \|$ satisfies the **triangle inequality**, i.e. for every $v, w \in V$ we have

$$\|v + w\| \leq \|v\| + \|w\|.$$

A vector space equipped with a norm is called a **normed space**.

Remark. Note that by homogeneity of norm we always have

$$\begin{gathered}\|0\| = \|0 \cdot 0\| = |0|\|0\| = 0,\\ \| - v\| = \|(-1)v\| = |-1|\|v\| = \|v\|.\end{gathered}$$

It is also easy to show by induction that

$$\|v_1 + \cdots + v_k\| \leq \|v_1\| + \cdots + \|v_k\|.$$

***Example* B.2.** Every inner product space is a normed space with the norm induced by the inner product.

Theorem B.3. *On every normed space, $d(v, w) := \|v - w\|$ is a metric.*

Remark. We always equip a normed space with this metric and its induced topology.

Proof. First note that $d(v, w) = \|v - w\| \geq 0$, and $d(v, v) = \|v - v\| = \|0\| = 0$. Now if $d(v, w) = 0$ then $\|v - w\| = 0$; thus $v - w = 0$. Hence $v = w$. We also have

$$d(v, w) = \|v - w\| = \|(-1)(w - v)\| = |-1|\|w - v\| = d(w, v).$$

Finally, we have

$$\begin{aligned} d(v, u) &= \|v - u\| = \|v - w + w - u\| \\ &\leq \|v - w\| + \|w - u\| = d(v, w) + d(w, u). \end{aligned}$$ ■

Remark. Note that metrics induced by norms are homogeneous and translation invariant, i.e.

$$d(cv, cw) = |c|d(v, w), \qquad d(v + v_0, w + v_0) = d(v, w). \tag{$*$}$$

Also, every norm can be expressed in terms of its metric as $\|v\| = d(v, 0)$.

***Exercise* B.4.** Suppose d is a metric on a (real or complex) vector space V that satisfies $(*)$. Show that $\|v\| = d(v, 0)$ is a norm on V that induces d.

Theorem B.5. *In every normed space $(V, \| \, \|)$ we have*

$$\big|\|v\| - \|w\|\big| \leq \|v - w\|.$$

Hence the norm is Lipschitz continuous.

Proof. We have $\|v\| \leq \|v - w\| + \|w\|$. So $\|v\| - \|w\| \leq \|v - w\|$. By switching v, w we get the desired result.

Note that by the triangle inequality we have $\|v\| \leq \|v - w\| + \|w\|$. Therefore $\|v\| - \|w\| \leq \|v - w\|$. By switching v, w we get

$$\|w\| - \|v\| \leq \|w - v\| = \|v - w\| \implies \|v\| - \|w\| \geq -\|v - w\|.$$

Therefore $\big|\|v\| - \|w\|\big| \leq \|v - w\|$, as desired. ■

Definition B.6. Two norms $\| \, \|_1$, $\| \, \|_2$ on a vector space V are **equivalent** if there exist $c, C > 0$ such that

$$c\|v\|_1 \leq \|v\|_2 \leq C\|v\|_1$$

for all $v \in V$.

Remark. Equivalent norms induce equivalent metrics, hence they induce the same topology.

Every subspace of a normed space is itself a normed space. The product of normed spaces can also be made into a normed space as the next theorem shows.

Theorem B.7. *Suppose $(V_1, \| \|_1), \dots, (V_n, \| \|_n)$ are normed spaces. Then on the* ***product space***

$$\prod_{i=1}^{n} V_i := V_1 \times \cdots \times V_n$$

there are three equivalent norms

$$\begin{aligned} \|v\|_2 &:= \left[\sum_{i=1}^{n} \|v_i\|_i^2 \right]^{\frac{1}{2}}, \\ \|v\|_1 &:= \sum_{i=1}^{n} \|v_i\|_i, \\ \|v\|_\infty &:= \max_{i \le n} \{\|v_i\|_i\}, \end{aligned}$$

where $v = (v_1, \dots, v_n) \in \prod_{i \le n} V_i$.

Remark. These norms induce the corresponding product metrics.

Proof. All of these functions satisfy the first two conditions of a norm obviously. Also it is easy to see that $\| \|_1$ satisfies the triangle inequality. For $\| \|_\infty$ we have

$$\begin{aligned} \|v + w\|_\infty &= \max_{i \le n} \{\|v_i + w_i\|_i\} \le \max_{i \le n} \{\|v_i\|_i + \|w_i\|_i\} \\ &\le \max_{i \le n} \{\|v_i\|_i\} + \max_{i \le n} \{\|w_i\|_i\} = \|v\|_\infty + \|w\|_\infty. \end{aligned}$$

Finally for $\| \|_2$ we have

$$\begin{aligned} \|v + w\|_2 &= \left[\sum_{i=1}^{n} \|v_i + w_i\|_i^2 \right]^{\frac{1}{2}} \le \left[\sum_{i=1}^{n} \big(\|v_i\|_i + \|w_i\|_i \big)^2 \right]^{\frac{1}{2}} \\ &\le \left[\sum_{i=1}^{n} \|v_i\|_i^2 \right]^{\frac{1}{2}} + \left[\sum_{i=1}^{n} \|w_i\|_i^2 \right]^{\frac{1}{2}} = \|v\|_2 + \|w\|_2. \end{aligned}$$

Here we applied the triangle inequality for the standard norm on $\mathbb{R}^n$. Thus it only remains to show that these norms are equivalent. Let $a_i := \|v_i\|_i$. Then

$$\max \{a_i\} \le \big(\sum a_i^2 \big)^{\frac{1}{2}} \le \sum a_i \le n \max \{a_i\} \le n \big(\sum a_i^2 \big)^{\frac{1}{2}} \le n \sum a_i.$$

The second inequality above follows from squaring its both sides. ■

Remark. A necessary and sufficient condition for a norm to be induced by an inner product is that it satisfies the parallelogram law, i.e. for every vectors u, v we have

$$\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2.$$

However, not every norm is induced by an inner product. For example $\| \|_1, \| \|_\infty$ on $\mathbb{R}^n$ are not induced by an inner product, since they do not satisfy the parallelogram law (why?).

Theorem B.8. *The vector addition $+ : V \times V \to V$ and the scalar multiplication $\cdot : \mathbb{R} \times V \to V$ are continuous.*

Proof. Let $\| \|_1$ be the norm on $V \times V$. Then to see the continuity of the addition note that

$$\|v + w - (v_0 + w_0)\| \leq \|v - v_0\| + \|w - w_0\| = \|(v, w) - (v_0, w_0)\|_1.$$

Hence addition is Lipschitz continuous.

Now let $\| \|_1$ be the norm on $\mathbb{R} \times V$. Suppose $|c - c_0| < 1$. Then for $C = \max\{\|v_0\|, |c_0| + 1\}$ we have

$$\begin{aligned} \|cv - c_0v_0\| &= \|cv - cv_0 + cv_0 - c_0v_0\| \leq |c|\|v - v_0\| + |c - c_0|\|v_0\| \\ &\leq C(\|v - v_0\| + |c - c_0|) = C\|(c, v) - (c_0, v_0)\|_1. \end{aligned}$$

So scalar multiplication is continuous. ■

Definition B.9. A **Banach space** is a normed space which is complete as a metric space. A **Hilbert space** is a Banach space whose norm is induced by an inner product.

Remark. Closed subspaces of Banach spaces, and products of finitely many Banach spaces are Banach spaces, as they are complete.

B.2 Bounded Linear Maps

Definition B.10. A linear map $T : V \to W$ between normed spaces is called **bounded** if there exists a constant $C > 0$ such that

$$\|Tv\|_W \leq C\|v\|_V$$

for all $v \in V$.

Remark. It is easy to see that bounded linear maps form a subspace of the space of linear maps. We denote this subspace by $B(V, W)$.

Remark. Note that the composition of bounded maps is bounded.

Theorem B.11. *For a linear map $T : V \to W$ between normed spaces the following are equivalent*

(i) *T is bounded.*

(ii) *T is Lipschitz (hence it is uniformly continuous everywhere).*

(iii) *T is continuous at one point.*

Proof. If T is bounded we have

$$\|Tv - Tw\| = \|T(v - w)\| \leq C\|v - w\|,$$

so T is Lipschitz.

If T is continuous at v_0 then there is $\epsilon > 0$ such that when $\|v - v_0\| < \epsilon$ we have $\|Tv - Tv_0\| < 1$. Now for an arbitrary nonzero v we have

$$\Big\|\Big(\frac{\epsilon v}{2\|v\|} + v_0\Big) - v_0\Big\| = \frac{\epsilon}{2\|v\|}\|v\| < \epsilon.$$

Thus $\|T(\frac{\epsilon v}{2\|v\|})\| = \|T(\frac{\epsilon v}{2\|v\|} + v_0) - Tv_0\| < 1$. Hence $\|Tv\| < \frac{2}{\epsilon}\|v\|$. We can obviously allow v to be 0 in this inequality. Therefore T is bounded. ■

Definition B.12. Let V be a normed space over the field F, where F is either $\mathbb{R}$ or $\mathbb{C}$. We equip F with its standard norm. Then the **dual space** of the normed space V is

$$V^* = B(V, F).$$

Elements of V^* are called (continuous or bounded) **functionals** on V.

Definition B.13. The **operator norm** of a bounded linear map $T : V \to W$ is

$$\|T\| := \sup_{v \neq 0} \frac{\|Tv\|_W}{\|v\|_V}.$$

Remark. It is easy to see that equivalent norms on V, W give rise to equivalent operator norms.

Theorem B.14. *The operator norm makes $B(V, W)$ into a normed space.*

Proof. It is obvious that $\|T\|$ is nonnegative and finite for bounded maps. Suppose $\|T\| = 0$. Then $\|Tv\| = 0$ for all nonzero v, so $T = 0$. We also have

$$\|cT\| = \sup_{v \neq 0} \frac{\|cTv\|}{\|v\|} = \sup_{v \neq 0} \frac{|c|\|Tv\|}{\|v\|} = |c| \sup_{v \neq 0} \frac{\|Tv\|}{\|v\|} = |c|\|T\|.$$

In addition

$$\begin{aligned}\|T+S\| &= \sup_{v\neq 0}\frac{\|Tv+Sv\|}{\|v\|} \leq \sup_{v\neq 0}\frac{\|Tv\|+\|Sv\|}{\|v\|} \\ &\leq \sup_{v\neq 0}\frac{\|Tv\|}{\|v\|}+\sup_{v\neq 0}\frac{\|Sv\|}{\|v\|} = \|T\|+\|S\|.\end{aligned}$$

Here we used the fact that $\sup_{a\in A,\, b\in B}\{a+b\} \leq \sup_{a\in A}\{a\}+\sup_{b\in B}\{b\}$. ■

Theorem B.15. *$B(V,W)$ is a Banach space when W is a Banach space. As a result, V^* is always a Banach space.*

Proof. Let T_n be a Cauchy sequence of linear maps. Then for all $v \in V$ we have

$$\|T_m v - T_n v\| \leq \|T_m - T_n\|\|v\| \underset{m,n\to\infty}{\longrightarrow} 0.$$

Thus $(T_n v)$ is a Cauchy sequence in W. Thus it converges to a unique element $Tv \in W$. Now define $T : V \to W$ by $T(v) = Tv$. We must show that T is linear. We have

$$T(v+cw) = \lim(T_n(v+cw)) = \lim(T_n v + cT_n w) = Tv + cTw.$$

Next we must show that T is bounded. First note that $|\|T_m\| - \|T_n\|| \leq \|T_m - T_n\|$. Hence the sequence of operator norms is also Cauchy and it converges to some $r \in \mathbb{R}$. Now for large n we have $\|T_n\| < r+1$. So $\frac{\|Tv\|}{\|v\|} = \lim \frac{\|T_n v\|}{\|v\|} < r+1$ for all nonzero v. Thus T is bounded.

Finally we need to show that $T_n \to T$. Let n be large enough such that $\|T_m - T_n\| < \frac{\epsilon}{2}$ for $m \geq n$. For each v we can also find $m(v) \geq n$ such that $\|T\frac{v}{\|v\|} - T_{m(v)}\frac{v}{\|v\|}\| < \frac{\epsilon}{2}$. Then

$$\|T-T_n\| = \sup_{v\neq 0}\frac{\|Tv-T_n v\|}{\|v\|} \leq \sup_{v\neq 0}\frac{\|Tv-T_{m(v)}v\|+\|T_{m(v)}v-T_n v\|}{\|v\|} < \epsilon.$$

Therefore $T_n \to T$ as desired. ■

B.3 Finite Dimensional Normed Spaces

Theorem B.16. *Let $(V, \|\ \|)$ be a normed space. Then any linear map $T : \mathbb{R}^n \to V$ is continuous. Furthermore, if T is a bijection it is a homeomorphism.*

Proof. Let $\{e_i\}$ be the standard basis of $\mathbb{R}^n$. Then for $v = \sum v_i e_i$ we have

$$\|Tv\| = \|\sum v_i Te_i\| \leq \sum |v_i|\|Te_i\| \leq (\sum \|Te_i\|)|v|.$$

Hence $\|Tv\| \le C|v|$ for some $C > 0$. Therefore T is continuous. As a result the set

$$A := T(\{v : |v| = 1\})$$

is a compact subset of V, being the continuous image of a compact set. Thus A is closed and its complement is open.

Now suppose T is bijective. Then $0 \in A^c$, since $Tv \neq 0$ when $v \neq 0$. So there is $r > 0$ such that $B_r(0) \subset A^c$. This means that for $\|v\| < r$ we have $|S(v)| \neq 1$, where $S = T^{-1}$. If $|S(v)| > 1$ for some v with $\|v\| < r$, then we have $\|\frac{v}{|S(v)|}\| < r$. But then we get $|S(\frac{v}{|S(v)|})| = \frac{|S(v)|}{|S(v)|} = 1$, which is a contradiction. Therefore

$$\|v\| < r \implies |S(v)| < 1.$$

Now for an arbitrary nonzero v we have $\|\frac{rv}{2\|v\|}\| = \frac{r}{2} < r$. So $|S(\frac{rv}{2\|v\|})| < 1$. Thus $|S(v)| < \frac{2}{r}\|v\|$. We can obviously allow v to be 0 in this inequality. Therefore $S = T^{-1}$ is also continuous. ■

Theorem B.17. *All norms on a finite dimensional space are equivalent.*

Proof. It suffices to show that all norms are equivalent to one particular norm. Let V be a finite dimensional space with some norm $\|\ \|$. Let $\{v_i\}$ be a basis for V, and let $[\ ] : V \to \mathbb{R}^n$ be the representation of vectors in this basis. Define $|v| := |[v]|$. It is easy to see that this is a norm. As $[\ ]^{-1}$ is a linear isomorphism, it is a homeomorphism. Therefore $[\ ]^{-1}, [\ ]$ are continuous, hence they are bounded. Thus there are $c, C > 0$ such that $|v| = |[v]| \le c\|v\|$, and $\|v\| = \|[[v]]^{-1}\| \le C|[v]| = C|v|$. ■

Theorem B.18. *All finite dimensional normed spaces are Banach spaces.*

Proof. Let $(V, \|\ \|)$ be a finite dimensional space. Let $[\ ] : V \to \mathbb{R}^n$ be the representation map in some basis. We know that $[\ ]^{-1}$ is a homeomorphism being a linear bijection. Let (v_i) be a Cauchy sequence in V. Then

$$|[v_i] - [v_j]| \le c\|v_i - v_j\| \underset{i.j\to\infty}{\longrightarrow} 0.$$

Hence $([v_i])$ is Cauchy in $\mathbb{R}^n$. So $[v_i] \to x$ for some $x \in \mathbb{R}^n$. Let $v \in V$ be the vector that satisfies $[v] = x$. Then we have $\|v_i - v\| \le C|[v_i] - x| \to 0$. ■

Theorem B.19. *A finite dimensional subspace of a normed space is closed.*

Proof. If a sequence in the subspace converges to some point in the space, then that sequence is Cauchy. Since the finite dimensional subspace is complete by the above theorem, the sequence must converge to some point in the subspace. Now the result follows from the uniqueness of limit. ■