ك

Pure 15
Topology

Chapter 10

Separation Axioms

INTRODUCTION

Many properties of a topological space X depend upon the distribution of the open sets in the space. Roughly speaking, a space is more likely to be separable, or first or second countable, if there are "few" open sets; on the other hand, an arbitrary function on X to some topological space is more likely to be continuous, or a sequence to have a unique limit, if the space has "many" open sets.

The *separation axioms* of Alexandroff and Hopf, discussed in this chapter, postulate the existence of "enough" open sets.

T_1-SPACES

A topological space X is a *T_1-space* iff it satisfies the following axiom:

[T_1] Given any pair of distinct points $a, b \in X$, each belongs to an open set which does not contain the other.

In other words, there exist open sets G and H such that

$$a \in G,\ b \notin G \quad \text{and} \quad b \in H,\ a \notin H$$

The open sets G and H are not necessarily disjoint.

Our next theorem gives a very simple characterization of T_1-spaces.

Theorem 10.1: A topological space X is a T_1-space if and only if every singleton subset $\{p\}$ of X is closed.

Since finite unions of closed sets are closed, the above theorem implies:

Corollary 10.2: $(X, \mathcal{T})$ is a T_1-space if and only if $\mathcal{T}$ contains the cofinite topology on X.

Example 1.1: Every metric space X is a T_1-space, since we proved that finite subsets of X are closed.

Example 1.2: Consider the topology $\mathcal{T} = \{X, \emptyset, \{a\}\}$ on the set $X = \{a, b\}$. Observe that X is the only open set containing b, but it also contains a. Hence $(X, \mathcal{T})$ does not satisfy **[T_1]**, i.e. $(X, \mathcal{T})$ is not a T_1-space. Note that the singleton set $\{a\}$ is not closed since its complement $\{a\}^c = \{b\}$ is not open.

Example 1.3: The cofinite topology on X is the coarsest topology on X for which $(X, \mathcal{T})$ is a T_1-space (Corollary 10.2). Hence the cofinite topology is also called the *T_1-topology*.

HAUSDORFF SPACES

A topological space X is a *Hausdorff space* or *T_2-space* iff it satisfies the following axiom:

[T_2] Each pair of distinct points $a, b \in X$ belong respectively to disjoint open sets.

In other words, there exist open sets G and H such that

$$a \in G, \quad b \in H \quad \text{and} \quad G \cap H = \emptyset$$

Observe that a Hausdorff space is always a T_1-space.

Example 2.1: We show that every metric space X is Hausdorff.

Let $a, b \in X$ be distinct points; hence by **[M$_4$]** $d(a,b) = \epsilon > 0$. Consider the open spheres $G = S(a, \frac{1}{3}\epsilon)$ and $H = S(b, \frac{1}{3}\epsilon)$, centered at a and b respectively. We claim that G and H are disjoint. For if $p \in G \cap H$, then $d(a,p) < \frac{1}{3}\epsilon$ and $d(p,b) < \frac{1}{3}\epsilon$; hence by the Triangle Inequality,

$$d(a,b) \leq d(a,p) + d(p,b) < \tfrac{1}{3}\epsilon + \tfrac{1}{3}\epsilon = \tfrac{2}{3}\epsilon$$

But this contradicts the fact that $d(a,b) = \epsilon$. Hence G and H are disjoint, i.e. a and b belong respectively to the disjoint open spheres G and H. Accordingly, X is Hausdorff.

We formally state the result in the preceding example, namely:

Theorem 10.3: Every metric space is a Hausdorff space.

Example 2.2: Let $\mathcal{T}$ be the cofinite topology, i.e. T_1-topology, on the real line $\mathbf{R}$. We show that $(\mathbf{R}, \mathcal{T})$ is not Hausdorff. Let G and H be any non-empty $\mathcal{T}$-open sets. Now G and H are infinite since they are complements of finite sets. If $G \cap H = \emptyset$, then G, an infinite set, would be contained in the finite complement of H; hence G and H are not disjoint. Accordingly, no pair of distinct points in $\mathbf{R}$ belongs, respectively, to disjoint $\mathcal{T}$-open sets. Thus T_1-spaces need not be Hausdorff.

As noted previously, a sequence $\langle a_1, a_2, \ldots \rangle$ of points in a topological space X could, in general, converge to more than one point in X. This cannot happen if X is Hausdorff:

Theorem 10.4: If X is a Hausdorff space, then every convergent sequence in X has a unique limit.

The converse of the above theorem is not true unless we add additional conditions.

Theorem 10.5: Let X be first countable. Then X is Hausdorff if and only if every convergent sequence has a unique limit.

Remark: The notion of a sequence has been generalized to that of a *net* (Moore-Smith sequence) and to that of a *filter* with the following results:

Theorem 10.4A: X is a Hausdorff space if and only if every convergent net in X has a unique limit.

Theorem 10.4B: X is a Hausdorff space if and only if every convergent filter in X has a unique limit.

The definitions of net and filter and the proofs of the above theorems lie beyond the scope of this text.

REGULAR SPACES

A topological space X is *regular* iff it satisfies the following axiom:

[R] If F is a closed subset of X and $p \in X$ does not belong to F, then there exist disjoint open sets G and H such that $F \subset G$ and $p \in H$.

A regular space need not be a T_1-space, as seen by the next example.

Example 3.1: Consider the topology $\mathcal{T} = \{X, \emptyset, \{a\}, \{b,c\}\}$ on the set $X = \{a,b,c\}$. Observe that the closed subsets of X are also $X, \emptyset, \{a\}$ and $\{b,c\}$ and that $(X, \mathcal{T})$ does satisfy **[R]**. On the other hand, $(X, \mathcal{T})$ is not a T_1-space since there are finite sets, e.g. $\{b\}$, which are not closed.

A regular space X which also satisfies the separation axiom **[$\mathbf{T_1}$]**, i.e. a regular T_1-space, is called a *T_3-space.*

Example 3.2: Let X be a T_3-space. Then X is also a Hausdorff space, i.e. a T_2-space. For let $a, b \in X$ be distinct points. Since X is a T_1-space, $\{a\}$ is a closed set; and since a and b are distinct, $b \notin \{a\}$. Accordingly, by **[R]**, there exist disjoint open sets G and H such that $\{a\} \subset G$ and $b \in H$. Hence a and b belong respectively to disjoint open sets G and H.

NORMAL SPACES

A topological space X is normal iff X satisfies the following axiom:

[N] If F_1 and F_2 are disjoint closed subsets of X, then there exist disjoint open sets G and H such that $F_1 \subset G$ and $F_2 \subset H$.

A normal space can also be characterized as follows:

Theorem 10.6: A topological space X is normal if and only if for every closed set F and open set H containing F there exists an open set G such that $F \subset G \subset \bar{G} \subset H$.

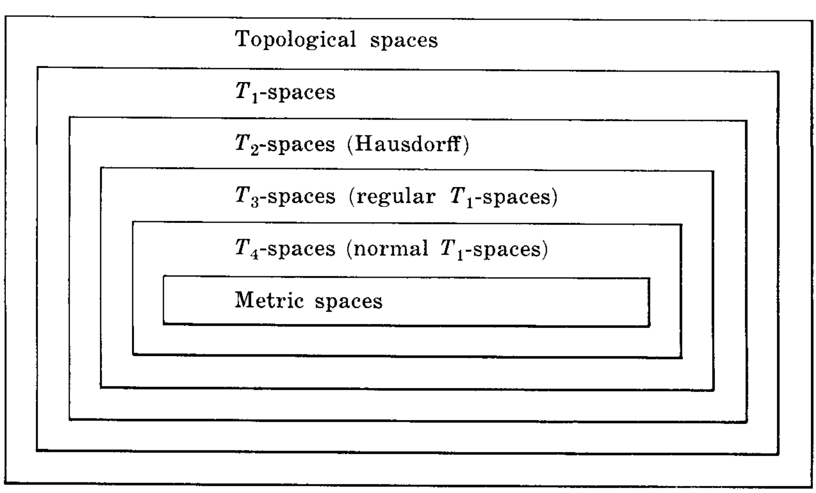
Example 4.1: Every metric space is normal by virtue of the Separation Theorem 8.8.

Example 4.2: Consider the topology $\mathcal{T} = \{X, \emptyset, \{a\}, \{b\}, \{a,b\}\}$ on the set $X = \{a,b,c\}$. Observe that the closed sets are X, $\emptyset$, $\{b,c\}$, $\{a,c\}$ and $\{c\}$. If F_1 and F_2 are disjoint closed subsets of $(X, \mathcal{T})$, then one of them, say F_1, must be the empty set $\emptyset$. Hence $\emptyset$ and X are disjoint open sets and $F_1 \subset \emptyset$ and $F_2 \subset X$. In other words, $(X, \mathcal{T})$ is a normal space. On the other hand, $(X, \mathcal{T})$ is not a T_1-space since the singleton set $\{a\}$ is not closed. Furthermore, $(X, \mathcal{T})$ is not a regular space since $a \notin \{c\}$, and the only open superset of the closed set $\{c\}$ is X which also contains a.

A normal space X which also satisfies the separation axiom **[$\mathbf{T_1}$]**, i.e. a normal T_1-space, is called a *T_4-space.*

Example 4.3: Let X be a T_4-space. Then X is also a regular T_1-space, i.e. T_3-space. For suppose F is a closed subset of X and $p \in X$ does not belong to F. By **[$\mathbf{T_1}$]**, $\{p\}$ is closed; and since F and $\{p\}$ are disjoint, by **[N]**, there exist disjoint open sets G and H such that $F \subset G$ and $p \in \{p\} \subset H$.

Now a metric space is both a normal space and a T_1-space, i.e. a T_4-space. The following diagram illustrates the relationship between the spaces discussed in this chapter.



URYSOHN'S LEMMA AND METRIZATION THEOREM

Next comes the classical result of Urysohn.

Theorem (Urysohn's Lemma) 10.7: Let F_1 and F_2 be disjoint closed subsets of a normal space X. Then there exists a continuous function $f: X \to [0,1]$ such that

$$f[F_1] = \{0\} \quad \text{and} \quad f[F_2] = \{1\}$$

One important consequence of Urysohn's Lemma gives a partial solution to the metrization problem as discussed in Chapter 8. Namely,

Urysohn's Metrization Theorem 10.8: Every second countable normal T_1-space is metrizable.

In fact, we will prove that every second countable normal T_1-space is homeomorphic to a subset of the Hilbert cube in $\mathbf{R}^\infty$.

FUNCTIONS THAT SEPARATE POINTS

Let $\mathcal{A} = \{f_i : i \in I\}$ be a class of functions from a set X into a set Y. The class $\mathcal{A}$ of functions is said to *separate points* iff for any pair of distinct points $a, b \in X$ there exists a function f in $\mathcal{A}$ such that $f(a) \neq f(b)$.

Example 5.1: Consider the class of real-valued functions

$$\mathcal{A} = \{f_1(x) = \sin x,\ f_2(x) = \sin 2x,\ f_3(x) = \sin 3x,\ \ldots\}$$

defined on $\mathbf{R}$. Observe that for every function $f_n \in \mathcal{A}$, $f_n(0) = f_n(\pi) = 0$. Hence the class $\mathcal{A}$ does not separate points.

Example 5.2: Let $C(X, \mathbf{R})$ denote the class of all real-valued continuous functions on a topological space X. We show that if $C(X, \mathbf{R})$ separates points, then X is a Hausdorff space. Let $a, b \in X$ be distinct points. By hypothesis, there exists a continuous function $f: X \to \mathbf{R}$ such that $f(a) \neq f(b)$. But $\mathbf{R}$ is a Hausdorff space; hence there exist disjoint open subsets G and H of $\mathbf{R}$ containing $f(a)$ and $f(b)$ respectively. Accordingly, the inverses $f^{-1}[G]$ and $f^{-1}[H]$ are disjoint, open and contain a and b respectively. In other words, X is a Hausdorff space.

We formally state the result in the preceding example.

Proposition 10.9: If the class $C(X, \mathbf{R})$ of all real-valued continuous functions on a topological space X separates points, then X is a Hausdorff space.

COMPLETELY REGULAR SPACES

A topological space X is *completely regular* iff it satisfies the following axiom:

[CR] If F is a closed subset of X and $p \in X$ does not belong to F, then there exists a continuous function $f: X \to [0,1]$ such that $f(p) = 0$ and $f[F] = \{1\}$.

We show later that

Proposition 10.10: A completely regular space is also regular.

A completely regular space X which also satisfies **[T₁]**, i.e. a completely regular T_1-space, is called a *Tychonoff space*. By virtue of Urysohn's Lemma, a T_4-space is a Tychonoff space and, by Proposition 10.10, a Tychonoff space is a T_3-space. Hence a Tychonoff space, i.e. a completely regular T_1-space, is sometimes called a $T_{3\frac{1}{2}}$-space.

One important property of Tychonoff spaces is the following:

Theorem 10.11: The class $C(X, \mathbf{R})$ of all real-valued continuous functions on a completely regular T_1-space X separates points.

Solved Problems

T_1-SPACES

1. Prove Theorem 10.1: A topological space X is a T_1-space if and only if every singleton subset of X is closed.

Solution:

Suppose X is a T_1-space and $p \in X$. We show that $\{p\}^c$ is open. Let $x \in \{p\}^c$. Then $x \neq p$, and so by **[T$_1$]**

$$\exists \text{ an open set } G_x \quad \text{such that} \quad x \in G_x \text{ but } p \notin G_x$$

Hence $x \in G_x \subset \{p\}^c$, and hence $\{p\}^c = \bigcup\{G_x : x \in \{p\}^c\}$. Accordingly $\{p\}^c$, a union of open sets, is open and $\{p\}$ is closed.

Conversely, suppose $\{p\}$ is closed for every $p \in X$. Let $a, b \in X$ with $a \neq b$. Now $a \neq b \Rightarrow b \in \{a\}^c$; hence $\{a\}^c$ is an open set containing b but not containing a. Similarly $\{b\}^c$ is an open set containing a but not containing b. Accordingly, X is a T_1-space.

2. Show that the property of being a T_1-space is hereditary, i.e. every subspace of a T_1-space is also a T_1-space.

Solution:

Let $(X, \mathcal{T})$ be a T_1-space and let $(Y, \mathcal{T}_Y)$ be a subspace of $(X, \mathcal{T})$. We show that every singleton subset $\{p\}$ of Y is a $\mathcal{T}_Y$-closed set or, equivalently, that $Y \setminus \{p\}$ is $\mathcal{T}_Y$-open. Since $(X, \mathcal{T})$ is a T_1-space, $X \setminus \{p\}$ is $\mathcal{T}$-open. But

$$p \in Y \subset X \quad \Rightarrow \quad Y \cap (X \setminus \{p\}) = Y \setminus \{p\}$$

Hence by definition of subspace, $Y \setminus \{p\}$ is a $\mathcal{T}_Y$-open set. Thus $(Y, \mathcal{T}_Y)$ is also a T_1-space.

3. Show that a finite subset of a T_1-space X has no accumulation points.

Solution:

Suppose $A \subset X$ has n elements, say $A = \{a_1, \ldots, a_n\}$. Since A is finite it is closed and therefore contains all of its accumulation points. But $\{a_2, \ldots, a_n\}$ is also finite and hence closed. Accordingly, the complement $\{a_2, \ldots, a_n\}^c$ of $\{a_2, \ldots, a_n\}$ is open, contains a_1, and contains no points of A different from a_1. Hence a_1 is not an accumulation point of A. Similarly, no other point of A is an accumulation point of A and so A has no accumulation points.

4. Show that every finite T_1-space X is a discrete space.

Solution:

Every subset of X is finite and therefore closed. Hence every subset of X is also open, i.e. X is a discrete space.

5. Prove: Let X be a T_1-space. Then the following are equivalent:

(i) $p \in X$ is an accumulation point of A.

(ii) Every open set containing p contains an infinite number of points of A.

Solution:

By definition of an accumulation point of a set, (ii) $\Rightarrow$ (i); hence we only need to prove that (i) $\Rightarrow$ (ii).

Suppose G is an open set containing p and only containing a finite number of points of A different from p; say

$$B = (G \setminus \{p\}) \cap A = \{a_1, a_2, \ldots, a_n\}$$

Now B, a finite subset of a T_1-space, is closed and so B^c is open. Set $H = G \cap B^c$. Then H is open, $p \in H$ and H contains no points of A different from p. Hence p is not an accumulation point of A and so (i) $\Rightarrow$ (ii).

6. Let X be a T_1-space and let $\mathcal{B}_p$ be a local base at $p \in X$. Show that if $q \in X$ is distinct from p, then some member of $\mathcal{B}_p$ does not contain q.

Solution:

Since $p \neq q$ and X satisfies $[\mathbf{T_1}]$, $\exists$ an open set $G \subset X$ containing p but not containing q. Now $\mathcal{B}_p$ is a local base at p, so G is a superset of some $B \in \mathcal{B}_p$ and B also does not contain q.

7. Let X be a T_1-space which satisfies the first axiom of countability. Show that if $p \in X$ is an accumulation point of $A \subset X$, then there exists a sequence of distinct terms in A converging to p.

Solution:

Let $\mathcal{B} = \{B_n\}$ be a nested local base at p. Set $B_{i_1} = B_1$. Since p is a limit point of A, B_{i_1} contains a point $a_1 \in A$ different from p. By the preceding problem,

$$\exists\ B_{i_2} \in \mathcal{B} \quad \text{such that} \quad a_1 \notin B_{i_2}$$

Similarly B_{i_2} contains a point $a_2 \in A$ different from p and, since $a_1 \notin B_{i_2}$, different from a_1. Again by the preceding problem,

$$\exists\ B_{i_3} \in \mathcal{B} \quad \text{such that} \quad a_2 \notin B_{i_3}$$

Furthermore, $$a_2 \in B_{i_2},\ a_2 \notin B_{i_3} \quad \Rightarrow \quad B_{i_2} \supset B_{i_3}$$

Continuing in this manner we obtain a subsequence $\{B_{i_1}, B_{i_2}, \ldots\}$ of $\mathcal{B}$ and a sequence $\langle a_1, a_2, \ldots\rangle$ of distinct terms in A with $a_1 \in B_{i_1}$, $a_2 \in B_{i_2}$, ... But $\{B_{i_n}\}$ is also a nested local base at p; hence $\langle a_n \rangle$ converges to p.

HAUSDORFF SPACES

8. Show that the property of being a Hausdorff space is hereditary, i.e. every subspace of a Hausdorff space is also Hausdorff.

Solution:

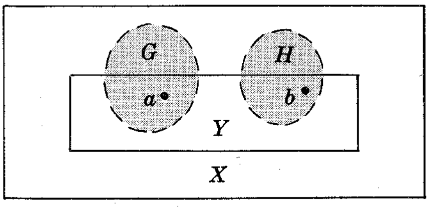
Let $(X, \mathcal{T})$ be a Hausdorff space and let $(Y, \mathcal{T}_Y)$ be a subspace of $(X, \mathcal{T})$. Furthermore, let $a, b \in Y \subset X$ with $a \neq b$. By hypothesis, $(X, \mathcal{T})$ is Hausdorff; hence

$$\exists\ G, H \in \mathcal{T} \quad \text{such that} \quad a \in G,\ b \in H \text{ and } G \cap H = \emptyset$$

By definition of a subspace, $Y \cap G$ and $Y \cap H$ are $\mathcal{T}_y$-open sets. Furthermore,

$$a \in G,\ a \in Y \quad \Rightarrow \quad a \in Y \cap G$$
$$b \in H,\ b \in Y \quad \Rightarrow \quad b \in Y \cap H$$
$$G \cap H = \emptyset \quad \Rightarrow \quad (Y \cap G) \cap (Y \cap H) = Y \cap (G \cap H) = Y \cap \emptyset = \emptyset$$

(as indicated in the diagram below). Accordingly $(Y, \mathcal{T}_y)$ is also a Hausdorff space.



9. Let $\mathcal{T}$ be the topology on the real line $\mathbf{R}$ generated by the open-closed intervals $(a, b]$. Show that $(\mathbf{R}, \mathcal{T})$ is Hausdorff.

Solution:

Let $a, b \in \mathbf{R}$ with $a \neq b$, say $a < b$. Choose $G = (a-1, a]$ and $H = (a, b]$. Then

$$G, H \in \mathcal{T}, \quad a \in G, \quad b \in H \quad \text{and} \quad G \cap H = \emptyset$$

Hence $(X, \mathcal{T})$ is Hausdorff.

10. Prove Theorem 10.4: Let X be a Hausdorff space. Then every convergent sequence in X has a unique limit.

Solution:

Suppose $\langle a_1, a_2, \ldots\rangle$ converges to a and b, and suppose $a \neq b$. Since X is Hausdorff, $\exists$ open sets G and H such that

$$a \in G, \quad b \in H \quad \text{and} \quad G \cap H = \emptyset$$

By hypothesis, $\langle a_n\rangle$ converges to a; hence

$$\exists\ n_0 \in \mathbf{N} \quad \text{such that} \quad n > n_0 \quad \text{implies} \quad a_n \in G$$

i.e. G contains all except a finite number of the terms of the sequence. But G and H are disjoint; hence H can only contain those terms of the sequence which do not belong to G and there are only a finite number of these. Accordingly, $\langle a_n\rangle$ cannot converge to b. But this violates the hypothesis; hence $a = b$.

11. Prove Theorem 10.5: Let X be a first countable space. Then the following are equivalent: (i) X is Hausdorff. (ii) Every convergent sequence has a unique limit.

Solution:

By the preceding problem, (i) $\Rightarrow$ (ii); hence we need only show that (ii) $\Rightarrow$ (i). Suppose X is not Hausdorff. Then $\exists\ a, b \in X$, $a \neq b$, with the property that every open set containing a has a non-empty intersection with every open set containing b.

Now let $\{G_n\}$ and $\{H_n\}$ be nested local bases at a and b respectively. Then $G_n \cap H_n \neq \emptyset$ for every $n \in \mathbf{N}$, and so

$$\exists\ \langle a_1, a_2, \ldots\rangle \quad \text{such that} \quad a_1 \in G_1 \cap H_1,\ a_2 \in G_2 \cap H_2,\ \ldots$$

Accordingly, $\langle a_n\rangle$ converges to both a and b. In other words, (ii) $\Rightarrow$ (i).

NORMAL SPACES AND URYSOHN'S LEMMA

12. Prove Theorem 10.6: Let X be a topological space. Then the following conditions are equivalent: (i) X is normal. (ii) If H is an open superset of a closed set F, then there exists an open set G such that $F \subset G \subset \bar{G} \subset H$.

Solution:

(i) $\Rightarrow$ (ii). Let $F \subset H$, with F closed and H open. Then H^c is closed, and $F \cap H^c = \emptyset$. But X is normal; hence

$$\exists \text{ open sets } G, G^* \quad \text{such that} \quad F \subset G,\ H^c \subset G^* \quad \text{and} \quad G \cap G^* = \emptyset$$

But $$G \cap G^* = \emptyset \;\Rightarrow\; G \subset G^{*c} \quad \text{and} \quad H^c \subset G^* \;\Rightarrow\; G^{*c} \subset H$$

Furthermore, G^{*c} is closed; hence $F \subset G \subset \bar{G} \subset G^{*c} \subset H$.

(ii) $\Rightarrow$ (i). Let F_1 and F_2 be disjoint closed sets. Then $F_1 \subset F_2^c$, and F_2^c is open. By (ii),

$$\exists \text{ an open set } G \quad \text{such that} \quad F_1 \subset G \subset \bar{G} \subset F_2^c$$

But $$\bar{G} \subset F_2^c \;\Rightarrow\; F_2 \subset \bar{G}^c \quad \text{and} \quad G \subset \bar{G} \;\Rightarrow\; G \cap \bar{G}^c = \emptyset$$

Furthermore, $\bar{G}^c$ is open. Thus $F_1 \subset G$ and $F_2 \subset \bar{G}^c$ with $G, \bar{G}^c$ disjoint open sets; hence X is normal.

13. Let $\mathcal{B}$ be a base for a normal T_1-space X. Show that for each $G_i \in \mathcal{B}$ and any point $p \in G_i$, there exists a member $G_j \in \mathcal{B}$ such that $p \in \bar{G}_j \subset G_i$.

Solution:

Since X is a T_1-space, $\{p\}$ is closed; hence G_i is an open superset of the closed set $\{p\}$. By Theorem 10.6,

$$\exists \text{ an open set } G \quad \text{such that} \quad \{p\} \subset G \subset \bar{G} \subset G_i$$

Since $p \in G$, there is a member G_j of the base $\mathcal{B}$ such that $p \in G_j \subset G$; so $p \in \bar{G}_j \subset \bar{G}$. But $\bar{G} \subset G_i$; hence $p \in \bar{G}_j \subset G_i$.

14. Let D be the set of dyadic fractions (fractions whose denominators are powers of 2) in the unit interval $[0, 1]$, i.e.,

$$D = \{\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{3}{4}, \tfrac{1}{8}, \tfrac{3}{8}, \tfrac{5}{8}, \tfrac{7}{8}, \tfrac{1}{16}, \ldots, \tfrac{15}{16}, \ldots\}$$

Show that D is dense in $[0, 1]$.

Solution:

To show that $\bar{D} = [0, 1]$, it is sufficient to show that any open interval $(a - \delta,\ a + \delta)$ centered at any point $a \in [0, 1]$ contains a point of D. Observe that $\lim_{n \to \infty} \frac{1}{2^n} = 0$; hence there exists a power $q = 2^{n_0}$ such that $0 < 1/q < \delta$. Consider the intervals

$$\left[0, \frac{1}{q}\right], \quad \left[\frac{1}{q}, \frac{2}{q}\right], \quad \left[\frac{2}{q}, \frac{3}{q}\right], \quad \ldots, \quad \left[\frac{q-2}{q}, \frac{q-1}{q}\right], \quad \left[\frac{q-1}{q}, 1\right]$$

Since $[0, 1]$ is the union of the above intervals, one of them, say $\left[\frac{m}{q}, \frac{m+1}{q}\right]$ contains a, i.e. $\frac{m}{q} \le a \le \frac{m+1}{q}$. But $\frac{1}{q} < \delta$; hence

$$a - \delta < \frac{m}{q} \le a < a + \delta$$

In other words, the open interval $(a - \delta,\ a + \delta)$ contains the point m/q which belongs to D. Thus D is dense in $[0, 1]$.

15. Prove Theorem (Urysohn's Lemma) 10.7: Let F_1 and F_2 be disjoint closed subsets of a normal space X. Then there exists a continuous function $f: X \to [0, 1]$ such that $f[F_1] = \{0\}$ and $f[F_2] = \{1\}$.

Solution:

By hypothesis, $F_1 \cap F_2 = \emptyset$; hence $F_1 \subset F_2^c$. In particular, since F_2 is a closed set, F_2^c is an *open* superset of the closed set F_1. By Theorem 10.4, there exists an open set $G_{1/2}$ such that

$$F_1 \subset G_{1/2} \subset \bar{G}_{1/2} \subset F_2^c$$

Observe that $G_{1/2}$ is an open superset of the closed set F_1, and F_2^c is an open superset of the closed set $\bar{G}_{1/2}$. Hence, by Theorem 10.4, there exist open sets $G_{1/4}$ and $G_{3/4}$ such that

$$F_1 \subset G_{1/4} \subset \bar{G}_{1/4} \subset G_{1/2} \subset \bar{G}_{1/2} \subset G_{3/4} \subset \bar{G}_{3/4} \subset F_2^c$$

We continue in this manner and obtain for each $t \in D$, where D is the set of dyadic fractions in $[0, 1]$, an open set G_t with the property that if $t_1, t_2 \in D$ and $t_1 < t_2$ then $\bar{G}_{t_1} \subset G_{t_2}$.

Define the function f on X as follows:

$$f(x) = \begin{cases} \inf\{t : x \in G_t\} & \text{if } x \notin F_2 \\ 1 & \text{if } x \in F_2 \end{cases}$$

Observe that, for every $x \in X$, $0 \le f(x) \le 1$, i.e. f maps X into $[0, 1]$. Observe also that $F_1 \subset G_t$ for all $t \in D$; hence $f[F_1] = \{0\}$. Moreover, by definition, $f[F_2] = \{1\}$. Consequently, the only thing left for us to prove is that f is continuous.

Now f is continuous if the inverses of the sets $[0, a)$ and $(b, 1]$ are open subsets of X (see Problem 7, Chapter 7). We claim that

$$f^{-1}[[0, a)] = \bigcup\{G_t : t < a\} \tag{1}$$

$$f^{-1}[(b, 1]] = \bigcup\{\bar{G}_t^c : t > b\} \tag{2}$$

Then each is the union of open sets and is therefore open.

We first prove (*1*). Let $x \in f^{-1}[[0,a)]$. Then $f(x) \in [0,a)$, i.e. $0 \leq f(x) < a$. Since D is dense in $[0,1]$, there exists $t_x \in D$ such that $f(x) < t_x < a$. In other words,

$$f(x) \quad = \quad \inf\{t : x \in G_t\} \quad < \quad t_x \quad < \quad a$$

Accordingly $x \in G_{t_x}$ where $t_x < a$. Hence $x \in \bigcup\{G_t : t < a\}$. We have just shown that every element in $f^{-1}[[0,a)]$ also belongs to $\bigcup\{G_t : t < a\}$, i.e.,

$$f^{-1}[[0,a)] \quad \subset \quad \bigcup\{G_t : t < a\}$$

On the other hand, suppose $y \in \bigcup\{G_t : t < a\}$. Then $\exists\ t_y \in D$ such that $t_y < a$ and $y \in G_{t_y}$. Therefore

$$f(y) \quad = \quad \inf\{t : y \in G_t\} \quad \leq \quad t_y \quad < \quad a$$

Hence y also belongs to $f^{-1}[[0,a)]$. In other words,

$$\bigcup\{G_t : t < a\} \quad \subset \quad f^{-1}[[0,a)]$$

The above two results imply (*1*).

We now prove (*2*). Let $x \in f^{-1}[(b,1]]$. Then $f(x) \in (b,1]$, i.e. $b < f(x) \leq 1$. Since D is dense in $[0,1]$, there exist $t_1, t_2 \in D$ such that $b < t_1 < t_2 < f(x)$. In other words,

$$f(x) \quad = \quad \inf\{t : x \in G_t\} \quad > \quad t_2$$

Hence $x \notin G_{t_2}$. Observe that $t_1 < t_2$ implies $\bar{G}_{t_1} \subset G_{t_2}$. Hence x does not belong to $\bar{G}_{t_1}$ either. Accordingly, $x \in \bar{G}^c_{t_1}$ where $t_1 > b$; hence $x \in \bigcup\{\bar{G}^c_t : t > b\}$. Consequently,

$$f^{-1}[(b,1]] \quad \subset \quad \bigcup\{\bar{G}^c_t : t > b\}$$

On the other hand, let $y \in \bigcup\{\bar{G}^c_t : t > b\}$. Then there exists $t_y \in D$ such that $t_y > b$ and $y \in \bar{G}^c_{t_y}$; hence y does not belong to $\bar{G}_{t_y}$. But $t < t_y$ implies $G_t \subset G_{t_y} \subset \bar{G}_{t_y}$; hence $y \notin G_t$ for every t less than t_y. Consequently,

$$f(y) \quad = \quad \inf\{t : y \in G_t\} \quad \geq \quad t_y \quad > \quad b$$

Hence $y \in f^{-1}((b,1])$. In other words,

$$\bigcup\{\bar{G}^c_t : t > b\} \quad \subset \quad f^{-1}[(b,1]]$$

The above two results imply (*2*). Hence f is continuous and Urysohn's Lemma is proven.

16. Prove Urysohn's Metrization Theorem 10.8: Every second countable normal T_1-space X is metrizable. (In fact, X is homeomorphic to a subset of the Hilbert cube $\mathbf{I}$ of $\mathbf{R}^\infty$.)

Solution:

If X is finite, then X is a discrete space and hence X is homeomorphic to any subset of H with an equivalent number of points. If X is infinite, then X contains a denumerable base $\mathcal{B} = \{G_1, G_2, G_3, \ldots\}$ where none of the members of $\mathcal{B}$ is X itself.

By a previous problem, for each G_i in $\mathcal{B}$ there exists some G_j in $\mathcal{B}$ such that $\bar{G}_j \subset G_i$. The class of *all* such pairs $\langle G_j, G_i \rangle$, where $\bar{G}_j \subset G_i$, is denumerable; hence we can denote them by $P_1, P_2, \ldots$ where $P_n = \langle G_{j_n}, G_{i_n} \rangle$. Observe that $\bar{G}_{j_n} \subset G_{i_n}$ implies that $\bar{G}_{j_n}$ and $G^c_{i_n}$ are disjoint closed subsets of X. Hence by Urysohn's Lemma there exists a function $f_n : X \to [0,1]$ such that $f_n[\bar{G}_{j_n}] = \{0\}$ and $f_n[G^c_{i_n}] = \{1\}$.

Now define a function $f : X \to \mathbf{I}$ as follows:

$$f(x) \quad = \quad \left\langle \frac{f_1(x)}{2}, \frac{f_2(x)}{2^2}, \frac{f_3(x)}{2^3}, \ldots \right\rangle$$

Observe that, for all $n \in N$, $0 \leq f_n(x) \leq 1$ implies $\left|\frac{f_n(x)}{2^n}\right| \leq \frac{1}{n}$; hence $f(x)$ is a point in the Hilbert cube $\mathbf{I}$. (Recall that $\mathbf{I} = \{\langle a_n \rangle : a_n \in \mathbf{R},\ n \in \mathbf{N},\ 0 \leq a_n \leq 1/n\}$, see Page 129.)

We now show that f is one-to-one. Let x and y be distinct points in X. Since X is a T_1-space, there exists a member G_i of the base $\mathcal{B}$ such that $x \in G_i$ but $y \notin G_i$. By a previous problem, there exists a pair $P_m = \langle G_j, G_i \rangle$ such that $x \in \bar{G}_j \subset G_i$. By definition, $f_m(x) = 0$ since $x \in \bar{G}_j$, and $f_m(y) = 1$ since $y \notin G_i$, i.e. $y \in G^c_i$. Hence $f(x) \neq f(y)$ since they differ in the mth coordinate. Thus f is one-to-one.

We now prove that f is continuous. Let $\epsilon > 0$. Observe that f is continuous at $p \in X$ if there exists an open neighborhood G of p such that $x \in G$ implies $||f(x) - f(p)|| < \epsilon$ or, equivalently, $||f(x) - f(p)||^2 < \epsilon^2$. Recall that

$$||f(x) - f(p)||^2 = \sum_{n=1}^{\infty} \frac{|f_n(x) - f_n(p)|^2}{2^{2n}}$$

Furthermore, since the values of f_n lie in $[0,1]$, $(|f_n(x) - f_n(p)|^2)/2^{2n} \leq 1/2^{2n}$. Note that $\sum_n 1/2^{2n}$ converges; hence there exists an $n_0 = n_0(\epsilon)$, which is independent of x and p, such that

$$||f(x) - f(p)||^2 = \sum_{n=1}^{n_0} \frac{|f_n(x) - f_n(p)|^2}{2^{2n}} + \frac{\epsilon^2}{2}$$

Now each function $f_n : X \to [0,1]$ is continuous; hence there exists an open neighborhood G_n of p such that $x \in G_n$ implies $|f_n(x) - f_n(p)|^2 < \epsilon^2/2n_0$. Let $G = G_1 \cap \cdots \cap G_{n_0}$. Since G is a finite intersection of open neighborhoods of p, G is also an open neighborhood of p. Furthermore, if $x \in G$ then

$$||f(x) - f(p)||^2 = \sum_{n=1}^{\infty} \frac{|f_n(x) - f_n(p)|^2}{2^{2n}} < n_0\left(\frac{\epsilon^2}{2n_0}\right) + \frac{\epsilon^2}{2} = \epsilon^2$$

Hence f is continuous.

Now let Y denote the range of f, i.e. $Y = f[X] \subset \mathbf{I}$. We want to prove that $f^{-1}: Y \to X$ is also continuous. Observe that continuity in Y is equivalent to sequential continuity; hence f^{-1} is continuous at $f(p) \in Y$ if for every sequence $\langle f(y_n)\rangle$ converging to $f(p)$, the sequence $\langle y_n\rangle$ converges to p.

Suppose f^{-1} is not continuous, i.e. suppose $\langle y_n\rangle$ does not converge to p. Then there exists an open neighborhood G of p such that G does not contain an infinite number of the terms of $\langle y_n\rangle$. Hence we can choose a subsequence $\langle x_n\rangle$ of $\langle y_n\rangle$ such that all the terms of $\langle x_n\rangle$ lie outside of G. Since $p \in G$, there exists a member G_i in the base $\mathcal{B}$ such that $p \in G_i \subset G$. Furthermore, by a previous problem, there exists a pair $P_m = \langle G_j, G_i\rangle$ such that $p \in \bar{G}_j \subset G_i \subset G$. Observe that, for all $n \in \mathbf{N}$, $x_n \notin G$; hence $x_n \in G_i^c$. Accordingly, $f_m(p) = 0$ and $f_m(x_n) = 1$. Then $|f_m(x_n) - f_m(p)|^2 = 1$ and

$$||f(x_n) - f(p)||^2 = \sum_{k=1}^{\infty} \frac{|f_k(x_n) - f_k(p)|^2}{2^{2k}} \geq \frac{1}{2^{2m}}$$

In other words, for every $n \in \mathbf{N}$, $||f(x_n) - f(p)|| > 1/2^m$. Therefore the sequence $\langle f(x_n)\rangle$ does not converge to $f(p)$. But this contradicts the fact that every subsequence of $\langle f(y_n)\rangle$ should also converge to $f(p)$. Hence f^{-1} is continuous. Hence f is a homeomorphism and X is homeomorphic to a subset of the Hilbert cube. Accordingly, X is metrizable.

REGULAR AND COMPLETELY REGULAR SPACES

17. Prove Proposition 10.10: A completely regular space X is also regular.

Solution:

Let F be a closed subset of X and suppose $p \in X$ does not belong to F. By hypothesis, X is completely regular; hence there exists a continuous function $f: X \to [0,1]$ such that $f(p) = 0$ and $f[F] = \{1\}$. But $\mathbf{R}$ and its subspace $[0,1]$ are Hausdorff spaces; hence there are disjoint open sets G and H containing 0 and 1 respectively. Accordingly, their inverses $f^{-1}[G]$ and $f^{-1}[H]$ are disjoint, open and contain p and F respectively. In other words, X is also regular.

18. Prove Theorem 10.11: *The class $C(X, \mathbf{R})$ of all real-valued continuous functions on a completely regular T_1-space X separates points.*

Solution:

Let a and b be distinct points in X. Since X is a T_1-space, $\{b\}$ is a closed set. Also, since a and b are distinct, $a \notin \{b\}$. By hypothesis, X is completely regular; hence there exists a real-valued continuous function f on X such that $f(a) = 0$ and $f[\{b\}] = \{1\}$. Accordingly, $f(a) \neq f(b)$.

19. Let $(Y, \mathcal{T}_Y)$ be a subspace of $(X, \mathcal{T})$ and let $p \in Y$ and $A \subset Y \subset X$. Show that if p does not belong to the $\mathcal{T}_Y$-closure of A, then $p \notin \bar{A}$, the $\mathcal{T}$-closure of A.

Solution:

Now, by a property of subspaces (see Problem 89, Chapter 5),

$$\mathcal{T}_Y\text{-closure of } A \quad = \quad Y \cap \bar{A}$$

But $p \in Y$ and $p \notin \mathcal{T}_Y$-closure of A; hence $p \notin \bar{A}$. (Observe that, in particular, if F is a $\mathcal{T}_Y$-closed subset of Y and $p \notin F$, then $p \notin \bar{F}$.)

20. Show that the property of being a regular space is hereditary, i.e. every subspace of a regular space is regular.

Solution:

Let $(X, \mathcal{T})$ be a regular space and let $(Y, \mathcal{T}_Y)$ be a subspace of $(X, \mathcal{T})$. Furthermore, let $p \in Y$ and let F be a $\mathcal{T}_Y$-closed subset of Y such that $p \notin F$. Now by Problem 19, $p \notin \bar{F}$, the $\mathcal{T}$-closure of F. By hypothesis, $(X, \mathcal{T})$ is regular; hence

$$\exists\ G, H \in \mathcal{T} \quad \text{such that} \quad \bar{F} \subset G,\ p \in H \ \text{ and } \ G \cap H = \emptyset$$

But $Y \cap G$ and $Y \cap H$ are $\mathcal{T}_Y$-open subsets of Y, and

$$F \subset Y,\ F \subset \bar{F} \subset G \quad \Rightarrow \quad F \subset Y \cap G$$

$$p \in Y,\ p \in H \quad \Rightarrow \quad p \in Y \cap H$$

$$G \cap H = \emptyset \quad \Rightarrow \quad (Y \cap G) \cap (Y \cap H) = \emptyset$$

Accordingly, $(Y, \mathcal{T}_Y)$ is also regular.

Supplementary Problems

T_1-SPACES

21. Show that the property of being a T_1-space is topological.

22. Show, by a counterexample, that the image of a T_1-space under a continuous map need not be T_1.

23. Let $(X, \mathcal{T})$ be a T_1-space and let $\mathcal{T} \precsim \mathcal{T}^*$. Show that $(X, \mathcal{T}^*)$ is also a T_1-space.

24. Prove: X is a T_1-space if and only if every $p \in X$ is the intersection of all open sets containing it, i.e. $\{p\} = \bigcap\{G : G \text{ open}, p \in G\}$.

25. A topological space X is called a T_0-space if it satisfies the following axiom:

[$\mathbf{T_0}$] For any pair of distinct points in X, there exists an open set containing one of the points but not the other.

(i) Give an example of a T_0-space which is not a T_1-space.

(ii) Show that every T_1-space is also a T_0-space.

26. Let X be a T_1-space containing at least two points. Show that if $\mathcal{B}$ is a base for X then $\mathcal{B} \setminus \{X\}$ is also a base for X.

HAUSDORFF SPACES

27. Show that the property of being a Hausdorff space is topological.

28. Let $(X, \mathcal{T})$ be a Hausdorff space and let $\mathcal{T} \precsim \mathcal{T}^*$. Show that $(X, \mathcal{T}^*)$ is also a Hausdorff space.

29. Show that if $a_1, \ldots, a_m$ are distinct points in a Hausdorff space X, then there exists a disjoint class $\{G_1, \ldots, G_m\}$ of open subsets of X such that $a_1 \in G_1, \ldots, a_m \in G_m$.

30. Prove: Let X be an infinite Hausdorff space. Then there exists an infinite disjoint class of open subsets of X.

31. Prove: Let $f: X \to Y$ and $g: X \to Y$ be continuous functions from a topological space X into a Hausdorff space Y. Then $A = \{x : f(x) = g(x)\}$ is a closed subset of X.

NORMAL SPACES

32. Show that the property of being a normal space is topological.

33. Let $\mathcal{T}$ be the topology on the real line $\mathbf{R}$ generated by the closed-open intervals $[a, b)$. Show that $(\mathbf{R}, \mathcal{T})$ is a normal space.

34. Let $\mathcal{T}$ be the topology on the plane $\mathbf{R}^2$ generated by the half-open rectangles,

$$[a, b) \times [c, d) \quad = \quad \{\langle x, y\rangle : a \leqq x < b,\ c \leqq y < d\}$$

Furthermore, let A consist of the points on the line $Y = \{\langle x, y\rangle : x + y = 1\} \subset \mathbf{R}^2$ whose coordinates are rational and let $B = Y \setminus A$.

(i) Show that A and B are closed subsets of $(\mathbf{R}^2, \mathcal{T})$.

(ii) Show that there exist no disjoint $\mathcal{T}$-open subsets G and H of $\mathbf{R}^2$ such that $A \subset G$ and $B \subset H$; and so $(\mathbf{R}^2, \mathcal{T})$ is not normal.

35. Let A be a closed subset of a normal T_1-space. Show that A with the relative topology is also a normal T_1-space.

36. Let X be an ordered set and let $\mathcal{T}$ be the order topology on X, i.e. $\mathcal{T}$ is generated by the subsets of X of the form $\{x : x < a\}$ and $\{x : x > a\}$. Show that $(X, \mathcal{T})$ is a normal space.

37. Prove: Let X be a normal space. Then X is regular if and only if X is completely regular.

URYSOHN'S LEMMA

38. Prove: If for every two disjoint closed subsets F_1 and F_2 of a topological space X, there exists a continuous function $f: X \to [0, 1]$ such that $f[F_1] = \{0\}$ *and* $f[F_2] = \{1\}$, *then X is a normal space.* (Note that this is the converse of Urysohn's Lemma.)

39. Prove the following generalization of Urysohn's Lemma: Let F_1 and F_2 be disjoint closed subsets of a normal space X. Then there exists a continuous function $f: X \to [a, b]$ such that $f[F_1] = \{a\}$ and $f[F_2] = \{b\}$.

40. Prove the Tietze Extension Theorem: Let F be a closed subset of a normal space X and let $f: F \to [a, b]$ be a real continuous function. Then f has a continuous extension $f^*: X \to [a, b]$.

41. Prove Urysohn's Lemma using the Tietze Extension Theorem.

REGULAR AND COMPLETELY REGULAR SPACES

42. Show that the property of being a regular space is topological.

43. *Show that the property of being* completely regular is topological.

44. Show that the property of being a completely regular space is hereditary, that is, every subspace of a completely regular space is also completely regular.

45. Prove: Let X be a regular Lindelöf space. Then X is normal.

Answers to Supplementary Problems

25. (i) Let $X = \{a, b\}$ and $\mathcal{T} = \{X, \{a\}, \emptyset\}$.

Chapter 11

Compactness

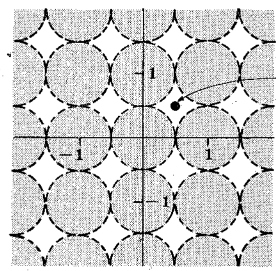
COVERS

Let $\mathcal{A} = \{G_i\}$ be a class of subsets of X such that $A \subset \cup_i G_i$ for some $A \subset X$. Recall that $\mathcal{A}$ is then called a *cover* of A, and an *open cover* if each G_i is open. Furthermore, if a finite subclass of $\mathcal{A}$ is also a cover of A, i.e. if

$$\exists\ G_{i_1}, \ldots, G_{i_m} \in \mathcal{A} \quad \text{such that} \quad A \subset G_{i_1} \cup \cdots \cup G_{i_m}$$

then $\mathcal{A}$ is said to be *reducible to a finite cover*, or contains a *finite subcover.*

Example 1.1: Consider the class $\mathcal{A} = \{D_p : p \in \mathbf{Z} \times \mathbf{Z}\}$, where D_p is the open disc in the plane $\mathbf{R}^2$ with radius 1 and center $p = \langle m, n\rangle$, m and n integers. Then $\mathcal{A}$ is a cover of $\mathbf{R}^2$, i.e. every point in $\mathbf{R}^2$ belongs to at least one member of $\mathcal{A}$. On the other hand, the class of open discs $\mathcal{B} = \{D_p^* : p \in \mathbf{Z} \times \mathbf{Z}\}$, where D_p has center p and radius $\frac{1}{2}$, is not a cover of $\mathbf{R}^2$. For example, the point $\langle \frac{1}{2}, \frac{1}{2}\rangle \in \mathbf{R}^2$ does not belong to any member of $\mathcal{B}$, as shown in the figure.



$\langle \frac{1}{2}, \frac{1}{2}\rangle$

$\mathcal{B}$ is displayed

Example 1.2: Consider the classical

Heine-Borel Theorem: Let $A = [a, b]$ be a closed and bounded interval and let $\{G_i\}$ be a class of open sets such that $A \subset \cup_i G_i$. Then one can select a finite number of the open sets, say $G_{i_1}, \ldots, G_{i_m}$, so that $A \subset G_{i_1} \cup \cdots \cup G_{i_m}$.

By virtue of the above terminology, the Heine-Borel Theorem can be restated as follows:

Heine-Borel Theorem: Every open cover of a closed and bounded interval $A = [a, b]$ is reducible to a finite cover.

COMPACT SETS

The concept of *compactness* is no doubt motivated by the property of a closed and bounded interval as stated in the classical Heine-Borel Theorem. Namely,

Definition: A subset A of a topological space X is *compact* if every open cover of A is reducible to a finite cover.

In other words, if A is compact and $A \subset \cup_i G_i$, where the G_i are open sets, then one can select a finite number of the open sets, say $G_{i_1}, \ldots, G_{i_m}$, so that $A \subset G_{i_1} \cup \cdots \cup G_{i_m}$.

Example 2.1: By the Heine-Borel Theorem, every closed and bounded interval $[a, b]$ on the real line $\mathbf{R}$ is compact.

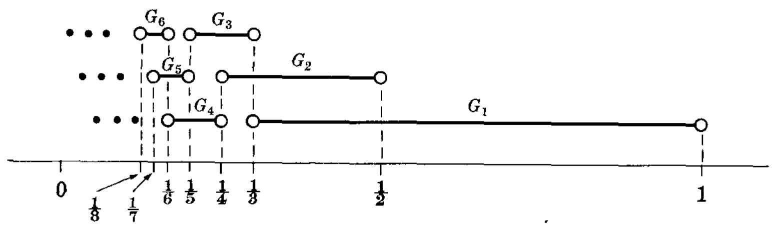
Example 2.2: Let A be any finite subset of a topological space X, say $A = \{a_1, \ldots, a_m\}$. Then A is necessarily compact. For if $\mathcal{G} = \{G_i\}$ is an open cover of A, then each point in A belongs to one of the members of $\mathcal{G}$, say $a_1 \in G_{i_1}, \ldots, a_m \in G_{i_m}$. Accordingly, $A \subset G_{i_1} \cup G_{i_2} \cup \cdots \cup G_{i_m}$.

Since a set A is compact iff every open cover of A contains a finite subcover, we only have to exhibit one open cover of A with no finite subcover to prove that A is not compact.

Example 2.3: The open interval $A = (0,1)$ on the real line $\mathbf{R}$ with the usual topology is not compact. Consider, for example, the class of open intervals

$$\mathcal{G} = \{(\tfrac{1}{3}, 1), (\tfrac{1}{4}, \tfrac{1}{2}), (\tfrac{1}{5}, \tfrac{1}{3}), (\tfrac{1}{6}, \tfrac{1}{4}), \ldots\}$$

Observe that $A = \cup_{n=1}^{\infty} G_n$, where $G_n = \left(\frac{1}{n+2}, \frac{1}{n}\right)$; hence $\mathcal{G}$ is an open cover of A.



But $\mathcal{G}$ contains no finite subcover. For let

$$\mathcal{G}^* = \{(a_1, b_1), (a_2, b_2), \ldots, (a_m, b_m)\}$$

be any finite subclass of $\mathcal{G}$. If $\epsilon = \min(a_1, \ldots, a_m)$ then $\epsilon > 0$ and

$$(a_1, b_1) \cup \cdots \cup (a_m, b_m) \subset (\epsilon, 1)$$

But $(0, \epsilon]$ and $(\epsilon, 1)$ are disjoint; hence $\mathcal{G}^*$ is not a cover of A, and so A is not compact.

Example 2.4: We show that a continuous image of a compact set is also compact, i.e. if the function $f: X \to Y$ is continuous and A is a compact subset of X, then its image $f[A]$ is a compact subset of Y. For suppose $\mathcal{G} = \{G_i\}$ is an open cover of $f[A]$, i.e. $f[A] \subset \cup_i G_i$. Then

$$A \subset f^{-1}[f[A]] \subset f^{-1}[\cup_i G_i] = \cup_i f^{-1}[G_i]$$

Hence $\mathcal{H} = \{f^{-1}[G_i]\}$ is a cover of A. Now f is continuous and each G_i is an open set, so each $f^{-1}[G_i]$ is also open. In other words, $\mathcal{H}$ is an open cover of A. But A is compact, so $\mathcal{H}$ is reducible to a finite cover, say

$$A \subset f^{-1}[G_{i_1}] \cup \cdots \cup f^{-1}[G_{i_m}]$$

Accordingly,

$$f[A] \subset f[f^{-1}[G_{i_1}] \cup \cdots \cup f^{-1}[G_{i_m}]] \subset G_{i_1} \cup \cdots \cup G_{i_m}$$

Thus $f[A]$ is compact.

We formally state the result in Example 2.4:

Theorem 11.1: Continuous images of compact sets are compact.

Compactness is an *absolute property* of a set. Namely,

Theorem 11.2: Let A be a subset of a topological space $(X, \mathcal{T})$. Then A is compact with respect to $\mathcal{T}$ if and only if A is compact with respect to the relative topology $\mathcal{T}_A$ on A.

Accordingly, we can frequently limit our investigation of compactness to those topological spaces which are themselves compact, i.e. to *compact spaces.*

SUBSETS OF COMPACT SPACES

A subset of a compact space need not be compact. For example, the closed unit interval $[0,1]$ is compact by the Heine-Borel Theorem, but the open interval $(0,1)$ is a subset of $[0,1]$ which, by Example 2.3 above, is not compact. We do, however, have the following

Theorem 11.3: Let F be a closed subset of a compact space X. Then F is also compact.

Proof: Let $\mathcal{G} = \{G_i\}$ be an open cover of F, i.e. $F \subset \cup_i G_i$. Then $X = (\cup_i G_i) \cup F^c$, that is, $\mathcal{G}^* = \{G_i\} \cup \{F^c\}$ is a cover of X. But F^c is open since F is closed, so $\mathcal{G}^*$ is an open cover of X. By hypothesis, X is compact; hence $\mathcal{G}^*$ is reducible to a finite cover of X, say

$$X = G_{i_1} \cup \cdots \cup G_{i_m} \cup F^c, \qquad G_{i_k} \in \mathcal{G}$$

But F and F^c are disjoint; hence

$$F \subset G_{i_1} \cup \cdots \cup G_{i_m}, \qquad G_{i_k} \in \mathcal{G}$$

We have just shown that any open cover $\mathcal{G} = \{G_i\}$ of F contains a finite subcover, i.e. F is compact.

FINITE INTERSECTION PROPERTY

A class $\{A_i\}$ of sets is said to have the *finite intersection property* if every finite subclass $\{A_{i_1}, \ldots, A_{i_m}\}$ has a non-empty intersection, i.e. $A_{i_1} \cap \cdots \cap A_{i_m} \neq \emptyset$.

Example 3.1: Consider the following class of open intervals:

$$\mathcal{A} = \{(0,1), (0,\tfrac{1}{2}), (0,\tfrac{1}{3}), (0,\tfrac{1}{4}), \ldots\}$$

Now $\mathcal{A}$ has the finite intersection property, for

$$(0,a_1) \cap (0,a_2) \cap \cdots \cap (0,a_m) = (0,b)$$

where $b = \min(a_1, \ldots, a_m) > 0$. Observe that $\mathcal{A}$ itself has an empty intersection.

Example 3.2: Consider the following class of closed infinite intervals:

$$\mathcal{B} = \{\ldots, (-\infty,-2], (-\infty,-1], (-\infty,0], (-\infty,1], (-\infty,2], \ldots\}$$

Note that $\mathcal{B}$ has an empty intersection, i.e. $\bigcap\{B_n : n \in \mathbf{Z}\} = \emptyset$ where $B_n = (-\infty, n]$. But any finite subclass of $\mathcal{B}$ has a non-empty intersection. In other words, $\mathcal{B}$ satisfies the finite intersection property.

With the above terminology, we can now state the notion of compactness in terms of the closed subsets of a topological space.

Theorem 11.4: A topological space X is compact if and only if every class $\{F_i\}$ of closed subsets of X which satisfies the finite intersection property has, itself, a non-empty intersection.

COMPACTNESS AND HAUSDORFF SPACES

Here we relate the concept of compactness to the separation property of Hausdorff spaces.

Theorem 11.5: Every compact subset of a Hausdorff space is closed.

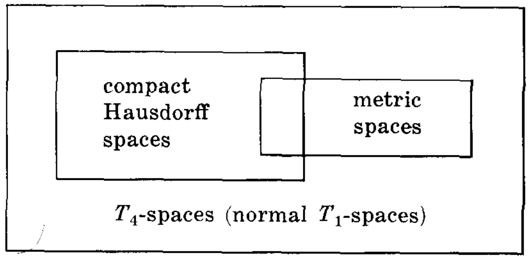
The above theorem is not true in general; for example, finite sets are always compact and yet there exist topological spaces whose finite subsets are not all closed.

Theorem 11.6: Let A and B be disjoint compact subsets of a Hausdorff space X. Then there exist disjoint open sets G and H such that $A \subset G$ and $B \subset H$.

In particular, suppose X is both Hausdorff and compact and F_1 and F_2 are disjoint closed subsets of X. By Theorem 11.3, F_1 and F_2 are compact and, by Theorem 11.6, F_1 and F_2 are subsets, respectively, of disjoint open sets. In other words,

Corollary 11.7: Every compact Hausdorff space is normal.

Thus metric spaces and compact Hausdorff spaces are both contained in the class of T_4-spaces, i.e. normal T_1-spaces.

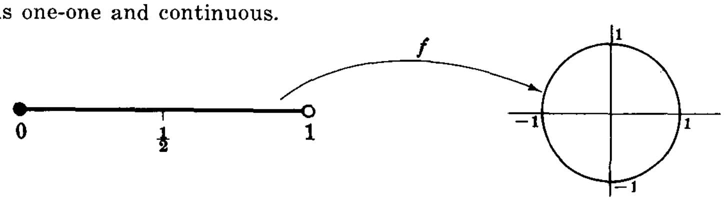


The following theorem plays a very important role in geometry.

Theorem 11.8: Let f be a one-one continuous function from a compact space X into a Hausdorff space Y. Then X and $f[X]$ are homeomorphic.

The next example shows that the above theorem is not true in general.

Example 4.1: Let f be the function from the half-open interval $X = [0,1)$ into the plane $\mathbf{R}^2$ defined by $f(t) = \langle \cos 2\pi t, \sin 2\pi t \rangle$. Observe that f maps X onto the unit circle and that f is one-one and continuous.



But the half-open interval $[0,1)$ is not homeomorphic to the circle. For example, if we delete the point $t = \frac{1}{2}$ from X, X will not be connected; but if we delete any point from a circle, the circle is still connected. The reason that Theorem 11.8 does not apply in this case is that X is not compact.

Example 4.2: Let f be a one-one continuous function from the closed unit interval $I = [0,1]$ into Euclidean n-space $\mathbf{R}^n$. Observe that I is compact by the Heine-Borel Theorem and that $\mathbf{R}^n$ is a metric space and therefore Hausdorff. By virtue of Theorem 11.8, I and $f[I]$ are homeomorphic.

SEQUENTIALLY COMPACT SETS

A subset A of a topological space X is *sequentially compact* iff every sequence in A contains a subsequence which converges to a point in A.

Example 5.1: Let A be a finite subset of a topological space X. Then A is necessarily sequentially compact. For if $\langle s_1, s_2, \ldots \rangle$ is a sequence in A, then at least one of the elements in A, say a_0, must appear an infinite number of times in the sequence. Hence $\langle a_0, a_0, a_0, \ldots \rangle$ is a subsequence of $\langle s_n \rangle$, it converges, and furthermore it converges to the point a_0 belonging to A.

Example 5.2: The open interval $A = (0,1)$ on the real line $\mathbf{R}$ with the usual topology is not sequentially compact. Consider, for example, the sequence $\langle s_n \rangle = \langle \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots \rangle$ in A. Observe that $\langle s_n \rangle$ converges to 0 and therefore every subsequence also converges to 0. But 0 does not belong to A. In other words, the sequence $\langle s_n \rangle$ in A does not contain a subsequence which converges to a point in A, i.e. A is not sequentially compact.

In general, there exist compact sets which are not sequentially compact and vice versa, although in metric spaces, as we show later, they are equivalent.

Remark: Historically, the term *bicompact* was used to denote a compact set, and the term compact was used to denote a sequentially compact set.

COUNTABLY COMPACT SETS

A subset A of a topological space X is *countably compact* iff every infinite subset B of A has an accumulation point in A. This definition is no doubt motivated by the classical

Bolzano-Weierstrass Theorem: Every bounded infinite set of real numbers has an accumulation point.

Example 6.1: Every bounded closed interval $A = [a, b]$ is countably compact. For if B is an infinite subset of A, then B is also bounded and, by the Bolzano-Weierstrass Theorem, B has an accumulation point p. Furthermore, since A is closed, the accumulation point p of B belongs to A, i.e. A is countably compact.

Example 6.2: The open interval $A = (0,1)$ is not countably compact. For consider the infinite subset $B = \{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots\}$ of $A = (0,1)$. Observe that B has exactly one limit point which is 0 and that 0 does not belong to A. Hence A is not countably compact.

The general relationship between compact, sequentially compact and countably compact sets is given in the following diagram and theorem.

$$\text{compact} \quad \rightarrow \quad \text{countably compact} \quad \leftarrow \quad \text{sequentially compact}$$

Theorem 11.9: Let A be a subset of a topological space X. If A is compact or sequentially compact, then A is also countably compact.

The next example shows that neither arrow in the above diagram can be reversed.

Example 6.3: Let $\mathcal{T}$ be the topology on $\mathbf{N}$, the set of positive integers, generated by the following sets:

$$\{1, 2\}, \ \{3, 4\}, \ \{5, 6\}, \ \ldots$$

Let A be a non-empty subset of $\mathbf{N}$, say $n_0 \in A$. If n_0 is odd, then $n_0 + 1$ is a limit point of A; and if n_0 is even, then $n_0 - 1$ is a limit point of A. In either case, A has an accumulation point. Accordingly, $(\mathbf{N}, \mathcal{T})$ is countably compact.

On the other hand, $(\mathbf{N}, \mathcal{T})$ is not compact since

$$\mathcal{A} = \{\{1,2\}, \{3,4\}, \{5,6\}, \ldots\}$$

is an open cover of $\mathbf{N}$ with no finite subcover. Furthermore, $(\mathbf{N}, \mathcal{T})$ is not sequentially compact, since the sequence $\langle 1, 2, 3, \ldots \rangle$ contains no convergent subsequence.

LOCALLY COMPACT SPACES

A topological space X is *locally compact* iff every point in X has a compact neighborhood.

Example 7.1: Consider the real line $\mathbf{R}$ with the usual topology. Observe that each point $p \in \mathbf{R}$ is interior to a closed interval, e.g. $[p - \delta, p + \delta]$, and that the closed interval is compact by the Heine-Borel Theorem. Hence $\mathbf{R}$ is a locally compact space. On the other hand, $\mathbf{R}$ is not a compact space; for example, the class

$$\mathcal{A} = \{\ldots, (-3, -1), (-2, 0), (-1, 1), (0, 2), (1, 3), \ldots\}$$

is an open cover of $\mathbf{R}$ but contains no finite subcover.

Thus we see, by the above example, that a locally compact space need not be compact. On the other hand, since a topological space is always a neighborhood of each of its points, the converse is true. That is,

Proposition 11.10: Every compact space is locally compact.

COMPACTIFICATION

A topological space X is said to be *embedded* in a topological space Y if X is homeomorphic to a subspace of Y. Furthermore, if Y is a compact space, then Y is called a *compactification* of X. Frequently, the compactification of a space X is accomplished by adjoining one or more points to X and then defining an appropriate topology on the enlarged set so that the enlarged space is compact and contains X as a subspace.

Example 8.1: Consider the real line $\mathbf{R}$ with the usual topology $\mathcal{U}$. We adjoin two new points, denoted by ∞ and $-\infty$, to $\mathbf{R}$ and call the enlarged set $\mathbf{R}^* = \mathbf{R} \cup \{-\infty, \infty\}$ the *extended real line*. The order relation in $\mathbf{R}$ can be extended to $\mathbf{R}^*$ by defining $-\infty < a < \infty$ for any $a \in \mathbf{R}$. The class of subsets of $\mathbf{R}^*$ of the form

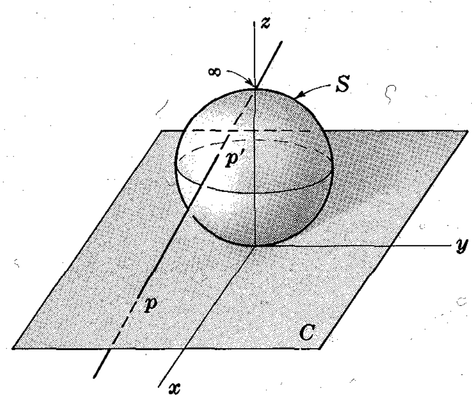
$$(a, b) = \{x : a < x < b\}, \quad (a, \infty] = \{x : a < x\} \quad \text{and} \quad [-\infty, a) = \{x : x < a\}$$

is a base for a topology $\mathcal{U}^*$ on $\mathbf{R}^*$. Furthermore, $(\mathbf{R}^*, \mathcal{U}^*)$ *is a compact space and contains* $(\mathbf{R}, \mathcal{U})$ as a subspace, and so it is a compactification of $(\mathbf{R}, \mathcal{U})$.

Recall that the real line $\mathbf{R}$ with the usual topology is homeomorphic to any open interval (a, b) of real numbers. The above space $(\mathbf{R}^*, \mathcal{U}^*)$ can, in fact, be shown to be homeomorphic to any closed interval $[a, b]$ which is compact by the classical Heine-Borel Theorem.

Example 8.2: Let C denote the $\langle x, y\rangle$-plane in Euclidian 3-space $\mathbf{R}^3$, and let S denote the sphere with center $\langle 0, 0, 1\rangle$ on the z-axis and radius 1. The line passing through the "north pole" $\infty = \langle 0, 0, 2\rangle \in S$ and any point $p \in C$ intersects the sphere S in exactly one point p' distinct from ∞, as shown in the figure.

Let $f : C \to S$ be defined by $f(p) = p'$. Then f is, in fact, a homeomorphism from the plane C, which is not compact, onto the subset $S \setminus \{\infty\}$ of the sphere S, and S is compact. Hence S is a compactification of C.



Now let $(X, \mathcal{T})$ be any topological space. We shall define the *Alexandrov* or *one-point compactification* of $(X, \mathcal{T})$ which we denote by $(X_\infty, \mathcal{T}_\infty)$. Here:

(1) $X_\infty = X \cup \{\infty\}$, where ∞, called the *point at infinity*, is distinct from every other point in X.

(2) $\mathcal{T}_\infty$ consists of the following sets:

  (i) each member of the topology $\mathcal{T}$ on X,

  (ii) the complement in X_∞ of any closed and compact subset of X.

We formally state:

Proposition 11.11: The above class $\mathcal{T}_\infty$ is a topology on X_∞, and $(X_\infty, \mathcal{T}_\infty)$ is a compactification of $(X, \mathcal{T})$.

In general, the space $(X_\infty, \mathcal{T}_\infty)$ may not possess properties similar to those of the original space. There does exist one important relationship between the two spaces; namely,

Theorem 11.12: *If $(X, \mathcal{T})$ is a locally compact Hausdorff space, then $(X_\infty, \mathcal{T}_\infty)$ is a compact* Hausdorff space.

Using Urysohn's lemma we obtain an important result used in measure and integration theory:

Corollary 11.13: Let E be a compact subset of a *locally compact Hausdorff space* X, *and* let E be a subset of an open set $G \neq X$. Then there exists a continuous function $f : X \to [0, 1]$ such that $f[E] = \{0\}$ and $f[G^c] = \{1\}$.

COMPACTNESS IN METRIC SPACES

Compactness in metric spaces can be summarized by the following

Theorem 11.14: Let A be a subset of a metric space X. Then the following statements are equivalent: (i) A is compact, (ii) A is countably compact, and (iii) A is sequentially compact.

Historically, metric spaces were investigated before topological spaces; hence the above theorem gives the main reason that the terms compact and sequentially compact are sometimes used synonymously.

The proof of the above theorem requires the introduction of two auxiliary metric concepts which are interesting in their own right: that of a *totally bounded set* and that of a *Lebesgue number* for a cover.

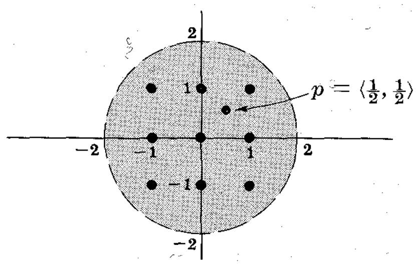
TOTALLY BOUNDED SETS

Let A be a subset of a metric space X and let $\epsilon > 0$. A finite set of points $N = \{e_1, e_2, \ldots, e_m\}$ is called an ϵ*-net* for A if for every point $p \in A$ there exists an $e_{i_0} \in N$ with $d(p, e_{i_0}) < \epsilon$.

Example 9.1: Let $A = \{\langle x, y\rangle : x^2 + y^2 < 4\}$, i.e. A is the open disc centered at the origin and of radius 2. If $\epsilon = 3/2$, then the set

$$N = \{\langle 1,-1\rangle, \langle 1,0\rangle, \langle 1,1\rangle, \langle 0,-1\rangle, \langle 0,0\rangle, \langle 0,1\rangle, \langle -1,-1\rangle, \langle -1,0\rangle, \langle -1,1\rangle\}$$

is an ϵ-net for A. On the other hand, if $\epsilon = \frac{1}{2}$, then N is not an ϵ-net for A. For example, $p = \langle \frac{1}{2}, \frac{1}{2}\rangle$ belongs to A but the distance between p and any point in N is greater than $\frac{1}{2}$.



A is shaded

N is displayed

Recall that the *diameter* of A, $d(A)$, is defined by $d(A) = \sup\{d(a, a') : a, a' \in A\}$ and that A is bounded if $d(A) < \infty$.

Definition: A subset A of a metric space X is *totally bounded* if A possesses an ϵ-net for every $\epsilon > 0$.

A totally bounded set can also be described as follows:

Proposition 11.15: A set A is totally bounded if and only if for every $\epsilon > 0$ there exists a decomposition of A into a finite number of sets, each with diameter less than ϵ.

We first show that a bounded set need not be totally bounded.

Example 9.2: Let A be the subset of Hilbert Space, i.e. of l_2-space, consisting of the following points:

$$e_1 = \langle 1, 0, 0, \ldots\rangle$$
$$e_2 = \langle 0, 1, 0, \ldots\rangle$$
$$e_3 = \langle 0, 0, 1, \ldots\rangle$$
$$\cdots$$

Observe that $d(e_i, e_j) = \sqrt{2}$ if $i \neq j$. Hence A is bounded; in fact,

$$d(A) = \sup\{d(e_i, e_j) : e_i, e_j \in A\} = \sqrt{2}$$

On the other hand, A is not totally bounded. For if $\epsilon = \frac{1}{2}$, the only non-empty subsets of A with diameter less than ϵ are the singleton sets, i.e. sets with one point. Accordingly, the infinite set A cannot be decomposed into a finite number of disjoint subsets each with diameter less than $\frac{1}{2}$.

The converse of the previous statement is true. Namely,

Proposition 11.16: Totally bounded sets are bounded.

One relationship between compactness and total boundedness is as follows:

Lemma 11.17: Sequentially compact sets are totally bounded.

LEBESGUE NUMBERS FOR COVERS

Let $\mathcal{A} = \{G_i\}$ be a cover for a subset A of a metric space X. A real number $\delta > 0$ is called a *Lebesgue number* for the cover if for each subset of A with diameter less than δ there is a member of the cover which contains A.

One relationship between compactness and Lebesgue number for a cover is as follows:

Lemma (Lebesgue) 11.18: Every open cover of a sequentially compact subset of a metric space has a (positive) Lebesgue number.

Solved Problems

COMPACT SPACES

1. Let $\mathcal{T}$ be the cofinite topology on any set X. Show that $(X, \mathcal{T})$ is a compact space.

Solution:

Let $\mathcal{G} = \{G_i\}$ be an open cover of X. Choose $G_0 \in \mathcal{G}$. Since $\mathcal{T}$ is the cofinite topology, G_0^c is a finite set, say $G_0^c = \{a_1, \ldots, a_m\}$. Since $\mathcal{G}$ is a cover of X,

$$\text{for each } a_k \in G_0^c \quad \exists \; G_{i_k} \in \mathcal{G} \quad \text{such that} \quad a_k \in G_{i_k}$$

Hence $G_0^c \subset G_{i_1} \cup \cdots \cup G_{i_m}$ and $X = G_0 \cup G_0^c = G_0 \cup G_{i_1} \cup \cdots \cup G_{i_m}$. Thus X is compact.

2. Show that any infinite subset A of a discrete topological space X is not compact.

Solution:

Recall that A is not compact if we can exhibit an open cover of A with no finite subcover. Consider the class $\mathcal{A} = \{\{a\} : a \in A\}$ of singleton subsets of A. Observe that: (i) $\mathcal{A}$ is a cover of A; in fact $A = \bigcup\{\{a\} : a \in A\}$. (ii) $\mathcal{A}$ is an open cover of A since all subsets of a discrete space are open. (iii) No proper subclass of $\mathcal{A}$ is a cover of A. (iv) $\mathcal{A}$ is infinite since A is infinite. Accordingly, the open cover $\mathcal{A}$ of A contains no finite subcover, so A is not compact.

Since finite sets are always compact, we have also proven that a subset of *a discrete space is* compact if and only if it is finite.

3. Prove Theorem 11.2: Let A be a subset of a topological space $(X, \mathcal{T})$. Then the following are equivalent:

(i) A is compact with respect to $\mathcal{T}$.

(ii) A is compact with respect to the relative topology $\mathcal{T}_A$ on A.

Solution:

(i) $\Rightarrow$ (ii): Let $\{G_i\}$ be a $\mathcal{T}_A$-open cover of A. By definition of the relative topology,

$$\exists\; H_i \in \mathcal{T} \quad \text{such that} \quad G_i = A \cap H_i \subset H_i$$

Hence
$$A \subset \cup_i G_i \subset \cup_i H_i$$

and therefore $\{H_i\}$ is a $\mathcal{T}$-open cover of A. By (i), A is $\mathcal{T}$-compact, so $\{H_i\}$ contains a finite subcover, say

$$A \subset H_{i_1} \cup \cdots \cup H_{i_m}, \qquad H_{i_k} \in \{H_i\}$$

But then
$$A \subset A \cap (H_{i_1} \cup \cdots \cup H_{i_m}) = (A \cap H_{i_1}) \cup \cdots \cup (A \cap H_{i_m}) = G_{i_1} \cup \cdots \cup G_{i_m}$$

Thus $\{G_i\}$ contains a finite subcover $\{G_{i_1}, \ldots, G_{i_m}\}$ and $(A, \mathcal{T}_A)$ is compact.

(ii) $\Rightarrow$ (i): Let $\{H_i\}$ be a $\mathcal{T}$-open cover of A. Set $G_i = A \cap H_i$; then

$$A \subset \cup_i H_i \Rightarrow A \subset A \cap (\cup_i H_i) = \cup_i (A \cap H_i) = \cup_i G_i$$

But $G_i \in \mathcal{T}_A$, so $\{G_i\}$ is a $\mathcal{T}_A$-open cover of A. By hypothesis, A is $\mathcal{T}_A$-compact; thus $\{G_i\}$ contains a finite subcover $\{G_{i_1}, \ldots, G_{i_m}\}$. Accordingly,

$$A \subset G_{i_1} \cup \cdots \cup G_{i_m} = (A \cap H_{i_1}) \cup \cdots \cup (A \cap H_{i_m}) = A \cap (H_{i_1} \cup \cdots \cup H_{i_m}) \subset H_{i_1} \cup \cdots \cup H_{i_m}$$

Thus $\{H_i\}$ is reducible to a finite cover $\{H_{i_1}, \ldots, H_{i_m}\}$ and therefore A is compact with respect to $\mathcal{T}$.

4. Let $(Y, \mathcal{T}^*)$ be a subspace of $(X, \mathcal{T})$ and let $A \subset Y \subset X$. Show that A is $\mathcal{T}$-compact if and only if A is $\mathcal{T}^*$-compact.

Solution:

Let $\mathcal{T}_A$ and $\mathcal{T}_A^*$ be the relative topologies on A. Then, by the preceding problem, A is $\mathcal{T}$- or $\mathcal{T}^*$-compact if and only if A is $\mathcal{T}_A$- or $\mathcal{T}_A^*$-compact; but $\mathcal{T}_A = \mathcal{T}_A^*$.

5. Prove that the following statements are equivalent:

(i) X is compact.

(ii) For every class $\{F_i\}$ of closed subsets of X, $\cap_i F_i = \emptyset$ implies $\{F_i\}$ contains a finite subclass $\{F_{i_1}, \ldots, F_{i_m}\}$ with $F_{i_1} \cap \cdots \cap F_{i_m} = \emptyset$.

Solution:

(i) $\Rightarrow$ (ii): Suppose $\cap_i F_i = \emptyset$. Then, by DeMorgan's Law,

$$X = \emptyset^c = (\cap_i F_i)^c = \cup_i F_i^c$$

so $\{F_i^c\}$ is an open cover of X, since each F_i is closed. But by hypothesis, X is compact; hence

$$\exists\; F_{i_1}^c, \ldots, F_{i_m}^c \in \{F_i^c\} \quad \text{such that} \quad X = F_{i_1}^c \cup \cdots \cup F_{i_m}^c$$

Thus by DeMorgan's Law,

$$\emptyset = X^c = (F_{i_1}^c \cup \cdots \cup F_{i_m}^c)^c = F_{i_1}^{cc} \cap \cdots \cap F_{i_m}^{cc} = F_{i_1} \cap \cdots \cap F_{i_m}$$

and we have shown that (i) $\Rightarrow$ (ii).

(ii) $\Rightarrow$ (i): Let $\{G_i\}$ be an open cover of X, i.e. $X = \cup_i G_i$. By DeMorgan's Law,

$$\emptyset = X^c = (\cup_i G_i)^c = \cap_i G_i^c$$

Since each G_i is open, $\{G_i^c\}$ is a class of closed sets and, by above, has an empty intersection. Hence by hypothesis,

$$\exists\ G_{i_1}^c, \ldots, G_{i_m}^c \in \{G_i^c\} \quad \text{such that} \quad G_{i_1}^c \cap \cdots \cap G_{i_m}^c = \emptyset$$

Thus by DeMorgan's Law,

$$X = \emptyset^c = (G_{i_1}^c \cap \cdots \cap G_{i_m}^c)^c = G_{i_1}^{cc} \cup \cdots \cup G_{i_m}^{cc} = G_{i_1} \cup \cdots \cup G_{i_m}$$

Accordingly, X is compact and so (ii) $\Rightarrow$ (i).

6. Prove Theorem 11.4: A topological space X is compact if and only if every class $\{F_i\}$ of closed subsets of X which satisfies the finite intersection property has, itself, a non-empty intersection.

Solution:

Utilizing the preceding problem, it suffices to show that the following statements are equivalent, where $\{F_i\}$ is any class of closed subsets of X:

(i) $$F_{i_1} \cap \cdots \cap F_{i_m} \neq \emptyset \quad \forall\, i_1, \ldots, i_m \quad \Rightarrow \quad \cap_i F_i \neq \emptyset$$

(ii) $$\cap_i F_i = \emptyset \quad \Rightarrow \quad \exists\, i_1, \ldots, i_m \text{ s.t. } F_{i_1} \cap \cdots \cap F_{i_m} = \emptyset$$

But these statements are contrapositives.

COMPACTNESS AND HAUSDORFF SPACES

7. Prove: Let A be a compact subset of a Hausdorff space X and suppose $p \in X \setminus A$. Then

$$\exists \text{ open sets } G, H \quad \text{such that} \quad p \in G,\ A \subset H,\ G \cap H = \emptyset$$

Solution:

Let $a \in A$. Since $p \notin A$, $p \neq a$. By hypothesis, X is Hausdorff; hence

$$\exists \text{ open sets } G_a, H_a \quad \text{such that} \quad p \in G_a,\ a \in H_a,\ G_a \cap H_a = \emptyset$$

Hence $A \subset \bigcup\{H_a : a \in A\}$, i.e. $\{H_a : a \in A\}$ is an open cover of A. But A is compact, so

$$\exists\ H_{a_1}, \ldots, H_{a_m} \in \{H_a\} \quad \text{such that} \quad A \subset H_{a_1} \cup \cdots \cup H_{a_m}$$

Now let $H = H_{a_1} \cup \cdots \cup H_{a_m}$ and $G = G_{a_1} \cap \cdots \cap G_{a_m}$. H and G are open since they are respectively the union and finite intersection of open sets. Furthermore, $A \subset H$ and $p \in G$ since p belongs to each G_{a_i} individually.

Lastly we claim that $G \cap H = \emptyset$. Note first that $G_{a_i} \cap H_{a_i} = \emptyset$ implies that $G \cap H_{a_i} = \emptyset$. Thus, by the distributive law,

$$G \cap H = G \cap (H_{a_1} \cup \cdots \cup H_{a_m}) = (G \cap H_{a_1}) \cup \cdots \cup (G \cap H_{a_m}) = \emptyset \cup \cdots \cup \emptyset = \emptyset$$

Thus the proof is complete.

8. Let A be a compact subset of a Hausdorff space X. Show that if $p \notin A$, then there is an open set G such that $p \in G \subset A^c$.

Solution:

By Problem 7 there exist open sets G and H such that $p \in G$, $A \subset H$ and $G \cap H = \emptyset$.

Hence $G \cap A = \emptyset$, and $p \in G \subset A^c$.

9. Prove Theorem 11.5: Let A be a compact subset of a Hausdorff space X. Then A is closed.

Solution:

We prove, equivalently, that A^c is open. Let $p \in A^c$, i.e. $p \notin A$. Then by Problem 8 there exists an open set G_p such that $p \in G_p \subset A^c$. Hence $A^c = \bigcup\{G_p : p \in A^c\}$.

Thus A^c is open as it is the union of open sets, or, A is closed.

10. Prove Theorem 11.6: Let A and B be disjoint compact subsets of a Hausdorff space X. Then there exist disjoint open sets G and H such that $A \subset G$ and $B \subset H$.

Solution:

Let $a \in A$. Then $a \notin B$, for A and B are disjoint. By hypothesis, B is compact; hence by Problem 1 there exist open sets G_a and H_a such that

$$a \in G_a, \quad B \subset H_a \quad \text{and} \quad G_a \cap H_a = \emptyset$$

Since $a \in G_a$, $\{G_a : a \in A\}$ is an open cover of A. Since A is compact, we can select a finite number of the open sets, say $G_{a_1}, \ldots, G_{a_m}$, so that $A \subset G_{a_1} \cup \cdots \cup G_{a_m}$. Furthermore, $B \subset H_{a_1} \cap \cdots \cap H_{a_m}$ since B is a subset of each individually.

Now let $G = G_{a_1} \cup \cdots \cup G_{a_m}$ and $H = H_{a_1} \cap \cdots \cap H_{a_m}$. Observe, by the above, that $A \subset G$ and $B \subset H$. In addition, G and H are open as they are the union and finite intersection respectively of open sets. The theorem is proven if we show that G and H are disjoint. First observe that, for each i, $G_{a_i} \cap H_{a_i} = \emptyset$ implies $G_{a_i} \cap H = \emptyset$. Hence, by the distributive law,

$$G \cap H = (G_{a_1} \cup \cdots \cup G_{a_m}) \cap H = (G_{a_1} \cap H) \cup \cdots \cup (G_{a_m} \cap H) = \emptyset \cup \cdots \cup \emptyset = \emptyset$$

Thus the theorem is proven.

11. Prove Theorem 11.8: Let f be a one-one continuous function from a compact space X into a Hausdorff space Y. Then X and $f[X]$ are homeomorphic.

Solution:

Now $f : X \to f[X]$ is onto and, by hypothesis, one-one and continuous, so $f^{-1} : f[X] \to X$ exists. We must show that f^{-1} is continuous. Recall that f^{-1} is continuous if, for every closed subset F of X, $(f^{-1})^{-1}[F] = f[F]$ is a closed subset of $f[X]$. By Theorem 11.3, the closed subset F of the compact space X is also compact. Since f is continuous, $f[F]$ is a compact subset of $f[X]$. But the subspace $f[X]$ of the Hausdorff space Y is also Hausdorff; hence by Theorem 11.5, $f[F]$ is closed. Accordingly, f^{-1} is continuous, so $f : X \to f[X]$ is a homeomorphism, and X and $f[X]$ are homeomorphic.

12. Let $(X, \mathcal{T})$ be compact and let $(X, \mathcal{T}^*)$ be Hausdorff. Show that if $\mathcal{T}^* \subset \mathcal{T}$, then $\mathcal{T}^* = \mathcal{T}$.

Solution:

Consider the function $f : (X, \mathcal{T}) \to (X, \mathcal{T}^*)$ defined by $f(x) = x$, i.e. the identity function on X. Now f is one-one and onto. Furthermore, f is continuous since $\mathcal{T}^* \subset \mathcal{T}$. Thus by the preceding problem, f is a homeomorphism and therefore $\mathcal{T}^* = \mathcal{T}$.

SEQUENTIALLY AND COUNTABLY COMPACT SETS

13. Show that a continuous image of a sequentially compact set is sequentially compact.

Solution:

Let $f : X \to Y$ be a continuous function and let A be a sequentially compact subset of X. We want to show that $f[A]$ is a sequentially compact subset of Y. Let $\langle b_1, b_2, \ldots \rangle$ be a sequence in $f[A]$. Then

$$\exists\ a_1, a_2, \ldots \in A \quad \text{such that} \quad f(a_n) = b_n, \quad \forall\ n \in \mathbf{N}$$

But A is sequentially compact, so the sequence $\langle a_1, a_2, \ldots \rangle$ contains a subsequence $\langle a_{i_1}, a_{i_2}, \ldots \rangle$ which converges to a point $a_0 \in A$. Now f is continuous and hence sequentially continuous, so

$$\langle f(a_{i_1}), f(a_{i_2}), \ldots \rangle = \langle b_{i_1}, b_{i_2}, \ldots \rangle \quad \text{converges to} \quad f(a_0) \in f[A]$$

Thus $f[A]$ is sequentially compact.

14. Let $\mathcal{T}$ be the topology on X which consists of $\emptyset$ and the complements of countable subsets of X. Show that every infinite subset of X is not sequentially compact.

Solution:

Recall (Example 7.3, Page 71) that a sequence in $(X, \mathcal{T})$ converges iff it is of the form

$$\langle a_1, a_2, \ldots, a_{n_0}, p, p, p, \ldots \rangle$$

that is, is constant from some term on. Hence if A is an infinite subset of X, there exists a sequence $\langle b_n \rangle$ in A with distinct terms. Thus $\langle b_n \rangle$ does not contain any convergent subsequence, and A is not sequentially compact.

15. Show that: (i) a continuous image of a countably compact set need not be countably compact; (ii) a closed subset of a countably compact space is countably compact.

Solution:

(i) Let $X = (\mathbf{N}, \mathcal{T})$ where $\mathcal{T}$ is the topology on the positive integers $\mathbf{N}$ generated by the sets $\{1,2\}, \{3,4\}, \{5,6\}, \ldots$. By Example 6.3, X is countably compact. Let $Y = (\mathbf{N}, \mathcal{D})$ where $\mathcal{D}$ is the discrete topology on $\mathbf{N}$. Now Y is not countably compact. On the other hand, the function $f: X \to Y$ which maps $2n$ and $2n-1$ onto n for $n \in \mathbf{N}$ is continuous and maps the countably compact set X onto the non-countably compact set Y.

(ii) Suppose X is countably compact and suppose F is a closed subset of X. Let A be an infinite subset of F. Since $F \subset X$, A is also an infinite subset of X. By hypothesis, X is countably compact; then A has an accumulation point $p \in X$. Since $A \subset F$, p is also an accumulation point of F. But F is closed and so contains its accumulation points; hence $p \in F$. We have shown that any infinite subset A of F has an accumulation point $p \in F$, that is, that F is countably compact.

16. Prove: Let X be compact. Then X is also countably compact.

Solution:

Let A be a subset of X with no accumulation points in X. Then each point $p \in X$ belongs to an open set G_p which contains at most one point of A. Observe that the class $\{G_p : p \in X\}$ is an open cover of the compact set X and, hence, contains a finite subcover, say $\{G_{p_1}, \ldots, G_{p_m}\}$.

Hence $$A \subset X \subset G_{p_1} \cup \cdots \cup G_{p_m}$$

But each G_{p_i} contains at most one point of A; hence A, a subset of $G_{p_1} \cup \cdots \cup G_{p_m}$, can contain at most m points, i.e. A is finite. Accordingly, every infinite subset of X contains an accumulation point in X, i.e. X is countably compact.

17. Prove: Let X be sequentially compact. Then X is also countably compact.

Solution:

Let A be any infinite subset of X. Then there exists a sequence $\langle a_1, a_2, \ldots \rangle$ in A with distinct terms. Since X is sequentially compact, the sequence $\langle a_n \rangle$ contains a subsequence $\langle a_{i_1}, a_{i_2}, \ldots \rangle$ (also with distinct terms) which converges to a point $p \in X$. Hence every open neighborhood of p contains an infinite number of the terms of the convergent subsequence $\langle a_{i_n} \rangle$. But the terms are distinct; hence every open neighborhood of p contains an infinite number of points in A. Accordingly, $p \in X$ is an accumulation point of A. In other words, X is countably compact.

Remark: Note that Problems 16 and 17 imply Theorem 11.9.

18. Prove: Let $A \subset X$ be sequentially compact. Then every countable open cover of A is reducible to a finite cover.

Solution:

We may assume A is infinite, for otherwise the proof is trivial. We prove the contrapositive, i.e. assume $\exists$ a countable open cover $\{G_i : i \in \mathbf{N}\}$ with no finite subcover. We define the sequence $\langle a_1, a_2, \ldots \rangle$ as follows.

Let n_1 be the smallest positive integer such that $A \cap G_{n_1} \neq \emptyset$. Choose $a_1 \in A \cap G_{n_1}$. Let n_2 be the least positive integer larger than n_1 such that $A \cap G_{n_2} \neq \emptyset$. Choose

$$a_2 \in (A \cap G_{n_2}) \setminus (A \cap G_{n_1})$$

Such a point always exists, for otherwise G_{n_1} covers A. Continuing in this manner, we obtain the sequence $\langle a_1, a_2, \ldots \rangle$ with the property that, for every $i \in \mathbf{N}$,

$$a_i \in A \cap G_{n_i}, \quad a_i \notin \cup_{j=1}^{n-1} (A \cap G_{n_j}) \quad \text{and} \quad n_i > n_{i-1}$$

We claim that $\langle a_i \rangle$ has no convergent subsequence in A. Let $p \in A$. Then

$$\exists\ G_{i_0} \in \{G_i\} \quad \text{such that} \quad p \in G_{i_0}$$

Now $A \cap G_{i_0} \neq \emptyset$, since $p \in A \cap G_{i_0}$; hence

$$\exists\ j_0 \in \mathbf{N} \quad \text{such that} \quad G_{n_{j_0}} = G_{i_0}$$

But by the choice of the sequence $\langle a_1, a_2, \ldots \rangle$

$$i > j_0 \quad \Rightarrow \quad a_i \notin G_{i_0}$$

Accordingly, since G_{i_0} is an open set containing p, no subsequence of $\langle a_i \rangle$ converges to p. But p was arbitrary, so A is not sequentially compact.

COMPACTNESS IN METRIC SPACES

19. Prove Lemma 11.17: Let A be a sequentially compact subset of a metric space X. Then A is totally bounded.

Solution:

We prove the contrapositive of the above statement, i.e. if A is not totally bounded, then A is not sequentially compact. If A is not totally bounded then there exists an $\epsilon > 0$ such that A possesses no (finite) ϵ-net. Let $a_1 \in A$. Then there exists a point $a_2 \in A$ with $d(a_1, a_2) \geqq \epsilon$, for otherwise $\{a_1\}$ would be an ϵ-net for A. Similarly, there exists a point $a_3 \in A$ with $d(a_1, a_3) \geqq \epsilon$ and $d(a_2, a_3) \geqq \epsilon$, for otherwise $\{a_1, a_2\}$ would be an ϵ-net for A. Continuing in this manner, we arrive at a sequence $\langle a_1, a_2, \ldots \rangle$ with the property that $d(a_i, a_j) \geqq \epsilon$ for $i \neq j$. Thus the sequence $\langle a_n \rangle$ cannot contain any subsequence which converges. In other words, A is not sequentially compact.

20. Prove Lemma (Lebesgue) 11.18: Let $\mathcal{A} = \{G_i\}$ be an open cover of a sequentially compact set A. Then $\mathcal{A}$ has a (positive) Lebesgue number.

Solution:

Suppose $\mathcal{A}$ does not have a Lebesgue number. Then for each positive integer $n \in \mathbf{N}$ there exists a subset B_n of A with the property that

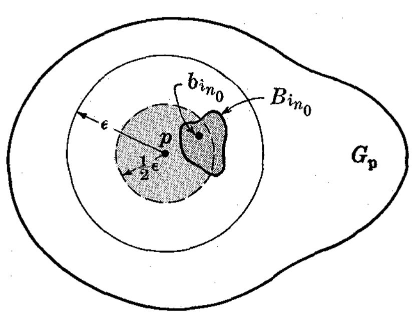
$$0 < d(B_n) < 1/n \quad \text{and} \quad B_n \not\subset G_i \quad \text{for every } G_i \text{ in } \mathcal{A}$$

For each $n \in \mathbf{N}$, choose a point $b_n \in B_n$. Since A is sequentially compact, the sequence $\langle b_1, b_2, \ldots \rangle$ contains a subsequence $\langle b_{i_1}, b_{i_2}, \ldots \rangle$ which converges to a point $p \in A$.

Since $p \in A$, p belongs to an open set G_p in the cover $\mathcal{A}$. Hence there exists an open sphere $S(p, \epsilon)$, with center p and radius ϵ, such that $p \in S(p, \epsilon) \subset G_p$. Since $\langle b_{i_n} \rangle$ converges to p, there exists a positive integer i_{n_0} such that

$$d(p, b_{i_{n_0}}) < \tfrac{1}{2}\epsilon, \quad b_{i_{n_0}} \in B_{i_{n_0}} \quad \text{and} \quad d(B_{i_{n_0}}) < \tfrac{1}{2}\epsilon$$

Using the Triangle Inequality we get $B_{i_{n_0}} \subset S(p, \epsilon) \subset G_p$. But this contradicts the fact that $B_{i_{n_0}} \not\subset G_i$ for every G_i in the cover $\mathcal{A}$. Accordingly $\mathcal{A}$ does possess a Lebesgue number.



$B_{i_{n_0}}$ is shaded

21. Prove: Let A be a countably compact subset of a metric space X. Then A is also sequentially compact.

Solution:

Let $\langle a_1, a_2, \ldots\rangle$ be a sequence in A. If the set $B = \{a_1, a_2, \ldots\}$ is finite, then one of the points, say a_{i_0}, satisfies $a_{i_0} = a_j$ for infinitely many $j \in \mathbf{N}$. Hence $\langle a_{i_0}, a_{i_0}, \ldots\rangle$ is a subsequence of $\langle a_n\rangle$ which converges to the point a_{i_0} in A.

On the other hand, suppose $B = \{a_1, a_2, \ldots\}$ is infinite. By hypothesis, A is countably compact. Hence the infinite subset B of A contains an accumulation point p in A. But X is a metric space; hence we can choose a subsequence $\langle a_{i_1}, a_{i_2}, \ldots\rangle$ of the sequence $\langle a_n\rangle$ which converges to the point p in A. In other words, A is sequentially compact.

22. Prove Theorem 11.14: Let A be a subset of a metric space X. Then the following are equivalent: (i) A is compact, (ii) A is countably compact, and (iii) A is sequentially compact.

Solution:

Recall (see Theorem 11.8) that (i) implies (ii) in every topological space; hence it is true for a metric space. In the preceding problem we proved that (ii) implies (iii). Accordingly, the theorem is proven if we show that (iii) implies (i).

Let A be sequentially compact, and let $\mathcal{A} = \{G_i\}$ be an open cover of A. We want to show that A is compact, i.e. that $\mathcal{A}$ possesses a finite subcover. By hypothesis, A is sequentially compact; hence, by Lemma 11.18, the cover $\mathcal{A}$ possesses a Lebesgue number $\delta > 0$. In addition, by Lemma 11.17, A is totally bounded. Hence there is a decomposition of A into a finite number of subsets, say $B_1, \ldots, B_m$, with $d(B_i) < \delta$. But δ is a Lebesgue number for $\mathcal{A}$; hence there are open sets $G_{i_1}, \ldots, G_{i_m} \in \mathcal{A}$ such that

$$B_1 \subset G_{i_1}, \quad \ldots, \quad B_m \subset G_{i_m}$$

Accordingly,
$$A \subset B_1 \cup B_2 \cup \cdots \cup B_m \subset G_{i_1} \cup G_{i_2} \cup \cdots \cup G_{i_m}$$

Thus $\mathcal{A}$ possesses a finite subcover $\{G_{i_1}, \ldots, G_{i_m}\}$, i.e. A is compact.

23. Let A be a compact subset of a metric space (X, d). Show that for any $B \subset X$ there is a point $p \in A$ such that $d(p, B) = d(A, B)$.

Solution:

Let $d(A, B) = \epsilon$. Since $d(A, B) = \inf\{d(a, b) : a \in A,\ b \in B\}$, for every positive integer $n \in \mathbf{N}$,

$$\exists\ a_n \in A,\ b_n \in B \quad \text{such that} \quad \epsilon \leqq d(a_n, b_n) < \epsilon + 1/n$$

Now A is compact and hence sequentially compact; so the sequence $\langle a_1, a_2, \ldots\rangle$ has a subsequence which converges to a point $p \in A$. We claim that $d(p, B) = d(A, B) = \epsilon$.

Suppose $d(p, B) > \epsilon$, say $d(p, B) = \epsilon + \delta$ where $\delta > 0$. Since a subsequence of $\langle a_n\rangle$ converges to p,

$$\exists\ n_0 \in \mathbf{N} \quad \text{such that} \quad d(p, a_{n_0}) < \tfrac{1}{2}\delta \quad \text{and} \quad d(a_{n_0}, b_{n_0}) < \epsilon + 1/n_0 < \epsilon + \tfrac{1}{2}\delta$$

Then
$$d(p, a_{n_0}) + d(a_{n_0}, b_{n_0}) < \tfrac{1}{2}\delta + \epsilon + \tfrac{1}{2}\delta = \epsilon + \delta = d(p, B) \leqq d(p, b_{n_0})$$

But this contradicts the Triangle Inequality; hence $d(p, B) = d(A, B)$.

24. Let A be a compact subset of a metric space (X, d) and let B be a closed subset of X such that $A \cap B = \emptyset$. Show that $d(A, B) > 0$.

Solution:

Suppose $d(A, B) = 0$. Then, by the preceding problem,

$$\exists\ p \in A \quad \text{such that} \quad d(p, B) = d(A, B) = 0$$

But B is closed and therefore contains all points whose distance from B is zero. Thus $p \in B$ and so $p \in A \cap B$. But this contradicts the hypothesis; hence $d(A, B) > 0$.

25. Prove: Let f be a continuous function from a compact metric space (X, d) into a metric space (Y, d^*). Then f is *uniformly continuous*, i.e. for every $\epsilon > 0$ there exists a $\delta > 0$ such that

$$d(x, y) < \delta \quad \Rightarrow \quad d^*(f(x), f(y)) < \epsilon$$

(*Remark*: Uniform continuity is a stronger condition than continuity, in that the δ above depends only upon the ϵ and not also on any particular point.)

Solution:

Let $\epsilon > 0$. Since f is continuous, for each point $p \in X$ there exists an open sphere $S(p, \delta_p)$ such that

$$x \in S(p, \delta_p) \quad \Rightarrow \quad f(x) \in S(f(p), \tfrac{1}{2}\epsilon)$$

Observe that the class $\mathcal{A} = \{S(p, \delta_p) : p \in X\}$ is an open cover of X. By hypothesis, X is compact and hence also sequentially compact. Therefore the cover $\mathcal{A}$ possesses a Lebesgue number $\delta > 0$.

Now let $x, y \in X$ with $d(x, y) < \delta$. But $d(x, y) = d\{x, y\} < \delta$ implies $\{x, y\}$ is contained in a member $S(p_0, \delta_{p_0})$ of the cover A. Now

$$x, y \in S(p_0, \delta_{p_0}) \quad \Rightarrow \quad f(x), f(y) \in S(f(p_0), \tfrac{1}{2}\epsilon)$$

But the sphere $S(f(p_0), \frac{1}{2}\epsilon)$ has diameter ϵ. Accordingly,

$$d(x, y) < \delta \quad \Rightarrow \quad d^*(f(x), f(y)) < \epsilon$$

In other words, f is uniformly continuous.

Supplementary Problems

COMPACT SPACES

26. Prove: If E is compact and F is closed, then $E \cap F$ is compact.

27. Let $A_1, \ldots, A_m$ be compact subsets of a topological space X. Show that $A_1 \cup \cdots \cup A_m$ is also compact.

28. Prove that compactness is a topological property.

29. Prove Proposition 11.11: The class $\mathcal{T}_\infty$ is a topology on X_∞ and $(X_\infty, \mathcal{T}_\infty)$ is a compactification of $(X, \mathcal{T})$. (Here $(X_\infty, \mathcal{T}_\infty)$ is the Alexandrov one-point compactification of $(X, \mathcal{T})$.)

30. Prove Theorem 11.12: If $(X, \mathcal{T})$ is a locally compact Hausdorff space, then $(X_\infty, \mathcal{T}_\infty)$ is a compact Hausdorff space.

SEQUENTIALLY AND COUNTABLY COMPACT SPACES

31. Show that sequential compactness is a topological property.

32. Prove: A closed subset of a sequentially compact space is sequentially compact.

33. Show that countable compactness is a topological property.

34. Suppose $(X, \mathcal{T})$ is countably compact and $\mathcal{T}^* \leq \mathcal{T}$. Show that $(X, \mathcal{T}^*)$ is also countably compact.

35. Prove: Let X be a topological space such that every countable open cover of X is reducible to a finite cover. Then X is countably compact.

36. Prove: Let X be a T_1-space. Then X is countably compact if and only if every countable open cover of X is reducible to a finite cover.

37. Prove: Let X be a second countable T_1-space. Then X is compact if and only if X is countably compact.

TOTALLY BOUNDED SETS

38. Prove Proposition 11.15: A set A is totally bounded if and only if for every $\epsilon > 0$ there exists a decomposition of A into a finite number of sets each with diameter less than ϵ.

39. Prove Proposition 11.16: Totally bounded sets are bounded.

40. Show that every subset of a totally bounded set is totally bounded.

41. Show that if A is totally bounded then $\bar{A}$ is also totally bounded.

42. Prove: Every totally bounded metric space is separable.

COMPACTNESS AND METRIC SPACES

43. Prove: A compact subset of a metric space X is closed and bounded.

44. Prove: Let $f : X \to Y$ be a continuous function from a compact space X into a metric space Y. Then $f[X]$ is a bounded subset of Y.

45. *Prove: A subset A of the real line R is compact if and only if A is closed and bounded.*

46. Prove: Let A be a compact subset of a metric space X. Then the derived set A' of A is compact.

47. Prove: The Hilbert cube $\mathbf{I} = \{\langle a_n \rangle : 0 \leq a_n \leq 1/n\}$ is a compact subset of $\mathbf{R}^\infty$.

48. Prove: Let A and B be compact subsets of a metric space X. Then there exist $a \in A$ and $b \in B$ such that $d(a, b) = d(A, B)$.

LOCALLY COMPACT SPACES

49. Show that local compactness is a topological property.

50. Show that every discrete space is locally compact.

51. Show that every indiscrete space is locally compact.

52. Show that the plane $\mathbf{R}^2$ with the usual topology is locally compact.

53. Prove: Let A be a closed subset of a locally compact space $(X, \mathcal{T})$. Then A with the relative topology is locally compact.

Chapter 13

Connectedness

SEPARATED SETS

Two subsets A and B of a topological space X are said to be *separated* if (i) A and B are disjoint, and (ii) neither contains an accumulation point of the other. In other words, A and B are separated iff

$$A \cap \bar{B} = \emptyset \quad \text{and} \quad \bar{A} \cap B = \emptyset$$

Example 1.1: Consider the following intervals on the real line $\mathbf{R}$:

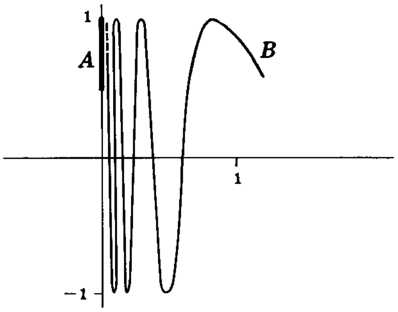
$$A = (0,1), \quad B = (1,2) \quad \text{and} \quad C = [2,3)$$

Now A and B are separated since $\bar{A} = [0,1]$ and $\bar{B} = [1,2]$, and so $A \cap \bar{B}$ and $\bar{A} \cap B$ are empty. On the other hand, B and C are not separated since $2 \in C$ is a limit point of B; thus:

$$\bar{B} \cap C = [1,2] \cap [2,3) = \{2\} \neq \emptyset$$

Example 1.2: Consider the following subsets of the plane $\mathbf{R}^2$:

$$A = \{\langle 0,y\rangle : \tfrac{1}{2} \leq y \leq 1\}$$
$$B = \{\langle x,y\rangle : y = \sin(1/x),\ 0 < x \leq 1\}$$



Now each point in A is an accumulation point of B; hence A and B are not separated sets.

CONNECTED SETS

Definition: A subset A of a topological space X is *disconnected* if there exist open subsets G and H of X such that $A \cap G$ and $A \cap H$ are disjoint non-empty sets whose union is A. In this case, $G \cup H$ is called a *disconnection* of A. A set is *connected* if it is not disconnected.

Observe that $$A = (A \cap G) \cup (A \cap H) \quad \text{iff} \quad A \subset G \cup H$$

and $$\emptyset = (A \cap G) \cap (A \cap H) \quad \text{iff} \quad G \cap H \subset A^c$$

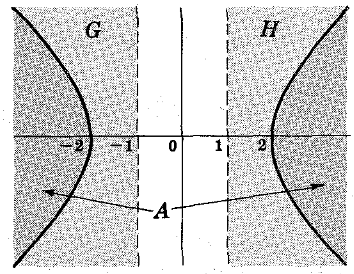
Therefore $G \cup H$ is a disconnection of A if and only if

$$A \cap G \neq \emptyset, \quad A \cap H \neq \emptyset, \quad A \subset G \cup H, \quad \text{and} \quad G \cap H \subset A^c$$

Note that the empty set $\emptyset$ and singleton sets $\{p\}$ are always connected.

Example 2.1: The following subset of the plane $\mathbf{R}^2$ is disconnected:

$$A = \{\langle x, y\rangle : x^2 - y^2 \geqq 4\}$$



For the two open half-planes

$$G = \{\langle x,y\rangle : x < -1\} \quad \text{and} \quad H = \{\langle x,y\rangle : x > 1\}$$

form a disconnection of A as indicated in the diagram above.

Example 2.2: Consider the following topology on $X = \{a, b, c, d, e\}$:

$$\mathcal{T} = \{X, \emptyset, \{a,b,c\}, \{c,d,e\}, \{c\}\}$$

Now $A = \{a, d, e\}$ is disconnected. For let $G = \{a,b,c\}$ and $H = \{c,d,e\}$; then $A \cap G = \{a\}$ and $A \cap H = \{d, e\}$ are non-empty disjoint sets whose union is A. (Observe that G and H are not disjoint.)

The basic relationship between connectedness and separation follows:

Theorem 13.1: A set is connected if and only if it is not the union of two non-empty separated sets.

The following proposition is very useful.

Proposition 13.2: If A and B are connected sets which are not separated, then $A \cup B$ is connected.

Example 2.3: Let A and B be the subsets of the plane $\mathbf{R}^2$ defined and illustrated in Example 1.2. We show later that A and B are each connected. But A and B are not separated; hence, by the previous proposition, $A \cup B$ is a connected set.

CONNECTED SPACES

Connectedness, like compactness, is an absolute property of a set; namely,

Theorem 13.3: Let A be a subset of a topological space $(X, \mathcal{T})$. Then A is connected with respect to $\mathcal{T}$ if and only if A is connected with respect to the relative topology $\mathcal{T}_A$ on A.

Accordingly, we can frequently limit our investigation of connectedness to those topological spaces which are themselves connected, i.e. to *connected spaces*.

Example 3.1: Let X be a topological space which is disconnected, and let $G \cup H$ be a disconnection of X; then

$$X = (X \cap G) \cup (X \cap H) \quad \text{and} \quad (X \cap G) \cap (X \cap H) = \emptyset$$

But $X \cap G = G$ and $X \cap H = H$; thus X is disconnected if and only if there exist non-empty open sets G and H such that

$$X = G \cup H \quad \text{and} \quad G \cap H = \emptyset$$

In view of the discussion in the above example, we can give a simple characterization of connected spaces.

Theorem 13.4: A topological space X is connected if and only if (i) X is not the union of two non-empty disjoint open sets; or, equivalently, (ii) X and $\emptyset$ are the only subsets of X which are both open and closed.

Example 3.2: Consider the following topology on $X = \{a, b, c, d, e\}$:

$$\mathcal{T} = \{X, \emptyset, \{a\}, \{c, d\}, \{a, c, d\}, \{b, c, d, e\}\}$$

Now X is disconnected; for $\{a\}$ and $\{b, c, d, e\}$ are complements and hence both open and closed. In other words,

$$X = \{a\} \cup \{b, c, d, e\}$$

is a disconnection of X. Observe that the relative topology on the subset $A = \{b, d, e\}$ is $\{A, \emptyset, \{d\}\}$. Accordingly, A is connected since A and $\emptyset$ are the only subsets of A both open and closed in the relative topology.

Example 3.3: The real line $\mathbf{R}$ with the usual topology is a connected space since $\mathbf{R}$ and $\emptyset$ are the only subsets of $\mathbf{R}$ which are both open and closed.

Example 3.4: Let f be a continuous function from a connected space X into a topological space Y. Thus $f: X \to f[X]$ is continuous (where $f[X]$ has the relative topology).

We show that $f[X]$ is connected. Suppose $f[X]$ is disconnected; say G and H form a disconnection of $f[X]$. Then

$$f[X] = G \cup H \quad \text{and} \quad G \cap H = \emptyset$$

and so $$X = f^{-1}[G] \cup f^{-1}[H] \quad \text{and} \quad f^{-1}[G] \cap f^{-1}[H] = \emptyset$$

Since f is continuous, $f^{-1}[G]$ and $f^{-1}[H]$ are open subsets of X and hence form a disconnection of X, which is impossible. Thus if X is connected, so is $f[X]$.

We state the result of the preceding example as a theorem.

Theorem 13.5: Continuous images of connected sets are connected.

Example 3.5: Let X be a disconnected space; say, $G \cup H$ is a disconnection of X. Then the function $f(x) = \begin{cases} 0 & \text{if } x \in G \\ 1 & \text{if } x \in H \end{cases}$ is a continuous function from X onto the discrete space $Y = \{0, 1\}$.

On the other hand, by Theorem 13.5, a continuous image of a connected space X cannot be the disconnected discrete space $Y = \{0, 1\}$. In other words,

Lemma 13.6: A topological space X is connected if and only if the only continuous functions from X into $Y = \{0, 1\}$ are the constant functions, $f(x) = 0$ or $f(x) = 1$.

CONNECTEDNESS ON THE REAL LINE

The connected sets of real numbers can be simply described as follows:

Theorem 13.7: A subset E of the real line $\mathbf{R}$ containing at least two points is connected if and only if E is an interval.

Recall that the intervals on the real line $\mathbf{R}$ are of the following form:

$$(a, b),\ (a, b],\ [a, b),\ [a, b], \qquad \text{finite intervals}$$

$$(-\infty, a),\ (-\infty, a],\ (a, \infty),\ [a, \infty),\ (-\infty, \infty), \qquad \text{infinite intervals}$$

An interval E can be characterized by the following property:

$$a,\ b \in E,\ a < x < b \quad \Rightarrow \quad x \in E$$

Since the continuous image of a connected set is connected, we have the following generalization of the Weierstrass Intermediate Value Theorem (see Page 53, Theorem 4.9):

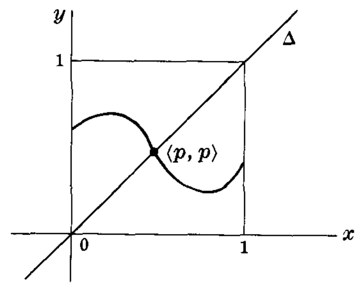
Theorem 13.8: Let $f: X \to \mathbf{R}$ be a real continuous function defined on a connected set X. Then f assumes as a value each number between any two of its values.

Example 4.1: An interesting application of the theory of connectedness is the following "fixed-point theorem": Let $I = [0,1]$ and let $f : I \to I$ be continuous; then $\exists p \in I$ such that $f(p) = p$.

This theorem can be interpreted geometrically. Note first that the graph of $f : I \to I$ lies in the unit square

$$I^2 = \{\langle x, y\rangle : 0 \leq x \leq 1,\ 0 \leq y \leq 1\}$$

The theorem then states that the graph of f, which connects a point on the left edge of the square to a point on the right edge of the square, must intersect the diagonal line Δ at, say, $\langle p, p\rangle$ as indicated in the diagram.



COMPONENTS

A *component* E of a topological space X is a maximal connected subset of X; that is E is connected and E is not a proper subset of any connected subset of X. Clearly E is non-empty. The central facts about the components of a space are contained in the following theorem.

Theorem 13.9: The components of a topological space X form a partition of X, i.e. they are disjoint and their union is X. Every connected subset of X is contained in some component.

Thus each point $p \in X$ belongs to a unique component of X, called the *component of* p.

Example 5.1: If X is connected, then X has only one component: X itself.

Example 5.2: Consider the following topology on $X = \{a, b, c, d, e\}$:

$$\mathcal{T} = \{X,\ \emptyset,\ \{a\},\ \{c,d\},\ \{a,c,d\},\ \{b,c,d,e\}\}$$

The components of X are $\{a\}$ and $\{b,c,d,e\}$. Any other connected subset of X, such as $\{b,d,e\}$ (see Example 3.2), is a subset of one of the components.

The statement in Example 5.1 is used to prove that connectedness is product invariant; that is,

Theorem 13.10: The product of connected spaces is connected.

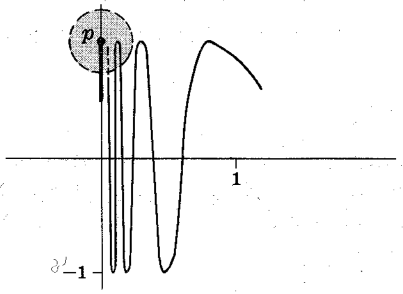
Corollary 13.11: Euclidean m-space $\mathbf{R}^m$ is connected.

LOCALLY CONNECTED SPACES

A topological space X is *locally connected at* $p \in X$ iff every open set containing p contains a connected open set containing p, i.e. if the open connected sets containing p form a local base at p. X is said to be *locally connected* if it is locally connected at each of its points or, equivalently, if the open connected subsets of X form a base for X.

Example 6.1: Every discrete space X is locally connected. For if $p \in X$, then $\{p\}$ is an open connected set containing p which is contained in every open set containing p. Note that X is not connected if X contains more than one point.

Example 6.2: Let A and B be the subsets of the plane $\mathbf{R}^2$ of Example 1.2. Now $A \cup B$ is a connected set. But $A \cup B$ is not locally connected at $p = \langle 0, 1\rangle$. For example, the open disc with center p and radius $\frac{1}{4}$ does not contain any connected neighborhood of p.



PATHS

Let $I = [0,1]$, the closed unit interval. A *path* from a point a to a point b in a topological space X is a continuous function $f: I \to X$ with $f(0) = a$ and $f(1) = b$. Here a is called the *initial point* and b is called the *terminal point* of the path.

Example 7.1: For any $p \in X$, the constant function $e_p: I \to X$ defined by $e_p(s) = p$ is continuous and hence a path. It is called the *constant path* at p.

Example 7.2: Let $f: I \to X$ be a path from a to b. Then the function $\hat{f}: I \to X$ defined by $\hat{f}(s) = f(1-s)$ is a path from b to a.

Example 7.3: Let $f: I \to X$ be a path from a to b and let $g: I \to X$ be a path from b to c. Then the juxtaposition of the two paths f and g, denoted by $f * g$, is the function $f * g: I \to X$ defined by

$$(f * g)(s) = \begin{cases} f(2s) & \text{if } 0 \leq s \leq \frac{1}{2} \\ g(2s-1) & \text{if } \frac{1}{2} \leq s \leq 1 \end{cases}$$

which is a path from a to c obtained by following the path f from a to b and then following g from b to c.

ARCWISE CONNECTED SETS

A subset E of a topological space X is said to be *arcwise connected* if for any two points $a, b \in E$ there is a path $f: I \to X$ from a to b which is contained in E, i.e. $f[I] \subset E$. The maximal arcwise connected subsets of X, called *arcwise connected components*, form a partition of X. The relationship between connectedness and arcwise connectedness follows:

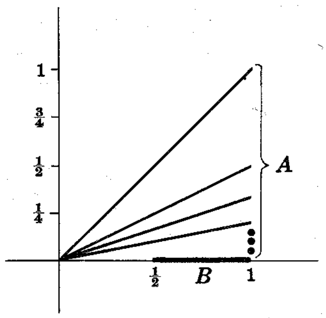
Theorem 13.12: Arcwise connected sets are connected.

The converse of this theorem is not true, as seen in the next example.

Example 8.1: Consider the following subsets of the plane $\mathbf{R}^2$:

$$A = \{\langle x, y\rangle : 0 \leq x \leq 1,\ y = x/n,\ n \in \mathbf{N}\}$$
$$B = \{\langle x, 0\rangle : \tfrac{1}{2} \leq x \leq 1\}$$

Here A consists of the points on the line segments joining the origin $\langle 0,0\rangle$ to the points $\langle 1, 1/n\rangle$, $n \in \mathbf{N}$; and B consists of the points on the x-axis between $\frac{1}{2}$ and 1. Now A and B are both arcwise connected, hence also connected. Furthermore, A and B are not separated since each $p \in B$ is a limit point of A; and so $A \cup B$ is connected. But $A \cup B$ is not arcwise connected; in fact, there exists no path from any point in A to any point in B.



Example 8.2: Let A and B be the subsets of the plane $\mathbf{R}^2$ defined in Example 1.2. Now A and B are continuous images of intervals and are therefore connected. Moreover, A and B are not separated sets and so $A \cup B$ is connected. But $A \cup B$ is not arcwise connected; in fact, there exists no path from a point in A to a point in B.

The topology of the plane $\mathbf{R}^2$ is an essential part of the theory of functions of a complex variable. In this case, a *region* is defined as an open connected subset of the plane. The following theorem plays an important role in this theory.

Theorem 13.13: An open connected subset of the plane $\mathbf{R}^2$ is arcwise connected.

HOMOTOPIC PATHS

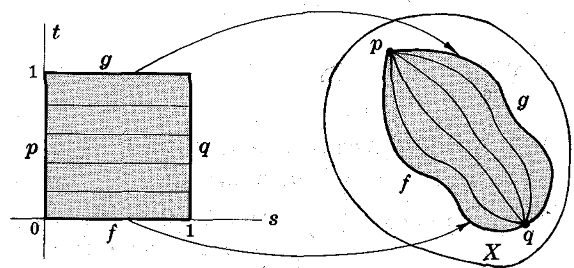
Let $f: I \to X$ and $g: I \to X$ be two paths with the same initial point $p \in X$ and the same terminal point $q \in X$. Then f is said to be *homotopic* to g, written $f \simeq g$, if there exists a continuous function

$$H: I^2 \to X$$

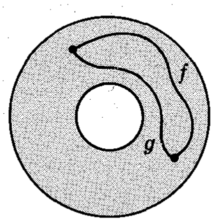
such that

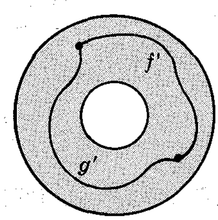
$$H(s,0) = f(s) \qquad H(0,t) = p$$
$$H(s,1) = g(s) \qquad H(1,t) = q$$

as indicated in the adjacent diagram. We then say that f can be continuously deformed into g. The function H is called a *homotopy* from f to g.



Example 9.1: Let X be the set of points between two concentric circles (called an *annulus*). Then the paths f and g in the diagram on the left below are homotopic, whereas the paths f' and g' in the diagram on the right below are not homotopic.





Example 9.2: Let $f: I \to X$ be any path. Then $f \simeq f$, i.e. f is homotopic to itself. For the function $H: I^2 \to X$ defined by

$$H(s,t) = f(s)$$

is a homotopy from f to f.

Example 9.3: Let $f \simeq g$ and, say, $H: I^2 \to X$ is a homotopy from f to g. Then the function $\hat{H}: I^2 \to X$ defined by

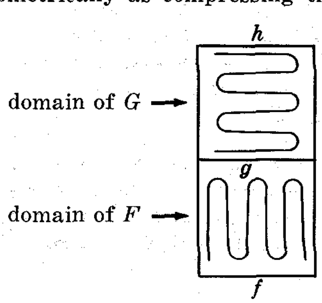
$$\hat{H}(s,t) = H(s, 1-t)$$

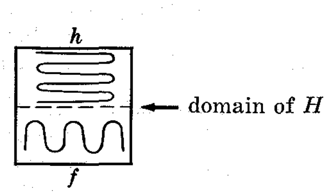
is a homotopy from g to f, and so $g \simeq f$.

Example 9.4: Let $f \simeq g$ and $g \simeq h$; say, $F: I^2 \to X$ is a homotopy from f to g and $G: I^2 \to X$ is a homotopy from g to h. The function $H: I^2 \to X$ defined by

$$H(s,t) = \begin{cases} F(s, 2t) & \text{if } 0 \leq t \leq \frac{1}{2} \\ G(s, 2t-1) & \text{if } \frac{1}{2} \leq t \leq 1 \end{cases}$$

is a homotopy from f to h, and so $f \simeq h$. The homotopy H can be interpreted geometrically as compressing the domains of F and G into one square.



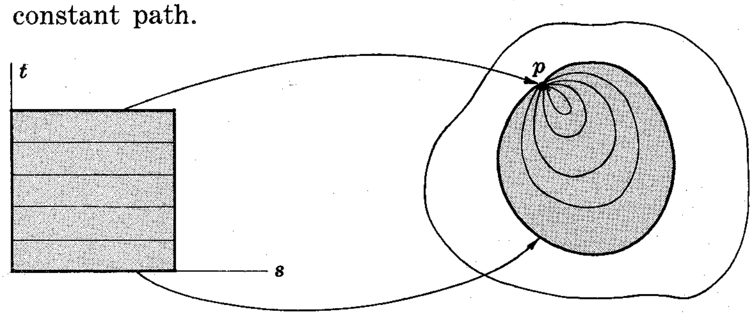


The previous three relations imply the following proposition:

Proposition 13.14: The homotopy relation is an equivalence relation in the collection of all paths from a to b.

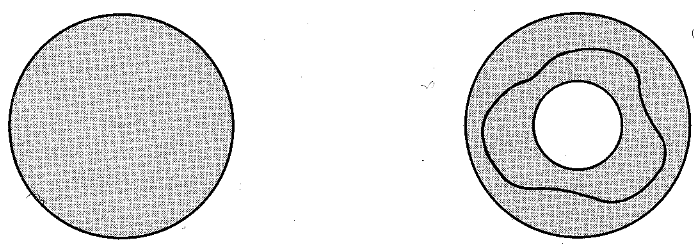
SIMPLY CONNECTED SPACES

A path $f: I \to X$ with the same initial and terminal point, say $f(0) = f(1) = p$, is called a *closed path at* $p \in X$. In particular, the constant path $e_p: I \to X$ defined by $e_p(s) = p$ is a closed path at p. A closed path $f: I \to X$ is said to be *contractable to a point* if it is homotopic to the constant path.



A topological space is *simply connected* iff every closed path in X is contractable to a point.

Example 10.1: An open disc in the plane $\mathbf{R}^2$ is simply connected, whereas an annulus is not simply connected since there are closed curves, as indicated in the diagram, that are not contractable to a point.



simply connected — not simply connected

Solved Problems

SEPARATED SETS

1. Show that if A and B are non-empty separated sets, then $A \cup B$ is disconnected.

Solution:

Since A and B are separated, $A \cap \bar{B} = \emptyset$ and $\bar{A} \cap B = \emptyset$. Let $G = \bar{B}^c$ and $H = \bar{A}^c$. Then G and H are open and

$$(A \cup B) \cap G = A \quad \text{and} \quad (A \cup B) \cap H = B$$

are non-empty disjoint sets whose union is $A \cup B$. Thus G and H form a disconnection of $A \cup B$, and so $A \cup B$ is disconnected.

2. Let $G \cup H$ be a disconnection of A. Show that $A \cap G$ and $A \cap H$ are separated sets.

Solution:

Now $A \cap G$ and $A \cap H$ are disjoint; hence we need only show that each set contains no accumulation point of the other. Let p be an accumulation point of $A \cap G$, and suppose $p \in A \cap H$. Then H is an open set containing p and so H contains a point of $A \cap G$ distinct from p, i.e. $(A \cap G) \cap H \neq \emptyset$. But

$$(A \cap G) \cap (A \cap H) = \emptyset = (A \cap G) \cap H$$

Accordingly, $p \notin A \cap H$.

Similarly, if p is an accumulation point of $A \cap H$, then $p \notin A \cap G$. Thus $A \cap G$ and $A \cap H$ are separated sets.

3. Prove Theorem 13.1: A set A is connected if and only if A is not the union of two non-empty separated sets.

Solution:

We show, equivalently, that A is disconnected if and only if A is the union of two non-empty separated sets. Suppose A is disconnected, and let $G \cup H$ be a disconnection of A. Then A is the union of non-empty sets $A \cap G$ and $A \cap H$ which are, by the preceding problem, separated. On the other hand, if A is the union of two non-empty separated sets, then A is disconnected by Problem 1.

CONNECTED SETS

4. Let $G \cup H$ be a disconnection of A and let B be a connected subset of A. Show that either $B \cap H = \emptyset$ or $B \cap G = \emptyset$, and so either $B \subset G$ or $B \subset H$.

Solution:

Now $B \subset A$, and so

$$A \subset G \cup H \Rightarrow B \subset G \cup H \quad \text{and} \quad G \cap H \subset A^c \Rightarrow G \cap H \subset B^c$$

Thus if both $B \cap G$ and $B \cap H$ are non-empty, then $G \cup H$ forms a disconnection of B. But B is connected; hence the conclusion follows.

5. Prove Proposition 13.2: If A and B are connected sets which are not separated, then $A \cup B$ is connected.

Solution:

Suppose $A \cup B$ is disconnected and suppose $G \cup H$ is a disconnection of $A \cup B$. Since A is a connected subset of $A \cup B$, either $A \subset G$ or $A \subset H$ by the preceding problem. Similarly, either $B \subset G$ or $B \subset H$.

Now if $A \subset G$ and $B \subset H$ (or $B \subset G$ and $A \subset H$), then, by Problem 2,

$$(A \cup B) \cap G = A \quad \text{and} \quad (A \cup B) \cap H = B$$

are separated sets. But this contradicts the hypothesis; hence either $A \cup B \subset G$ or $A \cup B \subset H$, and so $G \cup H$ is not a disconnection of $A \cup B$. In other words, $A \cup B$ is connected.

6. Prove: Let $\mathcal{A} = \{A_i\}$ be a class of connected subsets of X such that no two members of $\mathcal{A}$ are separated. Then $B = \cup_i A_i$ is connected.

Solution:

Suppose B is not connected and $G \cup H$ is a disconnection of B. Now each $A_i \in \mathcal{A}$ is connected and so (Problem 4) is contained in either G or H and disjoint from the other. Furthermore, any two members $A_{i_1}, A_{i_2} \in \mathcal{A}$ are not separated and so, by Proposition 13.2, $A_{i_1} \cup A_{i_2}$ is connected; then $A_{i_1} \cup A_{i_2}$ is contained in G or H and disjoint from the other. Accordingly, all the members of $\mathcal{A}$, and hence $B = \cup_i A_i$, must be contained in either G or H and disjoint from the other. But this contradicts the fact that $G \cup H$ is a disconnection of B; hence B is connected.

7. Prove: Let $\mathcal{A} = \{A_i\}$ be a class of connected subsets of X with a non-empty intersection. Then $B = \cup_i A_i$ is connected.

Solution:

Since $\cap_i A_i \neq \emptyset$, any two members of $\mathcal{A}$ are not disjoint and so are not separated; hence, by the preceding problem, $B = \cup_i A_i$ is connected.

8. Let A be a connected subset of X and let $A \subset B \subset \bar{A}$. Show that B is connected and hence, in particular, $\bar{A}$ is connected.

Solution:

Suppose B is disconnected and suppose $G \cup H$ is a disconnection of B. Now A is a connected subset of B and so, by Problem 4, either $A \cap H = \emptyset$ or $A \cap G = \emptyset$; say, $A \cap H = \emptyset$. Then H^c is a closed superset of A and therefore $A \subset B \subset \bar{A} \subset H^c$. Consequently, $B \cap H = \emptyset$. But this contradicts the fact that $G \cup H$ is a disconnection of B; hence B is connected.

CONNECTED SPACES

9. Let X be a topological space. Show that the following conditions are equivalent:

(i) X is disconnected.

(ii) There exists a non-empty proper subset of X which is both open and closed.

Solution:

(i) $\Rightarrow$ (ii): Suppose $X = G \cup H$ where G and H are non-empty and open. Then G is a non-empty proper subset of X and, since $G = H^c$, G is both open and closed.

(ii) $\Rightarrow$ (i): Suppose A is a non-empty proper subset of X which is both open and closed. Then A^c is also non-empty and open, and $X = A \cup A^c$. Accordingly, X is disconnected.

10. Prove Theorem 13.3: Let A be a subset of a topological space $(X, \mathcal{T})$ and let $\mathcal{T}_A$ be the relative topology on A. Then A is $\mathcal{T}$-connected if and only if A is $\mathcal{T}_A$-connected.

Solution:

Suppose A is disconnected with $G \cup H$ forming a $\mathcal{T}$-disconnection of A. Now $G, H \in \mathcal{T}$ and so $A \cap G, A \cap H \in \mathcal{T}_A$. Accordingly, $A \cap G$ and $A \cap H$ form a $\mathcal{T}_A$-disconnection of A; hence A is $\mathcal{T}_A$-disconnected.

On the other hand, suppose A is $\mathcal{T}_A$-disconnected, say G^* and H^* form a $\mathcal{T}_A$-disconnection of A. Then $G^*, H^* \in \mathcal{T}_A$ and so

$$\exists\ G, H \in \mathcal{T} \quad \text{such that} \quad G^* = A \cap G \ \text{ and } \ H^* = A \cap H$$

But $$A \cap G^* = A \cap A \cap G = A \cap G \quad \text{and} \quad A \cap H^* = A \cap A \cap H = A \cap H$$

Hence $G \cup H$ is a $\mathcal{T}$-disconnection of A and so A is $\mathcal{T}$-disconnected.

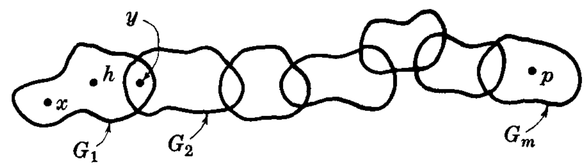
11. Let $p, q \in X$. The subsets $A_1, \ldots, A_m$ of X are said to form a *simple (finite) chain* joining p to q if A_1 (and only A_1) contains p, A_m (and only A_m) contains q, and $A_i \cap A_j = \emptyset$ iff $|i-j| > 1$.

Prove: Let X be connected and let $\mathcal{A}$ be an open cover of X. Then any pair of points in X can be joined by a simple chain consisting of members of $\mathcal{A}$.

Solution:

Let p be any arbitrary point in X and let H consist of those points in X which can be joined to p by some simple chain consisting of members of $\mathcal{A}$. Now $H \neq \emptyset$, since $p \in H$. We claim that H is both open and closed and so $H = X$ since X is connected.

Let $h \in H$. Then $\exists\ G_1, \ldots, G_m \in \mathcal{A}$ which form a simple chain from h to p. But if $x \in G_1 \setminus G_2$, then $G_1, \ldots, G_m$ form a simple chain from x to p; and if $y \in G_1 \cap G_2$, then $G_2, \ldots, G_m$ form a simple chain from y to p, as indicated in the diagram below.



Thus G_1 is a subset of H, i.e. $h \subset G_1 \subset H$. Hence H is a neighborhood of each of its points, and so H is open.

Now let $g \in H^c$. Since $\mathcal{A}$ is a cover of X, $\exists\ G \in \mathcal{A}$ such that $g \in G$, and G is open. If $G \cap H \neq \emptyset$, $\exists\ h \in G \cap H \subset H$ and so $\exists\ G_1, \ldots, G_m \in \mathcal{A}$ forming a simple chain from h to p. But then either $G, G_k, \ldots, G_m$, where we consider the maximum k for which G intersects G_k, or $G_1, \ldots, G_m$ form a simple chain from g to p, and so $g \in H$, a contradiction. Hence $G \cap H = \emptyset$, and so $g \in G \subset H^c$. Thus H^c is an open set, and so $H^{cc} = H$ is closed.

12. Prove Theorem 13.7: Let E be a subset of the real line $\mathbf{R}$ containing at least two points. Then E is connected if and only if E is an interval.

Solution:

Suppose E is not an interval; then

$$\exists\ a, b \in E,\ p \notin E \quad \text{such that} \quad a < p < b$$

Set $G = (-\infty, p)$ and $H = (p, \infty)$. Then $a \in G$ and $b \in H$, and hence $E \cap G$ and $E \cap H$ are non-empty disjoint sets whose union is E. Thus E is disconnected.

Now suppose E is an interval and, furthermore, assume E is disconnected; say, G and H form a disconnection of E. Set $A = E \cap G$ and $B = E \cap H$; then $E = A \cup B$. Now A and B are non-empty; say, $a \in A$, $b \in B$, $a < b$ and $p = \sup\{A \cap [a, b]\}$. Since $[a, b]$ is a closed set, $p \in [a, b]$ and hence $p \in E$.

Suppose $p \in A = E \cap G$. Then $p < b$ and $p \in G$. Since G is an open set

$$\exists\ \delta > 0 \quad \text{such that} \quad p + \delta \in G \quad \text{and} \quad p + \delta < b$$

Hence $p + \delta \in E$ and so $p + \delta \in A$. But this contradicts the definition of p, i.e. $p = \sup\{A \cap [a, b]\}$. Therefore $p \notin A$.

On the other hand, suppose $p \in B = E \cap H$. Then, in particular, $p \in H$. Since H is an open set,

$$\exists\ \delta^* > 0 \quad \text{such that} \quad [p - \delta^*, p] \subset H \quad \text{and} \quad a < p - \delta^*$$

Hence $[p - \delta^*, p] \subset E$ and so $[p - \delta^*, p] \subset B$. Accordingly, $[p - \delta^*, p] \cap A = \emptyset$. But then $p - \delta^*$ is an upper bound for $A \cap [a, b]$, which is impossible since $p = \sup\{A \cap [a, b]\}$. Hence $p \notin B$. But this contradicts the fact that $p \in E$, and so E is connected.

13. Prove (see Example 4.1): Let $I = [0, 1]$ and let $f: I \to I$ be continuous. Then $\exists\ p \in I$ such that $f(p) = p$.

Solution:

If $f(0) = 0$ or $f(1) = 1$, the theorem follows; hence we can assume that $f(0) > 0$ and $f(1) < 1$. Since f is continuous, the graph of the function

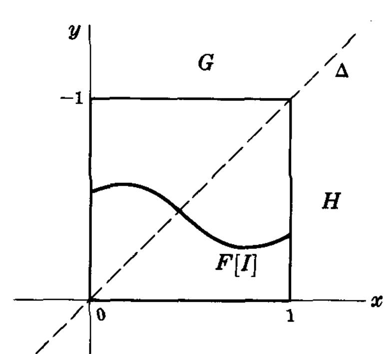
$$F: I \to \mathbf{R}^2 \quad \text{defined by} \quad F(x) = \langle x, f(x) \rangle$$

is also continuous.

Set $G = \{\langle x, y \rangle : x < y\}$, $H = \{\langle x, y \rangle : y < x\}$; then $\langle 0, f(0) \rangle \in G$, $\langle 1, f(1) \rangle \in H$. Hence if $F[I]$ does not contain a point of the diagonal

$$\Delta = \{\langle x, y \rangle : x = y\} = \mathbf{R}^2 \setminus (G \cup H)$$

then $G \cup H$ is a disconnection of $F[I]$. But this contradicts the fact that $F[I]$, the continuous image of a connected set, is connected; hence $F[I]$ contains a point $\langle p, p \rangle \in \Delta$, and so $f(p) = p$.



COMPONENTS

14. Show that every component E is closed.

Solution:

Now E is connected and so, by Problem 6, $\bar{E}$ is connected, $E \subset \bar{E}$. But E, a component, is a maximal connected set; hence $E = \bar{E}$, and so E is closed.

15. Prove: Let $p \in X$ and let $\mathcal{A}_p = \{A_i\}$ be the class of connected subsets of X containing p. Furthermore, let $C_p = \cup_i A_i$. Then: (i) C_p is connected. (ii) If B is a connected subset of X containing p, then $B \subset C_p$. (iii) C_p is a maximal connected subset of X, i.e. a component.

Solution:

(i) Since each $A_i \in \mathcal{A}_p$ contains p, $p \in \cap_i A_i$ and so, by Problem 7, $C_p = \cup_i A_i$ is connected.

(ii) If B is a connected subset of X containing p, then $B \in \mathcal{A}_p$ and so $B \subset C_p = \cup \{A_i : A_i \in \mathcal{A}_p\}$.

(iii) Let $C_p \subset D$, where D is connected. Then $p \in D$ and hence, by (ii), $D \subset C_p$; that is, $C_p = D$. Therefore C_p is a component.

16. Prove Theorem 13.9: The components of X form a partition of X. Every connected subset of X is contained in some component.

Solution:

Consider the class $\mathcal{C} = \{C_p : p \in X\}$ where C_p is defined as in the preceding problem. We claim that $\mathcal{C}$ consists of the components of X. By the preceding problem, each $C_p \in \mathcal{C}$ is a component. On the other hand, if D is a component, then D contains some point $p_0 \in X$ and so $D \subset C_{p_0}$. But D is a component; hence $D = C_{p_0}$.

We now show that $\mathcal{C}$ is a partition of X. Clearly, $X = \cup \{C_p : p \in X\}$; hence we need only show that distinct components are disjoint or, equivalently, if $C_p \cap C_q \neq \emptyset$, then $C_p = C_q$. Let $a \in C_p \cap C_q$. Then $C_p \subset C_a$ and $C_q \subset C_a$, since C_p and C_q are connected sets containing a. But C_p and C_q are components; hence $C_p = C_a = C_q$.

Lastly, if E is a non-empty connected subset of X, then E contains a point $p_0 \in X$ and so $E \subset C_{p_0}$ by the preceding problem. If $E = \emptyset$, then E is contained in every component.

17. Show that if X and Y are connected spaces, then $X \times Y$ is connected. Hence a finite product of connected spaces is connected.

Solution:

Let $p = \langle x_1, y_1 \rangle$ and $q = \langle x_2, y_2 \rangle$ be any pair of points in $X \times Y$. Now $\{x_1\} \times Y$ is homeomorphic to Y and is therefore connected. Similarly, $X \times \{y_2\}$ is connected.

But $\{x_1\} \times Y \cap X \times \{y_2\} = \{\langle x_1, y_2 \rangle\}$; hence $\{x_1\} \times Y \cup X \times \{y_2\}$ is connected. Accordingly, p and q belong to the same component. But p and q were arbitrary; hence $X \times Y$ has one component and is therefore connected.

18. Prove Theorem 13.10: The product of connected spaces is connected, i.e. connectedness is a product invariant property.

Solution:

Let $\{X_i : i \in I\}$ be a collection of connected spaces and let $X = \prod_i X_i$ be the product space. Furthermore, let $p = \langle a_i : i \in I \rangle \in X$ and let $E \subset X$ be the component of p. We claim that every point $x = \langle x_i : i \in I \rangle \in X$ belongs to the closure of E and hence belongs to E since E is closed. Now let

$$G = \prod \{X_i : i \neq i_1, \ldots, i_m\} \times G_{i_1} \times \cdots \times G_{i_m}$$

be any basic open set containing $x \in X$. Now

$$H = \prod \{\{a_i\} : i \neq i_1, \ldots, i_m\} \times X_{i_1} \times \cdots \times X_{i_m}$$

is homeomorphic to $X_{i_1} \times \cdots \times X_{i_m}$ and hence connected. Furthermore, $p \in H$ and so H is a subset of E, the component of p. But $G \cap H$ is non-empty; hence G contains a point of E. Accordingly, $x \in \bar{E} = E$. Thus X has one component and is therefore connected.

ARCWISE CONNECTED SETS

19. Let $f: I \to X$ be any path in X. Show that $f[I]$, the range of f, is connected.

Solution:

$I = [0,1]$ is connected and f is continuous; hence, by Theorem 13.5, $f[I]$ is connected.

20. Prove: Continuous images of arcwise connected sets are arcwise connected.

Solution:

Let $E \subset X$ be arcwise connected and let $f: X \to Y$ be continuous. We claim that $f[E]$ is arcwise connected. For let $p, q \in f[E]$. Then $\exists\, p^*, q^* \in E$ such that $f(p^*) = p$ and $f(q^*) = q$. But E is arcwise connected and so

$$\exists \text{ a path } g: I \to X \quad \text{such that} \quad g(0) = p^*,\ g(1) = q^* \text{ and } g[I] \subset E$$

Now the composition of continuous functions is continuous and so $f \circ g: I \to Y$ is continuous. Furthermore,

$$f \circ g(0) = f(p^*) = p, \quad f \circ g(1) = f(q^*) = q \quad \text{and} \quad f \circ g[I] = f[g[I]] \subset f[E]$$

Thus $f[E]$ is arcwise connected.

21. Prove Theorem 13.12: Every arcwise connected set A is connected.

Solution:

If A is empty, then A is connected. Suppose A is not empty; say, $p \in A$. Now A is arcwise connected and so, for each $a \in A$, there is a path $f_a: I \to A$ from p to a. Furthermore,

$$a \in f_a[I] \subset A \quad \text{and so} \quad A = \bigcup\{f_a[I] : a \in A\}$$

But $p \in f_a[I]$, for every $a \in A$; hence $\bigcap\{f_a[I] : a \in A\}$ is non-empty. Moreover, each $f_a[I]$ is connected and so, by Problem 7, A is connected.

22. Prove: Let $\mathcal{A}$ be a class of arcwise connected subsets of X with a non-empty intersection. Then $B = \bigcup\{A : A \in \mathcal{A}\}$ is arcwise connected.

Solution:

Let $a, b \in B$. Then

$$\exists\, A_a, A_b \in \mathcal{A} \quad \text{such that} \quad a \in A_a,\ b \in A_b$$

Now $\mathcal{A}$ has a non-empty intersection; say, $p \in \bigcap\{A : A \in \mathcal{A}\}$. Then $p \in A_a$ and, since A_a is arcwise connected, there is a path $f: I \to A_a \subset B$ from a to p. Similarly, there is a path $g: I \to A_b \subset B$ from p to b. The juxtaposition of the two paths (see Example 7.3) is a path from a to b contained in B. Hence B is arcwise connected.

23. Show that an open disc D in the plane $\mathbf{R}^2$ is arcwise connected.

Solution:

Let $p = \langle a_1, b_1\rangle,\ q = \langle a_2, b_2\rangle \in D$. The function $f: I \to \mathbf{R}^2$ defined by

$$f(t) = \langle a_1 + t(a_2 - a_1),\ b_1 + t(b_2 - b_1)\rangle$$

is a path from p to q which is contained in D. (Geometrically, $f[I]$ is the line segment connecting p and q.) Hence D is arcwise connected.

24. Prove Theorem 13.13: Let E be a non-empty open connected subset of the plane $\mathbf{R}^2$. Then E is arcwise connected.

Solution:

Method 1.

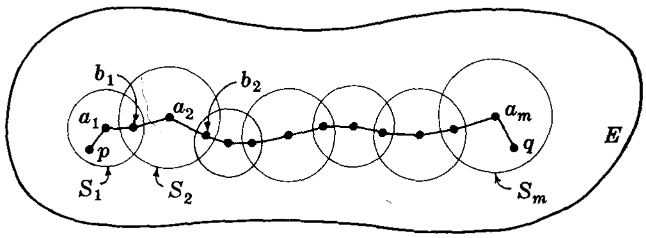
Let $p \in E$ and let G consist of those points in E which can be joined to p by a path in E. We claim that G is open. For let $q \in G \subset E$. Now E is open and so $\exists$ an open disc D with center q such that $q \in D \subset E$. But D is arcwise connected; hence each point $x \in D$ can be joined to q which can be joined to p. Hence each point $x \in D$ can be joined to p, and so $q \in D \subset G$. Accordingly, G is open.

Now set $H = E \setminus G$, i.e. H consists of those points in E which cannot be joined to E by a path in E. We claim that H is open. For let $q^* \in H \subset E$. Since E is open, $\exists$ an open disc D^* with center q^* such that $q^* \in D^* \subset E$. Since D^* is arcwise connected, each $x \in D^*$ cannot be joined to p with a path in E, and so $q^* \in D^* \subset H$. Hence H is open.

But E is connected and therefore E cannot be the union of two non-empty disjoint open sets. Then $H = \emptyset$, and so $E = G$ is arcwise connected.

Method 2.

Since E is open, E is the union of open discs. But E is connected; hence, by Problem 11, $\exists$ open discs $S_1, \ldots, S_m \subset E$ which form a simple chain joining any $p \in E$ to any $q \in E$. Let a_i be the center of S_i and let $b_i \in S_i \cap S_{i+1}$. Then the polygonal arc joining p to a_1 to b_1 to a_2, etc., is contained in the union of the discs and hence is contained in E. Thus E is arcwise connected.



TOTALLY DISCONNECTED SPACES

25. A topological space X is said to be *totally disconnected* if for each pair of points $p, q \in X$ there exists a disconnection $G \cup H$ of X with $p \in G$ and $q \in H$. Show that the real line $\mathbf{R}$ with the topology $\mathcal{T}$ generated by the open closed intervals $(a, b]$ is totally disconnected.

Solution:

Let $p, q \in R$; say, $p < q$. Then $G = (-\infty, p]$ and $H = (p, \infty)$ are open disjoint sets whose union is $\mathbf{R}$, i.e. $G \cup H$ is a disconnection of $\mathbf{R}$. But $p \in G$ and $q \in H$; hence $(\mathbf{R}, \mathcal{T})$ is totally disconnected.

26. Show that the set $\mathbf{Q}$ of rational numbers with the relative usual topology is totally disconnected.

Solution:

Let $p, q \in \mathbf{Q}$; say, $p < q$. Now there exists an irrational number a such that $p < a < q$.

Set $G = \{x \in Q : x < a\}$ and $H = \{x \in Q : x > a\}$. Then $G \cup H$ is a disconnection of $\mathbf{Q}$, and $p \in G$ and $q \in H$. Thus $\mathbf{Q}$ is totally disconnected.

27. Prove: The components of a totally disconnected space X are the singleton subsets of X.

Solution:

Let E be a component of X and suppose $p, q \in E$ with $p \neq q$. Since X is totally disconnected, there exists a disconnection $G \cup H$ of X such that $p \in G$ and $q \in H$. Consequently, $E \cap G$ and $E \cap H$ are non-empty and so $G \cup H$ is a disconnection of E. But this contradicts the fact that E is a component and so is connected. Hence E consists of exactly one point.

LOCALLY CONNECTED SPACES

28. Prove: Let E be a component in a locally connected space X. Then E is open.

Solution:

Let $p \in E$. Since X is locally connected, p belongs to at least one open connected set G_p. But E is the component of p; hence

$$p \in G_p \subset E \quad \text{and so} \quad E = \bigcup\{G_p : p \in E\}$$

Therefore E is open, as it is the union of open sets.

29. Prove: Let X and Y be locally connected. Then $X \times Y$ is locally connected.

Solution:

Now X is locally connected iff X possesses a base $\mathcal{B}$ consisting of connected sets. Similarly, Y possesses a base $\mathcal{B}^*$ consisting of connected sets. But $X \times Y$ is a finite product; hence

$$\{G \times H : G \in \mathcal{B},\ H \in \mathcal{B}^*\}$$

is a base for the product space $X \times Y$. Now each $G \times H$ is connected since G and H are connected. In other words, $X \times Y$ possesses a base consisting of connected sets and so $X \times Y$ is locally connected.

30. Prove: Let $\{X_i\}$ be a collection of connected locally connected spaces. Then the product space $X = \prod_i X_i$ is locally connected.

Solution:

Let G be an open subset of X containing $p = \langle a_i : i \in I\rangle \in X$. Then there exists a member of the defining base

$$B = G_{i_1} \times \cdots \times G_{i_m} \times \prod\{X_i : i \neq i_1, \ldots, i_m\}$$

such that $p \in B \subset G$, and so $a_{i_k} \in G_{i_k}$. Now each coordinate space is locally connected, and so there exists connected open subsets $H_{i_k} \subset X_{i_k}$ such that

$$a_{i_1} \in H_{i_1} \subset G_{i_1}, \ \ldots, \ a_{i_m} \in H_{i_m} \subset G_{i_m}$$

Set

$$H = H_{i_1} \times \cdots \times H_{i_m} \times \prod\{X_i : i \neq i_1, \ldots, i_m\}$$

Since each X_i is connected and each H_{i_k} is connected, H is also connected. Furthermore, H is open and $p \in H \subset B \subset G$. Accordingly, X is locally connected.

Supplementary Problems

CONNECTED SPACES

31. Show that if $(X, \mathcal{T})$ is connected and $\mathcal{T}^* \lesssim \mathcal{T}$, then $(X, \mathcal{T}^*)$ is connected.

32. Show that if $(X, \mathcal{T})$ is disconnected and $\mathcal{T} \lesssim \mathcal{T}^*$, then $(X, \mathcal{T}^*)$ is disconnected.

33. Show that every indiscrete space is connected.

34. Show, by a counterexample, that connectedness is not a hereditary property.

35. Prove: If $A_1, A_2, \ldots$ is a sequence of connected sets such that A_1 and A_2 are not separated, A_2 and A_3 are not separated, etc., then $A_1 \cup A_2 \cup \cdots$ is connected.

36. Prove: Let E be a connected subset of a T_1-space containing more than one element. Then E is infinite.

37. Prove: A topological space X is connected if and only if every non-empty proper subset of X has a non-empty boundary.

COMPONENTS

38. Determine the components of a discrete space.

39. Determine the components of a cofinite space.

40. Show that any pair of components are separated.

41. Prove: If X has a finite number of components, then each component is both open and closed.

42. Prove: If E is a non-empty connected subset of X which is both open and closed, then E is a component.

43. Prove: Let E be a component of Y and let $f: X \to Y$ be continuous. Then $f^{-1}[E]$ is a union of components of X.

44. Prove: Let X be a compact space. If the components of X are open, then there are only a finite number of them.

ARCWISE CONNECTED SETS

45. Show that an indiscrete space is arcwise connected.

46. Prove: The arcwise connected components of X form a partition of X.

47. Prove: Every component of X is partitioned by arcwise connected components.

MISCELLANEOUS PROBLEMS

48. Show that an indiscrete space is simply connected.

49. Show that a totally disconnected space is Hausdorff.

50. Prove: Let G be an open subset of a locally connected space X. Then G is locally connected.

51. Let $A = \{a, b\}$ be discrete and let $I = [0, 1]$. Show that the product space $X = \prod \{A_i : A_i = A, i \in I\}$ is not locally connected. Hence locally connectedness is not product invariant.

52. Show that "simply connected" is a topological property.

53. Prove: Let X be locally connected. Then X is connected if and only if there exists a simple chain of connected sets joining any pair of points in X.

Chapter 14

Complete Metric Spaces

CAUCHY SEQUENCES

Let X be a metric space. A sequence $\langle a_1, a_2, \ldots \rangle$ in X is a *Cauchy sequence* iff for every $\epsilon > 0$,

$$\exists\, n_0 \in \mathbf{N} \quad \text{such that} \quad n, m > n_0 \Rightarrow d(a_n, a_m) < \epsilon$$

Hence, in the case that X is a normed space, $\langle a_n \rangle$ is a Cauchy sequence iff for every $\epsilon > 0$,

$$\exists\, n_0 \in \mathbf{N} \quad \text{such that} \quad n, m > n_0 \Rightarrow \|a_n - a_m\| < \epsilon$$

Example 1.1: Let $\langle a_n \rangle$ be a convergent sequence; say $a_n \to p$. Then $\langle a_n \rangle$ is necessarily a Cauchy sequence since, for every $\epsilon > 0$,

$$\exists\, n_0 \in \mathbf{N} \quad \text{such that} \quad n > n_0 \Rightarrow d(a_n, p) < \tfrac{1}{2}\epsilon$$

Hence, by the Triangle Inequality,

$$n, m > n_0 \Rightarrow d(a_n, a_m) \le d(a_n, p) + d(a_m, p) < \tfrac{1}{2}\epsilon + \tfrac{1}{2}\epsilon = \epsilon$$

In other words, $\langle a_n \rangle$ is a Cauchy sequence.

We state the result of Example 1.1 as a proposition.

Proposition 14.1: Every convergent sequence in a metric space is a Cauchy sequence.

The converse of Proposition 14.1 is not true, as seen in the next example.

Example 1.2: Let $X = (0, 1)$ with the usual metric. Then $\langle \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots \rangle$ is a sequence in X which is Cauchy but which does not converge in X.

Example 1.3: Let d be the trivial metric on any set X and let $\langle a_n \rangle$ be a Cauchy sequence in (X, d). Recall that d is defined by

$$d(a, b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{if } a \neq b \end{cases}$$

Let $\epsilon = \frac{1}{2}$. Then, since $\langle a_n \rangle$ is Cauchy, $\exists\, n_0 \in \mathbf{N}$ such that

$$n, m > n_0 \Rightarrow d(a_n, a_m) < \tfrac{1}{2} \Rightarrow a_n = a_m$$

In other words, $\langle a_n \rangle$ is of the form $\langle a_1, a_2, \ldots, a_{n_0}, p, p, p, \ldots \rangle$, i.e. constant from some term on.

Example 1.4: Let $\langle p_1, p_2, \ldots \rangle$ be a Cauchy sequence in Euclidean m-space $\mathbf{R}^m$; say,

$$p_1 = \langle a_1^{(1)}, \ldots, a_1^{(m)} \rangle, \quad p_2 = \langle a_2^{(1)}, \ldots, a_2^{(m)} \rangle, \quad \ldots$$

The projections of $\langle p_n \rangle$ into each of the m coordinate spaces, i.e.,

$$\langle a_1^{(1)}, a_2^{(1)}, a_3^{(1)}, \ldots \rangle, \ \ldots, \ \langle a_1^{(m)}, a_2^{(m)}, a_3^{(m)}, \ldots \rangle \qquad (1)$$

are Cauchy sequences in $\mathbf{R}$, for, let $\epsilon > 0$. Since $\langle p_n \rangle$ is Cauchy, $\exists\, n_0 \in \mathbf{N}$ such that

$$r, s > n_0 \Rightarrow d(p_r, p_s)^2 = |a_r^{(1)} - a_s^{(1)}|^2 + \cdots + |a_r^{(m)} - a_s^{(m)}|^2 < \epsilon^2$$

Hence, in particular,

$$r, s > n_0 \Rightarrow |a_r^{(1)} - a_s^{(1)}|^2 < \epsilon^2, \ \ldots, \ |a_r^{(m)} - a_s^{(m)}|^2 < \epsilon^2$$

In other words, each of the m sequences in (*1*) is a Cauchy sequence.

COMPLETE METRIC SPACES

Definition: A metric space (X, d) is *complete* if every Cauchy sequence $\langle a_n \rangle$ in X converges to a point $p \in X$.

Example 2.1: By the fundamental Cauchy Convergence Theorem (see Page 52), the real line $\mathbf{R}$ with the usual metric is complete.

Example 2.2: Let d be the trivial metric on any set X. Now (see Example 1.3) a sequence $\langle a_n \rangle$ in X is Cauchy iff it is of the form $\langle a_1, a_2, \ldots, a_{n_0}, p, p, p, \ldots \rangle$, which clearly converges to $p \in X$. Thus every trivial metric space is complete.

Example 2.3: The open unit interval $X = (0,1)$ with the usual metric is not complete since (see Example 1.2) the sequence $\langle \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots \rangle$ in X is Cauchy but does not converge to a point in X.

Remark: Examples 2.1 and 2.3 show that completeness is not a topological property; for $\mathbf{R}$ is homeomorphic to $(0,1)$ even though $\mathbf{R}$ is complete and $(0,1)$ is not.

Example 2.4: Euclidean m-space $\mathbf{R}^m$ is complete. For, let $\langle p_1, p_2, \ldots \rangle$ be a Cauchy sequence in $\mathbf{R}^m$ where

$$p_1 = \langle a_1^{(1)}, \ldots, a_1^{(m)} \rangle, \quad p_2 = \langle a_2^{(1)}, \ldots, a_2^{(m)} \rangle, \quad \ldots$$

Then (see Example 1.4) the projections of $\langle p_n \rangle$ into the m coordinate spaces are Cauchy; and since $\mathbf{R}$ is complete, they converge:

$$\langle a_1^{(1)}, a_2^{(1)}, \ldots \rangle \to b_1, \quad \ldots, \quad \langle a_1^{(m)}, a_2^{(m)}, \ldots \rangle \to b_m$$

Thus $\langle p_n \rangle$ converges to the point $q = \langle b_1, \ldots, b_m \rangle \in \mathbf{R}^m$, since each of the m projections converges to the projection of q (see Page 169, Theorem 12.7).

PRINCIPLE OF NESTED CLOSED SETS

Recall that the diameter of a subset A of a metric space X, denoted by $d(A)$, is defined by $d(A) = \sup\{d(a,a') : a, a' \in A\}$ and that a sequence of sets, $A_1, A_2, \ldots,$ is said to be nested if $A_1 \supset A_2 \supset \cdots$.

The next theorem gives a characterization of complete metric spaces analogous to the Nested Interval Theorem for the real numbers.

Theorem 14.2: A metric space X is complete if and only if every nested sequence of non-empty closed sets whose diameters tend to zero has a non-empty intersection.

In other words, if $A_1 \supset A_2 \supset \cdots$ are non-empty closed subsets of a complete metric space X such that $\lim_{n\to\infty} d(A_n) = 0$, then $\cap_{n=0}^{\infty} A_n \neq \emptyset$; and vice versa.

The next examples show that the conditions $\lim_{n\to\infty} d(A_n) = 0$ and that the A_i are closed, are both necessary in Theorem 14.2.

Example 3.1: Let X be the real line $\mathbf{R}$ and let $A_n = [n, \infty)$. Now X is complete, the A_n are closed, and $A_1 \supset A_2 \supset \cdots$. But $\cap_{n=1}^{\infty} A_n$ is empty. Observe that $\lim_{n\to\infty} d(A_n) \neq 0$.

Example 3.2: Let X be the real line $\mathbf{R}$ and let $A_n = (0, 1/n]$. Now X is complete, $A_1 \supset A_2 \supset \cdots$, and $\lim_{n\to\infty} d(A_n) = 0$. But $\cap_{n=1}^{\infty} A_n$ is empty. Observe that the A_n are not closed.

COMPLETENESS AND CONTRACTING MAPPINGS

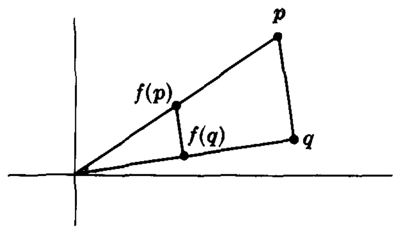
Let X be a metric space. A function $f: X \to X$ is called a *contracting mapping* if there exists a real number α, $0 \leq \alpha < 1$, such that, for every $p, q \in X$,

$$d(f(p), f(q)) \leq \alpha\, d(p,q) < d(p,q)$$

Thus, in a contracting mapping, the distance between the images of any two points is less than the distance between the points.

Example 4.1: Let f be the function on Euclidean 2-space $\mathbf{R}^2$, i.e. $f: \mathbf{R}^2 \to \mathbf{R}^2$, defined by $f(p) = \frac{1}{2}p$. Then f is contracting, for

$$d(f(p), f(q)) = \|f(p) - f(q)\| = \|\tfrac{1}{2}p - \tfrac{1}{2}q\| = \tfrac{1}{2}\|p - q\| = \tfrac{1}{2}d(p, q)$$



If X is a complete metric space, then we have the following "fixed point" theorem which has many applications in analysis.

Theorem 14.3: If f is a contracting mapping on a complete metric space X, then there exists a unique point $p \in X$ such that $f(p) = p$.

COMPLETIONS

A metric space X^* is called a *completion* of a metric space X if X^* is complete and X is isometric to a dense subset of X^*.

Example 5.1: The set $\mathbf{R}$ of real numbers is a completion of the set $\mathbf{Q}$ of rational numbers, since $\mathbf{R}$ is complete and $\mathbf{Q}$ is a dense subset of $\mathbf{R}$.

We now outline one particular construction of a completion of an arbitrary metric space X. Let $C[X]$ denote the collection of all Cauchy sequences in X and let $\sim$ be the relation in $C[X]$ defined by

$$\langle a_n \rangle \sim \langle b_n \rangle \quad \text{iff} \quad \lim_{n\to\infty} d(a_n, b_n) = 0$$

Thus, under "$\sim$" we identify those Cauchy sequences which "should" have the same "limit".

Lemma 14.4: The relation $\sim$ is an equivalence relation in $C[X]$.

Now let X^* denote the quotient set $C[X]/\sim$, i.e. X^* consists of equivalence classes $[\langle a_n \rangle]$ of Cauchy sequences $\langle a_n \rangle \in C[X]$. Let e be the function defined by

$$e([\langle a_n \rangle], [\langle b_n \rangle]) = \lim_{n\to\infty} d(a_n, b_n)$$

where $[\langle a_n \rangle], [\langle b_n \rangle] \in X^*$.

Lemma 14.5: The function e is well-defined, i.e. $\langle a_n \rangle \sim \langle a_n^* \rangle$ and $\langle b_n \rangle \sim \langle b_n^* \rangle$ implies $\lim_{n\to\infty} d(a_n, b_n) = \lim_{n\to\infty} d(a_n^*, b_n^*)$.

In other words, e does not depend upon the particular Cauchy sequence chosen to represent any equivalence class. Furthermore,

Lemma 14.6: The function e is a metric on X^*.

Now for each $p \in X$, the sequence $\langle p, p, p, \ldots \rangle \in C[X]$, i.e. is Cauchy. Set

$$\hat{p} = [\langle p, p, \ldots \rangle] \quad \text{and} \quad \hat{X} = \{\hat{p} : p \in X\}$$

Then $\hat{X}$ is a subset of X^*.

Lemma 14.7: X is isometric to $\hat{X}$, and $\hat{X}$ is dense in X^*.

Lemma 14.8: Every Cauchy sequence in X^* converges, and so X^* is a completion of X.

Lastly, we show that

Lemma 14.9: If Y^* is any completion of X, then Y^* is isometric to X^*.

The previous lemmas imply the following fundamental result.

Theorem 14.10: Every metric space X has a completion and all completions of X are isometric.

In other words, up to isometry, there exists a unique completion of any metric space.

BAIRE'S CATEGORY THEOREM

Recall that a subset A of a topological space X is *nowhere dense* in X iff the interior of the closure of A is empty:

$$\text{int}(\bar{A}) = \emptyset$$

Example 6.1: The set $\mathbf{Z}$ of integers is a nowhere dense subset of the real line $\mathbf{R}$. For $\mathbf{Z}$ is closed, i.e. $\mathbf{Z} = \bar{\mathbf{Z}}$, and its interior is empty; hence

$$\text{int}(\bar{\mathbf{Z}}) = \text{int}(\mathbf{Z}) = \emptyset$$

Similarly every finite subset of $\mathbf{R}$ is nowhere dense in $\mathbf{R}$.

On the other hand, the set $\mathbf{Q}$ of rational numbers is not nowhere dense in $\mathbf{R}$ since the closure of $\mathbf{Q}$ is $\mathbf{R}$ and so

$$\text{int}(\bar{\mathbf{Q}}) = \text{int}(\mathbf{R}) = \mathbf{R} \neq \emptyset$$

A topological space X is said to be of *first category* (or *meager* or *thin*) if X is the countable union of nowhere dense subsets of X. Otherwise X is said to be of *second category* (or *non-meager* or *thick*).

Example 6.2: The set $\mathbf{Q}$ of rational numbers is of first category since the singleton subsets $\{p\}$ of $\mathbf{Q}$ are nowhere dense in $\mathbf{Q}$, and $\mathbf{Q}$ is the countable union of singleton sets.

In view of Baire's Category Theorem, which follows, the real line $\mathbf{R}$ is of second category.

Theorem (Baire) 14.11: Every complete metric space X is of second category.

COMPLETENESS AND COMPACTNESS

Let A be a subset of a metric space X. Now A is compact iff A is sequentially compact iff every sequence $\langle a_n \rangle$ in A has a convergent subsequence $\langle a_{i_n} \rangle$. But, by Example 1.1, $\langle a_{i_n} \rangle$ is a Cauchy sequence. Hence it is reasonable to expect that the notion of completeness is related to the notion of compactness and its related concept: total boundedness.

We state two such relationships:

Theorem 14.12: A metric space X is compact if and only if it is complete and totally bounded.

Theorem 14.13: Let X be a complete metric space. Then $A \subset X$ is compact if and only if A is closed and totally bounded.

CONSTRUCTION OF THE REAL NUMBERS

The real numbers can be constructed from the rational numbers by the method described in this chapter. Specifically, let $\mathbf{Q}$ be the set of rational numbers and let $\mathbf{R}$ be the collection of equivalence classes of Cauchy sequences in $\mathbf{Q}$:

$$\mathbf{R} = \{[\langle a_n \rangle] : \langle a_n \rangle \text{ is a Cauchy sequence in } \mathbf{Q}\}$$

Now **R** with the appropriate metric is a complete metric space.

Remark: Let X be a normed vector space. The construction in this chapter gives us a complete metric space X^*. We can then define the following operations of vector addition, scalar multiplication and norm in X^* so that X^* is, in fact, a complete normed vector space, called a *Banach space*:

(i) $[\langle a_n \rangle] + [\langle b_n \rangle] \equiv [\langle a_n + b_n \rangle]$ (ii) $k[\langle a_n \rangle] \equiv [\langle k a_n \rangle]$ (iii) $\|[\langle a_n \rangle]\| = \lim_{n\to\infty} \|a_n\|$

Solved Problems

CAUCHY SEQUENCES

1. Show that every Cauchy sequence $\langle a_n \rangle$ in a metric space X is totally bounded (hence also bounded).

Solution:

Let $\epsilon > 0$. We want to show that there is a decomposition of $\{a_n\}$ into a finite number of sets, each with diameter less than ϵ. Since $\langle a_n \rangle$ is Cauchy, $\exists\, n_0 \in \mathbf{N}$ such that

$$n, m > n_0 \;\Rightarrow\; d(a_n, a_m) < \epsilon$$

Accordingly, $B = \{a_{n_0+1}, a_{n_0+2}, \ldots\}$ has diameter at most ϵ. Thus $\{a_1\}, \ldots, \{a_{n_0}\}, B$ is a finite decomposition of $\{a_n\}$ into sets with diameter less than ϵ, and so $\langle a_n \rangle$ is totally bounded.

2. Let $\langle a_1, a_2, \ldots \rangle$ be a sequence in a metric space X, and let

$$A_1 = \{a_1, a_2, \ldots\}, \quad A_2 = \{a_2, a_3, \ldots\}, \quad A_3 = \{a_3, a_4, \ldots\}, \quad \ldots$$

Show that $\langle a_n \rangle$ is a Cauchy sequence if and only if the diameters of the A_n tend to zero, i.e. $\lim_{n\to\infty} d(A_n) = 0$.

Solution:

Suppose $\langle a_n \rangle$ is a Cauchy sequence. Let $\epsilon > 0$. Then

$$\exists\, n_0 \in \mathbf{N} \quad \text{such that} \quad n, m > n_0 \;\Rightarrow\; d(a_n, a_m) < \epsilon$$

Accordingly, $\quad n > n_0 \;\Rightarrow\; d(A_n) < \epsilon \quad$ and so $\quad \lim_{n\to\infty} (A_n) = 0$

On the other hand, suppose $\lim_{n\to\infty} d(A_n) = 0$. Let $\epsilon > 0$. Then

$$\exists\, n_0 \in \mathbf{N} \quad \text{such that} \quad d(A_{n_0+1}) < \epsilon$$

Hence $\quad n, m > n_0 \;\Rightarrow\; a_n, a_m \in A_{n_0+1} \;\Rightarrow\; d(a_n, a_m) < \epsilon$

and so $\langle a_n \rangle$ is a Cauchy sequence.

3. Let $\langle a_1, a_2, \ldots \rangle$ be a Cauchy sequence in X and let $\langle a_{i_1}, a_{i_2}, \ldots \rangle$ be a subsequence of $\langle a_n \rangle$. Show that $\lim_{n\to\infty} d(a_n, a_{i_n}) = 0$.

Solution:

Let $\epsilon > 0$. Since $\langle a_n \rangle$ is a Cauchy sequence,

$$\exists\, n_0 \in \mathbf{N} \quad \text{such that} \quad n, m > n_0 - 1 \;\Rightarrow\; d(a_n, a_m) < \epsilon$$

Now $i_{n_0} \geqq n_0 > n_0 - 1$ and therefore $d(a_{n_0}, a_{i_{n_0}}) < \epsilon$. In other words, $\lim_{n\to\infty} d(a_n, a_{i_n}) = 0$.

4. Let $\langle a_1, a_2, \ldots \rangle$ be a Cauchy sequence in X and let $\langle a_{i_1}, a_{i_2}, \ldots \rangle$ be a subsequence of $\langle a_n \rangle$ converging to $p \in X$. Show that $\langle a_n \rangle$ also converges to p.

Solution:

By the Triangle Inequality, $d(a_n, p) \leq d(a_n, a_{i_n}) + d(a_{i_n}, p)$ and therefore

$$\lim_{n\to\infty} d(a_n, p) \leq \lim_{n\to\infty} d(a_n, a_{i_n}) + \lim_{n\to\infty} d(a_{i_n}, p)$$

Since $a_{i_n} \to p$, $\lim_{n\to\infty} d(a_{i_n}, p) = 0$ and, by the preceding problem, $\lim_{n\to\infty} d(a_n, a_{i_n}) = 0$. Then

$$\lim_{n\to\infty} d(a_n, p) = 0 \quad \text{and so} \quad a_n \to p$$

5. Let $\langle b_1, b_2, \ldots \rangle$ be a Cauchy sequence in a metric space X, and let $\langle a_1, a_2, \ldots \rangle$ be a sequence in X such that $d(a_n, b_n) < 1/n$ for every $n \in \mathbf{N}$.
(i) Show that $\langle a_n \rangle$ is also a Cauchy sequence in X.
(ii) Show that $\langle a_n \rangle$ converges to, say, $p \in X$ if and only if $\langle b_n \rangle$ converges to p.

Solution:

(i) By the Triangle Inequality,

$$d(a_m, a_n) \leq d(a_m, b_m) + d(b_m, b_n) + d(b_n, a_n)$$

Let $\epsilon > 0$. Then $\exists\, n_1 \in \mathbf{N}$ such that $1/n_1 < \epsilon/3$. Hence

$$n, m > n_1 \Rightarrow d(a_m, a_n) < \epsilon/3 + d(b_m, b_n) + \epsilon/3$$

By hypothesis, $\langle b_1, b_2, \ldots \rangle$ is a Cauchy sequence; hence

$$\exists\, n_2 \in \mathbf{N} \quad \text{such that} \quad n, m > n_2 \Rightarrow d(b_m, b_n) < \epsilon/3$$

Set $n_0 = \max\{n_1, n_2\}$. Then

$$n, m > n_0 \Rightarrow d(a_m, a_n) < \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon$$

Thus $\langle a_n \rangle$ is a Cauchy sequence.

(ii) By the Triangle Inequality, $d(b_n, p) \leq d(b_n, a_n) + d(a_n, p)$; hence

$$\lim_{n\to\infty} d(b_n, p) \leq \lim_{n\to\infty} d(b_n, a_n) + \lim_{n\to\infty} (a_n, p)$$

But $\lim_{n\to\infty} d(b_n, a_n) \leq \lim_{n\to\infty} (1/n) = 0$. Hence, if $a_n \to p$, $\lim_{n\to\infty} d(b_n, p) \leq \lim_{n\to\infty} (a_n, p) = 0$ and so $\langle b_n \rangle$ also converges to p.

Similarly, if $b_n \to p$ then $a_n \to p$.

COMPLETE SPACES

6. Prove Theorem 14.2: The following are equivalent: (i) X is a complete metric space. (ii) Every nested sequence of non-empty closed sets whose diameters tend to zero has a non-empty intersection.

Solution:

(i) $\Rightarrow$ (ii):

Let $A_1 \supset A_2 \supset \cdots$ be non-empty closed subsets of X such that $\lim_{n\to\infty} d(A_n) = 0$. We want to prove that $\cap_n A_n \neq \emptyset$. Since each A_i is non-empty, we can choose a sequence

$$\langle a_1, a_2, \ldots \rangle \quad \text{such that} \quad a_1 \in A_1,\ a_2 \in A_2,\ \ldots$$

We claim that $\langle a_n \rangle$ is a Cauchy sequence. Let $\epsilon > 0$. Since $\lim_{n\to\infty} d(A_n) = 0$,

$$\exists\, n_0 \in \mathbf{N} \quad \text{such that} \quad d(A_{n_0}) < \epsilon$$

But the A_i are nested; hence

$$n, m > n_0 \Rightarrow A_n, A_m \subset A_{n_0} \Rightarrow a_n, a_m \in A_{n_0} \Rightarrow d(a_n, a_m) < \epsilon$$

Thus $\langle a_n \rangle$ is Cauchy.

Now X is complete and so $\langle a_n \rangle$ converges to, say, $p \in X$. We claim that $p \in \cap_n A_n$. Suppose not, i.e. suppose

$$\exists\ k \in \mathbf{N} \quad \text{such that} \quad p \notin A_k$$

Since A_k is a closed set, the distance between p and A_k is non-zero; say, $d(p, A_k) = \delta > 0$. Then A_k and the open sphere $S = S(p, \frac{1}{2}\delta)$ are disjoint. Hence

$$n > k \quad \Rightarrow \quad a_n \in A_k \quad \Rightarrow \quad a_n \notin S(p, \tfrac{1}{2}\delta)$$

This is impossible since $a_n \to p$. In other words, $p \in \cap_n A_n$ and so $\cap_n A_n$ is non-empty.

(ii) $\Rightarrow$ (i):

Let $\langle a_1, a_2, \ldots \rangle$ be a Cauchy sequence in X. We want to show that $\langle a_n \rangle$ converges. Set

$$A_1 = \{a_1, a_2, \ldots\}, \quad A_2 = \{a_2, a_3, \ldots\}, \quad \ldots$$

i.e. $A_k = \{a_n : n \geq k\}$. Then $A_1 \supset A_2 \supset \cdots$ and, by Problem 2, $\lim_{n \to \infty} d(A_n) = 0$. Furthermore, since $d(\bar{A}) = d(A)$, where $\bar{A}$ is the closure of A, $\bar{A}_1 \supset \bar{A}_2 \supset \cdots$ is a sequence of non-empty closed sets whose diameters tend to zero. Therefore, by hypothesis, $\cap_n \bar{A}_n \neq \emptyset$; say, $p \in \cap_n \bar{A}_n$. We claim that the Cauchy sequence $\langle a_n \rangle$ converges to p.

Let $\epsilon > 0$. Since $\lim_{n \to \infty} d(\bar{A}) = 0$,

$$\exists\ n_0 \in \mathbf{N} \quad \text{such that} \quad d(\bar{A}_{n_0}) < \epsilon$$

and so

$$n > n_0 \quad \Rightarrow \quad a_n, p \in \bar{A}_{n_0} \quad \Rightarrow \quad d(a_n, p) < \epsilon$$

In other words, $\langle a_n \rangle$ converges to p.

7. Let X be a metric space and let $f: X \to X$ be a contracting mapping on X, i.e. there exists $\alpha \in \mathbf{R}$, $0 \leq \alpha < 1$, such that, for every $p, q \in X$, $d(f(p), f(q)) \leq \alpha\, d(p, q)$. Show that f is continuous.

Solution:

We show that f is continuous at each point $x_0 \in X$. Let $\epsilon > 0$. Then

$$d(x, x_0) < \epsilon \quad \Rightarrow \quad d(f(x), f(x_0)) \leq \alpha\, d(x, x_0) \leq \alpha\epsilon < \epsilon$$

and so f is continuous.

8. Prove Theorem 14.3: Let f be a contracting mapping on a complete metric space X, say

$$d(f(a), f(b)) \leq \alpha\, d(a, b), \qquad 0 \leq \alpha < 1$$

Then there exists one and only one point $p \in X$ such that $f(p) = p$.

Solution:

Let a_0 be any point in X. Set

$$a_1 = f(a_0), \;\; a_2 = f(a_1) = f^2(a_0), \;\; \ldots, \;\; a_n = f(a_{n-1}) = f^n(a_0), \;\; \ldots$$

We claim that $\langle a_1, a_2, \ldots \rangle$ is a Cauchy sequence. First notice that

$$\begin{aligned} d(f^{s+t}(a_0), f^t(a_0)) &\leq \alpha\, d(f^{s+t-1}(a_0), f^{t-1}(a_0)) \leq \cdots \leq \alpha^t d(f^s(a_0), a_0) \\ &\leq \alpha^t [d(a_0, f(a_0)) + d(f(a_0), f^2(a_0)) + \cdots + d(f^{s-1}(a_0), f^s(a_0))] \end{aligned}$$

But $d(f^{i+1}(a_0), f^i(a_0)) \leq \alpha^i d(f(a_0), a_0)$ and so

$$\begin{aligned} d(f^{s+t}(a_0), f^t(a_0)) &\leq \alpha^t d(f(a_0), a_0)(1 + \alpha + \alpha^2 + \cdots + \alpha^{s-1}) \\ &\leq \alpha^t d(f(a_0), a_0)[1/(1-\alpha)] \end{aligned}$$

since $(1 + \alpha + \alpha^2 + \cdots + \alpha^{s-1}) \leq 1/(1-\alpha)$.

Now let $\epsilon > 0$ and set

$$\delta = \begin{cases} \epsilon(1-\alpha) & \text{if } d(f(a_0), a_0) = 0 \\ \epsilon(1-\alpha)/d(f(a_0), a_0) & \text{if } d(f(a_0), a_0) \neq 0 \end{cases}$$

Since $\alpha < 1$, $\exists\, n_0 \in \mathbf{N}$ such that $\alpha^{n_0} < \delta$

Hence if $r \geqq s > n_0$,

$$d(a_s, a_r) \leqq \alpha^s [1/(1-\alpha)]\, d(f(a_0), a_0) < \delta[1/(1-\alpha)]\, d(f(a_0), a_0) \leqq \epsilon$$

and so $\langle a_n \rangle$ is a Cauchy sequence.

Now X is complete and so $\langle a_n \rangle$ converges to, say, $p \in X$. We claim that $f(p) = p$; for f is continuous and hence sequentially continuous, and so

$$f(p) = f\left(\lim_{n\to\infty} a_n\right) = \lim_{n\to\infty} f(a_n) = \lim_{n\to\infty} a_{n+1} = p$$

Lastly, we show that p is unique. Suppose $f(p) = p$ and $f(q) = q$; then

$$d(p, q) = d(f(p), f(q)) \leqq \alpha\, d(p, q)$$

But $\alpha < 1$; hence $d(p, q) = 0$, i.e. $p = q$.

COMPLETIONS

9. Show that $\langle a_n \rangle \sim \langle b_n \rangle$ if and only if they are both subsequences of some Cauchy sequence $\langle c_n \rangle$.

Solution:

Suppose $\langle a_n \rangle \sim \langle b_n \rangle$, i.e. $\lim_{n\to\infty} d(a_n, b_n) = 0$. Define $\langle c_n \rangle$ by

$$c_n = \begin{cases} a_{\frac{1}{2}n} & \text{if } n \text{ is even} \\ b_{\frac{1}{2}(n+1)} & \text{if } n \text{ is odd} \end{cases}$$

Thus $\langle c_n \rangle = \langle b_1, a_1, b_2, a_2, \ldots \rangle$. We claim $\langle c_n \rangle$ is a Cauchy sequence. For, let $\epsilon > 0$; now

$$\exists\, n_1 \in \mathbf{N} \text{ such that } m, n > n_1 \Rightarrow d(a_m, a_n) < \tfrac{1}{2}\epsilon$$

$$\exists\, n_2 \in \mathbf{N} \text{ such that } m, n > n_2 \Rightarrow d(b_m, b_n) < \tfrac{1}{2}\epsilon$$

$$\exists\, n_3 \in \mathbf{N} \text{ such that } n > n_3 \Rightarrow d(a_n, b_n) < \tfrac{1}{2}\epsilon$$

Set $n_0 = \max(n_1, n_2, n_3)$. We claim that

$$m, n > 2n_0 \Rightarrow d(c_m, c_n) < \epsilon$$

Note that $m > 2n_0 \Rightarrow \tfrac{1}{2}m > n_1, n_3;\ \tfrac{1}{2}(m+1) > n_2, n_3$

Thus

$$m, n \text{ even} \Rightarrow c_m = a_{\frac{1}{2}m},\ c_n = a_{\frac{1}{2}n} \Rightarrow d(c_m, c_n) < \tfrac{1}{2}\epsilon < \epsilon$$

$$m, n \text{ odd} \Rightarrow c_m = b_{\frac{1}{2}(m+1)},\ c_n = b_{\frac{1}{2}(n+1)} \Rightarrow d(c_m, c_n) < \tfrac{1}{2}\epsilon < \epsilon$$

$$m \text{ even}, n \text{ odd} \Rightarrow c_m = a_{\frac{1}{2}m},\ c_n = b_{\frac{1}{2}(n+1)} \Rightarrow$$

$$d(c_m, c_n) \leqq d(a_{\frac{1}{2}m}, b_{\frac{1}{2}m}) + d(b_{\frac{1}{2}m}, b_{\frac{1}{2}(n+1)}) < \tfrac{1}{2}\epsilon + \tfrac{1}{2}\epsilon = \epsilon$$

and so $\langle c_n \rangle$ is a Cauchy sequence.

Conversely, if there exists a Cauchy sequence $\langle c_n \rangle$ for which $\langle a_n \rangle = \langle c_{j_n} \rangle$ and $\langle b_n \rangle = \langle c_{k_n} \rangle$, then

$$\lim_{n\to\infty} d(a_n, b_n) = \lim_{n\to\infty} d(c_{j_n}, c_{k_n}) = 0$$

since $\langle c_n \rangle$ is Cauchy and $n \to \infty$ implies $j_n, k_n \to \infty$.

10. Prove Lemma 14.5: The function e is well-defined, i.e. $\langle a_n \rangle \sim \langle a_n^* \rangle$ and $\langle b_n \rangle \sim \langle b_n^* \rangle$ implies $\lim_{n\to\infty} d(a_n, b_n) = \lim_{n\to\infty} d(a_n^*, b_n^*)$.

Solution:

Set $r = \lim d(a_n, b_n)$ and $r^* = \lim d(a_n^*, b_n^*)$, and let $\epsilon > 0$. Note that

$$d(a_n, b_n) \leq d(a_n, a_n^*) + d(a_n^*, b_n^*) + d(b_n^*, b_n)$$

Now $\exists\, n_1 \in \mathbf{N}$ such that $n > n_1 \Rightarrow d(a_n, a_n^*) < \epsilon/3$

$\exists\, n_2 \in \mathbf{N}$ such that $n > n_2 \Rightarrow d(b_n, b_n^*) < \epsilon/3$

$\exists\, n_3 \in \mathbf{N}$ such that $n > n_3 \Rightarrow |d(a_n^*, b_n^*) - r^*| < \epsilon/3$

Accordingly, if $n > \max(n_1, n_2, n_3)$, then

$$d(a_n, b_n) < r^* + \epsilon \quad \text{and so} \quad \lim_{n\to\infty} d(a_n, b_n) = r \leq r^* + \epsilon$$

But this inequality holds for every $\epsilon > 0$; hence $r \leq r^*$. In the same manner we may show that $r^* \leq r$; thus $r = r^*$.

11. Let $\langle a_n \rangle$ be a Cauchy sequence in X. Show that $\alpha = [\langle a_n \rangle] \in X^*$ is the limit of the sequence $\langle \hat{a}_1, \hat{a}_2, \ldots \rangle$ in $\hat{X}$. (Here $\hat{X} = \{\hat{p} = [\langle p, p, p, \ldots \rangle] : p \in X\}$.)

Solution:

Since $\langle a_n \rangle$ is a Cauchy sequence in X,

$$\lim_{m\to\infty} e(\hat{a}_m, \alpha) = \lim_{m\to\infty}\left(\lim_{n\to\infty} d(a_m, a_n)\right) = \lim_{\substack{m\to\infty\\ n\to\infty}} d(a_m, a_n) = 0$$

Accordingly, $\langle \hat{a}_n \rangle \to \alpha$.

12. Prove Lemma 14.7: X is isometric to $\hat{X}$, and $\hat{X}$ is dense in X^*.

Solution:

For every $p, q \in X$,

$$e(\hat{p}, \hat{q}) = \lim_{n\to\infty} d(p, q) = d(p, q)$$

and so X is isometric to $\hat{X}$. We show that $\hat{X}$ is dense in X^* by showing that every point in X^* is the limit of a sequence in $\hat{X}$. Let $\alpha = [\langle a_1, a_2, \ldots \rangle]$ be an arbitrary point in X^*. Then $\langle a_n \rangle$ is a Cauchy sequence in X and so, by the preceding problem, α is the limit of the sequence $\langle \hat{a}_1, \hat{a}_2, \ldots \rangle$ in $\hat{X}$. Thus $\hat{X}$ is dense in X^*.

13. Prove Lemma 14.8: Every Cauchy sequence in (X^*, e) converges, and so (X^*, e) is a completion of X.

Solution:

Let $\langle \alpha_1, \alpha_2, \ldots \rangle$ be a Cauchy sequence in X^*. Since $\hat{X}$ is dense in X^*, for every $n \in \mathbf{N}$,

$$\exists\, \hat{a}_n \in \hat{X} \quad \text{such that} \quad e(\hat{a}_n, \alpha_n) < 1/n$$

Then (Problem 5) $\langle \hat{a}_1, \hat{a}_2, \ldots \rangle$ is also a Cauchy sequence and, by Problem 12, $\langle \hat{a}_1, \hat{a}_2, \ldots \rangle$ converges to $\beta = [\langle a_1, a_2, \ldots \rangle] \in X^*$. Hence (Problem 5) $\langle \alpha_n \rangle$ also converges to β and therefore (X^*, e) is complete.

14. Prove Lemma 14.9: If Y^* is a completion of X, then Y^* is isometric to X^*.

Solution:

We can assume X is a subspace of Y^*. Hence, for every $p \in Y^*$, there exists a sequence $\langle a_1, a_2, \ldots \rangle$ in X converging to p; and in particular, $\langle a_n \rangle$ is a Cauchy sequence. Let $f: Y^* \to X^*$ be defined by

$$f(p) = [\langle a_1, a_2, \ldots \rangle]$$

Now if $\langle a_1^*, a_2^*, \ldots \rangle \in X$ also converges to p, then

$$\lim_{n\to\infty} d(a_n, a_n^*) = 0 \quad \text{and so} \quad [\langle a_n \rangle] = [\langle a_n^* \rangle]$$

In other words, f is well-defined.

Furthermore, f is onto. For if $[\langle b_1, b_2, \ldots \rangle] \in X^*$, then $\langle b_1, b_2, \ldots \rangle$ is a Cauchy sequence in $X \subset Y^*$ and, since Y^* is complete, $\langle b_n \rangle$ converges to, say, $q \in Y^*$. Accordingly, $f(q) = [\langle b_n \rangle]$.

Now let $p, q \in Y^*$ with, say, sequences $\langle a_n \rangle$ and $\langle b_n \rangle$ in X converging, respectively, to p and q. Then

$$e(f(p), f(q)) = e([\langle a_n \rangle], [\langle b_n \rangle]) = \lim_{n \to \infty} d(a_n, b_n) = d\left(\lim_{n \to \infty} a_n, \lim_{n \to \infty} b_n\right) = d(p, q)$$

Consequently, f is an isometry between Y^* and X^*.

BAIRE'S CATEGORY THEOREM

15. Let N be a nowhere dense subset of X. Show that $\overline{N}^c$ is dense in X.

Solution:

Suppose $\overline{N}^c$ is not dense in X, i.e. $\exists\, p \in X$ and an open set G such that

$$p \in G \quad \text{and} \quad G \cap \overline{N}^c = \emptyset$$

Then $p \in G \subset \overline{N}$ and so $p \in \text{int}(\overline{N})$. But this is impossible since N is nowhere dense in X, i.e. $\text{int}(\overline{N}) = \emptyset$. Therefore $\overline{N}^c$ is dense in X.

16. Let G be an open subset of the metric space X and let N be nowhere dense in X. Show that there exist $p \in X$ and $\delta > 0$ such that $S(p, \delta) \subset G$ and $S(p, \delta) \cap N = \emptyset$.

Solution:

Set $H = G \cap \overline{N}^c$. Then $H \subset G$ and $H \cap N = \emptyset$. Furthermore, H is non-empty since G is open and $\overline{N}^c$ is dense in X; say, $p \in H$. But H is open since G and $\overline{N}^c$ are open; hence $\exists\, \delta > 0$ such that $S(p, \delta) \subset H$. Consequently, $S(p, \delta) \subset G$ and $S(p, \delta) \cap N = \emptyset$.

17. Prove Theorem 14.11: Every complete metric space X is of second category.

Solution:

Let $M \subset X$ and let M be of first category. We want to show that $M \neq X$, i.e. $\exists\, p \in X$ such that $p \notin M$. Since M is of first category, $M = N_1 \cup N_2 \cup \cdots$ where each N_i is nowhere dense in X.

Since N_1 is nowhere dense in X, $\exists\, a_1 \in X$ and $\delta_1 > 0$ such that $S(a_1, \delta_1) \cap N_1 = \emptyset$. Set $\epsilon_1 = \delta_1/2$. Then

$$\overline{S(a_1, \epsilon_1)} \cap N_1 = \emptyset$$

Now $S(a_1, \epsilon_1)$ is open and N_2 is nowhere dense in X, and so, by Problem 16,

$$\exists\, a_2 \in X \text{ and } \delta_2 > 0 \quad \text{such that} \quad S(a_2, \delta_2) \subset S(a_1, \epsilon_1) \subset \overline{S(a_1, \epsilon_1)} \quad \text{and} \quad S(a_2, \delta_2) \cap N_2 = \emptyset$$

Set $\epsilon_2 = \delta_2/2 \leq \epsilon_1/2 = \delta_1/4$. Then

$$\overline{S(a_2, \epsilon_2)} \subset \overline{S(a_1, \epsilon_1)} \quad \text{and} \quad \overline{S(a_2, \epsilon_2)} \cap N_2 = \emptyset$$

Continuing in this manner, we obtain a nested sequence of closed sets

$$\overline{S(a_1, \epsilon_1)} \supset \overline{S(a_2, \epsilon_2)} \supset \overline{S(a_3, \epsilon_3)} \supset \cdots$$

such that, for every $n \in N$, $\overline{S(a_n, \epsilon_n)} \cap N_n = \emptyset$ and $\epsilon_n \leq \delta_1/2^n$

Thus $\lim_{n \to \infty} \epsilon_n \leq \lim_{n \to \infty} \delta_1/2^n = 0$ and so, by Theorem 14.2,

$$\exists\, p \in X \quad \text{such that} \quad p \in \cap_{n=1}^{\infty} \overline{S(a_n, \epsilon_n)}$$

Furthermore, for every $n \in N$, $p \notin N_n$ and so $p \notin M$.

COMPLETENESS AND COMPACTNESS

18. Show that every compact metric space X is complete.

Solution:

Let $\langle a_1, a_2, \ldots \rangle$ be a Cauchy sequence in X. Now X is compact and so sequentially compact; hence $\langle a_n \rangle$ contains a subsequence $\langle a_{i_1}, a_{i_2}, \ldots \rangle$ which converges to, say, $p \in X$. But (Problem 4) $\langle a_n \rangle$ also converges to p. Hence X is complete.

19. Let E be a totally bounded subset of a metric space X. Show that every sequence $\langle a_n \rangle$ in E contains a Cauchy subsequence.

Solution:

Since E is totally bounded, we can decompose E into a finite number of subsets of diameter less than $\epsilon_1 = 1$. One of these sets, call it A_1, must contain an infinite number of the terms of the sequence; hence

$$\exists\ i_1 \in \mathbf{N} \quad \text{such that} \quad a_{i_1} \in A_1$$

Now A_1 is totally bounded and can be decomposed into a finite number of subsets of diameter less than $\epsilon_2 = \frac{1}{2}$. Similarly, one of these sets, call it A_2, must contain an infinite number of the terms of the sequence; hence

$$\exists\ i_2 \in \mathbf{N} \quad \text{such that} \quad i_2 > i_1 \text{ and } a_{i_2} \in A_2$$

Furthermore, $A_2 \subset A_1$.

We continue in this manner and obtain a nested sequence of sets

$$E \supset A_1 \supset A_2 \supset \cdots \quad \text{with} \quad d(A_n) < 1/n$$

and a subsequence $\langle a_{i_1}, a_{i_2}, \ldots \rangle$ of $\langle a_n \rangle$ with $a_{i_n} \in A_n$. We claim that $\langle a_{i_n} \rangle$ is a Cauchy sequence. For, let $\epsilon > 0$; then

$$\exists\ n_0 \in \mathbf{N} \quad \text{such that} \quad 1/n_0 < \epsilon \quad \text{and so } d(A_{n_0}) < \epsilon$$

Therefore $$i_n, i_m > i_{n_0} \quad \Rightarrow \quad a_{i_n}, a_{i_m} \in A_{n_0} \quad \Rightarrow \quad d(a_{i_m}, d_{i_n}) < \epsilon$$

20. Prove Theorem 14.12: A metric space X is compact if and only if X is complete and totally bounded.

Solution:

Suppose X is compact. Then, by Problem 15, X is complete and, by Lemma 11.17, Page 158, X is totally bounded.

On the other hand, suppose X is complete and totally bounded. Let $\langle a_1, a_2, \ldots \rangle$ be a sequence in X. Then, by the preceding problem, $\langle a_n \rangle$ contains a Cauchy subsequence $\langle a_{i_n} \rangle$ which converges since X is complete. Thus X is sequentially compact and therefore compact.

21. Prove Theorem 14.13: Let A be a subset of a complete metric space X. Then the following are equivalent: (i) A is compact. (ii) A is closed and totally bounded.

Solution:

If A is compact, then by Theorem 11.5 and Lemma 11.17 it is closed and totally bounded.

Conversely, suppose A is closed and totally bounded. Now a closed subset of a complete space is complete, and so A is complete and totally bounded. Hence, by the preceding problem, A is compact.

Supplementary Problems

COMPLETE METRIC SPACES

22. Let (X, d) be a metric space and let e be the metric on X defined by $e(a, b) = \min\{1, d(a, b)\}$. Show that $\langle a_n \rangle$ is a Cauchy sequence in (X, d) if and only if $\langle a_n \rangle$ is a Cauchy sequence in (X, e).

23. Show that every finite metric space is complete.

24. Prove: Every closed subspace of a complete metric space is complete.

25. Prove that Hilbert Space (l_2-space) is complete.

26. Prove: Let $\mathcal{B}(X, \mathbf{R})$ be the collection of bounded real-valued functions defined on X with norm

$$||f|| = \sup\{|f(x)| : x \in X\}$$

Then $\mathcal{B}(X, \mathbf{R})$ is complete.

27. Prove: A metric space X is complete if and only if every infinite totally bounded subset of X has an accumulation point.

28. Show that a countable union of first category sets is of first category.

29. Show that a metric space X is totally bounded if and only if every sequence in X contains a Cauchy subsequence.

30. Show that if X is isometric to Y and X is complete, then Y is complete.

MISCELLANEOUS PROBLEM

31. Prove: Every normed vector space X can be densely embedded in a Banach space, i.e. a complete normed vector space. (*Hint*: See Remark on Page 199).

Chapter 15

Function Spaces

FUNCTION SPACES

Let X and Y be arbitrary sets, and let $\mathcal{F}(X, Y)$ denote the collection of all functions from X into Y. Any subcollection of $\mathcal{F}(X, Y)$ with some topology $\mathcal{T}$ is called a *function space*.

We can identify $\mathcal{F}(X, Y)$ with a product set as follows: Let Y_x denote a copy of Y indexed by $x \in X$, and let $\mathbf{F}$ denote the product of the sets Y_x, i.e.,

$$\mathbf{F} = \prod \{Y_x : x \in X\}$$

Recall that $\mathbf{F}$ consists of all points $p = \langle a_x : x \in X \rangle$ which assign to each $x \in X$ the element $a_x \in Y_x = Y$, i.e. $\mathbf{F}$ consists of all functions from X into Y, and so $\mathbf{F} = \mathcal{F}(X, Y)$.

Now for each element $x \in X$, the mapping e_x from the function set $\mathcal{F}(X, Y)$ into Y defined by

$$e_x(f) = f(x)$$

is called the *evaluation mapping* at x. (Here f is any function in $\mathcal{F}(X, Y)$, i.e. $f : X \to Y$.) Under our identification of $\mathcal{F}(X, Y)$ with $\mathbf{F}$, the evaluation mapping e_x is precisely the projection mapping π_x from $\mathbf{F}$ into the coordinate space $Y_x = Y$.

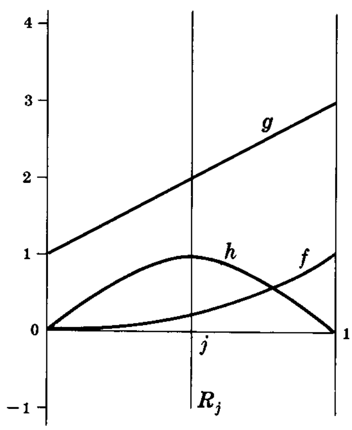
Example 1.1: Let $\mathcal{F}(I, \mathbf{R})$ be the collection of all real-valued functions defined on $I = [0, 1]$, and let $f, g, h \in \mathcal{F}(I, \mathbf{R})$ be the functions

$$f(x) = x^2, \quad g(x) = 2x + 1, \quad h(x) = \sin \pi x$$

Consider the evaluation function $e_j : \mathcal{F}(I, \mathbf{R}) \to \mathbf{R}$ at, say, $j = \frac{1}{2}$. Then

$$e_j(f) = f(j) = f(\tfrac{1}{2}) = \tfrac{1}{4}$$
$$e_j(g) = g(j) = g(\tfrac{1}{2}) = 2$$
$$e_j(h) = h(j) = h(\tfrac{1}{2}) = 1$$

Graphically, $e_j(f)$, $e_j(g)$ and $e_j(h)$ are the points where the graphs of f, g and h intersect the vertical line R_j through $x = j$.



POINT OPEN TOPOLOGY

Let X be an arbitrary set and let Y be a topological space. We first investigate the product topology $\mathcal{T}$ on $\mathcal{F}(X, Y)$ where we identify $\mathcal{F}(X, Y)$ with the product set $\mathbf{F} = \prod \{Y_x : x \in X\}$ as above. Recall that the defining subbase $\mathcal{S}$ of the product topology on $\mathbf{F}$ consists of all subsets of $\mathbf{F}$ of the form

$$\pi_{x_0}^{-1}[G] = \{f : \pi_{x_0}(f) \in G\}$$

where $x_0 \in X$ and G is an open subset of the coordinate space $Y_{x_0} = Y$. But $\pi_{x_0}(f) = e_{x_0}(f) = f(x_0)$, where e_{x_0} is the evaluation mapping at $x_0 \in X$. Hence the defining subbase $\mathcal{S}$ of the product topology $\mathcal{T}$ on $\mathcal{F}(X, Y)$ consists of all subsets of $\mathcal{F}(X, Y)$ of the form

$\{f : f(x_0) \in G\}$, i.e. all functions which map an arbitrary point $x_0 \in X$ into an arbitrary open set G of Y. We call this product topology on $\mathcal{F}(X, Y)$, appropriately, the *point open topology*.

Alternatively, we can define the point open topology on $\mathcal{F}(X, Y)$ to be the coarsest topology on $\mathcal{F}(X, Y)$ with respect to which the evaluation functions $e_x : \mathcal{F}(X, Y) \to Y$ are continuous. This definition corresponds directly to the definition of the product topology.

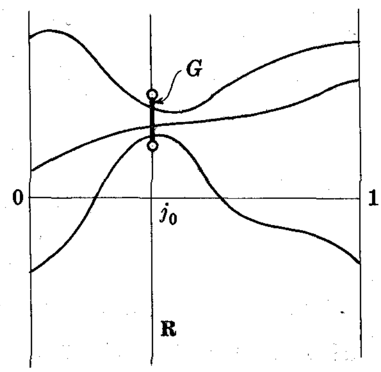
Example 2.1: Let $\mathcal{T}$ be the point open topology on $\mathcal{F}(I, \mathbf{R})$ where $I = [0, 1]$. As above, members of the defining subbase of $\mathcal{T}$ are of the form

$$\{f : f(j_0) \in G\}$$

where $j_0 \in I$ and G is an open subset of $\mathbf{R}$. Graphically, the above subbase element consists of all functions passing through the open set G on the vertical real line $\mathbf{R}$ through the point j_0 on the horizontal axis. Recall that this is identical to the subbase element of the product space

$$X = \prod \{R_i : i \in I\}$$

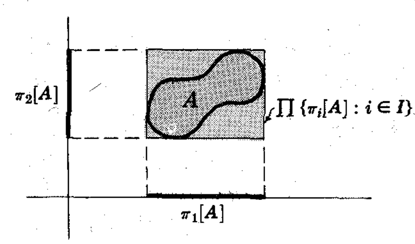
illustrated in Chapter 12, Page 170.



Example 2.2: If A is a subset of a product space $\prod \{X_i : i \in I\}$, then A is a subset of the product of its projections, i.e.

$$A \subset \prod \{\pi_i[A] : i \in I\}$$

(as indicated in the diagram).



Thus $A \subset \prod \{\overline{\pi_i[A]} : i \in I\}$ where $\overline{\pi_i[A]}$ is the closure of $\pi_i[A]$. Accordingly, if $\mathcal{A} = \mathcal{A}(X, Y)$ is a subcollection of $\mathcal{F}(X, Y)$, then

$$\mathcal{A} \subset \prod \{\overline{\pi_x[\mathcal{A}]} : x \in X\} = \prod \{\overline{e_x[\mathcal{A}]} : x \in X\}$$

and $\overline{e_x[\mathcal{A}]} = \overline{\{f(x) : f \in \mathcal{A}\}}$. By the Tychonoff Product Theorem, if $\overline{\{f(x) : x \in X\}}$ is compact for every $x \in X$, then $\prod \{\overline{\pi_x[\mathcal{A}]} : x \in X\}$ is a compact subset of the product space $\prod \{Y_x : x \in X\}$.

Recall that a closed subset of a compact set is compact. Hence the result of Example 2.2 implies

Theorem 15.1: Let $\mathcal{A}$ be a subcollection of $\mathcal{F}(X, Y)$. Then $\mathcal{A}$ is compact with respect to the point open topology on $\mathcal{F}(X, Y)$ if (i) $\mathcal{A}$ is a closed subset of $\mathcal{F}(X, Y)$ and (ii) for every $x \in X$, $\overline{\{f(x) : f \in \mathcal{A}\}}$ is compact in Y.

In the case that Y is Hausdorff we have the following stronger result:

Theorem 15.2: Let Y be a Hausdorff space and let $\mathcal{A} \subset \mathcal{F}(X, Y)$. Then $\mathcal{A}$ is compact with respect to the point open topology if and only if $\mathcal{A}$ is closed and, for every $x \in X$, $\overline{\{f(x) : f \in \mathcal{A}\}}$ is compact.

POINTWISE CONVERGENCE

Let $\langle f_1, f_2, \ldots \rangle$ be a sequence of functions from an arbitrary set X into a topological space Y. The sequence $\langle f_n \rangle$ is said to converge *pointwise* to a function $g : X \to Y$ if, for every $x_0 \in X$,

$$\langle f_1(x_0), f_2(x_0), \ldots \rangle \text{ converges to } g(x_0), \quad \text{i.e. } \lim_{n \to \infty} f_n(x_0) = g(x_0)$$

In particular, if Y is a metric space then $\langle f_n \rangle$ converges pointwise to g iff for every $\epsilon > 0$ and every $x_0 \in X$,

$$\exists\ n_0 = n_0(x_0, \epsilon) \in \mathbf{N} \quad \text{such that} \quad n > n_0 \ \Rightarrow\ d(f_n(x_0), g(x_0)) < \epsilon$$

Note that the n_0 depends upon the ϵ and also upon the point x_0.

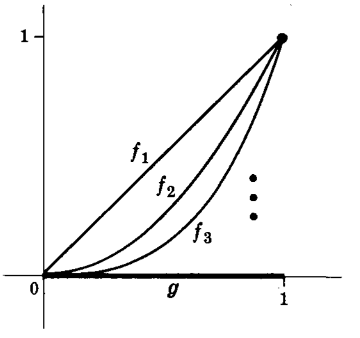
Example 3.1: Let $\langle f_1, f_2, \ldots \rangle$ be the sequence of functions from $I = [0,1]$ into $\mathbf{R}$ defined by

$$f_1(x) = x, \ f_2(x) = x^2, \ f_3(x) = x^3, \ \ldots$$

Then $\langle f_n \rangle$ converges pointwise to the function $g : I \to \mathbf{R}$ defined by

$$g(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x = 1 \end{cases}$$

Observe that the limit function g is not continuous even though each of the functions f_i is continuous.



The notion of pointwise convergence is related to the point open topology as follows:

Theorem 15.3: A sequence of functions $\langle f_1, f_2, \ldots \rangle$ in $\mathcal{F}(X,Y)$ converges to $g \in \mathcal{F}(X,Y)$ with respect to the point open topology on $\mathcal{F}(X,Y)$ if and only if $\langle f_n \rangle$ converges pointwise to g.

In view of the above theorem, the point open topology on $\mathcal{F}(X,Y)$ is also called the *topology of pointwise convergence.*

Remark: Recall that metrizability is not invariant under uncountable products; therefore, the topology of pointwise convergence of real-valued functions defined on $[0,1]$ is not a metric topology. The theory of topological spaces, as a generalization of metric spaces, was first motivated by the study of pointwise convergence of functions.

UNIFORM CONVERGENCE

Let $\langle f_1, f_2, \ldots \rangle$ be a sequence of functions from an arbitrary set X into a metric space (Y,d). Then $\langle f_n \rangle$ is said to converge uniformly to a function $g : X \to Y$ if, for every $\epsilon > 0$,

$$\exists\ n_0 = n_0(\epsilon) \in \mathbf{N} \quad \text{such that} \quad n > n_0 \ \Rightarrow\ d(f_n(x), g(x)) < \epsilon, \ \forall\, x \in X$$

In particular, $\langle f_n \rangle$ converges pointwise to g; that is, uniform convergence implies pointwise convergence. Observe that the n_0 depends only on the ϵ, whereas, in pointwise convergence, the n_0 depends on both the ϵ and the point x.

In the case where X is a topological space, we have the following classical result:

Proposition 15.4: Let $\langle f_1, f_2, \ldots \rangle$ be a sequence of continuous functions from a topological space X into a metric space Y. If $\langle f_n \rangle$ converges uniformly to $g : X \to Y$, then g is continuous.

Example 4.1: Let $f_1, f_2, \ldots$ be the following continuous functions from $I = [0,1]$ into $\mathbf{R}$:

$$f_1(x) = x, \ f_2(x) = x^2, \ f_3(x) = x^3, \ \ldots$$

Now, by Example 3.1, $\langle f_n \rangle$ converges pointwise to $g : I \to \mathbf{R}$ defined by

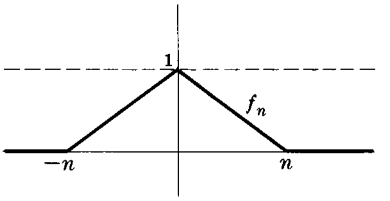
$$g(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x = 1 \end{cases}$$

Since g is not continuous, $\langle f_n \rangle$ does not converge uniformly to g.

Example 4.2: Let $\langle f_1, f_2, \ldots\rangle$ be the following sequence of functions in $\mathcal{F}(\mathbf{R}, \mathbf{R})$:

$$f_n(x) = \begin{cases} 1 - \frac{1}{n}|x| & \text{if } |x| < n \\ 0 & \text{if } |x| \geq n \end{cases}$$

Now $\langle f_n\rangle$ converges pointwise to the constant function $g(x) = 1$. But $\langle f_n\rangle$ does not converge uniformly to g. For, let $\epsilon = \frac{1}{2}$. Note that, for every $n \in \mathbf{N}$, there exist points $x_0 \in \mathbf{R}$ with $f_n(x_0) = 0$ and so $|f_n(x_0) - g(x_0)| = 1 > \epsilon$.



Let $\mathcal{B}(X, Y)$ denote the collection of all bounded functions from an arbitrary set X into a metric space (Y, d), and let e be the metric on $\mathcal{B}(X, Y)$ defined by

$$e(f, g) = \sup\{d(f(x), g(x)) : x \in X\}$$

This metric has the following property:

Theorem 15.5: Let $\langle f_1, f_2, \ldots\rangle$ be a sequence of functions in $\mathcal{B}(X, Y)$. Then $\langle f_n\rangle$ converges to $g \in \mathcal{B}(X, Y)$ with respect to the metric e if and only if $\langle f_n\rangle$ converges uniformly to g.

In view of the above theorem, the topology on $\mathcal{B}(X, Y)$ induced by the above metric is called the *topology of uniform convergence.*

Remark: The concept of uniform convergence defined in the case of a metric space Y cannot be defined for a general topological space. However, the notion of uniform convergence can be generalized to a collection of spaces, called *uniform spaces,* which lie between topological spaces and metric spaces.

THE FUNCTION SPACE $C[0, 1]$

The vector space $C[0, 1]$ of all continuous functions from $I = [0, 1]$ into $\mathbf{R}$ with norm defined by

$$||f|| = \sup\{|f(x)| : x \in I\}$$

is one of the most important function spaces in analysis. Note that the above norm induces the topology of uniform convergence.

Since $I = [0, 1]$ is compact, each $f \in C[0, 1]$ is *uniformly continuous*; that is,

Proposition 15.6: Let $f: [0, 1] \to \mathbf{R}$ be continuous. Then for every $\epsilon > 0$,

$$\exists\ \delta = \delta(\epsilon) > 0 \quad \text{such that} \quad |x_0 - x_1| < \delta \Rightarrow |f(x_0) - f(x_1)| < \epsilon$$

Uniform continuity (like uniform convergence) is stronger than continuity in that the δ depends only on ϵ and not on any particular point.

One consequence of Proposition 15.4 follows:

Theorem 15.7: $C[0, 1]$ is a complete normed vector space.

We shall use the Baire Category Theorem for complete metric spaces to prove the following interesting result:

Proposition 15.8: There exists a continuous function $f: [0, 1] \to \mathbf{R}$ which is nowhere differentiable.

Remark: All the results proven here for $C[0, 1]$ are also true for the space $C[a, b]$ of all continuous functions on the closed interval $[a, b]$.

UNIFORM BOUNDEDNESS

In establishing necessary and sufficient conditions for subsets of function spaces to be compact, we are led to the concepts of *uniform boundedness* and *equicontinuity* which are interesting in their own right.

A collection of real valued functions $\mathcal{A} = \{f_i : X \to \mathbf{R}\}$ defined on an arbitrary set X is said to be *uniformly bounded* if

$$\exists\, M \in \mathbf{R} \quad \text{such that} \quad |f(x)| \leq M, \ \forall f \in \mathcal{A}, \ \forall x \in X$$

That is, each function $f \in \mathcal{A}$ is bounded and there is one bound which holds for all of the functions.

In particular if $\mathcal{A} \subset C[0,1]$, then uniform boundedness is equivalent to

$$\exists\, M \in \mathbf{R} \quad \text{such that} \quad \|f\| \leq M, \ \forall f \in \mathcal{A}$$

or, $\mathcal{A}$ is a bounded subset of $C[0,1]$.

Example 5.1: Let $\mathcal{A}$ be the following subset of $\mathcal{F}(\mathbf{R},\mathbf{R})$:

$$\mathcal{A} = \{f_1(x) = \sin x,\ f_2(x) = \sin 2x,\ \ldots\}$$

Then $\mathcal{A}$ is uniformly bounded. For, let $M = 1$; then, for every $f \in \mathcal{A}$ and every $x \in \mathbf{R}$, $|f(x)| \leq M$. See Fig. (*a*) below.

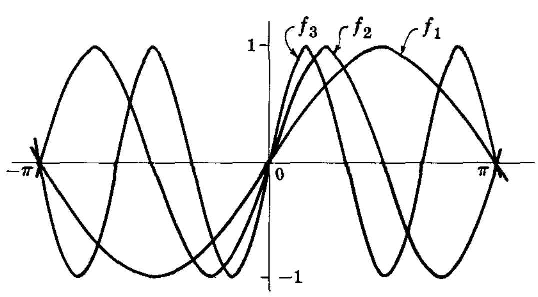


Fig. (*a*)

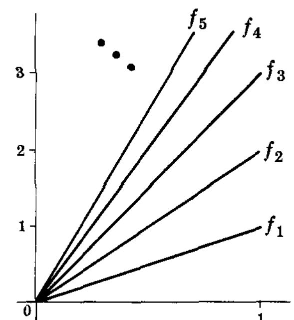


Fig. (*b*)

Example 5.2: Let $\mathcal{A} \subset C[0,1]$ be defined as follows (see Fig. (*b*) above):

$$\mathcal{A} = \{f_1(x) = x,\ f_2(x) = 2x,\ f_3(x) = 3x,\ \ldots\}$$

Although each function in $C[0,1]$, and in particular in $\mathcal{A}$, is bounded, $\mathcal{A}$ is not uniformly bounded. For if M is any real number, however large, $\exists\, n_0 \in \mathbf{N}$ with $n_0 > M$ and hence $f_{n_0}(1) = n_0 > M$.

EQUICONTINUITY. ASCOLI'S THEOREM

A collection of real-valued functions $\mathcal{A} = \{f_i : X \to \mathbf{R}\}$ defined on an arbitrary metric space X is said to be *equicontinuous* if for every $\epsilon > 0$,

$$\exists\, \delta = \delta(\epsilon) > 0 \quad \text{such that} \quad d(x_0, x_1) < \delta \ \Rightarrow\ |f(x_0) - f(x_1)| < \epsilon, \quad \forall f \in \mathcal{A}$$

Note that δ depends only on ϵ and not on any particular point or function. It is clear that each $f \in \mathcal{A}$ is uniformly continuous.

Theorem (Ascoli) 15.9: Let $\mathcal{A}$ be a closed subset of the function space $C[0,1]$. Then $\mathcal{A}$ is compact if and only if $\mathcal{A}$ is uniformly bounded and equicontinuous.

COMPACT OPEN TOPOLOGY

Let X and Y be arbitrary sets and let $A \subset X$ and $B \subset Y$. We shall write $F(A,B)$ for the class of functions from X into Y which carry A into B:

$$F(A,B) = \{f \in \mathcal{F}(X,Y) : f[A] \subset B\}$$

Example 6.1: Let $\mathcal{S}$ be the defining subbase for the point open topology on $\mathcal{F}(X,Y)$. Recall that the members of $\mathcal{S}$ are of the form

$$\{f \in \mathcal{F}(X,Y) : f(x) \in G\}, \quad \text{where } x \in X,\ G \text{ an open subset of } Y$$

Following the above notation, we denote this set by $F(x,G)$ and we can then define $\mathcal{S}$ by

$$\mathcal{S} = \{F(x,G) : x \in X,\ G \subset Y \text{ open}\}$$

Now let X and Y be topological spaces and let $\mathcal{A}$ be the class of compact subsets of X and $\mathcal{G}$ be the class of open subsets of Y. The topology $\mathcal{T}$ on $\mathcal{F}(X,Y)$ generated by

$$\mathcal{S} = \{F(A,G) : A \in \mathcal{A},\ G \in \mathcal{G}\}$$

is called the *compact open topology* on $\mathcal{F}(X,Y)$, and $\mathcal{S}$ is a defining subbase for $\mathcal{T}$.

Since singleton subsets of X are compact, $\mathcal{S}$ contains the members of the defining subbase for the point open topology on $\mathcal{F}(X,Y)$. Thus:

Theorem 15.10: The point open topology on $\mathcal{F}(X,Y)$ is coarser than the compact open topology on $\mathcal{F}(X,Y)$.

Recall that the point open topology is the coarsest topology with respect to which the evaluation mappings are continuous. Hence,

Corollary 15.11: The evaluation functions $e_x : \mathcal{F}(X,Y) \to Y$ are continuous relative to the compact open topology on $\mathcal{F}(X,Y)$.

TOPOLOGY OF COMPACT CONVERGENCE

Let $\langle f_1, f_2, \ldots \rangle$ be a sequence of functions from a topological space X into a metric space (Y,d). The sequence $\langle f_n \rangle$ is said to *converge uniformly on compacta* to $g : X \to Y$ if for every compact subset $E \subset X$ and every $\epsilon > 0$,

$$\exists\ n_0 = n_0(E,\epsilon) \in \mathbf{N} \quad \text{such that} \quad n > n_0 \ \Rightarrow\ d(f_n(x), g(x)) < \epsilon, \quad \forall x \in E$$

In other words, $\langle f_n \rangle$ converges uniformly on compacta to g iff, for every compact subset $E \subset X$, the restriction of $\langle f_n \rangle$ to E converges uniformly to the restriction of g to E, i.e.,

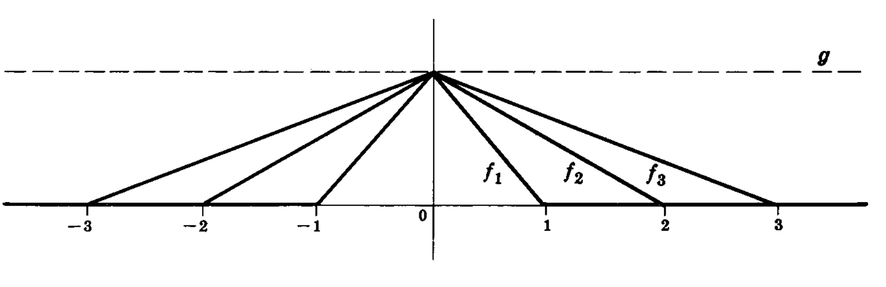
$$\langle f_1|E, f_2|E, \ldots \rangle \quad \text{converges uniformly to} \quad g|E$$

Now uniform convergence implies uniform convergence on compacta and, since singleton sets are compact, uniform convergence on compacta implies pointwise convergence.

Example 7.1: Let $\langle f_1, f_2, \ldots \rangle$ be the sequence in $\mathcal{F}(\mathbf{R},\mathbf{R})$ defined by

$$f_n(x) = \begin{cases} 1 - \frac{1}{n}|x| & \text{if } |x| < n \\ 0 & \text{if } |x| \geqq n \end{cases}$$

Now $\langle f_n \rangle$ converges pointwise to the constant function $g(x) = 1$ but $\langle f_n \rangle$ does not converge uniformly to g (see Example 4.2). However, since every compact subset E of R is bounded, $\langle f_n \rangle$ does converge uniformly on compacta to g.



Theorem 15.12: Let $C(X, Y)$ be the collection of continuous functions from a topological space X into a metric space (Y, d). Then a sequence of functions $\langle f_n \rangle$ in $C(X, Y)$ converges to $g \in C(X, Y)$ with respect to the compact open topology if and only if $\langle f_n \rangle$ converges uniformly on compacta to g.

In view of the preceding theorem, the compact open topology is also called the *topology of compact convergence.*

FUNCTIONALS ON NORMED SPACES

Let X be a normed vector space (over $\mathbf{R}$). A real-valued function f with domain X, i.e. $f: X \to \mathbf{R}$, is called a *functional.*

Definition: A functional f on X is *linear* if

(i) $f(x+y) = f(x) + f(y),\ \forall x, y \in X,$ and (ii) $f(kx) = k[f(x)],\ \forall x \in X,\ k \in \mathbf{R}$

A linear functional f on X is *bounded* if

$$\exists\, M > 0 \quad \text{such that} \quad |f(x)| \leq M\,||x||,\ \forall x \in X$$

Here M is called a *bound* for f.

Example 8.1: Let X be the space of all continuous real-valued functions on $[a, b]$ with norm $||f|| = \sup\{|f(x)| : x \in [a, b]\}$, i.e. $X = C[a, b]$. Let $\mathbf{I}: X \to \mathbf{R}$ be defined by

$$\mathbf{I}(f) = \int_a^b f(t)\, dt$$

Then $\mathbf{I}$ is a linear functional; for

$$\mathbf{I}(f+g) = \int_a^b (f(t)+g(t))\, dt = \int_a^b f(t)\, dt + \int_a^b g(t)\, dt = \mathbf{I}(f) + \mathbf{I}(g)$$

$$\mathbf{I}(kf) = \int_a^b (kf)(t)\, dt = \int_a^b k[f(t)]\, dt = k\int_a^b f(t)\, dt = k\,\mathbf{I}(f)$$

Furthermore, $M = b - a$ is a bound for $\mathbf{I}$ since

$$\mathbf{I}(f) = \int_a^b f(t)\, dt \leq M \sup\{|f(t)|\} = M\,||f||$$

Proposition 15.13: Let f and g be bounded linear functionals on X and let $k \in \mathbf{R}$. Then $f+g$ and $k \cdot f$ are also bounded linear functionals on X.

Thus (by Proposition 8.14, Page 119) the collection X^* of all bounded linear functionals on X is a linear vector space.

Proposition 15.14: The following function on X^* is a norm:

$$||f|| = \sup\{|f(x)|/||x|| : x \neq 0\}$$

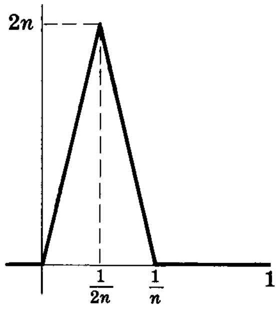
Observe that if M is a bound for f, i.e. $|f(x)| \leq M\,||x||,\ \forall x \in X,$ then in particular, for $x \neq 0$, $|f(x)|/||x|| \leq M$ and so $||f|| \leq M$. In fact, $||f||$ could have been defined equivalently by

$$||f|| = \inf\{M : M \text{ is a bound for } f\}$$

Remark: The normed space of all bounded linear functionals on X is called the *dual space* of X.

Solved Problems

POINTWISE CONVERGENCE, POINT OPEN TOPOLOGY



1. Let $\langle f_1, f_2, \ldots \rangle$ be the sequence of functions in $\mathcal{F}(I, \mathbf{R})$, where $I = [0,1]$, defined by

$$f_n(x) = \begin{cases} 4n^2x & \text{if } 0 \leq x \leq 1/2n \\ -4n^2x + 4n & \text{if } 1/2n < x < 1/n \\ 0 & \text{if } 1/n \leq x \leq 1 \end{cases}$$

Show that $\langle f_n \rangle$ converges pointwise to the constant function $g(x) = 0$.

Solution:

Now $f_n(0) = 0$ for every $n \in \mathbf{N}$, and so $\lim_{n\to\infty} f_n(0) = g(0) = 0$. On the other hand, if $x_0 > 0$, then $\exists\, n_0 \in \mathbf{N}$ such that $1/n_0 < x_0$; hence

$$n > n_0 \Rightarrow f_n(x_0) = 0 \Rightarrow \lim_{n\to\infty} f_n(x_0) = g(x_0) = 0$$

Thus $\langle f_n \rangle$ converges pointwise to the zero function.

Observe that $\int_0^1 f_n(x)\,dx = 1$, for every $n \in \mathbf{N}$, and $\int_0^1 g(x)\,dx = 0$

Thus, in this case, the limit of the integrals does not equal the integral of the limit, i.e.,

$$\lim_{n\to\infty} \int_0^1 f_n(x)\,dx \neq \int_0^1 \lim_{n\to\infty} f_n(x)\,dx$$

2. Let $C(I, \mathbf{R})$ denote the class of continuous real valued functions on $I = [0,1]$ with norm

$$\|f\| = \int_0^1 |f(x)|\,dx$$

Give an example of a sequence $\langle f_1, f_2, \ldots \rangle$ in $C(I, \mathbf{R})$ such that $f_n \to g$ in the above norm but $\langle f_n \rangle$ does not converge to g pointwise.

Solution:

Let $\langle f_n \rangle$ be defined by $f_n(x) = x^n$. Then

$$\lim_{n\to\infty} \|f_n\| = \lim_{n\to\infty} \int_0^1 x^n\,dx = \lim_{n\to\infty} 1/(n+1) = 0$$

Hence $\langle f_n \rangle$ converges to the zero function $g(x) = 0$ in the above norm. On the other hand, $\langle f_n \rangle$ converges pointwise (see Example 3.1) to the function f defined by $f(x) = 0$ if $0 \leq x < 1$ and $f(x) = 1$ if $x = 1$. Note $f \neq g$.

3. Show that if Y is T_1, T_2, regular, or connected, then $\mathcal{F}(X, Y)$ with the point open topology also has that property.

Solution:

Since the point open topology on $\mathcal{F}(X, Y)$ is the product topology, $\mathcal{F}(X, Y)$ inherits any product invariant property of Y. By previous results, the above properties are product invariant.

4. Prove Theorem 15.2: Let Y be Hausdorff and let $\mathcal{A}$ be a subset of $\mathcal{F}(X, Y)$ with the point open topology. Then the following are equivalent: (i) $\mathcal{A}$ is compact. (ii) $\mathcal{A}$ is closed and $\overline{\{f(x) : f \in \mathcal{A}\}}$ is compact in Y, for every $x \in X$.

Solution:

By Theorem 15.1, (ii) $\Rightarrow$ (i) and so we need only show that (i) $\Rightarrow$ (ii). Since Y is Hausdorff and T_2 is product invariant, $\mathcal{F}(X, Y)$ is also Hausdorff. Now by Theorem 11.5 a compact subset of a Hausdorff space is closed; hence $\mathcal{A}$ is closed. Furthermore, each evaluation map $e_x : \mathcal{F}(X, Y) \to Y$ is continuous with respect to the point open topology; hence, for each $x \in X$,

$$e_x[\mathcal{A}] = \{f(x) : f \in \mathcal{A}\}$$

is compact in Y and, since Y is Hausdorff, closed. In other words, $\overline{\{f(x) : f \in \mathcal{A}\}} = \{f(x) : f \in \mathcal{A}\}$ is compact.

5. Prove Theorem 15.3: Let $\mathcal{T}$ be the point open topology on $\mathcal{F}(X, Y)$ and let $\langle f_1, f_2, \ldots \rangle$ be a sequence in $\mathcal{F}(X, Y)$. Then the following are equivalent: (i) $\langle f_n \rangle$ converges to $g \in \mathcal{F}(X, Y)$ with respect to $\mathcal{T}$. (ii) $\langle f_n \rangle$ converges pointwise to g.

Solution:

Method 1.

We identify $\mathcal{F}(X, Y)$ with the product set $\mathbf{F} = \prod \{Y_x : x \in X\}$ and $\mathcal{T}$ with the product topology. Then by Theorem 12.7 the sequence $\langle f_n \rangle$ in $\mathbf{F}$ converges to $g \in \mathbf{F}$ if and only if, for every projection π_x,

$$\langle \pi_x(f_n) \rangle = \langle e_x(f_n) \rangle = \langle f_n(x) \rangle \quad \text{converges to} \quad \pi_x(g) = e_x(g) = g(x)$$

In other words, $\quad f_n \to g$ with respect to $\mathcal{T}$ $\quad$ iff $\quad \lim f_n(x) = g(x), \ \forall x \in X$

i.e. iff $\langle f_n \rangle$ converges pointwise to g.

Method 2.

(i) $\Rightarrow$ (ii): Let x_0 be an arbitrary point in X and let G be an open subset of Y containing $g(x_0)$, i.e. $g(x_0) \in G$. Then

$$g \in F(x_0, G) = \{f \in \mathcal{F}(X, Y) : f(x_0) \in G\}$$

and so $F(x_0, G)$ is a $\mathcal{T}$-open subset of $\mathcal{F}(X, Y)$ containing g. By (i), $\langle f_n \rangle$ converges to g with respect to $\mathcal{T}$; hence

$$\exists\, n_0 \in \mathbf{N} \quad \text{such that} \quad n > n_0 \ \Rightarrow\ f_n \in F(x_0, G)$$

Accordingly, $$n > n_0 \ \Rightarrow\ f_n(x_0) \in G \ \Rightarrow\ \lim_{n \to \infty} f_n(x_0) = g(x_0)$$

But x_0 was arbitrary; hence $\langle f_n \rangle$ converges pointwise to g.

(ii) $\Rightarrow$ (i): Let $F(x_0, G) = \{f : f(x_0) \in G\}$ be any member of the defining subbase for $\mathcal{T}$ which contains g. Then $g(x_0) \in G$. By (ii), $\langle f_n \rangle$ converges pointwise to g; hence

$$\exists\, n_0 \in \mathbf{N} \quad \text{such that} \quad n > n_0 \ \Rightarrow\ f_n(x_0) \in G$$

and so $$n > n_0 \ \Rightarrow\ f_n \in F(x_0, G) \ \Rightarrow\ \langle f_n \rangle \ \mathcal{T}\text{-converges to } g$$

UNIFORM CONVERGENCE

6. Prove Proposition 15.4: Let $\langle f_1, f_2, \ldots \rangle$ be a sequence of continuous functions from a topological space X into a metric space Y, and let $\langle f_n \rangle$ converge uniformly to $g : X \to Y$. Then g is continuous.

Solution:

Let $x_0 \in X$ and let $\epsilon > 0$. Then g is continuous at x_0 if $\exists$ an open set $G \subset X$ containing x_0 such that

$$x \in G \ \Rightarrow\ d(g(x), g(x_0)) < \epsilon$$

Now $\langle f_n \rangle$ converges uniformly to g, and so

$$\exists\ m \in \mathbf{N} \quad \text{such that} \quad d(f_m(x), g(x)) < \tfrac{1}{3}\epsilon, \quad \forall\, x \in X$$

Hence, by the Triangle Inequality,

$$d(g(x), g(x_0)) \leq d(g(x), f_m(x)) + d(f_m(x), f_m(x_0)) + d(f_m(x_0), g(x_0)) < d(f_m(x), f_m(x_0)) + \tfrac{2}{3}\epsilon$$

Since f_m is continuous, $\exists$ an open set $G \subset X$ containing x_0 such that

$$x \in G \Rightarrow d(f_m(x), f_m(x_0)) < \tfrac{1}{3}\epsilon \quad \text{and so} \quad x \in G \Rightarrow d(g(x), g(x_0)) < \epsilon$$

Thus g is continuous.

7. Let $\langle f_1, f_2, \ldots \rangle$ be a sequence of real, continuous functions defined on $[a, b]$ and converging uniformly to $g: [a, b] \to \mathbf{R}$. Show that

$$\lim_{n \to \infty} \int_a^b f_n(x)\, dx = \int_a^b g(x)\, dx$$

Observe (Problem 1) that this statement is not true in the case of pointwise convergence.

Solution:

Let $\epsilon > 0$. We need to show that

$$\exists\ n_0 \in \mathbf{N} \quad \text{such that} \quad n > n_0 \Rightarrow \left| \int_a^b f_n(x)\, dx - \int_a^b g(x)\, dx \right| < \epsilon$$

Now $\langle f_n \rangle$ converges uniformly to g, and so $\exists\, n_0 \in \mathbf{N}$ such that

$$n > n_0 \Rightarrow |f_n(x) - g(x)| < \epsilon/(b-a), \quad \forall\, x \in [a, b]$$

Hence, if $n > n_0$,

$$\begin{aligned} \left| \int_a^b f_n(x)\, dx - \int_a^b g(x)\, dx \right| &= \left| \int_a^b (f_n(x) - g(x))\, dx \right| \\ &\leq \int_a^b |f_n(x) - g(x)|\, dx \\ &< \int_a^b \epsilon/(b-a)\, dx = \epsilon \end{aligned}$$

8. Prove Theorem 15.5: Let $\langle f_1, f_2, \ldots \rangle$ be a sequence in $\mathcal{B}(X, Y)$ with metric

$$e(f, g) = \sup\{d(f(x), g(x)) : x \in X\}$$

Then the following are equivalent: (i) $\langle f_n \rangle$ converges to $g \in \mathcal{F}(X, Y)$ with respect to e. (ii) $\langle f_n \rangle$ converges uniformly to g.

Solution:

(i) ⇒ (ii): Let $\epsilon > 0$. Since $\langle f_n \rangle$ converges to g with respect to e,

$$\exists\ n_0 \in \mathbf{N} \quad \text{such that} \quad n > n_0 \Rightarrow e(f_n, g) < \epsilon$$

Therefore,

$$n > n_0 \Rightarrow d(f_n(x), g(x)) \leq \sup\{d(f_n(x), g(x)) : x \in X\} = e(f_n, g) < \epsilon, \quad \forall\, x \in X$$

that is, $\langle f_n \rangle$ converges uniformly to g.

(ii) ⇒ (i): Let $\epsilon > 0$. Since $\langle f_n \rangle$ converges uniformly to g,

$$\exists\ n_0 \in \mathbf{N} \quad \text{such that} \quad n > n_0 \Rightarrow d(f_n(x), g(x)) < \epsilon/2, \quad \forall\, x \in X$$

Therefore, $\quad n > n_0 \Rightarrow \sup\{d(f_n(x), g(x)) : x \in X\} \leq \epsilon/2 < \epsilon$

that is, $n > n_0$ implies $e(f_n, g) < \epsilon$, and so $\langle f_n \rangle$ converges to g with respect to e.

THE FUNCTION SPACE $C[0,1]$

9. Prove Proposition 15.6: Let $f: I \to \mathbf{R}$ be continuous on $I = [0,1]$. Then for every $\epsilon > 0$,

$$\exists\ \delta = \delta(\epsilon) > 0 \quad \text{such that} \quad |x-y| < \delta \ \Rightarrow\ |f(x)-f(y)| < \epsilon$$

i.e. f is uniformly continuous.

Solution:

Let $\epsilon > 0$. Since f is continuous, for every $p \in I$,

$$\exists\ \delta_p > 0 \quad \text{such that} \quad |x-p| < \delta_p \ \Rightarrow\ |f(x)-f(p)| < \tfrac{1}{2}\epsilon \qquad (1)$$

For each $p \in I$, set $S_p = I \cap (p - \frac{1}{2}\delta_p, p + \frac{1}{2}\delta_p)$. Then $\{S_p : p \in I\}$ is an open cover of I and, since I is compact, a finite number of the S_p also cover I; say, $I = S_{p_1} \cup \cdots \cup S_{p_m}$. Set

$$\delta = \tfrac{1}{2} \min(\delta_{p_1}, \ldots, \delta_{p_m})$$

Suppose $|x-y| < \delta$. Then $x \in S_{p_k}$ for some k, and so $|x-p_k| < \frac{1}{2}\delta_{p_k} < \delta_{p_k}$ and

$$|y-p_k| \leq |y-x| + |x-p_k| < \delta + \tfrac{1}{2}\delta_{p_k} \leq \tfrac{1}{2}\delta_{p_k} + \tfrac{1}{2}\delta_{p_k} = \delta_{p_k}$$

Hence by (1), $\quad |f(x)-f(p_k)| < \frac{1}{2}\epsilon \quad$ and $\quad |f(y)-f(p_k)| < \frac{1}{2}\epsilon$

Thus by the Triangle Inequality,

$$|f(x)-f(y)| \leq |f(x)-f(p_k)| + |f(p_k)-f(y)| < \tfrac{1}{2}\epsilon + \tfrac{1}{2}\epsilon = \epsilon$$

10. Let $\langle f_1, f_2, \ldots \rangle$ be a Cauchy sequence in $C[0,1]$. Show that, for each $x_0 \in I = [0,1]$, $\langle f_1(x_0), f_2(x_0), \ldots \rangle$ is a Cauchy sequence in $\mathbf{R}$.

Solution:

Let $x_0 \in I$ and let $\epsilon > 0$. Since $\langle f_n \rangle$ is Cauchy, $\exists\ n_0 \in \mathbf{N}$ such that

$$m, n > n_0 \ \Rightarrow\ \|f_n - f_m\| = \sup\{|f_n(x) - f_m(x)| : x \in I\} < \epsilon$$
$$\Rightarrow\ |f_n(x_0) - f_m(x_0)| < \epsilon$$

Hence $\langle f_n(x_0) \rangle$ is a Cauchy sequence.

11. Prove Theorem 15.7: $C[0,1]$ is a complete normed vector space.

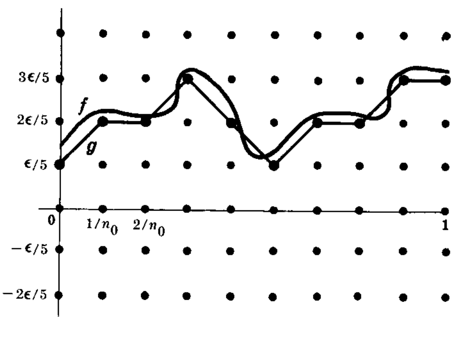
Solution:

Let $\langle f_1, f_2, \ldots \rangle$ be a Cauchy sequence in $C[0,1]$. Then, for every $x_0 \in I$, $\langle f_n(x_0) \rangle$ is a Cauchy sequence in $\mathbf{R}$ and, since $\mathbf{R}$ is complete, converges. Define $g: I \to \mathbf{R}$ by $g(x) = \lim_{n\to\infty} f_n(x)$. Then (see Problem 32) $\langle f_n \rangle$ converges uniformly to g. But, by Proposition 15.4, g is continuous, i.e. $g \in C[0,1]$; hence $C[0,1]$ is complete.

12. Let $f \in C[0,1]$ and let $\epsilon > 0$. Show that $\exists\ n_0 \in \mathbf{N}$ and points

$$p_0 = (0, \epsilon k_0/5), \ \ldots,$$
$$p_i = (i/n_0, \epsilon k_i/5), \ \ldots,$$
$$p_{n_0} = (1, \epsilon k_{n_0}/5)$$

where $k_0, \ldots, k_{n_0}$ are integers such that, if g is the polygonal arc connecting the p_i, then $\|f - g\| < \epsilon$ (see adjacent diagram). In other words, the piecewise linear (or polygonal) functions are dense in $C[0,1]$.



Solution:

Now f is uniformly continuous on $[0,1]$ and so

$$\exists\ n_0 \in \mathbf{N} \quad \text{such that} \quad |a-b| \leq 1/n_0 \ \Rightarrow\ |f(a)-f(b)| < \epsilon/5 \qquad (1)$$

Consider the following subset of $I \times \mathbf{R}$:

$$A = \{\langle x, y\rangle : x = i/n_0,\ y = k\epsilon/5 \text{ where } i = 0, \ldots, n_0;\ k \in \mathbf{Z}\}$$

Choose $p_i = \langle x_i, y_i\rangle \in A$ such that $\quad y_i \leq f(x_i) < y_i + \epsilon/5$

Then $\quad |f(x_i) - g(x_i)| = |f(x_i) - y_i| < \epsilon/5 \quad$ and by (1), $\quad |f(x_i) - f(x_{i+1})| < \epsilon/5$

as indicated in the diagram above.

Observe that

$$|g(x_i) - g(x_{i+1})| \leq |g(x_i) - f(x_i)| + |f(x_i) - f(x_{i+1})| + |f(x_{i+1}) - g(x_{i+1})| < \epsilon/5 + \epsilon/5 + \epsilon/5 = 3\epsilon/5$$

Since g is linear between x_i and x_{i+1},

$$x_i \leq z \leq x_{i+1} \Rightarrow |g(x_i) - g(z)| \leq |g(x_i) - g(x_{i+1})| < 3\epsilon/5$$

Now for any point $z \in I$, $\exists\, x_k$ satisfying $x_k \leq z \leq x_{k+1}$. Hence

$$|f(z) - g(z)| \leq |f(z) - f(x_k)| + |f(x_k) - g(x_k)| + |g(x_k) - g(z)| < \epsilon/5 + \epsilon/5 + 3\epsilon/5 = \epsilon$$

But z was an arbitrary point in I; hence $||f - g|| < \epsilon$.

13. Let m be an arbitrary positive integer and let $A_m \subset C[0,1]$ consist of those functions f with the property that

$$\exists\, x_0 \in \left[0, 1 - \frac{1}{m}\right] \quad \text{such that} \quad \left|\frac{f(x_0 + h) - f(x_0)}{h}\right| \leq m, \quad \forall\, h \in \left(0, \frac{1}{m}\right)$$

Show that A_m is a closed subset of $C[0,1]$. (Notice that every function f in $C[0,1]$ which is differentiable at a point belongs to some A_m for m sufficiently large.)

Solution:

Let $g \in \bar{A}_m$. We want to show that $g \in A_m$, i.e. $\bar{A}_m = A_m$. Since $g \in \bar{A}_m$, there exists a sequence $\langle f_1, f_2, \ldots\rangle$ in A_m converging to g. Now for each f_i there exists a point x_i such that

$$x_i \in \left[0, 1 - \frac{1}{m}\right] \quad \text{and} \quad \left|\frac{f_i(x_i + h) - f_i(x_i)}{h}\right| \leq m, \quad \forall\, h \in \left(0, \frac{1}{m}\right) \tag{1}$$

But $\langle x_n\rangle$ is a sequence in a compact set $\left[0, 1 - \frac{1}{m}\right]$ and so has a subsequence $\langle x_{i_n}\rangle$ which converges to, say, $x_0 \in \left[0, 1 - \frac{1}{m}\right]$.

Now $f_n \to g$ implies $f_{i_n} \to g$, and so (Problem 30), passing to the limit in (1), gives

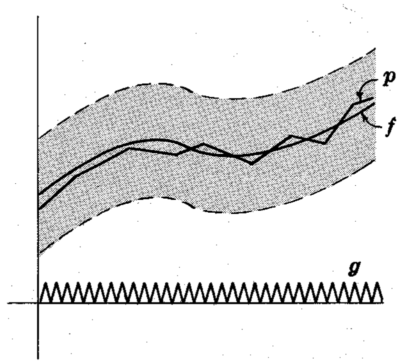
$$\left|\frac{g(x_0 + h) - g(x_0)}{h}\right| \leq m, \quad \forall\, h \in \left(0, \frac{1}{m}\right)$$

Hence $g \in A_m$, and A_m is closed.

14. Let $A_m \subset C[0,1]$ be defined as in Problem 13. Show that A_m is nowhere dense in $C[0,1]$.

Solution:

A_m is nowhere dense in $C[0,1]$ iff $\text{int}(\bar{A}_m) = \text{int}(A_m) = \emptyset$. Let $S = S(f, \delta)$ be any open sphere in $C[0,1]$. We claim that S contains a point not belonging to A_m, and so $\text{int}(\bar{A}_m) = \emptyset$.



By Problem 12, there exists a polygonal arc $p \in C[0,1]$ such that $||f - p|| < \frac{1}{2}\delta$. Let g be a saw-tooth function with magnitude less than $\frac{1}{2}\delta$ and slope sufficiently large (Problem 33). Then the function $h = p + g$ belongs to $C[0,1]$ but does not belong to A_m. Furthermore,

$$||f - h|| \leq ||f - p|| + ||g|| < \tfrac{1}{2}\delta + \tfrac{1}{2}\delta = \delta$$

so $h \in S$ and the proof is complete.

15. Let $A_m \subset C[0,1]$ be defined as in Problem 13. Show that $C[0,1] \neq \cup_{m=1}^{\infty} A_m$.

Solution:

Since A_m is nowhere dense in $C[0,1]$, $B = \cup_{m=1}^{\infty} A_m$ is of the first category. But, by Baire's Category Theorem, $C[0,1]$, a complete space, is of the second category. Hence $C[0,1] \neq B$.

16. Prove Proposition 15.8: There exists a continuous function $f: [0,1] \to \mathbf{R}$ which is nowhere differentiable.

Solution:

Let $f \in C[0,1]$ have a derivative at, say, x_0 and suppose $|f'(x_0)| = t$. Then

$$\exists\ \epsilon > 0 \quad \text{such that} \quad \left|\frac{f(x_0+h) - f(x_0)}{h}\right| \leq t+1, \quad \forall\, h \in (-\epsilon, \epsilon)$$

Now choose $m_0 \in \mathbf{N}$ so that $t+1 \leq m_0$ and $1/m_0 < \epsilon$. Then $f \in A_{m_0}$. Thus $\cup_{m=1}^{\infty} A_m$ contains all functions which are differentiable at some point of I.

But by the preceding problem, $C[0,1] \neq \cup_{m=1}^{\infty} A_m$ and so there exists a function in $C[0,1]$ which is nowhere differentiable.

17. Prove Theorem (Ascoli) 15.9: Let $\mathcal{A}$ be a closed subset of $C[0,1]$. Then the following are equivalent: (i) $\mathcal{A}$ is compact. (ii) $\mathcal{A}$ is uniformly bounded and equicontinuous.

Solution:

(i) ⇒ (ii): Since $\mathcal{A}$ is compact it is a bounded subset of $C[0,1]$ and is thus uniformly bounded as a set of functions. Now we need only show that $\mathcal{A}$ is equicontinuous.

Let $\epsilon > 0$. Since $\mathcal{A}$ is compact, it has a finite $\epsilon/3$-net, say, $\mathcal{B} = \{f_1, \ldots, f_t\}$. Hence, for any $f \in \mathcal{A}$,

$$\exists\ f_{i_0} \in \mathcal{B} \quad \text{such that} \quad \|f - f_{i_0}\| = \sup\{|f(x) - f_{i_0}(x)| : x \in I\} \leq \epsilon/3$$

Therefore, for any $x, y \in I = [0,1]$,

$$\begin{aligned} |f(x) - f(y)| &= |f(x) - f_{i_0}(x) + f_{i_0}(x) - f_{i_0}(y) + f_{i_0}(y) - f(y)| \\ &\leq |f(x) - f_{i_0}(x)| + |f_{i_0}(x) - f_{i_0}(y)| + |f_{i_0}(y) - f(y)| \\ &\leq \epsilon/3 + |f_{i_0}(x) - f_{i_0}(y)| + \epsilon/3 = |f_{i_0}(x) - f_{i_0}(y)| + 2\epsilon/3 \end{aligned}$$

Now each $f_i \in \mathcal{B}$ is uniformly continuous and so

$$\exists\ \delta_i > 0 \quad \text{such that} \quad |x - y| < \delta_i \Rightarrow |f_i(x) - f_i(y)| < \epsilon/3$$

Set $\delta = \min\{\delta_1, \ldots, \delta_t\}$. Then, for any $f \in \mathcal{A}$,

$$|x - y| < \delta \Rightarrow |f(x) - f(y)| \leq |f_{i_0}(x) + f_{i_0}(y)| + 2\epsilon/3 < \epsilon/3 + 2\epsilon/3 = \epsilon$$

Thus $\mathcal{A}$ is equicontinuous.

(ii) ⇒ (i): Since $\mathcal{A}$ is a closed subset of the complete space $C[0,1]$, we need only show that $\mathcal{A}$ is totally bounded. Let $\epsilon > 0$. Since $\mathcal{A}$ is equicontinuous,

$$\exists\ n_0 \in \mathbf{N} \quad \text{such that} \quad |a - b| < 1/n_0 \Rightarrow |f(a) - f(b)| < \epsilon/5, \quad \forall\, f \in \mathcal{A}$$

Now for each $f \in \mathcal{A}$, we can construct, by Problem 12, a polygonal arc p_f such that $\|f - p_f\| < \epsilon$ and p_f connects points belonging to

$$A = \{\langle x, y\rangle : x = 0, 1/n_0, 2/n_0, \ldots, 1;\ y = n\epsilon/5,\ n \in \mathbf{Z}\}$$

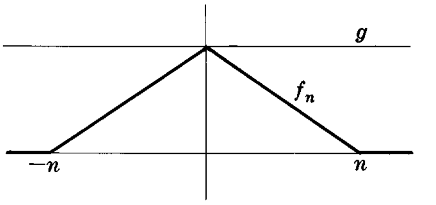
We claim that $\mathcal{B} = \{p_f : f \in \mathcal{A}\}$ is finite and hence a finite ϵ-net for $\mathcal{A}$.

Now $\mathcal{A}$ is uniformly bounded, and so $\mathcal{B}$ is uniformly bounded. Therefore only a finite number of the points in A will appear in the polygonal arcs in $\mathcal{B}$. Hence there can only be a finite number of arcs in $\mathcal{B}$. Thus $\mathcal{B}$ is a finite ϵ-net for $\mathcal{A}$, and so $\mathcal{A}$ is totally bounded.

COMPACT CONVERGENCE

18. Let $\langle f_1, f_2, \ldots\rangle$ in $\mathcal{F}(\mathbf{R},\mathbf{R})$ be defined by

$$f_n(x) = \begin{cases} 1 - \frac{1}{n}|x| & \text{if } |x| < n \\ 0 & \text{if } |x| \geq n \end{cases}$$



Show that $\langle f_n\rangle$ converges uniformly on compacta to the constant function $g(x) = 1$.

Solution:

Let E be a compact subset of $\mathbf{R}$ and let $0 < \epsilon < 1$. Since E is compact, it is bounded; say, $E \subset (-M, M)$ for $M > 0$. Now

$$\exists\, n_0 \in \mathbf{N} \quad \text{such that} \quad n_0 > M/\epsilon, \text{ or, } M/n_0 < \epsilon$$

Therefore, $$n > n_0 \Rightarrow |f_n(x) - g(x)| = \frac{1}{n}|x| < M/n_0 < \epsilon, \quad \forall x \in E$$

Hence $\langle f_n\rangle$ converges uniformly to g on E.

19. Show: If Y is Hausdorff, then the compact open topology on $\mathcal{F}(X, Y)$ is also Hausdorff.

Solution:

Method 1. Let $f, g \in \mathcal{F}(X, Y)$ with $f \neq g$. Then $\exists\, p \in X$ such that $f(p) \neq g(p)$. Now Y is Hausdorff, hence $\exists$ open subsets G and H of Y such that $f(p) \in G$, $g(p) \in H$ and $G \cap H = \emptyset$. Hence

$$f \in F(p, G), \quad g \in F(p, H) \quad \text{and} \quad F(p, G) \cap F(p, H) = \emptyset$$

But the singleton set $\{p\}$ is compact, and so $F(p, G)$ and $F(p, H)$ belong to the compact open topology on $\mathcal{F}(X, Y)$. Accordingly, $\mathcal{F}(X, Y)$ is Hausdorff.

Method 2. The compact open topology is finer than the point open topology, which is Hausdorff since T_2 is a product invariant property. Hence the compact open topology is also Hausdorff.

20. Prove Theorem 15.12: Let $\langle f_1, f_2, \ldots\rangle$ be a sequence in $\mathcal{C}(X, Y)$, the collection of all continuous functions from a topological space X into a metric space (Y, d). Then the following are equivalent:

(i) $\langle f_n\rangle$ converges uniformly on compacta to $g \in \mathcal{C}(X, Y)$.

(ii) $\langle f_n\rangle$ converges to g with respect to the compact open topology $\mathcal{T}$ on $\mathcal{C}(X, Y)$.

Solution:

(i) $\Rightarrow$ (ii):

Let $F(E, G)$ be an open subbase element of $\mathcal{T}$ containing g; hence $g[E] \subset G$ where E is compact and G is open. Since g is continuous, $g[E]$ is compact. Furthermore, $g[E] \cap G^c = \emptyset$ and so (see Page 164) the distance between the compact set $g[E]$ and the closed set G^c is greater than zero; say, $d(g[E], G^c) = \epsilon > 0$. Since $\langle f_n\rangle$ converges uniformly on compacta to g,

$$\exists\, n_0 \in \mathbf{N} \quad \text{such that} \quad n > n_0 \Rightarrow d(f_n(x), g(x)) < \epsilon, \quad \forall x \in E$$

Therefore, $$d(f_n(x), g[E]) \leq d(f_n(x), g(x)) < \epsilon, \quad \forall x \in E$$

and so, for every $x \in E$, $f_n(x) \notin G^c$. In other words,

$$n > n_0 \Rightarrow f_n[E] \subset G \Rightarrow f_n \in F(E, G)$$

Accordingly, $\langle f_n\rangle$ converges to g with respect to the compact open topology $\mathcal{T}$.

(ii) $\Rightarrow$ (i):

Let E be a compact subset of X and let $\epsilon > 0$. We want to show that $\langle f_n\rangle$ converges uniformly on E to g, i.e.,

$$\exists\, n_0 \in \mathbf{N} \quad \text{such that} \quad n > n_0 \Rightarrow d(f_n(x), g(x)) < \epsilon, \quad \forall x \in E$$

Since E is compact and g is continuous, $g[E]$ is compact. Let $\mathcal{B} = \{p_1, \ldots, p_t\}$ be a finite $\epsilon/3$-net for $g[E]$. Consider the open spheres

$$S_1 = S(p_1, \epsilon/3), \ \ldots, \ S_t = S(p_t, \epsilon/3) \quad \text{and} \quad G_1 = S(p_1, 2\epsilon/3), \ \ldots, \ G_t = S(p_t, 2\epsilon/3)$$

Hence $\bar{S}_1 \subset G_1, \ldots, \bar{S}_t \subset G_t$. Furthermore, since $\mathcal{B}$ is an $\epsilon/3$-net for $g[E]$,

$$g[E] \subset \bar{S}_1 \cup \cdots \cup \bar{S}_t \quad \text{and so} \quad E \subset g^{-1}[\bar{S}_1] \cup \cdots \cup g^{-1}[\bar{S}_t]$$

Now set $\quad E_i = E \cap g^{-1}[\bar{S}_i] \quad$ and so $\quad E = E_1 \cup \cdots \cup E_t \quad$ and $\quad g[E_i] \subset \bar{S}_i \subset G_i$

We claim that the E_i are compact. For g is continuous and so $g^{-1}[\bar{S}_i]$, the inverse of a closed set, is closed; hence $E_i = E \cap g^{-1}[\bar{S}_i]$, the intersection of a compact and a closed set, is compact.

Now $g[E_i] \subset G_i$ and so the $F(E_i, G_i)$ are $\mathcal{T}$-open subsets of $\mathcal{F}(X, Y)$ containing g; hence $\cap_{i=1}^t F(E_i, G_i)$ is also a $\mathcal{T}$-open set containing g. But $\langle f_n \rangle$ converges to g with respect to $\mathcal{T}$; hence

$$\exists \ n_0 \in \mathbf{N} \quad \text{such that} \quad n > n_0 \ \Rightarrow \ f_n \in \cap_{i=1}^t F(E_i, G_i) \ \Rightarrow \ f_n[E_1] \subset G_1, \ldots, f_n[E_t] \subset G_t$$

Now let $x \in E$. Then $x \in E_{i_0}$ and so, for $n > n_0$,

$$f_n(x) \in f_n[E_{i_0}] \subset G_{i_0} \ \Rightarrow \ d(f_n(x), p_{i_0}) < 2\epsilon/3$$

and

$$g(x) \in g[E_{i_0}] \subset \bar{S}_{i_0} \ \Rightarrow \ d(g(x), p_{i_0}) \leq \epsilon/3$$

Therefore, by the Triangle Inequality,

$$n > n_0 \ \Rightarrow \ d(f_n(x), g(x)) \leq d(f_n(x), p_{i_0}) + d(p_{i_0}, g(x)) < 2\epsilon/3 + \epsilon/3 = \epsilon, \quad \forall x \in E$$

FUNCTIONALS ON NORMED SPACES

21. Show that if f is a linear functional on X, then $f(0) = 0$.

Solution:

Since f is linear and $0 = 0 + 0$,

$$f(0) = f(0 + 0) = f(0) + f(0)$$

Adding $-f(0)$ to both sides gives $f(0) = 0$.

22. Show that a bounded linear functional f on X is uniformly continuous.

Solution:

Let M be a bound for f and let $\epsilon > 0$. Set $\delta = \epsilon/M$. Then

$$||x - y|| < \delta \ \Rightarrow \ |f(x) - f(y)| = |f(x - y)| \leq M \, ||x - y|| < \epsilon$$

23. Prove Proposition 15.13: Let f and g be bounded linear functionals on X and let $c \in \mathbf{R}$. Then $f + g$ and $c \cdot f$ are also bounded linear functionals on X.

Solution:

Let M and M^* be bounds for f and g respectively. Then

$$(f + g)(x + y) = f(x + y) + g(x + y) = f(x) + f(y) + g(x) + g(y) = (f + g)(x) + (f + g)(y)$$

$$(f + g)(kx) = f(kx) + g(kx) = k\,f(x) + k\,g(x) = k[f(x) + g(x)] = k\,(f + g)(x)$$

$$|(f + g)(x)| = |f(x) + g(x)| \leq |f(x)| + |g(x)| \leq M\,||x|| + M^*\,||x|| = (M + M^*)\,||x||$$

Thus $f + g$ is a bounded linear functional.

Furthermore,

$$(c\cdot f)(x+y) \;=\; c\,f(x+y) \;=\; c\,[f(x)+f(y)] \;=\; c\,f(x)+c\,f(y) \;=\; (c\cdot f)(x)+(c\cdot f)(y)$$

$$(c\cdot f)(kx) \;=\; c\,f(kx) \;=\; ck\,f(x) \;=\; kc\,f(x) \;=\; k\,(c\cdot f)(x)$$

$$|(c\cdot f)(x)| \;=\; |c\,f(x)| \;=\; |c|\,|f(x)| \;\leqq\; |c|\,(M\,||x||) \;=\; (|c|\,M)\,||x||$$

and so $c\cdot f$ is a bounded linear functional.

24. Prove Proposition 15.14: The following function on X^* is a norm:

$$||f|| \;=\; \sup\,\{|f(x)|/||x|| : x \neq 0\}$$

Solution:

If $f=0$, then $f(x)=0$, $\forall\, x\in X$, and so $||f|| \;=\; \sup\,\{0\} \;=\; 0$. If $f\neq 0$, then $\exists\, x_0 \neq 0$ such *that* $f(x_0)\neq 0$, *and so*

$$||f|| \;=\; \sup\,\{|f(x)|/||x||\} \;\geqq\; |f(x_0)|/||x_0|| \;>\; 0$$

Thus the axiom $[\mathbf{N_1}]$ (see Page 118) is satisfied.

Now

$$||k\cdot f|| \;=\; \sup\,\{|(k\cdot f)(x)|/||x||\} \;=\; \sup\,\{|k[f(x)]|/||x||\}$$
$$=\; \sup\,\{|k|\,|f(x)|/||x||\} \;=\; |k|\,\sup\,\{|f(x)|/||x||\} \;=\; |k|\,||f||$$

Hence axiom $[\mathbf{N_2}]$ is satisfied.

Furthermore,

$$||f+g|| \;=\; \sup\,\{|f(x)+g(x)|/||x||\} \;\leqq\; \sup\,\{(|f(x)|+|g(x)|)/||x||\}$$
$$\leqq\; \sup\,\{|f(x)|/||x||\} + \sup\,\{|g(x)|/||x||\} \;=\; ||f||+||g||$$

and so axiom $[\mathbf{N_3}]$ is satisfied.

Supplementary Problems

CONVERGENCE OF SEQUENCES OF FUNCTIONS

25. Let $\langle f_1, f_2, \ldots\rangle$ be the sequence of real-valued functions with domain $I = [0,1]$ defined by $f_n(x) = x^n/n$.

(i) Show that $\langle f_n\rangle$ converges pointwise to the constant function $g(x)=0$, i.e. for every $x\in I$, $\lim_{n\to\infty} f_n(x) \;=\; 0$.

(ii) Show that

$$\lim_{n\to\infty} \frac{d}{dx} f_n(x) \;\neq\; \frac{d}{dx} \lim_{n\to\infty} f_n(x)$$

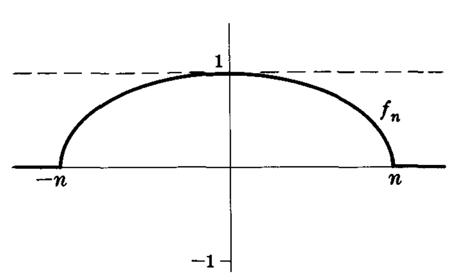
26. Let $\langle f_1, f_2, \ldots\rangle$ be a sequence of real-valued differentiable functions with domain $[a,b]$ which converge uniformly to g. Prove:

$$\frac{d}{dx} \lim_{n\to\infty} f_n(x) \;=\; \lim_{n\to\infty} \frac{d}{dx} f_n(x)$$

(Observe, by the preceding problem, that this result does not hold in the case of pointwise convergence.)

27. Let $f_n : \mathbf{R}\to\mathbf{R}$ be defined by

$$f_n(x) \;=\; \begin{cases} \dfrac{1}{n}\sqrt{n^2-x^2} & \text{if } |x|<n \\ 0 & \text{if } |x|\geqq n \end{cases}$$



(i) Show that $\langle f_n\rangle$ does not converge uniformly to the constant function $g(x)=1$.

(ii) Prove that $\langle f_n\rangle$ converges uniformly on compacta to the constant function $g(x)=1$.

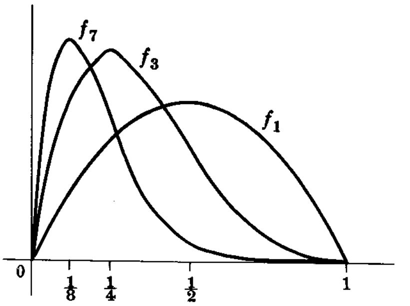
28. Let $\langle f_1, f_2, \ldots\rangle$ be the sequence of functions with domain $I = [0,1]$ defined by $f_n(x) = nx(1-x)^n$.

(i) Show that $\langle f_n\rangle$ converges pointwise to the constant function $g(x) = 0$.

(ii) Show that $\langle f_n\rangle$ does not converge uniformly to $g(x) = 0$.

(iii) Show that, in this case,

$$\lim_{n\to\infty} \int_0^1 f_n(x)\,dx \quad = \quad \int_0^1 \left[\lim_{n\to\infty} f_n(x)\right] dx$$



29. Let $\langle f_1, f_2, \ldots\rangle$ be the sequence in $\mathcal{F}(\mathbf{R},\mathbf{R})$ defined by $f_n(x) = \dfrac{n+1}{n}\,x$.

(i) Show that $\langle f_n\rangle$ converges uniformly on compacta to the function $g(x) = x$.

(ii) Show that $\langle f_n\rangle$ does not converge uniformly to $g(x) = x$.

30. Let $\langle f_1, f_2, \ldots\rangle$ be a sequence of (Riemann) integrable functions on $I = [0,1]$. The sequence $\langle f_n\rangle$ is said to *converge in the mean* to the function g if

$$\lim_{n\to\infty} \int_0^1 |f_n(x) - g(x)|^2\,dx \quad = \quad 0$$

(i) Show that if $\langle f_n\rangle$ converges uniformly to g, then $\langle f_n\rangle$ converges in the mean to g.

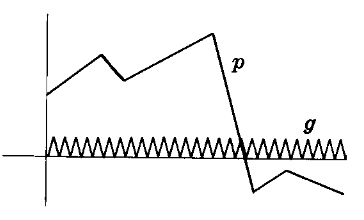
(ii) Show, by a counterexample, that convergence in the mean does not necessarily imply pointwise convergence.

THE FUNCTION SPACE $C[0,1]$

31. Show that $C[a,b]$ is isometric and hence homeomorphic to $C[0,1]$.

32. Prove: Let $\langle f_n\rangle$ converge to g in $C[0,1]$ and let $x_n \to x_0$. Then $\lim_{n\to\infty} f_n(x_n) = g(x_0)$.

33. Let p be a polygonal arc in $C[0,1]$ and let $\delta > 0$. Show that there exists a sawtooth function g with magnitude less than $\frac{1}{2}\delta$, i.e. $||g|| < \frac{1}{2}\delta$, such that $p+g$ does not belong to A_m (see Problem 14).



34. Let $\langle f_n\rangle$ be a Cauchy sequence in $C[0,1]$ and let $\langle f_n\rangle$ converge pointwise to g. Then $\langle f_n\rangle$ converges uniformly to g.

UNIFORM CONTINUITY

35. Show that $f(x) = 1/x$ is not uniformly continuous on the open interval $(0,1)$.

36. Define uniform continuity for a function $f: X \to Y$ where X and Y are arbitrary metric spaces.

37. Prove: Let f be a continuous function from a compact metric space X into a metric space Y. Then f is uniformly continuous.

FUNCTIONALS ON NORMED SPACES

38. Let f be a bounded linear functional on a normed space X. Show that

$$\sup\{|f(x)|/||x|| : x \neq 0\} \quad = \quad \inf\{M : M \text{ is a bound for } f\}$$

39. Show that if f is a continuous linear functional on X then f is bounded.

40. Prove: The dual space X^* of any normed space X is complete.

APPENDIX

Properties of the Real Numbers

FIELD AXIOMS

The set of real numbers, denoted by **R**, plays a dominant role in mathematics and, in particular, in analysis. In fact, many concepts in topology are abstractions of properties of sets of real numbers. The set **R** can be characterized by the statement that **R** is a *complete, Archimedean ordered field.* In this appendix we investigate the order relation in **R** which is used in defining the usual topology on **R** (see Chapter 4). We now state the field axioms of **R** which, with their consequences, are assumed throughout the text.

Definition: A set F of two or more elements, together with two operations called addition (+) and multiplication ($\cdot$), is a field if it satisfies the following axioms:

[A₁] Closure: $a, b \in F \Rightarrow a+b \in F$

[A₂] Associative Law: $a, b, c \in F \Rightarrow (a+b)+c = a+(b+c)$

[A₃] (Additive) Identity: $\exists\, 0 \in F$ such that $0+a = a+0 = a, \ \forall a \in F$

[A₄] (Additive) Inverse: $a \in F \Rightarrow \exists\, -a \in F$ such that $a+(-a) = (-a)+a = 0$

[A₅] Commutative Law: $a, b \in F \Rightarrow a+b = b+a$

[M₁] Closure: $a, b \in F \Rightarrow a \cdot b \in F$

[M₂] Associative Law: $a, b, c \in F \Rightarrow (a \cdot b) \cdot c = a \cdot (b \cdot c)$

[M₃] (Multiplicative) Identity: $\exists\, 1 \in F,\ 1 \neq 0$ such that $1 \cdot a = a \cdot 1 = a, \ \forall a \in F$

[M₄] (Multiplicative) Inverse: $a \in F,\ a \neq 0 \Rightarrow \exists\, a^{-1} \in F$ such that $a \cdot a^{-1} = a^{-1} \cdot a = 1$

[M₅] Commutative Law: $a, b \in F \Rightarrow a \cdot b = b \cdot a$

[D₁] Left Distributive Law: $a, b, c \in F \Rightarrow a \cdot (b+c) = a \cdot b + a \cdot c$

[D₂] Right Distributive Law: $a, b, c \in F \Rightarrow (b+c) \cdot a = b \cdot a + c \cdot a$

Here $\exists$ reads "there exists", $\forall$ reads "for every", and $\Rightarrow$ reads "implies".

The following algebraic properties of the real numbers follow directly from the field axioms.

Proposition A.1: Let F be a field. Then:

(i) The identity elements 0 and 1 are unique.

(ii) The following cancellation laws hold:

(1) $a+b = a+c \Rightarrow b = c$, (2) $a \cdot b = a \cdot c,\ a \neq 0 \Rightarrow b = c$

(iii) The inverse elements $-a$ and a^{-1} are unique.

(iv) For every $a, b \in F$,

(1) $a \cdot 0 = 0$, (2) $a \cdot (-b) = (-a) \cdot b = -(a \cdot b)$, (3) $(-a) \cdot (-b) = a \cdot b$

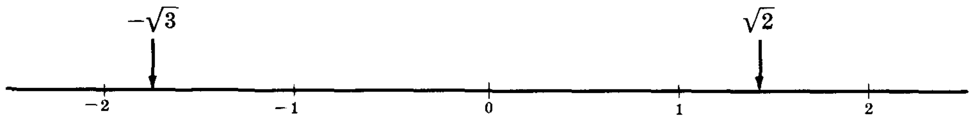
Subtraction and division (by a non-zero element) are defined in a field as follows:

$$b - a \equiv b + (-a) \quad \text{and} \quad \frac{b}{a} \equiv b \cdot a^{-1}$$

Remark: A non-empty set together with two operations which satisfy all the axioms of a field except possibly $[M_3]$, $[M_4]$ and $[M_5]$ is called a *ring*. The set **Z** of integers under addition and multiplication, for example, is a ring but not a field.

REAL LINE

We assume the reader is familiar with the geometric representation of **R** by means of points on a straight line as in the figure below. Notice that a point, called the origin, is chosen to represent 0 and another point, usually to the right of 0, is chosen to represent 1. Then there is a natural way to pair off the points on the line and the real numbers, i.e. each point will represent a unique real number and each real number will be represented by a unique point. For this reason we refer to **R** as the *real line* and use the words point and number interchangeably.



The real line **R**

SUBSETS OF R

The symbols **Z** and **N** are used to denote the following subsets of **R**:

$$\mathbf{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}, \qquad \mathbf{N} = \{1, 2, 3, 4, \ldots\}$$

The elements in **Z** are called *rational integers* or, simply, *integers*; and the elements in **N** are called *positive integers* or *natural numbers*.

The symbol **Q** is used to denote the set of *rational numbers*. The rational numbers are those real numbers which can be expressed as the ratio of two integers provided the second is non-zero:

$$\mathbf{Q} = \{x \in \mathbf{R} : x = p/q;\ p,q \in \mathbf{Z},\ q \neq 0\}$$

Now each integer is also a rational number since, e.g., $-5 = 5/-1$; hence **Z** is a subset of **Q**. In fact we have the following hierarchy of sets:

$$\mathbf{N} \subset \mathbf{Z} \subset \mathbf{Q} \subset \mathbf{R}$$

The *irrational numbers* are those real numbers which are not rational; thus $\mathbf{Q}^c$, the complement (relative to **R**) of the set **Q** of rational numbers, denotes the set of irrational numbers.

POSITIVE NUMBERS

Those numbers to the right of 0 on the real line **R**, i.e. on the same side as 1, are the *positive numbers*; those numbers to the left of 0 are the *negative numbers*. The following axioms completely characterize the set of positive numbers:

$[P_1]$ If $a \in \mathbf{R}$, then exactly one of the following is true: a is positive; $a = 0$; $-a$ is positive.

$[P_2]$ If $a, b \in \mathbf{R}$ are positive, then their sum $a + b$ and their product $a \cdot b$ are also positive.

It follows that a is positive if and only if $-a$ is negative.

Example 1.1: We show, using only $[P_1]$ and $[P_2]$, that the real number 1 is positive. By $[P_1]$, either 1 or -1 is positive. Assume that -1 is positive and so, by $[P_2]$, the product $(-1)(-1) = 1$ is also positive. But this contradicts $[P_1]$ which states that 1 and -1 cannot both be positive. Hence the assumption that -1 is positive is false, and 1 is positive.

Example 1.2: The real number -2 is negative. For, by Example 1.1, 1 is positive and so, by $[P_2]$, the sum $1+1 = 2$ is positive. Therefore, by $[P_1]$, -2 is not positive, i.e. -2 is negative.

Example 1.3: We show that the product $a \cdot b$ of a positive number a and a negative number b is negative. For if b is negative then, by $[P_1]$, $-b$ is positive and so, by $[P_2]$, the product $a \cdot (-b)$ is also positive. But $a \cdot (-b) = -(a \cdot b)$. Thus $-(a \cdot b)$ is positive and so, by $[P_1]$, $a \cdot b$ is negative.

ORDER

We define an order relation in **R**, using the concept of positiveness.

Definition: The real number a is *less than* the real number b, written $a < b$, if the difference $b - a$ is positive.

Geometrically speaking, if $a < b$ then the point a on the real line lies to the left of the point b.

The following notation is also used:

$b > a$,	read b is greater than a,	means $a < b$
$a \leq b$,	read a is less than or equal to b,	means $a < b$ or $a = b$
$b \geq a$,	read b is greater than or equal to a,	means $a \leq b$

Example 2.1: $2 < 5$; $-6 \leq -3$; $4 \leq 4$; $5 > -8$

Example 2.2: A real number x is positive iff $x > 0$, and x is negative iff $x < 0$.

Example 2.3: The notation $2 < x < 7$ means $2 < x$ and also $x < 7$; hence x will lie between 2 and 7 on the real line.

The axioms $[P_1]$ and $[P_2]$ which define the positive real numbers are used to prove the following theorem.

Theorem A.2: Let a, b and c be real numbers. Then:

(i) either $a < b$, $a = b$ or $b < a$;

(ii) if $a < b$ and $b < c$, then $a < c$;

(iii) if $a < b$, then $a + c < b + c$;

(iv) if $a < b$ and c is positive, then $ac < bc$; and

(v) if $a < b$ and c is negative, then $ac > bc$.

Corollary A.3: The set **R** of real numbers is totally ordered by the relation $a \leq b$.

ABSOLUTE VALUE

The absolute value of a real number x, denoted by $|x|$, is defined by

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

Observe that the absolute value of any number is always non-negative, i.e. $|x| \geq 0$ for every $x \in \mathbf{R}$.

Geometrically speaking, the absolute value of x is the distance between the point x on the real line and the origin, i.e. the point 0. Furthermore, the distance between any two points $a, b \in \mathbf{R}$ is $|a-b| = |b-a|$.

Example 3.1: $|-2| = 2$, $|7| = 7$, $|-\pi| = \pi$, $|-\sqrt{2}| = \sqrt{2}$

Example 3.2: $|3-8| = |-5| = 5$ and $|8-3| = |5| = 5$

Example 3.3: The statement $|x| < 5$ can be interpreted to mean that the distance between x and the origin is less than 5; hence x must lie between -5 and 5 on the real line. In other words,

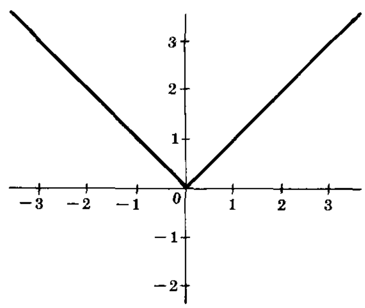
$$|x| < 5 \quad \text{and} \quad -5 < x < 5$$

have identical meaning and, similarly,

$$|x| \leqq 5 \quad \text{and} \quad -5 \leqq x \leqq 5$$

have identical meaning.

The graph of the function $f(x) = |x|$, i.e. the absolute value function, lies entirely in the upper half plane since $f(x) \geqq 0$ for every $x \in \mathbf{R}$ (see diagram below).



Graph of $f(x) = |x|$

The central facts about the absolute value function are the following:

Proposition A.4: Let a, b and c be real numbers. Then:

(i) $|a| \geqq 0$, and $|a| = 0$ iff $a = 0$;

(ii) $|ab| = |a|\,|b|$;

(iii) $|a+b| \leqq |a| + |b|$;

(iv) $|a-b| \geqq \big|\,|a| - |b|\,\big|$; and

(v) $|a-c| \leqq |a-b| + |b-c|$.

LEAST UPPER BOUND AXIOM

Chapter 14 discusses the concept of completeness for general metric spaces. For the real line **R**, we may use the definition: **R** *is complete means that* **R** *satisfies the following axiom:*

[LUB] (Least Upper Bound Axiom): If A is a set of real numbers bounded from above, then A has a least upper bound, i.e. $\sup(A)$ exists.

Example 4.1: *The set* **Q** *of rational numbers does not satisfy the Least Upper Bound Axiom.* For let

$$A = \{q \in \mathbf{Q} : q > 0,\ q^2 < 2\}$$

i.e., A consists of those rational numbers which are greater than 0 and less than $\sqrt{2}$. Now A is bounded from above, e.g. 5 is an upper bound for A. But A does not have a least upper bound, i.e. there exists no rational number m such that $m = \sup(A)$. Observe that m cannot be $\sqrt{2}$ since $\sqrt{2}$ does not belong to **Q**.

We use the Least Upper Bound Axiom to prove that **R** is Archimedean ordered:

Theorem (Archimedean Order Axiom) A.5: The set $\mathbf{N} = \{1, 2, 3, \ldots\}$ of positive integers is not bounded from above.

In other words, there exists no real number which is greater than every positive integer. One consequence of this theorem is:

Corollary A.6: There is a rational number between any two distinct real numbers.

NESTED INTERVAL PROPERTY

The *nested interval property* of **R**, contained in the next theorem, is an important consequence of the Least Upper Bound Axiom, i.e. the completeness of **R**.

Theorem (Nested Interval Property) A.7: Let $I_1 = [a_1, b_1]$, $I_2 = [a_2, b_2]$, ... be a sequence of nested closed (bounded) intervals, i.e. $I_1 \supset I_2 \supset \ldots$. Then there exists at least one point common to every interval, i.e.

$$\cap_{i=1}^{\infty} I_i \neq \emptyset$$

It is necessary that the intervals in the theorem be closed and bounded, or else the theorem is not true as seen by the following two examples.

Example 5.1: Let $A_1, A_2, \ldots$ be the following sequence of open-closed intervals:

$$A_1 = (0, 1],\quad A_2 = (0, 1/2],\quad \ldots,\quad A_k = (0, 1/k],\quad \ldots$$

Now the sequence of intervals is nested, i.e. each interval contains the succeeding interval: $A_1 \supset A_2 \supset \cdots$. But the intersection of the intervals is empty, i.e.,

$$A_1 \cap A_2 \cap \cdots \cap A_k \cap \cdots = \emptyset$$

Thus there exists no point common to every interval.

Example 5.2: Let $A_1, A_2, \ldots$ be the following sequence of closed infinite intervals:

$$A_1 = [1, \infty),\quad A_2 = [2, \infty),\quad \ldots,\quad A_k = [k, \infty),\quad \ldots$$

Now $A_1 \supset A_2 \supset \cdots$, i.e. the sequence of intervals is nested. But there exists no point common to every interval, i.e.,

$$A_1 \cap A_2 \cap \cdots \cap A_k \cap \cdots = \emptyset$$

Solved Problems

FIELD AXIOMS

1. Prove Proposition A.1(iv): For every $a, b \in F$,

$$(1)\ a0 = 0, \quad (2)\ a(-b) = (-a)b = -ab, \quad (3)\ (-a)(-b) = ab$$

Solution:

(1) $a0 = a(0+0) = a0 + a0$. Adding $-a0$ to both sides gives $0 = a0$.

(2) $0 = a0 = a(b + (-b)) = ab + a(-b)$. Hence $a(-b)$ is the negative of ab, that is, $a(-b) = -ab$. Similarly, $(-a)b = -ab$.

(3) $0 = (-a)0 = (-a)(b + (-b)) = (-a)b + (-a)(-b) = -ab + (-a)(-b)$. Adding ab to both sides gives $ab = (-a)(-b)$.

2. Show that multiplication distributes over subtraction in a field F, i.e. $a(b-c) = ab - ac$.

Solution: $a(b-c) = a(b+(-c)) = ab + a(-c) = ab + (-ac) = ab - ac$

3. Show that a field F has no zero divisors, i.e. $ab = 0 \Rightarrow a = 0$ or $b = 0$.

Solution:

Suppose $ab = 0$ and $a \neq 0$. Then a^{-1} exists and so $b = 1b = (a^{-1}a)b = a^{-1}(ab) = a^{-1}0 = 0$.

INEQUALITIES AND POSITIVE NUMBERS

4. Rewrite so that x is alone between the inequality signs:
(i) $3 < 2x - 5 < 7$, (ii) $-7 < -2x + 3 < 5$.

Solution:

We use Theorem A.2:

(i) By (iii), we can add 5 to each side of $3 < 2x - 5 < 7$ to get $8 < 2x < 12$. By (iv), we can multiply each side by $\frac{1}{2}$ to obtain $4 < x < 6$.

(ii) Add -3 to each side to get $-10 < -2x < 2$. By (v), we can multiply each side by $-\frac{1}{2}$ and reverse the inequalities to obtain $-1 < x < 5$.

5. Prove that $\frac{1}{2}$ is a positive number.

Solution:

By $[\mathbf{P}_1]$, either $-\frac{1}{2}$ is positive or $\frac{1}{2}$ is positive. Suppose $-\frac{1}{2}$ is positive and so, by $[\mathbf{P}_2]$, $(-\frac{1}{2}) + (-\frac{1}{2}) = -1$ is also positive. But by Example 1.1, 1 is positive and not -1. Thus we have a contradiction, and so $\frac{1}{2}$ is positive.

6. Prove Theorem A.2(ii): *If $a < b$ and $b < c$, then $a < c$.*

Solution:

By definition, $a < b$ means $b - a$ is positive; and $b < c$ means $c - b$ is positive. Now, by $[\mathbf{P}_2]$, the sum $(b-a) + (c-b) = c - a$ is positive and so, by definition, $a < c$.

7. Prove Theorem A.2(v): If $a < b$ and c is negative, then $ac > bc$.

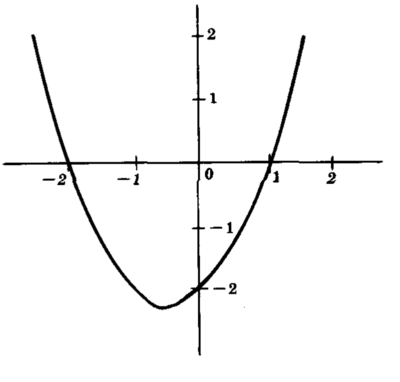
Solution:

By definition, $a < b$ means $b - a$ is positive. By $[\mathbf{P}_1]$, if c is negative then $-c$ is positive, and so, by $[\mathbf{P}_2]$, the product $(b-a)(-c) = ac - bc$ is also positive. Hence, by definition, $bc < ac$ or, equivalently, $ac > bc$.

8. Determine all real numbers x such that $(x-1)(x+2) < 0$.

Solution:

We must find all values of x such that $y = (x-1)(x+2)$ is negative. Since the product of two numbers is negative iff one is positive and the other is negative, y is negative if (i) $x - 1 < 0$ and $x + 2 > 0$, or (ii) $x - 1 > 0$ and $x + 2 < 0$. If $x - 1 > 0$ and $x + 2 < 0$, then $x > 1$ and $x < -2$, which is impossible. Thus y is negative iff $x - 1 < 0$ and $x + 2 > 0$, or $x < 1$ and $x > -2$, that is, if $-2 < x < 1$.



Observe that the graph of $y = (x-1)(x+2)$ crosses the x-axis at $x = 1$ and $x = -2$ (as shown on the right). Furthermore, the graph lies below the x-axis iff y is negative, that is, iff $-2 < x < 1$.

ABSOLUTE VALUES

9. Evaluate: (i) $|1-3| + |-7|$, (ii) $|-1-4| - 3 - |3-5|$, (iii) $||-2| - |-6||$.

Solution:

(i) $|1-3| + |-7| = |-2| + |-7| = 2 + 7 = 9$

(ii) $|-1-4| - 3 - |3-5| = |-5| - 3 - |-2| = 5 - 3 - 2 = 0$

(iii) $||-2| - |-6|| = |2-6| = |-4| = 4$

10. Rewrite without the absolute value sign: (i) $|x-2| < 5$, (ii) $|2x+3| < 7$.

Solution:

(i) $-5 < x-2 < 5$ or $-3 < x < 7$

(ii) $-7 < 2x+3 < 7$ or $-10 < 2x < 4$ or $-5 < x < 2$

11. Rewrite using the absolute value sign: (i) $-2 < x < 6$, (ii) $4 < x < 10$.

Solution:

First rewrite each inequality so that a number and its negative appear at the ends of the inequality:

(i) Add -2 to each side of $-2 < x < 6$ to obtain $-4 < x-2 < 4$ which is equivalent to $|x-2| < 4$.

(ii) Add -7 to each side of $4 < x < 10$ to obtain $-3 < x-7 < 3$ which is equivalent to $|x-7| < 3$.

12. Prove Proposition A.4(iii): $|a+b| \leq |a| + |b|$.

Solution:

Method 1.

Since $|a| = \pm a$, $-|a| \leq a \leq |a|$; also $-|b| \leq b \leq |b|$. Then, adding,

$$-(|a| + |b|) \leq a + b \leq |a| + |b|$$

Therefore,

$$|a+b| \leq ||a| + |b|| = |a| + |b|$$

since $|a| + |b| \geq 0$.

Method 2.

Now $ab \leq |ab| = |a|\,|b|$ implies $2ab \leq 2\,|a|\,|b|$, and so

$$(a+b)^2 = a^2 + 2ab + b^2 \leq a^2 + 2\,|a|\,|b| + b^2 = |a|^2 + 2\,|a|\,|b| + |b|^2 = (|a| + |b|)^2$$

But $\sqrt{(a+b)^2} = |a+b|$ and so, by the square root of the above, $|a+b| \leq |a| + |b|$.

13. Prove Proposition A.4(v): $|a-c| \leq |a-b| + |b-c|$.

Solution:

$$|a-c| = |(a-b) + (b-c)| \leq |a-b| + |b-c|$$

LEAST UPPER BOUND AXIOM

14. Prove Theorem (Archimedean Order Axiom) A.5: The subset $\mathbf{N} = \{1, 2, 3, \ldots\}$ of $\mathbf{R}$ is not bounded from above.

Solution:

Suppose $\mathbf{N}$ is bounded from above. By the Least Upper Bound Axiom, $\sup(\mathbf{N})$ exists, say $b = \sup(\mathbf{N})$. Then $b-1$ is not an upper bound for $\mathbf{N}$ and so

$$\exists\ n_0 \in \mathbf{N} \quad \text{such that} \quad b - 1 < n_0 \quad \text{or} \quad b < n_0 + 1$$

But $n_0 \in \mathbf{N}$ implies $n_0 + 1 \in \mathbf{N}$, and so b is not an upper bound for $\mathbf{N}$, a contradiction. Hence $\mathbf{N}$ is not bounded from above.

15. Prove: Let a and b be positive real numbers. Then there exists a positive integer $n_0 \in \mathbf{N}$ such that $b < n_0 a$. In other words, some multiple of a is greater than b.

Solution:

Suppose n_0 does not exist, that is, $na < b$ for every $n \in \mathbf{N}$. Then, since a is positive, $n < b/a$ for every $n \in \mathbf{N}$, and so b/a is an upper bound for $\mathbf{N}$. This contradicts Theorem A.5 (Problem 14), and so n_0 does exist.

16. Prove: If a is a positive real number, i.e. $0 < a$, then there exists a positive integer $n_0 \in \mathbf{N}$ such that $0 < 1/n_0 < a$.

Solution:

Suppose n_0 does not exist, i.e. $a \leqq 1/n$ for every $n \in \mathbf{N}$. Then, multiplying both sides by the positive number n/a, we have $n \leqq 1/a$ for every $n \in \mathbf{N}$. Hence $\mathbf{N}$ is bounded by $1/a$, an impossibility. Consequently, n_0 does exist.

17. Prove Corollary A.6: There is a rational number q between any two distinct real numbers a and b.

Solution:

One of the real numbers, say a, is less than the other, i.e. $a < b$. If a is negative and b is positive, then the rational number 0 lies between them, i.e. $a < 0 < b$. We now prove the corollary for the case where a and b are both positive; the case where a and b are negative is proven similarly, and the case where a or b is zero follows from Problem 16.

Now $a < b$ means $b - a$ is positive and so, by the preceding problem,

$$\exists\ n_0 \in \mathbf{N} \quad \text{such that} \quad 0 < 1/n_0 < b - a \quad \text{or} \quad a + (1/n_0) < b$$

We claim that there is an integral multiple of n_0 which lies between a and b. Notice that $1/n_0 < b$ since $1/n_0 < a + (1/n_0) < b$. By Problem 15, some multiple of $1/n_0$ is greater than b. Let m_0 be the least positive integer such that $m_0/n_0 \geqq b$; hence $(m_0 - 1)/n_0 < b$. We claim that

$$a < \frac{m_0 - 1}{n_0} < b$$

Otherwise $\quad \dfrac{m_0 - 1}{n_0} \leqq a \quad$ and so $\quad \dfrac{m_0 - 1}{n_0} + \dfrac{1}{n_0} = \dfrac{m_0}{n_0} \leqq a + \dfrac{1}{n_0} < b$

which contradicts the definition of m_0. Thus $(m_0 - 1)/n_0$ is a rational number between a and b.

NESTED INTERVAL PROPERTY

18. Prove Theorem A.7 (Nested Interval Property): Let $I_1 = [a_1, b_1]$, $I_2 = [a_2, b_2]$, ... be a sequence of nested closed (bounded) intervals, i.e. $I_1 \supset I_2 \supset \cdots$. Then there exists at least one point common to every interval.

Solution:

Now $I_1 \supset I_2 \supset \cdots$ implies that $a_1 \leqq a_2 \leqq \cdots$ and $\cdots \leqq b_2 \leqq b_1$. We claim that

$$a_m < b_n \quad \text{for every} \quad m, n \in \mathbf{N}$$

for, $m > n$ implies $a_m < b_m \leqq b_n$ and $m \leqq n$ implies $a_m \leqq a_n < b_n$. Thus each b_n is an upper bound for the set $A = \{a_1, a_2, \ldots\}$ of left end points. By the Least Upper Bound Axiom of $\mathbf{R}$, $\sup(A)$ exists; say, $p = \sup(A)$. Now $p \leqq b_n$, for every $n \in \mathbf{N}$, since each b_n is an upper bound for A and p is the least upper bound. Furthermore, $a_n \leqq p$ for every $n \in \mathbf{N}$, since p is an upper bound for $A = \{a_1, a_2, \ldots\}$. But

$$a_n \leqq p \leqq b_n \quad \Rightarrow \quad p \in I_n = [a_n, b_n]$$

Hence p is common to every interval.

19. Suppose, in the preceding problem, that the lengths of the intervals tend to zero, i.e. $\lim_{n\to\infty} (b_n - a_n) = 0$. Show that there would then exist exactly one point common to every interval. Recall that $\lim_{n\to\infty} (b_n - a_n) = 0$ means that, for every $\epsilon > 0$,

$$\exists\ n_0 \in \mathbf{N} \quad \text{such that} \quad n > n_0 \Rightarrow (b_n - a_n) < \epsilon$$

Solution:

Suppose p_1 and p_2 belong to every interval. If $p_1 \neq p_2$, then $|p_1 - p_2| = \delta > 0$. Since $\lim_{n\to\infty} (b_n - a_n) = 0$, there exists an interval $I_{n_0} = [a_{n_0}, b_{n_0}]$ such that the length of I_{n_0} is less than the distance $|p_1 - p_2| = \delta$ between p_1 and p_2. Accordingly, p_1 and p_2 cannot both belong to I_{n_0}, a contradiction. Thus $p_1 = p_2$, i.e. only one point can belong to every interval.

Supplementary Problems

FIELD AXIOMS

20. Show that the Right Distributive Law **[D$_2$]** is a consequence of the Left Distributive Law **[D$_1$]** and the Commutative Law **[M$_5$]**.

21. Show that the set **Q** of rational numbers under addition and multiplication is a field.

22. Show that the following set A of real numbers under addition and multiplication is a field:

$$A = \{a + b\sqrt{2} : a, b \text{ rational}\}$$

23. Show that the set $A = \{\ldots, -4, -2, 0, 2, 4, \ldots\}$ of even integers under addition and multiplication satisfies all the axioms of a field except **[M$_3$]**, **[M$_4$]** and **[M$_5$]**, that is, is a ring.

INEQUALITIES AND POSITIVE NUMBERS

24. Rewrite so that x is alone between the inequality signs:
(i) $4 < -2x < 10$, (ii) $-1 < 2x - 3 < 5$, (iii) $-3 < 5 - 2x < 7$.

25. Prove: The product of any two negative numbers is positive.

26. Prove Theorem A.2(iii): If $a < b$, then $a + c < b + c$.

27. Prove Theorem A.2(iv): If $a < b$ and c is positive, then $ac < bc$.

28. Prove Corollary A.3: The set **R** of real numbers is totally ordered by the relation $a \leq b$.

29. Prove: If $a < b$ and c is positive, then: (i) $\frac{a}{c} < \frac{b}{c}$, (ii) $\frac{c}{b} < \frac{c}{a}$.

30. Prove: $\sqrt{ab} \leq (a+b)/2$. More generally, prove $\sqrt[n]{a_1 a_2 \cdots a_n} \leq (a_1 + a_2 + \cdots + a_n)/n$.

31. Prove: Let a and b be real numbers such that $a < b + \epsilon$ for every $\epsilon > 0$. Then $a \leq b$.

32. Determine all real values of x such that: (i) $x^3 + x^2 - 6x > 0$, (ii) $(x-1)(x+3)^2 \leq 0$.

ABSOLUTE VALUES

33. Evaluate: (i) $|-2| + |1-4|$, (ii) $|3-8| - |1-9|$, (iii) $\big||-4| - |2-7|\big|$.

34. Rewrite, using the absolute value sign: (i) $-3 < x < 9$, (ii) $2 \leq x \leq 8$, (iii) $-7 < x < -1$.

35. Prove: (i) $|-a| = |a|$, (ii) $a^2 = |a|^2$, (iii) $|a| = \sqrt{a^2}$, (iv) $|x| < a$ iff $-a < x < a$.

Lecture in Application of Linear Algebra

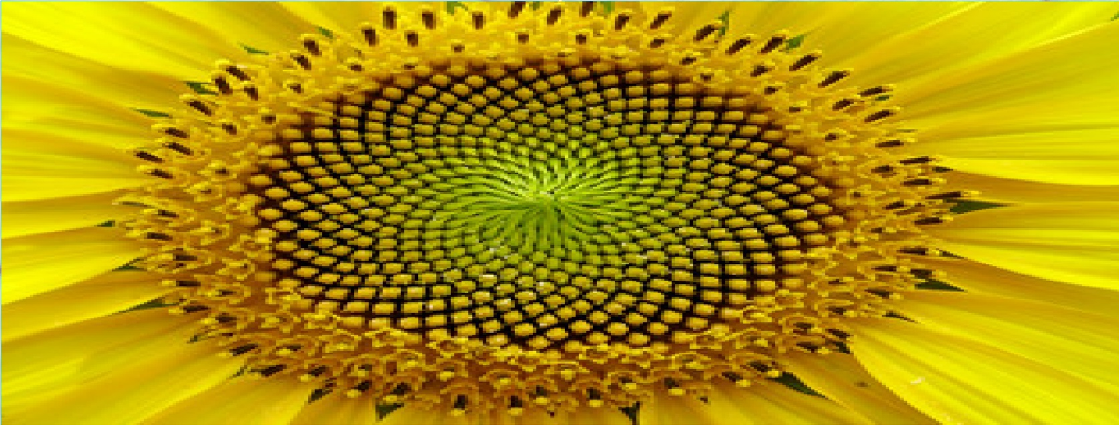
Preparing by

Dr. Amr M. Elrawy

DEPARTMENT OF MATHEMATICS, FACULTY OF SCIENCE, SOUTH VALLEY UNIVERSITY

PREPARING BY DR. AMR M. ELRAWY

First release, 2021



Contents

Part I

Topics in Linear Algebra

Chapte 1

Linear Transformations

1. Linear transformations

The central objective of linear algebra is the analysis of linear functions defined on a finite-dimensional vector space.In this chapter, we define the concept of a linear function or transformation.

Definition 1.0.1 Let V and W be real vector spaces (their dimensions can be different), and let T be a function with domain V and range in W (written $T : V \to W$). We say T is a linear transformation if
(i) For all $\mathbf{x}, \mathbf{y} \in V, T(\mathbf{x}+\mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y})$($T$ is additive)
(ii) For all $\mathbf{x} \in V, r \in \mathbb{K}, T(r\mathbf{x}) = rT(\mathbf{x})$($T$ is homogeneous).

R If $V = W$, then T can be called a linear operator.

R The homogeneity and additivity properties of a linear transformation $T : V \to W$ can be used in combination to show that if $\mathbf{x}$ and $\mathbf{y}$

are vectors in V and r and s are any scalars, then

$$T(r\mathbf{x}+s\mathbf{y}) = rT(\mathbf{x})+sT(\mathbf{y}).$$

■ **Example 1.1** Show that $T: R^2 \to R^2$, defined by

$$T((x_1,x_2)) = (x_1+x_2, x_1-x_2)$$

is a linear transformation.
Solution:
Let $u=(x_1,x_2)$, $v=(y_1,y_2) \in R^2$, then

$(i) T(u+v) = T((x_1+y_1, x_2+y_2))$
$= (x_1+y_1+x_2+y_2, x_1+y_1-x_2-y_2)$
$= (x_1+x_2, x_1-x_2)+(y_1+y_2, y_1-y_2)$
$= T(u)+T(v).$

$(ii) T(ru) = T(r(x_1, x_2+y))$
$= (rx_1+rx_2, rx_1-rx_2)$
$= r(x_1+x_2, x_1-x_2).$
$= rT(u).$

Therefore, T is linear transformation. ■

■ **Example 1.2** Show that $T: R^2 \to R^2$, defined by

$$T((x_1,x_2)) = (x_1+x_2, x_1-x_2+1)$$

is not linear transformation.
Solution: let $(0,0),(1,1) \in R^2$, then

$$T((0,0)+(1,1)) = T(1,1) = (2,1) \neq (2,2).$$

Therefore, T is not linear transformation. ■

Theorem 1.0.1 If $T: V \to W$ is a linear transformation, then:
(a) $T(\mathbf{0}) = \mathbf{0}$.
$(b) T(\mathbf{u}-\mathbf{v}) = T(\mathbf{u})-T(\mathbf{v})$ for all $\mathbf{u}$ and $\mathbf{v}$ in V.

Proof. Let $\mathbf{u}$ be any vector in V. Since $0\mathbf{u} = \mathbf{0}$, it follows from the homogeneity property in Definition 1 that

$$T(\mathbf{0}) = T(0\mathbf{u}) = 0T(\mathbf{u}) = \mathbf{0}$$

which proves (a) We can prove part (b) by rewriting $T(\mathbf{u} - \mathbf{v})$ as

$$\begin{aligned} T(\mathbf{u} - \mathbf{v}) &= T(\mathbf{u} + (-1)\mathbf{v}) \\ &= T(\mathbf{u}) + (-1)T(\mathbf{v}) \\ &= T(\mathbf{u}) - T(\mathbf{v}) \end{aligned}$$

We leave it for you to justify each step. ■

■ **Example 1.3** Show that $T : V \to W$, defined by $T(v) = 0$ for every v in V is a linear transformation called the **zero transformation**.
Solution:

$$T(u+v) = 0,\ T(u) = 0,\ T(v) = 0, \text{ and } T(ku) = 0.$$

Therefore,

$$T(u+v) = T(u) + T(v) \text{ and } T(ku) = kT(u).$$

■

■ **Example 1.4** Show that $T : V \to V$, defined by $T(v) = v$ for every v in V is a linear transformation called the **identity operator** on V.
Solution:

$$T(u+v) = u+v,\ T(u) = u,\ T(v) = v, \text{ and } T(ku) = ku.$$

Therefore,

$$T(u+v) = T(u) + T(v) \text{ and } T(ku) = kT(u).$$

■

■ **Example 1.5** Show that $T : V \to V$, defined by $T(v) = kv$ for every v in V and k any scalar is a linear transformation.
Solution:

$$T(u+v) = k(u+v) = ku+kv = T(u)+T(v),$$

and,

$$T(ru) = k(ru) = rku = rT(u).$$

■

R In the above Example 1.5. If $0 < k < 1$, then T is called **the contraction** of V with factor k, and if $k > 1$, it is called **the dilation** of V with factor k.

■ **Example 1.6** Let M_{nn} be the vector space of $n \times n$ matrices. In each part determine whether the transformation is linear.
(a) $T_1(A) = A^T$.
(b) $T_2(A) = \det(A)$.

Solution:

(a)

$$(i) T_1(A+B) = (A+B)^T = A^T + B^T = T_1(A) + T_1(B).$$
$$(ii) T_1(kA) = (kA)^T = kA^T = kT_1(A),$$

so T_1 is linear.

(b) T_2 is not linear for

$$T_2(A+B) = \det(A+B) \neq \det(A) + \det(B) = T_2(A) + T_2(B).$$

■

1.1 Finding linear transformations from images of basis vectors

In this section, we show how to find the linear transformations from images of basis vectors.

If $T_A : R^n \to R^m$ is multiplication by A, and if $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ are the standard basis vectors for R^n, then A can be expressed as

$$A = [T(\mathbf{e}_1)\,|T(\mathbf{e}_2)| \cdots |\, T(\mathbf{e}_n)]$$

and we say A is *a matrix transformation*.
It follows from this that the image of any vector $\mathbf{v} = (c_1, c_2, \ldots, c_n)$ in R^n under multiplication by A can be expressed as

$$T_A(\mathbf{v}) = c_1 T_A(\mathbf{e}_1) + c_2 T_A(\mathbf{e}_2) + \cdots + c_n T_A(\mathbf{e}_n)$$

This formula tells us that for a matrix transformation the image of any vector is expressible as a linear combination of the images of the standard basis vectors. This is a special case of the following more general result.

Theorem 1.1.1 Let $T : V \to W$ be a linear transformation, where V is finite-dimensional. If $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is a basis for V,| then the image of any vector $\mathbf{v}$ in V can be expressed as

$$T(\mathbf{v}) = c_1 T(\mathbf{v}_1) + c_2 T(\mathbf{v}_2) + \cdots + c_n T(\mathbf{v}_n)$$

where $c_1, c_2, \ldots, c_n$ are the coefficients required to express $\mathbf{v}$ as a linear combination of the vectors in the basis S.

Proof. we write $\mathbf{v}$ as $\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n$ and use the linearity of T ■

■ **Example 1.7** Let $T : R^3 \to R^2$ be the linear transformation for which

$$T(\mathbf{v}_1) = (1,0), \quad T(\mathbf{v}_2) = (2,-1), \quad T(\mathbf{v}_3) = (4,3).$$

Find a formula for $T(x_1, x_2, x_3)$, and then use that formula to compute $T(2,-3,5)$ by using the basis $S = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ for R^3, where

$$\mathbf{v}_1 = (1,1,1), \quad \mathbf{v}_2 = (1,1,0), \quad \mathbf{v}_3 = (1,0,0).$$

Solution:
We first need to express $\mathbf{x} = (x_1, x_2, x_3)$ as a linear combination of $\mathbf{v}_1, \mathbf{v}_2$, and $\mathbf{v}_3$. If we write

$$(x_1, x_2, x_3) = c_1(1,1,1) + c_2(1,1,0) + c_3(1,0,0)$$

then on equating corresponding components, we obtain

$$\begin{aligned} c_1 + c_2 + c_3 &= x_1 \\ c_1 + c_2 &= x_2 \\ c_1 &= x_3 \end{aligned}$$

which yields $c_1 = x_3, c_2 = x_2 - x_3, c_3 = x_1 - x_2$, so

$$\begin{aligned} (x_1, x_2, x_3) &= x_3(1,1,1) + (x_2 - x_3)(1,1,0) + (x_1 - x_2)(1,0,0) \\ &= x_3\mathbf{v}_1 + (x_2 - x_3)\mathbf{v}_2 + (x_1 - x_2)\mathbf{v}_3 \end{aligned}$$

Thus

$$\begin{aligned} T(x_1, x_2, x_3) &= x_3 T(\mathbf{v}_1) + (x_2 - x_3) T(\mathbf{v}_2) + (x_1 - x_2) T(\mathbf{v}_3) \\ &= x_3(1,0) + (x_2 - x_3)(2,-1) + (x_1 - x_2)(4,3) \\ &= (4x_1 - 2x_2 - x_3, 3x_1 - 4x_2 + x_3) \end{aligned}$$

From this formula we obtain

$$T(2,-3,5) = (9,2,3).$$

■

1.2 Kernel and range

In this section, we define and study the kernel and range of linear transformations.

Recall that if A is an $m \times n$ matrix, then the null space of A consists of all vectors x in R^n such that $Ax = 0$, and the column space of A consists of all vectors b in R^m for which there is at least one vector x in R^n such that $Ax = b$. From the viewpoint of matrix transformations, the null space of A consists of all vectors in R^n that multiplication by A maps into 0, and the column space of A consists of all vectors in R^m that are images of at least one vector in R^n under multiplication by A. The following definition extends these ideas to general linear transformations.

Definition 1.2.1 If $T : V \to W$ is a linear transformation, then the set of vectors in V that T maps into $\mathbf{0}$ is called the kernel of T and is denoted by $\ker(T)$. The set of all vectors in W that are images under T of at least one vector in V is called the range of T and is denoted by $R(T)$

■ **Example 1.8** Let $T : V \to W$ be the zero transformation. Find $Ker(T)$ and $R(T)$.

Solution:

Since T maps every vector in V into $\mathbf{0}$, it follows that $\ker(T) = V$. Moreover, since $\mathbf{0}$ is the only image under T of vectors in V, it follows that $R(T) = \{\mathbf{0}\}$.
■

■ **Example 1.9** Let $I : V \to V$ be the identity operator. Find $Ker(T)$ and $R(T)$.

Solution:

Since $I(\mathbf{v}) = \mathbf{v}$ for all vectors in V, every vector in V is the image of some vector (namely, itself); thus $R(I) = V$. Since the only vector that I maps into $\mathbf{0}$ is $\mathbf{0}$, it follows that $\ker(I) = \{\mathbf{0}\}$. ■

1.2.1 Properties of kernel and range

Theorem 1.2.1 If $T : V \to W$ is a linear transformation, then:
(a) The kernel of T is a subspace of V.
(b) The range of T is a subspace of W.

Proof. (a) To show that $\ker(T)$ is a subspace, we must show that it contains at least one vector and is closed under addition and scalar multiplication. Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be vectors in $\ker(T)$, and let k be any scalar. Then

$$T(\mathbf{v}_1+\mathbf{v}_2)=T(\mathbf{v}_1)+T(\mathbf{v}_2)=\mathbf{0}+\mathbf{0}=\mathbf{0}$$

so $\mathbf{v}_1+\mathbf{v}_2$ is in $\ker(T)$. Also,

$$T(k\mathbf{v}_1)=kT(\mathbf{v}_1)=k\mathbf{0}=\mathbf{0}$$

so $k\mathbf{v}_1$ is in $\ker(T)$.

(b) To show that $R(T)$ is a subspace of W, we must show that it contains at least one vector and is closed under addition and scalar multiplication. However, it contains at least the zero vector of W since $T(\mathbf{0})=(\mathbf{0})$. To prove that it is closed under addition and scalar multiplication, we must show that if $\mathbf{w}_1$ and | $\mathbf{w}_2$ are vectors in $R(T)$, and if k is any scalar, then there exist vectors a and $\mathbf{b}$ in V for which

$$T(\mathbf{a})=\mathbf{w}_1+\mathbf{w}_2 \quad \text{and} \quad T(\mathbf{b})=k\mathbf{w}_1$$

But the fact that $\mathbf{w}_1$ and $\mathbf{w}_2$ are in $R(T)$ tells us there exist vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ in V such that

$$T(\mathbf{v}_1)=\mathbf{w}_1 \quad \text{and} \quad T(\mathbf{v}_2)=\mathbf{w}_2$$

The following computations complete the proof by showing that the vectors $\mathbf{a}=\mathbf{v}_1+\mathbf{v}_2$ and $\mathbf{b}=k\mathbf{v}_1$ satisfy the equations in (4):

$$\begin{aligned}
&T(\mathbf{a})=T(\mathbf{v}_1+\mathbf{v}_2)=T(\mathbf{v}_1)+T(\mathbf{v}_2)=\mathbf{w}_1+\mathbf{w}_2\\
&T(\mathbf{b})=T(k\mathbf{v}_1)=kT(\mathbf{v}_1)=k\mathbf{w}_1
\end{aligned}$$

■

1.3 Rank and nullity of linear transformations

In this section, we defined the notions of rank and nullity for an $m \times n$ matrix. Also, we proved that the sum of the rank and nullity is n.

Definition 1.3.1 Let $T : V \to W$ be a linear transformation. If the range of T is finite dimensional, then its dimension is called the rank of T; and if the kernel of T is finite-dimensional, then its dimension is called the nullity of T. The rank of T is denoted by rank (T) and the nullity of T by nullity (T).

Theorem 1.3.1 If $T : V \to W$ is a linear transformation from a finite-dimensional vector space V to a vector space W, then the range of T is finite-dimensional, and

$$\text{rank}(T) + \text{nullity}(T) = \dim(V).$$

Proof. Assume that V is n -dimensional. We must show that

$$\dim(R(T)) + \dim(\ker(T)) = n$$

We will give the proof for the case where $1 \leq \dim(\ker(T)) < n$. The cases where $\dim(\ker(T)) = 0$ and $\dim(\ker(T)) = n$ are left as exercises. Assume $\dim(\ker(T)) = r$ and let $\mathbf{v}_1, \ldots, \mathbf{v}_r$ be a basis for the kernel. Since $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ is linearly independent, Theorem 4.5.5(b) states that there are $n-r$ vectors, $\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n$, such that the extended set $\{\mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ is a basis for V. To complete the proof, we will show that the $n-r$ vectors in the set $S = \{T(\mathbf{v}_{r+1}), \ldots, T(\mathbf{v}_n)\}$ form a basis for the range of T. It will then follow that

$$\dim(R(T)) + \dim(\ker(T)) = (n-r) + r = n$$

First we show that S spans the range of T. If $\mathbf{b}$ is any vector in the range of T, then $\mathbf{b} = T(\mathbf{v})$ for some vector $\mathbf{v}$ in V. Since $\{\mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ is a basis for V, the vector v can be written in the form

$$\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_r\mathbf{v}_r + c_{r+1}\mathbf{v}_{r+1} + \cdots + c_n\mathbf{v}_n$$

Since $\mathbf{v}_1,\ldots,\mathbf{v}_r$ lie in the kernel of T, we have $T(\mathbf{v}_1)=\cdots=T(\mathbf{v}_r)=\mathbf{0}$, so

$$\mathbf{b}=T(\mathbf{v})=c_{r+1}T(\mathbf{v}_{r+1})+\cdots+c_nT(\mathbf{v}_n)$$

Thus S spans the range of T. Finally, we show that S is a linearly independent set and consequently forms a basis for the range of T. Suppose that some linear combination of the vectors in S is zero; that is,

$$k_{r+1}T(v_{r+1})+\cdots+k_nT(\mathbf{v}_n)=0 \tag{1.1}$$

We must show that $k_{r+1}=\cdots=k_n=0$. Since T is linear, (3.5) can be rewritten as

$$T(k_{r+1}\mathbf{v}_{r+1}+\cdots+k_n\mathbf{v}_n)=\mathbf{0}$$

which says that $k_{r+1}\mathbf{v}_{r+1}+\cdots+k_n\mathbf{v}_n$ is in the kernel of T. This vector can therefore be written as a linear combination of the basis vectors $\{v_1,\ldots,v_r\}$, say

$$k_{r+1}\mathbf{v}_{r+1}+\cdots+k_n\mathbf{v}_n=k_1\mathbf{v}_1+\cdots+k_r\mathbf{v}_r$$

Thus,

$$k_1\mathbf{v}_1+\cdots+k_r\mathbf{v}_r-k_{r+1}\mathbf{v}_{r+1}-\cdots-k_n\mathbf{v}_n=\mathbf{0}$$

Since $\{v_1,\ldots,v_n\}$ is linearly independent, all of the k 's are zero; in particular, $k_{r+1}=\cdots=k_n=0$, which completes the proof. ■

R In the special case where A is an $m\times n$ matrix and $T_A:R^n\to R^m$ is multiplication by A, the kernel of T_A is the null space of A, and the range of T_A is the column space o A. Thus, it follows from Theorem 1.3.1 that

$$\operatorname{rank}(T_A)+\operatorname{nullity}(T_A)=n$$

.

1.4 Composition linear transformations

In this section, we define one to one and onto linear transformations. Also, we discussed composition linear transformations.

Definition 1.4.1 If $T : V \to W$ is a linear transformation from a vector space V to a vector space W, then T is said to be one-to-one if T maps distinct vectors in V into distinct vectors in W, i.e.,

$$\forall u, v \in V, T(u) = T(v) \Rightarrow u = v.$$

Definition 1.4.2 If $T : V \to W$ is a linear transformation from a vector space V to a vector space W, then T is said to be onto (or onto W) if every vector in W is the image of at least one vector in V.

■ **Example 1.10** Let $T : R^2 \to R^2$ be a linear transformation define as

$$T(x,y) = (x, x+y).$$

(1) Find Ker T?
(2) Is T one-one?
(3) Is T onto?
Solution:
(1) Let $u = (x,y) \in KerT$, then

$$\begin{aligned} T(u) = O &\Rightarrow (x, x+y) = (0,0) \\ &\Rightarrow x = 0, x+y = 0 \\ &\Rightarrow x = y = 0, \end{aligned}$$

thus, $KerT = \{(0,0)\}$.
(2) Let $u = (x,y), v = (s,t) \in R^2$ and

$$\begin{aligned} T(u) = T(v) &\Rightarrow (x, x+y) = (s, s+t) \\ &\Rightarrow x = s, x+y = s+t \\ &\Rightarrow x = s, y = t \\ &\Rightarrow (x,y) = (s,t). \end{aligned}$$

thus, T is one-one.
(3) Let $v = (s,t) \in R^2$ and $v = T(u)$ for all $u = (x,y) \in R^2$, then

$$\begin{aligned} v = T(u) &\Rightarrow (s,t) = T(x,y) \\ &\Rightarrow (s,t) = (x, x+y) \\ &\Rightarrow x = s, y = t - s \\ &\Rightarrow (x,y) \in R^2. \end{aligned}$$

thus, T is onto. ■

■ **Example 1.11** Let $T : R^2 \to R^3$ be a linear transformation define as

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

(1) Find Ker T?
(2) Is T one-one?
(3) Is T onto?
Solution:
(1) Let $u = (x,y) \in KerT$, then

$$\begin{aligned} T(u) = O &\Rightarrow (x, x+y, z) = (0,0,0) \\ &\Rightarrow x = 0, x+y = 0, z = 0 \\ &\Rightarrow x = y = z = 0, \end{aligned}$$

thus, $KerT = \{(0,0)\}$.
(2) Let $u = (x,y), v = (s,t) \in R^2$ and

$$\begin{aligned} T(u) = T(v) &\Rightarrow (x, x+y, z) = (s, s+t, e) \\ &\Rightarrow x = s, x+y = s+t, z = e \\ &\Rightarrow x = s, y = t, z = e \\ &\Rightarrow (x,y) = (s,t) \\ &\Rightarrow u = v. \end{aligned}$$

thus, T is one-one.

(3) Let $v = (s,t,e) \in R^3$ and $v = T(u)$ for all $u = (x,y) \in R^2$, then

$$\begin{aligned} v = T(u) &\Rightarrow (s,t,e) = T(x,y) \\ &\Rightarrow (s,t,e) = (x, x+y, z) \\ &\Rightarrow x = s, y = t - s, z = e \\ &\Rightarrow x, y, z \in R. \end{aligned}$$

thus, T is onto. ■

Theorem 1.4.1 If $T : V \to W$ is a linear transformation, then the following statements are equivalent.
(a) T is one-to-one.
(b) $\ker(T) = \{\mathbf{0}\}$.

Proof. $(a) \Rightarrow (b)$ Since T is linear, we know that $T(\mathbf{0}) = \mathbf{0}$. Since T is one-to-one, there can be no other vectors in V that map into $\mathbf{0}$, so $\ker(T) = \{\mathbf{0}\}$.
$(b) \Rightarrow (a)$ Assume that $\ker(T) = \{0\}$. If $\mathbf{u}$ and $\mathbf{v}$ are distinct vectors in V, then $\mathbf{u} - \mathbf{v} \neq \mathbf{0}$. This implies that $T(\mathbf{u} - \mathbf{v}) \neq \mathbf{0}$, for otherwise $\ker(T)$ would contain a nonzero vector. Since T is linear, it follows that

$$T(\mathbf{u}) - T(\mathbf{v}) = T(\mathbf{u} - \mathbf{v}) \neq \mathbf{0}$$

so T maps distinct vectors in V into distinct vectors in W and hence is one-to-one. ■

Definition 1.4.3 If $T_1 : U \to V$ and $T_2 : V \to W$ are linear transformations, then the composition of T_2 with T_1, denoted by $T_2 \circ T_1$ (which is read " T_2 circle T_1"), is the function defined by the formula

$$(T_2 \circ T_1)(\mathbf{u}) = T_2(T_1(\mathbf{u}))$$

where $\mathbf{u}$ is a vector in U.

Theorem 1.4.2 Let $T_1 : U \to V$ and $T_2 : V \to W$ be a linear transformations, then $(T_2 \circ T_1) : U \to W$ is also a linear transformation.

Proof. Suppose that $\mathbf{u}$ and $\mathbf{v}$ are vectors in U and c is a scalar, then it follows from (1) and the linearity of T_1 and T_2 that

$$\begin{aligned}(T_2 \circ T_1)(\mathbf{u}+\mathbf{v}) &= T_2(T_1(\mathbf{u}+\mathbf{v})) = T_2(T_1(\mathbf{u})+T_1(\mathbf{v})) \\ &= T_2(T_1(\mathbf{u}))+T_2(T_1(\mathbf{v})) \\ &= (T_2 \circ T_1)(\mathbf{u})+(T_2 \circ T_1)(\mathbf{v})\end{aligned}$$

and

$$\begin{aligned}(T_2 \circ T_1)(c\mathbf{u}) &= T_2(T_1(c\mathbf{u})) = T_2(cT_1(\mathbf{u})) \\ &= cT_2(T_1(\mathbf{u})) = c(T_2 \circ T_1)(\mathbf{u})\end{aligned}$$

Thus, $T_2 \circ T_1$ satisfies the two requirements of a linear transformation. ■

■ **Example 1.12** Let $T_1 : P_1 \to P_2$ and $T_2 : P_2 \to P_2$ be the linear transformations given by the formulas

$$T_1(p(x)) = xp(x) \text{ and } T_2(p(x)) = p(2x+4)$$

Then find $(T_2 \circ T_1)$.
<u>Solution:</u>
The composition $(T_2 \circ T_1) : P_1 \to P_2$ is given by the formula

$$(T_2 \circ T_1)(p(x)) = T_2(T_1(p(x))) = T_2(xp(x)) = (2x+4)p(2x+4)$$

In particular, if $p(x) = c_0 + c_1x$, then

$$\begin{aligned}(T_2 \circ T_1)(p(x)) &= (T_2 \circ T_1)(c_0+c_1x) = (2x+4)(c_0+c_1(2x+4)) \\ &= c_0(2x+4)+c_1(2x+4)^2.\end{aligned}$$

■

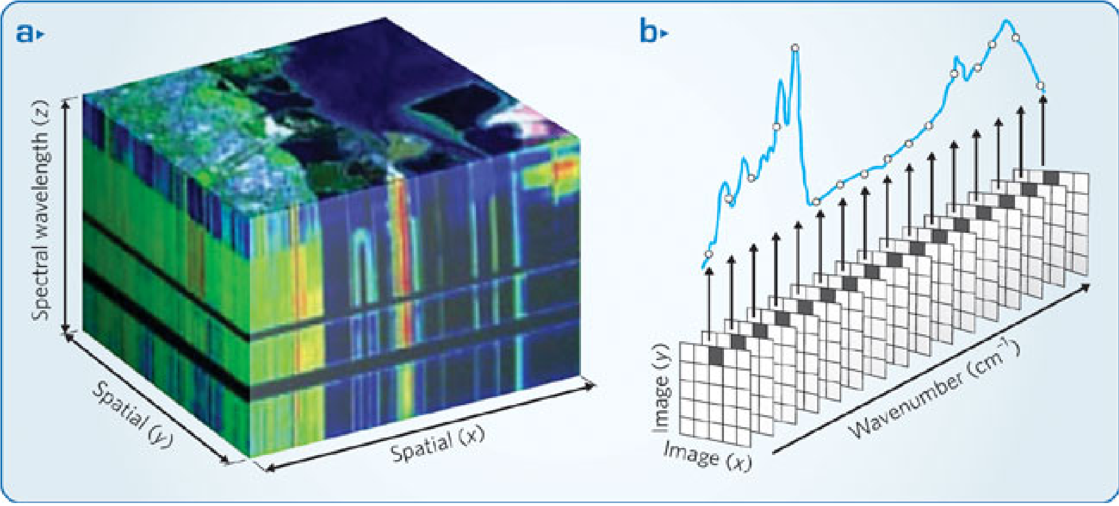
1.5 Exercises

1. Suppose that T is a mapping whose domain is the vector space M_{22}. In each part, determine whether T is a linear transformation, and if so, find its kernel.
   (a) $T(A) = A^2$.
   (b) $T(A) = \text{tr}(A)$.
   (c) $T(A) = A + A^T$.
   (d) $T(A) = (A)_{11}$
   (e) $T(A) = 0_{2\times 2}$
   (f) $T(A) = cA$
2. Determine whether the mapping T is a linear transformation, and if so, find its kernel.
   (a) $T : R^3 \to R$, where $T(\mathbf{u}) = \|\mathbf{u}\|$.
   (b) $T : R^3 \to R^3$, where v_0 is a fixed vector in R^3 and $T(\mathbf{u}) = \mathbf{u} \times \mathbf{v}_0$.
   (c) $T : M_{22} \to M_{23}$, where B is a fixed 2×3 matrix and $T(A) = AB$.
   (d) $T : M_{22} \to R$, where
   (i) $T\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = 3a - 4b + c - d$
   (ii) $T\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = a^2 + b^2$
   (e) $T : P_2 \to P_2$, where
   (i) $T\left(a_0 + a_1x + a_2x^2\right) = a_0 + a_1(x+1) + a_2(x+1)^2$
   (ii) $T\left(a_0 + a_1x + a_2x^2\right) = (a_0+1) + (a_1+1)x + (a_2+1)x^2$
   (f) $T : F(-\infty,\infty) \to F(-\infty,\infty)$, where (a) $T(f(x)) = 1 + f(x)$ (b) $T(f(x)) = f(x+1)$
3. Let $T : P_2 \to P_3$ be the linear transformation defined by $T(p(x)) = xp(x)$. Which of the following are in $\ker(T)$?
   (a) x^2
   (b) 0
   (c) $1+x$
   (d) $-x$
4. Let $T : M_{22} \to M_{22}$ be the dilation operator with factor $k = 3$.
   (a) Find $T\left(\begin{bmatrix} 1 & 2 \\ -4 & 3 \end{bmatrix}\right)$.
   (b) Find the rank and nullity of T.

5. Let $T : P_2 \to P_2$ be the contraction operator with factor $k = 1/4$
   (a) Find $T\left(1+4x+8x^2\right)$.
   (b) Find the rank and nullity of T.
6. Determine whether the linear transformation is one-to-one and onto by finding its kernel:
   (a) $T : R^2 \to R^2$, where $T(x,y) = (y,x)$.
   (b) $T : R^2 \to R^3$, where $T(x,y) = (x,y,x+y)$.
   (c) $T : R^3 \to R^2$, where $T(x,y,z) = (x+y+z, x-y-z)$.
   (d) $T : R^2 \to R^3$, where $T(x,y) = (x-y, y-x, 2x-2y)$.
   (e) $T : R^2 \to R^2$, where $T(x,y) = (0, 2x+3y)$.
   (f) $T : R^2 \to R^2$, where $T(x,y) = (x+y, x-y)$.

Chapte 2

Eigenvalues and Eigenvectors



2. Eigenvalues and Eigenvectors

In this chapter, we'll look at the "eigenvalues" and "eigenvectors" of scalars and vectors, words derived from the German word eigen, which means "own," "peculiar to," "characteristic," or "individual." The fundamental definition was first used in the study of rotational motion, but it was later applied to distinguish various types of surfaces and to explain solutions to differential equations.

Definition 2.0.1 If A is an $n \times n$ matrix, then a nonzero vector $\mathbf{x}$ in R^n is called an eigenvector of A (or of the matrix operator T_A) if $A\mathbf{x}$ is a scalar multiple of $\mathbf{x}$; that is,

$$A\mathbf{x} = \lambda \mathbf{x}$$

for some scalar λ. The scalar λ is called an **eigenvalue** of A (or of T_A), and $\mathbf{x}$ is said to be an **eigenvector** corresponding to λ.

Computing Eigenvalues and Eigenvectors

Our next goal is to establish a general method for determining the eigenvalues and eigenvectors of an $n \times n$ matrix A. We will begin with the problem of finding the eigenvalues of A. Note first that the equation $A\mathbf{x} = \lambda\mathbf{x}$ can be rewritten as $A\mathbf{x} = \lambda I\mathbf{x}$, or equivalently, as

$$(\lambda I - A)\mathbf{x} = \mathbf{0}$$

For λ to be an eigenvalue of A this equation must have a nonzero solution for $\mathbf{x}$. The coefficient matrix $\lambda I - A$ has a zero determinant. Thus, we have the following result.

Theorem 2.0.1 If A is an $n \times n$ matrix, then λ is an eigenvalue of A if and only if it satisfies the equation

$$\det(\lambda I - A) = 0 \tag{2.1}$$

This is called the characteristic equation of A.

■ **Example 2.1** Finding eigenvalues of the matrix

$$A = \begin{bmatrix} 3 & 0 \\ 8 & -1 \end{bmatrix}.$$

Solution:
The eigenvalues of A are the solutions of the equation

$$\det(\lambda I - A) = 0,$$

which we can write as

$$\begin{vmatrix} \lambda - 3 & 0 \\ -8 & \lambda + 1 \end{vmatrix} = 0,$$

from which we obtain

$$(\lambda - 3)(\lambda + 1) = 0.$$

Thus, the eigenvalues of A are $\lambda = 3$ and $\lambda = -1$. ■

When the determinant $\det(\lambda I - A)$ in (2.1) is expanded, the characteristic equation of A takes the form

$$\lambda^n + c_1\lambda^{n-1} + \cdots + c_n = 0 \tag{2.2}$$

where the left side of this equation is a polynomial of degree n in which the coefficient of λ^n is 1. The polynomial

$$p(\lambda) = \lambda^n + c_1\lambda^{n-1} + \cdots + c_n \tag{2.3}$$

is called the **characteristic polynomial** of A.

■ **Example 2.2** Recall Example (4.3), the characteristic polynomial of the 2×2 matrix is

$$p(\lambda) = (\lambda - 3)(\lambda + 1) = \lambda^2 - 2\lambda - 3$$

which is a polynomial of degree 2 . ■

Since a polynomial of degree n has at most n distinct roots, it follows from (2.2) that the characteristic equation of an $n \times n$ matrix A has at most n distinct solutions and consequently the matrix has at most n distinct eigenvalues. Since some of these solutions may be complex numbers, it is possible for a matrix to have complex eigenvalues, even if that matrix itself has real entries.

■ **Example 2.3** Find the eigenvalues of

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 4 & -17 & 8 \end{bmatrix}$$

Solution:
The characteristic polynomial of A is

$$\det(\lambda I - A) = \det \begin{bmatrix} \lambda & -1 & 0 \\ 0 & \lambda & -1 \\ -4 & 17 & \lambda - 8 \end{bmatrix} = \lambda^3 - 8\lambda^2 + 17\lambda - 4$$

The eigenvalues of A must therefore satisfy the cubic equation

$$\lambda^3 - 8\lambda^2 + 17\lambda - 4 = 0,$$

and we can be rewritten the above equation as

$$(\lambda - 4)(\lambda^2 - 4\lambda + 1) = 0,$$

Thus, the eigenvalues of A are

$$\lambda = 4, \lambda = 2 + \sqrt{3}, \; and \; \lambda = 2 - \sqrt{3}.$$

■

■ **Example 2.4** Find the eigenvalues of the upper triangular matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{bmatrix}.$$

Solution:
The determinant of a triangular matrix is the product of the entries on the main diagonal, we obtain

$$\det(\lambda I - A) = \det \begin{bmatrix} \lambda - a_{11} & -a_{12} & -a_{13} & -a_{14} \\ 0 & \lambda - a_{22} & -a_{23} & -a_{24} \\ 0 & 0 & \lambda - a_{33} & -a_{34} \\ 0 & 0 & 0 & \lambda - a_{44} \end{bmatrix}$$
$$= (\lambda - a_{11})(\lambda - a_{22})(\lambda - a_{33})(\lambda - a_{44})$$

Thus, the characteristic equation is

$$(\lambda - a_{11})(\lambda - a_{22})(\lambda - a_{33})(\lambda - a_{44}) = 0$$

and the eigenvalues are

$$\lambda = a_{11}, \quad \lambda = a_{22}, \quad \lambda = a_{33}, \quad \lambda = a_{44}$$

which are precisely the diagonal entries of A. ■

Theorem 2.0.2 If A is an $n \times n$ triangular matrix (upper triangular, lower triangular, or diagonal), then the eigenvalues of A are the entries on the main diagonal of A

Theorem 2.0.3 If A is an $n \times n$ matrix, the following statements are equivalent.
(a) λ is an eigenvalue of A.
(b) λ is a solution of the characteristic equation $\det(\lambda I - A) = 0$.
(c) The system of equations $(\lambda I - A)\mathbf{x} = \mathbf{0}$ has nontrivial solutions.
(d) There is a nonzero vector $\mathbf{x}$ such that $A\mathbf{x} = \lambda\mathbf{x}$.

Finding Eigenvectors and Bases for Eigenspaces:
Now that we know how to find the eigenvalues of a matrix, we will consider the problem of finding the corresponding eigenvectors. By definition, the eigenvectors of A corresponding to an eigenvalue λ are the nonzero vectors that satisfy

$$(\lambda I - A)x = O.$$

Thus, we can find the eigenvectors of A corresponding to λ by finding the nonzero vector x which it is a solution of the system $(\lambda I - A)x = O$.

■ **Example 2.5** Find bases for the eigenvectors of the matrix

$$A = \begin{bmatrix} -1 & 3 \\ 2 & 0 \end{bmatrix}.$$

Solution:
The characteristic equation of A is

$$\begin{vmatrix} \lambda+1 & -3 \\ -2 & \lambda \end{vmatrix} = \lambda(\lambda+1) - 6 = (\lambda-2)(\lambda+3) = 0$$

so the eigenvalues of A are $\lambda = 2$ and $\lambda = -3$.
Thus, there are two eigenvectors of A, one for each eigenvalue. By definition, suppose that

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

is an eigenvector of A corresponding to an eigenvalue λ if and only if $(\lambda I - A)\mathbf{x} = 0$, that is,

$$\begin{bmatrix} \lambda+1 & -3 \\ -2 & \lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

In the case where $\lambda = 2$ this equation becomes

$$\begin{bmatrix} 3 & -3 \\ -2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

whose general solution is

$$x_1 = t, \quad x_2 = t.$$

Since this can be written in matrix form as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} t \\ t \end{bmatrix} = t \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

it follows that

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

is a basis for the eigenvectors corresponding to $\lambda = 2$. We leave it for you to follow the pattern of these computations and show that

$$\begin{bmatrix} -\frac{3}{2} \\ 1 \end{bmatrix}$$

is a basis for the eigenspace corresponding to $\lambda = -3$. ■

■ **Example 2.6** Find bases for the eigenvectors of

$$A = \begin{bmatrix} 0 & 0 & -2 \\ 1 & 2 & 1 \\ 1 & 0 & 3 \end{bmatrix}.$$

Solution:
The characteristic equation of A is

$$\lambda^3 - 5\lambda^2 + 8\lambda - 4 = 0,$$

or in factored form,

$$(\lambda - 1)(\lambda - 2)^2 = 0.$$

Thus, the distinct eigenvalues of A are $\lambda = 1$ and $\lambda = 2$, so there are two eigenvectors of A. By definition,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

is an eigenvector of A corresponding to λ if and only if $\mathbf{x}$ is a nontrivial solution of $(\lambda I - A)\mathbf{x} = 0$, or in matrix form,

$$\begin{bmatrix} \lambda & 0 & 2 \\ -1 & \lambda - 2 & -1 \\ -1 & 0 & \lambda - 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

In the case where $\lambda = 2$, the above equation becomes

$$\begin{bmatrix} 2 & 0 & 2 \\ -1 & 0 & -1 \\ -1 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Solving this system using Gaussian elimination yields

$$x_1 = -s, \quad x_2 = t, \quad x_3 = s$$

Thus, the eigenvectors of A corresponding to $\lambda = 2$ are the nonzero vectors of the form

$$\mathbf{x} = \begin{bmatrix} -s \\ t \\ s \end{bmatrix} = \begin{bmatrix} -s \\ 0 \\ s \end{bmatrix} + \begin{bmatrix} 0 \\ t \\ 0 \end{bmatrix} = s \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} + t \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Sin

$$\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

are linearly independent (why?), these vectors form a basis for the eigenvector corresponding to $\lambda = 2$.
If $\lambda = 1$, then $(\lambda I - A)x = O$ becomes

$$\begin{bmatrix} 1 & 0 & 2 \\ -1 & -1 & -1 \\ -1 & 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Solving this system yields (verify)

$$x_1 = -2s, \quad x_2 = s, \quad x_3 = s$$

Thus, the eigenvectors corresponding to $\lambda = 1$ are the nonzero vectors of the form

$$\begin{bmatrix} -2s \\ s \\ s \end{bmatrix} = s \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} \text{ so that } \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$$

is a basis for the eigenspace corresponding to $\lambda = 1$. ■

The next theorem establishes a relationship between the eigenvalues and the invertibility of a matrix.

Theorem 2.0.4 A square matrix A is invertible if and only if $\lambda = 0$ is not an eigenvalue of A.

Proof. Assume that A is an $n \times n$ matrix and observe first that $\lambda = 0$ is a solution of the characteristic equation

$$\lambda^n + c_1\lambda^{n-1} + \cdots + c_n = 0$$

if and only if the constant term c_n is zero. Thus, it suffices to prove that A is invertible if and only if $c_n \neq 0$. But

$$\det(\lambda I - A) = \lambda^n + c_1\lambda^{n-1} + \cdots + c_n$$

or, on setting $\lambda = 0$,

$$\det(-A) = c_n \quad \text{or} \quad (-1)^n \det(A) = c_n$$

It follows from the last equation that $\det(A) = 0$ if and only if $c_n = 0$, and this in turn implies that A is invertible if and only if $c_n \neq 0$. ■

2.1 Exercises

A- Find eigenvalue and eigenvectors of the following:

1. $A = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix}.$

2. $A = \begin{bmatrix} 5 & -1 \\ 1 & 3 \end{bmatrix}.$

3. $A = \begin{bmatrix} 4 & 0 & 1 \\ 2 & 3 & 2 \\ 1 & 0 & 4 \end{bmatrix}.$

4. $A = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}.$

B- Find the characteristic equation, the eigenvalues, and bases for the eigenvectors of the matrix.

(a) $\begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix}.$

(b) $\begin{bmatrix} -2 & -7 \\ 1 & 2 \end{bmatrix}.$

(c) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$

(d) $\begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}.$

(e) $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$

(f) $\begin{bmatrix} 2 & -3 \\ 0 & 2 \end{bmatrix}.$

(i) $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$

(u) $\begin{bmatrix} 1 & 2 \\ -2 & -1 \end{bmatrix}.$

C- Find the characteristic equation, the eigenvalues, and bases for the eigenvectors of the matrix.

1. $\begin{bmatrix} 4 & 0 & 1 \\ -2 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix}$.
2. $\begin{bmatrix} 1 & 0 & -2 \\ 0 & 0 & 0 \\ -2 & 0 & 4 \end{bmatrix}$.
3. $\begin{bmatrix} 6 & 3 & -8 \\ 0 & -2 & 0 \\ 1 & 0 & -3 \end{bmatrix}$.
4. $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$.
5. $\begin{bmatrix} 4 & 0 & -1 \\ 0 & 3 & 0 \\ 1 & 0 & 2 \end{bmatrix}$.
6. $\begin{bmatrix} 1 & -3 & 3 \\ 3 & -5 & 3 \\ 6 & -6 & 4 \end{bmatrix}$.

2.2 Diagonalization

Definition 2.2.1 If A and B are square matrices, then we say that B is similar to A if there is an invertible matrix P such that

$$B = P^{-1}AP.$$

Note that if B is similar to A, then it is also true that A is similar to B since we can express A as $A = Q^{-1}BQ$ by taking $Q = P^{-1}$. This being the case, we will usually say that A and B are similar matrices if either is similar to the other.

Definition 2.2.2 A square matrix A is said to be diagonalizable if it is similar to some diagonal matrix; that is, if there exists an invertible matrix P such that $P^{-1}AP$ is diagonal. In this case the matrix P is said to diagonalize A.

Theorem 2.2.1 (a) If $\lambda_1, \lambda_2, \ldots, \lambda_k$ are distinct eigenvalues of a matrix A, and if $v_1, v_2, \ldots, v_k$ are corresponding eigenvectors, then $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ is a linearly independent set.
(b) An $n \times n$ matrix with n distinct eigenvalues is diagonalizable.

A Procedure for Diagonalizing an $n \times n$ Matrix

Step 1 . Determine first whether the matrix is actually diagonalizable by searching for n linearly independent eigenvectors. One way to do this is to find a basis for each eigenvector and count the total number of vectors obtained. If there is a total of n vectors, then the matrix is diagonalizable, and if the total is less than n, then it is not.
Step 2. If you ascertained that the matrix is diagonalizable, then form the matrix $P = [\mathbf{p}_1\ \mathbf{p}_2 \ldots. \mathbf{p}_n]$ whose column vectors are the n basis vectors you obtained in Step 1.
Step 3. P^{-1} AP will be a diagonal matrix whose successive diagonal entries are the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ that correspond to the successive columns of P.

■ **Example 2.7** Find a matrix P that diagonalizes

$$A=\begin{bmatrix} 0 & 0 & -2 \\ 1 & 2 & 1 \\ 1 & 0 & 3 \end{bmatrix}.$$

Solution:
The characteristic equation of A to be

$$(\lambda-1)(\lambda-2)^2=0$$

and we found the following bases for the eigenvalues:

$$\lambda=2: \quad \mathrm{p}_1=\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \quad \mathrm{P}_2=\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}: \quad \lambda=1: \quad \mathrm{p}_3=\begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$$

There are three basis vectors in total, so the matrix

$$P=\begin{bmatrix} -1 & 0 & -2 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

diagonalizes A. As a check, you should verify that

$$P^{-1}AP=\begin{bmatrix} 1 & 0 & 2 \\ 1 & 1 & 1 \\ -1 & 0 & -1 \end{bmatrix}\begin{bmatrix} 0 & 0 & -2 \\ 1 & 2 & 1 \\ 1 & 0 & 3 \end{bmatrix}\begin{bmatrix} -1 & 0 & -2 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}=\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

In general, there is no preferred order for the columns of P. Since the i th diagonal entry of $P^{-1}AP$ is an eigenvalue for the i th column vector of P, changing the order of the columns of P just changes the order of the eigenvalues on the diagonal of $P^{-1}AP$. Thus, had we written

$$P=\begin{bmatrix} -1 & -2 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

in the preceding example, we would have obtained

$$P^{-1}AP = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

■

■ **Example 2.8** Show that the following matrix is not diagonalizable:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ -3 & 5 & 2 \end{bmatrix}.$$

Solution:
The characteristic polynomial of A is

$$\det(\lambda I - A) = \begin{vmatrix} \lambda - 1 & 0 & 0 \\ -1 & \lambda - 2 & 0 \\ 3 & -5 & \lambda - 2 \end{vmatrix} = (\lambda - 1)(\lambda - 2)^2$$

so the characteristic equation is

$$(\lambda - 1)(\lambda - 2)^2 = 0$$

and the distinct eigenvalues of A are $\lambda = 1$ and $\lambda = 2$. We leave it for you to show that bases for the eigenvalues are

$$\lambda = 1 : \quad \mathbf{p}_1 = \begin{bmatrix} \frac{1}{8} \\ -\frac{1}{8} \\ 1 \end{bmatrix}; \quad \lambda = 2 : \quad \mathbf{p}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Since A is a 3×3 matrix and there are only two basis vectors in total A is not diagonalizable. ■

2.2.1 Eigenvalues of Powers of a Matrix

Suppose that λ is an eigenvalue of A and $\mathbf{x}$ is a corresponding eigenvector. Then

$$A^2\mathbf{x} = A(A\mathbf{x}) = A(\lambda\mathbf{x}) = \lambda(A\mathbf{x}) = \lambda(\lambda\mathbf{x}) = \lambda^2\mathbf{x}$$

which shows not only that λ^2 is a eigenvalue of A^2 but that $\mathbf{x}$ is a corresponding eigenvector. In general, we have the following result.

Theorem 2.2.2 If k is a positive integer, λ is an eigenvalue of a matrix A, and $\mathbf{x}$ is a corresponding eigenvector, then λ^k is an eigenvalue of A^k and $\mathbf{x}$ is a corresponding eigenvector.

■ **Example 2.9** In Example 2.8 we found the eigenvalues and corresponding eigenvectors of the matrix

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ -3 & 5 & 2 \end{bmatrix}$$

Do the same for A^7.
Solution:
We know from Example 2.8 that the eigenvalues of A are $\lambda = 1$ and $\lambda = 2$, so the eigenvalues of A^7 are $\lambda = 1^7 = 1$ and $\lambda = 2^7 = 128$. The eigenvectors $\mathbf{p}_1$ and $\mathbf{p}_2$ obtained in Example 2.8 corresponding to the eigenvalues $\lambda = 1$ and $\lambda = 2$ of A are also the eigenvectors corresponding to the eigenvalues $\lambda = 1$ and $\lambda = 128$ of A^7. ■

Computing Powers of a Matrix;

Suppose that A is a diagonalizable $n \times n$ matrix, that P diagonalizes A, and that

$$P^{-1}AP = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} = D$$

Squaring both sides of this equation yields

$$\left(P^{-1}AP\right)^2 = \begin{bmatrix} \lambda_1^2 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n^2 \end{bmatrix} = D^2$$

We can rewrite the left side of this equation as

$$\left(P^{-1}AP\right)^2 = P^{-1}APP^{-1}AP = P^{-1}AIAP = P^{-1}A^2P$$

from which we obtain the relationship $P^{-1}A^2P = D^2$. More generally, if k is a positive integer, then a similar computation will show that

$$P^{-1}A^kP = D^k = \begin{bmatrix} \lambda_1^k & 0 & \cdots & 0 \\ 0 & \lambda_2^k & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n^k \end{bmatrix}$$

which we can rewrite as

$$A^k = PD^kP^{-1} = P \begin{bmatrix} \lambda_1^k & 0 & \cdots & 0 \\ 0 & \lambda_2^k & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n^k \end{bmatrix} P^{-1}.$$

■ **Example 2.10** Find A^{13}, where

$$A = \begin{bmatrix} 0 & 0 & -2 \\ 1 & 2 & 1 \\ 1 & 0 & 3 \end{bmatrix}.$$

Solution:
Recall Example 2.7 that the matrix A is diagonalized by

$$P = \begin{bmatrix} -1 & 0 & -2 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

and that

$$D = P^{-1}AP = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus, it follows that

$$A^{13} = PD^{13}P^{-1} = \begin{bmatrix} -1 & 0 & -2 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2^{13} & 0 & 0 \\ 0 & 2^{13} & 0 \\ 0 & 0 & 1^{13} \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 \\ 1 & 1 & 1 \\ -1 & 0 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} -8190 & 0 & -16382 \\ 8191 & 8192 & 8191 \\ 8191 & 0 & 16383 \end{bmatrix}.$$

■

2.3 Exercise

A- Show that A and B are not similar matrices

1. $A = \begin{bmatrix} 1 & 1 \\ 3 & 2 \end{bmatrix}$. $B = \begin{bmatrix} 1 & 0 \\ 3 & -2 \end{bmatrix}$

2. $A = \begin{bmatrix} 4 & -1 \\ 2 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 4 & 1 \\ 2 & 4 \end{bmatrix}$

3. $A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 2 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

4. $A = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 0 & 2 \\ 3 & 0 & 3 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix}$

B- Find a matrix P that diagonalizes A, and check your work by computing $P^{-1}AP$.

5. $A = \begin{bmatrix} 1 & 0 \\ 6 & -1 \end{bmatrix}$

6. $A = \begin{bmatrix} -14 & 12 \\ -20 & 17 \end{bmatrix}$.

7. $A = \begin{bmatrix} 2 & 0 & -2 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}$. 8. $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$.

9. Let

$$A = \begin{bmatrix} 4 & 0 & 1 \\ 2 & 3 & 2 \\ 1 & 0 & 4 \end{bmatrix}.$$

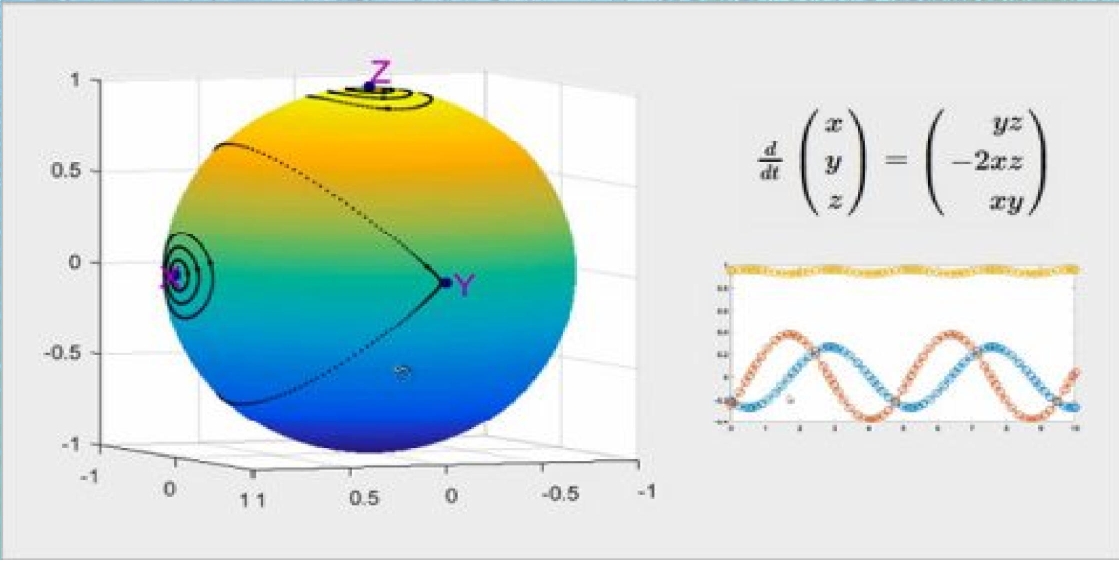
(a) Find the eigenvalues of A.
(b) For each eigenvalue λ, find the rank of the matrix $\lambda I - A$.
(c) Is A diagonalizable? Justify your conclusion.

Part II

Applications of Linear Algebra

Chapte 3

Differential Equations



3. Differential Equations

Many laws of physics, chemistry, biology, engineering, and economics are described in terms of "differential equations"—that is, equations involving functions and their derivatives. In this section we will illustrate one way in which matrix diagonalization can be used to solve systems of differential equations.

Recall from calculus that a differential equation is an equation involving unknown functions and their derivatives. The order of a differential equation is the order of the highest derivative it contains. The simplest differential equations are the first-order equations of the form

$$y' = ay \tag{3.1}$$

where $y = f(x)$ is an unknown differentiable function to be determined, $y' = dy/dx$ is its derivative, and a is a constant. As with most differential equations, this equation has infinitely many solutions; they are the functions of the form

$$y = ce^{ax} \tag{3.2}$$

where c is an arbitrary constant. That every function of this form is a solution of 3.1 follows from the computation

$$y' = cae^{ax} = ay$$

and that these are the only solution is shown in the exercises. Accordingly, we call 3.2 the general solution of 3.1. As an example, the general solution of the differential equation $y' = 5y$ is

$$y = ce^{5x} \tag{3.3}$$

Often, a physical problem that leads to a differential equation imposes some conditions that enable us to isolate one particular solution from the general solution. For example, if we require that solution 3.3 of the equation $y' = 5y$ satisfy the added condition

$$y(0) = 6 \tag{3.4}$$

(that is, $y = 6$ when $x = 0$), then on substituting these values in 3.3, we obtain $6 = ce^0 = c$, from which we conclude that

$$y = 6e^{5x}$$

is the only solution $y' = 5y$ that satisfies 3.4.

A condition such as 3.4, which specifies the value of the general solution at a point, is called an **initial condition**, and the problem of solving a differential equation subject to an initial condition is called an **initial-value problem.**

3.1 First-Order Linear Systems

A systems of differential equations of the form

$$
\begin{aligned}
y_1' &= a_{11}y_1 + a_{12}y_2 + \cdots + a_{1n}y_n \\
y_2' &= a_{21}y_1 + a_{22}y_2 + \cdots + a_{2n}y_n \\
&\vdots \\
y_n' &= a_{n1}y_1 + a_{n2}y_2 + \cdots + a_{nn}y_n
\end{aligned}
\tag{3.5}
$$

where $y_1 = f_1(x), y_2 = f_2(x), \ldots, y_n = f_n(x)$ are functions to be determined, and the a_{ij} s are constants.

By using matrix notation, (3.5) can be written as

$$
\begin{bmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}
$$

or

$$
y' = Ay \tag{3.6}
$$

where the notation y' denotes the vector obtained by differentiating each component of y.

We call (3.5) or its matrix form (3.6) a constant coefficient first-order homogeneous linear system. It is of first order because all derivatives are of that order, it is linear because differentiation and matrix multiplication are linear transformations, and it is homogeneous because

$$
y_1 = y_2 = \cdots = y_n = 0
$$

is a solution regardless of the values of the coefficients As expected, this is called the trivial solution.

■ **Example 3.1** Write the following system in matrix form:

$$
\begin{aligned}
y_1' &= 3y_1 \\
y_2' &= -2y_2 \\
y_3' &= 5y_3
\end{aligned}
$$

Solution:

$$\begin{bmatrix} y_1' \\ y_2' \\ y_3' \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

or

$$y' = \begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 5 \end{bmatrix} \mathrm{y}$$

■

■ **Example 3.2** Solve the system in the above Example (3.1)
Solution:
Since the above system involves only one unknown function, we can solve the equations individually. then the solutions are

$$\begin{aligned} y_1 &= c_1 e^{\mathrm{t}x} \\ y_2 &= c_2 e^{-2x} \\ y_3 &= c_3 e^{\mathrm{sx}} \end{aligned}$$

or, in matrix notation,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} c_1 e^{sx} \\ c_2 e^{-2} \\ c_s e^{4x} \end{bmatrix}$$

■

■ **Example 3.3** Find a solution of the system in the above Example (3.1) that satisfies the initial conditions $y_1(0) = 1$, $y_2(0) = 4$, and $y_3(0) = -2$
Solution:
From the given initial conditions, we obtain

$$1 = y_1(0) = c_1 e^0 = c_1$$

$$4 = y_2(0) = c_2 e^0 = c_2$$

$$-2 = y_3(0) = cse^0 = c_3$$

so the solution satisfying these conditions is

$$y_1 = e^{3x}, \quad y_2 = 4e^{-2x}, \quad y_3 = -2e^{4x}$$

or, in matrix notation,

$$\mathrm{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} e^{3x} \\ 4e^{-2x} \\ -2e^{4x} \end{bmatrix}.$$

■

3.2 Solve First-Order Linear System by Diagonalization

The basic idea for solving a system

$$y' = Ay$$

whose coefficient matrix A is not diagonal is to introduce a new unknown vector u that is related to the unknown vector y by an equation of the form

$$y = Pu$$

in which P is an invertible matrix that diagonalizes A. Of course, such a matrix may or may not exist, but if it does, then we can rewrite the equation

$$y' = Ay$$

as

$$P\mathrm{u}' = A(P\mathrm{u})$$

or alternatively as

$$\mathbf{u}' = \left(P^{-1}AP\right)\mathbf{u}$$

Since P is assumed to diagonalizes A, this equation has the form

$$\mathbf{u}' = D\mathbf{u}$$

where D is diagonal. We can now solve this equation for u using the method of Example (3.1), and then obtain y by matrix multiplication using the relationship $y = P\mathrm{u}$.

In summary, we have the following procedure for solving a system $y' = Ay$ in the case were A is diagonalizable.

A Procedure for Solving $y' = Ay$ If A Is Diaganalizable

Step 1. Find a matrix P that dizgonalizes A.
Step 2. Make the substitutions $\mathrm{y} = P\mathrm{u}$ and $y' = P\mathrm{u}'$ to obtain a new "diagonal system"

$$\mathrm{u}' = D\mathrm{u},$$

where $D = P^{-1}AP$.
Step 3. Solve $u' = D\mathbf{u}$.
Step 4. Determine y from the equation $\mathrm{y} = P\mathrm{u}$.

■ **Example 3.4** Let a system

$$\begin{aligned} y_1' &= y_1 + y_2 \\ y_2' &= 4y_1 - 2y_2 \end{aligned}$$

then
(a) Solve this system.
(b) Find the solution that satisfies the initial conditions $y_1(0) = 1, y_2(0) = 6$.
Solution:
(a) The coefficient matrix for the system is

$$A = \begin{bmatrix} 1 & 1 \\ 4 & -2 \end{bmatrix}.$$

Since A will be diagonalized by any matrix P whose columns are linearly independent eigenvectors of A.
Now

$$\begin{aligned} \det(\lambda I - A) &= \begin{vmatrix} \lambda - 1 & -1 \\ -4 & \lambda + 2 \end{vmatrix} \\ &= \lambda^2 + \lambda - 6 \\ &= (\lambda + 3)(\lambda - 2), \end{aligned}$$

the eigenvalues of A are $\lambda = 2$ and $\lambda = -3$.
By definition,

$$\mathrm{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

is an eigenvector of A corresponding to λ if and only if x is a nontrivial solution of

$$\begin{bmatrix} \lambda-1 & -1 \\ -4 & \lambda+2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

If $\lambda = 2$, this system becomes

$$\begin{bmatrix} 1 & -1 \\ -4 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Solving this system yields $x_1 = t, x_2 = t$, so

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} t \\ t \end{bmatrix} = t \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Thus,

$$P_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

is a basis for the eigenvector corresponding to $\lambda = 2$.
Similarly, you can show that

$$\mathrm{P}_2 = \begin{bmatrix} -\frac{1}{4} \\ 1 \end{bmatrix}$$

is a basis for the eigenvector corresponding to $\lambda = -3$. Thus,

$$P = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

diagonalizes A, and

$$D = P^{-1}AP = \begin{bmatrix} 2 & 0 \\ 0 & -3 \end{bmatrix}$$

Thas, as noted in Step 2 of the procedure stated above, the substitution

$$y = Pu \text{ and } y' = P'_{\mathrm{u}}$$

yields the diagonal system"

$$\begin{aligned} u' &= Du \\ &= \begin{bmatrix} 2 & 0 \\ 0 & -3 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \end{aligned}$$

or

$$\begin{aligned} u_1' &= 2u_1 \\ u_2' &= -3u_2 \end{aligned}$$

and the solution of this system is

$$\begin{aligned} u_1 &= c_1 e^{2x} \\ u_2 &= c_2 e^{-1a} \end{aligned} \quad \text{or } u = \begin{bmatrix} c_1 e^{2x} \\ c_2 e^{-3x} \end{bmatrix}$$

so the equation $\mathbf{y} = P\mathbf{u}$ yields, as the solution for $\mathbf{y}$,

$$\begin{aligned} \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} &= \begin{bmatrix} 1 & -\frac{1}{4} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} c_1 e^{2x} \\ c_2 e^{-3x} \end{bmatrix} \\ &= \begin{bmatrix} c_1 e^{2x} - \frac{1}{4} c_2 e^{-3x} \\ c_1 e^{2x} + c_2 e^{-3x} \end{bmatrix} \end{aligned}$$

or

$$\begin{aligned} y_1 &= c_1 e^{2x} - \tfrac{1}{4} c_2 e^{-3x} \\ y_2 &= c_1 e^{2x} + c_2 e^{-3x} \end{aligned}$$

(b) If we substitute the given initial conditions in the above system, we obtain

$$\begin{aligned} c_1 - \tfrac{1}{4} c_2 &= 1 \\ c_1 + c_2 &= 6 \end{aligned}$$

Solving this system, we obtain $c_1 = 2, c_2 = 4$, so the solution with the initial conditions is

$$\begin{aligned} y_1 &= 2e^{2x} - e^{-3x} \\ y_2 &= 2e^{2x} + 4e^{-3x} \end{aligned}$$

■

3.3 Exercises

1. (a) Solve the system

$$\begin{aligned} y_1' &= y_1 + 4y_2 \\ y_2' &= 2y_1 + 3y_2 \end{aligned}$$

(b) Find the solution that satisfies the initial conditions $y_1(0) = 0, y_2(0) = 0$

2. (a) Solve the system

$$\begin{aligned} y_1' &= y_1 + 3y_2 \\ y_2' &= 4y_1 + 5y_2 \end{aligned}$$

(b) Find the solution that satisfies the conditions $y_1(0) = 2$, $y_2(0) = 1$

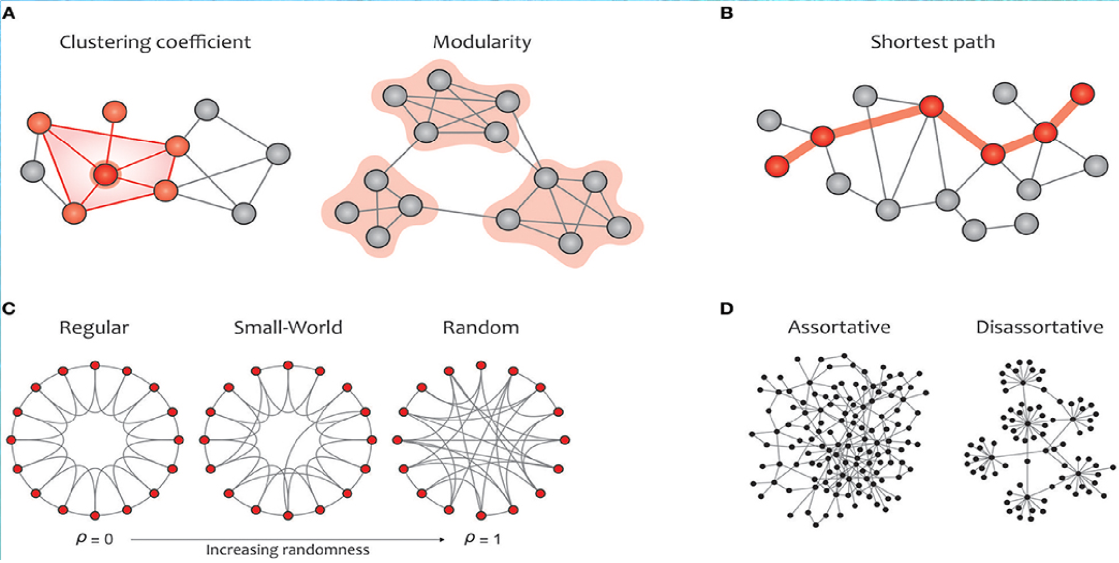
3. (a) Solve the system

$$\begin{aligned} y_1' &= 4y_1 \qquad + y_3 \\ y_2' &= -2y_1 + y_2 \\ y_3' &= -2y_1 \qquad + y_3 \end{aligned}$$

(b) Find the solution that satisfies the initial conditions $y_1(0) = -1, y_2(0) = 1, y_3(0) = 0$

Chapte 4

Graph Theory



4. Graph Theory

4.1 Directed Graphs

Definition 4.1.1 A **directed graph** is a finite set of elements, $\{P_1, P_2, ..., P_n\}$, together with a finite collection of ordered pairs (P_i, P_j) of distinct elements of this set, with no ordered pair being repeated. The elements of the set are called **vertices**, and the ordered pairs are called **directed edges**, of the directed graph.

We use the notation $P_i \to P_j$ (which is read "P_i is connected to P_j") to indicate that the directed edge (P_i, P_j) belongs to the directed graph. Geometrically, we can visualize a directed graph by representing the vertices as points in the plane and representing the directed edge $P_i \to P_j$ by drawing a line or arc from vertex P_i to vertex P_j, with an arrow pointing from P_i to P_j. If both $P_i \to P_j$ and $P_j \to P_i$ hold (denoted $P_i \leftrightarrow P_j$), we draw a single line between P_i and P_j with two oppositely pointing arrows. See Figure (4.1)

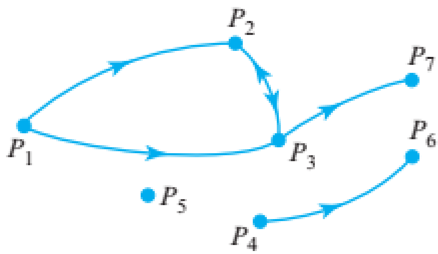


Figure 4.1:

With a directed graph having n vertices, we may associate an $n \times n$ matrix $M = [m_{ij}]$, called **the vertex matrix** of the directed graph. Its elements are defined by

$$m_{ij} = \begin{cases} 1, & \text{if } P_i \to P_j \\ 0, & \text{otherwise} \end{cases}$$

R Vertex matrices have the following two properties:
(i) All entries are either 0 or 1.
(ii) All diagonal entries are 0.
Conversely, any matrix with these two properties determines a unique directed graph having the given matrix as its vertex matrix.

■ **Example 4.1** Find the corresponding vertex matrices for directed graphs in Figure (4.2)

Solution:

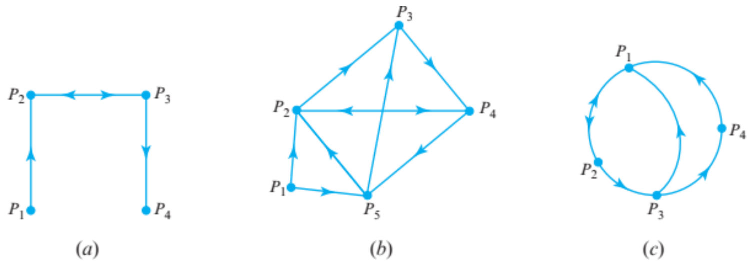


Figure 4.2:

(a)

$$M = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

(b)

$$M = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

(c)

$$M = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

■

■ **Example 4.2** Let M be the vertex matrix define as follow

$$M = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Find the corresponding directed graphs for M.
Solution:
the corresponding directed graphs for M explain in Figure (4.3). ■

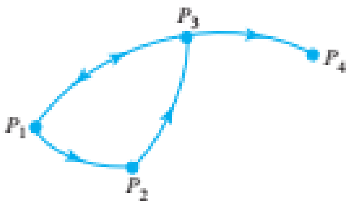


Figure 4.3:

■ **Example 4.3** A certain family consists of a mother, father, daughter, and two sons. The family members have influence, or power, over each other in the following ways: the mother can influence the daughter and the oldest son; the father can influence the two sons; the daughter can influence the father; the oldest son can influence the youngest son; and the youngest son can influence the mother. Find the directed graph and vertex matrix of this model.
Solution:
Figure (4.4) is the resulting directed graph, where we have used obvious letter designations for the five family members.

The vertex matrix of this directed graph is

$$\begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

■

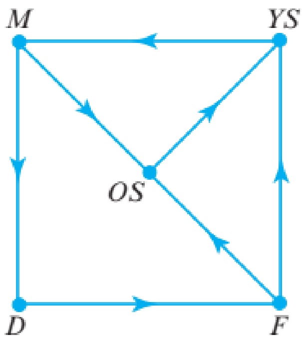


Figure 4.4:

In Example (4.3) the father cannot directly influence the mother; that is, $F \to M$ is not true. But he can influence the youngest son, who can then influence the mother. We write this as $F \to YS \to M$ and call it a 2-step connection from F to M. Analogously, we call $M \to D$ a 1-step connection, $F \to OS \to YS \to M$ a 3-step connection, and so forth.

Now we show a technique for finding the number of all possible r-step connections $(r = 1, 2, \ldots)$ from one vertex P_i to another vertex P_j of an arbitrary directed graph.

The number of 1-step connections from P_i to P_j is simply m_{ij}. That is, there is either zero or one 1-step connection from P_i to P_j, depending on whether m_{ij} is zero or one. For the number of 2 -step connections, we consider the square of the vertex matrix. If we let $m_{ij}^{(2)}$ be the (i, j) -th element of M^2, we have

$$m_{ij}^{(2)} = m_{i1}m_{1j} + m_{i2}m_{2j} + \cdots + m_{in}m_{nj}. \tag{4.1}$$

If $m_{i1} = m_{1j} = 1$, there is a 2 -step connection $P_i \to P_1 \to P_j$ from P_i to P_j. But if either m_{i1} or m_{1j} is zero, such a 2 -step connection is not possible. Thus $P_i \to P_1 \to P_j$ is a 2 -step connection if and only if $m_{i1}m_{1j} =$

1. Similarly, for any $k = 1, 2, \ldots, n$, $P_i \to P_k \to P_j$ is a 2-step connection from P_i to P_j if and only if the term $m_{ik}m_{kj}$ on the right side of (4.1) is one; otherwise, the term is zero. Thus, the right side of (4.1) is the total number of two 2 -step connections from P_i to P_j.

In general, we have the following result.

Theorem 4.1.1 Let M be the vertex matrix of a directed graph and let $m_{ij}^{(r)}$ be the (i,j) -th element of M^r. Then $m_{ij}^{(r)}$ is equal to the number of r -step connections from P_i to P_j.

■ **Example 4.4** Figure 4.5 is the route map of a small airline that services the four cities P_1, P_2, P_3, P_4. Find a vertex matrix and r -step connections from P_4 to P_3.

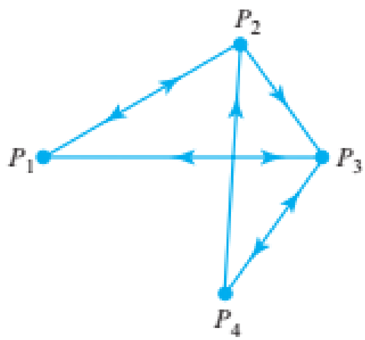


Figure 4.5:

Solution:

As a directed graph, its vertex matrix is

$$M = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

We have that

$$M^2 = \begin{bmatrix} 2 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 2 & 2 & 0 \\ 2 & 0 & 1 & 1 \end{bmatrix},$$

and

$$M^3 = \begin{bmatrix} 1 & 3 & 3 & 1 \\ 2 & 2 & 3 & 1 \\ 4 & 0 & 2 & 2 \\ 1 & 3 & 3 & 1 \end{bmatrix}$$

If we are interested in connections from city P_4 to city P_3, by using Theorem 4.1.1 we find their number.
since $m_{43} = 1$, there is one 1-step connection; because $m_{43}^{(2)} = 1$, there is one 2-step connection; and because $m_{43}^{(3)} = 3$, there are three 3-step connections.
Now we verify this, from Figure 3.5 we find
1-step connections from P_4 to $P_3 : P_4 \to P_3$.
2-step connections from P_4 to $P_3 : P_4 \to P_2 \to P_3$.
3-step connections from P_4 to P_3 :

$$P_4 \to P_3 \to P_4 \to P_3.$$
$$P_4 \to P_2 \to P_1. \to P_3$$
$$P_4 \to P_3 \to P_1. \to P_3.$$

■

4.2 cliques

Definition 4.2.1 A subset of a directed graph is called a clique if it satisfies the following three axioms:
(i) The subset contains at least three vertex.
(ii) For each pair of vertices P_i and P_j in the subset, both $P_i \rightarrow P_j$ and $P_j \rightarrow P_i$ are true.
(iii) The subset is as large as possible; that is, it is not possible to add another vertex the subset and still satisfy condition (ii).

■ **Example 4.5** The directed graph illustrated in Figure 4.6 which might represent the route map of an airline. Find the set of cliques.

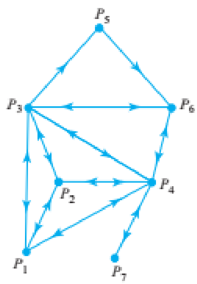


Figure 4.6:

Solution:
The directed graph has two cliques:

$$\{P_1, P_2, P_3, P_4\},$$

and

$$\{P_3, P_4, P_6\}.$$

■

Cliques can be identified by inspecting simple directed graphs. However, a systematic method for detecting cliques in large directed graphs would be ideal. For this reason, t it will be helpful to define a matrix

$$S = [s_{ij}],$$

related to a given directed graph as follows:

$$s_{ij} = \begin{cases} 1, & \text{if } P_i \leftrightarrow P_j \\ 0, & \text{otherwise} \end{cases}$$

The matrix S determines a directed graph that is the same as the given directed graph, with the exception that the directed edges with only one arrow are deleted.

The matrix S may be obtained from the vertex matrix M of the original directed graph by setting $s_{ij} = 1$ if $m_{ij} = m_{ji} = 1$ and setting $s_{ij} = 0$ otherwise.

The following theorem, which uses the matrix S, is helpful for identifying cliques.

Theorem 4.2.1 Let $s_{ij}^{(3)}$ be the (i, j)-th element of S^3. Then a vertex P_i belongs to some clique if and only if $s_{ii}^{(3)} \neq 0$

Proof. If $s_{ii}^{(3)} \neq 0$, then there is at least one 3-step connection from P_i to itself in the modified directed graph determined by S. Suppose it is $P_i \to P_j \to P_k \to P_i$. In the modified directed graph, all directed relations are two-way, so we also have the connections $P_i \leftrightarrow P_j \leftrightarrow P_k \leftrightarrow P_i$. But this means that $\{P_i, P_j, P_k\}$ is either a clique or a subset of a clique. In either case, P_i must belong to some clique. The converse statement, "if P_i belongs to a clique, then $s_{ii}^{(3)} \neq 0$," follows in a similar manner. ■

■ **Example 4.6** Suppose that a directed graph has as its vertex matrix

$$M = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Find S and show that the directed graph has not clique.
<u>Solution:</u>

$$S = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

and

$$S^3 = \begin{bmatrix} 0 & 3 & 0 & 2 \\ 3 & 0 & 2 & 0 \\ 0 & 2 & 0 & 1 \\ 2 & 0 & 1 & 0 \end{bmatrix}.$$

Because all diagonal entries of S^3 are zero, it follows from Theorem 4.2.1 that the directed graph has no cliques. ■

■ **Example 4.7** Suppose that a directed graph has as its vertex matrix

$$M = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Find S and show that the directed graph has not clique.
<u>Solution:</u>

$$S = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix},$$

and

$$S^3 = \begin{bmatrix} 2 & 4 & 0 & 4 & 3 \\ 4 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 4 & 3 & 0 & 2 & 1 \\ 3 & 1 & 0 & 1 & 0 \end{bmatrix}.$$

The nonzero diagonal entries of S^3 are $s_{11}^{(3)}, s_{22}^{(3)}$, and $s_{44}^{(3)}$. Consequently, in the given directed graph, P_1, P_2, and P_4 belong to cliques. Because a clique must contain at least three vertices, the directed graph has only one clique, $\{P_1, P_2, P_4\}$. ∎

4.3 Dominance-Directed Graphs

Definition 4.3.1 A dominance-directed graph is a directed graph such that for any distinct pair of vertices P_i and P_j, either $P_i \to P_j$ or $P_j \to P_i$, but not both.

An example of a directed graph satisfying this definition is a league of n sports teams that play each other exactly one time, as in one round of a round-robin tournament in which no ties are allowed. If $P_i \to P_j$ means that team P_i beat team P_j in their single match, it is easy to see that the definition of a dominance-directed group is satisfied. For this reason, dominance-directed graphs are sometimes called **tournaments**.

Theorem 4.3.1 In any dominance-directed graph, there is at least one vertex from which there is a 1-step or 2 -step connection to any other vertex.

Proof. Suppose that a vertex (there may be several) with the largest total number of 1-step and 2-step connections to other vertices in the graph. By renumbering the vertices, we may assume that P_1 is such a vertex. Suppose there is some vertex P_i such that there is no 1-step or 2-step connection from P_1 to P_i. Then, in particular, $P_1 \to P_i$ is not true, so that by definition of a dominance-directed graph, it must be that $P_i \to P_1$. Next, let P_k be any vertex such that $P_1 \to P_k$ is true. Then we cannot have $P_k \to P_i$, as then $P_1 \to P_k \to P_i$ would be a 2-step connection from P_1 to P_i. Thus, it must be that $P_i \to P_k$. That is, P_i has 1-step connections to all the vertices to which P_1 has 1-step connections. The vertex P_i must then also have 2-step connections to all the vertices to which P_1 has 2-step connections.
But because, in addition, we have that $P_i \to P_1$, this means that P_i has more l-step and 2-step connections to other vertices than does P_1. However, this contradicts the way in which P_1 was chosen. Hence, there can be no vertex P_i to which P_1 has no 1-step or 2-step connection. ■

The sum of the entries in the ith row of M is the total number of 1-step connections from P_i to other vertices, and the sum of the entries of the i th row of M^2 is the total number of 2 -step connections from P_i to other vertices. Consequently, the sum of the entries of the i th row of the matrix

$$A = M + M^2,$$

is the total number of 1-step and 2-step connections from P_i to other vertices. In other words, a row of

$$A = M + M^2,$$

with the largest row sum identifies a vertex having the property stated in Theorem 4.3.1

■ **Example 4.8** Suppose that five baseball teams play each other exactly once, and the results are as indicated in the dominance-directed graph of Figure 4.7. Find M and A of the graph.

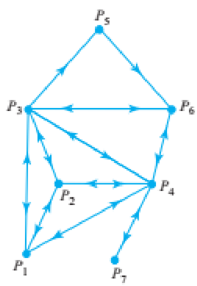


Figure 4.7:

Solution:

The vertex matrix of the graph is

$$M = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix},$$

so

$$A = M + M^2 = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 2 & 3 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 2 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 1 & 1 & 2 & 0 \\ 2 & 0 & 3 & 3 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 2 & 3 & 0 \end{bmatrix}.$$

Then the row sums of A are

$$\begin{aligned} \text{1st row sum } &= 4 \\ \text{2nd row sum } &= 9 \\ \text{3rd row sum } &= 2 \\ \text{4th row sum } &= 4 \\ \text{5th row sum } &= 7 \end{aligned}$$

Since the second row has the largest row sum, the vertex P_2 must have a 1-step or 2-step connection to any other vertex. This is easily verified from Figure 3.7. ■

Definition 4.3.2 The **power** of a vertex of a dominance-directed graph is the total number of 1-step and 2-step connections from it to other vertices. Alternatively, the power of a vertex P_i is the sum of the entries of the i th row of the matrix $A = M + M^2$, where M is the vertex matrix of the directed graph.

Example 4.9 Let us rank the five baseball teams in Example 4.8 according to their powers. From the calculations for the row sums in that example, we have

$$\text{Power of team } P_1 = 4.$$
$$\text{Power of team } P_2 = 9.$$
$$\text{Power of team } P_3 = 2.$$
$$\text{Power of team } P_4 = 4.$$
$$\text{Power of team } P_5 = 7$$

Thus, the ranking of the teams according to their powers would be

First P_2.
Second P_5.
Third P_1.
And tied for thrird P_4.
Last P_3.

■

4.4 Exercises

A- Draw a diagram of the directed graph corresponding to each of the following vertex matrices.

$$(a) \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

$$(b) \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

B- Five baseball teams play each other one time with the following results:

A beats B, C, D
B beats C, E
C beats D, E
D beats B
E beats A, D

Rank the five baseball teams in accordance with the powers of the vertices they correspond to in the dominance-directed graph representing the outcomes of the games.

C- For the dominance-directed graph illustrated in Figure 4.8 construct the vertex matrix and find the power of each vertex

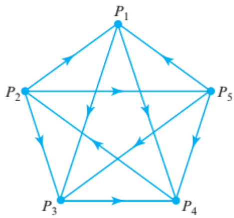
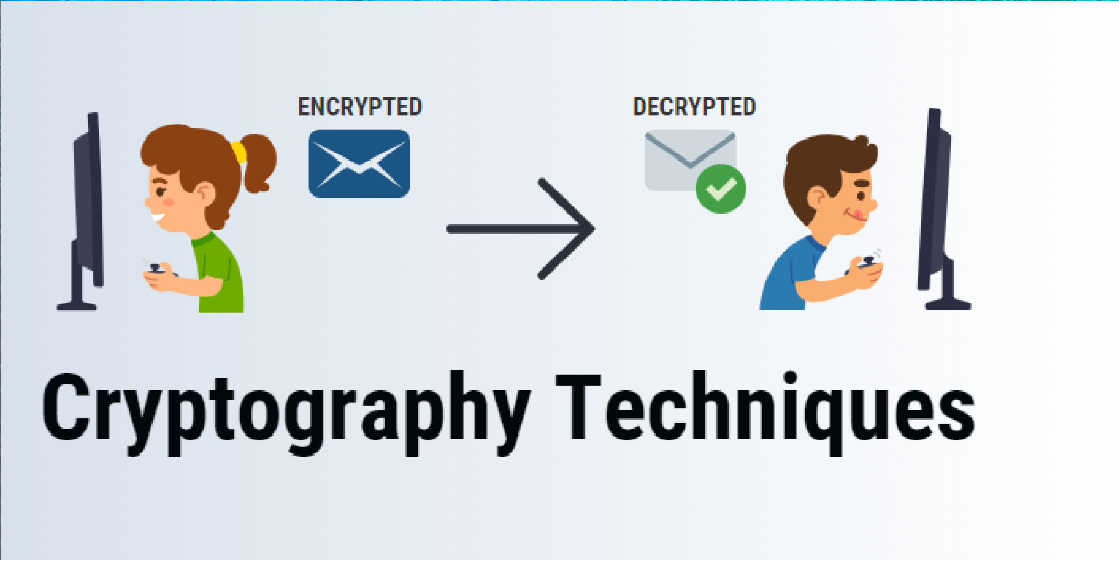


Figure 4.8:

Chapte 5

Cryptography



5. Cryptography

In this chapter, we introduce a method of encoding and decoding messages. Also, we examine modular arithmetic and explaine how Gaussian elimination can sometimes be used to break an opponent's code.

5.1 Ciphers

Secret codes date to the earliest days of written communication, there has been a recent surge of interest in the subject because of the need to maintain the privacy of information transmitted over public lines of communication.

Definition 5.1.1 **Cryptography** is study encoding and decoding of secret messages.

R In the language of cryptography:
(i) Codes are called **ciphers**.
(ii) Uncoded messages are called **plaintext**.
(iii) Coded messages are called **ciphertext**.
(iv) The process of converting from plaintext to ciphertext is called **enciphering**.
(v) The reverse process of converting from ciphertext to plaintext is called deciphering.

> The simplest ciphers, called substitution ciphers, are those that replace each letter of the alphabet by a different letter.

For example, in the substitution cipher

Plain	A	B	C	D	E	F	J	H	I	J	K	L	M
Cipher	D	E	F	G	H	I	J	K	L	M	N	O	P
Plain	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Cipher	Q	R	S	T	U	V	W	X	Y	Z	A	B	C

the plaintext letter A is replaced by D, the plaintext letter B by E, and so forth. With this cipher the plaintext message

ROME WAS NOT BUILT IN A DAY

becomes

URPH ZDV QRW EXLOW LQ D GDB

5.2 Hill Ciphers

In this section we will study a class of polygraphic systems based on matrix transformations.
Now, we assume that each plaintext and ciphertext letter except Z is assigned the numerical value that specifies its position in the standard alphabet (Table 5.1). For reasons that will become clear later, Z is assigned a value of zero.

Plain	A	B	C	D	E	F	J	H	I	J	K	L	M
Cipher	1	2	3	4	5	6	7	8	9	10	11	12	13
Plain	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Cipher	14	15	16	17	18	19	20	21	22	23	24	25	0

Table 5.1:

Whenever an integer greater than 25 occurs, it will be replaced by the remainder that results when this integer is divided by 26. Because the remainder after division by 26 is one of the integers $0, 1, 2, \ldots, 25$, this procedure will always yield an integer with an alphabet equivalent.

In the simplest Hill ciphers, successive pairs of plaintext are transformed into ciphertext by the following procedure:

Step 1. Choose a 2×2 matrix with integer entries

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

to perform the encoding. Certain additional conditions on A will be imposed later.
Step 2. Group successive plaintext letters into pairs, adding an arbitrary "dummy" letter to fill out the last pair if the plaintext has no odd number of letters, and replace each plaintext letter by its numerical value.
Step 3. Successively convert each plaintext pair

$$P = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$$

into a column vector and form the product AP. We will call P a plaintext vector and AP the corresponding ciphertext vector.
Step 4. Convert each ciphertext vector into its alphabetic equivalent.

■ **Example 5.1** Use the matrix

$$\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix},$$

to obtain the Hill cipher for the plaintext message

I AM HIDING

Solution:
If we group the plaintext into pairs and add the dummy letter G to fill out the last pair, we obtain

$$IA \quad MH \quad ID \quad IN \quad GG$$

Form Table 5.1, we find

$$9 \quad 1 \quad 13 \quad 8 \quad 9 \quad 4 \quad 9 \quad 14 \quad 7 \quad 7$$

To encipher the pair IA, we form the matrix product

$$\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 9 \\ 1 \end{bmatrix} = \begin{bmatrix} 11 \\ 3 \end{bmatrix}.$$

From Table 5.1 , yields the ciphertext KC. To encipher the pair MH, we form the product

$$\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 13 \\ 8 \end{bmatrix} = \begin{bmatrix} 29 \\ 24 \end{bmatrix}$$

However, there is a problem here, because the number 29 has no alphabet equivalent (Table 5.1). To resolve this problem, we use the above remark. Thus, we replace 29 by 3, which is the remainder after dividing 29 by 26. It now follows from Table 5.1 that the ciphertext for the pair MH is CX. The computations for the remaining ciphertext vectors are

$$\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 9 \\ 4 \end{bmatrix} = \begin{bmatrix} 17 \\ 12 \end{bmatrix}$$
$$\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 9 \\ 14 \end{bmatrix} = \begin{bmatrix} 37 \\ 42 \end{bmatrix} \text{ or } \begin{bmatrix} 11 \\ 16 \end{bmatrix}$$
$$\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 7 \\ 7 \end{bmatrix} = \begin{bmatrix} 21 \\ 21 \end{bmatrix}$$

These correspond to the ciphertext pairs QL, KP, and UU, respectively. Therefore, the entire ciphertext message is

$$KC \quad CX \quad QL \quad KP \quad UU$$

which, in most situations, will be sent as a single string with no spaces:

KCCXQLKPUU

■

R

- The Hill cipher in Example 5.1 is referred to as a Hill 2-cipher because the plaintext was grouped in pairs and enciphered by a 2×2 matrix.

- The plaintext can also be divided into triples and enciphered using a 3×3 matrix of integer entries, which is known as a Hill 3-cipher.
- In general, for a Hill n-cipher, plaintext is grouped into sets of n letters and enciphered by an $n \times n$ matrix with integer entries.

5.3 Modular Arithmetic

A positive integer m is called the modulus in modular arithmetic, and any two integers whose difference is an integer multiple of the modulus are treated as "equal" or "equivalent" with respect to the modulus. To be more specific, we include the following definition.

Definition 5.3.1 If m is a positive integer and a and b are any integers, then we say that a is equivalent to b modulo m, written

$$a = b \pmod{m},$$

if $a - b$ is an integer multiple of m.

■ **Example 5.2**

$$\begin{aligned} 7 &= 2 \pmod{5} \\ 19 &= 3 \pmod{2} \\ -1 &= 25 \pmod{26} \\ 12 &= 0 \pmod{4} \end{aligned}$$

■

For any modulus m it ean be proved that every integer a is equivalent, modulo m, to exactly one of the integers

$$0, 1, 2, \ldots, m-1$$

We call this integer the residue of a modulo m, and we write

$$Z_n = \{0, 1, 2, \ldots, m-1\}$$

to denote the set of residues modulo m. If a is a non-negative integer, then its residue modulo m is simply the remainder that results when a is divided by m.
The residue can be found using the following theorem for any integer a.

Theorem 5.3.1 For any integer a and modulus m, let

$$R = \text{ remainder of } \frac{|a|}{m}$$

Then the residue r of a modulo m is given by

$$r = \begin{cases} R & \text{if } a \geq 0 \\ m - R & \text{if } a < 0 \text{ and } R \neq 0 \\ 0 & \text{if } a < 0 \text{ and } R = 0 \end{cases}$$

■ **Example 5.3** Find the residue modulo 26 of
(a) 87,
$(b) - 38,$
$(c) - 26.$
Solution:
(a) Dividing $|87| = 87$ by 26 yields a remainder of $R = 9$, so $r = 9$. Thus,

$$87 = 9 \pmod{26}.$$

(b) Dividing $|-38| = 38$ by 26 yields a remainder of $R = 12$, so $r = 26 - 12 = 14$. Thus

$$-38 = 14 (\bmod 26).$$

(c) Dividing $|-26| = 26$ by 26 yields a remainder of $R = 0$. Thus,

$$-26 = 0 \pmod{26}.$$

■

The next definition explain the multiplicative inverse.

Definition 5.3.2 If a is a number in Z_{m} , then number Z_m is called a reciprocal or multiplicative inverse of a modulo m if $aa^{-1} = a^{-1}a = 1 \pmod{m}$.

R It can be proved that if a and m have no common prime factors, then a has a unique reciprocal modulo m; conversely, if a and m have a common prime factor, then a has no reciprocal modulo m.

■ **Example 5.4** The number 3 has multiplicative inverse in modulo 26 because 3 and 26 have no common prime factors. This multiplicative inverse can be obtained by finding the number x in Z_{26} that satisfies the modular equation

$$3x = 1 \pmod{26}.$$

Although there are general methods for solving such modular equations, it would take us too far afield to study them. However, because 26 is relatively small, this equation can be solved by trying the possible solutions, 0 to 25, one at a time. With this approach we find that $x = 9$ is the solution, because

$$3 \cdot 9 = 27 = 1 \pmod{26}.$$

Thus,

$$3^{-1} = 9 \quad (\text{mod.}$$

■

■ **Example 5.5** The number 4 has no multiplicative inverse in modulo 26, because 4 and 26 have 2 as a common prime factor ■

For future reference, in Table 5.2 we provide the following multiplicative inverse in modulo 26:

a	1	3	5	7	9	11	15	17	19	21	23	25
a^{-1}	1	9	21	15	3	19	7	23	11	5	17	25

Table 5.2: multiplicative inverse in Modulo 26

5.4 Deciphering

Every useful cipher must have a procedure for decipherment. In the case of a Hill cipher, decipherment uses the inverse (mod 26) of the enciphering matrix. To be precise, if m is a positive integer, then a square matrix A with entries in Z_m is said to be invertible modulo m if there is a matrix B with entries in Z_m such that

$$AB = BA = I \pmod{m}.$$

Assume that

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

is invertible modulo 26 and this matrix is used in a Hill 2-cipher. If

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix},$$

is a plaintext vector, then

$$c = Ap \pmod{26},$$

is the corresponding ciphertext vector and

$$\mathbf{p} = A^{-1}\mathbf{c} \pmod{26}.$$

Thus, each plaintext vector can be recovered from the corresponding ciphertext vector by multiplying it on the left by $A^{-1} \pmod{26}$.

It's crucial to know which matrices are modulo 26 invertible and how to get their inverses in cryptography. We're now looking into these questions.

In ordinary arithmetic, a square matrix A is invertible if and only if $\det(A) \neq 0$.
Now, the following theorem is the analog of this result in modular arithmetic.

Theorem 5.4.1 A square matrix A with entries in Z_m is invertible modulo m if and only if the residue of det(A) modulo m has a multiplicative inverse in modulo m.

Since the residue of $\det(A)$ modulo m will have a multiplicative inverse in modulo m if and only if this residue and m have no common prime factors, then we have the following corollary.

Corollary 5.4.2 A square matrix A with entries in Z_m is invertible modulo m if and only if m and the residue of det(A) modulo m have no common prime factors.

The following corollary is useful in cryptography since the only prime factors of $m = 26$ are 2 and 13.

Corollary 5.4.3 A square matrix A with entries in Z_{26} is invertible modulo 26 if and only if the residue of $\det(A)$ modulo 26 is not divisible by 2 or 13.

It easy to verify that if

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

has entries in Z_{26} and the residue of $det(A) = ad - bc$ modulo 26 is not divisible by 2 or 13, then the inverse of A (mod 26) is given by

$$A^{-1} = (ad - bc)^{-1} \begin{bmatrix} a & b \\ c & d \end{bmatrix} (mod\ 26), \tag{5.1}$$

where $(ad - bc)^{-1}$ is the inverse of the residue of $ad - bc$ (mod 26).

■ **Example 5.6** Find the inverse of

$$A = \begin{bmatrix} 5 & 6 \\ 2 & 3 \end{bmatrix},$$

modulo 26.
Solution:

$$\det(A) = ad - bc = 5 \cdot 3 - 6 \cdot 2 = 3$$

so from Table 5.2.

$$(ad - bc)^{-1} = 3^{-1} = 9(\bmod 26)$$

Thus, from (5.1),

$$\begin{aligned} A^{-1} &= 9\begin{bmatrix} 3 & -6 \\ -2 & 5 \end{bmatrix} \\ &= \begin{bmatrix} 27 & -54 \\ -18 & 45 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 24 \\ 8 & 19 \end{bmatrix} \quad (\bmod 26). \end{aligned}$$

As a check,

$$\begin{aligned} AA^{-1} &= \begin{bmatrix} 5 & 6 \\ 2 & 3 \end{bmatrix}\begin{bmatrix} 1 & 24 \\ 8 & 19 \end{bmatrix} \\ &= \begin{bmatrix} 53 & 234 \\ 26 & 105 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (\text{mod}26). \end{aligned}$$

Similarly, $A^{-1}A = I$. ■

■ **Example 5.7** Decode the following Hill 2 -cipher, which was enciphered by the matrix in Example 5.6

GTNKGKDUSK

Solution: From Table 5.1 the numerical equivalent of this ciphertext is

7 20 14 11 7 11 4 21 19 11

To obtain the plaintext pairs, we multiply each ciphertext vector by the inverse of (obtained in Example 5.6):

$$\begin{bmatrix} 1 & 24 \\ 8 & 19 \end{bmatrix} \begin{bmatrix} 7 \\ 20 \end{bmatrix} = \begin{bmatrix} 487 \\ 436 \end{bmatrix} = \begin{bmatrix} 19 \\ 20 \end{bmatrix} \pmod{26}$$
$$\begin{bmatrix} 1 & 24 \\ 8 & 19 \end{bmatrix} \begin{bmatrix} 14 \\ 11 \end{bmatrix} = \begin{bmatrix} 278 \\ 321 \end{bmatrix} = \begin{bmatrix} 18 \\ 9 \end{bmatrix} \pmod{26}$$
$$\begin{bmatrix} 1 & 24 \\ 8 & 19 \end{bmatrix} \begin{bmatrix} 7 \\ 11 \end{bmatrix} = \begin{bmatrix} 271 \\ 265 \end{bmatrix} = \begin{bmatrix} 11 \\ 5 \end{bmatrix} \pmod{26}$$
$$\begin{bmatrix} 1 & 24 \\ 8 & 19 \end{bmatrix} \begin{bmatrix} 4 \\ 21 \end{bmatrix} = \begin{bmatrix} 508 \\ 431 \end{bmatrix} = \begin{bmatrix} 14 \\ 15 \end{bmatrix} \pmod{26}$$
$$\begin{bmatrix} 1 & 24 \\ 8 & 19 \end{bmatrix} \begin{bmatrix} 19 \\ 11 \end{bmatrix} = \begin{bmatrix} 283 \\ 361 \end{bmatrix} = \begin{bmatrix} 23 \\ 23 \end{bmatrix} \pmod{26}$$

Figure 5.1:

From Table 5.1, the alphabet equivalents of these vectors are

$$ST \quad RI \quad KE \quad NO \quad WW$$

which yields the message

STRIKE NOW

■

5.5 Breaking a Hill Cipher

Cryptographers are concerned with the security of their ciphers—that is, how easily they can be broken—because the aim of encrypting messages and information is to prevent "opponents" from knowing their contents (deciphered by their opponents). We'll wrap up this section with a look at one approach for cracking Hill ciphers.

Assume you can able to obtain any associated plaintext and ciphertext from an opponent's message. Examining any intercepted ciphertext, for example, you may be able to deduce that the message is a letter that starts with DEAR SIR. We'll demonstrate how, given a small amount of such data, it's possible to deduce a Hill code's deciphering matrix and thus gain access to the rest of the message.

The fact that a linear transformation is absolutely determined by its values at a basis is a fundamental result in linear algebra. According to this theory, if we have a Hill n-cipher, and if

$$P_1 \quad P_2 \quad \dots \quad P_n$$

are linearly independent plaintext vectors whose corresponding ciphertext vectors

$$AP_1 \quad AP_2 \quad \dots \quad AP_n$$

are known, then there is enough information available to determine the matrix A and hence A^{-1} (mod m).

The next theorem tells us that to find the transpose of the deciphering matrix A^{-1}, we must find a sequence of row operations that reduces C to I and then perform this same sequence of operations on P.

Theorem 5.5.1 Let $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n$ be linearly independent plaintext vectors, and let $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n$ be the corresponding ciphertext vectors in a Hill n-cipher. If

$$P = \begin{bmatrix} \mathbf{p}_1^T \\ \mathbf{p}_2^T \\ \vdots \\ \mathbf{p}_n^T \end{bmatrix}$$

is the $n \times n$ matrix with row vectors $\mathbf{p}_1^T, \mathbf{p}_2^T, \ldots, \mathbf{p}_n^T$ and if

$$C = \begin{bmatrix} \mathbf{c}_1^T \\ \mathbf{c}_2^T \\ \vdots \\ \mathbf{c}_n^T \end{bmatrix}$$

is the $n \times n$ matrix with row vectors $\mathbf{c}_1^T, \mathbf{c}_2^T, \ldots, \mathbf{c}_n^T$, then the sequence of elementary row operations that reduces C to I transforms P to $\left(A^{-1}\right)^T$.

The following example illustrates a simple algorithm for doing this.

■ **Example 5.8** The following Hill 2-cipher is intercepted:

IOSBTGXESPXHOPDE

Decipher the message, given that it starts with the word *DEAR*.
Solution:
Since the numerical equivalent of the known plaintext is

D	E	A	R
4	5	1	18

and the numerical equivalent of the corresponding ciphertext is

I	O	S	B
9	15	19	2

so the corresponding plaintext and ciphertext vectors are

$$\mathbf{p}_1 = \begin{bmatrix} 4 \\ 5 \end{bmatrix} \leftrightarrow \mathbf{c}_1 = \begin{bmatrix} 9 \\ 15 \end{bmatrix}$$
$$\mathbf{p}_2 = \begin{bmatrix} 1 \\ 18 \end{bmatrix} \leftrightarrow \mathbf{c}_2 = \begin{bmatrix} 19 \\ 2 \end{bmatrix}$$

We want to reduce

$$C = \begin{bmatrix} \mathbf{c}_1^T \\ \mathbf{c}_2^T \end{bmatrix} = \begin{bmatrix} 9 & 15 \\ 19 & 2 \end{bmatrix}$$

to I by elementary row operations and simultaneously apply these operations to

$$P = \begin{bmatrix} \mathbf{p}_1^T \\ \mathbf{p}_2^T \end{bmatrix} = \begin{bmatrix} 4 & 5 \\ 1 & 18 \end{bmatrix}$$

to obtain $\left(A^{-1}\right)^T$. It is possible to do this.by adjoining P to the right of C and applying row operations to the resulting matrix $[C \mid P]$ until the left side is reduced to I. The final matrix will then have the form

$$\left[I \mid \left(A^{-1}\right)^T\right].$$

The computations can be carried out as follows:

$$\left[\begin{array}{cc|cc} 9 & 15 & 4 & 5 \\ 19 & 2 & 1 & 18 \end{array}\right] \overset{9^{-1}r_1=3r_1}{\rightarrow}$$

$$\left[\begin{array}{cc|cc} 1 & 45 & 12 & 15 \\ 19 & 2 & 1 & 18 \end{array}\right] \rightarrow$$

We replaced 45 by its residue modulo 26.

$$\left[\begin{array}{cc|cc} 1 & 19 & 12 & 15 \\ 19 & 2 & 1 & 18 \end{array}\right] \overset{-19r_1+r_2}{\rightarrow}$$

$$\left[\begin{array}{cc|cc} 1 & 19 & 12 & 15 \\ 0 & -359 & -227 & -267 \end{array}\right] \rightarrow$$

$$\left[\begin{array}{rr|rr} 1 & 19 & 12 & 15 \\ 0 & 1 & 147 & 399 \end{array}\right] \overset{5^{-1}r_2=21r_2}{\rightarrow}$$

$$\left[\begin{array}{rr|rr} 1 & 19 & 12 & 15 \\ 0 & 1 & 17 & 9 \end{array}\right] \overset{-19r_2+r_2}{\rightarrow}$$

$$\left[\begin{array}{rr|rr} 1 & 0 & -311 & -156 \\ 0 & 1 & 17 & 9 \end{array}\right] \rightarrow$$

$$\left[\begin{array}{rr|rr} 1 & 0 & 1 & 0 \\ 0 & 1 & 17 & 9 \end{array}\right].$$

Thus,

$$\left(A^{-1}\right)^T = \begin{bmatrix} 1 & 0 \\ 17 & 9 \end{bmatrix},$$

so the deciphering matrix is

$$A^{-1} = \begin{bmatrix} 1 & 17 \\ 0 & 9 \end{bmatrix},$$

To decipher the message, we first group the ciphertext into pairs and find the numerical equivalent of each letter:

IO	*SB*	*TG*	*XE*	*SP*	*XH*	*OP*	*DE*
915	192	207	245	1916	248	1516	45

Next, we multiply successive ciphertext vectors on the left by A^{-1} and find the alphabet equivalents of the resulting plaintext pairs:

$$
\begin{array}{ll}
\begin{bmatrix}1 & 17\\0 & 9\end{bmatrix}\begin{bmatrix}9\\15\end{bmatrix}=\begin{bmatrix}4\\5\end{bmatrix} & \begin{matrix}D\\E\end{matrix}\\
\begin{bmatrix}1 & 17\\0 & 9\end{bmatrix}\begin{bmatrix}19\\2\end{bmatrix}=\begin{bmatrix}1\\18\end{bmatrix} & \begin{matrix}A\\R\end{matrix}\\
\begin{bmatrix}1 & 17\\0 & 9\end{bmatrix}\begin{bmatrix}20\\7\end{bmatrix}=\begin{bmatrix}9\\11\end{bmatrix} & \begin{matrix}I\\K\end{matrix}\\
\begin{bmatrix}1 & 17\\0 & 9\end{bmatrix}\begin{bmatrix}24\\5\end{bmatrix}=\begin{bmatrix}5\\19\end{bmatrix} & \begin{matrix}E\\S\end{matrix}\\
\begin{bmatrix}1 & 17\\0 & 9\end{bmatrix}\begin{bmatrix}19\\16\end{bmatrix}=\begin{bmatrix}5\\14\end{bmatrix} & \begin{matrix}E\\N\end{matrix}\\
\begin{bmatrix}1 & 17\\0 & 9\end{bmatrix}\begin{bmatrix}24\\8\end{bmatrix}=\begin{bmatrix}4\\20\end{bmatrix} & \begin{matrix}D\\T\end{matrix}\\
\begin{bmatrix}1 & 17\\0 & 9\end{bmatrix}\begin{bmatrix}15\\16\end{bmatrix}=\begin{bmatrix}1\\14\end{bmatrix} & \begin{matrix}A\\N\end{matrix}\\
\begin{bmatrix}1 & 17\\0 & 9\end{bmatrix}\begin{bmatrix}4\\5\end{bmatrix}=\begin{bmatrix}11\\19\end{bmatrix} & \begin{matrix}K\\S\end{matrix}
\end{array}
\pmod{26}
$$

Figure 5.2:

Finally, we construct the message from the plaintext pairs:

DE AR IK ES EN DT AN KS

DEARIKE SEND TANKS

■

5.6 Exercise

1. Obtain the Hill cipher of the message

DARK NIGHT

for each of the following enciphering matrices:

(a) $\begin{bmatrix} 1 & 3 \\ 2 & 1 \end{bmatrix}$.

(b) $\begin{bmatrix} 4 & 3 \\ 1 & 2 \end{bmatrix}$.

2. In each part determine whether the matrix is invertible modulo 26. If so, find its inverse modulo 26 and check your work by verifying that

$$AA^{-1} = A^{-1}A = I(\bmod 26)$$

(a) $A = \begin{bmatrix} 9 & 1 \\ 7 & 2 \end{bmatrix}$

(b) $A = \begin{bmatrix} 3 & 1 \\ 5 & 3 \end{bmatrix}$

(c) $A = \begin{bmatrix} 8 & 11 \\ 1 & 9 \end{bmatrix}$

(d) $A = \begin{bmatrix} 2 & 1 \\ 1 & 7 \end{bmatrix}$

(e) $A = \begin{bmatrix} 3 & 1 \\ 6 & 2 \end{bmatrix}$

(f) $A = \begin{bmatrix} 1 & 8 \\ 1 & 3 \end{bmatrix}$

3. Decode the message

SAKNOXAOJX

given that it is a Hill cipher with enciphering matrix

$$\begin{bmatrix} 4 & 1 \\ 3 & 2 \end{bmatrix}$$

4. A Hill 2 -cipher is intercepted that starts with the pairs

$$SLHK$$

Find the deciphering and enciphering matrices, given that the plaintext is known to start with the word $ARMY$.
5. Decode the following Hill 2 -cipher if the last four plaintext letters are known to be $ATOM$.

LNGIHGYBVRENJYQO

Wish you all the best,
Dr. A. Elrawy